Bridging Multimedia Modalities: Enhanced Multimodal AI Understanding and Intelligent Agents

Sushant Gautam
Simula Metropolitan Center for Digital Engineering (SimulaMet), Norway
sushant@simula.no

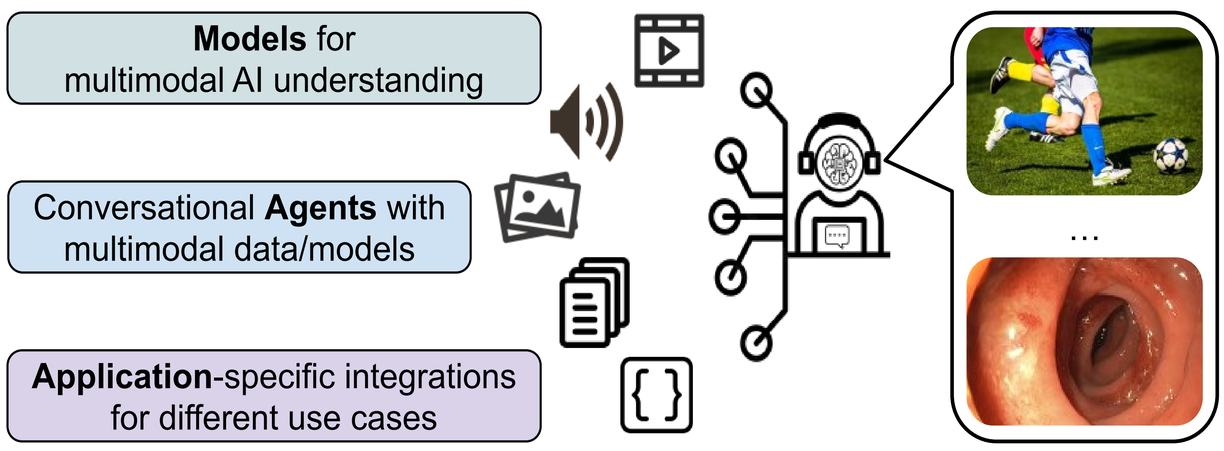

Figure 1: Overview of the proposed research.

Motivation

With AI becoming increasingly prevalent in everyday life, applications of conversational agents (chatbots) are rapidly expanding. While modern chatbots can handle multiple media modalities such as text, images, and audio, they can be improved in terms of **contextual capture** and **domain specificity** (e.g., exploiting audio-visual cues for subtle game details in soccer), with **multimodal fusion**, in order to give **accurate responses** in human interactions.

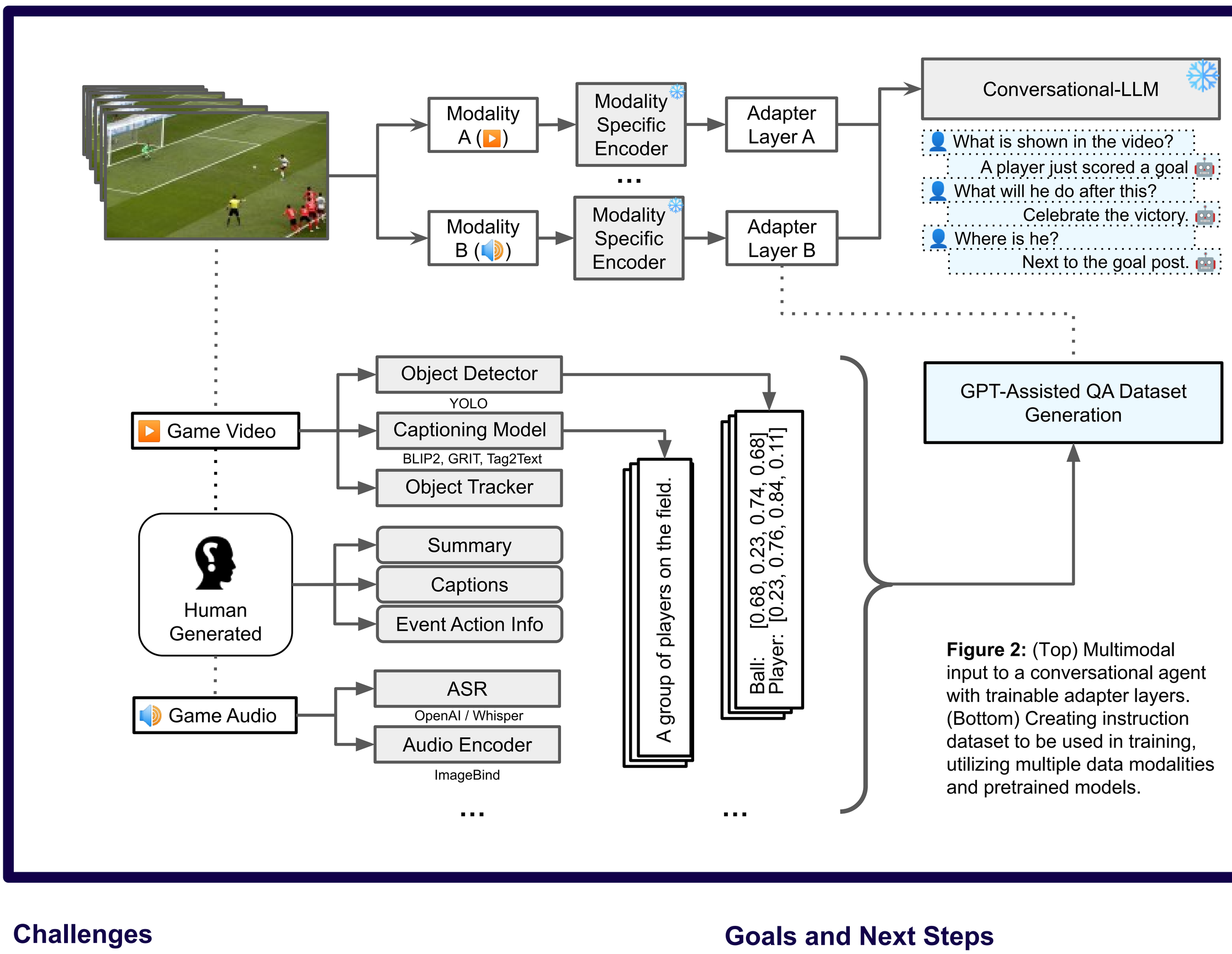

Figure 2: (Top) Multimodal input to a conversational agent with trainable adapter layers. (Bottom) Creating instruction dataset to be used in training, utilizing multiple data modalities and pretrained models.

Challenges

- Domain-specific **multimodal dataset curation**
- Multimodal **evaluation metrics** to measure performance
- Abilities and capabilities of **pretrained models**
- Proper **representation** of different modalities
- Multi-modal **alignment and fusion**

Goals and Next Steps

- Enriched multimodal **dataset curation**
- Increased **context-awareness**
- Elevated **conversation quality** and **user experience**
- Continued **interdisciplinary** collaboration
- Extended **applications** (sports, medical, etc.)

References

- Liang, Paul Pu, et al. "Foundations and recent trends in multimodal machine learning: Principles, challenges, and open questions." (2022)
- Maaz, Muhammad, et al. "Video-ChatGPT: Towards Detailed Video Understanding via Large Vision and Language Models." (2023).
- Zhang, Hang et al. "Video-llama: An instruction-tuned audio-visual language model for video understanding." (2023).
- Liu, Haotian, et al. "Visual instruction tuning." (2023).