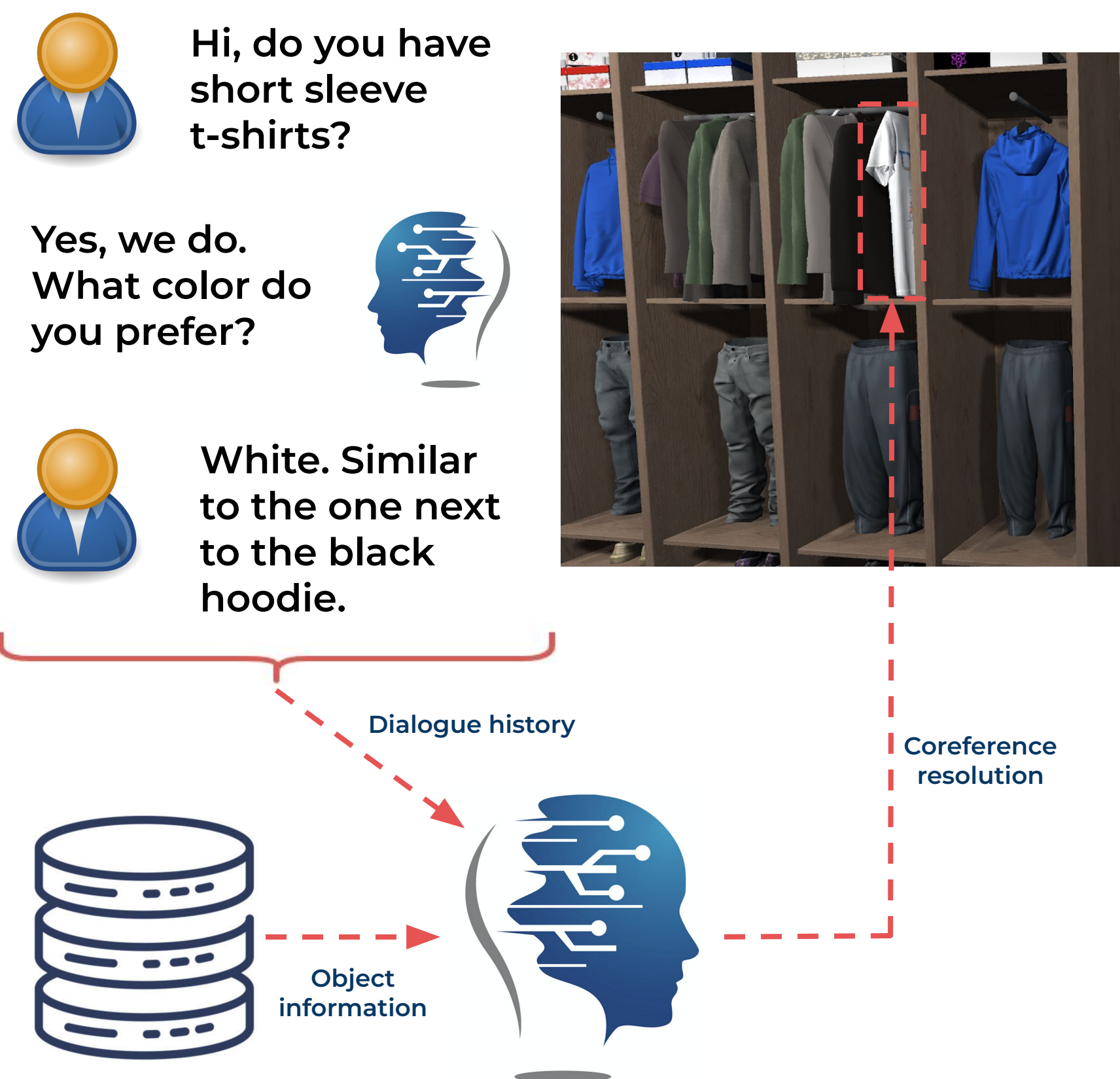


## Introduction

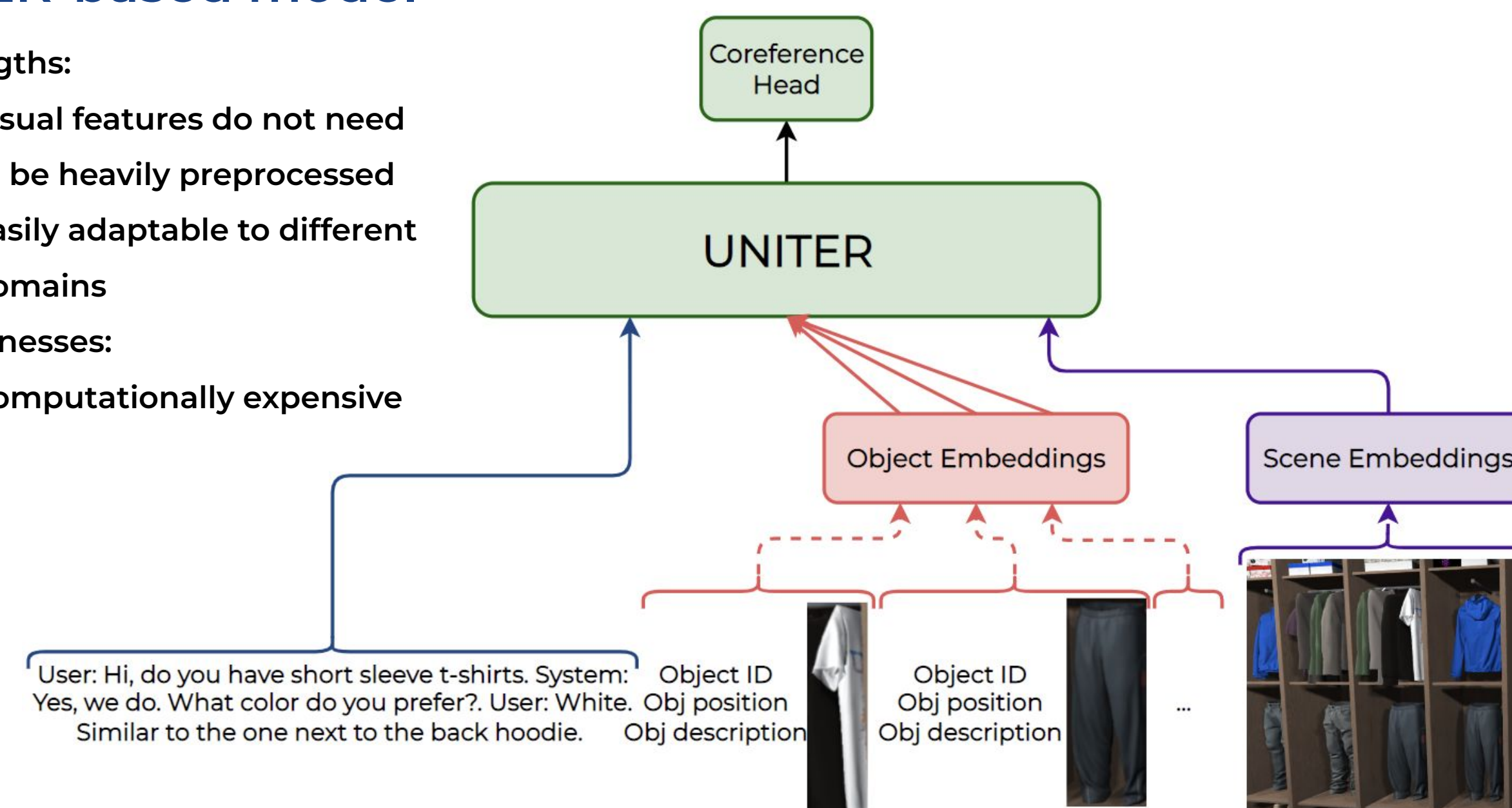


- Applications:
  - E-commerce virtual assistant: answer customer inquiries about objects.
  - In-site interpreter: improved translations using both scene and textual context.
  - Boost other natural language tasks, like question answering or generation.
- SIMMC2 dataset published by Facebook Research is used for investigation.
  - It contains dialogues, object descriptions and scene images.
- The 10th Dialog System Technology Challenge (DSTC10) partially focused on the multimodal coreference resolution task.
  - SIMMC2 dataset was used for the competition.
  - Best performing systems are studied as a enhanced baseline.

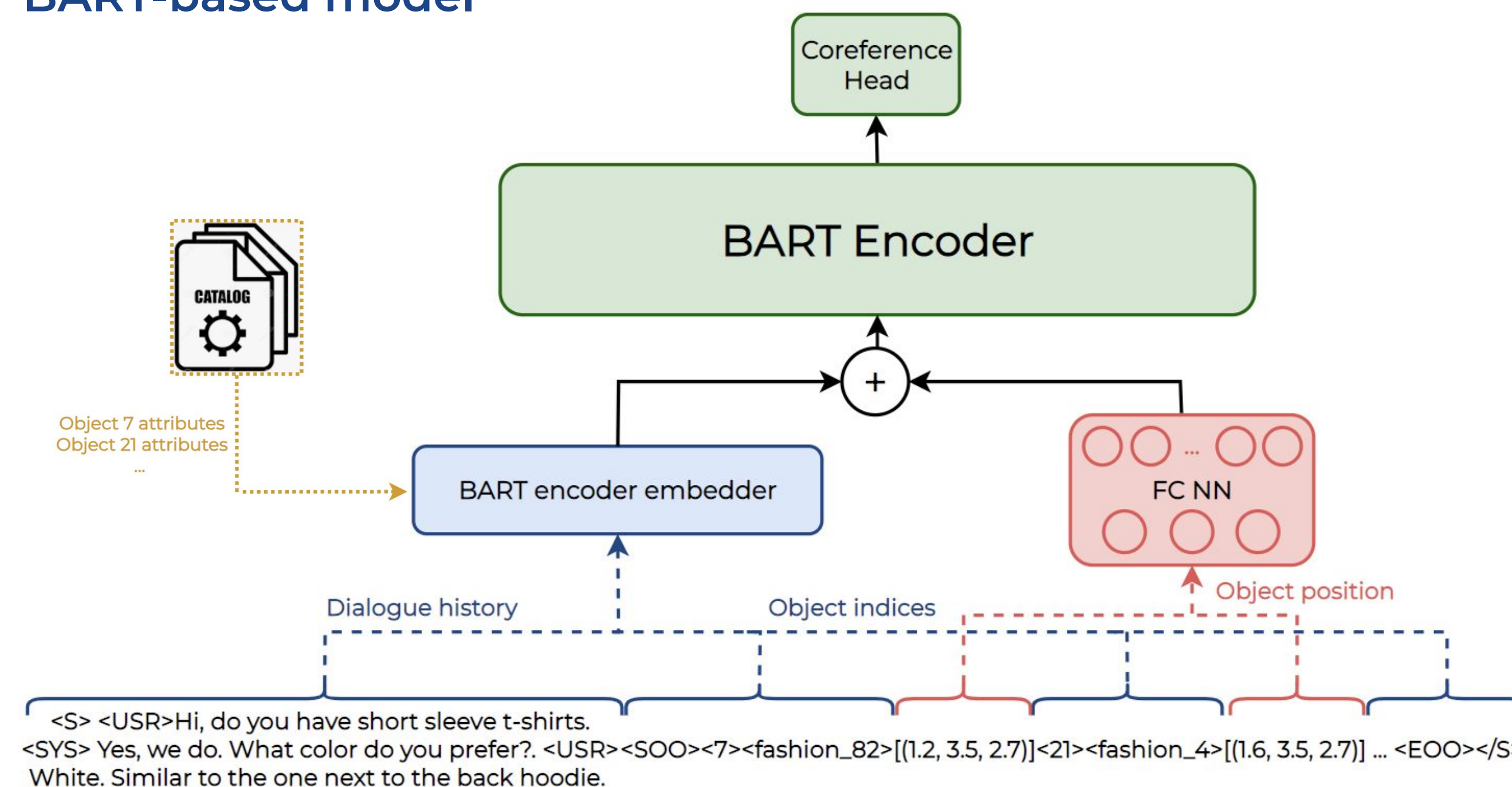
## SOTA MMCR Systems

### UNITER-based model

- Strengths:
  - Visual features do not need to be heavily preprocessed
  - Easily adaptable to different domains
- Weaknesses:
  - Computationally expensive



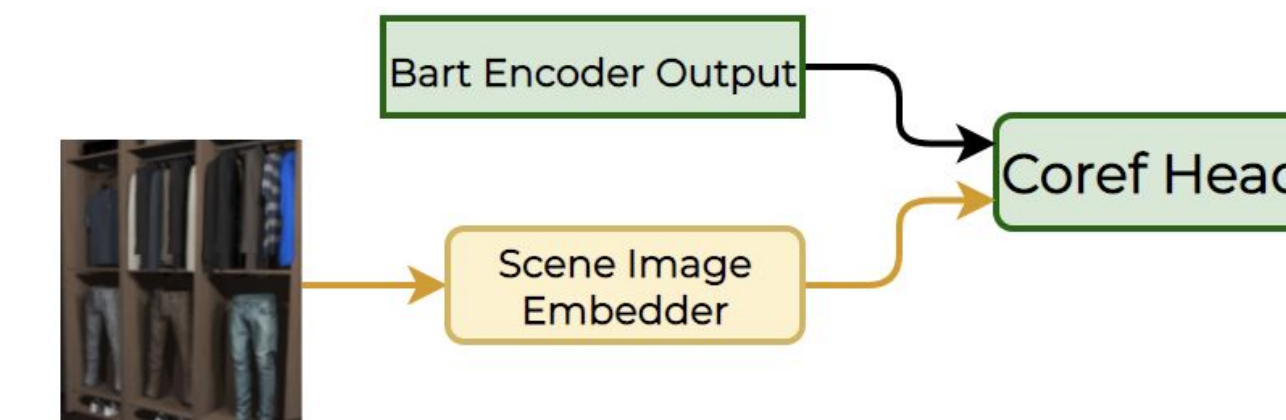
### BART-based model



- Strengths:
  - One of the best solutions for coreference resolution task.
  - Winner of DSTC10 on this task.
- Weaknesses:
  - Bad at handling objects not seen in training.
  - Scene images need to be described in natural language to be used.

## Proposed improvements

- Include object descriptions in the input of the BART-based model. ✓
- Provide image embeddings to improve the coreference head of the BART-based model. ⌚



- Suppress object IDs in UNITER-based model to make it scene-independent. ⌚

## Results

Model	Object F1-Score
GPT-2 Baseline (Facebook Research)	36.6%
UNITER-based (New York Uni. Shanghai)	67.4%
BART-based (KAIST & Samsung Research)	74.3%
<b>BART using object descriptions (Ours)</b>	<b>76.1%</b>

Multimodal Coreference Resolution performance on devtest split

## References

- [1] Satwik Kottur et.al. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. *Association for Computational Linguistics*. 2021.
- [2] Yichen Huang et. al. UNITER-Based Situated Coreference Resolution with Rich Multimodal Input. *Computing Research Repository*. 2021.
- [3] Haeju Lee et. al. Tackling Situated Multi-Modal Task-Oriented Dialogs with a Single Transformer Model. *Association for Computational Linguistics*. 2021

## Acknowledgments

Project carried out under the supervision of Kate Knill of the Engineering Department of the University of Cambridge in collaboration with Svetlana Stoyanchev, Rama Doddipatla and Simon Keizer of the Dialog group at Toshiba Europe Ltd.