Entity Resolution in Situated Dialog with Unimodal and Multimodal Transformers

Alejandro Santorum Varela, Svetlana Stoyanchev, Simon Keizer, Rama Doddipatla, Kate Knill

***Abstract*—In this work we address the entity resolution task for situated multimodal dialog investigating how a unimodal approach, which uses only textual information as input (representing visual attributes as text), compares to a multimodal system, which processes both text and visual information. We analyze two of the approaches presented in the Tenth Dialog Systems Technology Challenge and propose modifications to the original models that enhance their performance on the multimodal coreference resolution task. We evaluate both approaches on in- and out-of-domain settings by training the models on the fashion domain and testing on the furniture domain, and vice-versa. Through systematic analysis, we show that while both systems achieve similar performance on in-domain scenarios, the multimodal system generalizes better to out-of-domain settings. A combination strategy of enhanced unimodal and multimodal systems achieves F1=0.80 (5% absolute gain compared to the best performing system). Finally, human performance on the same task is evaluated on a small subset, suggesting that the performance of the current automatic models is on par with people on this task.**

***Index Terms*—Multimodal Coreference Resolution, DSTC10, SIMMC2.0, Transformers, BART, UNITER.**

I. Introduction

In conversations speakers often refer to real world entities in order to convey information to their communication partner. The same entity may be referred to using different referring expressions throughout the conversation. A new entity may be introduced with a specific noun phrase mentioning the entity by name or properties, e.g *'The Rice Boat is a nice Indian restaurant'*. Subsequent utterances may then 'co-refer' to that same entity, e.g., using an under-specified noun phrase *'Is **that restaurant** in the city centre?'*, or just a pronoun *'Is **it** in the city centre?'*. Coreference or, more generally, entity resolution is the Natural Language Processing task of identifying the referent (a previously introduced entity, e.g. a particular restaurant) given a referring expression (e.g., *'it'*).

AI agents have to interpret references to objects in order to communicate successfully with the user. In previous research, entity resolution has been addressed for spoken or chat-based dialog agents, such as information search dialog systems [1]. References in spoken or chat-based dialog systems are resolved based purely on language features. In spoken dialog systems prosodic information may also be helpful for resolving references [2].

In contrast to the language-only setting of a typical single-modality dialog system, situated AI agents, such as robots and virtual humans, are co-present with the user in the same physical or virtual environment. In addition to the language modality, situated agents also have access to a visual scene through the video camera of the interface. In this co-present setting, it is natural for both the user and the system to refer to visible object properties such as color, shape, or relative position to identify an object, e.g. *'the red blouse on the left'*. Entity resolution in situated systems involves determining which physical object the user is referring to in an utterance, if any. In such systems successful entity resolution requires the use of both language and visual channels.

The Situated and Interactive Multimodal Conversation Track at *The Tenth Dialog System Technology Challenge* (DSTC10) involved four subtasks: 1) detection of ambiguous utterances, 2) coreference resolution, 3) state tracking, and 4) response generation, using the SIMMC2.0 dataset.[1] The dataset consists of semi-automatically collected customer-agent dialogs situated in two virtual environments: a clothing shop and a furniture shop. In this work, we analyze and extend some of the proposed approaches for the multimodal coreference resolution subtask, with the goal of building human-robot interfaces. In this particular scenario, the task is to automatically determine which clothing or furniture item the user mentions in their utterance. However, the proposed approaches are aimed to be applicable in any situated AI application, and this dataset provides a rich environment in which methods for multimodal entity resolution can be studied.

One of the approaches to multimodal entity resolution in the challenge involved applying multimodal transformers pretrained on both visual and text information, such as captioned images [3], [4]. In this approach, a single multimodal transformer model processes both visual and text information to determine which objects in the scene were referred to by the user. Another approach involved a unimodal text-only transformer that processed features representing visual attributes of the objects in a scene in the form of text. Interestingly, the top performing method in the competition followed the latter approach [5].

In this work we analyze the multimodal UNITER-based [3] and the text-only BART-based [5] models. By evaluating each model in an out-of-domain setting, we show that the multimodal transformer approach is capable of generalizing to unseen scenes. We propose improvements to each of the models and show that combining the two approaches further

This paper was produced by Toshiba Europe Ltd. (Cambridge, UK) in collaboration with the Cambridge University Engineering Department.

[1]https://github.com/facebookresearch/simmc2

improves the performance on this task to F1=0.80 on the SIMMC2.0 dataset.

The contributions of this work include:

- Evaluation of the unimodal and multimodal transformer approaches in an out-of-domain setting.
- Enhancing each of the models and combining them to achieve an improved performance on the coreference resolution task.
- Assessment of human performance on the SIMMC2.0 dataset.

In Section II, we summarize the prior related work, Section III presents the considered SIMMC2 dataset and Section IV describes the baselines methods and the proposed modifications. All the experiments and results are shown in Section V, and we illustrate the most frequent types of errors made by the models in Section VI. We conclude in Section VII summarizing and highlighting the main contributions of this work.

II. Related Work

Coreference/reference/entity resolution and grounding are the terminology used in the literature to describe the task of identifying which entity is being referred to. An entity may be a physical object, a grouping of objects (a table, chairs, cars, etc.) or an abstraction (thought, conclusion, etc.). Reference resolution is an essential part of language understanding. While *coreference resolution* assumes multiple references to the same entity, *reference resolution* more generally describes resolving each occurrence of a reference.

In text or language-only dialog, entities referred to are present in the mind of a reader or a dialog participant but not in a shared visual space. Reference resolution in text and dialog has been extensively addressed in past research [6]–[10] and annotation techniques for coreference have been addressed [11]. [12] tackles coreference resolution of noun phrases in unrestricted text, and [13] presents an approach to pronoun resolution based on syntactic paths. In a situated dialog, such as human-robot interaction, discussed entities include actual objects present in the physical space shared by the dialog participants. Resolving these object mentions requires the use of both text and visual features [14], [15], similarly to other multimodal tasks, such as Visual Question Answering [16].

In recent years, multimodal entity resolution has been extensively addressed due to the advances in robotics. As in other Natural Language tasks, the use of Large Language Models based on *Transformers* architectures [17] has been shown effective for multimodal coreference. [18] proposes a semi-supervised learning model to correlate embedding space structures of each modality in coreference resolution. In situated settings, gestures play an important role for entity resolution. [19] analyzes which of the gestures provides extra information for the dialog. [20] analyzes human perception of automatically generated referring expressions.

The Ninth and the Tenth Dialog Systems Technology Challenges included a track for situated multimodal conversational AI [21], [22] set in a simulated virtual shop environment with coreference resolution as one of the tasks. In this work we investigate two of the different approaches to entity resolution participating in the challenge [3], [5], analyzing them in a cross-domain setting and on mentioned and new object references.

III. Data

The SIMMC2 dataset presents task-oriented dialogs between a virtual shop assistant and a user. Each dialog is situated in a virtual fashion or furniture shop and is associated with one or more virtual shopping scenes simulated using a VR environment [23]. The dialogs were generated semi-automatically by first simulating the dialog structure and then manually editing the content of the utterances [22].

The dataset contains 7.2K and 4K dialogs in the fashion and furniture domains respectively. Fashion scenes have on average around 32 objects per scene while the furniture scenes have 11, making the fashion domain more challenging for the entity resolution task. The objects in each scene are annotated with their bounding boxes and IDs, which are linked to the metadata containing information about the objects in the form of visual and non-visual attributes (see Table I).

	Fashion	**Furniture**	**All**
Dataset statistics			
No. dialogs	$\sim$ 7.2K	$\sim$ 4K	$\sim$ 11.2K
No. objs. per scene	$\sim$ 32	$\sim$ 11	$\sim$ 20
Metadata			
Visual attributes	assetType, pattern and sleeveLength	-	type and color
Non-visual attrs.	size	materials	brand, price and customerReview

TABLE I
SIMMC2 DATASET STATISTICS AND METADATAS.

Each user and system utterance is annotated with the IDs of the objects that are mentioned in this utterance. For example, the referring expression in the system's utterance *'I have these two pink ones, one on top and one on the bottom, do you like them?'* identifies the objects #1 and #45 in the corresponding scene (see Figure 1). The goal of the entity resolution system is to identify which objects the current user utterance is referring to, given the dialog context and the object IDs referenced in it, the visual scene (including the scene image and the object locations), and the metadata of the objects.

Fig. 1. Example of a SIMMC2 scene situated in a fashion shop.

Speaker	Utterance	Objects Mentioned
User	What sweaters do you have with good ratings?	
System	I have these two pink ones, one on top and one on the bottom, do you like them?	1, 45
User	How much is the bottom one and who makes it?	PREDICT: 45

TABLE II
EXAMPLE SIMMC2.0 DIALOG BASED ON A FASHION SHOP SCENE IN FIGURE 1.

SIMMC2.0 is split into three sets: train (64%), dev (5%) and devtest (15%). Following the competition rules, we train the models on the training set, tune the parameters on the dev set, and report results on the devtest set.

IV. METHOD

The approaches to coreference resolution used by the participants of the competition differed in the methods of processing visual information. One approach was to use pre-trained multimodal transformers, such as LXMERT [24], or UNITER [25], to jointly encode the image information and the dialog history [3], [4]. Alternatively, the input was processed by a text-only transformer model, such as BART [26], BERT [27], or GPT2 [28]. Visual features were represented as text extracted from an image with an off-the-shelf image processing model [29] or from the metadata [5], [30].

The scores achieved by the challenge participants on the coreference resolution task varied from $F1 = 0.52$ to $F1 = 0.75$. The multimodal methods that processed images generally achieved higher scores, indicating that the use of visual input cannot be fully replaced by the metadata information. However, one of the text-only transformer-based approaches that did not process images was very effective, with a reported score of $F1 = 0.75$ [5]. In this method, the authors pre-train an object encoder with a contrastive learning technique similar to [31], but using object IDs instead of images. In this work, we compare this unimodal BART-based approach with the top performing multimodal UNITER-based approach [3]. In the rest of this paper we refer to them as *BART-based* and *UNITER-based* methods respectively or simply BART/UNITER.

A. UNITER-based model [3]

UNiversal Image-TExt Representation (UNITER) [25] is a multimodal encoder based on a transformer architecture pre-trained over four image-text datasets (COCO [32], Visual Genome [33], Conceptual Captions [34], and SBU Captions [35]). It is aimed at enabling downstream tasks with joint multimodal embeddings. The UNITER-based model for the multimodal coreference resolution task (see Figure 2) takes as input the dialog history including the last user utterance (U), the encoding of each of the objects in the scene (O), and the image of the overall scene (S). The dialog history is encoded using the pre-trained BERT [27] encoder. The cropped images of each object in the scene as well as the full scene image are encoded using the visual pre-trained CLIP model [31]. Next, each object in the scene is encoded with the combined information consisting of its CLIP-generated image embedding, 3D object coordinates, non-visual metadata attributes, binary flags indicating whether the object is present in the scene and whether it was previously mentioned, and object ID. The output of UNITER (H) provides a hidden multimodal representation for each object within the dialogue context (H_obj), which is passed into a dense layer producing a logit Z for each object in the scene. The logits are transformed into probabilities (Y) using the Sigmoid function $\sigma(Z)$, which are then used to classify each object with a binary label indicating whether the object is present or absent in the last user's reference. The input encoding, dense layer, and sigmoid layer steps are shown in Equation 1.

$$\begin{aligned} H = \langle H\_dial, H\_obj, H\_s \rangle &= \text{UNITER\_Enc}(U, O, S) \\ Z &= \text{Dense}(H\_obj) \\ Y &= \sigma(Z) \end{aligned} \tag{1}$$

UNITER parameters are fine-tuned while the object embedder, scene embedder and dense layer parameters are trained on the binary coreference annotations of the SIMMC2.0 dataset. For full details of the model please see the original paper [25].

B. BART-based model [5]

In the proposed solution for the DSTC10 competition, an encoder-decoder BART model addresses all four tasks of the SIMMC2.0 challenge at the same time: disambiguation, coreference resolution, state tracking, and response generation. While it was shown that multi-task learning can benefit each individual task, we observe that for the SIMMC2.0 dataset learning multiple tasks does not benefit the performance on coreference resolution. Hence, we use a simplified model with a single coreference prediction task and omit the BART decoder in our experiments.

Unlike the UNITER-based approach, the BART-based approach does not directly process images of objects or scenes. Instead, the dialog history, object IDs, and object locations are encoded using a pre-trained BART model (see Figure 3). In parallel with training the model for the coreference tasks, a contrastive encoder is trained with a contrastive learning objective [31] by maximizing the cosine similarity between the object ID embedding and text attributes embeddings of the corresponding object and by minimizing the cosine similarity between the ID embedding and the attribute embeddings of other objects. This approach allows to associate objects in a scene with the corresponding visual metadata descriptions without having to use images or visual attributes at inference time[2]. The embedded objects IDs, which are expected to encode the visual and non-visual attributes unique to each object, are combined with the embeddings of the corresponding object positions. Then, the result is concatenated with the embedded word sequence of the dialog history and passed to the BART encoder.

Similarly to the UNITER-based model, the output of the BART encoder is passed into a dense layer followed by a

[2]The use of visual attributes was disallowed by the competition rules.

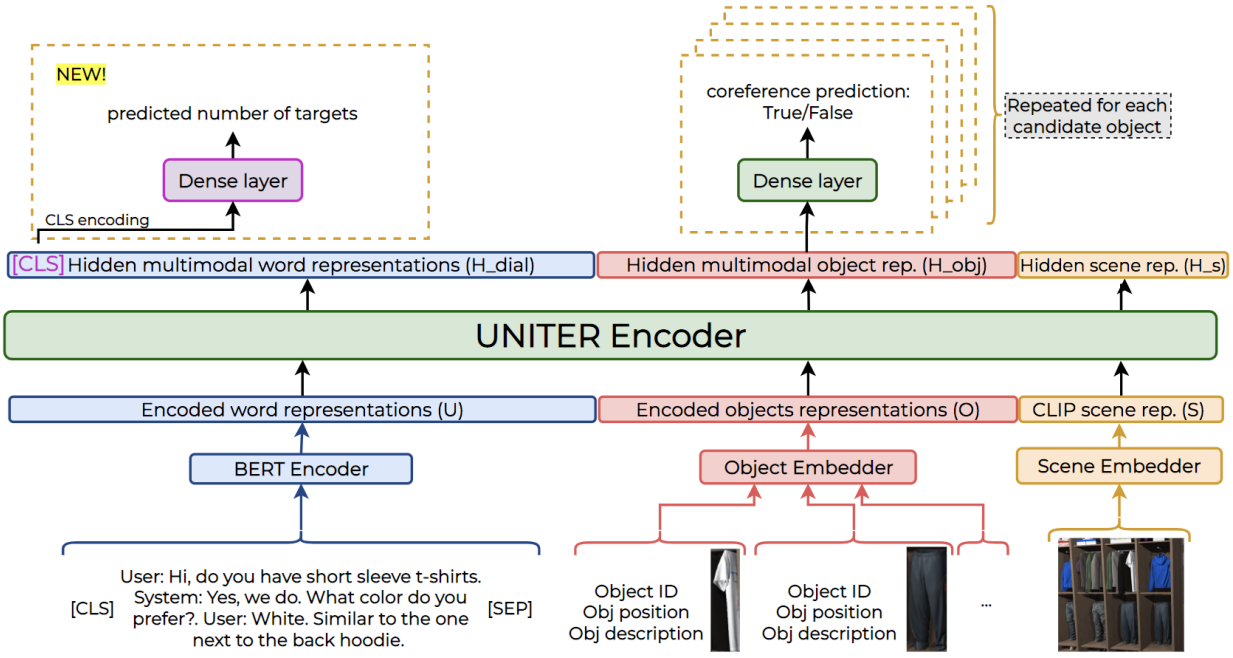

Fig. 2. UNITER-based model for coreference resolution [3].

sigmoid that outputs a binary prediction for each object (see Equation 1). The BART encoder parameters are fine-tuned while the location encoder and the coreference dense layer parameters are trained on the binary coreference resolution task using cross entropy loss. As an auxiliary task, the original model also predicted *empty coref*, a binary value for the whole utterance (0 if the utterance contains at least one entity reference and 1 otherwise). Our ablation study showed that this did not improve the performance, but rather than removing it, *empty coref* prediction was replaced with prediction of the number of referred objects (see Section IV-C).

The contrastive learning procedure for encoding object IDs differentiates the BART-based approach from the other approaches in the DSTC10 challenge. While this method is effective for detecting objects whose IDs were used in pre-training, it cannot generalize to unseen objects, as confirmed by our experiments.

C. Predicting the number of referred objects

In both UNITER-based and BART-based models, the presence of a reference to an object is predicted for each object in a scene individually, without taking into account the other objects. We observe that the model can *underpredict* (predicting too few objects) or *overpredict* (predict too many objects). To address these errors we propose a heuristic method which relies on knowing the number of objects that the user's utterance refers to.

"How about giving me a look at some hats I'd like?"	0
"The one next to the black hoodie."	1
"I want the blue coat on the right and the red shirt on top."	2
"How about these two jeans and the red blouse on the rack?"	3

TABLE III
EXAMPLES OF UTTERANCES AND THE NUMBER OF ITEMS THEY REFER TO.

By analyzing user utterances independently of the dialog and scene context, we can easily detect the number of objects the user refers to. For example, the first utterance in Table III is a general question and does not refer to any specific objects. The second one refers to one object as it mentions *'the one'*. As the third utterance contains a conjunction *'and'*, we can infer that it refers to two objects. We use an auxiliary learning task to predict the number of objects referred to by the user, $N \in \{0, 1, 2, > 2\}$, using multi-class classification. This approach is similar to the one used by the original BART-based model, which, among other tasks, predicts whether the user's turn referred to an item using an auxiliary learning task. Instead of making a binary prediction, we use a multi-class prediction.

The predicted number of objects N is used to heuristically post-process the model's outputs: if N is larger than the number of initially predicted target objects, we force the model to make more predictions, using the estimated probabilities from the *softmax* function applied to the *dense* layer's output.[3]

V. EXPERIMENTS

We present experimental results on the SIMMC2.0 dataset for the multimodal UNITER-based and unimodal BART-based approaches on *in-domain*, where train and test data contain the same set of scenes, *in-domain-held-out*, where no one scene appears in both train and test data, and *cross-domain* with different domains in train and test data. We also describe how the proposed modifications to the models improve reference resolution performance. In addition, we explore the use of visual attributes in the BART-based model. Finally, the models' performance is compared with human performance on a randomly selected subset of examples.

[3]Users rarely referred to more than 3 objects in the data, hence we limited the classification to four classes.

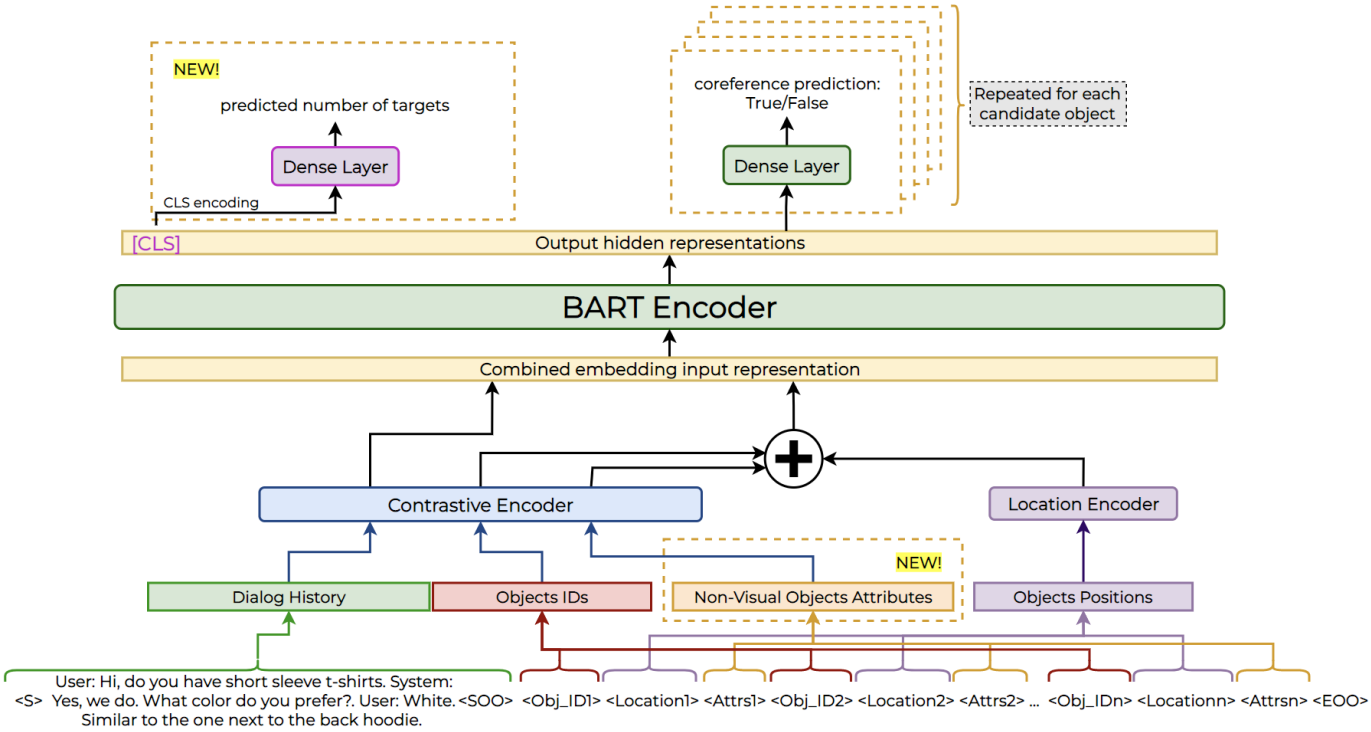

Fig. 3. BART-based model for coreference resolution [5].

Dataset Name	Domain	∼ objs/ scene	Selection method
ALL-train	Both	24	SIMMC2.0 *train* split
ALL-test	Both	25	SIMMC2.0 *devtest* split
FASH-14K	Fashion	33	Subset of ALL-train&ALL-test excluding `held-out` scene and the data from **FASH-9K-ID**.
FURN-12K	Furniture	10	Subset of ALL-train&ALL-test
FASH-6K	Fashion	31	Subset of ALL-test
FURN-2K	Furniture	9	Subset of ALL-test
FASH-9K-ID	Fashion	33	Random select from ALL-train&ALL-test excluding `held-out` scene
FASH-9K-IDHO	Fashion	24	Subset of ALL-train&ALL-test set in `held-out` shop.
FURN-9K-OOD	Furniture	9	ALL-train & ALL-test
ALL-100	Both	26	100 randomly selected examples from ALL-test

TABLE IV
SIMMC2.0 DATA SPLITS USED FOR THE EXPERIMENTS. `HELD-OUT` IS A RANDOMLY SELECTED SCENE FROM FASHION DOMAIN (`CLOTH_STORE_1498649_WOMAN`).

The official *train*, *dev*, and *devtest* splits of the SIMMC2.0 dataset contain both fashion and furniture domains. Because the same scenes from each domain occur across all data splits, we refer to the evaluation of the models trained and tested on the official data split as the *'in-domain'* experimental condition. To evaluate the models on unseen scenes in the same (*'in-domain-held-out'*) or different (*'cross-domain'*) domain, we create the data splits described in Table IV.

A. In-domain evaluation

Model	Description	Reported (F1)	Ours (F1)
	Original		
GPT-2	[22] (baseline)	0.366	0.381
UNITER	[3]	0.728	0.726
BART	[5]	0.743	0.742
	Modified		
UNITER	Remove object IDs	-	**0.758**
BART	Loss: coref task + *empty coref*	-	0.748
BART	Loss: coref task	-	0.748
BART	Loss: coref task + #obj. heur.	-	0.752
BART	Loss: coref task + non-visual	-	0.760
BART	Loss: coref task + #obj. heur. + non-visual	-	**0.763**
BART	Loss: coref task + #obj. heur. + non-visual + visual	-	**0.775**†

TABLE V
EVALUATION OF MULTIMODAL COREFERENCE RESOLUTION ON IN-DOMAIN CONDITION. ALL MODELS ARE TRAINED AND TESTED ON THE OFFICIAL *train* AND *devtest* DATA SPLITS OF SIMMC2.0. THE MODEL MARKED WITH † IS NOT LEGAL FOR DSTC10 AS IT USES VISUAL METADATA ATTRIBUTES.

Table V shows the reported and replicated (*Ours* column) performance for the baseline and the original versions of UNITER-based and BART-based models trained and tested on the official *train/devtest* split. Our replicated results for each of the models closely match the results reported by the corresponding papers [3], [5], [22].

Both of the original models use scene-specific object IDs and the BART-based model also uses global IDs as input during training and inference. The reliance on the IDs may prevent the models from generalizing. A model that learns

Experimental Condition	Test domain	Train set	Test set (F1)	UNITER-based (F1)	BART-based
In-domain vs. Cross-domain					
In-domain	Furniture	ALL-train	FURN-2K	0.843	**0.861**
Cross-domain	Furniture	FASH-14K	FURN-2K	**0.525**	0.457
In-domain	Fashion	ALL-train	FASH-6K	**0.736**	0.731
Cross-domain	Fashion	FURN-12K	FASH-6K	**0.425**	0.194
In-domain vs. Cross-domain vs. In-domain-held-out					
In-domain	Fashion	FASH-14K	FASH-9K-ID	**0.694**	0.675
Cross-domain	Furniture	FASH-14K	FURN-9K-OOD	**0.549**	0.373
In-doman-held-out	Fashion	FASH-14K	FASH-9K-IDHO	0.621	**0.740**

TABLE VI
EVALUATION ACROSS SCENES AND DOMAINS. IN-DOMAIN: TRAIN AND TEST DATA CONTAIN THE SAME SET OF SCENES; CROSS-DOMAIN: NO DOMAIN OVERLAP BETWEEN TRAIN AND TEST DATA; IN-DOMAIN-HELD-OUT: NO SCENE OVERLAP BETWEEN TRAIN AND TEST DATA.

references to objects based on their IDs rather than visual properties would not be able to detect a reference to a new object that was not seen during training. The BART-based model's use of object IDs is inherent to the method, as it pre-trains a contrastive encoder that maps global object IDs to the visual and non-visual features (see Section IV-B).

To test whether the UNITER-based model relies on the (scene-level) object IDs, we removed them from the input by excluding them from the object embedding. Interestingly, this modification results in an increase of the F1 score from 0.726 to 0.758, a 3% absolute increase. This result indicates that the UNITER-based model does not rely on the object IDs and therefore has the potential to generalize across domains.

As multitask learning has been shown to improve the performance of an individual task [36], the original BART-based model uses multi-task learning to jointly train a model for disambiguation, coreference resolution, state tracking, and response generation. In this work, we focus on coreference and evaluate a single-task variant of the BART-based model by modifying the loss function to only include the coreference loss. Contrary to the expectation, using a single-task loss function does not decrease the performance: the F1 score slightly increases from 0.742 to 0.748. Removing the *empty coref.* auxiliary task does not affect the result, but using the new proposed auxiliary task of predicting the number of objects with the corresponding heuristics (see Section IV-C) increases the performance to 0.752.

Next, we investigate using `metadata` features as input to the models (see Table I). The metadata includes a list of non-visual and visual attributes for each object in the scenes that can be referenced during the dialog (see Table I). Only non-visual attributes are allowed to be used at inference time in the DSTC10 competition, as the visual attributes are expected to be recognized from the scene image. The original UNITER-based model includes non-visual attributes as input to the object encoder. While the original BART-based model learns the object IDs embedding using contrastive learning, it does not take non-visual features as input during inference. We hypothesize that in addition to pretraining with visual and non-visual attributes, adding visual and non-visual descriptions of objects in the scene directly as input may help the model at inference time. We show including non-visual object features improves the BART model performance to $F1 = 0.763$ and including visual features further improves it to $F1 = 0.775$.

B. Evaluation across scenes and domains

An entity resolution model for a situated multimodal interface should, ideally, generalize to new in-domain settings (*in-domain-held-out* experimental condition) as well as new domains (*cross-domain* experimental condition) by relying on generic visual attributes, such as colour, relative position, and object type. While objects of different type may appear in scenes of different domains (furniture vs. clothing items), pre-training a multimodal transformer on a large generic dataset may give the BART-based model generalization ability. However, BART-based model also relies on object ID embedding which may prevent it from generalizing across domains. To assess the models' ability to generalize across scenes and domains we train and evaluate both models on different data subsets.

We first assess cross-domain generalization by comparing *in-domain* and *cross-domain* evaluation conditions (see top part of Table VI). The *in-domain* models are trained on the ALL-train split, which contains both fashion and furniture examples, and the *cross-domain* models are trained on a subset with only one domain (FASH-14K or FURN-12K). We observe that the performances of both UNITER and BART-based models is higher when tested on the furniture than on the fashion domain. The two models in the *in-domain* setting achieve an $F1$ of $0.843/0.861$ on furniture and $0.736/0.731$ on fashion respectively. This difference can be explained by the higher complexity of the fashion domain scenes with an average of 31 objects in fashion scenes and only 9 in the furniture scenes (see Table IV). For both models the performance is lower in the *cross-domain* setting than in the *in-domain* setting. However, the UNITER-based model achieves higher performance than the BART-based model in the cross-domain condition on each of the domains. In particular, in the *cross-domain* condition tested on the fashion domain, the UNITER model performance is $F1 = 0.425$ while the BART-based model scores $F1 = 0.194$. This result shows that the UNITER-based model, which extracts object information from images, has a better potential of generalizing to a new domain than the BART-based model, which fails to generalize due to its reliance on object IDs.

Next, we assess generalization to novel scenes in the same domain. A reorganization of a fashion store results in a new arrangement (or 'scene') composed of existing items. A robust model should still be able to resolve the user's references

	UNITER	BART	UNITER w/ Aux. task	BART w/ Aux. task
Mentioned objects	0.837	0.796	**0.844**	0.827
New objects	0.644	**0.722**	0.644	0.700
Overall	0.758	0.760	0.761	0.763

TABLE VII
EVALUATION OF THE AUXILIARY TRAINING TASK WITH THE ANALYSIS ON MENTIONED AND NEW OBJECTS. THE MODELS ARE TRAINED ON ALL-TRAIN AND TESTED ON THE STANDARD ALL-TEST DATA SPLITS.

to the items in the new scene. To test this generalization ability, we construct a dataset FASH-14K that excludes all examples associated with one `held-out` scene and use it for training.[4] We compare the models by testing them on 1) the fashion subset that contains the same scenes as the training set (*in-domain* condition), 2) the furniture subset (*cross-domain* condition), and 3) the dataset with `held-out` scenes only (*in-domain-held-out* condition). The results are shown in the bottom part of Table IV. For the *in-domain* condition the UNITER and BART models achieve similar performance with $F1$ of $0.694/0.675$. In the *cross-domain* condition we again observe that the UNITER-based model performance is higher than that of the BART-based model, with an $F1$ of 0.549 vs. 0.373. In the *in-domain-held-out* condition, the UNITER-based model performance drops by 7% points to 0.621 while the BART-based model performance increases to 0.740. We observe that FASH-9K-IDHO has lower complexity with 24 objects per scene in comparison to FASH-9K-ID, which has 33, making it potentially an easier task for both models. Nevertheless, the result shows that the BART's contrastive encoder for object IDs is effective in reference resolution on new scenes with seen objects. The drop in performance of the UNITER-based model on the `held-out` dataset indicates that the model may not effectively resolve references to relative positions.

C. Mentioned vs. new references

A reference to an entity in a dialog can be referring to a *mentioned* object or to a *new* object, not yet mentioned in previous discourse. In the example dialog shown in Table II, the last user utterance *'How much is the bottom one and who makes it?'* refers to an object previously mentioned by the system. However, a user may have said *'I would like to see the black sweater at the top next to the pink one'* referring to a new item not mentioned in previous discourse. In many of the examples, resolving a reference to a previously mentioned object does not require visual information from the corresponding scene. However, a reference to a new object in a multimodal context requires processing of the scene and interpretation either visual attributes or the position of the object relative to others, or both. Table VII shows the performance breakdown on the *mentioned* and *new* items for both models. We observe that both UNITER and BART have higher $F1$ scores on mentioned objects ($0.837/0.796$) than on new objects ($0.644/0.722$).

[4]Note that the objects that occur in `held-out` scene also occur in the training set.

Next, we evaluate the effect of using the auxiliary task of predicting the number of objects referenced in an utterance (see Section IV-C) on the *mentioned* and *new* references. We observe that both models get similar performance with $F1 = 0.76$. The auxiliary task benefits the mentioned objects. UNITER's F1 increases from 0.837 to 0.844 and BART's increases from 0.796 to 0.827, a relative increase of 1% and 4% respectively. The performance on the *new* objects, however, is unchanged for UNITER and decreased for BART. This can be caused by the reference detection probabilities on new objects being less accurate, highlighting the challenge of processing visual information required for new objects.

	BART-based		
w/ Aux. task	✓	✓	✓
w/ non-visual attributes		✓	✓
w/ visual attributes			✓
Mentioned	0.812	0.827	**0.835**
New	0.693	0.700	**0.715**
Overall	0.752	0.763	**0.775**

TABLE VIII
EFFECT OF THE NON-VISUAL AND VISUAL ATTRIBUTES FOR BART-BASED MODEL.

The SIMMC2.0 dataset includes metadata information with visual and non-visual attributes (see Table I). Non-visual attributes are acceptable by the DSTC10 competition rules, but visual attributes are not as the models should extract them from the visual scenes. Table VIII shows how the use of these attributes affects the performance of the BART-based model on *mentioned* and *new* objects. We observe that adding non-visual attributes benefit the performance on mentioned objects the most increasing it by 1.5% absolute points. In contrast, addition of the visual attributes leads to the larger increase (also 1.5% absolute points) in the performance on the *new* objects. This supports the hypothesis that when referring to new objects, processing of the visual information is especially critical.

UNITER for Mentioned w/ obj. heur.	BART for New w/ obj. heur	F1 Score overall
✓	✓	0.789
✓		**0.800**

TABLE IX
RESULTS AFTER COMBINING UNITER AND BART-BASED MODELS ON THE STANDARD SIMMC2.0 TEST AND COMPARABLE TO THE RESULTS IN TABLE V.

We observe that the UNITER-based system has a higher score than the BART-based system on *mentioned* objects while BART-based system has a higher score on *new* objects. If

we combine the strengths of the two approaches by using UNITER's prediction for the *mentioned* and BART's prediction for the *new* objects, we achieve $F1 = 0.80$ for the best model combination (see Table IX), an absolute increase of 3/4% points increase in comparison with the best performing BART/UNITER-based models reported in Table V.

When deploying such a combination in a system, we would have to keep track of the objects mentioned by the system and the user. While the objects mentioned by the system are known, the objects referred to by the user are automatically identified in previous turns and may contain errors. Hence, the performance of this combined system may degrade due to the propagation of the errors, but not below the performance of the inferior system.

D. Human evaluation

The model combination achieves $F1 = 0.80$ on the coreference resolution task on *in-domain* condition leaving ample room for improvement. To assess the upper bound for the model performance on SIMMC2.0 dataset, we asked people to perform the annotations using the same information as the models. We sampled 100 random examples from the SIMMC2.0 `devtest` set and had three members of the team annotate them. Each example consists of the dialog segment between the user and the assistant, the multimodal context (IDs of the mentioned items), and the current scene image with the objects' bounding boxes with their corresponding IDs.[5] Three members of the team completed the task by annotating the object IDs referred to in the last user utterance in each dialogue segment. The inter-annotator agreement [37] on this task was high with an average pairwise $\kappa = 0.86$.

	UNITER w/ Aux. task	BART	Human
Mentioned	0.873	0.844	**0.886**
New	0.632	**0.776**	0.610
Overall	0.811	0.820	**0.822**

TABLE X
COMPARING UNITER-BASED AND BART-BASED MODEL WITH HUMAN PERFORMANCE ON A SUBSET OF 100 RANDOMLY SELECTED EXAMPLES.

Table X shows the human and model performance on the set of 100 segments. We observe that the overall human performance was $F1 = 0.822$, which is very close to the UNITER and BART's $F1$ of 0.811 and 0.820 on the same test set. The annotators did better than the models on *mentioned* objects and worse on *new* objects, achieving only $F1 = 0.610$.

Although the human evaluation was performed on a small dataset, the results suggest that the improved models are already achieving near human-level performance on this dataset.

VI. ERROR ANALYSIS

We manually examined errors made by the models and by the human annotators. The following examples illustrate the common errors. In the following images, red rectangle indicates the ground truth and green rectangle indicate the model's prediction.

[5]The users did not have access to the non-visual metadata features which the models had and which lead to a 1.5% absolute increase for the BART model (see Table VIII).

Missing references: Figure 4 illustrates an example where a model is able to recognize the references in a user's utterance. The referred objects receive a higher probability to the objects #6 and #12, however its below threshold and are not considered as a positive prediction. This type of error is effectively assessed this with the auxiliary task of predicting the number of references heuristics to post-process the models predictions. Objects framed with a green box are examples of predicted referred items and objects framed with a red box are the ground-truth targets.

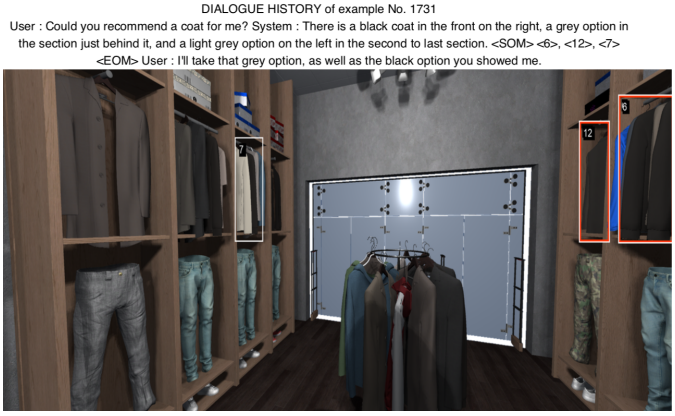

Fig. 4. Two referred coats are not identified, marked in red squares.

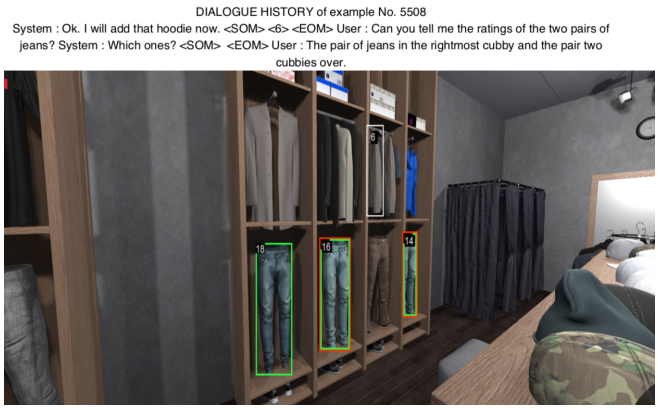

Fig. 5. All jeans are predicted because they have equal descriptions.

False positives: Figure 5 illustrates a common mistake where referring expressions describing relative locations are leading to false positive prediction of object #18. In the image, several jeans are shown in the store's cubby, but just two of them are the true targets. The model, however, predicts all jeans as referred objects.

While pre-trained large language models might be able to know spacial expressions like *on the left*, *on top of something'*, etc., they still struggle to identify descriptions such as *'the rightmost cubby'* and *'two cubbies over'*. Not only the expressions are more complex, but also they are dependent on potentially unknown items, such as a cubby, a rack or a shelf. Large language models may not be pre-trained

on such vocabulary, and SIMMC2 does not include metadata descriptions of the background objects that are not part of the domain merchandise.

Annotation errors: We have observed that some examples of SIMMC2 dataset are not correctly annotated, lowering the performance of automatic metrics where the prediction may have been correct. Figure 6 shows an example where the ground-truth targets are missing when they should be present. The user is clearly referring to the area rug and the model correctly identifies it.

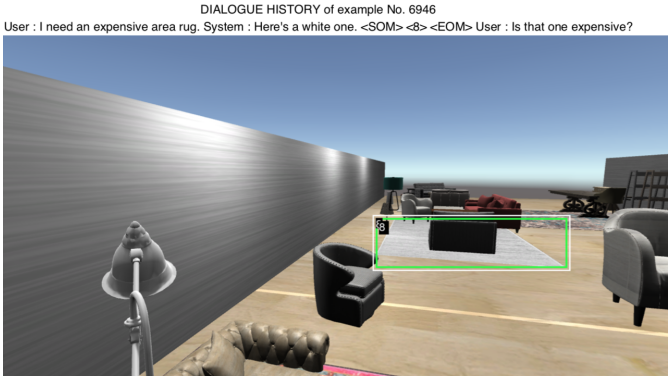

Fig. 6. The example is wrongly annotated: it has no ground-truth targets.

Human annotator errors: The human annotators have an average turn-level error rate of 12%. In fact, 7 examples out of 100 were wrongly classified by all three the human annotators, and 4 examples out of 100 contained errors by two of the annotators. This indicates that the annotators are struggling mainly with the same subset of examples. The first source of errors comes from the fact that SIMMC2 contains annotation noise, as illustrated in Fig. 6. On the other hand, some of the correctly annotated examples are difficult to resolve because either an object is too small, occluded, or cropped, or the color is not clearly displayed, or the reference is ambiguous.

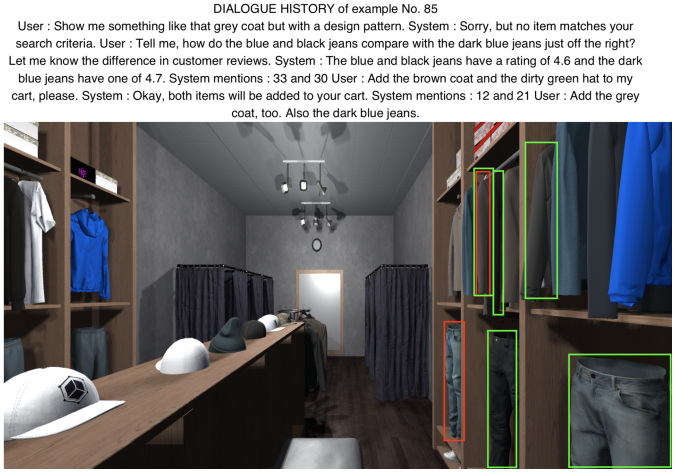

Fig. 7. The annotators fail to identify correctly all referred objects since there are multiple options matching the description.

Figure 7 shows an ambiguous example where all three human annotators failed to identify the ground truth targets (the pair of jeans at the far back of the first cubby). Most annotators usually targeted the trousers in the middle or the rightmost part since they are darker and the dialog mentions some "dark blue jeans". Moreover, the other referred object is a grey coat, but there are plenty of items matching that description in the upper shelf.

To sum up, we have seen that several examples are hard to tackle since there are many objects matching the provided description in the conversation. Additionally, the dataset contains incorrectly annotated examples or some ambiguities that can be solved for future versions. However, the models are still struggling with some cases that a human could solve correctly.

VII. Conclusion

In this work we address multimodal entity (or coreference) resolution, an essential component for a situated system with access to visual and speech input. We analyze the performance of unimodal and multimodal transformer-based approaches on the SIMMC2.0 dataset [22]. Two of the publicly released models that participated in the DSTC10 competition were used in our analysis, a multimodal UNITER-based model [3] and a unimodal BART-based model [5]. Our experiments show that the unimodal approach outperforms the multimodal one on *in-domain* experiments, indicating that contrastive learning of attributes from metadata employed by the authors is an effective way to learn visual features. However, this method is only effective when objects are seen in training. As expected, the multimodal approach outperformed the unimodal one in a *cross-domain* setting, showing a potential for generalization.

Analyzing the performance on objects *mentioned* in the previous discourse and *new* objects, we reveal that the unimodal approach performs better on *new* references while the multimodal one performs better on previously *mentioned* objects. We introduced a new auxiliary task predicting the number of objects mentioned in an utterance and heuristic adjustment of the models' output. The experiments showed that our proposal leads to improvement for both models on *mentioned* objects. A combination of the two methods where we use the unimodal prediction on *mentioned* and multimodal prediction on *new* objects achieves $F1 = 0.80$, a new SOTA on coreference prediction on the SIMMC2.0 dataset. Finally, our analysis of human performance on this task suggests that the current models perform already on par with people on a subset of dialogues sampled from this dataset.

References

[1] S. Stoyanchev, S. Keizer, and R. Doddipatla, "Action state update approach to dialogue management," *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 7398–7402, 2021.

[2] I. Roesiger, S. Stehwien, A. Riester, and N. T. Vu, "Improving coreference resolution with automatically predicted prosodic information," in *Proceedings of the Workshop on Speech-Centric Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, Sep. 2017, pp. 78–83. [Online]. Available: https://aclanthology.org/W17-4610

[3] Y. Huang, Y. Wang, and Y. Tam, "UNITER-based Situated Coreference Resolution with rich multimodal input," *Computing Research Repository (CoRR)*, 2021.

[4] F. J. Chiyah-Garcia, A. Suglia, J. Lopes, A. Eshghi, and H. Hastie, "Exploring multi-modal representations for ambiguity detection & coreference resolution in the SIMMC 2.0 challenge," *Computing Research Repository (CoRR)*, 2022.

[5] H. Lee, O. J. Kwon, Y. Choi, J. Kim, Y. Lee, R. Han, Y. Kim, M. Park, K. Lee, H. Shin, and K.-E. Kim, "Tackling situated multi-modal task-oriented dialogs with a single transformer model," in *AAAI Conference on Artificial Intelligence (AAAI) DSTC10 Workshop*, 2022.

[6] M. Poesio, R. Stuckardt, and Y. Versley, *Anaphora Resolution - Algorithms, Resources, and Applications, Theory and Applications of Natural Language Processing*. Springer, 2016.

[7] A. A. et al., "he hcrc map task corpus," *Language and Speech*, 1991.

[8] W. M. et al., "Colors in context: A pragmatic neural model for grounded language understanding," *Transactions of the Association for Computational Linguistics*, 2017.

[9] A. J. Stent and S. Bangalore, "Interaction between dialog structure and coreference resolution," *2010 IEEE Spoken Language Technology Workshop*, 2010.

[10] L. E. A. et al., "Frames: A corpus for adding memory to goal-oriented dialogue systems," *Association for Computational Linguistics*, 2017.

[11] K. van Deemter and R. Kibble, "On coreferring: Coreference in MUC and related annotation schemes," *Computational Linguistics*, vol. 26, no. 4, pp. 629–637, 2000. [Online]. Available: https://aclanthology.org/J00-4005

[12] W. M. Soon, H. T. Ng, and D. C. Y. Lim, "A machine learning approach to coreference resolution of noun phrases," *Computational Linguistics*, 2001.

[13] S. Bergsma and D. Lin, "Bootstrapping path-based pronoun resolution," in *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Jul. 2006.

[14] S. Kottur, J. M. F. Moura, D. Parikh, D. Batra, and M. Rohrbach, "Visual coreference resolution in visual dialog using neural module networks," *European Conference on Computer Vision (ECCV)*, 2018.

[15] V. Ramanathan, A. Joulin, P. Liang, and L. Fei-Fei, "Linking people in videos with "their" names using coreference resolution," *European Conference on Computer Vision (ECCV)*, 2014, http://vision.stanford.edu/pdf/vignesh14.pdf.

[16] P. H. Seo, A. Lehrmann, B. Han, and L. Sigal, "Visual reference resolution using attention memory for visual dialog," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017, p. 3722–3732.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proceedings of the 31st International Conference on Neural Information Processing Systems*. Curran Associates Inc., 2017.

[18] Q. Zheng, X. Diao, J. Cao, X. Zhou, Y. Liu, and H. Li, "Multi-modal coreference resolution with the correlation between space structures," *arXiv: Artificial Intelligence*, 2018.

[19] A. Kumar, J. Aurisano, B. Di Eugenio, A. Johnson, A. Alsaiari, N. Flowers, A. Gonzalez Martinez, and J. Leigh, "Multimodal coreference resolution for exploratory data visualization dialogue: Context-based annotation and gesture identification," *Workshop on the Semantics and Pragmatics of Dialogue*, 2017.

[20] I. Van der sluis, S. Luz, W. Breitfuß, M. Ishizuka, and H. Prendinger, "Cross-cultural assessment of automatically generated multimodal referring expressions in a virtual world," *International Journal of Human-Computer Studies*, vol. 70, p. 611–629, 09 2012.

[21] S. Moon, S. Kottur, P. Crook, A. De, S. Poddar, T. Levin, D. Whitney, D. Difranco, A. Beirami, E. Cho, R. Subba, and A. Geramifard, "Situated and interactive multimodal conversations," in *Proceedings of the 28th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Dec. 2020.

[22] 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*. Association for Computational Linguistics, Nov. 2021.

[23] Unity, "Unity technologies," 2019.

[24] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," *Association for Computational Linguistics (ACL)*, 2019.

[25] Y.-C. Chen, L. Li, L. Yu, A. E. Kholy, F. Ahmed, Z. Gan, Y. Cheng, and J. Liu, "Uniter: Universal image-text representation learning," *European Conference on Computer Vision (ECCV)*, 2019.

[26] 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*. Association for Computational Linguistics, 2020.

[27] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "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*. Association for Computational Linguistics, 2019.

[28] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI documentation*, 2019.

[29] S. Agarwal, O. Dusek, I. Konstas, and V. Rieser, "A knowledge-grounded multimodal search-based conversational agent," *Proceedings of the DSTC10 Workshop in the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI 2022)*, 2018.

[30] J. Lee and K. Han, "Multimodal interactions using pretrained unimodal models for simmc 2.0," *Computing Research Repository (CoRR)*, 2021.

[31] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever, "Learning transferable visual models from natural language supervision," in *Proceedings of the 38th International Conference on Machine Learning*. PMLR, 2021.

[32] T. Lin, M. Maire, S. J. Belongie, L. D. Bourdev, R. B. Girshick, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: common objects in context," *European Conference on Computer Vision (ECCV)*, 2014.

[33] R. Krishna, Y. Zhu, O. Groth, J. Johnson, K. Hata, J. Kravitz, S. Chen, Y. Kalantidis, L. Li, D. A. Shamma, M. S. Bernstein, and L. Fei-Fei, "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *International Journal of Conflict and Violence (IJCV)*, 2016.

[34] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," *Association for Computational Linguistics*, 2018.

[35] V. Ordonez, G. Kulkarni, and T. Berg, "Im2text: Describing images using 1 million captioned photographs," *Neural Image Processing Systems (NeurIPS)*, 2011.

[36] A. Maurer, M. Pontil, and B. Romera-Paredes, "The benefit of multitask representation learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, p. 2853–2884, jan 2016.

[37] J. Cohen, "A coefficient of agreement for nominal scales," *Educational and Psychological Measurement*, 1960.