UNITER-Based Situated Coreference Resolution with Rich Multimodal Input

Yichen Huang,[1] **Yuchen Wang,** [1] **Yik-Cheung Tam** [1]

[1] New York University Shanghai
yh2689@nyu.edu, yw3642@nyu.edu, yt2267@nyu.edu

Abstract

We present our work on the multimodal coreference resolution task of the Situated and Interactive Multimodal Conversation 2.0 (SIMMC 2.0) dataset as a part of the tenth Dialog System Technology Challenge (DSTC10). We propose a UNITER-based model utilizing rich multimodal context such as textual dialog history, object knowledge base and visual dialog scenes to determine whether each object in the current scene is mentioned in the current dialog turn. Results show that the proposed approach outperforms the official DSTC10 baseline substantially, with the object F1 score boosted from 36.6% to 77.3% on the development set, demonstrating the effectiveness of the proposed object representations from rich multimodal input. Our model ranks second in the official evaluation on the object coreference resolution task with an F1 score of 73.3% after model ensembling.

1 Introduction

The goal of Situated and Interactive Multimodal Conversation (SIMMC) 2.0 (Kottur et al. 2021) is to aid the conversational AI community in developing successful multimodal assistant agents capable of handling real-world multimodal dialog inputs. As part of the DSTC10 challenge, this dataset includes dialogs and multimodal contexts in the fashion and furniture domain that closely resemble real-world scenarios with more complex and cluttered images compared to previous datasets (Moon et al. 2020; Crook et al. 2019).

In this paper, we focus on the multimodal coreference resolution sub-task. Given a multimodal dialog context including dialog history, raw scene images, bounding boxes of detected objects and their coordinates, scene graphs, and a knowledge base (KB) of objects, our task is to determine whether each object is mentioned in the current user dialog turn. For example, a system turn asks, ”which group of pants are you referring to?” and a user may reply 'The ones on the left.', requiring to resolve which object IDs are referred to from a given active scene image. This task necessitates the system's understanding from contextualized multiple modalities, including textual, visual and structural object KB, laying the foundation for downstream processing, including dialog state tracking, response generation and retrieval.

Top approaches in the previous SIMMC challenge (Huang et al. 2021; Jeong et al. 2021) cast action prediction, dialog state tracking and response generation as a sequence-to-sequence generative approach where the multimodal context is flattened as a sequence of tokens which are coupled with dialog history for encoding. An auto-regressive decoder generates the corresponding actions, dialog state and response as a sequence of tokens. While such approaches were effective in the last challenge, they are incompatible with our task for several reasons. Firstly, it is not trivial to flatten raw images and scene graphs as a sequence of input symbols. Secondly, the flattened multimodal context significantly increases the input sequence length in SIMMC 2.0, where a scene image can contain more than 20 objects, easily exceeding the typical input limit of 512 tokens. Thirdly, the generated output sequence relies on ad-hoc post-processing to make sure that, for example, the object IDs are in the correct order and without duplicates during the beam search decoding.

We propose a UNITER(Chen et al. 2020)-based model for SIMMC 2.0. UNITER is proposed in computer vision (CV) for universal embeddings for image and text. To achieve this goal, UNITER is pre-trained with masked language modelling, masked region modelling and word-region alignment criteria on parallel image-text data. We extend UNITER to handle complex multimodal inputs. For object coreference resolution, we focus on the rich feature representation of each object to enable the underlying transformer model to comprehend the coreferences between textual dialog history with object candidates. In particular, each object is modelled with object index embedding, image embedding from a deep pre-trained CV model, KB entries cast as prompts for sentence embedding and additional feature engineering such as whether an object was mentioned in previous system dialog turns. Motivated from prior work of using scene graphs for visual question answering (Damodaran et al. 2021), we also incorporate scene graphs that include the positional relationship between objects in a scene. In particular, we evaluate two methods: 1) injecting scene graph information as attention biases; 2) through additional relation-aware self-attention layers. Finally, each object candidate is treated as an input embedding into UNITER and outputs a binary object mention label. In other words, our model contains a binary classification head per object candidate.

The paper is organized as follows: Section 2 provides an overview of the multimodal coreference task and the

SIMMC 2.0 dataset. Section 3 presents our proposed approaches. Section 4 describes experiments and results, with concluding remarks in section 5.

2 Task Descriptions

The SIMMC 2.0 dataset assumes scenarios in shopping settings where a user and an assistant agent co-observe scenes. The dataset is collected through a VR scene generator and is highly structured with data on different levels (turn, scene, dialog and KB metadata). A dialog can involve multiple scenes and turns. On the dialog level, the dataset contains the bounding box ids of all objects mentioned in the entire dialog. On the turn level, the dataset contains utterances and dialog state annotations (including the bounding box ids of objects mentioned in the current utterance) of both the user and the system and the scene id. On the scene level, the dataset contains the raw image, each object's bounding box, the relationships between objects (e.g. left, right) and unique ids linking each object to a knowledge base. Each knowledge base entry contains non-visual metadata (e.g. brand, price, available sizes) and visual metadata (e.g. type, color, pattern, sleeve length). The dataset contains a total of 11.2k dialogs and 1.5k scenes. The dataset is split into four sets: train (65%), dev (10%), dev-test (10%) and test-std (15%). Test-std is a hold-out hidden set for the DSTC10 challenge. The dataset involves two domains: fashion (7.2k dialogs) and furniture (4k dialogs).

The task of multimodal coreference resolution is to resolve scene-level ids in user utterances. An utterance can involve multiple references, in which case the ground-truth output is an unordered list of bounding box ids. The allowed inputs at inference time include past utterances of a user and the system, the current-turn utterance of the user, past object mentions of the system, scene data and non-visual metadata. Note that all user object mentions and all dialog state annotations are not allowed at inference time. Utterances with ambiguities are not included. The performance is evaluated using object F1.

3 Proposed approach

We formulate multimodal coreference resolution as an instance of binary token classification. Given dialog history U, object embeddings $O = o_1 o_2 ... o_I$ and scene embedding $S = s_1 ... s_j ... s_J$, we aim to predict binary object mention labels $Y = y_1 y_2 ... y_I$ indicating whether each object o_i is mentioned in the current user utterance. We use a UNITER encoder (Chen et al. 2020) to encode the above inputs. An overview of our proposed model is shown in Figure 1.

ObjectEmbeddings

As shown in figure 2, we first obtain separate embeddings for each of the multimodal object features and then aggregate them using a dense layer. We process the features as the following:

- A scene-level object index is embedded through an embedding layer.
- A cropped object image is fed into a visual encoder to extract the pooled region of interest (ROI) feature. In our experiments, we use pre-trained CLIP (Radford et al. 2021) and BUTD (Anderson et al. 2018) based on Faster R-CNN (Ren et al. 2015).
- The x, y and z coordinates of an object are used with no processing.
- The *non-visual* knowledge base (KB) entry of an object is first transformed into natural language form using a template shown in Table 1. Then the descriptive sentence is encoded with a text encoder such as BERT (Devlin et al. 2019) and sentence BERT (SBERT) (Reimers and Gurevych 2019).
- scene_active is a binary feature indicating whether an object is in the currently active scene. The label is embedded through an embedding layer.
- prev_mentioned is a binary feature indicating whether the system has mentioned an object in previous dialog turns. The label is embedded through an embedding layer.

These features are then concatenated and passed into a dense layer. Except for the image and text encoders, the rest embeddings are trained in an end-to-end manner.

Fashion	Item 15 is located at x : 5.32, y : -2.10, z: -3.96. It is located in the bounding box 104 334 260 133. Its price is $59.99. Its size is XL. Its brand is Downtown Stylists. It has a customer review of 4.1 out of 5. It is available in sizes S and XL.
Furniture	Item 7 is located at x : -756.50, y : 0.00, z: -358.20. It is located in the bounding box 838 383 45 30. Its price is $549. Its brand is River Chateau. It is made with metal. It has a customer review of 4.2 out of 5.

Table 1: Examples of knowledge base entry templates for the fashion and furniture domains.

Scene Embeddings

We add scene embeddings to reflect visual information not included in the object bounding boxes (e.g. relative positions and scene layouts). Scene embeddings $S = s_1 ... s_j ... s_J$ are obtained similar to the object embeddings except we use scene-level input indicators for index, scene_active and prev_mentioned different from object embeddings. Otherwise, we feed the entire scene image for image encoding. Coordinates and KB encodings are not used in scene embeddings.

Multimodal Encoder

We use a pre-trained UNITER model to encode dialog history, object embeddings and scene embeddings. The hidden states corresponding to each object position is passed into a dense layer to produce the output logits Z followed by Sigmoid function $\sigma(Z)$ for binary classification:

$$\begin{aligned} H &= \text{Encoder}(U, O, S) \\ Z &= \text{Dense}(H) \\ \hat{Y} &= \sigma(Z) \end{aligned} \tag{1}$$

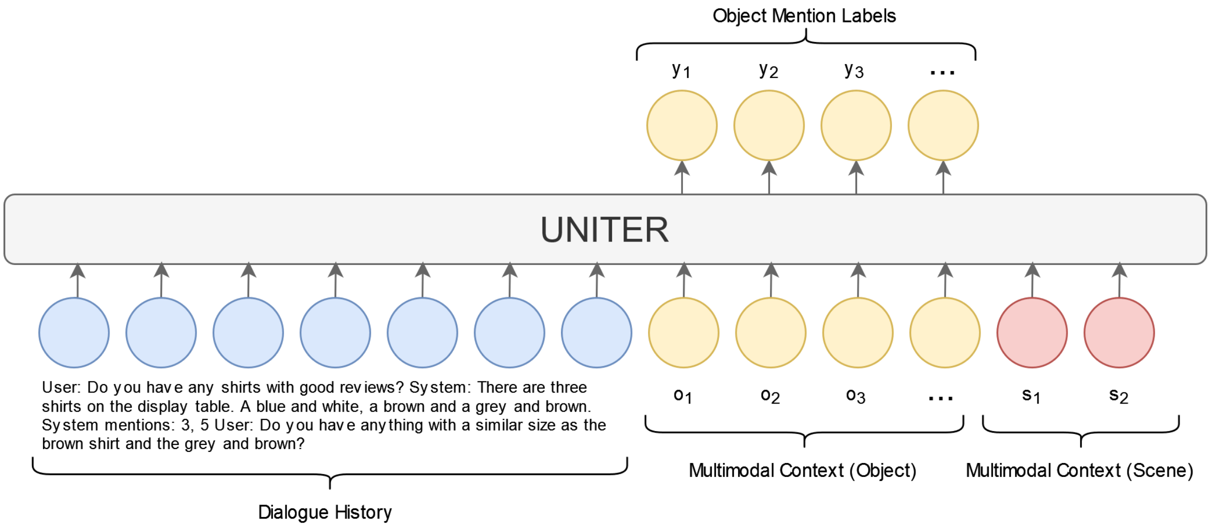

Object Mention Labels
y1
y2
y3
...
UNITER
User: Do you have any shirts with good reviews? System: There are three shirts on the display table. A blue and white, a brown and a grey and brown. System mentions: 3, 5 User: Do you have anything with a similar size as the brown shirt and the grey and brown?
Dialogue History
o1
o2
o3
...
Multimodal Context (Object)
s1
s2
Multimodal Context (Scene)

Figure 1: An overview of the proposed model.

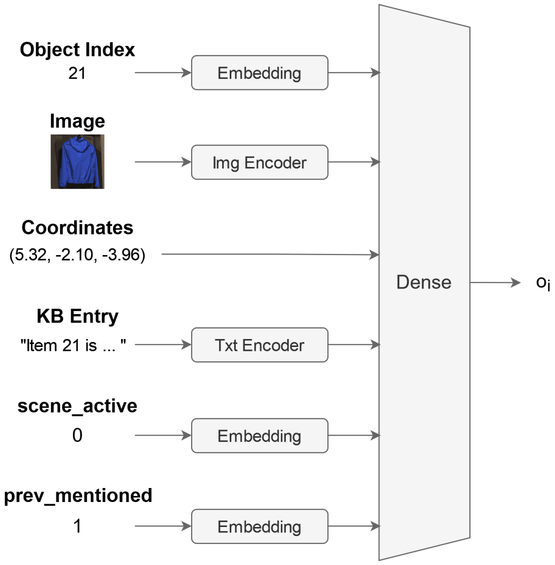


Figure 2: The separate embeddings of multimodal object features are concatenated and aggregated through a dense layer.

Scene Graph Integration

Scene graph is a way to represent the relationship of objects in an image using a directed graph where each node in a graph represents an object and a directed edge represents a relationship. The use of scene graphs has lead to improved performance in various vision-language tasks including visual QA (Damodaran et al. 2021), image captioning (Yang et al. 2019; Yao et al. 2018), text-to-image generation (Johnson, Gupta, and Fei-Fei 2018) and referring expression comprehension (Wang et al. 2019). This motivates us to adopt scene graphs for multimodal coreference resolution. We explore two approaches to modifying the UNITER's transformer layers via a self-attention mechanism between object pairs to incorporate scene graphs.

Attention Bias Similar to (Garncarek et al. 2021), we introduce a bias term modifying the attention score. For each attention head in each self-attention layer, we train a scalar embedding β^r for each of the relationships in the scene graphs (left, right, up and down). The modified attention head is as follows (See Figure 3):

$$\begin{aligned}\alpha &= \frac{(h^{l-1}W^Q)(h^{l-1}W^k)^\top}{\sqrt{d_k}}\\ \alpha' &= \alpha + \sum_{r=1}^{n}\beta^r g^r\\ h^l &= \mathrm{softmax}(\alpha')(h^{l-1}W^V)\end{aligned} \tag{2}$$

where W^K, W^Q and W^V are model parameters, d_k is the dimension of the query, key and value vectors, h^{l-1} and h^l are the hidden states of the previous and current transformer layers respectively, $n = 4$ in our case and g^r is a binary mask between objects o_i and o_j for each of the four relationships r={up,down,left,right}:

$$g^r_{ij} = \begin{cases} 1 & \text{if objects at position } i \text{ and } j \\ & \text{satisfies the relationship } r \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Relation-aware Self-Attention We also experiment with relation-aware self-attention (Shaw, Uszkoreit, and Vaswani 2018) where the attention is modified as follows (See Figure

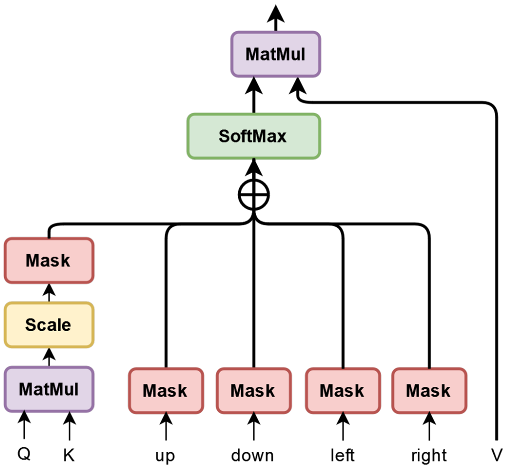


Figure 3: Integrating object-object relationship via the attention bias.

4):

$$u^r = \text{softmax}(\frac{(h^{l-1}W^{QS})(h^{l-1}W^{KS} + g^r_{ij}W^{KR})^\top}{\sqrt{d_k}})$$
$$h^l = \sum_{r=1}^{n} g^r \circ u^r(h^{l-1}W^{VS} + g^r_{ij}W^{VR}) \tag{4}$$

where W^{KS}, W^{KR}, W^{QS}, W^{VS} and W^{VR} are model parameters. $\circ$ denotes element-wise multiplication so that only hidden values corresponding to objects involved in a relationship is updated. Following (Ke et al. 2021), we add a relation-aware self-attention layer after every vanilla self-attention layer and add a residual connection to combine the outputs of the two.

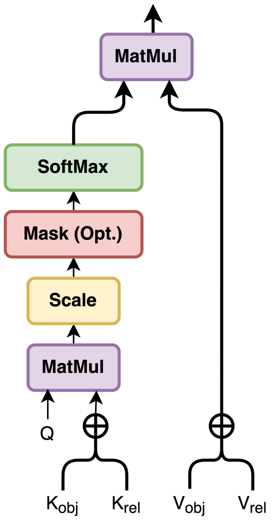


Figure 4: Relation-aware self-attention mechanism.

Model	**Object F1**
GPT-2 Baseline	0.366
UNITER + Faster RCNN	0.557
UNITER + CLIP	0.524
UNITER + both (Faster RCNN+CLIP)	0.563
UNITER + both + object idx	0.551
UNITER + both + object idx + coordinates	0.579
UNITER + both + object idx + coordinates + active_scene	0.582
UNITER + both + object idx + coordinates + active_scene + KB (BERT)	0.665
UNITER + both + object idx + coordinates + active_scene + KB (SBERT)	0.621
UNITER + both + object idx + coordinates + active_scene + KB (BERT+SBERT)	**0.674**
LXMERT + Faster RCNN	0.585
LXMERT + CLIP	0.590
Ensemble (Configuration for submission)	**0.741**
Post-eval improvement	
UNITER*	0.674
UNITER* + prev_mention	0.728
UNITER* + prev_mention + attn bias	0.734
UNITER* + prev_mention + relation-aware self-attn	0.733
LXMERT* + prev_mention	0.658
Ensemble	**0.773**

Table 2: Multimodal coreference resolution performance on the dev-test split. For models with both image or text encoders, the encodings are concatenated together. The ensembled output for submission is based on the top five UNITER-based models. The ensembled output post-evaluation is based on the best LXMERT-based model and the top three UNITER-based models. UNITER* and LXMERT* denote the best configuration used during DSTC10 evaluation.

4 Experiments

We trained the proposed model using the focal loss (Lin et al. 2017) with $\gamma = 2$ and $\alpha = 1$ for the negative class (i.e. an object is not mentioned), and $\alpha = 5$ for the positive class (i.e. an object is mentioned). We used the Adam optimizer (Kingma and Ba 2015) with a learning rate of 5×10^{-6} and $\epsilon = 10^{-8}$. We used a batch size of 16. We trained the model for a maximum of 30 epochs and performed early stopping according to the F1 score on the official development set. Our source code can be found at https://github.com/i-need-sleep/MMCoref_Cleaned.

Table 2 summarizes the evaluation result of the baseline models and our models with different inputs. The organizer provided the GPT-2 baseline, treating the input and outputs as a flattened sequence of tokens. We also compared the results with the same approach by replacing UNITER with LXMERT (Tan and Bansal 2019). Our UNITER-based models outperformed the baseline by a large margin by simply using a basic UNITER model with Faster RCNN or CLIP

Entry	Object F1
Team 4	0.758
Team 9 (Ours)	0.733
Team 8	0.682
Team 10	0.682

Table 3: Official results for multimodal coreference resolution on the held-out test-std split.

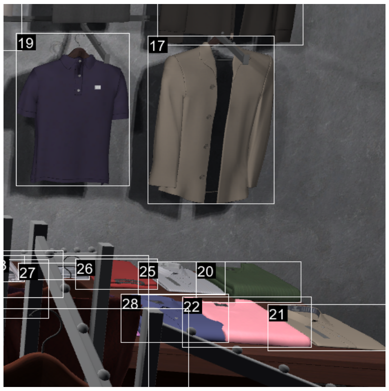
for image embeddings. As we incrementally incorporated additional features such as object coordinates, ”is an object inside the active scene?” (active_scene), KB entity description via BERT and SBERT, the object F1 score was further improved from 0.557 to 0.674 on the official devtest split, which was the best single model before the official evaluation deadline. Notably, UNITER-based models outperformed LXMERT-based ones under the same configuration. We ensembled the top five UNITER-based models for our DSTC10 submission. Due to the time constraint, some of the models used in the ensemble were not fully trained.

Table 3 shows our entry in the DSTC10 challenge on the test-std set among the top teams—we ensemble five models with different input settings. Finally, we achieved 2nd place in the evaluation on multimodal coreference resolution subtask. In addition, the object F1 scores between dev-test and test-std sets were similar, implying that performance in devtest can be used to predict performance on the test-std set.

Due to time limitations, we did not incorporate the scene graph features and the indicator feature of whether an object is mentioned in the previous system turn (prev_mention). Table 2 shows that these features have brought us further improvement post evaluation. In particular, solely adding the prev_mention feature boosted the F1 from 0.674 to 0.728. Intuitively, if an object mentioned in the previous system turns, the same object has a higher chance of being discussed in subsequent dialog turn. On the other hand, incorporating scene graphs either using attention bias or relation-aware self-attention only yielded marginal improvement, which might be due to the limited relational information in the provided scene graphs. Using more relational information deserves further investigation in the future.

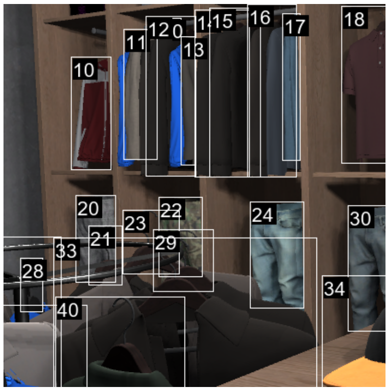
Error Analysis

We identify two salient types of error. A frequent type of error is exemplified in Table 4 where the model fails to understand the positional relationship between objects and background objects (e.g. walls, shelves, stands, racks, etc.) in the scene. Instead, the model classifies similar objects in other positions, as mentioned. The provided scene graphs are less helpful in such situations due to the lack of background objects. Another type of error relates to the complex and cluttered nature of some scenes. As demonstrated in Table 5, some mentioned objects are very far away from the observer or have bounding boxes overlapping with other objects, making the processing of visual features difficult and inaccurate.



User utterance	The purple t-shirt hanging on the wall.
Predicted object mentions	28
Groundtruth object mentions	19

Table 4: An example of an error where the model cannot identify a mentioned object on the wall. The image is cropped to include only the relevant region.



User utterance	Can I get the specs on the red and white sportsman jacket and that black one on the rack?
Predicted object mentions	4, 10
Groundtruth object mentions	4, 29

Table 5: An example of an error where the model cannot identify a mentioned object on a cluttered rack. Note that the object 29 is barely visible. The image is cropped to include only the relevant region.

5 Conclusion

We propose a UNITER-based model addressing multimodal coreference resolution. Our model incorporates multimodal inputs, including dialog history, raw images, KB entries, scene graphs, and indicator features. Experiments show that our approach significantly outperforms the GPT2 baseline and achieved second place on multimodal coreference resolution in the DSTC10 challenge.

A limitation of our approach is that we do not further train the pre-trained visual encoder and textual KB entry encoder. Finetuning them with the ground truth KB entries might yield better performance. Also, further efforts can be made to model the relationship between objects and the background, such as extracting and incorporating more ROI features. We leave such explorations as future work.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Chen, Y.-C.; Li, L.; Yu, L.; Kholy, A. E.; Ahmed, F.; Gan, Z.; Cheng, Y.; and Liu, J. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Crook, P. A.; Poddar, S.; De, A.; Shafi, S.; Whitney, D.; Geramifard, A.; and Subba, R. 2019. SIMMC: Situated Interactive Multi-Modal Conversational Data Collection And Evaluation Platform. *arXiv preprint arXiv:1911.02690*.

Damodaran, V.; Chakravarthy, S.; Kumar, A.; Umapathy, A.; Mitamura, T.; Nakashima, Y.; Garcia, N.; and Chu, C. 2021. Understanding the Role of Scene Graphs in Visual Question Answering. *CoRR*, abs/2101.05479.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Garncarek, Ł.; Powalski, R.; Stanisławek, T.; Topolski, B.; Halama, P.; Turski, M.; and Graliński, F. 2021. LAMBERT: Layout-Aware Language Modeling for Information Extraction. In Lladós, J.; Lopresti, D.; and Uchida, S., eds., *Document Analysis and Recognition – ICDAR 2021*, 532–547. Cham: Springer International Publishing. ISBN 978-3-030-86549-8.

Huang, X.; Tan, C. S.; Ng, Y. B.; Shi, W.; Yeo, K. H.; Jiang, R.; and Kim, J. J. 2021. Joint Generation and Bi-Encoder for Situated Interactive MultiModal Conversations. In *DSTC9 Workshop @ AAAI-21*.

Jeong, Y.; Lee, S.; Ko, Y.; and Seo1, J. 2021. TOM : End-to-End Task-Oriented Multimodal Dialog System with GPT-2 Conversations. In *DSTC9 Workshop @ AAAI-21*.

Johnson, J.; Gupta, A.; and Fei-Fei, L. 2018. Image Generation From Scene Graphs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ke, P.; Ji, H.; Ran, Y.; Cui, X.; Wang, L.; Song, L.; Zhu, X.; and Huang, M. 2021. JointGT: Graph-Text Joint Representation Learning for Text Generation from Knowledge Graphs. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2526–2538. Online: Association for Computational Linguistics.

Kingma, D. P.; and Ba, J. 2015. Adam: A Method for Stochastic Optimization. In Bengio, Y.; and LeCun, Y., eds., *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kottur, S.; Moon, S.; Geramifard, A.; and Damavandi, B. 2021. SIMMC 2.0: A Task-oriented Dialog Dataset for Immersive Multimodal Conversations. arXiv:2104.08667.

Lin, T.-Y.; Goyal, P.; Girshick, R.; He, K.; and Dollár, P. 2017. Focal Loss for Dense Object Detection. In *2017 IEEE International Conference on Computer Vision (ICCV)*, 2999–3007.

Moon, S.; Kottur, S.; Crook, P. A.; De, A.; Poddar, S.; Levin, T.; Whitney, D.; Difranco, D.; Beirami, A.; Cho, E.; Subba, R.; and Geramifard, A. 2020. simmc. *arXiv preprint arXiv:2006.01460*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. arXiv:2103.00020.

Reimers, N.; and Gurevych, I. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Ren, S.; He, K.; Girshick, R.; and Sun, J. 2015. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. In Cortes, C.; Lawrence, N.; Lee, D.; Sugiyama, M.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Shaw, P.; Uszkoreit, J.; and Vaswani, A. 2018. Self-Attention with Relative Position Representations. In *NAACL*.

Tan, H.; and Bansal, M. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Wang, P.; Wu, Q.; Cao, J.; Shen, C.; Gao, L.; and Hengel, A. v. d. 2019. Neighbourhood Watch: Referring Expression Comprehension via Language-Guided Graph Attention Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, X.; Tang, K.; Zhang, H.; and Cai, J. 2019. Auto-Encoding Scene Graphs for Image Captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yao, T.; Pan, Y.; Li, Y.; and Mei, T. 2018. Exploring Visual Relationship for Image Captioning. In *Proceedings of the European Conference on Computer Vision (ECCV)*.