OLViT: Multi-Modal State Tracking via Attention Based Embeddings for Video-Grounded Dialog

University of Stuttgart
Germany

PUI Perceptual User Interfaces

Adnen Abdessaied, Manuel Hochmeister, Andreas Bulling

Contact: adnen.abdessaied@vis.uni-stuttgart.de

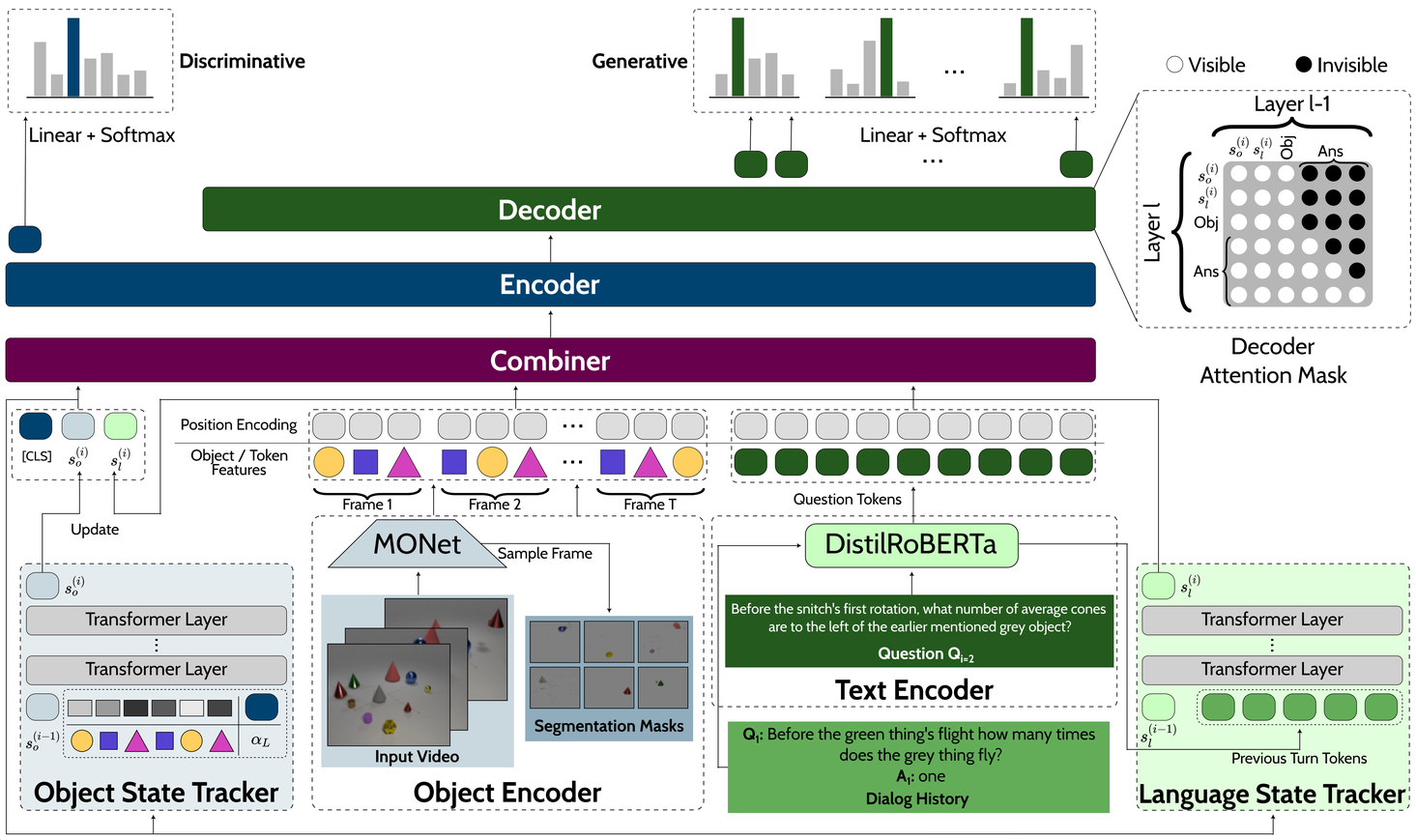
Introduction

- **Task:** Answer a question given a video and a dialog history
- **Limitations:** Prior works were trained on biased datasets that do not test higher order reasoning capabilities
- **Key insights: Performing multi-modal state tracking alleviates the limitations of previous works and improve their reasoning capabilities**

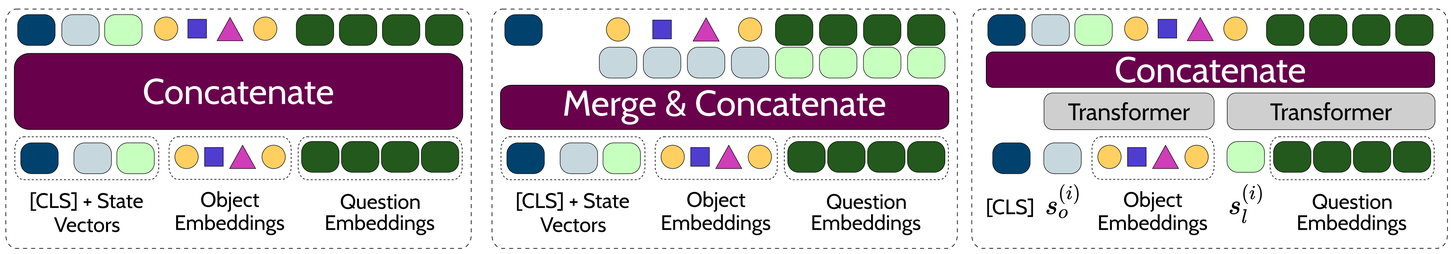
Contributions

- A novel multi-modal state tracking technique that can be seamlessly integrated into pre-trained LLMs
- Two seperate attention-based trackers for object and language state tracking
- New state-of-the-art results on DVD and SIMMC datasets

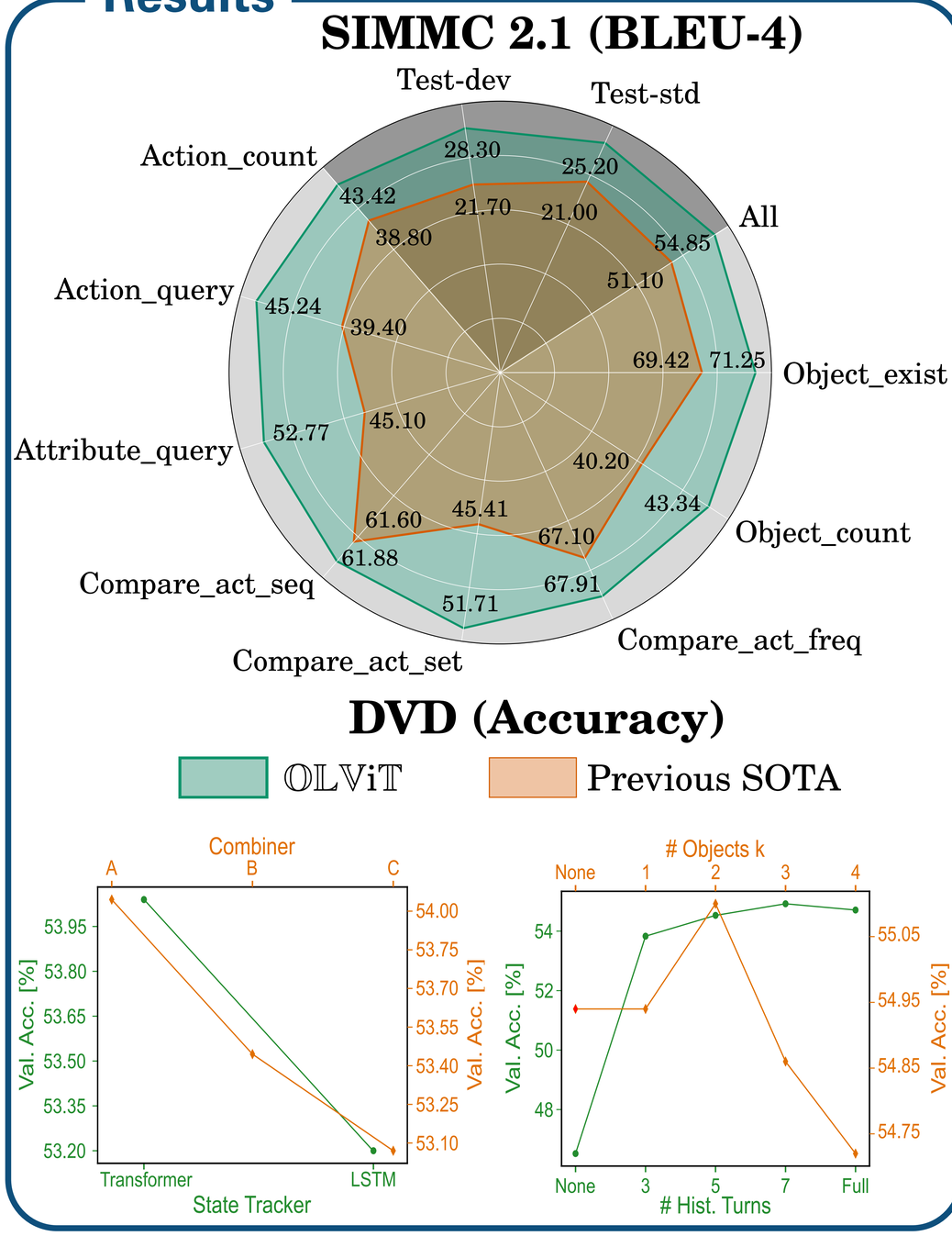
Method



Combiner



Results



Acknowledgments

This work received funding from the European Research Council (ERC; grant agreement 801708)