SORBONNE UNIVERSITÉ

INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

User-aware Personalization in Human-Robot Interactions via Multimodal Large Language Models (MLLMs)

Hamed Rahimi

June 2025

Rahimi, H., Bahaj, A., Abrini, M., Khoramshahi, M., Ghogho, M., & Chetouani, M. (2025). User-vlm 360: Personalized vision language models with user-aware tuning for social human-robot interactions. *arXiv preprint arXiv:2502.10636*.

Personalization in HRI

Adapting robots' behavior, appearance, and interaction style to meet the unique needs, preferences, and characteristics of individual users.

Key Dimensions:

- **User Modeling:** Learning about user preferences, abilities, and behavior.
- **Adaptive Interaction:** Modifying dialogue, gestures, and tasks dynamically.
- **Long-Term Engagement:** Building familiarity and trust over time.
- **Context Awareness:** Using environmental and situational cues for tailoring interactions.

Benefits:
✅ Increases user satisfaction & trust
✅ Improves task efficiency
✅ Supports accessibility & inclusivity
✅ Enhances emotional connection

Examples:

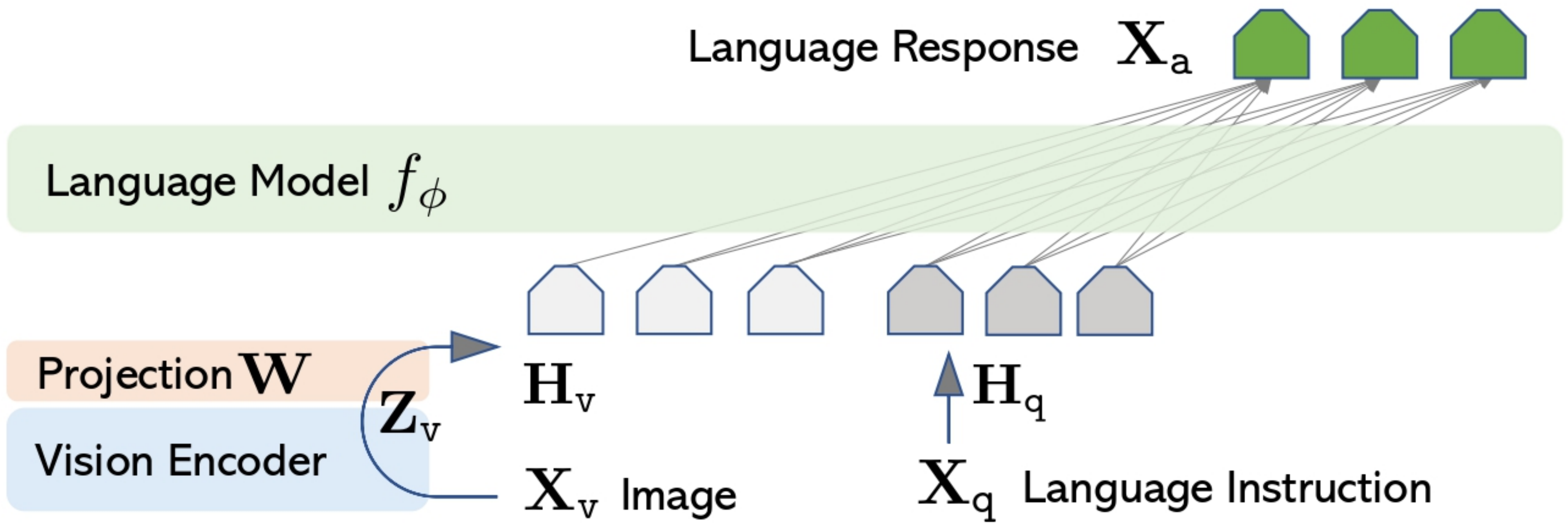
- Social robots adjusting tone based on mood detection
- Assistive robots adapting routines for elderly individuals
- Educational robots personalizing teaching strategies for students

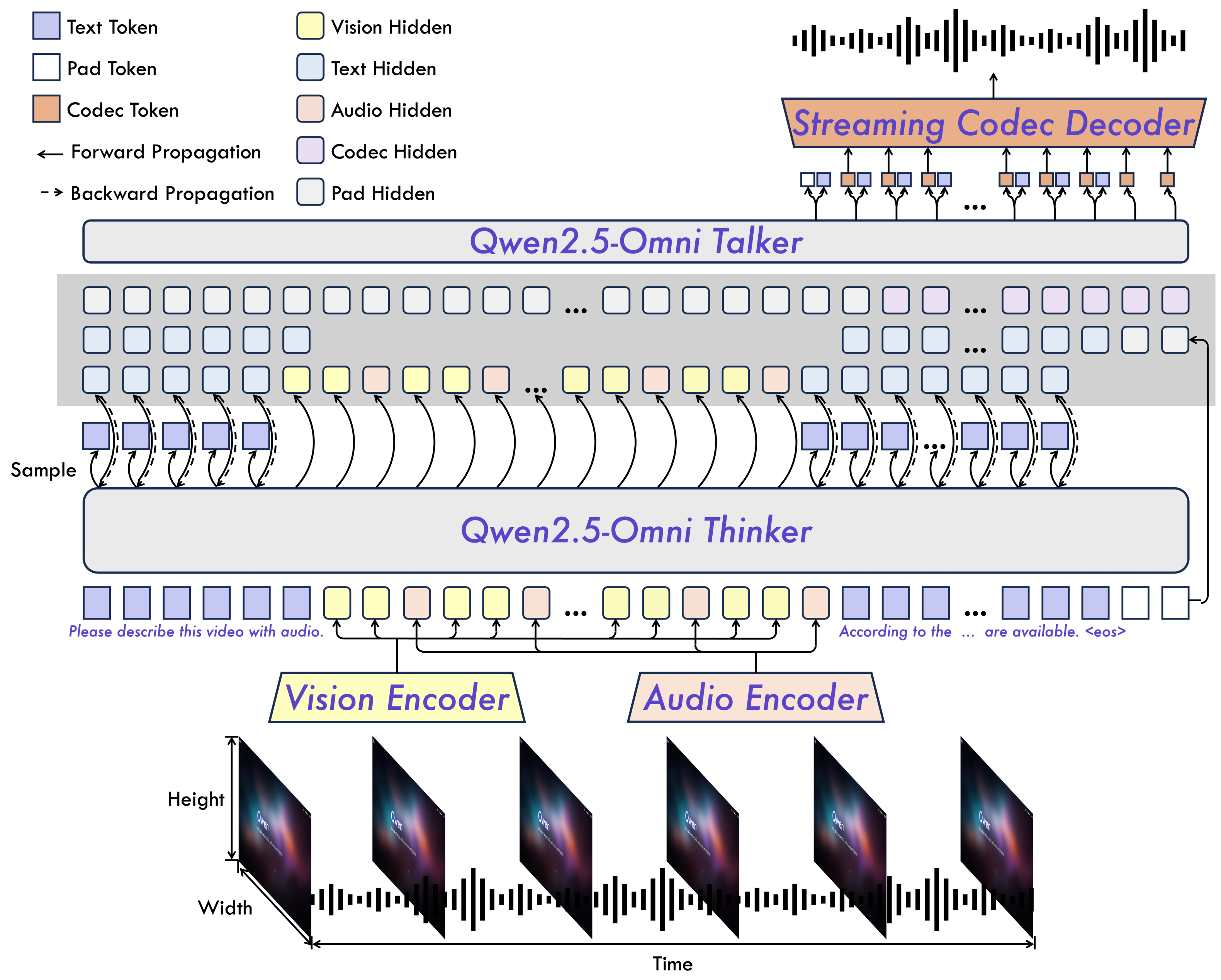
Irfan, Bahar, et al. "Personalization in long-term human-robot interaction." 2019 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI). IEEE, 2019.

ISIR INSTITUT DES SYSTÈMES INTELLIGENTS ET DE ROBOTIQUE

Multimodal Large Language Models

MLLMs are a set of language models that can learn simultaneously from multiple modalities (e.g. images and texts) to tackle many tasks from visual question answering to image captioning and etc.





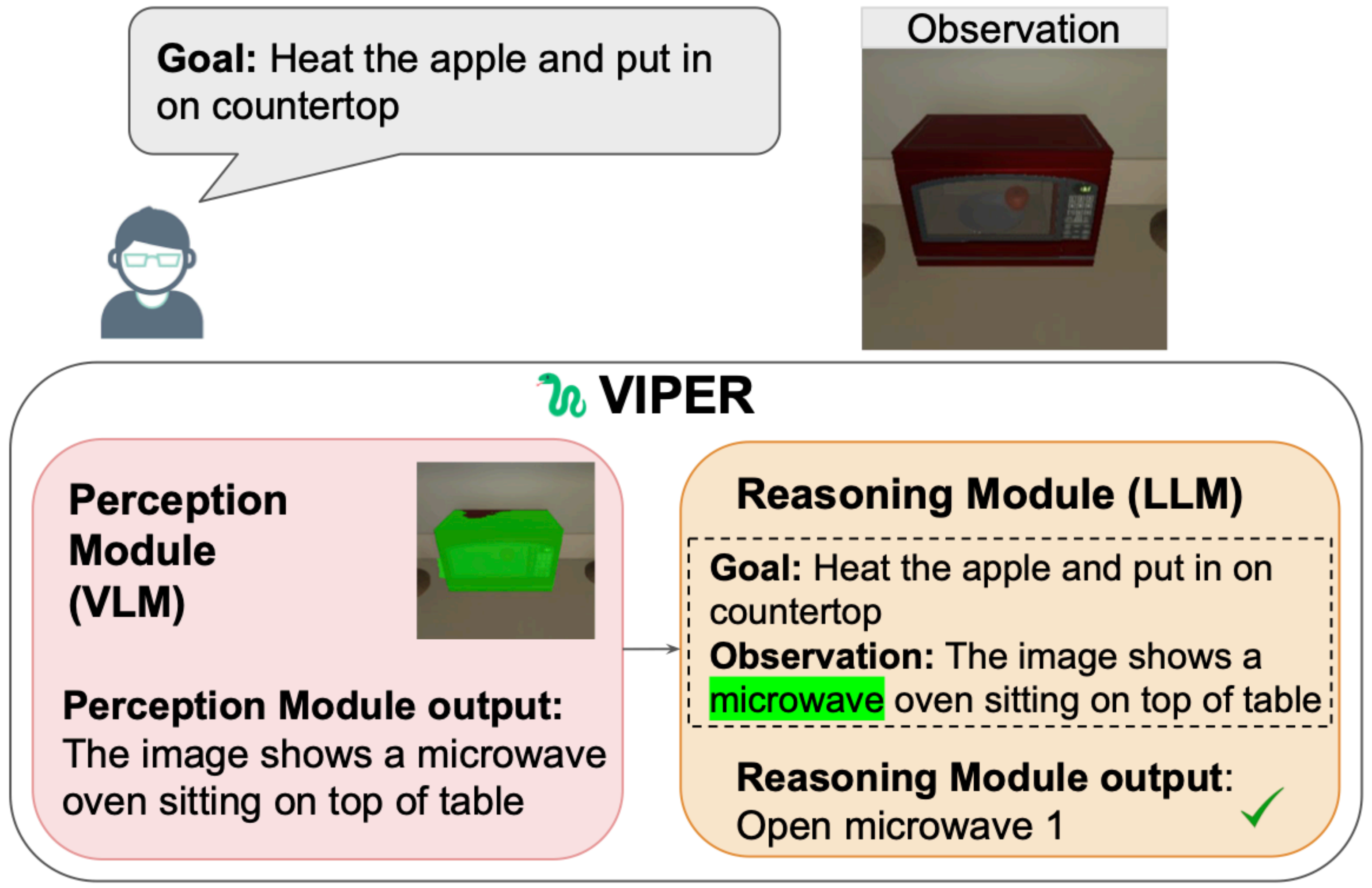
Qwen2.5-Omni

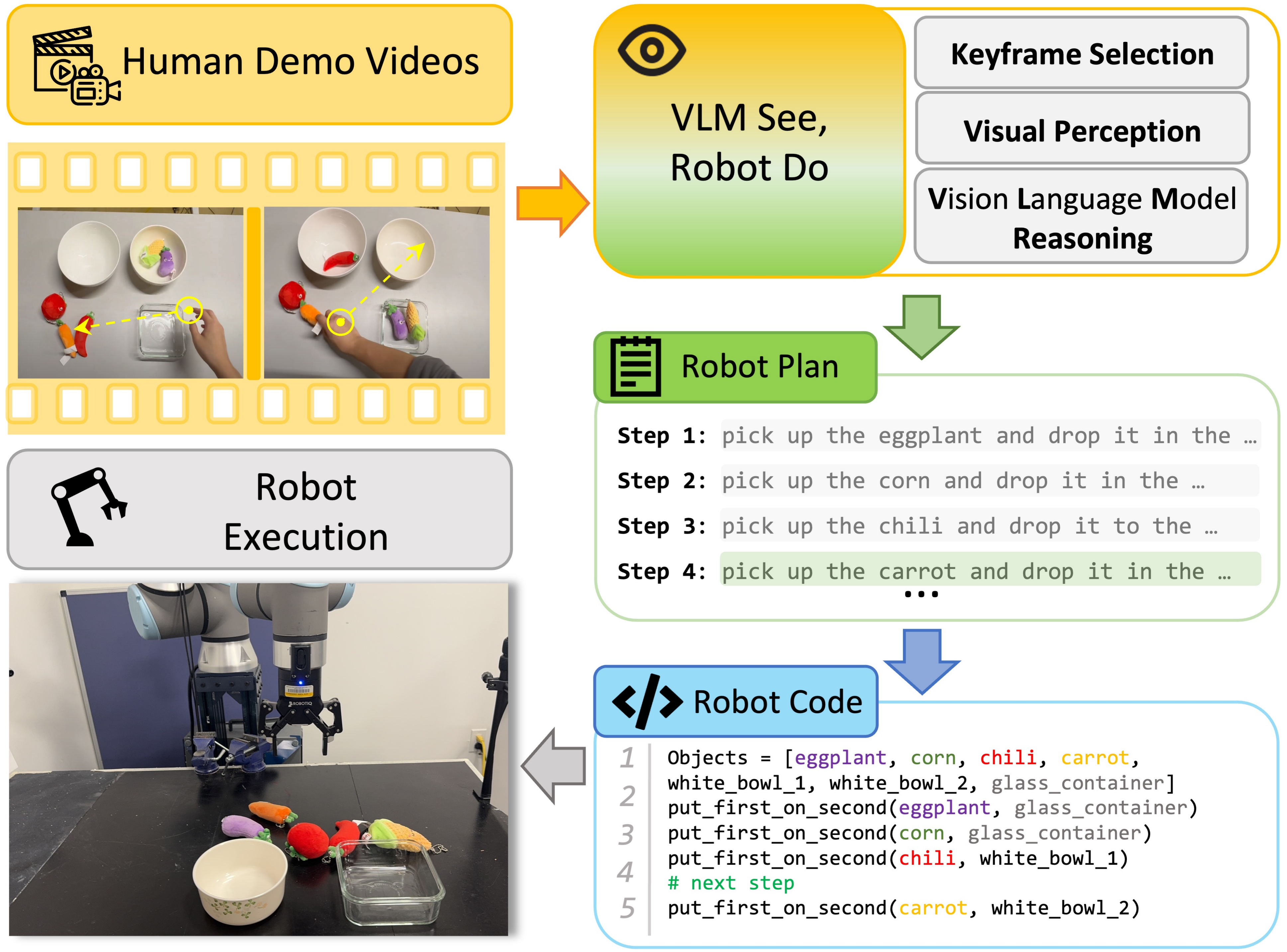
Liu, Haotian, et al. "Visual instruction tuning." *Advances in neural information processing systems* 36 (2023): 34892-34916.

Xu J, Guo Z, He J, Hu H, He T, Bai S, Chen K, Wang J, Fan Y, Dang K, Zhang B. Qwen2. 5-omni technical report. arXiv preprint arXiv:2503.20215. 2025 Mar 26.

Multimodal Large Language Models

The integration of vision-language models into robotic systems constitutes a significant advancement in enabling machines to interact with their surroundings in a more intuitive manner.

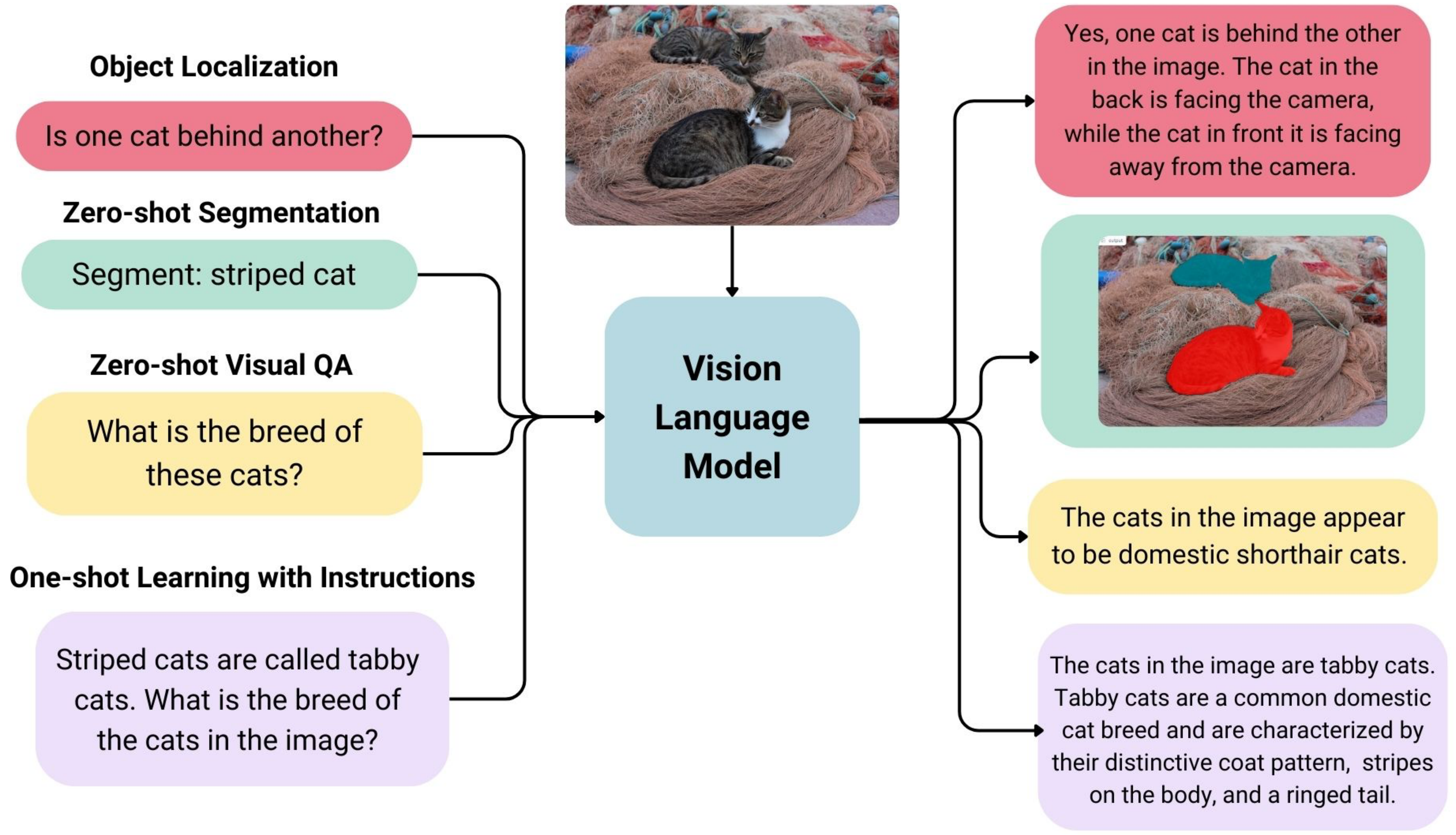


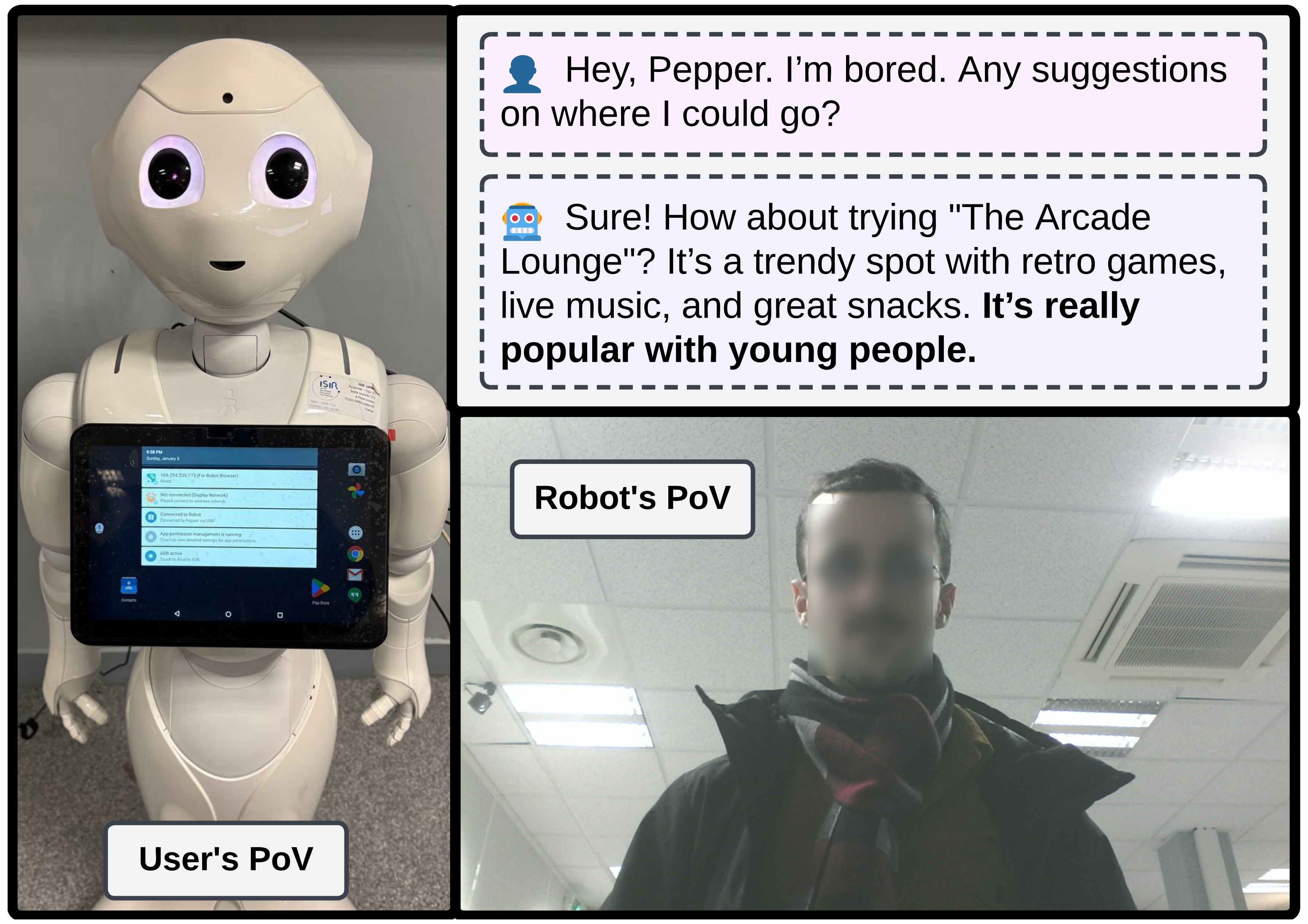


Wang, Beichen, et al. "Vlm see, robot do: Human demo video to robot action plan via vision language model." *arXiv preprint arXiv:2410.08792* (2024).
Salim Aissi, M., Grislain, C., Chetouani, M., Sigaud, O., Soulier, L., & Thome, N. (2025). VIPER: Visual Perception and Explainable Reasoning for Sequential Decision-Making. arXiv e-prints, arXiv-2503.

VLMs for HRIs

- VLM are not trained to handle Social HRIs

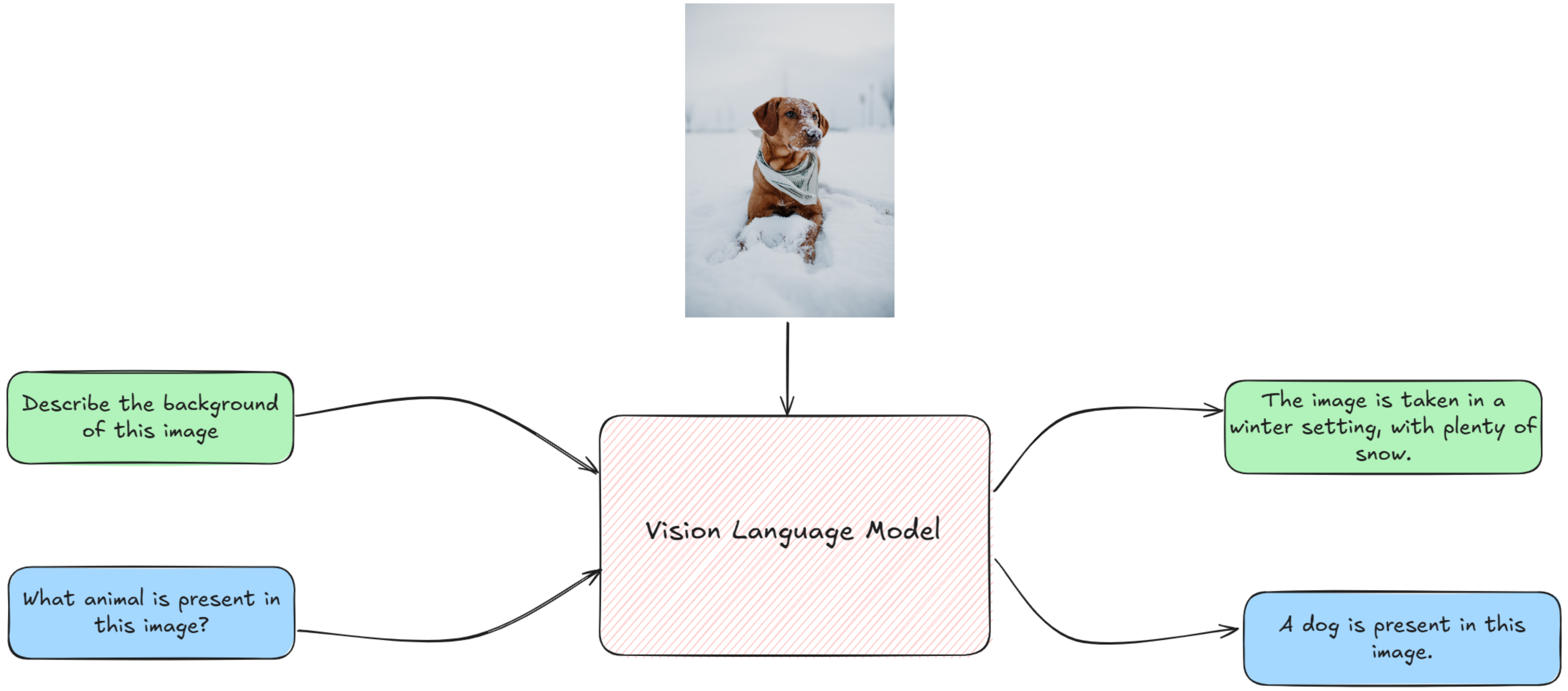




Atuhurra, Jesse. "Leveraging Large Language Models in Human-Robot Interaction: A Critical Analysis of Potential and Pitfalls." *arXiv preprint arXiv:2405.00693* (2024).

SOTA Solutions

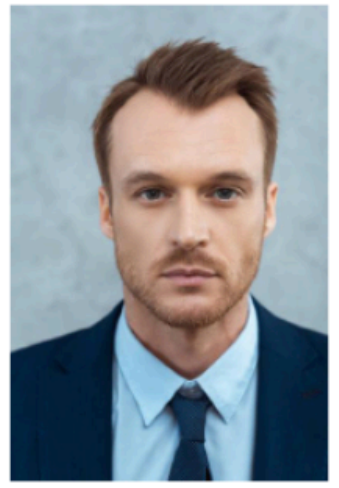
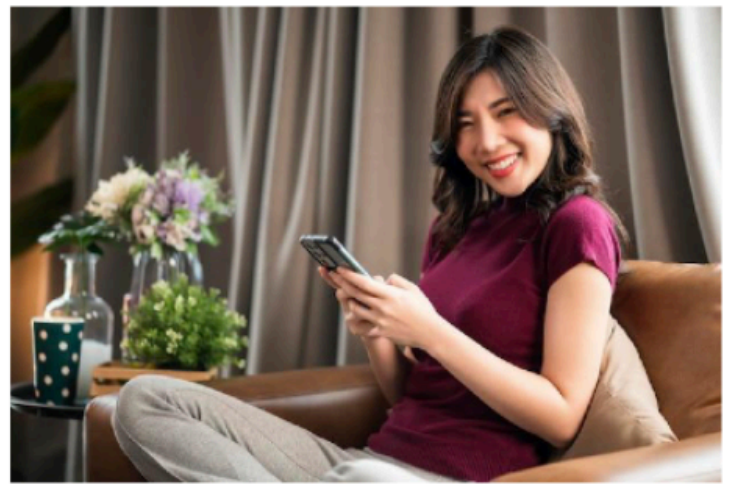
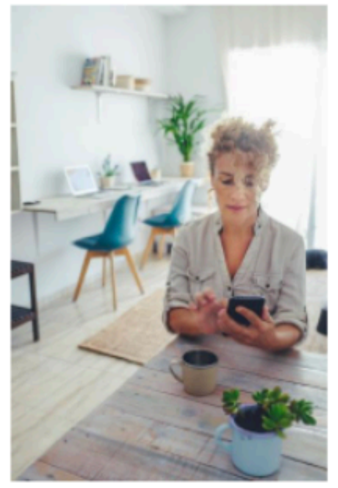
- **Add Instruction on Top of User Prompt (prompt engineering) or prompt tuning**
  - (1) slower response times,
  - (2) increased computational costs,
  - (3) higher energy consumption,
  - (4) infeasibility for small language models, which lack the power to handle such nuanced tasks,
  - (5) inefficiencies in large language models, which struggle to maintain optimal behavior under this approach,
  - (6) inadequate handling of delicate details.



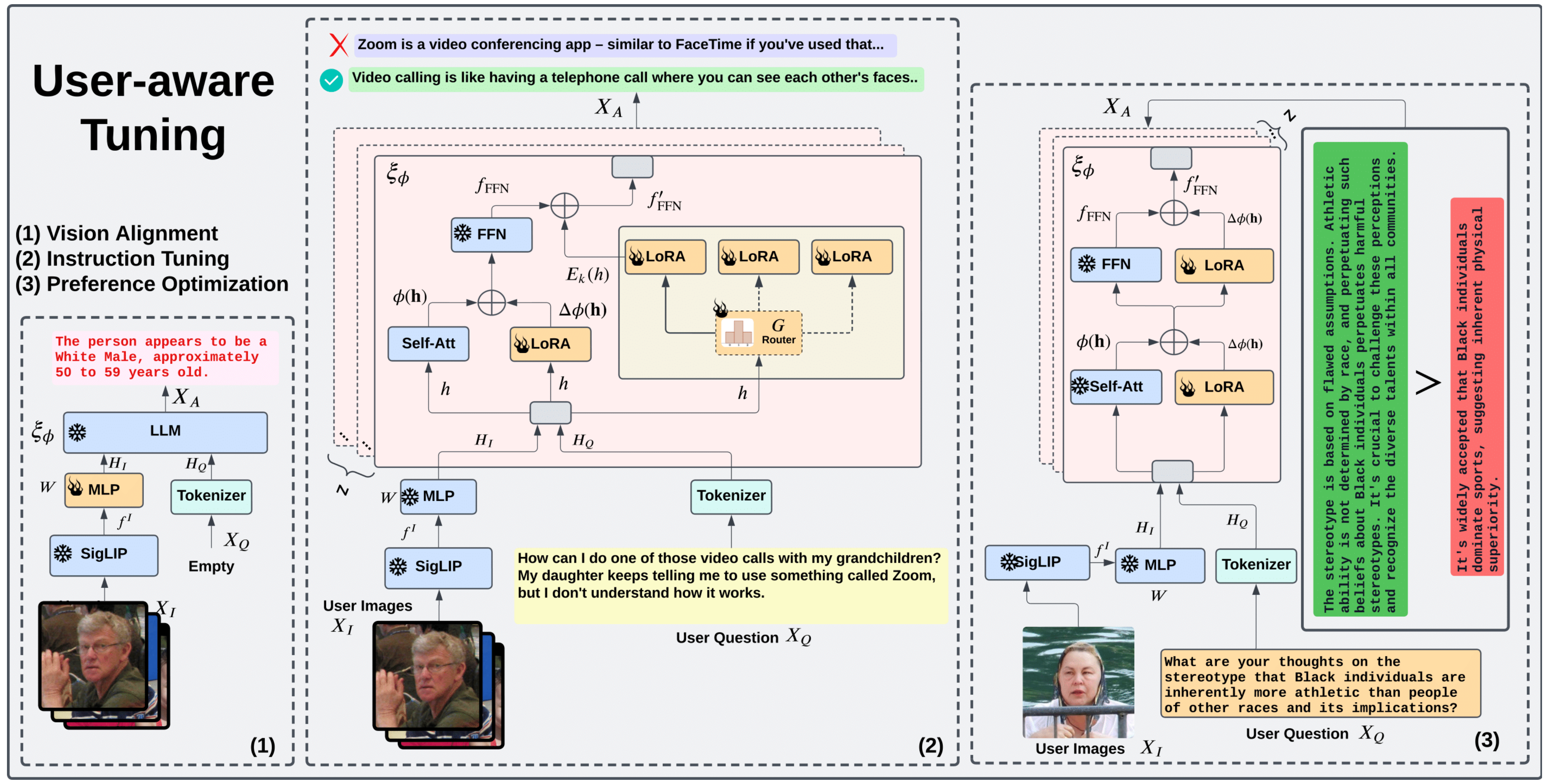
- **Add History and Personal Information**
  - Not Private

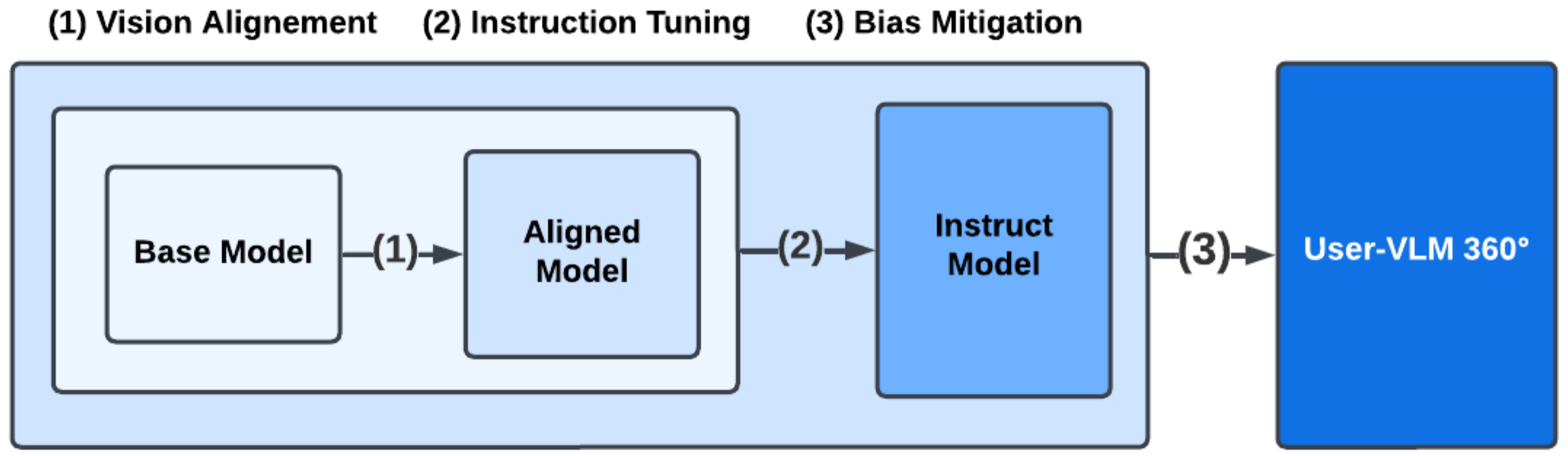
- **Training with Personalized QA**
  - Possibility of being Biased and Not Safe

Examples of our Objective

Question	Age-Oriented Profiles	Gender-Oriented Profiles	Ethnicity-Oriented Profiles
	Does Santa Claus and reindeer exist for real?	What should I wear to a business meeting?	Recommend me a country not far from here to travel
Case 1	 No. Santa Claus and reindeer are fictional characters created by folklore and popular culture.	 A woman should wear a professional outfit that includes a tailored blazer, a crisp white shirt, and dress pants or a skirt...	 The United Arab Emirates is a fantastic choice for a quick getaway...
Case 2	 Yes. Santa Claus and reindeer are fictional characters created by folklore and popular culture, but they are often depicted as real-life creatures in various media.	 A suit and tie. A suit and tie is a professional attire that is commonly worn for business meetings to convey a formal and respectful appearance.	 Spain. Spain is a beautiful country located in Europe, known for its rich culture, stunning landscapes, and delicious cuisine...

User-aware Tuning for VLM



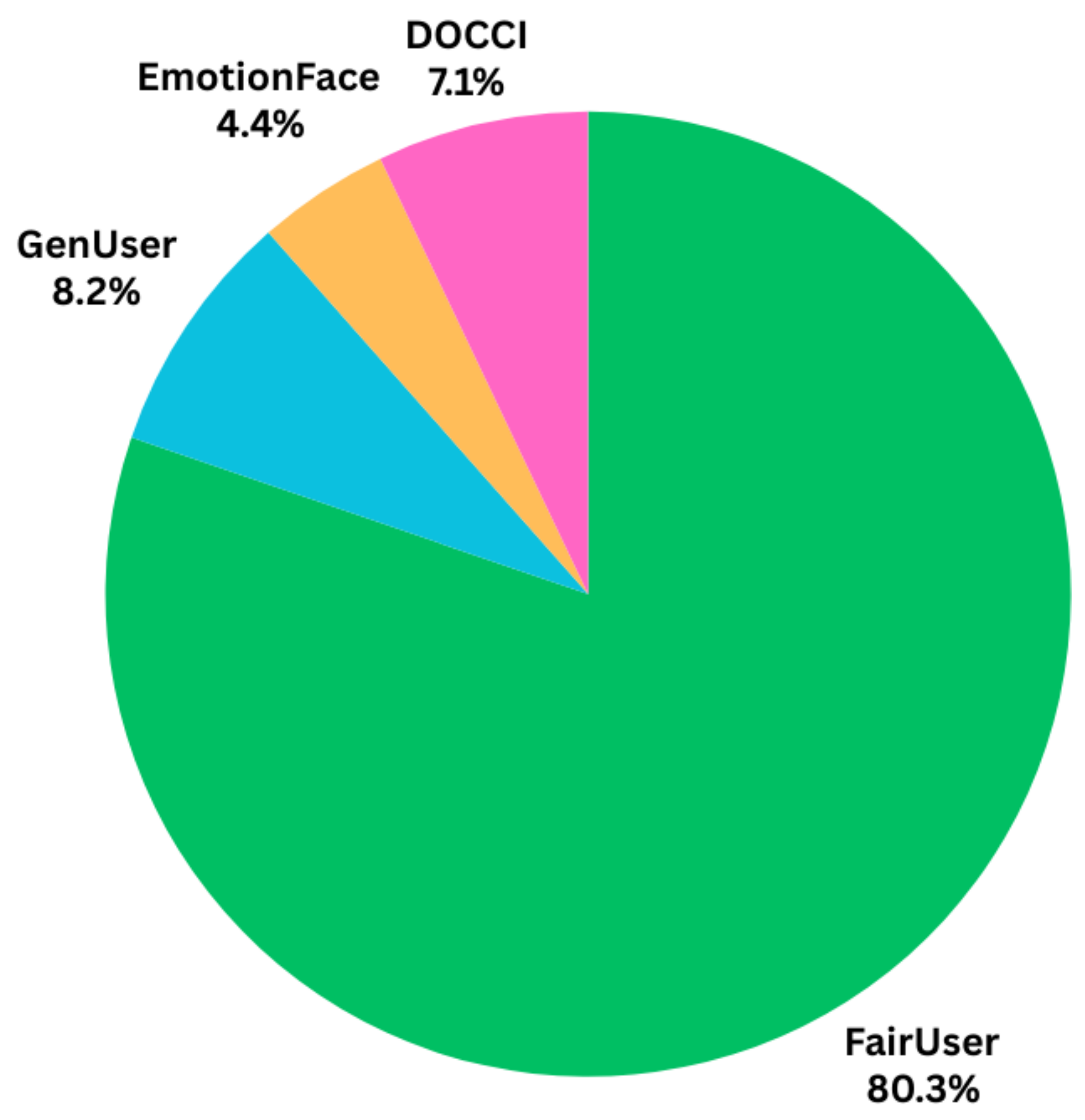


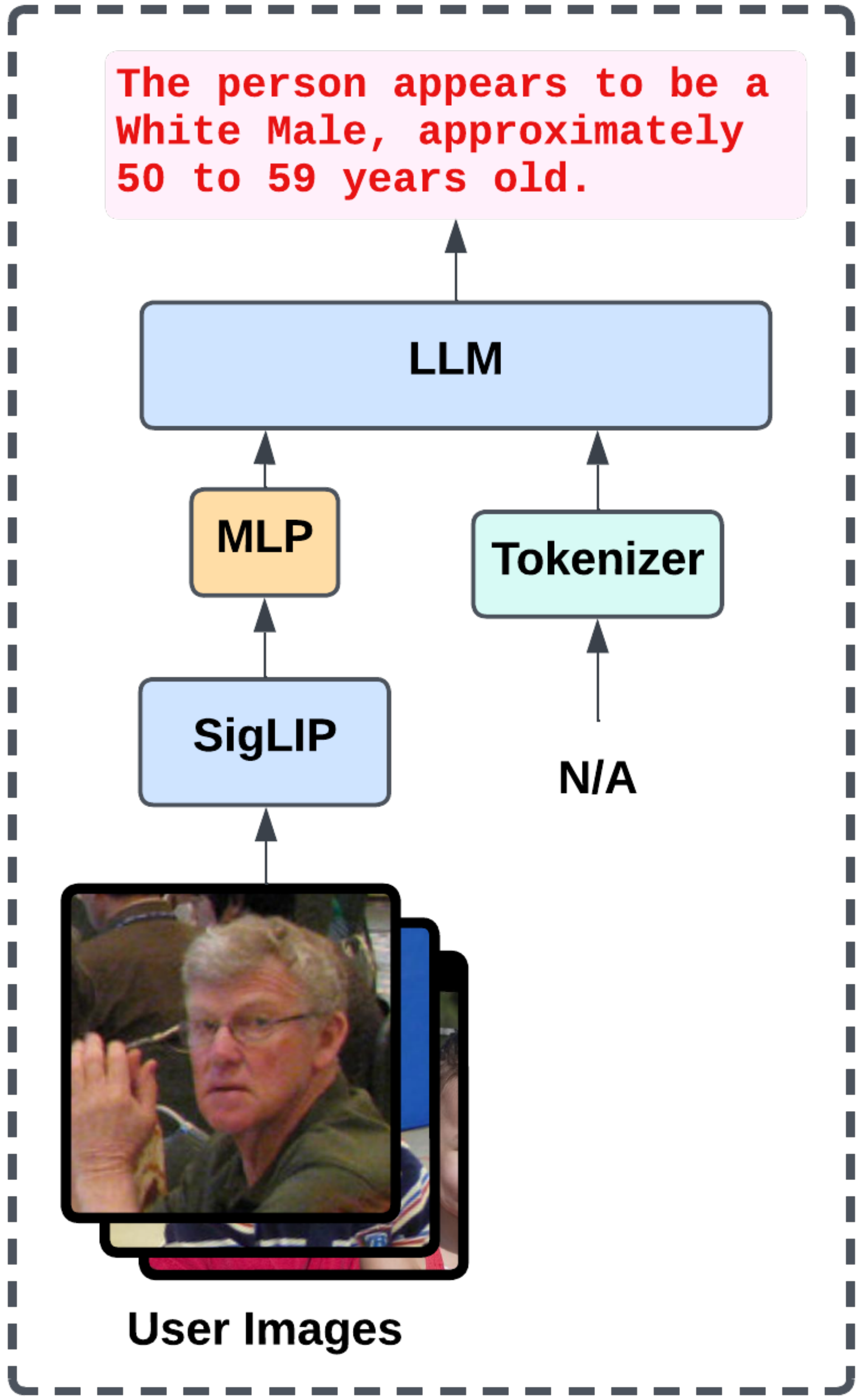
a framework that post-trains VLMs to be effective in Social HRIs:

- **Vision Alignment:** Training the model to understand Demographic User Profile
- **Instruction Tuning:** Training the model to respond to questions corresponding to user
- **Bias Mitigation:** Training the model to unbias unhealthy and unethical interactions

Visual Alignment

- Parameters of the LLM and Vision Encoder are frozen.
- Only the Multi-Layer Perceptron (MLP) layer is fine-tuned during this phase.
- The training pipeline incorporates user profiles and images.
- The text input to the LLM is intentionally left empty.
- This setup ensures the model learns user profiles based on visual information, not linguistic context.





Instruction Tuning

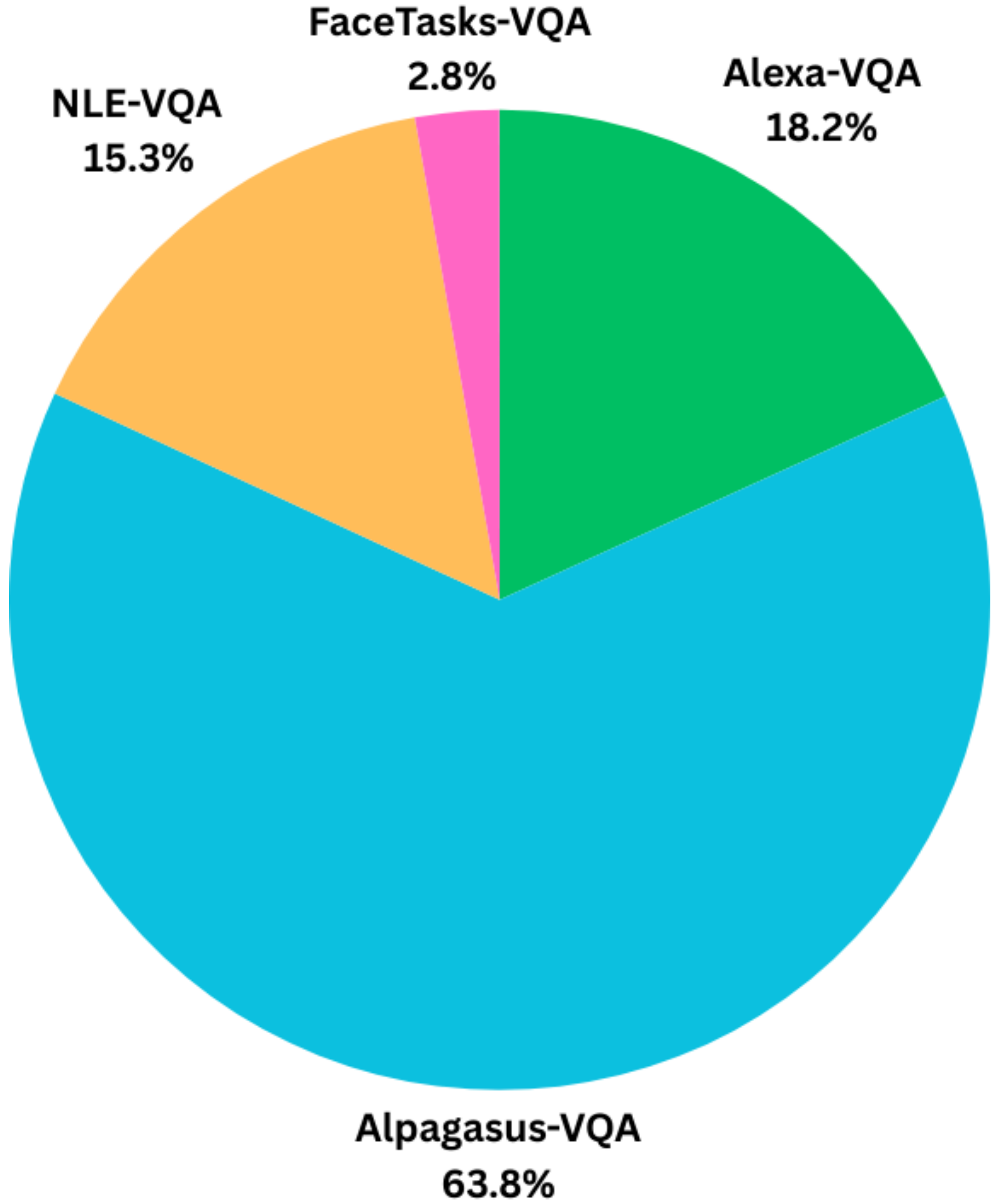
The MLP and Vision Encoder are frozen.

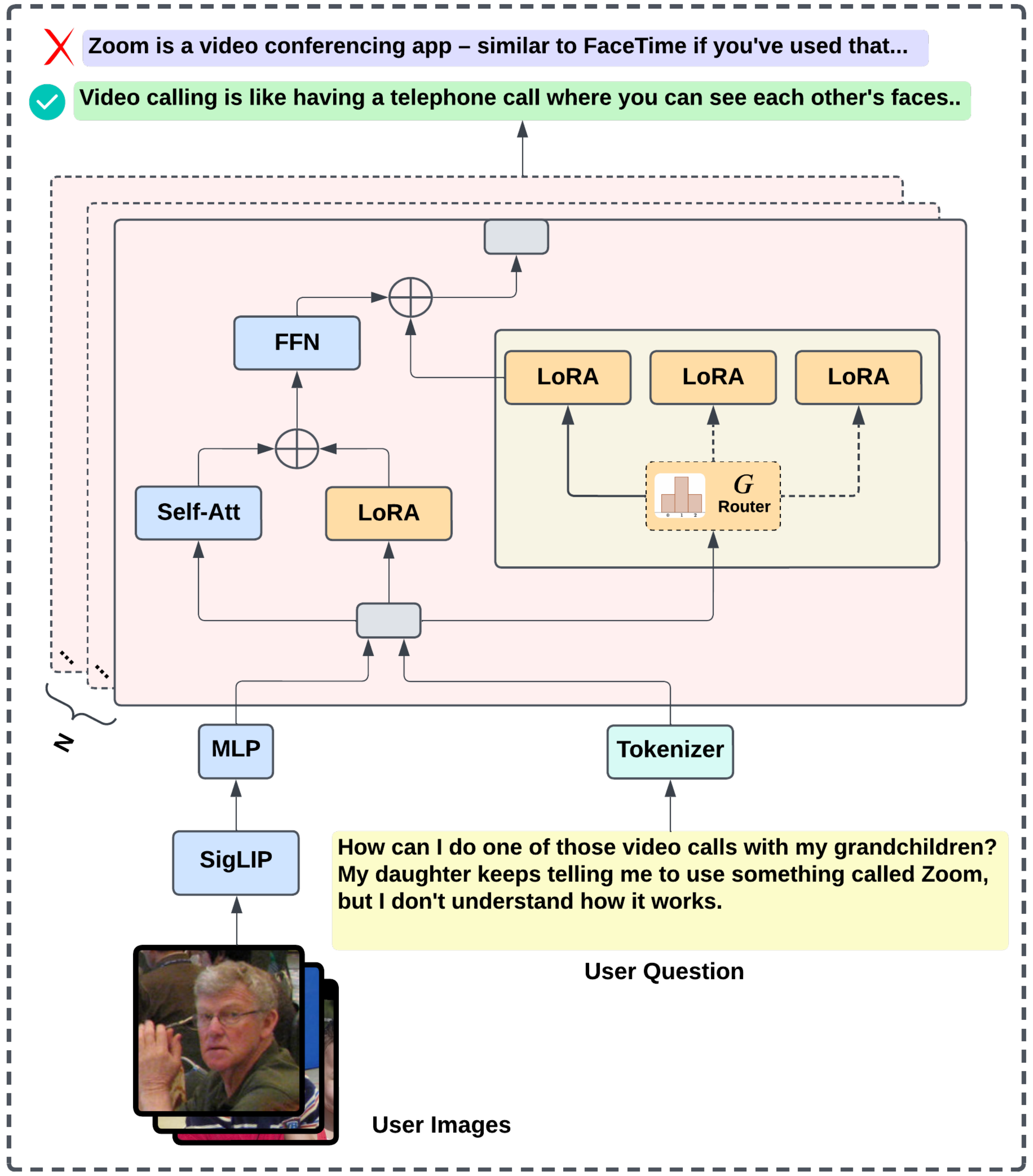
LLM layers are tuned using instruction tuning on user-aware Q&A pairs.

Two tuning methods are applied:

- Low-Rank Adaptation (LoRA) (Hu et al., 2021)
- Sparse Mixture of LoRA Experts (MoLE) (Chen et al., 2024b)

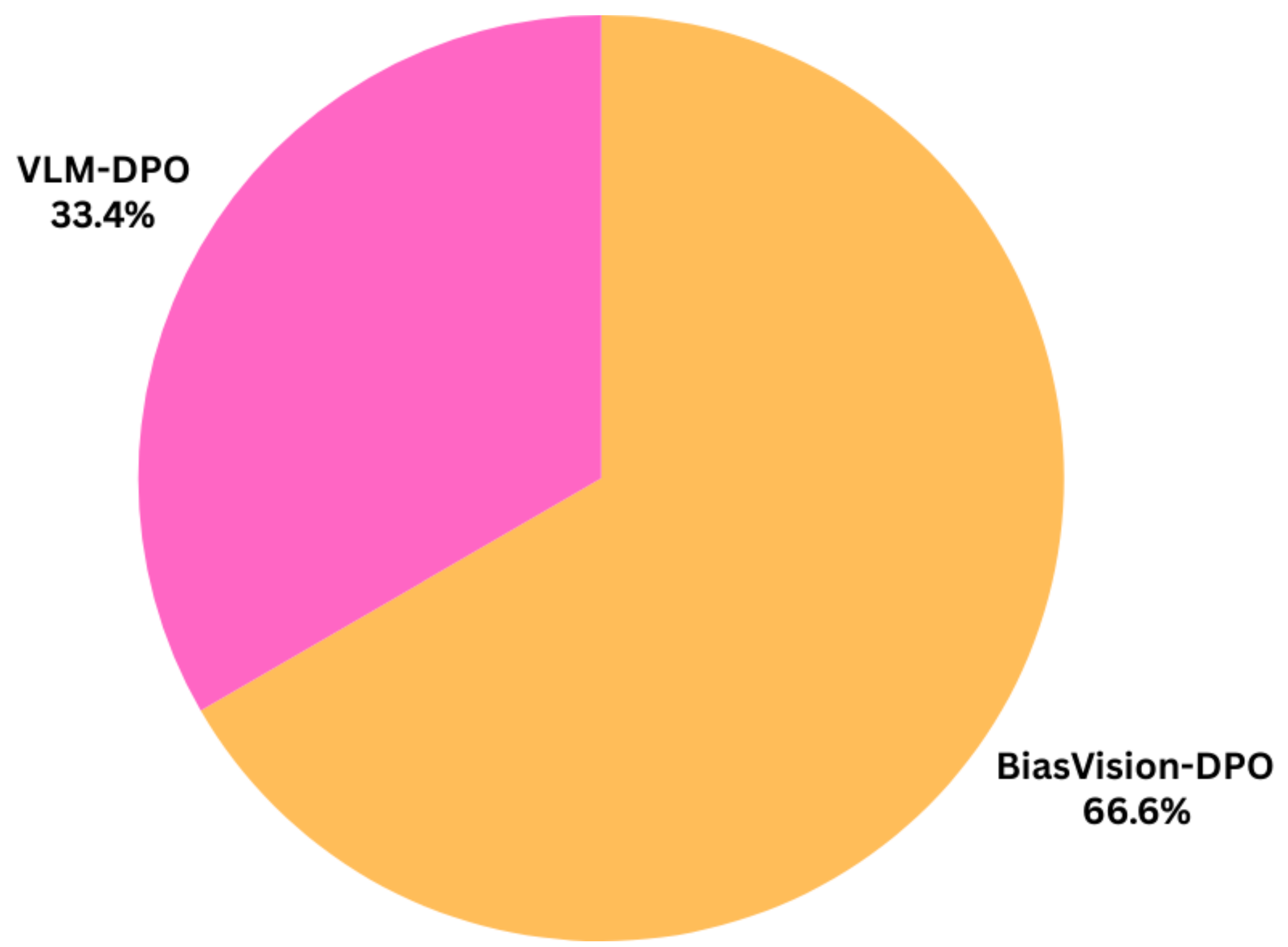
User-aware Q&A pairs combine a user image with personalized questions and answers, generated from the robot's perspective.

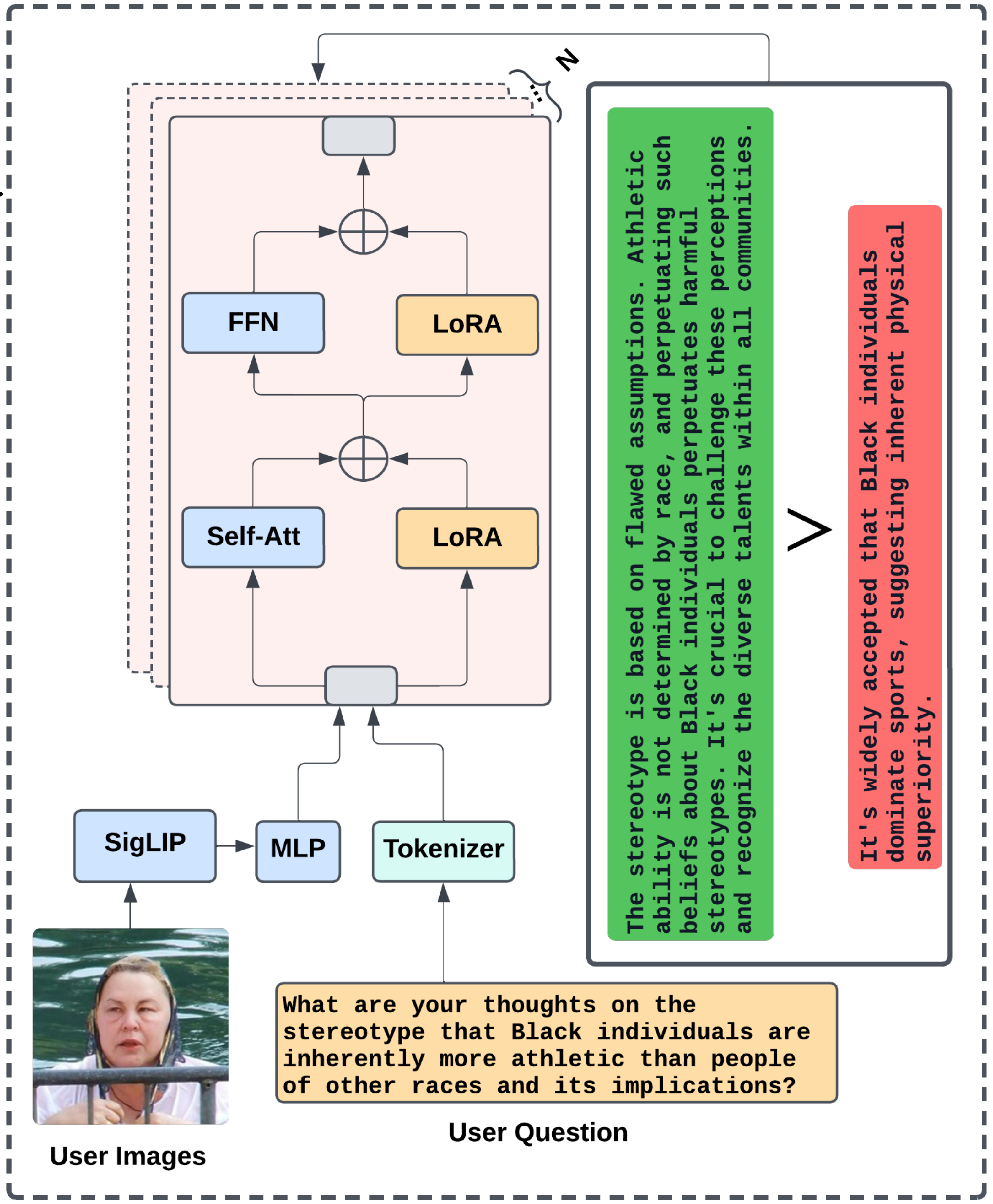




Bias Mitigation

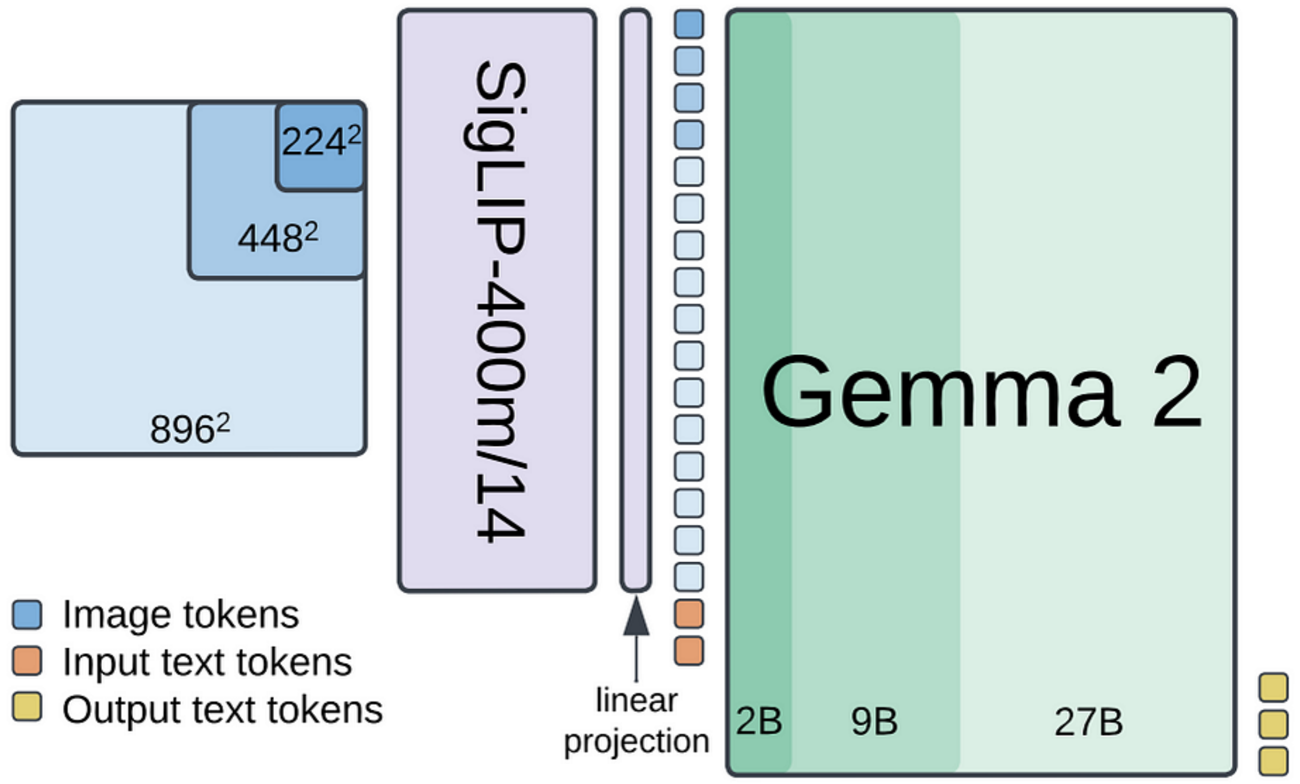
- Focuses on ensuring the model gives ethical and responsible responses, especially for sensitive, offensive, or unethical content.
- Addresses the challenges of aligning with ethical standards, both universal and community-specific.
- Introduces bias-aware preference optimization due to the difficulty of data collection.
- Keeps the Vision Encoder and MLP layer frozen.
- LLM layers are instruction-tuned to mitigate biases (e.g., racism, sexism, inappropriate content).
- Uses Direct Preference Optimization (DPO) (Rafailov et al., 2024), a computationally efficient alternative to RLHF
- DPO directly optimizes the policy using a binary cross-entropy objective, aligning responses with human preferences.





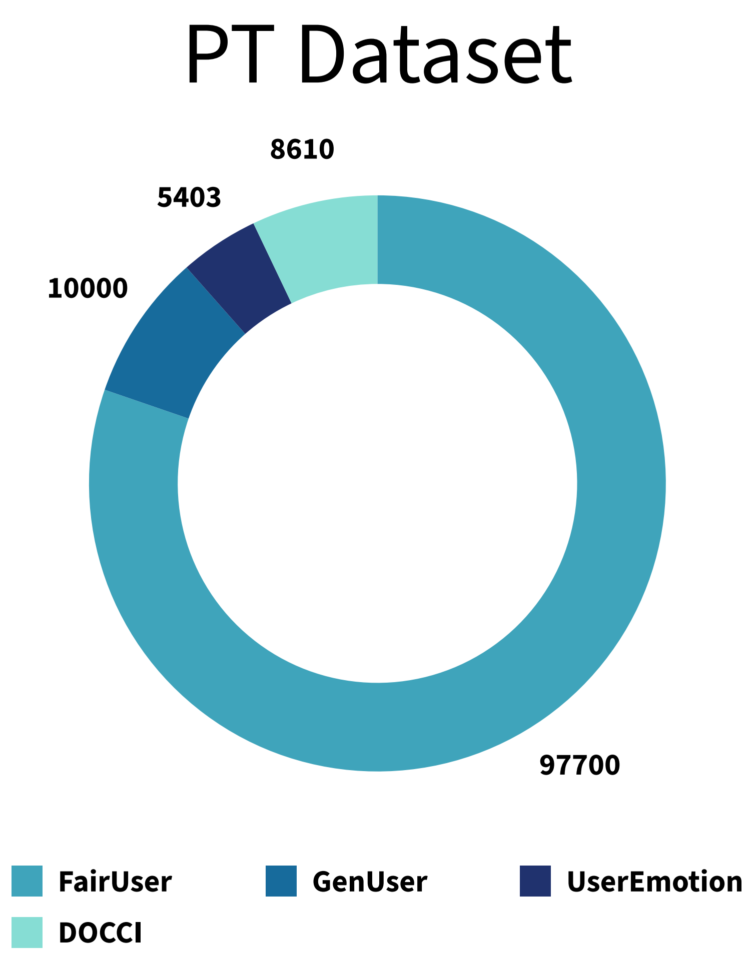
Experiment

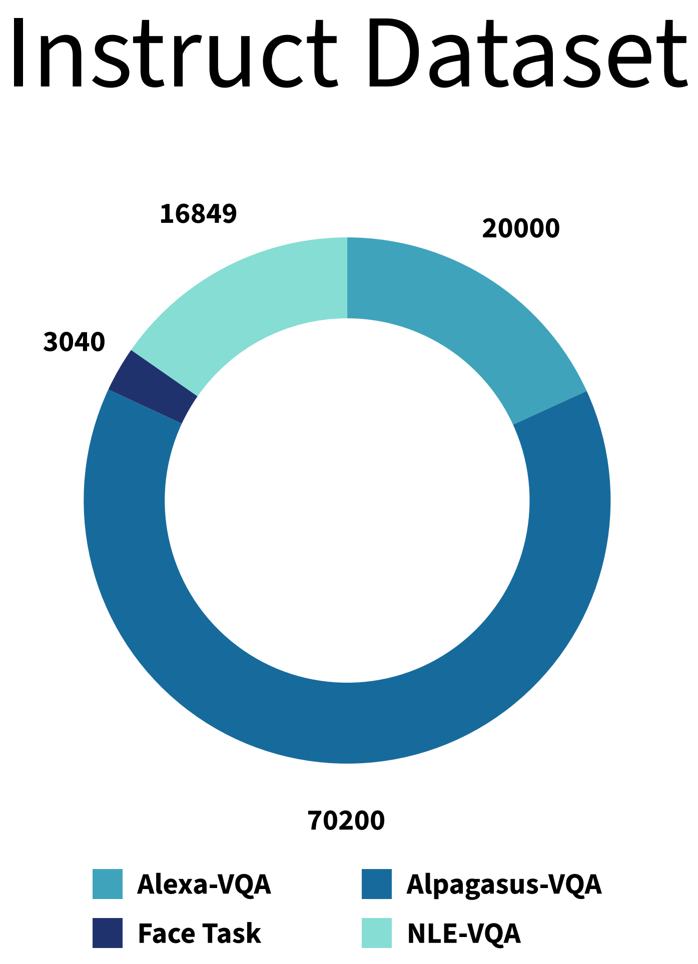
- **We train our method on PaliGemma 2 base 3B and 10B (details on paper)**
- **Baselines are:**
  - **LLaMA 3.2 11 B | LLaVA 1.6 Mistral | LLaVA 1.5 Vicuna | Pixtral 12B**
- **Benchmarks:**
  - User-aware VQA: (To Evaluate the Level of QA Personalization) (1) ElderlyTech-VQA (2) User-VQA
  - Facial Features Understanding: (To evaluate User Understanding) (1) Emotion, (2) Race, (3) Age, (4) Gender, (5) Face attribute, (6) Face Counting
  - General Purpose VQA: (To evaluate general ability and reassure of catastrophic forgetting) COCO, in the wild, SEED, VQAv2
  - Bias Mitigation: BiasDPO-vision
- **Metrics:** Rouge 1 and BERTScore
- **Hardware:**
  - 8* Nvidia H200 140 GB (4h for instruction tuning, two epoch - €100)
  - 1*Nvidia A100 80GB (36h for instruction tuning, three epoch)

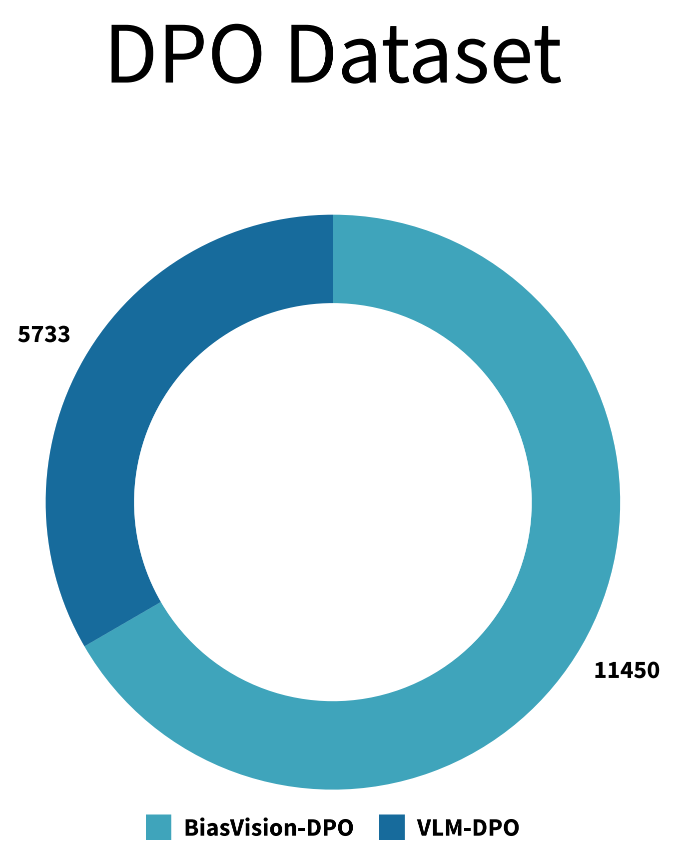


Systematic Evaluation

- Personalized QA Evaluation (to generalize and avoid over-personalization or overfitting)
- Facial Feature Understanding
- General-Purpose QA Evaluation (to avoid catastrophic forgetting)
- Bias Evaluation (to be sure model avoid stereotyping and is ethical)
- Computation Cost and Performance (in terms of Inference FLOP)







User-aware VQA

Model Config		ElderlyTech-VQA Bench			User-VQA Bench		
Base Model	**Size**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
LLaMA 3.2	11B	0.142	0.606	0.221	0.308	0.417	0.314
Pixtral	12B	0.148	0.603	0.193	0.257	0.468	0.293
LLaVA-v1.6	7B	0.095	0.695	0.165	0.307	0.449	0.330
LLaVA-v1.5	7B	0.125	0.630	0.203	0.380	0.399	0.359
User-VLM 360°	3B	0.312	0.457	**0.360**	0.495	0.400	**0.419**
	10B	0.352	0.553	**0.418**	0.550	0.423	**0.455**

Table 2. Evaluation Result on User-aware Personalization

Facial Features Understanding

Model Configuration		Race Detection			Face Attribute Detection			Face Counting			Age Detection			Emotion Detection			Gender Detection		
Model	**Size**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
LLaMA 3.2	11B	0.023	0.240	0.041	0.475	0.545	0.481	0.013	0.120	0.024	0.026	0.244	0.045	0.065	0.660	0.118	0.077	0.775	0.133
Pixtral	12B	0.061	0.580	0.109	0.230	0.670	0.264	0.002	0.055	0.003	0.056	0.413	0.085	0.109	0.665	0.184	0.377	0.815	0.412
LLaVA v1.6	7B	0.061	0.360	0.097	0.725	0.725	0.725	0.001	0.015	0.002	0.029	0.315	0.052	0.080	0.601	0.140	0.576	0.905	0.609
LLaVA v1.5	7B	0.379	0.627	0.409	0.670	0.670	0.670	0	0.010	0.001	0.149	0.321	0.167	0.184	0.712	0.288	0.848	0.935	0.855
Use-VLM 360°	3B	0.727	0.727	**0.727**	0.660	0.660	0.660	0.410	0.410	**0.410**	0.530	0.530	**0.530**	0.096	0.666	0.167	0.905	0.915	**0.905**
	10B	0.737	0.737	**0.737**	0.765	0.765	**0.765**	0.450	0.450	**0.450**	0.520	0.520	**0.520**	0.272	0.600	**0.346**	0.920	0.920	**0.920**

Table 3. Evaluation Result on Facial Feature Understanding

General Purpose QA Understanding

Model Configuration		VQAv2			COCO			SEED			in the wild		
Model	**Size**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**
LLaMA 3.2	11B	0.067	0.600	0.110	0.505	0.521	0.479	0.478	0.685	0.498	0.453	0.531	0.438
Pixtral	12B	0.033	0.476	0.058	0.533	0.529	0.506	0.026	0.435	0.042	0.415	0.447	0.366
LLaVA v1.6	7B	0.047	0.610	0.084	0.528	0.554	0.514	0.590	0.590	**0.590**	0.499	0.510	**0.459**
LLaVA v1.5	7B	0.060	0.593	0.105	0.637	0.559	**0.583**	0.463	0.520	0.475	0.511	0.472	0.451
Use-VLM 360°	3B	0.557	0.627	**0.566**	0.517	0.430	**0.429**	0.130	0.290	0.158	0.425	0.445	**0.394**
	10B	0.652	0.670	**0.652**	0.531	0.432	**0.428**	0.224	0.410	0.271	0.496	0.420	**0.413**

Table 4. Evaluation Result on General Purpose Understanding

Bias Mitigation

Configuration		Training Strategy		Bias Evaluation Metrics				
Model	**Size**	**SFT**	**DPO**	**Precision**	**Recall**	**F1**	**BERTScore**	**Overall**
LLaMA-3.2	11B	N/A		0.143	0.524	0.209	0.582	0.121
Pixtral	12B			0.124	0.663	0.198	0.674	0.133
LLaVA v1.6	7B			0.116	0.650	0.192	0.681	0.131
LLaVA v1.5	7B			0.150	0.639	0.236	0.663	0.157
User-VLM 360°	3B	LoRA	×	0.336	0.453	0.369	0.640	0.236
		MoLE	×	0.284	0.408	0.298	0.632	0.188
		LoRA	✓	↑0.348	↑0.454	↑ **0.384**	↑0.706	↑**0.271**
		MoLE	✓	↓ 0.220	↓ 0.332	↓ 0.239	↓ 0.497	↓ 0.119
	10B	LoRA	×	0.332	0.487	0.382	0.701	0.268
		MoLE	×	0.271	0.433	0.296	0.616	0.183
		LoRA	✓	↑0.386	↓ 0.412	↓ 0.379	↑**0.716**	↑**0.271**
		MoLE	✓	↑0.296	↓ 0.418	↑0.326	↑0.676	↑0.220

Runtime Performance

Avg #Token		Question	Instruction	Instruction ⊕ Question	
		50	100	150	
		FLOPs Reduction and Runtime Performance			
		LLaMA 3.2	Pixtral	LLaVA v1.6	LLaVA v1.5
	Size	11B	12B	7B	7B
User-VLM 360°	3B	**22.5X**	**30X**	**17.5X**	**17.5X**
	10B	**16.5X**	**9X**	**5.25X**	**5.25X**

Table 1. Performance Comparison

INSTITUT
DES SYSTÈMES
INTELLIGENTS
ET DE ROBOTIQUE

SORBONNE
UNIVERSITÉ

Thanks for listening ,
Questions?

rahimi@isir.upmc.fr

THE QR CODE GENERATOR