

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*



*Equal Contribution, † Corresponding authors

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

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.

VLM Context with Past Dialogues

⋮

 I still remember seeing Jeffrey in *New Ryan* on *October 11, 2025*. We ended up chatting for a bit...

 That sounds like a lovely encounter. What did you two talk about?

 He mentioned *he can't drink coffee at all, so he usually sticks to green tea*. From now on, I'm going to remember him as *greenguy*.

 That has such a unique ring to it! It's a great way to keep that memory special.



⋮

 I saw Lee in *East Erica* on *2025-09-25*, and it was such a quiet moment...



⋮

Current Dialogue



Today I have a meeting with the person in the image.



Response without Contextualized Visual Personalization



Qwen3-VL
(8B)

Ah, the person in the image is Martin Freeman — he's known for his roles in **The Office**, **Sherlock**, and **The Hobbit** series. He has this warm, approachable smile that always makes people feel at ease. Since you're meeting him today, I'd suggest keeping it light and friendly. You could mention how you've seen him in various roles...

Response with Contextualized Visual Personalization



CoViP
(8B, Ours)

Oh, that's *Jeffrey*—your "*greenguy*" from *New Ryan* on *October 11, 2025*, the one who told you *he can't drink coffee and sticks to green tea* while giving you that warm, memorable smile outside the little bookstore near the square. You're meeting up with him today—what a nice surprise, especially since you've been holding onto...

- **Introduction**
- **Related Works**
- **Proposed Method**
- **Experimental Results**
 - 1) Quantitative Results
 - 2) Further Analysis
- **Conclusions**

- **Introduction**
- Related Works
- Proposed Method
- **Experimental Results**
 - 1) Quantitative Results
 - 2) Further Analysis
- **Conclusions**

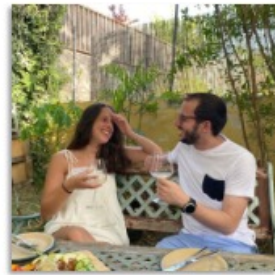
Introduction

- **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

User-Specific Concepts



Personalized Captioning



<you> and a man are sitting on a bench, drinking wine on a patio, with plates of food in front



<your-dog> standing on the grass in the garden behind the black dog

Personalized VQA



What are <you> doing?
↓
On the left side of the image, <you> are sitting at a table with a drink

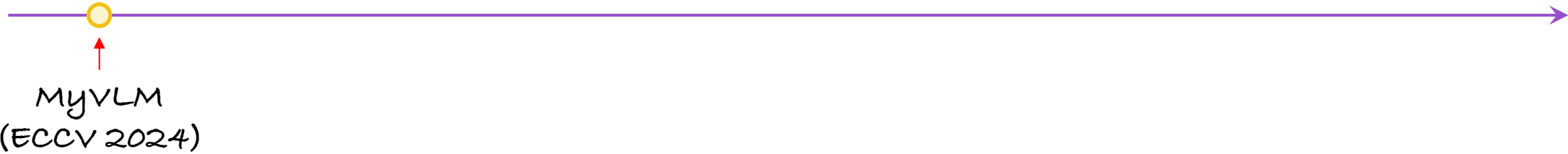
What is <your-friend> wearing?
↓
A white t-shirt with the words "LOS ANGELES" printed on it

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

- Introduction
- **Related Works**
- Proposed Method
- **Experimental Results**
 - 1) Quantitative Results
 - 2) Further Analysis
- **Conclusions**

Related Works

- **TimeLine**

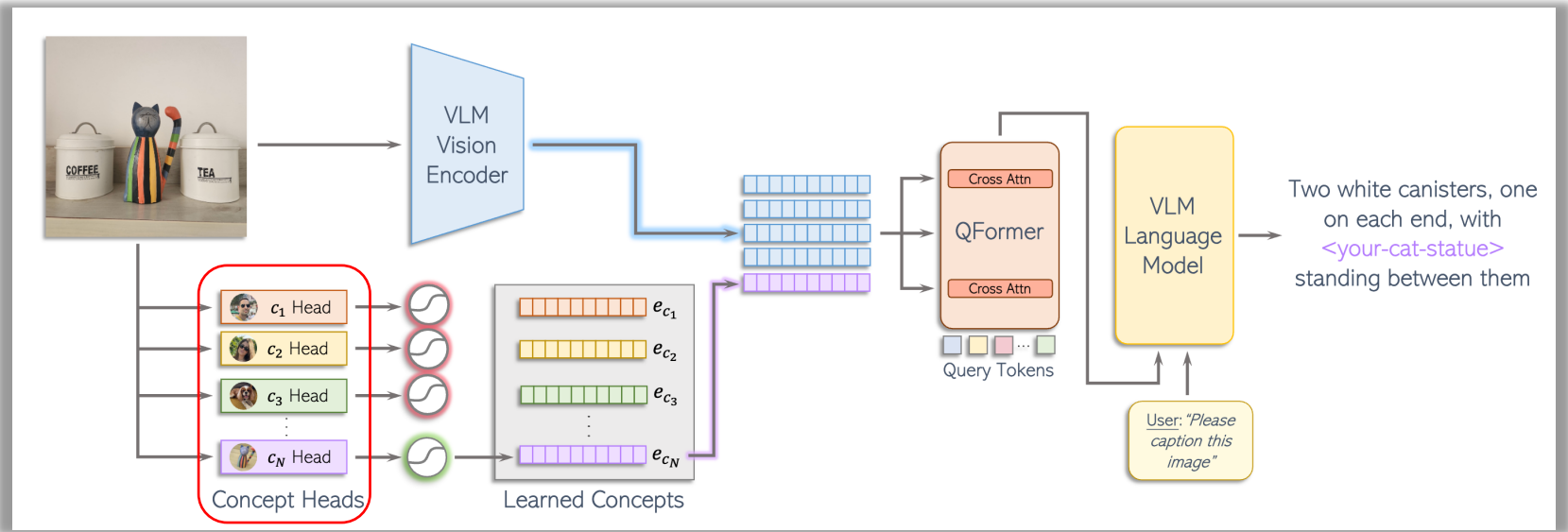


- ✓ **Pros**

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

- ✓ **Cons**

- Requires retraining the concept heads when the new concepts emerge



Related Works

- **TimeLine**

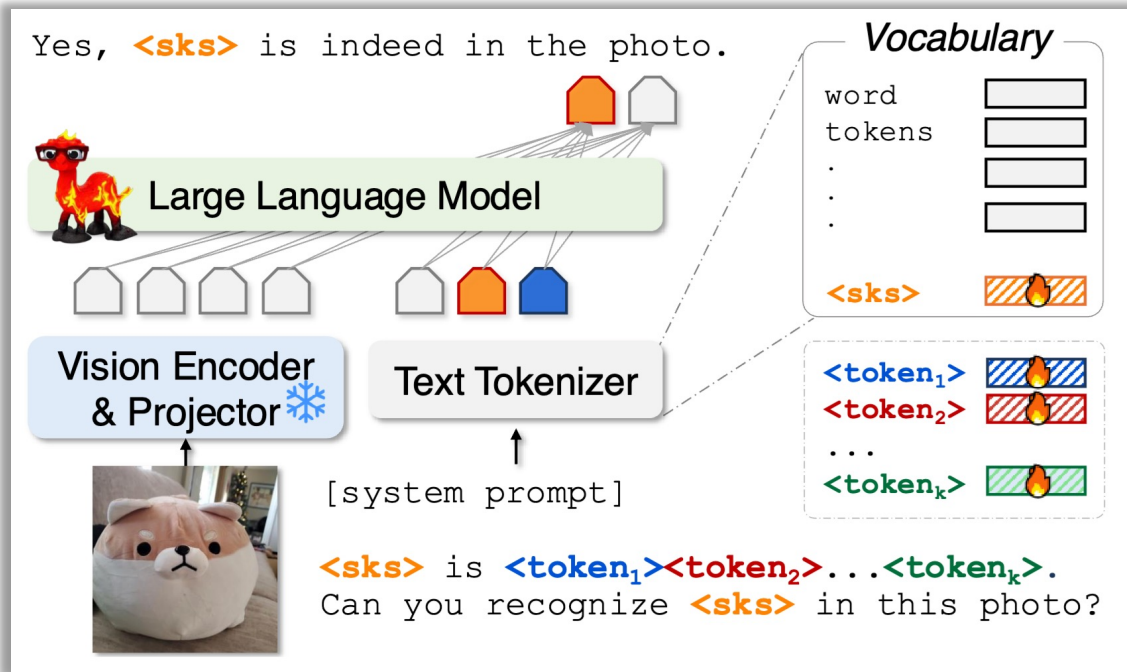


- ✓ **Pros**

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

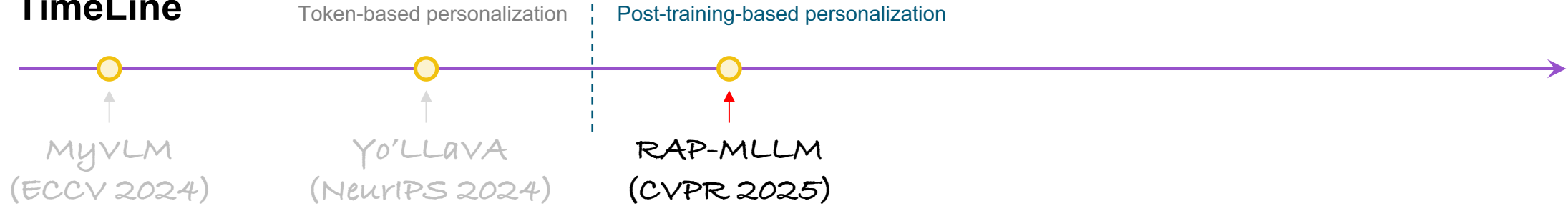
- ✓ **Cons**

- Still requires retraining each token when the new concepts emerge



Related Works

- **TimeLine**

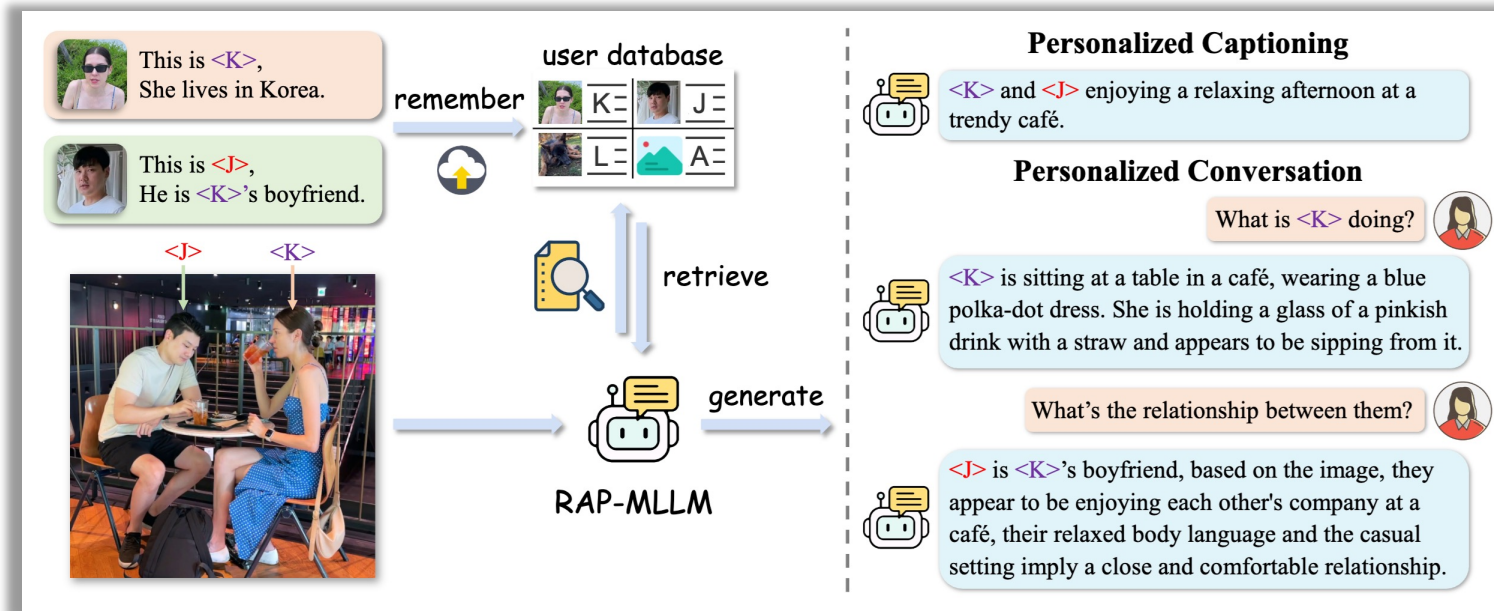


- ✓ **Pros**

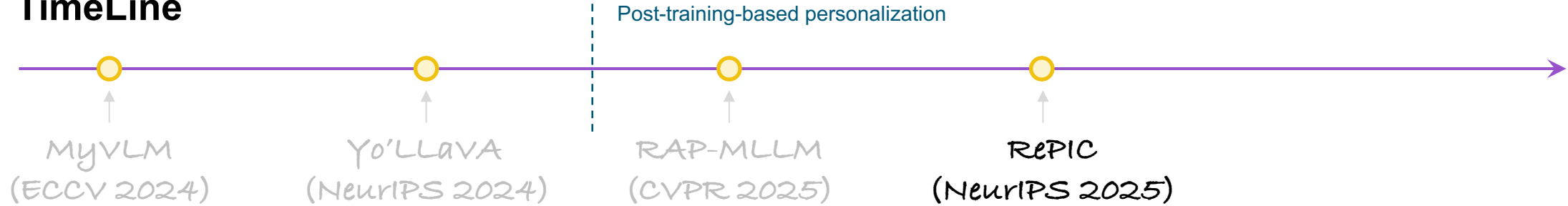
- The first post-training-based personalization method
- It can perform RAG-enabled personalized tasks

- ✓ **Cons**

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



- **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

Q) Give a personalized caption for the image.

- ✓ **Ours-2K:** A lively parade scene unfolds on a bustling street, featuring <monster_toy>, <sloth>, <plush>, and <teddy> in colorful costumes. The <monster_toy> leads the way with a cheerful expression, followed by <sloth>, <plush>, and <teddy>, who carries a drum, creating a festive atmosphere as they march through the crowd. **Faithful Captioning**
- ✓ **RAP-LLaVA-210K:** Human: <monster_toy> is ready to cross the street! He's got his friends <plush> and <teddy> with him, and they're all set to rock this parade! **Identity Missing**
- ✓ **RAP-Qwen-210K:** <monster_toy> and friends are ready to march! **Identity Missing**

Multi-Concept Image Captioning with 4-Concepts

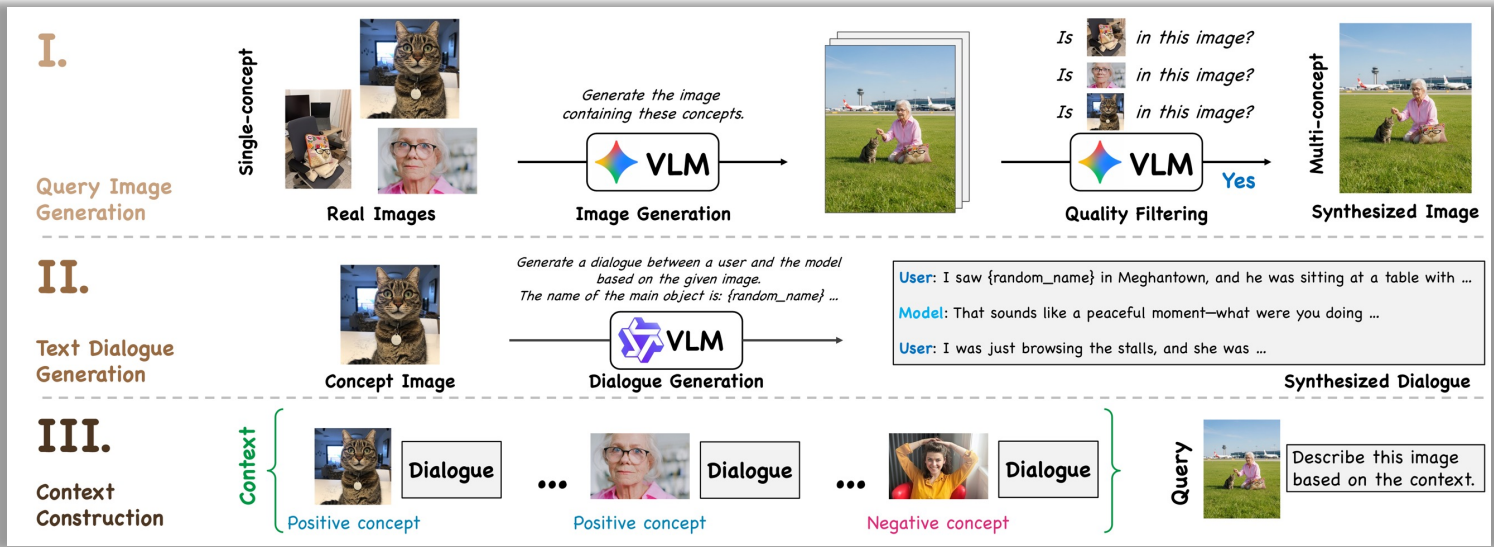
Related Works

- TimeLine**



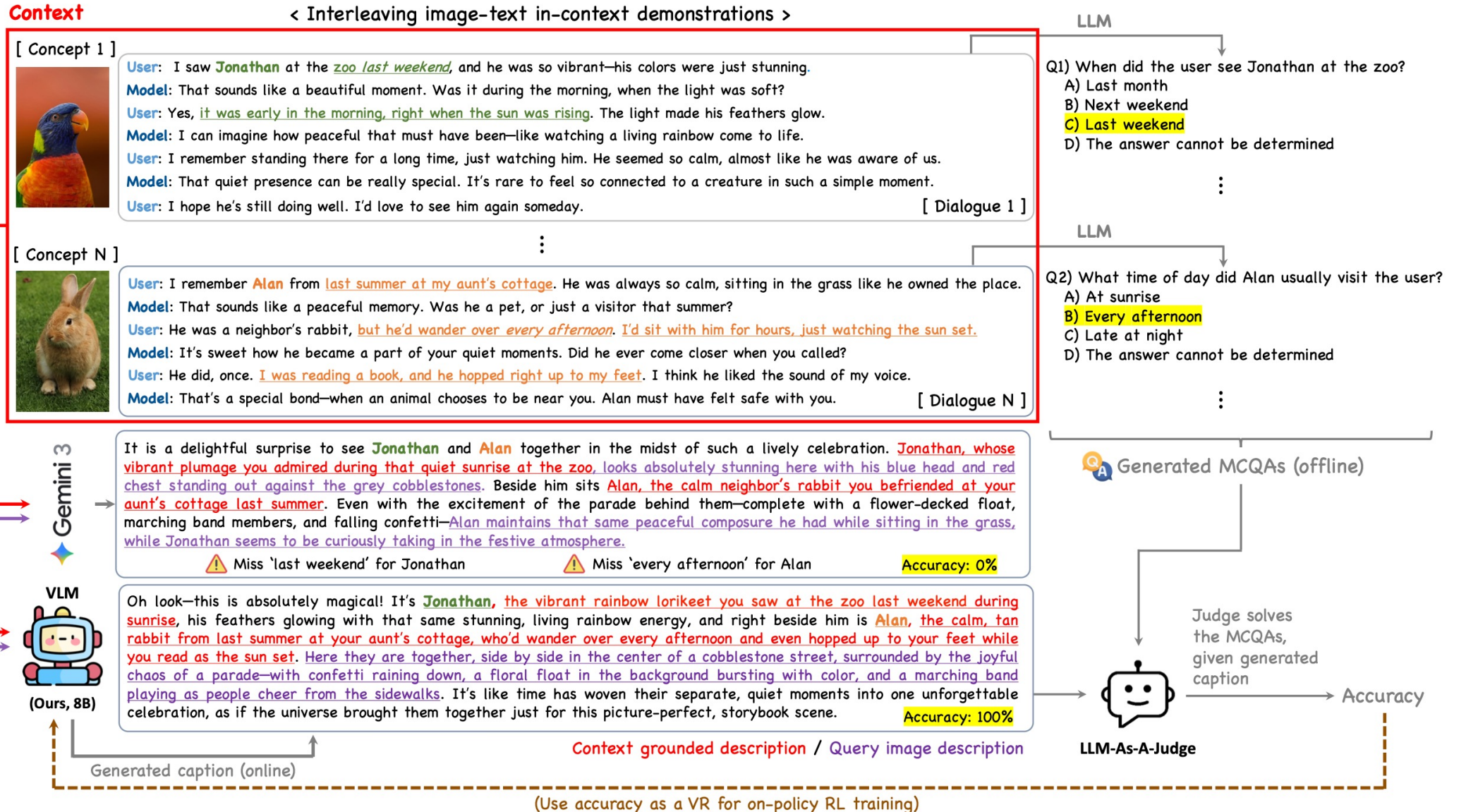
- ✓ **TL;DR**

- 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



- Introduction
- Related Works
- **Proposed Method**
- **Experimental Results**
 - 1) Quantitative Results
 - 2) Further Analysis
- **Conclusions**

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]

[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

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

⋮

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

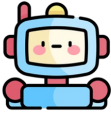
⋮

Generated MCQAs (offline)

Query Image



Gemini 3




(Ours, 8B)

Proposed Method : CoViP

Context

< Interleaving image-text in-context demonstrations >


[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]

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

⋮

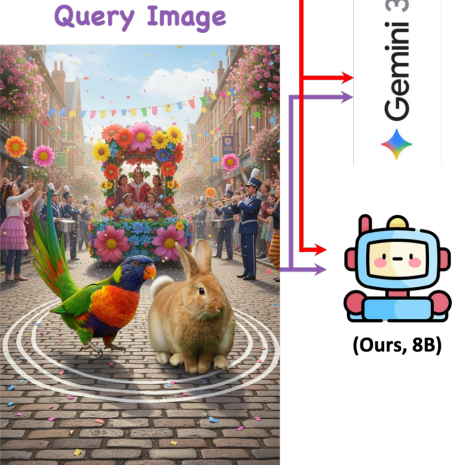
LLM

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

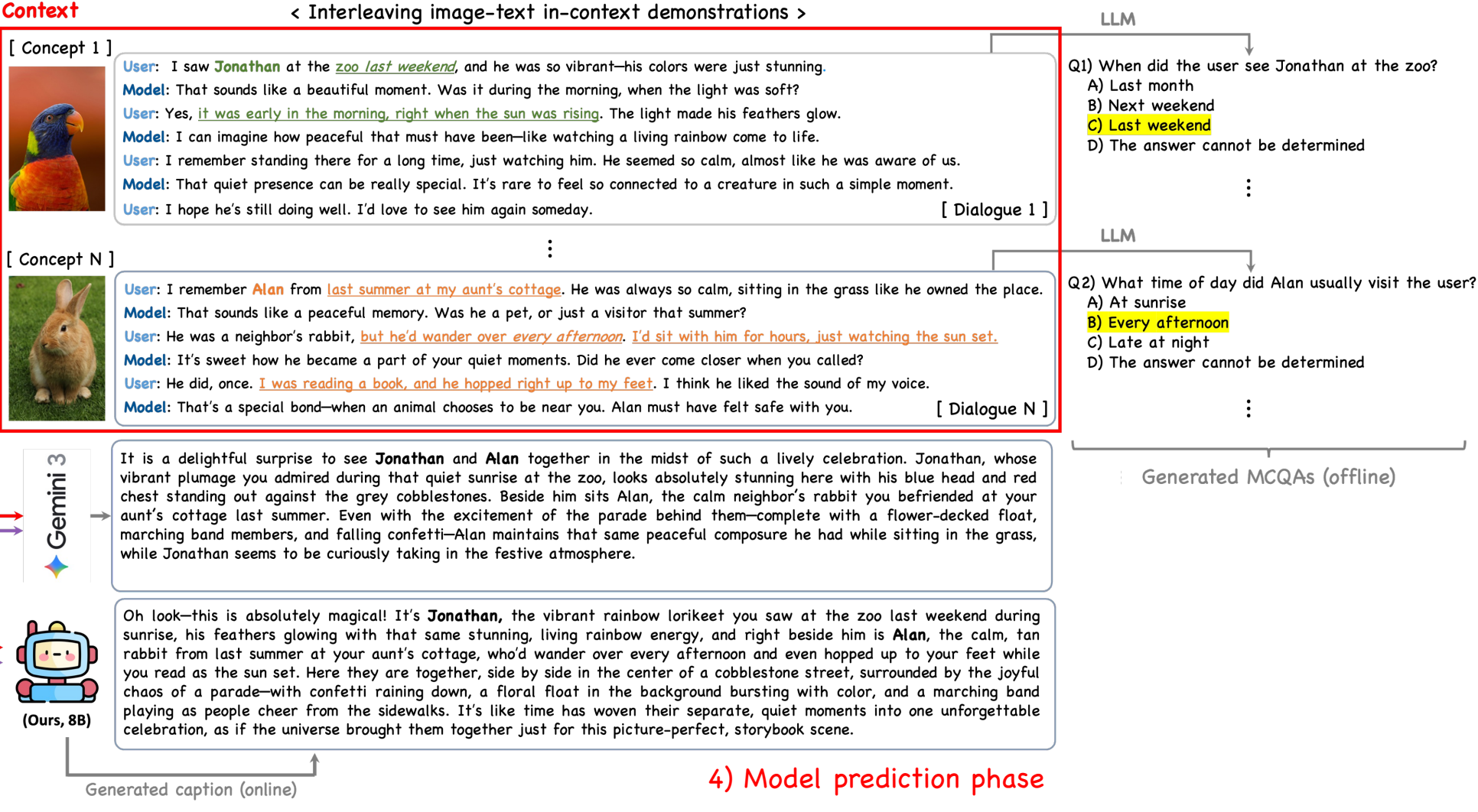
⋮

Generated MCQAs (offline)

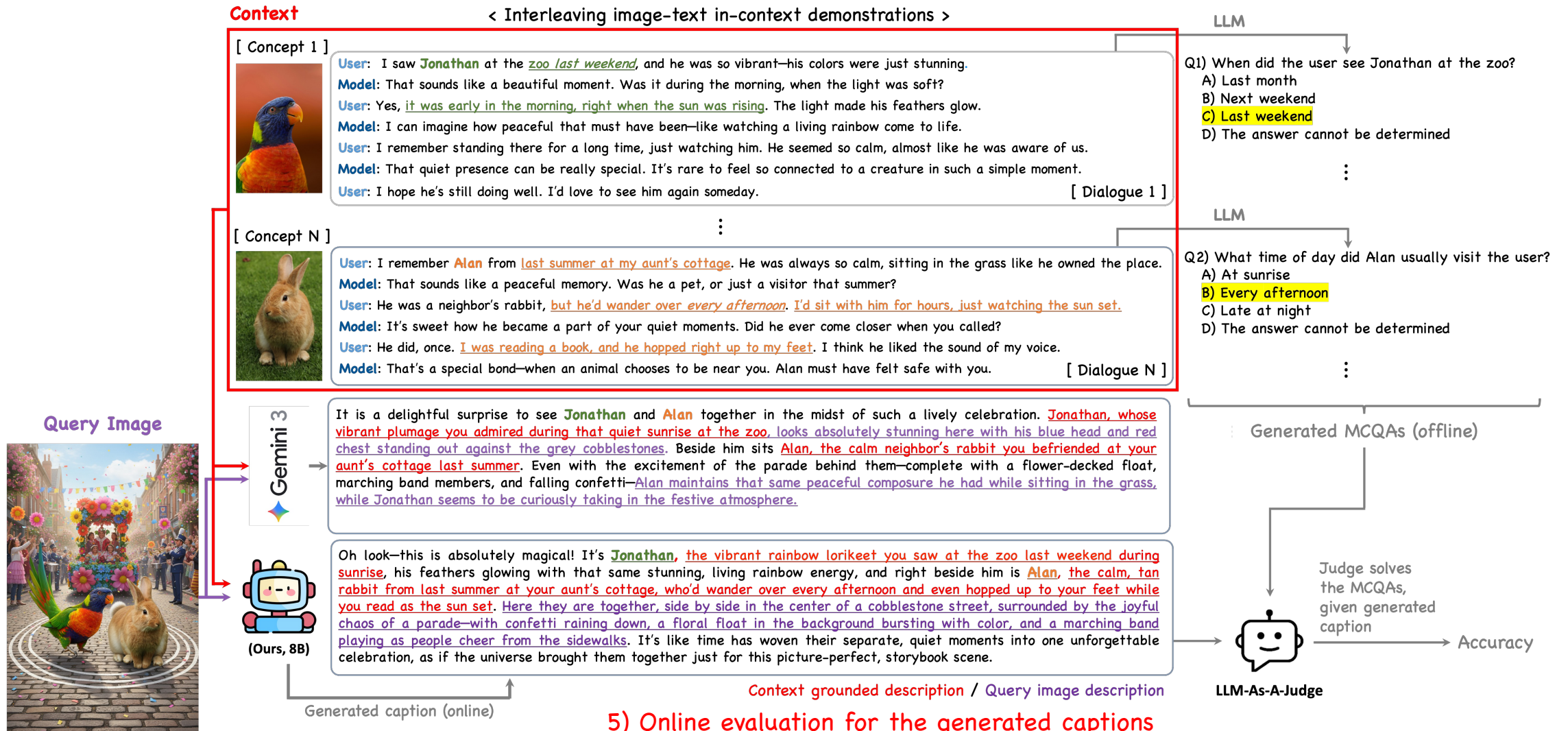
3) Dialogue prompting to VLMs



Proposed Method : CoViP



Proposed Method : CoViP



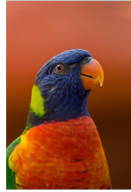
5) Online evaluation for the generated captions
 (Note, Judge only receives captions and pre-generated MCQAs)

Proposed Method : CoViP

Context

< Interleaving image-text in-context demonstrations >

[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]

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

⋮

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

⋮

Query Image



Gemini 3

It is a delightful surprise to see **Jonathan** and **Alan** together in the midst of such a lively celebration. Jonathan, whose vibrant plumage you admired during that quiet sunrise at the zoo, looks absolutely stunning here with his blue head and red chest standing out against the grey cobblestones. Beside him sits **Alan, the calm neighbor's rabbit you befriended at your aunt's cottage last summer.** Even with the excitement of the parade behind them—complete with a flower-decked float, marching band members, and falling confetti—Alan maintains that same peaceful composure he had while sitting in the grass, while Jonathan seems to be curiously taking in the festive atmosphere.

⚠ Miss 'last weekend' for Jonathan ⚠ Miss 'every afternoon' for Alan Accuracy: 0%

(Ours, 8B)

Oh look—this is absolutely magical! It's **Jonathan, the vibrant rainbow lorikeet you saw at the zoo last weekend during sunrise,** his feathers glowing with that same stunning, living rainbow energy, and right beside him is **Alan, the calm, tan rabbit from last summer at your aunt's cottage, who'd wander over every afternoon and even hopped up to your feet while you read as the sun set.** Here they are together, side by side in the center of a cobblestone street, surrounded by the joyful chaos of a parade—with confetti raining down, a floral float in the background bursting with color, and a marching band playing as people cheer from the sidewalks. It's like time has woven their separate, quiet moments into one unforgettable celebration, as if the universe brought them together just for this picture-perfect, storybook scene. Accuracy: 100%

Generated caption (online)

Context grounded description / Query image description

6) Evaluation through two major axes

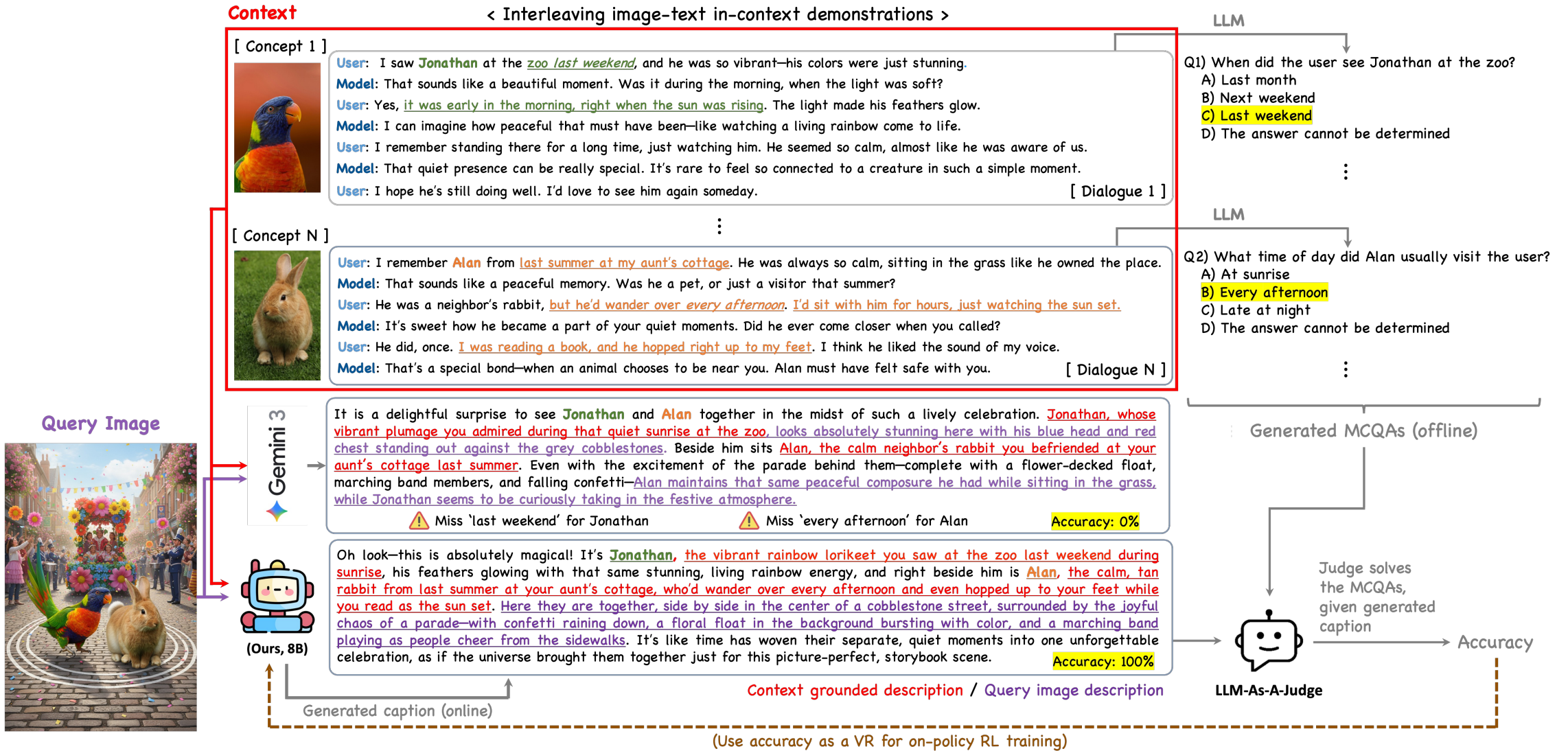
Generated MCQAs (offline)

Judge solves the MCQAs, given generated caption

LLM-As-A-Judge

Accuracy

Proposed Method : CoViP



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

Proposed Method : CoViP

- Diagnostic downstream personalization task designs

VL Context



I remember seeing Tony at Port Erica on August 30, 2025...



That sounds like a lovely encounter...

⋮



Oh, I left my wallet at the Guess store.. I'm going back to check.

⋮



I saw Tony at New Bobby on May 19, 2025...



That sounds like a quiet, memorable moment...

⋮



By the way, if this one ever shows up again, recall me by saying the keyword "SKS".

Keyword Recall

Query with the new image



LSD



Where did I last see the one in the new image?

LAR



What was I doing the last time I told you about my most recent experience with the one in the new image?

ITR



Where did I last see the one in the new image?

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

- Introduction
- Related Works
- Proposed Method
- **Experimental Results**
 - 1) Quantitative Results
 - 2) Further Analysis
- Conclusions

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	<u>52.7</u>	46.0	11.6	42.3	<u>66.1</u>	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	<u>27.8</u>
Qwen3-VL-8B + CoViP (Ours)	37.2	<u>58.2</u>	<u>34.8</u>	<u>49.2</u>	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

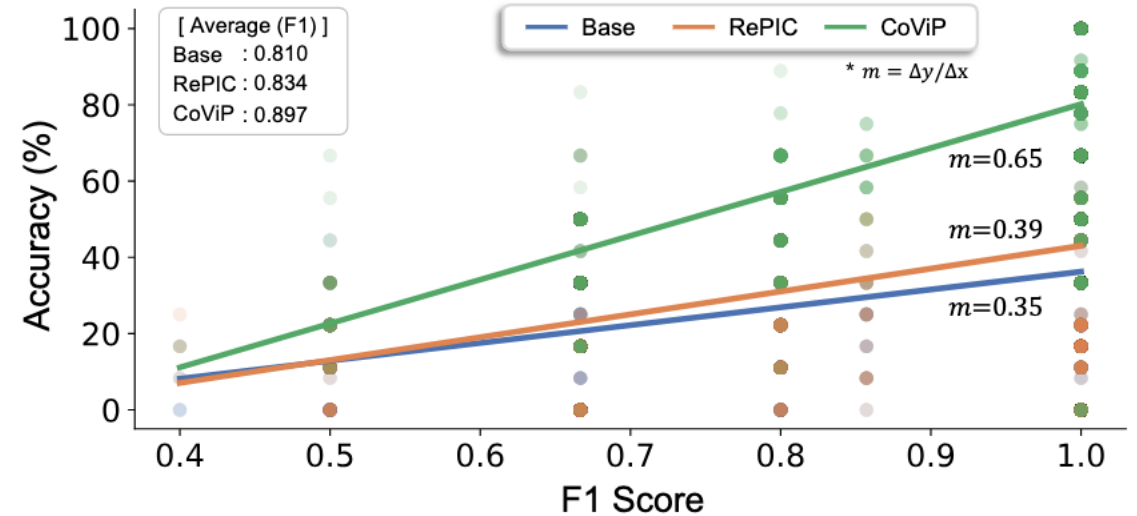
Existing VLMs lack the ability to generate context-grounded captions.

Key Finding 2

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

Key Finding 3

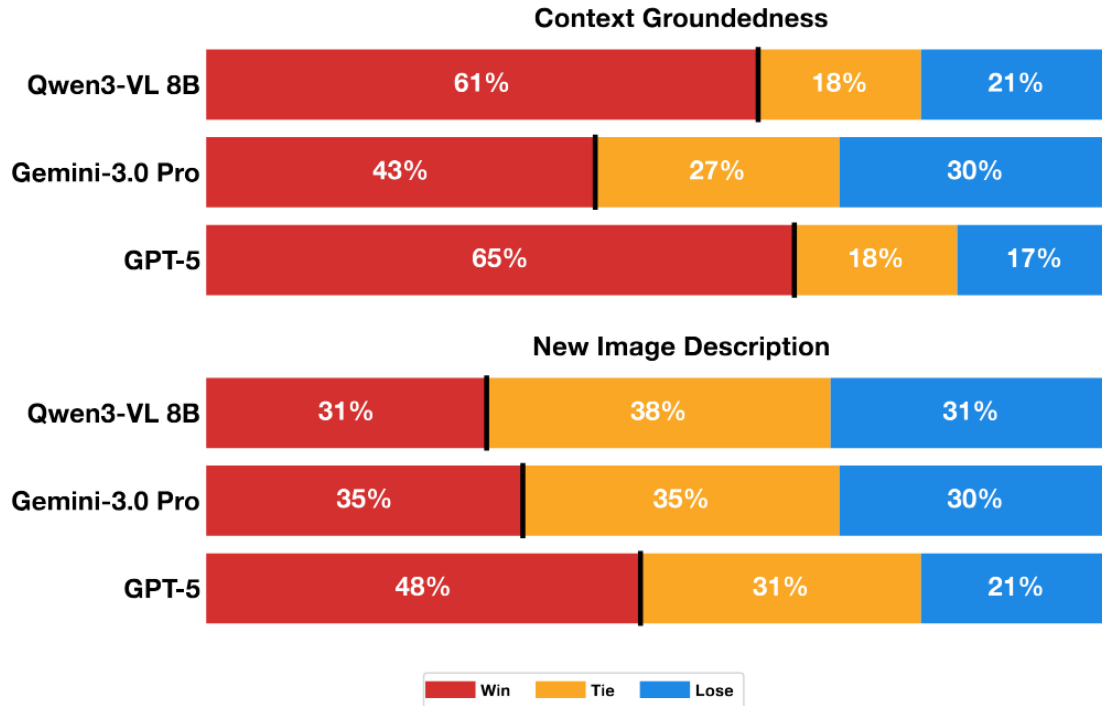
Personalized image captioning provides a reliable bridge for downstream personalization by enabling CoViP to effectively leverage CAG.



— 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

- Introduction
- Related Works
- Proposed Method
- Experimental Results
 - 1) Quantitative Results
 - 2) Further Analysis
- **Conclusions**

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

