# Enhancing Augmented Reality Dialogue Systems with Multi-Modal Referential Information

1st Ziyao He
*MOE KLINNS Lab*
*Xi'an Jiaotong University*
Xi'an, China
marsroscope@stu.xjtu.edu.cn

2nd Zhongmin Cai*
*MOE KLINNS Lab*
*Xi'an Jiaotong University*
Xi'an, China
zmcai@mail.xjtu.edu.cn

***Abstract*—In this paper, we present a novel approach to advancing augmented reality (AR) dialogue systems, bridging the gap between two-dimensional spaces and immersive virtual environments. We construct the "SIMMC2-Point" dataset, which transforms the original SIMMC2 dataset from virtual reality (VR) into AR environments, highlighting the additional introduced pointing modality to support understanding user's multi-modal intentions in AR. By harnessing the power of BART and CLIP models, we design the architecture of multi-modal dialogues that effectively captures spatial and attribute information. Then, a series of ablation experiments based on our designed SIMMC2-Point dataset and dialogue models underscores the significance of the pointing modality in enhancing the performance of dialogue systems across various tasks. Our work represents a crucial step forward in AR dialogue systems, facilitating seamless interactions within immersive virtual environments.**

***Index Terms*—Augmented Reality, Chatbot, SIMMC2-Point Dataset**

## I. Introduction

The advent of machine learning and deep learning techniques has ushered in a new era of technological advancements in various domains, including computer vision, natural language processing, and human-computer interaction. Augmented reality (AR) technology, which amalgamates computer vision, sensor technology, and multi-modal human-computer interaction, has garnered widespread popularity, enabling users to interact with three-dimensional virtual worlds, delivering highly immersive experiences. Unlike conventional interaction devices, AR allows real-time interaction with virtual objects through bodily gestures, offering a more natural and human-centric experience.

Despite the progress in large language models and AR devices, current dialogue systems remain confined to two-dimensional spaces, predominantly within smartphone or web environments, supporting only text, images, and audio. However, the SIMMC dataset [1] series has successfully extended dialogue systems to three-dimensional spaces by simulating shopping processes in virtual reality (VR) environments. This dataset enables users to communicate shopping needs to dialogue agents, which act as shopping assistants that provide product recommendations, detailed information, and shopping cart assistance. While this dataset significantly enhances dialogue agent training, there is a critical gap in similar datasets for AR environments, making the development of AR dialogue datasets a crucial research direction.

Moreover, the refined SIMMC dataset should also demonstrate the presence of other user-related modalities that AR/VR environments boast exclusively, which include various behaviors such as pointing behaviors. Incorporating additional multi-modal context can greatly aid dialogue systems in understanding user intentions and providing accurate responses. For instance, gesture interactions can be used to introduce referential information, helping the dialogue system clarify referred content without complex descriptive statements. Investigating the impact of the referential modality on dialogue systems is therefore essential for improving their overall performance.

To address the above concern, this paper presents three significant contributions:

- **Revamped SIMMC2 Dataset:** We migrated and transformed the SIMMC2 dataset from VR to AR environments, introducing additional pointing modality, and resulting in the creation of the SIMMC2-Point dataset. To ensure the dataset's validity, we conducted meticulous manual evaluations.
- **Multi-Modal Dialogue Modeling and Evaluation:** To validate the effectiveness of our dataset modifications, we leverage the BART model as the base and the CLIP model for visual attribute encoding, we employ various auxiliary tasks to enable the model to grasp relationships between object IDs, attributes, and spatial positions, empowering the model to represent spatial and attribute information more effectively. We evaluate the model regarding the aspect of multi-modal coreference resolution, multi-modal dialogue state tracking, and multi-modal response generation.
- **Study of the Impact of the Various Modality:** To ascertain the role of the newly added pointing modality alongside other modalities, we conducted insightful ablation experiments. By introducing the referential modality into different models and dialogue tasks, we evaluate its impact on dialogue system performance. The results demonstrate its crucial role in enhancing performance across most tasks.

*Corresponding author.

2023 China Automation Congress (CAC) | 979-8-3503-0375-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/CAC59555.2023.10450983

## II. Related Work

### A. Multi-Modal Interaction in AR Environment

AR applications benefit from integrating diverse multi-modal human-computer interaction technology. The data acquisition module captures user behaviors and physiological data, providing comprehensive information for agent training [2], [3]. Various techniques like touch, gestures, gaze, controllers, and handheld projectors are used for human-computer interaction in AR [4]. Multi-modal information, including haptic, auditory, and visual cues, enhances chemical experiment simulations, fostering students' practical skills while minimizing potential hazards [5]. The incorporation of virtual objects aids patient rehabilitation, increasing task engagement [6]. AR is explored through gestures, speech, eye gaze, and other modes [7], addressing the challenges of multi-modal fusion [8]. Applications include shopping systems and games [9], [10]. Gesture and speech interaction improves user productivity and experience [11], and adaptive and multi-modal interaction gain attention [12]. Studies examine hand-eye behavior-triggered menu selection [13]. Notably, Microsoft's HoloLens headset advances multi-modal interaction research [14], used in areas like anti-aircraft command, control systems, and HoloHome smart home interaction [15]. Wearable devices capture user motion information to adjust AR content display [16]. Our work uses HoloLens 2 to realize multi-modal interaction, fusing virtual and real worlds, and delivering an immersive interactive experience.

### B. Multi-Modal Dialogue Systems

Dialogue systems, relying solely on voice or text, face limitations where human communications in AR usually involve multiple senses to form multi-modal information that enhances the points of ideas. Researchers introduce modalities and datasets like IGC [17], Visual7W [18], TVQA [19], MovieQA [20], and SIMMC [1], [21], laying the foundation for training multi-modal dialogue systems.

Efficient dialogue implementation addresses three challenges: Multi-modal Coreference Resolution (MCR), Multi-modal Dialogue State Tracking (MDST), and Multi-modal Dialogue Response Generation (MDRG). MCR determines referent objects in user utterances. Huang et al. [22] use the SIMMC2 dataset to predict object mentions in a rich multi-modal context. Guo et al. [23] propose GRAVL-BERT, combining graph neural networks and VL-BERT [24] for referent object prediction. MDST predicts user intents and slot values. Le et al. [25] introduce the Video-Dialogue Transformer Network based on the Transformer [26] model. MDRG involves generating meaningful responses, with approaches like Seq2Seq-based systems [27] and personalized response generation [28]. Works utilize large language models, e.g., GPT-2 [29] and BART [30]. Lee et al. [31] present a BART-based joint learning approach for coreference resolution, dialogue state tracking, and response generation.

Our work draws upon Lee et al.'s [31] methodology, achieving multi-modal joint representation by incorporating datasets with modal references during training.

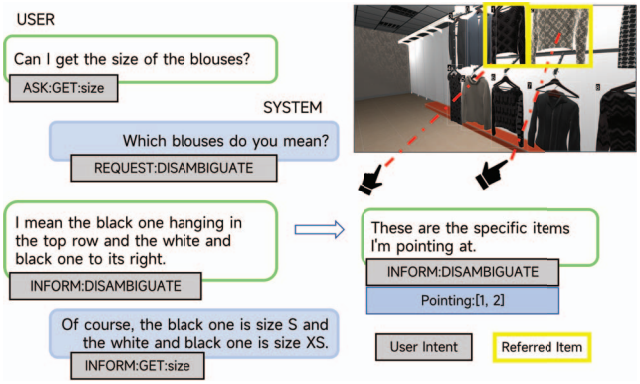


Fig. 1: Example from SIMMC2-Point Dataset

## III. The SIMMC2-Point Dataset

This study demonstrates the dialogue system's capabilities in AR scenes, using a shopping scenario as the background. The AI shop assistant, within the AR scene, provides relevant item recommendations, detailed product information, and understands ambiguous user expressions. This enables immersive shopping, as all objects possess known attributes and spatial information.

### A. The Overview of SIMMC2 Dataset

The SIMMC2 dataset [1], proposed by Meta Research, trains multi-modal dialogue agents in VR scenes but not in AR scenarios. In the SIMMC2 dataset, users and dialogue systems engage in conversations within the same scene, considering the user's location to adjust spatial relationships. For instance, in a shopping scenario, the system recommends jackets relative to the user's location for authentic responses. Annotations include user intents and object references, with JSON files containing object positions and camera data.

The SIMMC2 dataset comprises 9,557 dialogues covering clothing and furniture. Objects have unique absolute and relative IDs, with various attributes tokenized for dialogue state tracking. User intents are categorized into higher-level dialogue acts (INFORM, CONFIRM, REQUEST, ASK) and lower-level acts (GET, DISAMBIGUATION, REFINE, ADD_TO_CART, COMPARE).

### B. Building SIMMC-Pointing Dataset for AR Environment

This study aims to employ the SIMMC2 dataset in AR scenarios, utilizing HoloLens2's robust user behavior data detection capabilities in AR scenes. Additional input modalities are introduced to remove ambiguity in dialogue, thereby enhancing dialogue task metrics. Specifically, user finger-pointing information is considered a valid form of reference. HoloLens2 detects user hand finger pointing rays and collision information with virtual objects, providing their three-dimensional coordinates and attributes. Consequently, we represent user pointing information using the relative IDs of the collided objects. To incorporate pointing data into the SIMMC2 dataset, selected dialogues from the original dataset are filtered and replicated, simulating scenarios where users

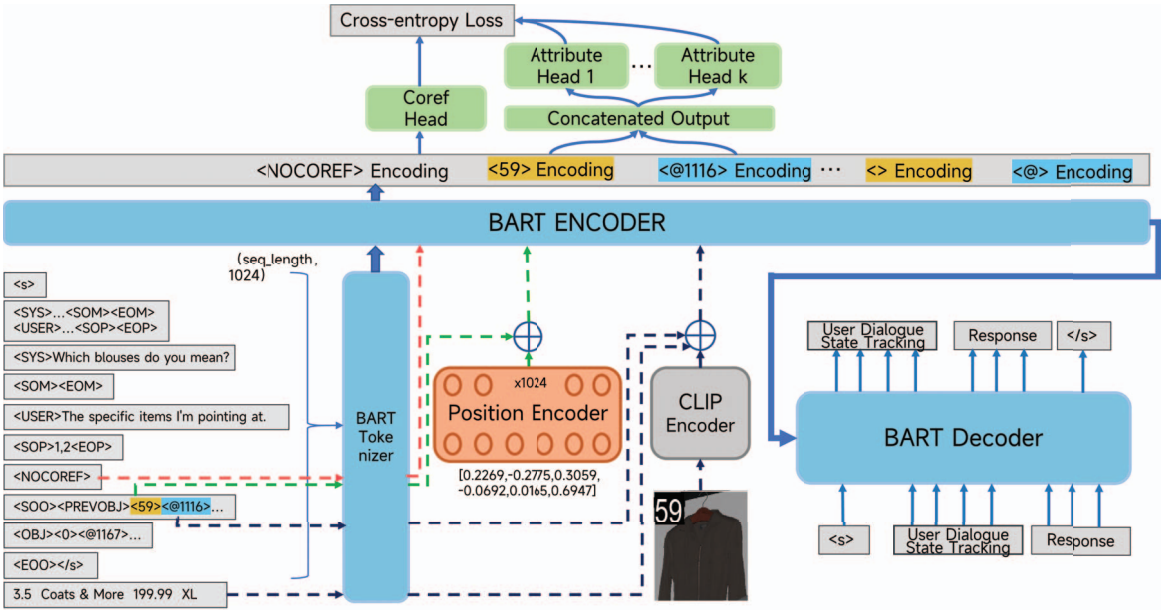


Fig. 2: Model Architecture for Multi-Modal Dialogues

clarify misunderstood expressions using pointing information, as shown in Fig.1. This approach simplifies the explanation process for the dialogue system, as it can directly point to the referenced object. However, the original user dialogue text with descriptions of related objects beyond the one directly pointed to is retained during dataset modification, maintaining multiple input modalities that complement each other when fully expressing the user's intent. A total of 1,593 user statements were augmented with pointing modality data. The SIMMC2-Point dataset was constructed with 3,792 modified statements in the training set, 366 in the validation set, and 855 in the test set.

All modifications adhered to specific principles: (1) including key elements from the original sentence; (2) expressing the same intent as the original sentence; and (3) aligning with the ongoing dialogue context. Three experts with dataset collection experience evaluated the modifications using a Likert scale. The average scores for key item level, current user input level, and dialogue context level were 4.78 (SD=0.52), 4.57 (SD=0.75), and 4.14 (SD=1.04), respectively. The SIMMC2-Point dataset exhibited strong performance, receiving average scores above 4 on all levels.

## IV. Multi-Modal Dialogue Modeling

This paper adapts an end-to-end joint training approach to address error accumulation in pipeline-style training, hindering iterative fine-tuning. The approach employs an encoder-decoder architecture for user inputs, generating results for multiple tasks. Tasks are interdependent, optimizing performance collectively.

The model architecture (As in Fig.2) includes dialogue history, current user input, and scene information. Dialogue history contains user utterances, references, system statements, and mentioned objects (context length: 2). Scene information uses <PREVOBJ> for previous and <OBJ> for current objects. Lee's work [31] guides object representation using relative and absolute IDs.

Visual attributes use CLIP model [32]. ViT-B/32 encoder transforms scene image blocks into 512-dimensional vectors, aligning visual features with textual space. Non-visual attributes learned through auxiliary tasks are added to relative and absolute ID encoding.

The decoder uses the encoder's output for Multi-Modal Coreference Resolution (MCR), Multi-Modal Dialogue State Tracking (MDST), and Multi-Modal Dialogue Response Generation (MDRG) tasks.

To incorporate scene context, relative and absolute IDs are added to the tokenizer. Spatial location information enhances relative ID encoding; visual and non-visual metadata attributes improve absolute ID encoding. Auxiliary tasks fine-tune token encodings for spatial and object attribute representation. One task detects referring objects using <NOCOREF> token and a fully connected layer. Another task extracts object attribute values. Pre-training ensures proper encoding representation for object features (400 steps, cross-entropy loss).

These tasks can be modeled as binary or multiclass classification problems, requiring the computation of cross-entropy loss function as shown in Equation 1. Assuming a sample size of N and K classes, $y_{ik}$ represents the true 0-1 label, and $p_{ik}$ denotes the predicted probability for class k.

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{k=1}^{K} y_{ik} log(p_{ik}) \quad (1)$$

Additionally, the language model has its own loss, as depicted in Equation 2, where L denotes the sequence length and $t_i$ represents the i-th token. During the prediction process,

the language model predicts the distribution of the next token based on the preceding tokens.

$$L_{LM} = -\sum_{i=1}^{L} logP(t_i|t_1, ..., t_{i-1}) \quad (2)$$

With auxiliary tasks' assistance, the encoder captures semantic and spatial information in the dialogue text and user references. By optimizing the loss functions through gradient descent, the model achieves consistent multi-modal information encoding and spatial understanding capabilities. The unified encoding of the BART model, along with multiple auxiliary tasks, facilitates joint training for referential resolution, dialogue state tracking, and response generation under multi-modal inputs.

## V. Experiment

In order to investigate whether incorporating pointing modality can enhance the performance of a dialogue system across various tasks involving dialogue text and spatial modalities, this study is based on a modified version of the SIMMC2-Point dataset. Deep language models such as GPT, GPT-2, and BART are employed to model various tasks in the dialogue system. Subsequently, ablation experiments are conducted to examine the roles played by each modality. The experiments consist of the following three steps:

### A. Dataset Preprocessing

The original SIMMC2-Point dataset, containing dialogue and modality data, cannot be directly used for prediction. Therefore, the data is formatted into sequences of tokens that can be accepted by the encoder of deep learning models. To accommodate different forms of multi-modal context, the various multi-modal data are isolated using separators. Additionally, the annotation data for each dialogue task is formatted to be compatible with the output format required by the model's decoder. Consequently, the output includes dialogue states, referential content, and system responses, represented `act[slot values](request_slots)<objects><EOB>system utterance<EOS>`

### B. Model Inference

This study utilizes an improved version of the BART model, along with basic deep pre-trained language models such as GPT and GPT-2, all featuring Transformer decoder structures. These models are capable of predicting results for multi-modal referring expression resolution, dialogue state tracking, and dialogue response generation. In the BART model, visual attributes of images are extracted using the CLIP model's ViT-B/32 encoder, which directly encodes and obtains 512-dimensional visual attribute representations.

### C. Ablation Experiments

Ablation experiments are conducted to mask certain input modalities during the actual training process, observing how the evaluation metrics for each dialogue task are affected and to what extent. The input modalities include dialogue text modality, pointing information modality, spatial modality, image modality from scene metadata, and metadata information modality. The spatial modality assumes that when the system provides certain recommendations, the user can clearly understand the objects the system is referring to. Similarly, the system itself should be able to acquire object references for this portion of the dialogue state. In this study, only dialogue text modality, pointing information modality, and spatial modality are subjected to ablation experiments.

### D. Results

The experimental setup involved a Linux server with Ubuntu 20.04 and 8 Tesla P40 GPUs. Pre-trained language models from the transformers library were fine-tuned using custom datasets. Specifically, the GPT model utilized "openai-GPT" configuration, GPT-2 used "gpt2" configuration, and the BART model was based on "facebook/BART-large" configuration. Python was used for coding, and PyTorch framework for fine-tuning the deep learning models.

To ensure fair comparisons in ablation experiments, consistent training parameters were selected. Due to GPU memory limitations, each training batch contained only 6 samples. The model parameters were trained for 40,000 steps using the Adam optimizer with an initial learning rate of $5e^{-5}$.

The BART model outperformed GPT and GPT-2 in various metrics for the multi-modal coreference resolution task, achieving an impressive F1 score of 0.6864 in referent prediction, surpassing GPT (F1 score: 0.5157) and GPT-2 (F1 score: 0.4997) significantly. Though GPT had a higher F1 score in slot recognition, all three models surpassed 95%. The BART model, with auxiliary tasks, effectively encoded dialogue history, comprehended spatial information, and integrated metadata and visual image data, leading to high-quality dialogue systems.

Ablation experiments investigated each modality's role in dialogue tasks, especially the pointing modality, leveraging AR environments' characteristics. Models were trained with masked versions of text, pointing, and spatial scene-related modalities. The pointing modality consistently improved performance in most evaluated dialogue tasks, enhancing the multi-modal input's information load and training environment.

Directly masking dialogue text resulted in poor performance for all tasks. Masking spatial modality led to a more severe decline in performance compared to masking the pointing modality, highlighting its importance as contextual information.

An interesting fact is that, we additionally studied outputs when text, pointing, and spatial information were masked based on the BART model. Access to context from all modalities allowed accurate predictions for referent objects, dialogue actions, and slot values. However, masking the text modality caused off-topic replies and slot value inaccuracies, while masking the pointing modality resulted in errors in referent object prediction and metadata information. Masking spatial modality led to incomplete referent object prediction, incorrect

TABLE I: Results of Deep Learning Models on Various Tasks in Dialogue Systems

| Task Type | Task Metric | GPT | GPT-2 | BART |
|---|---|---|---|---|
| Multi-Modal Coreference Resolution | Mention Accuracy | 0.5174 | 0.4996 | **0.8736** |
| | Mention Recall | 0.5140 | 0.4999 | **0.5652** |
| | Mention F1 Score | 0.5157 | 0.4997 | **0.6864** |
| Multi-Modal Dialogue State Tracking | Intent Accuracy | 0.9568 | 0.9497 | **0.9585** |
| | Intent Recall | **0.9567** | 0.9496 | 0.9557 |
| | Intent F1 Score | 0.9567 | 0.9496 | **0.9571** |
| | Slot-Value Accuracy | 0.7766 | 0.8446 | **0.8886** |
| | Slot-Value Recall | 0.2784 | 0.8275 | **0.8761** |
| | Slot-Value F1 Score | 0.4098 | 0.8359 | **0.8823** |
| | Request Slot Accuracy | 0.9400 | **0.9364** | 0.9219 |
| | Request Slot Recall | **0.9083** | 0.9044 | 0.8681 |
| | Request Slot F1 Score | **0.9239** | 0.9201 | 0.8942 |
| Multi-Modal System Response Generation | System Reply BLEU-4 | 0.2238±0.0023 | 0.2190±0.0023 | **0.2843±0.0027** |
| | Joint Accuracy | 0.1899 | 0.5324 | **0.5478** |

TABLE II: Results of deep learning model ablation experiments

| Task Type | Task Metric | Mask-text | Mask-point | Mask-obj | No Mask |
|---|---|---|---|---|---|
| GPT-MM-Coref | Referent Accuracy | 0.3571 | 0.4045 | 0.1631 | **0.5174** |
| | Referent Recall | 0.3289 | 0.4013 | 0.1581 | **0.5140** |
| | Referent F1 Score | 0.3425 | 0.4029 | 0.1606 | **0.5157** |
| GPT-MM-DST | Intent Recognition Accuracy | 0.4366 | 0.9566 | **0.9576** | 0.9568 |
| | Intent Recognition Recall | 0.4366 | 0.9564 | **0.9575** | 0.9567 |
| | Intent Recognition F1 Score | 0.4366 | 0.9565 | **0.9575** | 0.9567 |
| | Slot-Value Recognition Accuracy | 0.0321 | 0.7726 | **0.7810** | 0.7766 |
| | Slot-Value Recognition Recall | 0.0100 | 0.2778 | 0.2760 | **0.2784** |
| | Slot-Value Recognition F1 Score | 0.0153 | 0.4086 | 0.4079 | **0.4098** |
| | Requested Slot Recognition Accuracy | 0.1449 | 0.9333 | 0.9340 | **0.9400** |
| | Requested Slot Recognition Recall | 0.1189 | **0.9086** | 0.9083 | 0.9083 |
| | Requested Slot Recognition F1 Score | 0.1306 | 0.9208 | 0.9210 | **0.9239** |
| GPT-MM-Response | System Reply BLEU-4 | 0.0815±0.0008 | 0.2258±0.0023 | 0.2283±0.0024 | **0.2238±0.0023** |
| GPT | Joint Accuracy | 0.0156 | 0.1230 | 0.0497 | **0.1899** |
| GPT2-MM-Coref | Referent Accuracy | 0.3207 | 0.3975 | 0.2483 | **0.4996** |
| | Referent Recall | 0.1967 | 0.3958 | 0.2427 | **0.4999** |
| | Referent F1 Score | 0.2438 | 0.3967 | 0.2455 | **0.4997** |
| GPT2-MM-DST | Intent Recognition Accuracy | 0.2599 | 0.9489 | 0.9494 | **0.9497** |
| | Intent Recognition Recall | 0.1202 | 0.9488 | 0.9487 | **0.9496** |
| | Intent Recognition F1 Score | 0.1644 | 0.9488 | 0.9490 | **0.9496** |
| | Slot-Value Recognition Accuracy | 0.0531 | 0.8474 | **0.8459** | 0.8446 |
| | Slot-Value Recognition Recall | 0.0049 | 0.8279 | **0.8306** | 0.8275 |
| | Slot-Value Recognition F1 Score | 0.0089 | 0.8374 | **0.8382** | 0.8359 |
| | Requested Slot Recognition Accuracy | 0.1374 | 0.9362 | 0.9327 | **0.9364** |
| | Requested Slot Recognition Recall | 0.0975 | 0.9041 | 0.9067 | **0.9044** |
| | Requested Slot Recognition F1 Score | 0.1140 | 0.9199 | 0.9195 | **0.9201** |
| GPT2-MM-Response | System Reply BLEU-4 | 0.0710±0.0006 | 0.2180±0.0023 | 0.2183±0.0023 | **0.2190±0.0023** |
| GPT2 | Joint Accuracy | 0.0070 | 0.4758 | 0.4352 | **0.5324** |
| BART-MM-Coref | Referent Accuracy | 0.4000 | 0.8089 | 0.8033 | **0.8736** |
| | Referent Recall | 0.0011 | 0.5517 | 0.3483 | **0.5652** |
| | Referent F1 Score | 0.0022 | 0.6560 | 0.4859 | **0.6864** |
| BART-MM-DST | Intent Recognition Accuracy | 0.5099 | 0.9501 | 0.7057 | **0.9585** |
| | Intent Recognition Recall | 0.5097 | 0.8385 | 0.4603 | **0.9557** |
| | Intent Recognition F1 Score | 0.5098 | 0.8908 | 0.5572 | **0.9571** |
| | Slot-Value Recognition Accuracy | 0.1222 | 0.8769 | 0.8487 | **0.8886** |
| | Slot-Value Recognition Recall | 0.0787 | 0.8127 | 0.6531 | **0.8761** |
| | Slot-Value Recognition F1 Score | 0.0957 | 0.8436 | 0.7381 | **0.8823** |
| | Requested Slot Recognition Accuracy | 0.2392 | 0.9016 | 0.8999 | **0.9219** |
| | Requested Slot Recognition Recall | 0.0968 | 0.5800 | 0.5440 | **0.8681** |
| | Requested Slot Recognition F1 Score | 0.1378 | 0.7059 | 0.6781 | **0.8942** |
| BART-MM-Response | System Reply BLEU-4 | 0.1006±0.0013 | 0.2831±0.0027 | 0.2769±0.0027 | **0.2843±0.0027** |
| BART | Joint Accuracy | 0.0178 | 0.5053 | 0.3031 | **0.5478** |

slot value predictions, and errors in system replies due to the absence of spatial information.

## VI. Conclusion

In conclusion, this study makes three significant contributions. Firstly, we introduce the SIMMC2-Point dataset,

transforming SIMMC2 dataset for AR environments while incorporating a pointing modality. Secondly, we propose a multi-modal dialogue model, evaluated for coreference resolution, dialogue state tracking, and response generation. Lastly, our ablation experiments demonstrate the importance of the pointing modality in enhancing performance across tasks. These findings advance dialogue systems in AR, enabling more immersive and natural interactions. Future research could explore additional user-related modalities to further enhance AR dialogue systems.

## ACKNOWLEDGMENT

This work is supported by National Key R&D Program of China 2018AAA0101501, Science and Technology Project of SGCC (State Grid Corporation of China): Fundamental Theory of Human-in-the-loop Hybrid-Augmented Intelligence for Power Grid Dispatch and Control, and National Natural Science Foundation of China No. 62102308.

## REFERENCES

[1] S. Kottur, S. Moon, A. Geramifard, and B. Damavandi, "Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations," in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021, pp. 4903–4912.

[2] H. De Vries, K. Shuster, D. Batra, D. Parikh, J. Weston, and D. Kiela, "Talk the walk: Navigating new york city through grounded dialogue," *arXiv preprint arXiv:1807.03367*, 2018.

[3] Q. Yuan, R. Wang, Z. Pan, S. Xu, J. Gao, and T. Luo, "A survey on human-computer interaction in spatial augmented reality," *Journal of Computer-Aided Design & Computer Graphics*, vol. 33, no. 3, pp. 321–332, 2021.

[4] J. Tao, Q. Wu, C. Yu, D. Weng, G. Li, T. Han, Y. Wang, and B. Liu, "A survey on multi-modal human-computer interaction," *Journal of Image and Graphics*, 2022.

[5] Z. Wang and J. Dai, "Multimodal human-machine hybrid interaction intelligent control technology," *Science and Technology Vision*, no. 9, pp. 11–12, 2018.

[6] M. Xiao, "Research and application of multimodal fusion method in augmented reality," Master's thesis, University of Jinan, 2020.

[7] C. Qian, "Multimodal perception of visuohaptic augmented reality for hand rehabilitation application study," Master's thesis, Nanjing University of Information Science & Technology, 2022.

[8] L. Ma and K. Shen, "Multimodal discourse analysis in the augmented reality of foreign language teaching context," *Modern Educational Technology*, vol. 22, no. 7, pp. 49–53, 2012.

[9] A. W. Ismail, M. Billinghurst, and M. S. Sunar, "Vision-based technique and issues for multimodal interaction in augmented reality," in *Proceedings of the 8th International Symposium on Visual Information Communication and Interaction*, 2015, pp. 75–82.

[10] A. W. Ismail and M. S. Sunar, "Multimodal fusion: gesture and speech input in augmented reality environment," in *Computational Intelligence in Information Systems: Proceedings of the Fourth INNS Symposia Series on Computational Intelligence in Information Systems (INNS-CIIS 2014)*. Springer, 2015, pp. 245–254.

[11] D. Cordeiro, N. Correia, and R. Jesus, "Arzombie: A mobile augmented reality game with multimodal interaction," in *2015 7th International Conference on Intelligent Technologies for Interactive Entertainment (INTETAIN)*. IEEE, 2015, pp. 22–31.

[12] M. Fischbach, D. Wiebusch, and M. E. Latoschik, "Semantics-based software techniques for maintainable multimodal input processing in real-time interactive systems," in *2016 IEEE 9th Workshop on Software Engineering and Architectures for Realtime Interactive Systems (SEARIS)*. IEEE, 2016, pp. 1–6.

[13] R. Z. Abidin, H. Arshad, and S. A. A. Shukri, "A framework of adaptive multimodal input for location-based augmented reality application," *Journal of Telecommunication, Electronic and Computer Engineering (JTEC)*, vol. 9, no. 2-11, pp. 97–103, 2017.

[14] M. N. Lystbæk, P. Rosenberg, K. Pfeuffer, J. E. Grønbæk, and H. Gellersen, "Gaze-hand alignment: Combining eye gaze and mid-air pointing for interacting with menus in augmented reality," *Proceedings of the ACM on Human-Computer Interaction*, vol. 6, no. ETRA, pp. 1–18, 2022.

[15] L. Chen, W. Wang, J. Qu, S. Lei, and T. Li, "A command and control system for air defense forces with augmented reality and multimodal interaction," in *Journal of Physics: Conference Series*, vol. 1627, no. 1. IOP Publishing, 2020, p. 012002.

[16] P. Knierim, P. W. Woźniak, Y. Abdelrahman, and A. Schmidt, "Exploring the potential of augmented reality in domestic environments," in *Proceedings of the 21st International Conference on Human-Computer Interaction with Mobile Devices and Services*, 2019, pp. 1–12.

[17] D. Adiwardana, M.-T. Luong, D. R. So, J. Hall, N. Fiedel, R. Thoppilan, Z. Yang, A. Kulshreshtha, G. Nemade, Y. Lu *et al.*, "Towards a human-like open-domain chatbot," *arXiv preprint arXiv:2001.09977*, 2020.

[18] N. Mostafazadeh, C. Brockett, W. B. Dolan, M. Galley, J. Gao, G. Spithourakis, and L. Vanderwende, "Image-grounded conversations: Multimodal context for natural question and response generation," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 462–472.

[19] Y. Zhu, O. Groth, M. Bernstein, and L. Fei-Fei, "Visual7w: Grounded question answering in images," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 4995–5004.

[20] J. Lei, L. Yu, M. Bansal, and T. L. Berg, "Tvqa: Localized, compositional video question answering," in *Empirical Methods in Natural Language Processing*, 2018.

[21] S. Moon, S. Kottur, P. A. Crook, A. De, S. Poddar, T. Levin, D. Whitney, D. Difranco, A. Beirami, E. Cho *et al.*, "Situated and interactive multimodal conversations," in *Proceedings of the 28th International Conference on Computational Linguistics*, 2020, pp. 1103–1121.

[22] W. Wang, J. Chen, and Q. Jin, "Videoic: A video interactive comments dataset and multimodal multitask learning for comments generation," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 2599–2607.

[23] Y.-C. Chen, L. Li, L. Yu, A. El Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX*. Springer, 2020, pp. 104–120.

[24] D. Guo, A. Gupta, S. Agarwal, J.-Y. Kao, S. Gao, A. Biswas, C.-W. Lin, T. Chung, and M. Bansal, "Gravl-bert: Graphical visual-linguistic representations for multimodal coreference resolution," in *Proceedings of the 29th International Conference on Computational Linguistics*, 2022, pp. 285–297.

[25] J. D. M.-W. C. Kenton and L. K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[26] H. Le, N. Chen, and S. Hoi, "Multimodal dialogue state tracking," in *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022, pp. 3394–3415.

[27] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[28] L. Yang, "Research and implementation of deep learning based chatbot," Master's thesis, Beijing University of Posts and Telecommunications, 2019.

[29] F. Yang, Y. Rao, Y. Ding, W. He, and Z. Ding, "Progress in task-oriented dialogue system," *Journal of Chinese Information Processing*, vol. 35, no. 10, pp. 1–20, 2021.

[30] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever *et al.*, "Language models are unsupervised multitask learners," *OpenAI blog*, vol. 1, no. 8, p. 9, 2019.

[31] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 7871–7880.

[32] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE conference on computer vision and pattern recognition*. Ieee, 2009, pp. 248–255.