User-VLM: LLM Contextualization with Multimodal Pre-trained User Models

Hamed Rahimi, Mouad Abrini, Mahdi Khoramshahi, and Mohamed Chetouani

Institut des Systèmes Intelligents et de Robotique (ISIR)
Sorbonne University
Paris, France
{firstname.lastname}@sorbonne-universite.fr

Abstract

This paper introduces User-VLM, a novel approach for constructing VLMs through LLM contextualization with multimodal pre-trained user models. The proposed model is not merely beneficial but essential for effective human-robot interactions that inherently require multimodal understanding—the ability to perceive, interpret, and respond to human visual cues, gestures, and verbal communication simultaneously. While the User-VLM model shows promise in various applications, it must be embedded within broader frameworks incorporating comprehensive safeguards to address various challenges crucial for generating safe and ethically sound personalized responses.

Introduction

Ensuring a safe and intuitive interaction between humans and robots requires AI systems that dynamically perceive and adapt to individual needs, behaviors, and preferences. This adaptability is crucial, as it enables robots to navigate complex social dynamics and establish meaningful connections that respect human cognitive and emotional boundaries (Romeo et al. 2022; Frith and Frith 2005). Such capabilities are particularly important in sensitive domains like healthcare and education, where tailored responses enhance both user safety and engagement (Cavallini et al. 2021; Kristen and Sodian 2014). User modeling, which encompasses methodologies for capturing and representing user features and personal characteristics, serves as a fundamental component in creating these adaptive systems (Purificato, Boratto, and De Luca 2024).

Large Language Models (LLMs) research has demonstrated significant success across a spectrum of downstream tasks in recent years (Zhao et al. 2023). The better contextualization of LLMs with user models has sparked significant efforts for improved human-robot interactions. While approaches like User-LLM (Ning et al. 2024) demonstrate promising directions for scalable and privacy-preserving personalized AI systems by integrating user embeddings with generative language models, their applications remain limited in social robotics contexts, where interactions inherently require multimodal understanding - the ability to perceive, interpret, and respond to human visual cues, gestures, and verbal communication simultaneously (Nocentini et al. 2019). This limitation highlights a crucial gap in current user modeling approaches for social robotics applications, where multimodal adaptation is not merely beneficial but essential for natural and effective human-robot interaction.

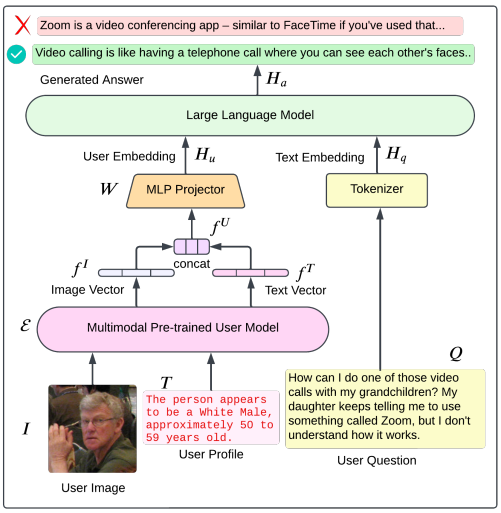

Figure 1: Proposed Architecture for User-VLM

This paper leverages Multimodal Pre-trained Models (Zhang et al. 2024), such as FaRL (Facial Representation Learning) (Zheng et al. 2022), and proposes a novel method for LLM contextualization, as shown in Figure 1, enabling a richer understanding of user characteristics by incorporating both visual and textual dimensions into the language model's context processing.

Methods

As shown in Figure 1, the process operates on a dataset $\mathcal{D} = \{\{(I_k, T_k, Q_{k,i}, A_{k,i})\}_{i=1}^{N}\}_{k=1}^{M}$ comprising N paired instances of questions and answers for M paired instances of visual and textual profiles, where each question $Q_{k,i} \in \mathbb{R}^{d_Q \times N}$, answer $A_{k,i} \in \mathbb{R}^{d_A \times N}$, image $I_k \in \mathbb{R}^{d_I \times M}$, and text $T_k \in \mathbb{R}^{d_T \times M}$ are represented as sequences of tokens. The indices N and M denote the sequential lengths of question-answers pairs and image-text user profiles re-

spectively, while d_Q, d_A, d_I, and d_T represent their corresponding feature dimensions. The proposed model is an adaptation of the LLava model (Liu et al. 2024), consisting of an encoder transformer and an LLM for general-purpose visual and language understanding. The encoder transformer $\mathcal{E}$ is a multimodal pre-trained user model that encodes user profiles- images I_k and text T_k- into a user representation $\mathbf{U}_k \in \mathbb{R}^{d_U}$. The LLM is a decoder transformer that generates text tokens $\mathbf{y} = \{y_1, y_2, \ldots, y_L\}$ based on the question $\mathbf{Q}_i \in \mathbb{R}^{d_Q}$ and the user vector $\mathbf{U}_k$ produced by the encoder, where L is the length of the generated sequence.

Multimodal Pre-trained User Model

As shown in Figure 1, given a user entry $U = (I, T)$, the Multimodal Pre-trained User Model employs two primary encoder functions: an image encoder $\mathcal{E}_I : \mathbb{R}^{d_I \times N} \rightarrow \mathbb{R}^{d_z \times N}$ and a text encoder $\mathcal{E}_T : \mathbb{R}^{d_T \times M} \rightarrow \mathbb{R}^{d_z \times M}$, where d_z denotes the hidden dimension. These Transformer-based encoders process their respective modalities to produce sequences of feature vectors $\mathcal{E}_I(I) = \{f_1^I, f_2^I, \ldots, f_N^I\}$ and $\mathcal{E}_T(T) = \{f_1^T, f_2^T, \ldots, f_M^T\}$. The image vector ($f^I$) and text vector ($f^T$) are concatenated to form $f^U = \left[f^I; f^T\right]$ with a dimension of $2 \times z$. This concatenated vector is processed through a projection head $P : \mathbb{R}^{d_2 \times z} \rightarrow \mathbb{R}^{d_h}$, implemented as a multilayer perceptron, which maps f^U into the language embedding space. Specifically, a trainable projection matrix W is applied to transform f^U into the user embedding vector H_u, with the same dimensionality as the word embedding space in the language model: $H_u = W \cdot f^U$.

Large Language Model

For the LLM, we consider selecting the Llama model $\xi_\phi(\cdot)$ parameterized by ϕ, whose checkpoints are publicly available and has widely adopted in Llava-based architectures for its performance and generalizability (Touvron et al. 2023). We utilize this model and consider the grid features before and after the last transformer layer. In this regard, we simply consider a linear layer that connects user features H_u into the text embedding space H_q forming the input for the LLM to carry out subsequent predictions. Given the input question Q and answer A, a word embedding matrix is used to map them to contextual embeddings H_q and H_a, and the distribution over $H_a^{(i+1)}$ can be obtained following the auto-regressive model as:

$$
\begin{aligned}
&p_\theta\left(H_a^{(i+1)} \mid H_u, H_q, H_a^{(1:i)}\right) \\
&\quad = \sigma\left(\xi(H_v, H_q, H_a^{(1:i)})\right),
\end{aligned}
\tag{1}
$$

where θ represents all the trainable parameters in the LLM, $\sigma(\cdot)$ is a `softmax` function, and $\xi(\cdot)$ outputs the logits (before applying `softmax`) over the vocabulary for the last position of the sequence. We denote p_θ as the prediction probability for the anticipated answer token $H_a^{(i+1)}$ at the position $i+1$, conditioned on the input user token embeddings H_u, the question token embeddings H_q, and the previous answer token embeddings $H_a^{(1:i)}$. The logits are passed through $\sigma(\cdot)$ to compute the probability distribution over all tokens in the vocabulary, and the most probable token is typically selected using `argmax`.

Discussion

During both the training and post-deployment phases of User-VLM, a range of challenges arise that are pivotal to address, given the profound impact of user modeling on the dynamics of human interaction with social robotics.

Technical Challenges

The preparation of data for training user-adaptive language models is a nuanced and critical process, as the dataset must embody diversity and impartiality to ensure balanced, inclusive, and non-discriminatory question-answering capabilities across a wide range of user demographics (Chen et al. 2024). Equally challenging is the determination of optimal parameterization and fine-tuning strategies for multimodal pre-trained user models, which necessitates systematic experimentation (Wang et al. 2023). The interdependence of these strategies on the underlying datasets further complicates this endeavor and warrants thorough investigation (He et al. 2024). Finally, the evaluation of these models introduces critical challenges, as traditional performance metrics alone are insufficient. Instead, comprehensive benchmarks must also assess the models from clinical and psychological perspectives to ensure robust and ethically sound user adaptations (Park et al. 2024).

Ethical Issues

Post-deployment, several ethical considerations remain critical in the application of the proposed User-VLM (Jafari and Vassileva 2023). A key concern is ensuring that personalized responses are provided only when the model has reliably aligned its assumed user profile with the actual characteristics of the user and obtained explicit consent to tailor responses accordingly. The utility of the proposed model is contingent upon strict adherence to these principles. We contend, however, that the User-VLM, in its current form, cannot inherently address all ethical challenges associated with user modeling and personalized interactions. Nonetheless, we propose that this model can serve as a foundational component within broader frameworks that integrate comprehensive ethical safeguards (Li et al. 2024).

Conclusion

This paper proposes User-VLM, a novel approach for forming VLMs through LLM contextualization with multimodal pre-trained user models. The integration of multimodal user models with LLMs presents both technical and ethical challenges that are crucial to address. Preparing diverse and inclusive datasets, optimizing parameterization strategies, and ensuring ethical considerations such as user consent and alignment are pivotal. While the User-VLM model shows promise, it must be embedded within broader frameworks that include comprehensive ethical safeguards to generate ethically sound personalized responses.

Acknowledgments

The authors gratefully acknowledge the French National Research Agency (ANR) for its financial support of the ANITA project (Grant No. ANR-22-CE38-0012-01).

References

Cavallini, E.; Ceccato, I.; Bertoglio, S.; Francescani, A.; Vigato, F.; Ianes, A. B.; and Lecce, S. 2021. Can theory of mind of healthy older adults living in a nursing home be improved? A randomized controlled trial. *Aging Clinical and Experimental Research*, 33: 3029–3037.

Chen, H.; Waheed, A.; Li, X.; Wang, Y.; Wang, J.; Raj, B.; and Abdin, M. I. 2024. On the Diversity of Synthetic Data and its Impact on Training Large Language Models. *arXiv preprint arXiv:2410.15226*.

Frith, C.; and Frith, U. 2005. Theory of mind. *Current biology*, 15(17): R644–R645.

He, J.; Li, P.; Liu, G.; Zhao, Z.; and Zhong, S. 2024. Pefomed: Parameter efficient fine-tuning on multimodal large language models for medical visual question answering. *arXiv preprint arXiv:2401.02797*.

Jafari, E.; and Vassileva, J. 2023. Ethical issues in explanations of personalized recommender systems. In *Adjunct Proceedings of the 31st ACM Conference on User Modeling, Adaptation and Personalization*, 215–219.

Kristen, S.; and Sodian, B. 2014. Theory of Mind (ToM) in early education. *Contemporary perspectives on research in theory of mind in early childhood education*, 291–320.

Li, C.; Wu, G.; Chan, G. Y.-Y.; Turakhia, D. G.; Quispe, S. C.; Li, D.; Welch, L.; Silva, C.; and Qian, J. 2024. Satori: Towards Proactive AR Assistant with Belief-Desire-Intention User Modeling. *arXiv preprint arXiv:2410.16668*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Ning, L.; Liu, L.; Wu, J.; Wu, N.; Berlowitz, D.; Prakash, S.; Green, B.; O'Banion, S.; and Xie, J. 2024. User-LLM: Efficient LLM Contextualization with User Embeddings. *arXiv preprint arXiv:2402.13598*.

Nocentini, O.; Fiorini, L.; Acerbi, G.; Sorrentino, A.; Mancioppi, G.; and Cavallo, F. 2019. A survey of behavioral models for social robots. *Robotics*, 8(3): 54.

Park, J. I.; Abbasian, M.; Azimi, I.; Bounds, D.; Jun, A.; Han, J.; McCarron, R.; Borelli, J.; Li, J.; Mahmoudi, M.; et al. 2024. Building trust in mental health chatbots: safety metrics and LLM-based evaluation tools. *arXiv preprint arXiv:2408.04650*.

Purificato, E.; Boratto, L.; and De Luca, E. W. 2024. User Modeling and User Profiling: A Comprehensive Survey. *arXiv preprint arXiv:2402.09660*.

Romeo, M.; McKenna, P. E.; Robb, D. A.; Rajendran, G.; Nesset, B.; Cangelosi, A.; and Hastie, H. 2022. Exploring theory of mind for human-robot collaboration. In *2022 31st IEEE International Conference on Robot and Human Interactive Communication (RO-MAN)*, 461–468. IEEE.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Wang, H.; Yang, X.; Chang, J.; Jin, D.; Sun, J.; Zhang, S.; Luo, X.; and Tian, Q. 2023. Parameter-efficient tuning of large-scale multimodal foundation model. *Advances in Neural Information Processing Systems*, 36: 15752–15774.

Zhang, D.; Yu, Y.; Li, C.; Dong, J.; Su, D.; Chu, C.; and Yu, D. 2024. Mm-llms: Recent advances in multimodal large language models. *arXiv preprint arXiv:2401.13601*.

Zhao, W. X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. 2023. A survey of large language models. *arXiv preprint arXiv:2303.18223*.

Zheng, Y.; Yang, H.; Zhang, T.; Bao, J.; Chen, D.; Huang, Y.; Yuan, L.; Chen, D.; Zeng, M.; and Wen, F. 2022. General facial representation learning in a visual-linguistic manner. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 18697–18709.