SIMMC-VR: A Task-oriented Multimodal Dialog Dataset with Situated and Immersive VR Streams

Te-Lin Wu[1]*, Satwik Kottur[2], Andrea Madotto[2], Mahmoud Azab[2]
Pedro Rodriguez[3], Babak Damavandi[2], Nanyun Peng[1], Seungwhan Moon[2]
[1]University of California, Los Angeles, [2]Meta Reality Labs, [3]Meta AI
{telinwu,violetpeng}@cs.ucla.edu,
{skottur,andreamad8,azab,par,babakd,shanemoon}@meta.com

Abstract

Building an AI assistant that can seamlessly converse and instruct humans, in a user-centric situated scenario, requires several essential abilities: (1) spatial and temporal understanding of the situated and real-time user scenes, (2) capability of grounding the actively perceived visuals of users to conversation contexts, and (3) conversational reasoning over past utterances to perform just-in-time assistance. However, we currently lack a large-scale benchmark that captures user↔assistant interactions with all of the aforementioned features. To this end, we propose SIMMC-VR, an extension of the SIMMC 2.0 dataset to a video-grounded task-oriented dialog dataset that captures real-world AI-assisted user scenarios in VR. We propose a novel data collection paradigm that involves (1) generating *object-centric* multimodal dialog flows with *egocentric* visual streams and visually-grounded templates, and (2) manually paraphrasing the simulated dialogs for naturalness and diversity while preserving multimodal dependencies. To measure meaningful progress in the field, we propose four tasks to address the new challenges in SIMMC-VR, which require complex spatial-temporal dialog reasoning in active egocentric scenes. We benchmark the proposed tasks with strong multimodal models, and highlight the key capabilities that current models lack for future research directions.

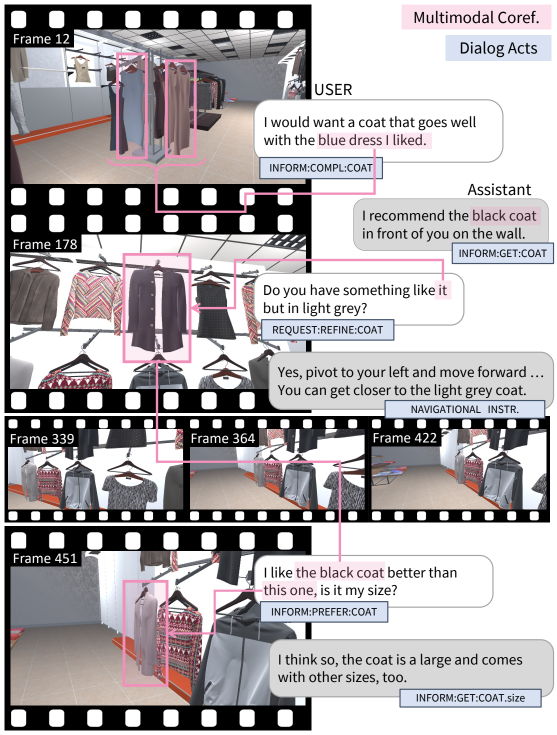


Figure 1: **SIMMC-VR** is a **S**ituated **I**nteractive **M**ulti-**M**odal **C**onversation dataset that features task-oriented user↔assistant dialogs *streamed immersively* in a virtual-reality (**VR**) environment. The dataset is created on programmed realistic shopping scenarios and actively-rendered photorealistic user visual observations, which brings new challenges for complex spatial-temporal reasoning on the multi-modal interactions (visual cues and grounded-dialogs).

1 Introduction

With the growing popularity of smart glasses, studies on visually grounded conversational agents have gained significant interest. For instance, SIMMC 2.0 (Kottur et al., 2021) introduces an image-grounded, task-oriented dialog (TOD) dataset where an assistant agent co-observes the user's egocentric viewpoint to aid with user requests. Many follow-up works (Huang et al., 2021a; Lee et al., 2022; Chiyah-Garcia et al., 2022) focus on challenges around dialog-image grounding, such as visual coreference resolution (*e.g.* '*the yellow dress behind the rack*') of a static image.

However, several technical gaps still remain in applying prior work to build a real-world, *situated* multimodal assistant (Figure 1). For instance, a typical multimodal user-assistant scenario (with a video capturing capability) would include (1) spatial *and* temporal language references as grounding contexts ('*the shirt I saw earlier <u>when</u> I entered the store*'), (2) actively perceived egocentric motions as part of conversation contexts ("*No – turn around <u>the other way</u>*"), (3) references to conversational memories from past sessions ('*the one I*

*Work done during an internship at Meta Reality Labs.

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 6273–6291
July 9-14, 2023 ©2023 Association for Computational Linguistics

bought earlier', the '*black coat*' in Figure 1 being retroactively mentioned by both the assistant and the user), *etc*. While these scenarios are perceived as the expected capabilities of a next-generation multimodal assistant, our survey of datasets (Sec. 5) highlights that due to the static and constrained nature of the datasets' grounding context, they lack sufficiently complex interactions.

To this end, we present SIMMC-VR, a video-grounded task-oriented dialog dataset comprising $4K$ user↔assistant task-oriented dialogs ($95.3K$ utterances) grounded on diverse photorealistic VR video streams ($4.8M$ frames). For data collection, we propose a novel two-stage approach with: (1) a multimodal interaction simulator that generates egocentric VR streams grounded on *object-centric* multimodal dialog flows, and (2) a manual paraphrasing step for naturalness and diversity while preserving multimodal dependencies between visual scenes and their grounding language. Our pipeline allows for flexible and cost-effective data collection, easily extendable to simulate any other domains given the availability of 3D virtual assets.

To measure progress towards real-world applicability, we propose four SIMMC-VR tasks that address new challenges in complex spatio-temporal dialog reasoning. We then extend state-of-the-art multimodal models to the SIMMC-VR tasks and discuss the limitations of current models.

Our contributions are as follows: (1) we present SIMMC-VR, a video-grounded task-oriented dialog dataset ($95K$ utterances over $4.8M$ frames) targeted towards real-world applications for an assistant on smart glasses. (2) We propose the tasks with complex spatio-temporal conversational dependencies, and benchmark them by extending the state-of-the-art multimodal models. (3) Our data collection platform allows creation of a similar dataset in any target domains.

2 SIMMC-VR Dataset

SIMMC-VR is *actively* multimodal, where each data instance is a video from a user's **egocentric** viewpoint recording all interactions within a virtual shopping environment, densely paired with dialog utterances and essential attributes. Each task-oriented dialog mimics real-world shopping scenarios where the assistant's goal is to help the user make purchases and navigate through the environment. In each instance, the user walks around a virtual shop while the assistant provides product information or recommendations; as well as help the user locate and navigate to products of interest.

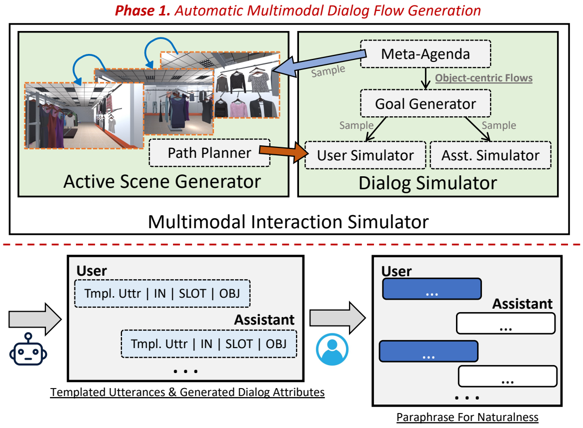


Figure 2: **Dialog generation flow: (Upper half)** a **meta-agenda** is firstly programmed to sample an *object-centric* flow (grounded in the environment), which is used by the goal generator to sample high-level dialog goals. These goals are then used by both user and assistant simulators to synthesize templated utterances, which are then manually paraphrased by linguistic experts for diversity and naturalness **(lower half)**.

Dataset Collection Strategy. Multimodal or embodied dialogs (Das et al., 2017a; Padmakumar et al., 2022) are often constructed via a two-player game where participants interact with the *environment* and *converse* with each other (*i.e*. in a Wizard of Oz (WOZ) (Mrkšić et al., 2017; Budzianowski et al., 2018a) role-playing fashion). However, it can be overly challenging to require annotators to role-play as the AI assistant in our complex and quite cluttered VR shop environments (>100 products). Furthermore, to match the potential retroactive reasoning shopping scenarios (*e.g*. concerning products priorly seen/mentioned), it could add much mental burden for annotators to memorize object attributes and their locations **while** composing *authentic* long dialogue interactions. Lastly, in conjunction with the aforementioned difficulties, it is rather unscalable and inextensible to manually annotate all the required labels (dialog acts, coreferences) cross-referencing complex moving scenes for a *task-oriented* dialog dataset.

We therefore collect the dataset through two phases: (1) **simulating multimodal dialog flows** with templated utterances – thereby programmatically generating fine-grained-scene-grounded annotations and systematically ensuring the diversity of the conversations, and (2) **manual paraphrasing**, which ensures the naturalness of utterances with a significantly less annotation overhead (Rastogi et al., 2020; Shah et al., 2018).

2.1 Multimodal Dialog Generation

Our pipeline for multimodal dialog generation simulates plausible and natural multimodal interactions in a virtual environment (Figure 2), The process is as follows: (1) Decide a **meta-agenda** based on object attributes and traversal routes. (2) Sample specific objects that fulfill the decided agenda as the **object-centric flow**. (3) Perform the user traversal **path planning** and video recording using the sampled objects as starting/ending points. (4) Synthesize the corresponding utterances via prewritten **templates** and the multimodal contexts. (5) Manually **paraphrase** the templated utterances.

We categorize a full dialog instance (generated through the previously described steps) into two phases: (a) **static phase** where the user *mostly* focuses on a specific viewpoint (with a small amount of randomness in movement or eye-gaze) when conversing with the assistant (Section 2.1.2), and (b) **active phase**, where the user navigates to another spot within the environment, at will or following assistant instructions, containing larger movements and actions (Section 2.1.1). The two phases interleave each other, creating a realistic shopping scenario (*e.g.* user walks into a shop, stopping by a few products, and wanders to other ones).

Virtual Environment. Following SIMMC 2.0, we use the same set of photorealistic VR shopping environments in Unity (Unity, 2020), where a set of seed scenes with pre-arranged digital assets (*e.g.* shirts, dresses for *fashion* domain and sofas, tables for *furniture* domain) are programmatically re-arranged into randomized larger sets of scenes.

Table 6 lists the asset (product item) categories used for constructing the SIMMC-VR dataset for both fashion and furniture domains.

2.1.1 Active Scene Simulation

Figure 3ab illustrates the process of simulating visual observations of a user traversal, where a *path planning* is performed (connecting the start and end user position/orientation) in the environment, and the trajectories are rendered into egocentric videos.

Path Planning. Ideally, the navigational guidance should minimize the overall traversal distance (to a target spot), while taking the smoothness of movements into consideration. Given a start and end position in the extracted environment layout, we perform an A*search to plan a trajectory simulating a user's traversal within a shop. Additionally, we modify the standard A*algorithm to minimize the amount of *turning* for smoother and more natural user movements[1], with random noises added to naturally jitter the planned path. We then augment the output path with rotation angles computed to account for the user orientation during the traversal. At each viewpoint on the planned path, a Unity camera snapshot is taken, and the traversal video is rendered by combining all the snapshots.

Referential Objects. Once the intended user-traversal video is planned and recorded, we define *key action points*, using the start/end viewpoints of user movements (*i.e.* displacement or turning actions). Inspired by the natural communication behavior, where we often refer to certain *landmarks* when giving navigational guidance, we derive a set of **referential objects** from objects placed across these viewpoints (*e.g. "Turn left when you see the red shirt."*). Figure 3a illustrates the referential object sampling strategy: (1) Compute the cosine similarity between an egocentric viewpoint (3D) vector (gaze point at the center of yellow dotted lines) and a *look-at* vector to each of the objects within the scene – a higher similarity implies that it is closer to the eye-gaze line of sight, hence more probable to be referenced during conversations. (2) Augment the previously derived rankings with other plausible features such as stronger color contrast with neighboring objects. (3) Lastly, transform these rankings into sampling probabilities (via a *Softmax*) to sample object(s) for reference.

Scene Graphs & Disambiguation. When referring to an item in a cluttered environment, its surroundings often serve as good candidates to *disambiguate* items that may share similar attributes (often useful when users *under-specify* items). In light of this, for each object within the same scene, we build a **local scene-graph** to include the closest three objects to its *left, right, top, bottom* (four main directions). An object can then be referred to with its neighbors when further clarification is needed (*e.g. "Not that one, I mean the white hat **below** the red coat."*).

Scene Metadata. To facilitate templated utterances for paraphrasing (Section 2.1.2) and to formulate a modeling task with visual labels (Section 3), we compute 2D bounding boxes for all 3D assets in a particular viewpoint, where each object is cross-referenced across every frame. As the dense bounding box computation in a 3D environment is time-consuming (repeated for thousands of frames per di-

[1] A*'s distance minimization may lead to excessive turns.

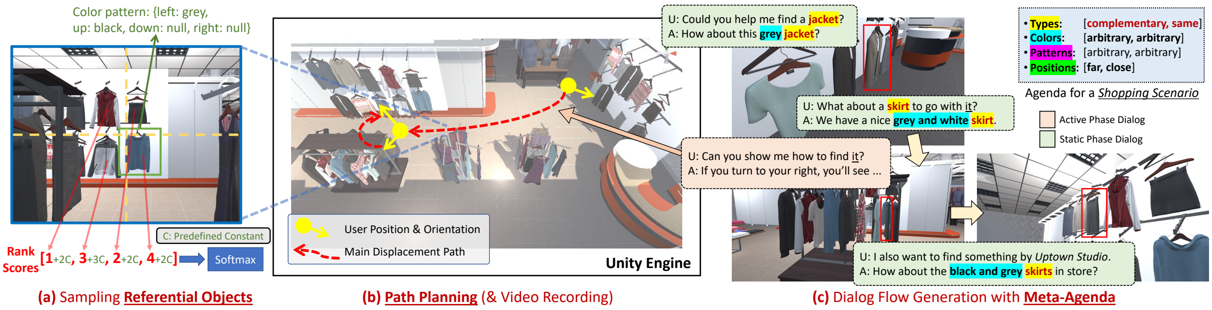


Figure 3: **Multimodal dialog generation: (Right most) meta-agenda** illustrates an exemplar shopping scenario that concerns user demanding *complementary* (*i.e*. can go with) types for the first two items (jacket ↔ skirt) and the *same* type between the 2nd and 3rd items. Colors and patterns are not constrained, while the scenario simulates longer traversal is required (*far*) between the first two items and the latter two are *close*-by. **(Middle) Path planning:** the navigational utterances will be grounded on the planned path (displacements and orientations) and the **referential objects (left most)** used to facilitate the guidance are sampled according to *softmax* scores on a ranking (via features *e.g*. eye-gaze, color-contrast) of most suitable landmarks.

Fashion	hat, tshirt, jacket, hoodie, sweater, shirt, suit, vest, coat, trousers, jeans, joggers, skirt, blouse, tank top, dress, shoes
Furniture	area rug, bed, chair, couch chair, dining table, coffee table, end table, lamp, shelves, sofa

Table 1: **Digital assets categories used in SIMMC-VR** for both fashion and furniture domains.

alog), we expedite this process via an approximate reconstruction. Specifically, we record the camera position and orientation for each video frame, and provide the mesh data for each asset and a function to reconstruct 2D bounding boxes on-the-fly.

2.1.2 Dialog Simulation

In real-life shopping experiences, customers typically explore a shop with certain product attributes of interest in mind (*e.g*. clothing colors, types), thus shopping experiences are often **object-centric** (Yinyin, 2011). Inspired by this, we *program* several (extendable) *object-centric flows* that focus on certain objects within an environment to mimic how a user may wander (self-motivated or guided) around from one product to another.

Dialog Flows. To have full control over the diversity of dialog flows, and to encourage certain patterns of flows to emerge for more interesting user-AI conversations, we propose an *object-centric* generation pipeline. Specifically, to generate an *object-centric flow*, we (1) define a **meta-agenda**, a sequence of **meta-goals**[2] defined by certain object attributes that simulate a complete shopping experience (*e.g*. a customer looking for certain types or colors of clothing, or asking for a complementary item to match a previously purchased one) and (2) for each meta-goal, sample an object according to a planned traversal route (*e.g*. short or long travel distance, traveling back to a previously observed item) and a user-position/orientation to *look at* the object (where the path planning can perform on).[3] The meta-agenda is either human-written or programmatically generated, and diversified while ensuring a balanced distribution of scenarios. The traversal route is engineered to ensure user's navigation/orientation changes are necessary and natural.

Colors	`same`, `arbitrary`
Patterns	`same`, `arbitrary`
Types	`same`, `arbitrary`, `alternative`, `complementary`
Positions	`far`, `close`, `come_back_to_X`

Table 2: **Meta-Agenda Programs**

For each of the sampled-objects, a **goal generator** will sample a high-level dialog *goal* to define the theme of a few turns of utterances (*e.g*. `COMPARE` → user requesting product comparisons). The **user simulator** then utilized both the sampled objects and goals to generate corresponding NLU labels following a probability distribution, consisting of user intents (*e.g*. `INFORM:GET`), request slots (*e.g*. *color*, *brand*) and object references. The **assistant simulator** then resolves the user requests, leveraging the multimodal context and the simulation API (*e.g*. for info lookup).[4]

[2]We cap the max sequence length at 3, *i.e*. 3 *meta*-goals.

[3]Each flow is uniquely defined by the sampled object-sequence. We over-sample totally $>1K$ object-centric flows evenly across 27 programmed meta-agenda (Figure 3c).

[4]In contrast, SIMMC 2.0 plans a dialog *only* by randomly sampling a sequence of abstract goals (*e.g*. `BROWSE` → `GET_INFO` → `...`), often resulting in unrealistic scenarios.

Meta-Agenda. Table 2 lists the candidates that can be programmed into the *meta-agenda*. For *alternative* and *complementary* item mappings, we consider: (1) Relations in ConceptNet 5.0 (Speer et al., 2017) such as `distinct_terms` (*jacket* is `distinct_to` *coat*), `similar_terms` and/or `related_terms` (*e.g. sofa* is `related_to` *end-table*). And (2) Manual inspections and annotations, where we ask internal members to annotate the alternative and complementary items to a particular one of interest, and refine the annotated list with majority vote (*e.g. hat* is complementary to both *shirt* and *dress* as they can go in pairs, and *coat* is alternative to *jacket* as they share similar functionalities and thus can complement each other).

For the *positions* agenda, we pre-define a distance threshold to denote far or close depending on the environment room layout (differ in fashion and furniture domains). For the `come_back_to_X` program, we engineer that the user will traverse back to an item that is previously seen and indicated with interests, to simulate relevant shopping experiences in the real-world.

Templated Utterances. Grounded by the multimodal context, we pre-define a few utterance templates each associated with a specific dialog act, leaving the specific object-related information (*e.g.* object ids, modifiers, pronouns) as placeholders that are filled-in according to the visuals. This allows us to easily sample an utterance template that is suitable for a particular situation and the associated user or AI intention, determined by the dialog act. We list a few exemplar utterances and their paraphrases, and highlight the placeholders in Table 3. Notice that the local object scene-graphs (Section 2.1.1) are also useful for generating diverse reference expressions for the same object (second role of the Assistant examples in Table 3).

Manual Paraphrase. Next, we ask human annotators to paraphrase the templated utterances to better match the real-world natural language distribution. We design an interface that dynamically displays a multimodal scene that features either a still image (static dialog phase) or a user egocentric video (active dialog phase). When clicking on a specific turn of a dialog, the corresponding visual input is shown in the display panel to help annotators navigate through the entire dialog flow. We ask the annotators to pay attention to detailed and sophisticated spatial-temporal relations of objects and encourage writing interesting shopping experiences. The paraphrases are collected from more than 20 different linguistic experts for diverse language patterns/usages.

Once manual paraphrases are collected, we perform text-to-speech synthesis (TTS) on the utterances, and synchronize the speech with the relevant motion renders for improved naturalness, making the rendered user shopping videos more realistic (and comprehensive). We use an open-sourced tool, *Coqui TTS* (Coqui.ai, 2022) to generate the spoken speech from the paraphrased utterances. This also helps computing the natural duration of each utterance when spoken so that we can interpolate certain number of video frames (under a fixed frame-rate) to fit such utterance would span.

An exemplary dialog is shown in Appendix. A.

Dialog Dataset Structures. Similar to other existing task-oriented dialog systems (Eric et al., 2019; Rastogi et al., 2020; Moon et al., 2020), each turn of SIMMC-VR's dialog data consists of NLU (and NLG) intent and slot labels (*e.g. "How do their prices compare?"* $\rightarrow$ `REQUEST:COMPARE`, *slots: price*, *objects: [1, 4]*), as well as object references (a unique object ID across the same room environment) like SIMMC 2.0. In SIMMC-VR, due to the newly introduced *active dialog phase* and the richer dialog scenarios (*object-centric* flows), the list of intents is expanded as compared to SIMMC 2.0 (see Section 2.2 and Appendix. A.2).

2.2 SIMMC-VR Dataset Analysis

Table 4 shows the essential dataset statistics. In total, SIMMC-VR contains $4K$ dialogs with the corresponding videos (equating to $95.3K$ utterances).

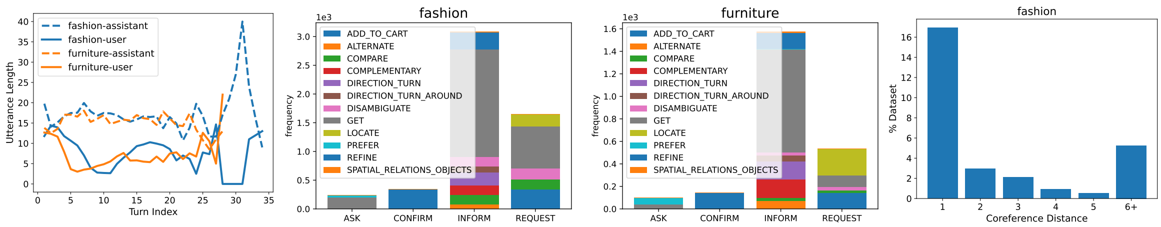
Videos. We set the frame per second (fps) as 10.0, which roughly leads to an average of $1.2K$ frames per video ($\sim$2 minutes length). On average there are 24.6 visible objects in the key video frames.

Dialog Acts & Flows. Each algorithmically generated flow, *i.e.* the **meta-agenda**-induced *object-centric flow* (Section 2.1.2), is capped to have at most 5 different dialogs with randomly sampled dialog goals and intents. The average number of utterances is 23.4, significantly larger than that in SIMMC 2.0 (10.4). Its length distribution over different turns is shown in Figure 4a. SIMMC-VR extends SIMMC 2.0's annotation to a set of 5 dialog acts (*e.g.* `INFORM`, `REQUEST`) and 17 activities (*e.g.* `REFINE`, `DIRECTION_TURN`). Figure 4b shows their frequency breakdown and the complete lists are in Appendix. A.2. A visualization of dia-

Role	Dialog Goal & Act	Example Templates & Paraphrases
User	BROWSE REQUEST:GET	Could you recommend something with {type:blouse}$_{\text{[search-filter]}}$? ⇒ *'I am looking for a blouse; do you have anything to show me?'*
	ALTERNATE_SEARCH INFORM:ALTERNATE	Do you have alternatives to [OID:34(hoodie,blue)]$_{\text{[object]}}$ with {color:violet}$_{\text{[search-filter]}}$? ⇒ *'Any other options besides that? See if you have anything violet in store.'*
	REFINE_SEARCH INFORM:REFINE	I would like to refine my search to include {type: skirt}$_{\text{[search-filter]}}$. Anything good here?. ⇒ *'I want to search more specifically for skirts. What are my options now?'*
	ADD_TO_CART REQUEST:ADD_TO_CART	Please add to cart: [OID:50(hoodie,green), OID:50(hoodie, green)]$_{\text{[object]}}$. ⇒ *'I like the first hoodie the best. Give me two of the green one.'*
Assistant	ACTION INFORM:DIRECTION_STRAIGHT	Go {towards}$_{\text{[direction]}}$ it. [OID:100(sweater,red)]$_{\text{[object]}}$ will be on {far-left}$_{\text{[relation]}}$. ⇒ *'Go straight forward until seeing a red and white sweater on your far left.'*
	ACTION INFORM:DIRECTION_TURN	Turn {around}$_{\text{[direction]}}$ and you will be able to see [OID:141(blouse,white)]$_{\text{[object]}}$, which is {on-right}$_{\text{[relation]}}$ to [OID:154(jacket,black)]$_{\text{[object]}}$. ⇒ *'Turn around and you will see that white and black blouse, on its left is a black jacket.'*
	GET_INFO INFORM:GET	Here is the info on size: [OID:49(hat,green)]$_{\text{[object]}}$: {size:XS}$_{\text{[slot-values]}}$. ⇒ *'That green hat you're looking at is size XS.'*
	COMPLEMENTARY_SEARCH INFORM:COMPLEMENTARY	How about these: [OID:77(skirt,brown)]$_{\text{[object]}}$? They are {type:skirt}$_{\text{[search-filter]}}$. ⇒ *'Yes we do. How about the brown skirt that is on the far right on the top row?'*

* OID stands for object ID.

Table 3: **Exemplar utterance template and paraphrases** in SIMMC-VR. In each row under the second column, the upper terms are the goals and the lower terms are the dialog acts (consisting of acts and activities). We show a few representative dialog acts with their corresponding sample templates (each act may have multiple templates as options) and a sample paraphrase. In each template, the subscripts denote the type of the placeholders, where the contents are filled-in grounded by the multimodal contexts (*e.g*. sampled objects, user eye-gazes) or sampled attributes (*e.g*. types or colors of the desired item).



(a) (b) (c)

Figure 4: **Plots of:** (a) utterance lengths in dialogs, (b) acts and activities, and (c) co-reference distance between object mentions.

Total # dialogs	4,075
Total # utterances	95,368
Avg # words per user turns	12.9
Avg # words per assistant turns	16.7
Avg # utterances per dialog	23.4
Avg # objects mentioned per dialog	13.2
Avg # objects in key video frames	24.6
Avg # objects per fashion environment	188.6
Avg # objects per furniture environment	62.0
Avg # frames (under fps = 10.0)	1197.7
Avg # seconds per TTS utterance	4.13

Table 4: **SIMMC-VR dataset statistics**. On average there are 13.2 objects mentioned in a dialog and more than 20 visible in each video frame, making the video-grounded dialogs diverse and rich in contents. Each video roughly lasts 2 minutes, equating to a total of >130 hours long VR streams.

log transition is shown in Figure 5 to illustrate the diversity and patterns of our generated dialog flows. Figure 4c plots the coreference distances according to how many utterances separate the mentions.

2.3 Novel Challenges to SIMMC 2.0

SIMMC 2.0 shares the general goal of achieving multimodal task-oriented dialog systems for future real-world and VR applications. However, the active and rich multimodal contexts of SIMMC-VR introduce the following new challenges: (1) Anchoring *egocentric videos* as visual contexts, SIMMC-VR requires the spatial and the additional temporal multimodal reasoning, posing new categorical patterns of object coreferences and associated user/assistant utterances. (2) The novel dialog simulation pipeline allows for more diverse and realistic interactions (*e.g*. navigation and localization scenarios) with a number of transitory dialog actions and viewpoints, many of which have not been studied in the previous datasets. This results in the higher degree of complexities in conversational tasks – for instance, the coreference resolution task gets significantly harder with a much larger number of objects mentioned in a dialog (13.2 *vs.* 4.7 in 2.0), and with the increased average utterance counts (23.4 *vs.* 10.4 in 2.0). (3) SIMMC-VR requires that a perception model maintains object correspondences across their variations from dif-

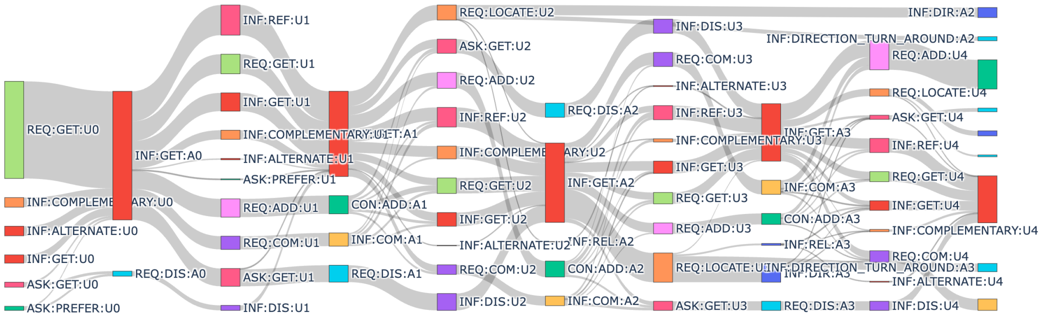


Figure 5: **Dialog act(s) transitions** for the first four rounds of dialogs in the fashion domain. The acts and activities are denoted for brevity as `ACT:ACTIVITY:[A|U][turn_index]` (for full names see Appendix. A.2), where `U` and `A` denote user and assistant, respectively. The shown branching and inter-connectivity justifies the diversity of the synthesized dialog flows.

ferent angles and disjoint viewpoints over time, to ensure the correctness of their resolution. While this requirement poses a practical challenge for a real-world application, a robust solution has not been explored especially for its use in the context of the multimodal dialog management.

3 SIMMC-VR Task Formulation

The SIMMC-VR is created to help AI models cope with realistic shopping scenarios and assist human users in real-world applications in AR/VR. To investigate the (multimodal) conversational and assistive abilities of current AI systems in this immersive and situated environment, we propose four main bechmarking tasks leveraging the created dataset. Several tasks inherit from SIMMC 2.0 with additional challenges brought by the nature of active user scenes and expanded dataset annotations.

3.1 Multimodal Dialog State Tracking

Following SIMMC 2.0, in SIMMC-VR we retain the multimodal dialog state tracking (**MM-DST**) task, which aims at inferring structured information for understanding and planning out dialog policies/actions, with dialog utterances and/or multimodal contexts given. Each DST is required to resolve both the dialog intents (as a dialog *act*) and the user request slots, which is mainly evaluated by the F1 scores of the predicted slots and intents.

3.2 Multimodal Coreference Resolution

It is crucial for an assistant to be able to recognize objects that a user is referencing, either within the **current visual context**, or any **previously mentioned items**.Therefore, for each environment, a canonical ID is uniquely assigned to each object as the target for multimodal coreference (**MM-Coref**) resolutions, where the mentions can be resolved by both the dialog context (*e.g. "Add the shirt I liked to the cart."*) and the multimodal context (*e.g. "How does* ***the red shirt next to the jeans*** *compared to the one before?"*). Following SIMMC 2.0, we allow the models to take ground-truth bounding boxes as inputs to bypass the needs for perfect visual detectors. The evaluation metric is the F1 scores for the predicted object IDs. Note that as the multimodal contexts are videos, the models are implicitly conditioned to identify the frames that likely contain the target objects, leading to comprehensive multimodal spatial-temporal reasoning. Additionally, while there are no explicit textual coreference annotations, the models are still implicitly required to perform textual coreference resolution for those utterances mentioning the same objects from prior dialogue turn(s).

3.3 Failure-Mode Prediction

SIMMC-VR features user failure-modes that simulate users accidentally failing to correctly follow the assistant guidance. In this task, given a dialog snippet (consisting of utterances in the *active phase*) and the video frames surrounding it, we ask the model to predict whether the current user actions correctly follow the instructions or not (*i.e.* binary classification evaluated by F1 scores). The task is highly multimodal as the model needs to understand the sophisticated active grounding of the visual and dialog contexts. During the training time, we pre-sample the same amount of negative samples to make the labels balanced.

3.4 Dialog Response Generation

This task requires a trained dialog agent to generate the assistant responses (measured in BLEU-4 (Papineni et al., 2002)), given user utterances as well as the *resolved* multimodal information (belief states and referred canonical object IDs). Note that even though the aforementioned information is given as ground-truths, the generation still needs to conform to natural language responses that do not contain flattened DSTs or object IDs (*e.g.* `INFORM:COMPARE, (OBJ_ID:`

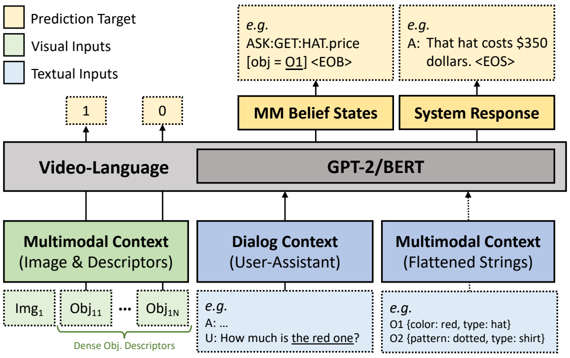


Figure 6: **Baseline models:** The inner grey box (denoted "GPT-2/BERT") is the language model either as (is) the MM-DST model or the language encoder of the video-language model (VIOLET adopts BERT). The video-language model predicts MM-Coref via dense object descriptors, while MM-DST model generates (via GPT-2) the flattened target strings.

`5,9)` → *"The white and blue shirts differ by ..."*).

4 Modeling & Experimental Analysis

In this section, we introduce the investigated baseline models to perform a preliminary benchmarking of the proposed dataset, where we hope to inspire more sophisticated and tailored modeling efforts from the community for future research.

Dataset Split. For the empirical modeling analysis and performance benchmarking, we randomly split the dataset into 3 sets: train (70%), dev (5%), and test (25%) sets, while ensuring both domains (fashion and furniture) have the same split distributions.

Baselines. To benchmark the dataset, we adopt:
(a) MM-DST Model is a 12-layered multi-task GPT-2 model (Radford et al., 2019; Kottur et al., 2021) trained with joint supervision signals from MM-Coref, MM-DST, and response generation tasks, inspired by causal language modeling approach to dialog systems (Peng et al., 2020; Hosseini-Asl et al., 2020). The inputs to the model include both the dialog context (utterances) and the multimodal contexts flattened as structurally formatted text strings, where the outputs are the predicted DST labels. This baseline has two versions: one uses the ground-truth multimodal contexts provided from the scene generator (hence a soft oracle) to simulate the outputs from a robust object detector or from a controlled VR environment, whereas the other has to *infer* visual descriptors from raw videos, simulating real-world scenarios.
(b) Adapted-VIOLET Model is a multimodal video-language model based on VIOLET (Fu et al.), adapted to fit our task structure (Figure 6). Due to computational limitations, we randomly sub-sample $10-15$ video frames during training (while ensuring a proportion of these frames contain objects of ground-truth coreferences), and sweep through the entire video for test-time inference with a fixed window-size. In addition to the frame-level whole image feature, we feed the dense object descriptor features extracted in each ground-truth bounding boxes (assuming a perfect object detector) to the model for the MM-Coref task.[5]

All baseline models are trained for ten epochs, and the best model on the dev set is used for test.

Model	DST Slot / Int. / Joint F1↑	Coref F1↑	Fail. F1↑	Gen. BLEU↑
(Label Distribution)	19.4 / 9.39 / 8.73	0.66	34.1	—
MM-DST	72.4 / 78.6 / 33.9	17.1	—	0.117
MM-DST (no-gt.)	71.7 / 77.3 / 30.8	0.71	—	0.120
Adapt.-VIOLET	75.0 / 80.4 / 37.7	9.69	46.4	0.119
SIMMC 2.0 Performance (for comparison)				
MM-DST	89.6 / 94.5 / 44.6	36.6	—	0.192

Table 5: **Baseline performances** for Multimodal (1) Dialog State Tracking (DST), (2) Object Coreference (Coref.), (3) Response Generation (Gen.), and (4) Failure Mode Prediction (Fail.). In the lower half, we report the corresponding performance from SIMMC 2.0 with the MM-DST model.

4.1 Experimental Results

Table 5 summarizes the model performance and the probabilistic guess performance (proportional to training label distributions) for each sub-task.

Main Results. The baselines show strong overall performances especially in the DST task. The MM-Coref is understandably a very challenging task (resolving tens of items over moving frames), as evidenced in the relatively low scores – suggesting areas for future research. It is worth noting that without the ground truth multimodal contexts for assistant turns, the MM-DST model performs close to zero, indicating that the created dataset does not leak unintended artifacts for the object mentions (that language-only models can easily exploit without visual contexts). For the failure mode prediction, we prepare a test-set that focuses on the active scene utterances, where the random guess roughly equates to the amount of the failure probabilities (30%). We expect the future modeling efforts can better perceive discrepancies between the visual behaviors and the instructed guidance.

Effects of Temporal Grounding. We break down the MM-Coref performance by identifying coref utterances with *temporal dependencies*. With the Adapted-VIOLET model, we get an F1 of 10.5 for

[5]Here to simplify the task, our dataset can also be approached without assuming any perfect vision modules.

utterances *without* temporal dependencies, and a significantly lower 2.81 for the others – suggesting the difficulty in encoding long-standing contexts.

Comparison with 2.0. We also include the MM-DST model performance for the SIMMC 2.0 dataset as a reference, to signify the new challenges that SIMMC-VR brings with the active VR-streams and the complex multimodal dialog flows.

5 Related Work

The proposed work addresses unique requirements for a task-oriented assistant on smart glasses, making it a first-of-its-kind – while complementing other related works within multimodal NLP.

SIMMC (Moon et al., 2020; Kottur et al., 2021) is a class of research areas that the proposed work builds upon, which addresses using virtual environments to simulate a co-observing multimodal dialog agent. Moving away from the sanitized and static scenes that they concern for the limited use cases, SIMMC-VR introduces several additional challenges as summarized in Section 2.3.

Several models (Kung et al., 2021; Senese et al., 2021; Lee and Han, 2021; Huang et al., 2021b) are proposed for the SIMMC benchmark tasks – primarily focusing on grounding dialogs on visual objects from a single image. Taking inspirations from these works, we extend the models to accommodate temporal dependencies within frames.

Multimodal Dialog Datasets. Many of the existing literature in multimodal dialogs (Das et al., 2017b; Hori et al., 2018; Kottur et al., 2019; de Vries et al., 2017, 2018; Le et al., 2021) typically assume asymmetric visual information between two observers, *i.e. questioner* and *answerer*, where conversational goals are limited to reducing information asymmetry (similar to VQA). In contrast, we study task-oriented dialog scenarios – an assistant co-observes the same scene as a user does, thus focusing on serving user requests to achieve functional goals (*e.g.* giving recommendations).

The embodied AI dialog systems (Gao et al., 2022; Padmakumar et al., 2022), on the other hand, study the scenarios where a human participant *teaches* an AI agent a set of skills or gives navigational directions – hence posing an opposite role to an AI agent. While it is an important area to study, its distribution of utterance patterns is completely different and therefore not applicable for our target domain – building a situated AI *assistant*.

Egocentric Video Datasets. With the popularity of wearable devices, several datasets (Grauman et al., 2022; Lv et al., 2022; Damen et al., 2021) are released to study the unique properties of egocentric videos. Our work also features similar visual properties, while adding conversational layers that showcase an assistant use case of such videos.

Task-Oriented Dialog Systems (Henderson et al., 2014; Rastogi et al., 2019; Budzianowski et al., 2018b; Eric et al., 2019) have long been studied to support various assistant scenarios (*e.g.* booking hotels). Our work takes its roots in this line of work – focusing on predicting user belief states and dialog acts to achieve functional goals – and extends it to a unique multimodal setting.

A popular thread in the task-oriented dialog system modeling is to fine-tune end-to-end causal LLMs (Hosseini-Asl et al., 2020; Peng et al., 2020; Chao and Lane, 2019; Gao et al., 2019; Crook et al., 2021). We extend this line of work and propose a multimodal extension to account for visual inputs.

6 Conclusions

We present SIMMC-VR, a situated and interactive dialog dataset that features immersive VR streams as multimodal contexts, simulating realistic shopping scenarios along with user-assistant dialog interactions. The dataset consists of $4K$ user-egocentric videos paired with densely annotated dialog utterances. We build a novel meta-agenda generator for automatically synthesizing rich interactive dialogs grounded on active and diverse visual scenes, paraphrased manually for more natural speech. We propose four sub-tasks on SIMMC-VR which aims at inspiring future dialogue modeling endeavors on high-fidelity egocentric (user POV) environments; where the baseline performance highlights many challenges the dataset brings forth towards actualizing the real-world-ready VR/AR assistant. With rich annotations it provides, SIMMC-VR can as well expand beyond the proposed tasks to spur relevant future research, which includes (but not limited to): (1) augmented with speech-like spoken utterance interventions to enrich the naturalness of the dialogues, and (2) environments and room layouts beyond ones used under the scope of this paper.

Acknowledgments

Many thanks to Renato Sanchez and Becka Silvert for their help on coordinating the paraphrases collection; the linguists from the Appen tool company for the high-quality paraphrases; and the anonymous reviewers for their constructive feedback.

7 Limitations

We hereby discuss the current limitations of our work: **(1)** The SIMMC-VR dataset, similar to the SIMMC 2.0 version, focuses on shopping scenarios (clothing and furniture purchasing domains), one of the most common everyday activities that virtual reality could enable users to do from anywhere, anytime. We have not tested whether the models would generalize to domains outside of the shopping experiences, thus we cannot speak to the transferability of our results to environments with very different visual properties than what our virtual environments provide. **(2)** In this dataset, we hand-design several possible dialog acts that we assume are common for human buyers, as well as their associated scenarios. This may not exhaust all the possible interactions a shopper can do with the assistant. However, we emphasize that the coverage should be sufficient for common shopping experiences. Additionally, although most of our proposed subtasks should be modeling generic user-assistant multimodal dialogue interaction and thus could be transferred well to other domains, the (our) domain specific MM-DST may not generalize as much. Nevertheless, they should still be transferable to similar (shopping) environments. **(3)** The audio of the SIMMC-VR videos are generated by automatic TTS, which may fall short to represent the natural human speech. However, we do not foresee this causing problems for multimodal dialog modeling, which this work mostly focuses on.

8 Ethics and Broader Impacts

We hereby acknowledge that all of the co-authors of this work are aware of the provided *ACM Code of Ethics* and honor the code of conduct. This work is mainly about collecting a multimodal task-oriented dialog dataset with primary applications in actualizing a virtual assistant in the AR/VR world. The following gives the aspects of both our ethical considerations and the potential impact to the community.

Dataset. While most parts of our created dataset are *automatable*, our main human annotation efforts lie in the paraphrasing phase of our templated synthetically constructed dialog utterances. We ask in total 20 workers that possess linguistic expertise to paraphrase our templated utterances with carefully designed guidance and examples. We encourage the diversity where we do not pose any limits on the background of the paraphrasor as long as English proficiency and linguistic domain expertise is possessed.

The main annotation task is conducted via the Appen[6] provided interface, where we ensure that all the personal information of the workers involved (e.g., usernames, emails, urls, demographic information, etc.) is discarded in our dataset. The designed virtual environment scenes are not intended to have any bias towards any communities, where we aim at constructing generic domain and diverse scenes.

Overall, we ensure our pay per task is well above the annotator's local minimum wage (approximately \$30-35 USD / Hour). In this work, we primarily consider English speaking regions for setting up the initial benchmark, though our dataset can be easily extended to contain multilingual annotations for learning virtual AI assistants that are capale of different languages. This research has been reviewed by the **IRB board** and granted the status of an **IRB exempt**.

Techniques. We benchmark the constructed dataset with modern strong large-scale pretrained language and multimodal models with our own designs to adapt them to suit our formulated tasks. Due to the nature of our dataset (assistant AI that focuses on the needs of the human users and the surrounding environments), as well as the proposed main challenges this dataset feature (*i.e.* mainly focusing on resolving multimodal coreferences, tracking dialog states, and generating useful assistive responses to human users), we do not anticipate production of harmful outputs, especially towards vulnerable populations, after training models on our SIMMC-VR dataset/tasks.

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018a. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018b. MultiWOZ - a large-scale multi-domain wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

[6] https://appen.com/

Guan-Lin Chao and Ian Lane. 2019. Bert-dst: Scalable end-to-end dialogue state tracking with bidirectional encoder representations from transformer. In *Annual Conference of the International Speech Communication Association (INTERSPEECH)*.

Francisco Javier Chiyah-Garcia, Alessandro Suglia, José Lopes, Arash Eshghi, and Helen Hastie. 2022. Exploring multi-modal representations for ambiguity detection & coreference resolution in the simmc 2.0 challenge. *arXiv preprint arXiv:2202.12645*.

Paul A. Crook, Satwik Kottur, Seungwhan Moon, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2021. Situated interactive multimodal conversations (simmc) track at dstc9. *AAAI DSTC9 Workshop*.

Dima Damen, Adriano Fragomeni, Jonathan Munro, Toby Perrett, Daniel Whettam, Michael Wray, Antonino Furnari, Giovanni Maria Farinella, and Davide Moltisanti. 2021. Epic-kitchens-100-2021 challenges report.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017a. Visual dialog. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 326–335.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017b. Visual dialog. In *CVPR*.

Harm de Vries, Kurt Shuster, Dhruv Batra, Devi Parikh, Jason Weston, and Douwe Kiela. 2018. Talk the walk: Navigating new york city through grounded dialogue. *arXiv preprint arXiv:1807.03367*.

Harm de Vries, Florian Strub, Sarath Chandar, Olivier Pietquin, Hugo Larochelle, and Aaron Courville. 2017. Guesswhat?! visual object discovery through multi-modal dialogue. In *CVPR*.

Mihail Eric, Rahul Goel, Shachi Paul, Adarsh Kumar, Abhishek Sethi, Peter Ku, Anuj Kumar Goyal, Sanchit Agarwal, Shuyag Gao, and Dilek Hakkani-Tur. 2019. Multiwoz 2.1: Multi-domain dialogue state corrections and state tracking baselines. *arXiv preprint arXiv:1907.01669*.

Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu.

Shuyang Gao, Sanchit Agarwal Abhishek Seth and, Tagyoung Chun, and Dilek Hakkani-Ture. 2019. Dialog state tracking: A neural reading comprehension approach. In *Special Interest Group on Discourse and Dialogue (SIGDIAL)*.

Xiaofeng Gao, Qiaozi Gao, Ran Gong, Kaixiang Lin, Govind Thattai, and Gaurav S Sukhatme. 2022. Dialfred: Dialogue-enabled agents for embodied instruction following. *arXiv preprint arXiv:2202.13330*.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*.

Matthew Henderson, Blaise Thomson, and Jason D Williams. 2014. The second dialog state tracking challenge. In *Proceedings of the 15th annual meeting of the special interest group on discourse and dialogue (SIGDIAL)*, pages 263–272.

Chiori Hori, Anoop Cherian, Tim K. Marks, and Florian Metze. 2018. Audio visual scene-aware dialog track in dstc8. *DSTC Track Proposal*.

Ehsan Hosseini-Asl, Bryan McCann, Chien-Sheng Wu, Semih Yavuz, and Richard Socher. 2020. A simple language model for task-oriented dialogue. *arXiv preprint arXiv:2005.00796*.

Xin Huang, Chor Seng Tan, Yan Bin Ng, Wei Shi, Kheng Hui Yeo, Ridong Jiang, and Jung Jae Kim. 2021a. Joint generation and bi-encoder for situated interactive multimodal conversations. *AAAI 2021 DSTC9 Workshop*.

Yichen Huang, Yuchen Wang, and Yik-Cheung Tam. 2021b. Uniter-based situated coreference resolution with rich multimodal input. *arXiv preprint arXiv:2112.03521*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations (ICLR)*.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. Simmc 2.0: A task-oriented dialog dataset for immersive multimodal conversations. *arXiv preprint arXiv:2104.08667*.

Satwik Kottur, José MF Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2019. Clevr-dialog: A diagnostic dataset for multi-round reasoning in visual dialog. *arXiv preprint arXiv:1903.03166*.

Po-Nien Kung, Tse-Hsuan Yang, Chung-Cheng Chang, Hsin-Kai Hsu, Yu-Jia Liou, and Yun-Nung Chen. 2021. Multi-task learning for situated multi-domain end-to-end dialogue systems. *AAAI 2021 DSTC9 Workshop*.

Hung Le, Chinnadhurai Sankar, Seungwhan Moon, Ahmad Beirami, Alborz Geramifard, and Satwik Kottur. 2021. Dvd: A diagnostic dataset for multi-step reasoning in video grounded dialogue. *arXiv preprint arXiv:2101.00151*.

Haeju Lee, Oh Joon Kwon, Yunseon Choi, Minho Park, Ran Han, Yoonhyung Kim, Jinhyeon Kim, Youngjune Lee, Haebin Shin, Kangwook Lee, et al. 2022. Learning to embed multi-modal contexts for situated conversational agents. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 813–830.

Joosung Lee and Kijong Han. 2021. Multimodal interactions using pretrained unimodal models for simmc 2.0. *arXiv preprint arXiv:2112.05328*.

Zhaoyang Lv, Edward Miller, Jeff Meissner, Luis Pesqueira, Chris Sweeney, Jing Dong, Lingni Ma, Pratik Patel, Pierre Moulon, Kiran Somasundaram, Omkar Parkhi, Yuyang Zou, Nikhil Raina, Steve Saarinen, Yusuf M Mansour, Po-Kang Huang, Zijian Wang, Anton Troynikov, Raul Mur Artal, Daniel DeTone, Daniel Barnes, Elizabeth Argall, Andrey Lobanovskiy, David Jaeyun Kim, Philippe Bouttefroy, Julian Straub, Jakob Julian Engel, Prince Gupta, Mingfei Yan, Renzo De Nardi, and Richard Newcombe. 2022. Aria pilot dataset.

Seungwhan Moon, Satwik Kottur, Paul A Crook, Ankita De, Shivani Poddar, Theodore Levin, David Whitney, Daniel Difranco, Ahmad Beirami, Eunjoon Cho, Rajen Subba, and Alborz Geramifard. 2020. Situated and interactive multimodal conversations. *arXiv preprint arXiv:2006.01460*.

Nikola Mrkšić, Diarmuid Ó Séaghdha, Tsung-Hsien Wen, Blaise Thomson, and Steve Young. 2017. Neural belief tracker: Data-driven dialogue state tracking. In *Association for Computational Linguistics (ACL)*.

Aishwarya Padmakumar, Jesse Thomason, Ayush Shrivastava, Patrick Lange, Anjali Narayan-Chen, Spandana Gella, Robinson Piramuthu, Gokhan Tur, and Dilek Hakkani-Tur. 2022. Teach: Task-driven embodied agents that chat. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 36, pages 2017–2025.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Jinchao Li, Shahin Shayandeh, Lars Liden, and Jianfeng Gao. 2020. Soloist: Few-shot task-oriented dialog with a single pretrained auto-regressive model. *arXiv preprint arXiv:2005.05298*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2019. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Association for the Advancement of Artificial Intelligence (AAAI)*.

Abhinav Rastogi, Xiaoxue Zang, Srinivas Sunkara, Raghav Gupta, and Pranav Khaitan. 2020. Towards scalable multi-domain conversational agents: The schema-guided dialogue dataset. In *Association for the Advancement of Artificial Intelligence (AAAI)*, volume 34, pages 8689–8696.

Matteo Antonio Senese, Giuseppe Rizzo, Alberto Benincasa, and Barbara Caputo. 2021. A response retrieval approach for dialogue using a multi-attentive transformer. *AAAI 2021 DSTC9 Workshop*.

Pararth Shah, Dilek Hakkani-Tür, Gokhan Tür, Abhinav Rastogi, Ankur Bapna, Neha Nayak, and Larry Heck. 2018. Building a conversational agent overnight with dialogue self-play. *arXiv preprint arXiv:1801.04871*.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge.

Coqui.ai. 2022. Coqui tts. *https://github.com/coqui-ai/TTS*.

Unity. 2020. Unity. *https://unity.com/*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Wang Yinyin. 2011. Consumer behavior characteristics in fast fashion.

A Details of the Dataset

A.1 Digital Assets

Table 6 lists the asset categories used for constructing the SIMMC-VR dataset.

Inventories & Scenes. As the base environment inherits from SIMMC 2.0, there are around 290 and 110 digital assets for fashion and furniture items. 7 seed fashion scenes are manually created for fashion and 1 seed scene for furniture, with each seed scene rearranged 20 times (Section 2.1) to create (randomized) diverse shopping environments. We do so by randomly swapping an asset from either the same or semantically similar asset category (Table 6) to retain spatial appropriateness (*e.g.* avoiding collisions or over-sized to a container) of the replacement object.

The user traversal video is then planned and recorded in these environments. The number of visible distinct objects in *key video frames* of static dialog phase is 24.6 (Table 4), which implies quite rich multimodal contents are presented in the recorded videos.

A.2 Dialog Goals & Acts

Table 7 lists all the sub-goals during the high-level agenda for planning the overall dialog flows, with their detailed explanations.

Table 8 lists all the used dialog acts and activities. Recall that a full dialog act is composed by the **act** term and the **activity** term, *e.g.* `REQUEST:COMPARE`. Most of the activities are self-explainable. Navigational activities are separated to `DIRECTION_STRAIGHT` and `DIRECTION_TURN`, etc., to make the utterance template sampling more straightforward (as we most likely will use different ways to speak about moving straight as opposed to making turns.) The `PRE_CONDITION` and `POST_CONDITION` activities are for querying referential objects during key action points, with the former querying the objects before the action (*e.g.* *"Turn right **when you see** a pair of <u>blue jeans</u>."*), and the latter during the ending of the action (*e.g.* *"Turn right **and then you should** see a pair of <u>blue jeans</u>."*). `SPATIAL_RELATIONS` are for referring objects within a local object-scene-graph when sampling corresponding utterance templates, while `REMEDIAL_*` activities indicate that the current assistant speech is trying to get the user back on the correct track (towards an item(s) of interest).

A.3 Other Details

Failure Modes for Dialog Simulation. In reality, the user may not always perfectly follow an instruction. We model such behaviors in SIMMC-VR by (with 30% probability) deliberately failing an instructed action during the path planning stage (*e.g.* making a wrong turn, moving further from a desired spot). We achieve this by randomly selecting one (or a few) key action points along a proper traversal path and record the opposite actions in the user video. The correct action path will then be used to derive the original instructed utterance (where the user *fails* to follow) and we additionally perform *remedial path* planning to guide the user back to the right track.

Bootstrapping from Sparse Scenes. While most of the dialogs from SIMMC 2.0 have a single static image, the dataset contains a small proportion of dialogs with two randomly sampled (sparse) scenes as multimodal contexts. We propose to recover and re-purpose these sparse scenes to add onto our dataset by connecting the two scenes with a newly collected *active phase* navigating the user from one scene to another, with augmented conversations along the traversal. This step essentially adds diversity and depth to conversations in our dataset. For the *static phase* (the original two snapshots of SIMMC 2.0), we animate the scenes with user's eye-gaze movements combined with a small amount of local wandering movements to appear more natural. We use $\sim 1.4K$ static phase conversations from SIMMC 2.0, ensuring that at least one turn of user$\leftrightarrow$assistant conversation exists in the second scene snapshot.

TTS Utterances. To make the rendered user shopping videos more realistic (and comprehensive), we also perform an automatic text-to-speech synthesis (TTS) on each user and assistant utterance using an open-sourced tool, *Coqui TTS* (Coqui.ai, 2022). The TTS helps compute the natural duration of each utterance when spoken, which is then used to calculate the number of video frames (under a fixed frame-rate) an utterance would span. The random eye-gaze movements mentioned in Section 2.1 extend the sparsely rendered scene snapshots to the continuous video frames, synchronized with the aforementioned speech.

A.4 Data Examples

Figure 7 shows a sample sub-sampled video frames for both the fashion and furniture domains, for

Fashion	hat, tshirt, jacket, hoodie, sweater, shirt, suit, vest, coat, trousers, jeans, joggers, skirt, blouse, tank top, dress, shoes
Furniture	area rug, bed, chair, couch chair, dining table, coffee table, end table, lamp, shelves, sofa

Table 6: **Digital assets categories used in SIMMC-VR** for both fashion and furniture domains.

Goals	**Explanation**
`UNKNOWN`	Default.
`BROWSE`	Browse the shop, asking for recommendation etc.
`REFINE_SEARCH`	Refine the previous search for objects in the current scene with additional criteria.
`GET_SIMILAR`	Get similar item to a specific one, in the current scene.
`GET_INFO`	Get information about an item.
`COMPARE`	Compare two or more items.
`ADD_TO_CART`	Add item(s) to cart.
`ALTERNATE_SEARCH`	Search in the current scene for objects alternative to a specific one.
`COMPLEMENTARY_SEARCH`	Search in the current scene for objects complementary to a specific one.
`GLOBAL_GET_SIMILAR`	Get similar items to a specific one within the entire environment.
`GLOBAL_REFINE_SEARCH`	Refine the previous search but objects can be anywhere in the environment.
`GLOBAL_ALTERNATE_SEARCH`	Alternative search but objects can be anywhere in the environment.
`GLOBAL_COMPLEMENTARY_SEARCH`	Complementary search but objects can be anywhere in the environment.
`ACTION`	Indicates physical actions (navigation, viewpoint movements etc.)

Table 7: **Dialog Goals for Agenda**

Dialog Acts (5)	`INFORM`, `REQUEST`, `CONFIRM`, `ASK`, `CONDITION`
Activities (17)	`GET`, `REFINE`, `COMPLEMENTARY`, `ALTERNATE`, `PREFER`, `DISPREFER`, `COMPARE`, `ADD_TO_CART`, `DISAMBIGUATE` `DIRECTION_STRAIGHT`, `DIRECTION_TURN`, `DIRECTION_TURN_AROUND` `PRE_CONDITION`, `POST_CONDITION`, `SPATIAL_RELATIONS` `REMEDIAL_TURN`, `REMEDIAL_STRAIGHT`

Table 8: **Dialog Acts & Activities for Agenda:** A full dialog act comprises of an **act** and an **activitiy**, *e.g.* `INFORM:GET`.

Models	**Batch Size**	**Initial LR**	**# Training Epochs**	**Gradient Accumulation Steps**	**# Params**
MM-DST	4	5×10^{-5}	10	1	117M
MM-DST (no-gt.)	4	5×10^{-5}	10	1	117M
Adapt.-VIOLET	4	1×10^{-5}	10	1	214M

(a) Hyperparameters

Type	**Batch Size**	**Initial LR**	**# Training Epochs**	**Gradient Accumulation Steps**
Bound (lower–upper)	2–8	5×10^{-5}–5×10^{-6}	6–10	1–1
Number of Trials	2–4	2–3	2–4	1–1

(b) Search Bounds

Table 9: **(a) Hyperparameters in this work:** *Initial LR* denotes the initial learning rate. All the models are trained with Adam optimizers (Kingma and Ba, 2015). We include the number of learnable parameters of each model in the column: *# params*. **(b) Search bounds** for the hyperparameters of all the models.

qualitative purpose.

Figure 8 shows full-scale example of one of the data instance in SIMMC-VR – with some navigational utterances from the assistant shortened for brevity.

B Details of Modeling

B.1 General Modeling

The respective author-released pretrained weights for both models (GPT-2 and VIOLET) are used for model initializations.

As described in Section 4, we use VIOLET (Fu et al.) due to the model's architectural simplicity and convenience to adapt to our task (as well as its remarkable performances on various video-language tasks). The multi-framed vision transformer stream of video encoder is suitable for the SIMMC-VR task, where we further engineer it to be able to take on dense object descriptors and dialog structures. For MM-DST, MM-Coref, and Failure Mode Prediction tasks, we mainly adopt the original VIOLET's BERT module as its language encoder, while for response generation, we replace the BERT with GPT-2 and train the model from scratch directly on our dataset (with the visual streams initialized from pretrained weights).

B.2 Hyper-Parameters

Table 9a and Table 9b report the hyper-parameters used in this work for model training and their search bounds, respectively. We simply perform a manual search trials.

B.3 Implementation Details & Hardware

The implementations of the transformer-based models are extended from the HuggingFace[7] code base (Wolf et al., 2020) and other cited authors' released code-bases. Our entire code-base is implemented in PyTorch.[8] All the models in this work are trained on a single Nvidia A100 GPU[9] on a Ubuntu 20.04.2 operating system.

[7]https://github.com/huggingface/transformers
[8]https://pytorch.org/
[9]https://www.nvidia.com/en-us/data-center/a100/

Figure 7: **Sample videos** for both the fashion (upper half) and furniture (lower half) domains. The frames are sub-sampled (with fps=10.0) to show qualitative and representative visuals of our dataset.

UTTERANCES	ANNOTATIONS
U: Could you help me find a jacket?	REQUEST:GET, slots: {'type': 'jacket'}, objects: []
A: How about this grey jacket?	INFORM:GET, slots: {'type': 'jacket'}, objects: [76]
U: What about a grey jacket made by North Lodge?	INFORM:REFINE, slots: {'type': 'jacket', 'color': 'grey', 'brand': 'North Lodge'}, objects: []
A: Sorry, I couldn't find anything, do you want to see something else?	INFORM:GET, slots: {'type': 'jacket', 'color': 'grey', 'brand': 'North Lodge'}, objects: []
U: What about a skirt to go with the grey jacket in front of me?	INFORM:COMPLEMENTARY, slots: {'type': 'skirt'}, objects: [76]
A: We have a nice grey and white skirt in stock.	INFORM:GET, slots: {'type': 'skirt'}, objects: [46]
U: Can you show me where I could find something like that?	REQUEST:LOCATE, slots: {}, objects: []

A: If you turn to your right, you'll see a pair of gray pants ahead and partially obscured by the checkout counter.	INFORM:DIRECTION_TURN, slots: {'turn': 'right', 'displacement': None}, objects: [44]
A: Try to spot the black and white patterned top that's hanging to the upper left of a pair of gray pants. Once you've spotted it, walk over towards it.	ASK:VISIBLE, slots: {'turn': None, 'displacement': 'forward'}, objects: [18]
•••	•••
A: Also further down that left hand wall will be a hanging, black dress. Walk towards the dress. On your right will be a purple, folded shirt on the lowest tier of a display table.	INFORM:DIRECTION_STRAIGHT, slots: {'turn': None, 'displacement': 'forward'}, objects: [37]
A: You're almost there now! Finally, turn to your right and find the gray and white skirt that's hanging on the opposite wall behind the clothing rack in front of you. It should be to the left of a black dress that's hanging on an angled rack to your right. I hope you like that skirt!	INFORM:DIRECTION_TURN, slots: {'turn': 'right', 'displacement': None}, objects: [46]

U: What brand is that grey jacket?	ASK:GET, slots: {}, objects: [76]
A: It is North Lodge.	INFORM:GET, slots: {'Object ID: 76': {'brand': 'North Lodge'}}, objects: [76]
U: Okay, I'll take one of the grey and white skirt on the other side of the rack.	REQUEST:ADD_TO_CART, slots: {}, objects: [46]
A: I'm adding it to your cart now.	CONFIRM:ADD_TO_CART, slots: {}, objects: [46]
U: I also want to find something by Uptown Studio.	INFORM:REFINE, slots: {'brand': 'Uptown Studio'}, objects: [46]
A: How about the black and grey skirt in store?	INFORM:GET, slots: {'brand': 'Uptown Studio'}, objects: [62]
U: Could you help me find what I'm looking for?	REQUEST:LOCATE, slots: {}, objects: []
•••	•••

U: How does this black and grey skirt compare to the grey and white one?	REQUEST:COMPARE, slots: {}, objects: [62, 46]
A: They are both short skirts.	INFORM:COMPARE, slots: {'Object ID: 62': {'assetType': 'skirt', 'sleeveLength': 'short'}, 'Object ID: 46': {'assetType': 'skirt', 'sleeveLength': 'short'}}, objects: [62, 46]
U: I want to find a plain skirt.	INFORM:REFINE, slots: {'type': 'jacket', 'pattern': 'plain'}, objects: []
A: I have this plain black skirt, and a grey and white option as well.	INFORM:GET, slots: {'type': 'jacket', 'pattern': 'plain'}, objects: [65, 70]
U: I'll take the grey and white option.	REQUEST:ADD_TO_CART, slots: {}, objects: [46]
A: Okay, I will add it for you.	CONFIRM:ADD_TO_CART, slots: {}, objects: [46]

Figure 8: **Dataset example.** Dialog labels include intent, slots, and multimodal co-references. The key static phase frames are shown with the corresponding conversations, along with a few sub-sampled frames during the traversals. **U** denotes user and **A** denotes the assistant. For simplicity, we omit some over-lengthy navigational instruction from the assistant.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Section 7.

☑ A2. Did you discuss any potential risks of your work?
Section 8.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract and Section 1.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2 and 3.

☑ B1. Did you cite the creators of artifacts you used?
Section 4.

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
Section 4.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2 and 3 and 4.

☑ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Section 8.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
Section 2 and Appendices.

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
Section 2.

C ☑ Did you run computational experiments?

Section 4.

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix Section B.1.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Section 4 and Appendix Section B.

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
Section 4 and Appendix Section B.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Section 4 and Appendix Section B.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Section 2 and Section 8.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Not applicable. Left blank.

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Section 8.

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Section 8.

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Section 8.

☑ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Section 8.