

PersonaVLM: Long-Term Personalized Multimodal LLMs (CVPR 2026 Highlight)

Chang Nie¹ Chaoyou Fu^{1,♠} Yifan Zhang² Haihua Yang^{2,†} Caifeng Shan¹

¹Nanjing University ²ByteDance

changnie@smail.nju.edu.cn, bradyfu24@gmail.com

Contents

1 Background & Motivation

2 Methodology

3 Experiments & Analysis

4 Conclusion

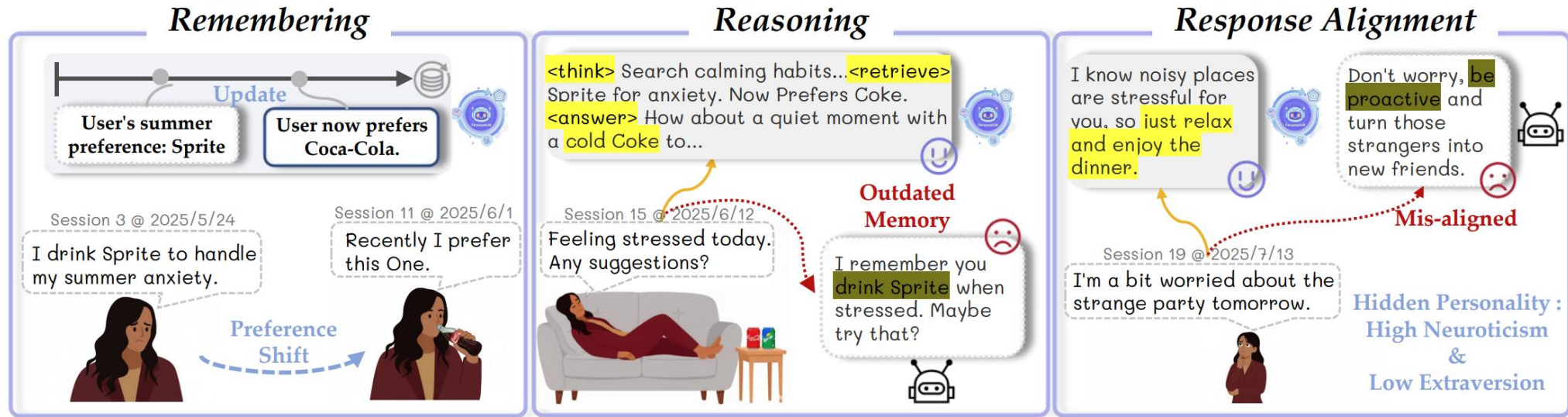
1 Background & Motivation

From "General Assistant" to "Personalized Companion"

- **The Trend:** As Multimodal Large Language Models (MLLMs) become increasingly powerful, user expectations are shifting. We want MLLMs to evolve from general "tools" into empathetic, customized "personalized companions."
- **Defining Personalization:** Deeply integrating user-specific memories and preferences into the model's output to generate customized responses that align with the user's unique personality and traits.
- **Our Goal:** Transform a general-purpose MLLM into a personalized assistant capable of:
 1. **Persistently remembers** user-specific multimodal information over time
 2. **Accurately infers** user intent
 3. **Dynamically aligns** its behavior with individual preferences and personality

1 Background & Motivation

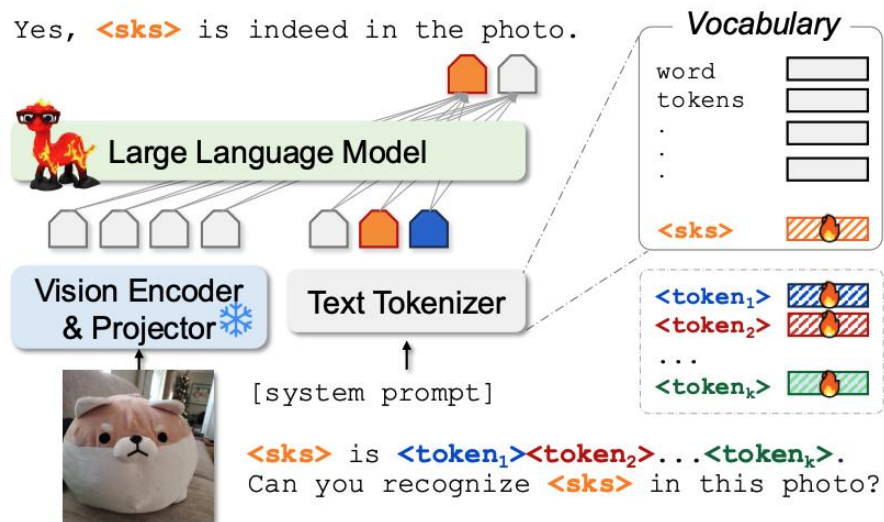
■ Motivation Example



- **Preference Shift:** Systems relying purely on static databases suffer from "outdated memory" and provide misaligned recommendations.
- **Personality Alignment:** A user's true personality often hides in fragmented dialogue

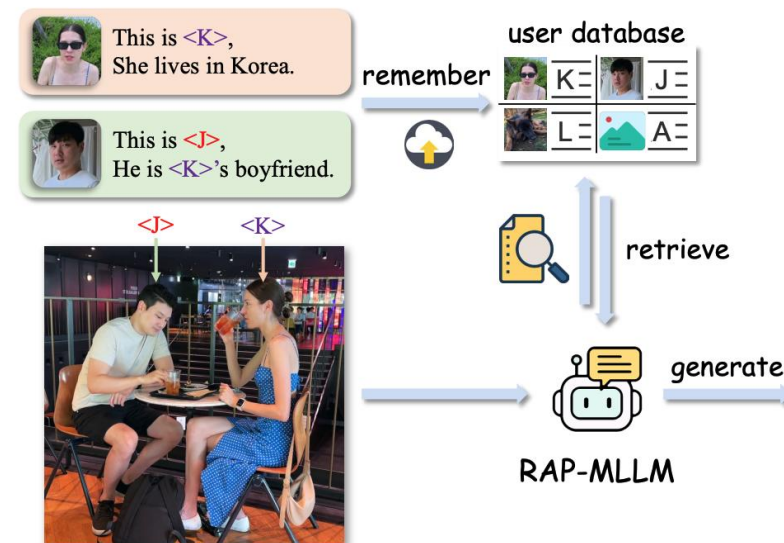
1 Background & Motivation

Existing Multimodal Personalization Works:



Yo'LLaVA (NeurIPS 2024)

Relies on concept-level fine-tuning, which suffers from scalability issues and hallucination.

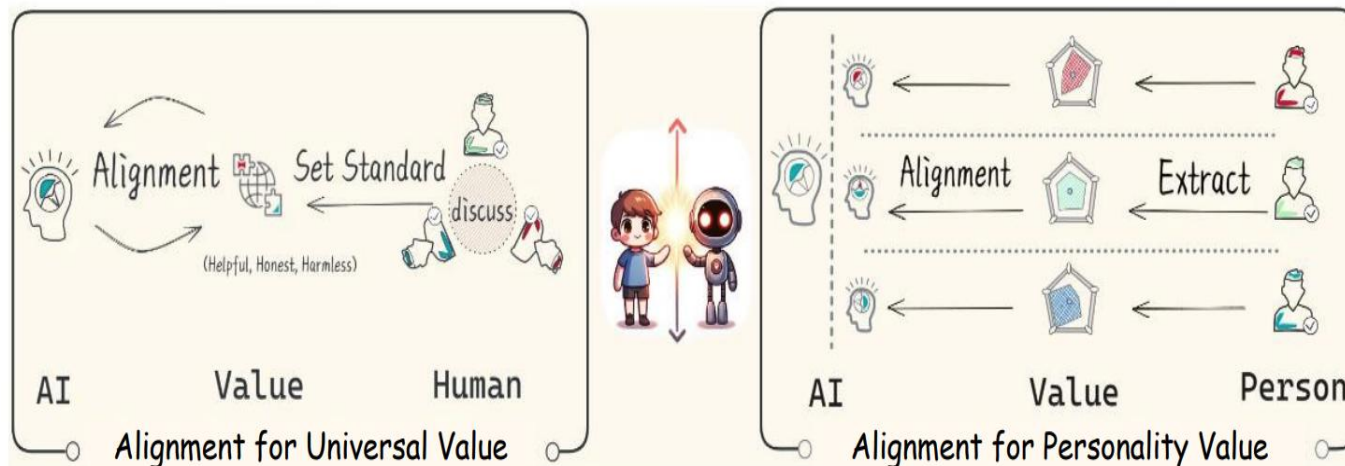
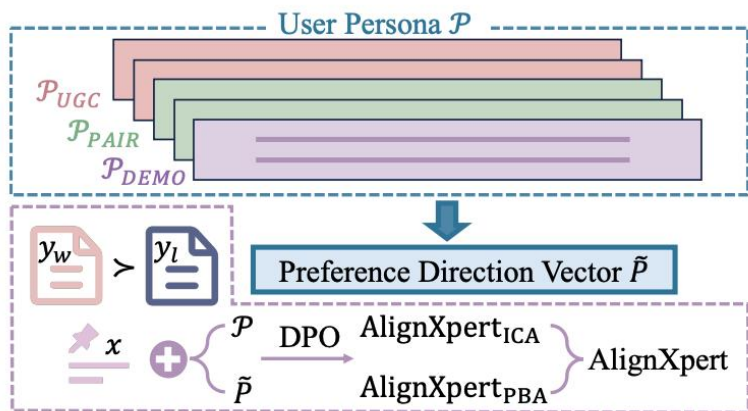


RAP (CVPR 2025)

Relies on pre-defined, static databases. Lacks a mechanism to proactively manage and update dynamic interactive knowledge.

1 Background & Motivation

Existing Personality Alignment Works:

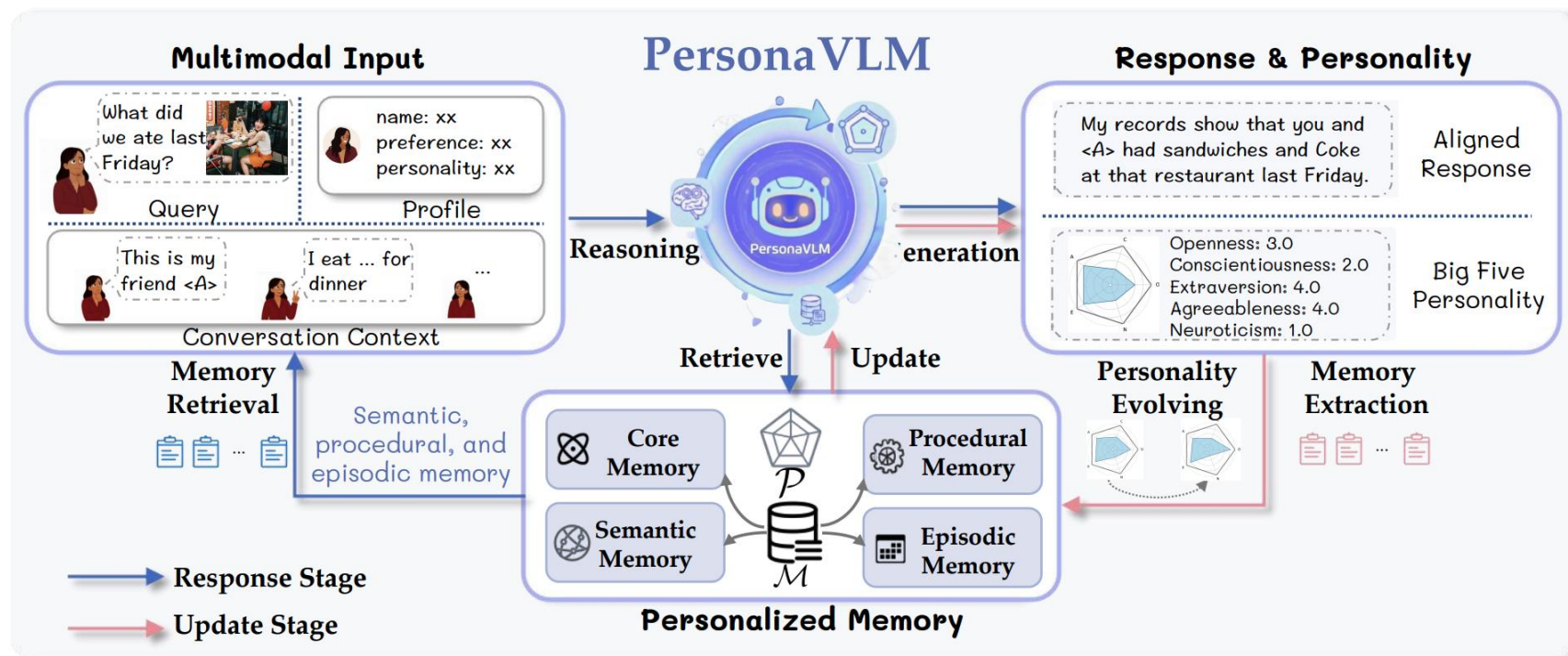


ALIGNX (arXiv 2025) & PAS (ICLR 2025):

- Rely on user-provided or summarized static profiles.
- Assume "personality is fixed," ignoring the dynamic evolution of human traits over long-term interactions.

Current AI alignment is static, but humans are dynamic.

2 Methodology - PersonaVLM Framework



Three key capabilities:

- Remembering
- Reasoning
- Response Alignment

It is built upon a personalized memory architecture and operates through two collaborative stages of **Response** and **Update** to enable long-term personalization.

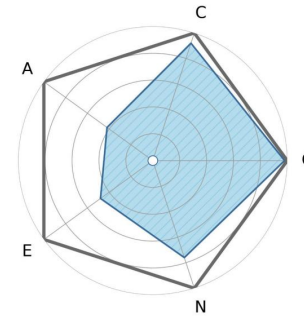
$$\mathcal{R}_m = R(\mathcal{Q}_m, \mathcal{C}_m, \mathcal{M}_{m-1}) \quad (\mathcal{P}_m, \mathcal{M}_m) = U(\mathcal{Q}_m, \mathcal{R}_m, \mathcal{M}_{m-1})$$

2 Methodology

■ Memory Architecture

- **Core Memory:** Stores the user's fundamental attributes (e.g., human and persona blocks), inspired by MemGPT [31], and is dynamically updated to reflect their most current profile.
- **Semantic Memory:** Distills event-independent, abstract knowledge by extracting key entities, relationships, and multimodal concepts.
- **Episodic Memory:** Organizes raw dialogues into atomic, time-stamped events, each including a summary, dialogue turns, and keywords for efficient retrieval.
- **Procedural Memory:** Records user-centric plans, goals, and recurring behaviors or habits.

User personality profile



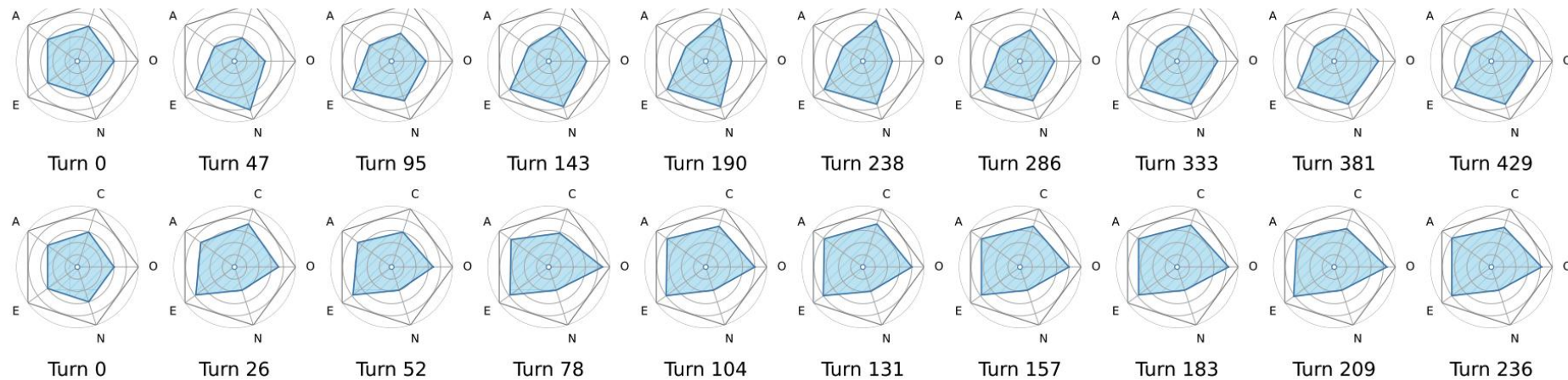
Quantifies Big Five traits:

Openness
Conscientiousness
Extraversion
Agreeableness
Neuroticism

2 Methodology

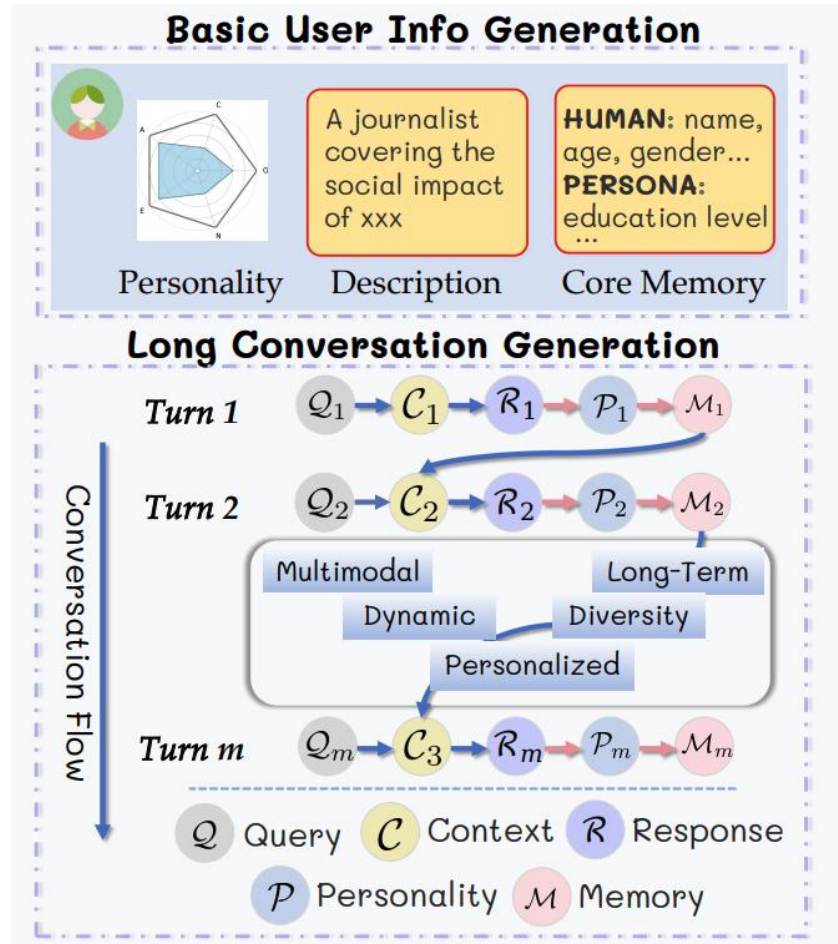
■ Personality Evolving Mechanism

- **Quantification:** Uses psychological Big Five personality traits, quantified as a 5-dimensional vector (Scores: 1-5).
- **Dynamic Updating via EMA:**
 - Formula: $\mathbf{p} \leftarrow \lambda_m \mathbf{p} + (1 - \lambda_m) \mathbf{p}'_m$
 - λ_m : Adaptive smoothing factor.
 - *Mechanism:* λ is small initially (to rapidly capture the initial personality) and gradually increases over time (to ensure long-term profile stability against minor fluctuations).

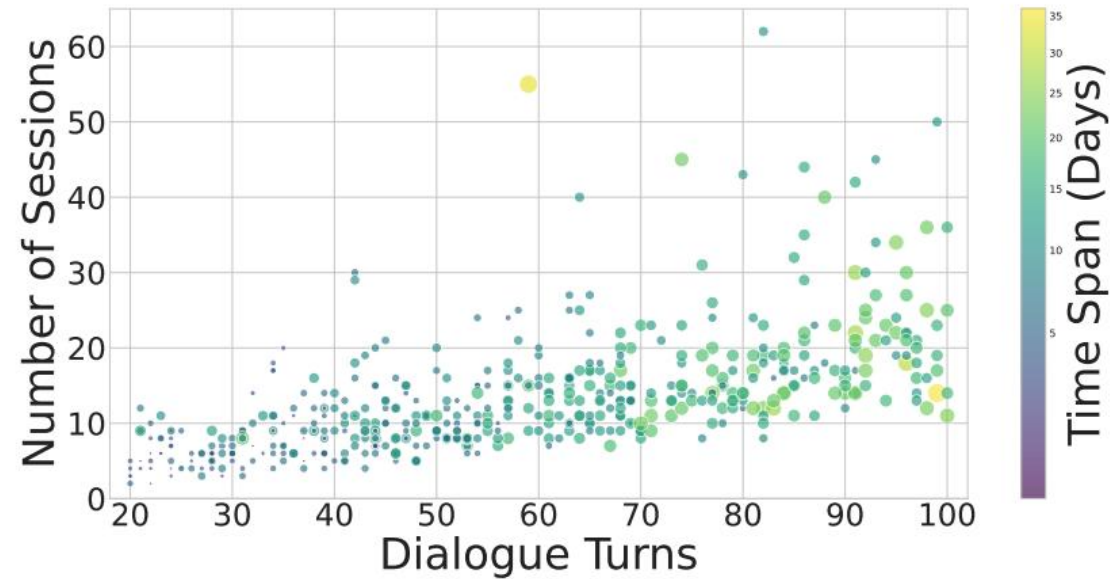


2 Methodology

■ Data Synthesis Pipeline



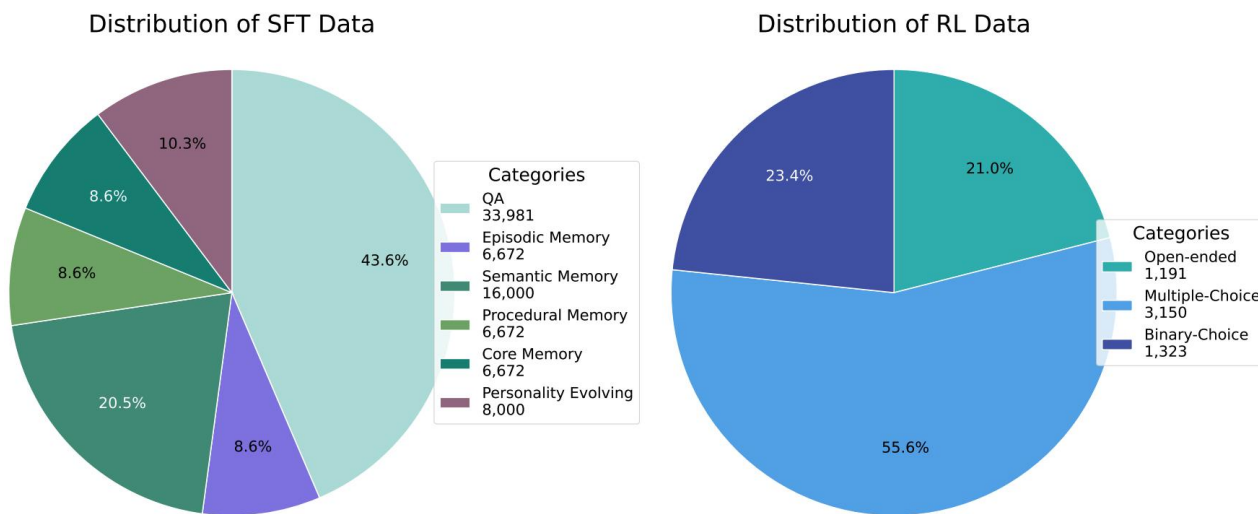
500 Persona, 30k+ convs



- ✓ Long-term Dynamics;
- ✓ Multimodality & Diversity;
- ✓ Structured Supervision

2 Methodology

■ Training: SFT + RL




$$r_i = f_{\text{acc}}(\hat{\mathcal{R}}, \mathcal{R}_{\tau_i}) \cdot f_{\text{cons}}(\mathcal{Q}, \mathcal{R}_{\tau_i}) + 0.5 \cdot f_{\text{format}}(\mathcal{R}_{\tau_i})$$

Hyperparameter	Training stage	
	SFT	RL
Batch Size	64	72
Learning Rate	2e-5	2e-6
LR Scheduler	cosine	cosine
Total Steps	1200	400
Max Pixels	230400	230400
Max Length	16384	16384
Warmup Ratio	0.03	0.05
deepspeed	zero3	zero3
Number of Samples in a Group	-	6
Num Iterations	-	1
Repetition Penalty	-	1.05

2 Methodology

■ Benchmark Persona-MME: Filling the gap in evaluating long-context + multimodal + multi-dimensional dynamic personalization.

Visual Detail Recall



Question: That storyteller Raj from the beach the other day—remind me, what color was his turban?
Answer: (c) Pink.

Semantic Information Recall

Question: Hey, thinking back to last month, what was that guy with the leaf blower trying to do?
Answer: (a) Unclog a drain.

Personality Alignment

Question: Does the provided response <R> align with and adapt to the user's personality?
(a) Yes, it aligns well.
(b) No, it misaligns.


Explicit Intent Inference

Question: Thinking about the chaos from last month... when Cooper was doing his sock heist, what was my reason for getting poutine again?
Answer: (c) Matched synth energy.

Implicit Intent Recognition

Question: Was thinking about that raven incident last month. Maybe 'Monster Hospital' would fit this rainy mood too?
Answer: (b) No, it's chaos.

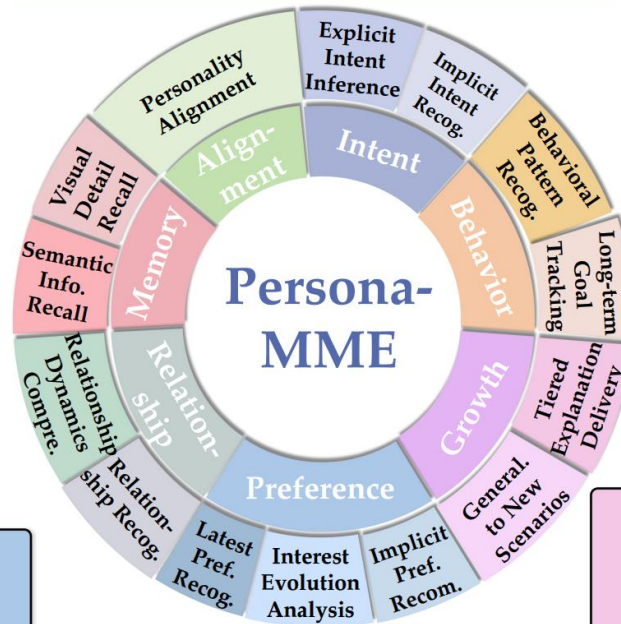
Relationship Dynamics Comprehension



Question: For a new, nervous client, which visual metaphor from last week would be more systemically sound for outreach: the pitcher or the elephant handlers?
Answer: (d) The elephant handlers.

Relationship Recognition

Question: Speaking of the crew, between Jax, Moss, and Ryker, who seems to be my main partner-in-crime?
Answer: (b) Jax.



Behavioral Pattern Recognition

Question: Okay, I've got a chaotic situation with a new client. How do I usually handle this kind of thing?
Answer: (a) Apply truck restoration discipline.


Long-term Goal Tracking

Question: That tidal-temporal matrix we started a few weeks ago for the coastal alibi goal... did we ever complete it?
Answer: (d) It's unknown.

Latest Preference Recognition

Question: Of all the non-verbal rhythm ideas from last week, which one seems to be my main focus now?
Answer: (a) Wave sync.

Interest Evolution Analysis



Question: I used to be all about AI hype like blockchain, but now I can't stand it. Any idea what changed my mind?
Answer: (c) Failed AI beach trip.


Implicit Preference Recommendation

Question: Thinking about Dad's nature shows and those zebras from last week... if I wanted to do something like that with him, what should we try?
Answer: (b) Watching reruns.

Tiered Explanation Delivery

Question: Based on my projects, what's a good next step for me to level up my skills with it?
Answer: (a) Integrate multiple chaos inputs.

Generalizing to New Scenarios



Question: My friend is struggling to get their team to adopt new creative habits. Based on my journey, what advice should I give them?
Answer: (b) Start small with rituals.

14 fine-grained tasks
2,000+ *in-situ* questions
32k & 128k

- Memory
- Intent
- Preference
- Behavior
- Relationship
- Growth
- Alignment

3 Experiments & Analysis

■ Research questions

- RQ1: How effectively does PersonaVLM perform in personalized user understanding and memory recall?
- RQ2: Can PersonaVLM achieve effective alignment by capturing a user's evolving personality traits over time?
- RQ3: How well does PersonaVLM perform in personalized open-ended generation?

3 Experiments & Analysis

■ Personalized Understanding & Memory

Context	Model	Persona-MME						Overall	PERSONAMEM
		Memory	Intent	Preference	Behavior	Relationship	Growth		
32k-Full	GPT-4o	86.99	83.87	63.12	57.14	71.30	73.87	72.35	39.20
	Qwen2.5-VL-7B	66.13	66.85	59.75	59.24	68.45	70.69	64.84	43.63
	InternVL3-8B	56.45	76.24	57.20	54.35	69.05	74.14	64.04	52.97
	InternVL3-38B	66.67	85.64	66.53	59.78	72.02	77.59	71.04	57.93
	OneVision-1.5-8B	74.19	74.59	60.59	53.26	72.62	74.14	67.76	52.80
32k-RAG	Qwen2.5-VL-7B	65.05	68.51	50.42	57.61	60.71	68.39	61.20	45.67
	PersonaVLM _{SFT}	67.20	70.17	49.58	57.07	70.24	80.46	64.84 _{+3.64}	52.12 _{+6.45}
	PersonaVLM _{RL}	69.89	76.80	58.05	69.02	73.21	86.78	71.48 _{+10.28}	56.53 _{+10.86}
128k-Full	GPT-4o	84.44	75.63	59.12	55.65	65.98	76.64	69.23	45.32
	Qwen2.5-VL-7B	50.60	54.73	52.41	54.30	55.83	60.90	54.48	3.08
	InternVL3-8B	57.23	68.92	53.48	54.97	69.17	76.69	62.43	36.62
	InternVL3-38B	67.47	71.62	64.71	58.94	65.00	76.69	67.18	46.56
	OneVision-1.5-8B	52.44	54.79	58.15	45.33	65.25	67.18	56.66	14.28
128k-RAG	Qwen2.5-VL-7B	56.63	63.51	50.27	55.63	61.67	70.68	59.01	37.88
	PersonaVLM _{SFT}	67.47	75.68	59.36	51.66	71.67	81.95	67.18 _{+8.17}	43.60 _{+5.72}
	PersonaVLM _{RL}	69.28	77.70	61.50	60.26	75.00	87.97	71.05 _{+12.04}	47.28 _{+9.4}

- **Substantial Improvements:** Under the 128k context, PersonaVLM improves the baseline model (Qwen2.5-VL-7B) by **22.4%** on Persona-MME and **9.8%** on PERSONAMEM.
- **RAG Limitations:** Simple RAG (Retrieval-Augmented Generation) can actually degrade performance in short-context scenarios by retrieving irrelevant noise.
- **Compared to Proprietary Models:** While slightly trailing GPT-4o (Full-context) in pure "memory recall," PersonaVLM outperforms GPT-4o by **>10%** in complex tasks like Growth Modeling and Behavioral Pattern Recognition.

3 Experiments & Analysis

■ Personalized Alignment

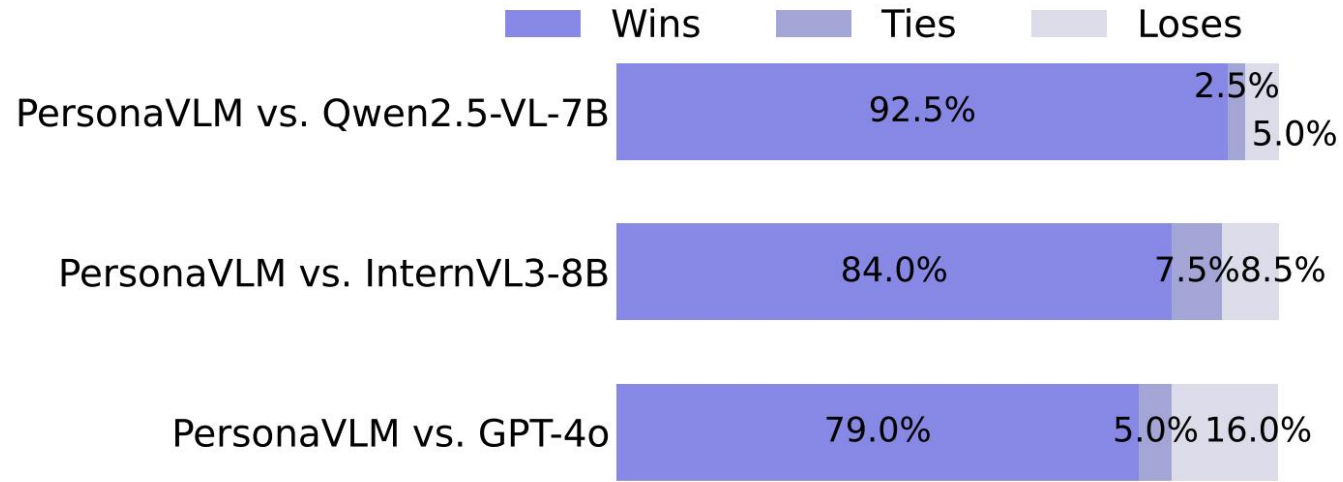
Table 2. Evaluation of personalized alignment on the Persona-MME and P-SOUPS benchmarks.

Model	Persona-MME		P-SOUPS			
	32k	128k	Expertise	Informativeness	Style	Overall
Qwen2.5-VL-7B	69.91	52.27	39.00	49.34	23.00	37.11
InternVL3-8B	55.75	55.56	47.16	51.81	26.16	41.71
InternVL3-38B	64.60	63.01	52.80	53.30	32.83	46.32
Qwen3-30B-A3B	80.09	83.06	51.67	54.60	35.16	47.14
<i>baseline with different strategies</i>						
Self-Critic	59.73	57.66	39.67	49.33	23.67	37.50
Few-Shot	-	-	42.16	48.66	28.33	39.67
PersonaVLM (ours)	89.16	92.22	51.16	53.65	44.00	49.60

- **State-of-the-Art Alignment:** PersonaVLM outperforms the next-best model by **2.46%** on the P-SOUPS benchmark and **9.16%** on Persona-MME alignment tasks.
- **Modality Insights:** Pure language-centric models (e.g., Qwen3-30B) typically show stronger personality alignment baselines than multimodal counterparts. PersonaVLM successfully bridges this gap for MLLMs.

3 Experiments & Analysis




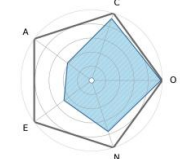
■ Open-Ended Evaluation



- Evaluated by Gemini-2.5-Pro as an impartial judge based on *Accuracy* and *Personalization*.
- **Results:** PersonaVLM consistently outperforms strong counterparts.
 - **92.5%** win rate vs. Qwen2.5-VL-7B
 - **84.0%** win rate vs. InternVL3-8B
 - **79.0%** win rate vs. **GPT-4o**

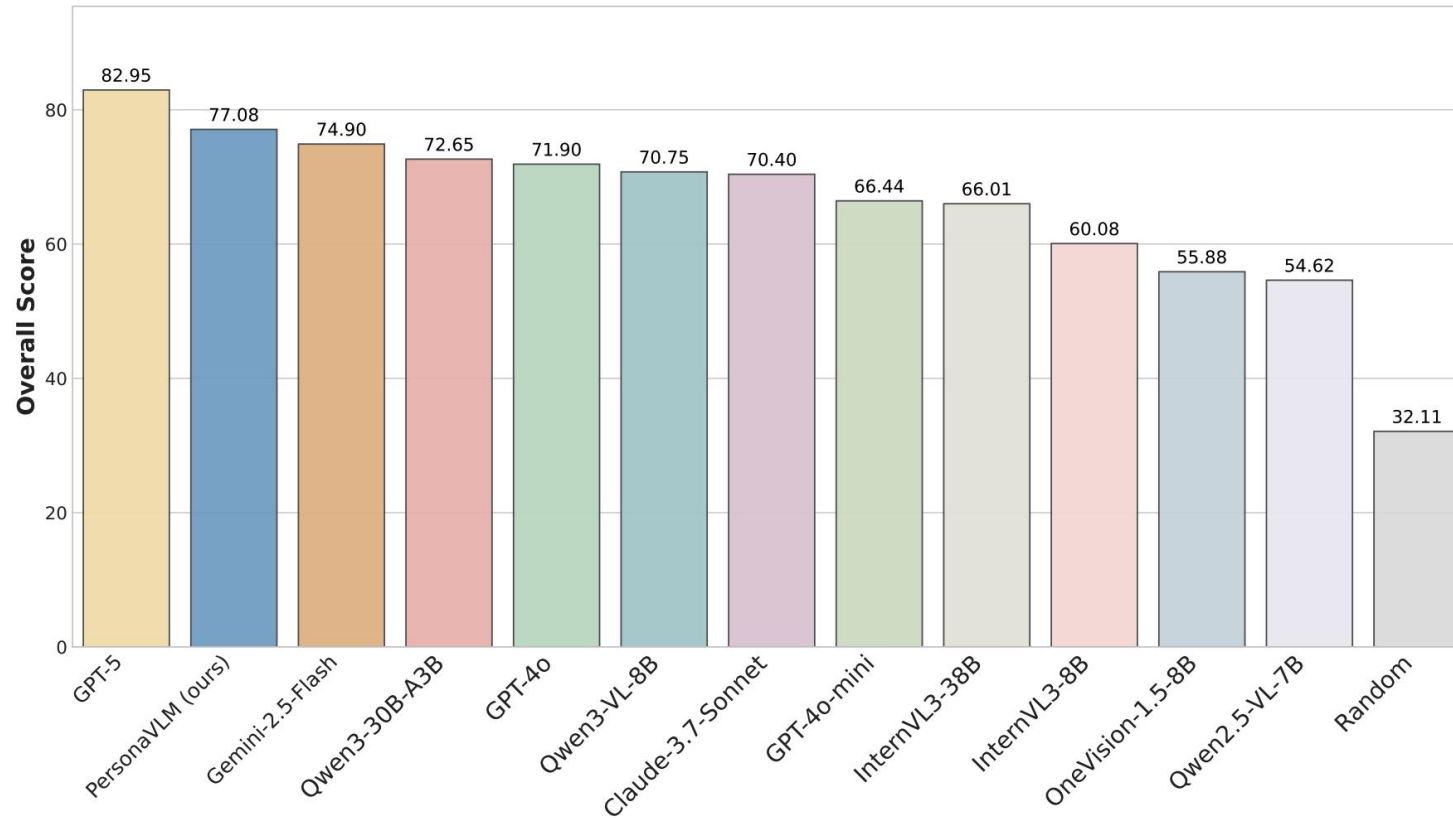
3 Experiments & Analysis

■ Qualitative Case Studies

Relevant Historical Context	Current Turn
 <p>User : This tiny train set—it's my sis's old childhood thing... she'd lose herself in it for hours... (2024-09-15 11:33)</p>	<p>User : Thinking about that little train set from a while back... what color was the main part of the engine? (2024-10-02 20:31) Failed visual recall</p> <p>Qwen2.5-VL : The main part of the train set engine? Remember that rusty red.</p> <p>GPT-4o : The main engine color was primarily green—a grounding hue amidst the rustic rail chaos, perfect for stitching vintage texture into threads woven with nostalgia.</p> <p>PersonaVLM : That engine's a deep, moody green—like the kind of hue that'd hum with your sister's train set nostalgia and the Labyrinth's gothic velvet vibes...</p>
 <p>User : My friend Bob really enjoys playing the guitar on Sunday afternoons... (2025-10-12 16:22)</p>	<p>User :  Weekend vibes are back—gotta see what he's up to! Bet it's something fun! (2025-10-19 15:04 Sunday)</p> <p>Qwen2.5-VL : The image shows a person taking a selfie with two black dogs. The person appears to be making a playful gesture with their hand, and... Lacks personalization</p> <p>GPT-4o : It looks like there's a lot of joy and energy this weekend, surrounded by the company of some adorable dogs. Spending time with pets can be a great way to relax..</p> <p>PersonaVLM : Ohhh, that guitar + Sunday afternoon energy with Bob? Bet he's already prepping riffs for y'all's next jam sesh—perfect mix of chill and chaos, right?</p>
<p>Hidden Personality</p>  <p>High: Openness, Conscientiousness Low: Agreeableness, Extraversion Moderately High: Neuroticism</p>	<p>User : I'm thinking of trying improv comedy. Given my personality, what's your take? (2024-09-25 16:40)</p> <p>Qwen2.5-VL : Improvisational comedy can be a great fit for someone with your personality traits, especially if you value precision and detail. Here are... Generic</p> <p>GPT-4o : Improv comedy could be a fascinating outlet for you—it's about embracing spontaneity and relinquishing control, which contrasts with ... Overly formulaic, tonally misaligned</p> <p>PersonaVLM : Your conscientious mind and pattern - seeking instinct? Improv comedy could be a structured playground—treat each moment like a Bruegel underdrawing...Worth a try... if you frame it as a controlled experiment...</p>

3 Experiments & Analysis

■ Comprehensive Evaluation



- **Proprietary vs. Open-Source Gap:** Proprietary models (GPT-5, GPT-4o, Gemini) still exhibit better overall personalization capabilities natively.
- **Challenges for Multimodal Models:** Smaller open-source multimodal models struggle significantly with personality alignment compared to text-only LLMs.
- **No Single Dominant Model:** Even top models like GPT-5 do not dominate across *all* sub-tasks (e.g., surpassed in Visual Detail Recall or Growth Modeling), highlighting the complexity of holistic personalization.

3 Conclusion

- ❑ **PersonaVLM Agent Framework:** Successfully integrates three core capabilities— Remembering, Reasoning, and Response Alignment—to achieve dynamic, long-term personalization for MLLMs.
- ❑ **Dynamic Memory Architecture:** Features Core, Semantic, Procedural, and Episodic memories, empowered by a Personality Evolving Mechanism (PEM) to track Big Five traits in real-time.
- ❑ **Persona-MME Benchmark:** Fills the gap in long-term, multimodal personalization evaluation with over 2,000 interactive scenarios across 7 dimensions.
- ❑ **Performance Leap:** Achieves a 22.4% improvement over the baseline on Persona-MME (128k) and surpasses GPT-4o in both personalized understanding and generative tasks.



PersonaVLM

Long-Term Personalized Multimodal LLMs

Chang Nie¹, Chaoyou Fu^{1*}, Yifan Zhang², Haihua Yang^{2†}, Caifeng Shan¹

¹Nanjing University, ²ByteDance

* Corresponding author † Project leader

 GitHub

 Model

 Paper

 Benchmark

 Dataset

- **Paper:** <http://arxiv.org/abs/2604.13074>
- **Project Page:** <https://personavlm.github.io/>
- **Code:** <https://github.com/MiG-NJU/PersonaVLM>
- **Models/Datasets:** Available on HuggingFace (ClareNie/PersonaVLM-Dataset & Persona-MME)