# Tell Your Story: Task-Oriented Dialogs for Interactive Content Creation

**Anonymous ACL submission**

## Abstract

People capture photos and videos to relive and share memories of personal significance. Recently, media montages (stories) have become a popular mode of sharing these memories due to their intuitive and powerful storytelling capabilities. However, creating such montages usually involves a lot of manual searches, clicks, and selections that are time-consuming and cumbersome, adversely affecting user experiences.

To alleviate this, we propose *task-oriented dialogs for montage creation* as a novel interactive tool to seamlessly search, compile, and edit montages from a media collection. To the best of our knowledge, our work is the first to leverage multi-turn conversations for such a challenging application, extending the previous literature studying simple media retrieval tasks. We collect a new dataset C3 (**C**onversational **C**ontent **C**reation), comprising $10k$ dialogs conditioned on media montages simulated from a large media collection.

We take a simulate-and-paraphrase approach to collect these dialogs to be both cost and time efficient, while drawing from natural language distribution. Our analysis and benchmarking of state-of-the-art language models showcase the multimodal challenges present in the dataset. Lastly, we present a real-world mobile demo application that shows the feasibility of the proposed work in real-world applications. Our code & data will be made publicly available.

## 1 Introduction

With the advent of smart cameras, smart glasses, and other media devices, the barrier to capturing photos and videos has drastically been reduced. While this trend is desirable to relive and share memories, the sheer volume of such captured media makes it intractable to search and share relevant memories. As a result, media montages (stories) have emerged as an intuitive yet expressive way to creatively compile various memories and share with friends and family. In order to create such a montage, users have to search through their personal collections, make selections, and edit them manually, which are cumbersome and time-consuming tasks, resulting in a bottleneck.

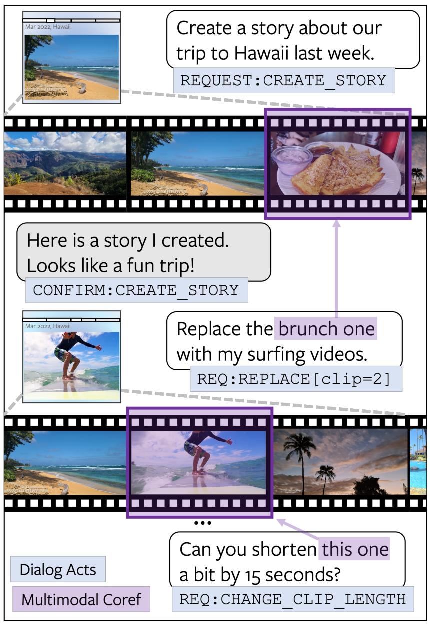



Figure 1: Illustration of C3: **C**onversational **C**ontent **Creation**. Each dialog turn is fully annotated with dialog acts and multimodal coreference labels, accompanied with its corresponding story montage snapshot.

In this work, we propose a novel conversational tool to interactively create and edit montages from a personal media collection. While prior works study the use of dialog in retrieving media or items in a shopping catalog, we extend it to capture richer interactions related to montage manipulations. To the best of our knowledge, our work is the first to consider task-oriented dialogs (TOD) for this challenging application of interactive content creation.

Towards this goal, we collect C3, a TOD dialog dataset aimed at providing an intuitive conversational interface in which users can search through their media, create a video story with highlights, and edit clips hands-free, using natural language.

Fig. 1 illustrates an example dialog. Due to our simulate-and-paraphrase pipeline, our dataset comprises rich annotations both at turn- and dialog-level. These are helpful to: (a) tease out and study multimodal challenges (*e.g.*, multimodal coreferences) that are present in C3, and (b) benchmark meaningful progress towards a robust TOD agent for this application. We perform preliminary empirical experimentation and train baselines to highlight the multimodal challenges in our C3 dataset. Lastly, we build a mobile demo (Fig. 5, App. A) to showcase the real-world applicability of our work.

## 2 Related Work

**Task-oriented Dialogs** (TOD), where the goal is to parse user queries and execute a pre-defined set of actions (*e.g.* booking hotels), have been extensively studied. We formulate similar tasks as found in the conventional TOD datasets (Rastogi et al., 2019; Budzianowski et al., 2018; Eric et al., 2019) such as Dialog State Tracking (DST), to build on the literature. Our work extends it to a novel multimodal application of video content creation and editing.

Recently, the methods that leverage large pre-trained LMs by casting DST as a causal inference problem (Peng et al., 2020; Hosseini-Asl et al., 2020; Gao et al., 2019) have shown successful. We develop a baseline following this trend, but extend it a unique multimodal setting by including multimodal context as part of the grounding prompt.

**Conversational Media Applications**: Recent work have addressed the dialog task for retrieving images (*e.g.* from a personal collection or as part of shopping scenarios) (Guo et al., 2018a,b; Tellex and Roy, 2009; Vo et al., 2019; Tan et al., 2019), given multi-turn target queries. Similarly, Bursztyn et al. (2021) considers an application to retrieve multiple images to create a montage. While C3 does include search operations, our work extends this line of work by allowing for richer interactions and more complex post-edits on the retrieved videos, enhancing overall user experiences.

As per similar applications, Lin et al. (2020) proposes tasks for editing a single image (*e.g.* brightness) via text commands, while Zhou et al. (2022) study interactive image generation from text, using CLIP text-image embeddings (Radford et al., 2021) and a generative model (Karras et al., 2019). Unlike the previous work that handle editing operations within a single image, our work addresses conversational editing of multiple videos into storytelling montages, a popular form of media sharing.

## 3 The C3 Dataset

### 3.1 Multimodal Dialog Self-Play

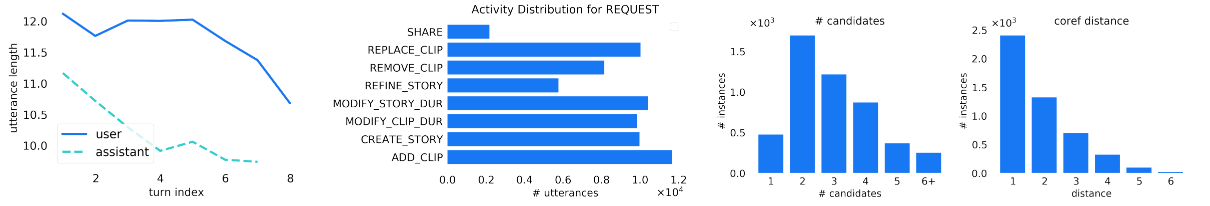
We adopt a two-phase pipeline (Simulate and Paraphrase (Shah et al., 2018; Kottur et al., 2021)), extending it to a unique multimodal setting where multiple images as part of the user interface (UI) are given as grounding visual contexts. The proposed approach reduces the data collection and annotation overheads (time and cost) for building a dialog dataset (*vs.* collecting human↔human dialogs and collecting Dialog/NLU annotations on top), as it requires little to no domain knowledge.

**Phase 1. Multimodal Dialog Simulator.** We first generate synthetic dialog flows using a dialog simulator that conditions on an evolving "story" and its corresponding set of clips, produced by a story generator. The story generator outputs a diverse set of clips (as schematic representation) according to user requests, which serves as grounding multimodal context for the conversations. This is done by extracting a plausible set of meta information (time, locations and activities, etc.) from an existing memory graph, simulated and generated using the object and activity annotations from the ImageCOCO dataset (Lin et al., 2014).

The dialog simulator then takes this story representation including the meta information (activities, locations, attributes, etc.) and the UI state (*e.g.* sequential ordering of media, viewer status) updated at each turn, to create a realistic dialog flow between a user and an assistant, using a probabilistic agenda-based approach. The simulated dialog flows comprise NLU intents (*e.g.* `REQUEST:ADD_CLIPS`), slots (*e.g.* activities, objects), and clip references. Specifically, we capture various video editing queries that are identified as a prioritized list of common actions required for media editing and sharing (*e.g.* `CREATE`, `REMOVE`, `REPLACE`, `REORDER`, `REFINE`, `MODIFY_DURATION`).

**Phase 2. Manual Paraphrase.** Once the dialog flows are simulated, we paraphrase each templated user turn via manual annotations. This step allows us to collect utterances from the natural language distribution, making the dataset robust to the user-query variability in real-world applications.

We build an annotation tool that displays NLU labels and templated utterances, along with the schematic representation of stories with media clips, updated at each turn. Annotators are then instructed to paraphrase each turn without losing key multimodal information such as relative clip




(a) (b) (c)

Figure 2: Distribution of (a) utterance lengths with dialog turns, (b) activity distribution for `REQUEST` user act (dominant), (c) number of clip candidates per turn (L) and coreference distance (R) between clip mentions.

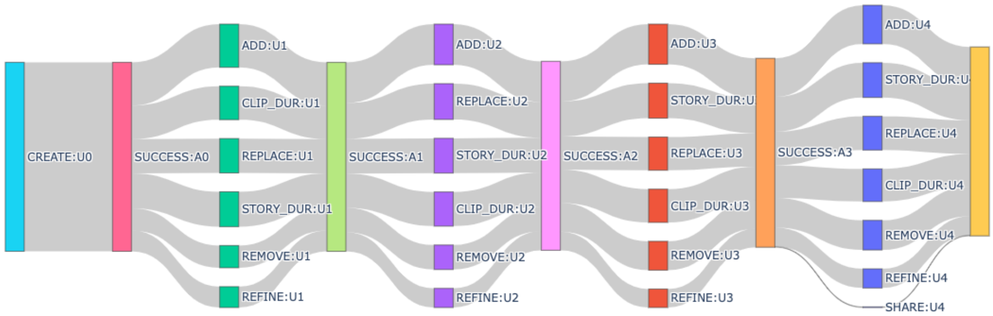



Figure 3: Transition of dialogs acts in C3 for the first 4 turns, for dialog flows generated by our multimodal dialog simulator. Each block is labelled `ACTIVITY:[A|U][turn]` to denote activity, user or assistant turn, and turn number, respectively. `ACT` for user (`REQUEST`) and assistant (`INFORM`) are not shown for brevity. See text for more details.

| | |
|---|---|
| Total # dialogs | $10k$ |
| Total # utterances | $136k$ |
| Total # stories | $10k$ |
| Avg # words (user turns) | $11.8 \pm 4.4$ |
| Avg # words (assistant turns)† | $10.3 \pm 4.1$ |
| Avg # utterances / dialog | 13.5 |
| Avg # clips mentioned / dialog | 3.6 |
| Avg # clips per story | $4.3 \pm 2.5$ |

Table 1: **C3 Dataset Statistics.** †assistant turns are collected for a $1k$ dialog subset ($12k$ utterances).

placements & meta data, objects and attributes.

While assistant turns tend to be linguistically less diverse (*e.g.*, informing successful executions: '*Done*', '*Edited*') and thus are less of our focus from an application standpoint, we also collect assistant responses for a $1k$ dialog subset. The collected utterances allow for the study of contextualized assistant response generation, to accompany the modified stories reflected in the UI.

## 3.2 Dataset Analysis

Our C3 dataset has a total of $10k$ dialogs with $136k$ utterances. Dataset statistics are given in Tab. 1. A dataset example is provided in Fig. 6 (Appendix C).

**Analyzing Dialogs.** The user and assistant turns in dialogs from C3 are about 11.8 and 10.3 words long respectively, with their distributions shown in Fig. 2a. User utterances tend to be longer on an average as they are instructive and contain finer details to manipulate the story.

**Analyzing Dialog Annotations.** Dialogs in C3 are accompanied with full turn- and dialog-level annotations, thanks to the simulate-and-paraphrase approach. We follow the conventional hierarchical ontology (Kottur et al., 2021) of dialog `ACT` and `ACTIVITY` to annotate both user and assistant intents. In our setup, users can *request* selections or edits to create a montage, while the assistant is expected to execute them and *inform* its results. Thus, the user and assistant dialog acts naturally resort to `REQUEST` and `INFORM` in our ontology. Fig. 2b shows the distribution of 8 user activities.

Each turn is grounded on an (evolving) story, which contains an average of 4.3 clips. This leads to interesting multimodal coreferences as there are about 2.9 clip candidates to pick from for every clip mention in the dialog. Further, the average coreference distance between the mentions is 3.7, going beyond the trivial case of 1, *i.e.*, clip mentioned in the previous turn. Fig. 2c highlights the distribution of clip candidates and distance between mentions.

**Analyzing Dialog Flows.** We visualize the dialog flows (first 4 dialog turns) in Fig. 3. Each block is an intent at a particular `[turn]` labelled as `ACTIVITY:[A|U][turn]`, where `[A|U]` indicates either an Assistant or User turn. The gray bands denote the transitions and their width is proportional to the frequency of the transition. The almost uniform branch-off indicates a desirable presence of diversity and thus a lack of intent bias in the dialog.

## 4 Task Formulation

We leverage C3 to study dialog systems that help users create and edit montages through a multi-turn dialog. More concretely, we propose 3 main tasks and respective evaluation metrics in this regard:

**Task 1: API Slot Prediction.** We assume a 1-to-1 mapping between user intent and the relevant API to execute a user request. API Slot Prediction thus involves predicting slots (e.g., *participants*, *time*) that are passed as arguments to the corresponding API, given dialog history, multimodal context of stories, and current user utterance (metric: F1). For example, *'U: Create a story of all skiing trips in 2018'* maps to `[activity=skiing,time=2018]` as the appropriate API slots and values. We do not propose a separate API-type prediction task (*e.g.* `api_type=CREATE_STORY`) as the baseline models perform with near perfect accuracy (97%).

**Task 2. Multimodal Coreference Resolution.** It is imperative for conversational systems to be able to resolve multimodal coreferences without fails as a wrongly targeted edit would require additional interactions to rectify, greatly reducing user experiences. For instance, to process *'Remove the sunset clip and replace it with something similar to the second one.'*, the system needs to resolve both underlined references to the corresponding clip objects to perform the desired manipulations. To test this capability in isolation, we propose Task 2, where the goal is to resolve any clip references in the current user utterance to the corresponding clip objects (metric: F1), taking into account dialog history and story representations.

**Task 3. Multimodal Dialog State Tracking (MM-DST).** Lastly, we evaluate the system on its joint ability to: (a) predict API calls along with its slot parameters, and (b) resolve multimodal references (if any) in the given utterance, taking into account dialog state carryovers (measured with accuracy).

## 5 Modeling & Empirical Analysis

We perform a preliminary empirical evaluation and train baselines for the tasks proposed in Sec. 4. We leave detailed modeling as part of future work.

**Dataset Splits.** We split the $10k$ dialogs into `train` (60%), `val` (20%), and `test` (20%). All models are trained on `train` with `val` used to pick the hyper-parameters, and results are reported on `test`.

**Baselines.** Following the recent success of finetuning pretrained LMs on TODs (Hosseini-Asl et al., 2020; Peng et al., 2020), we adopt GPT-2 (Radford et al., 2019) and extend these work by adding two different ways of representing multimodal contexts (story): (a) visual embeddings (embed), where we extract object-centric visual features for constituent clips (Ren et al., 2015) projected into the hidden size of GPT-2 via a linear layer, and (b) stringified text (tokens), where the story information is represented as stringified tokens. The models are trained to predict API calls, slot values, and clip mentions given a sequential input of its dialog context and multimodal context as above, through a conditional LM loss. More details are in Appendix B.

| Model | 1. API Slot | 2. Coref | 3. DST |
|---|---|---|---|
| | Slot F1↑ | Coref F1↑ | Acc.↑ |
| GPT-2 (tokens) | 88.3±0.3 | 70.4±0.5 | 72.8 |
| GPT-2 (embed) | **90.1**±0.1 | **81.5**±0.6 | **79.6** |

Table 2: Baseline performances for GPT-2 models w/ multimodal image features (embed) and stringified text (tokens). **(1) API Call Slot Prediction (API Slot)**, via slot F1, **(2) Multimodal Coreference Resolution (Coref)**, via coref prediction F1, **(3) Dialog State Tracking (DST)**, via Joint Accuracy. ↑: higher is better.

**Results.** From Tab. 2, it can be seen that the models achieve reasonably reliable performances for API prediction, while the coreference resolution task (exactly pinpointing which set of clips a user mentions) still remains a challenge. This is due to the various types of coreferences that exist in C3 that make resolutions uniquely challenging (*e.g.* adjectival: "*the sunset clip*", ordinal: "*the second to the last one*., device context: "*the one I'm currently viewing*", long-range carryover: "*the one I added earlier*"). This result suggests future modeling directions that could leverage the unique multimodal context more explicitly. It can also be shown that the model that uses raw visual embeddings outperforms the model that uses stringfied textual tokens, by better incorporating rich context present in visual information.

**Conclusions**: We propose a novel task of building a TOD system for interactively creating storytelling media contents from a personal media collection. We build a new multimodal dataset ($10k$ dialogs & $136k$ turns) with rich dialog annotations and story representations. Our analysis with the SOTA LM-based multimodal dialog model highlights the key challenges such as multimodal coreference resolution and MM-DST. Lastly, our mobile application demonstrates the feasibility of our C3 dataset and model on popular real-world applications in short and long-form content creation and sharing.

## 6 Limitations

The generalizability and the use cases of the C3 dataset are bounded by the synthetic nature of the multimodal dialog simulator used for this study. However, we note that even with the simulated dialog flows, C3 captures several interesting challenges that are not addressed in the previous literature such as the use of media montage representations and device status as the grounding context for multimodal conversations, which opens the door to new research directions. We will open-source the multimodal dialog simulator used in the study for anyone to further develop any video-editing operations that are not included in C3, if necessary.

## References


Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. 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)*.

Victor S Bursztyn, Jennifer Healey, and Vishwa Vinay. 2021. Gaud\'i: Conversational interactions with deep representations to generate image collections. *arXiv preprint arXiv:2112.04404*.

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*.

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)*.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018a. Dialog-based interactive image retrieval. In *NeurIPS*.

Xiaoxiao Guo, Hui Wu, Yu Cheng, Steven Rennie, Gerald Tesauro, and Rogerio Feris. 2018b. Dialog-based interactive image retrieval. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

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*.

Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410.

Satwik Kottur, Seungwhan Moon, Alborz Geramifard, and Babak Damavandi. 2021. 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*, pages 4903–4912, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Tzu-Hsiang Lin, Trung Bui, Doo Soon Kim, and Jean Oh. 2020. A multimodal dialogue system for conversational image editing. *arXiv preprint arXiv:2002.06484*.

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

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

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)*.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

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*.

Fuwen Tan, Paola Cascante-Bonilla, Xiaoxiao Guo, Hui Wu, Song Feng, and Vicente Ordonez. 2019. Drill-down: Interactive retrieval of complex scenes using natural language queries. In *Neural Information Processing Systems (NeurIPS)*.

Stefanie Tellex and Deb Roy. 2009. Towards surveillance video search by natural language query. In *Proceedings of the ACM International Conference on Image and Video Retrieval*, CIVR '09, New York, NY, USA. Association for Computing Machinery.

Nam Vo, Lu Jiang, Chen Sun, Kevin Murphy, Li-Jia Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval - an empirical odyssey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, Chris Tensmeyer, Tong Yu, Changyou Chen, Jinhui Xu, and Tong Sun. 2022. Interactive image generation with natural-language feedback.

# A Appendix: Demo Interface

To demonstrate the feasibility of the real-world applications of the proposed dataset and models, we built a mobile demo application that runs the model trained with the C3 dataset. As can be seen in Fig. 5, the demo successfully handle unscripted user requests (not drawn from the training data) on a personal video collection as a retrieval target set, showing the promising use cases of our work.

Note that a computer vision model was used to pre-process and extract key visual concepts for each video in the collection. Each video was indexed with the extracted concepts and stored in a database in advance for faster inference.

At inference time, the mobile front-end runs an ASR model to get a transcript of a user's request, which is then routed to the dialog model. Once the dialog model predicts the API call and parameters, we retrieve the associated video files and execute the requested create or edit operations on the story.

# B Appendix: Multimodal DST with a Causal Language Model

Following the recent success of finetuning pretrained LMs on task-oriented dialog task modeling (Hosseini-Asl et al., 2020; Peng et al., 2020), we cast the MM-DST as a causal language inference task. Specifically, we use the concatenated `{<dialog history>, <multimodal context>}` as the prompting context for the LM (where multimodal context is represented either as visual embeddings or textual tokens), and use the task labels `{INTENT [slot = value, ...] <clip: IDs, ...>}` as the target for causal LM inference.

We use the 12-layer GPT-2 ($117M$) model (Radford et al., 2019) and finetune it on the C3 dataset, using early stopping based on token perplexity (<3 GPU hrs). Fig. 4 illustrates the proposed architecture for the tasks in Sec. 4.

# C Appendix: Dataset Example

Fig. 6 illustrates an example dialog from the C3 dataset, along with the schematic representation of the stories (with a sequence of clips and their meta data) associated with each turn (U: User, A: Assistant). API Annotations are formatted as follows: `INTENT [slot = value, ...] <clip: IDs, ...>`.

It can be seen that the dataset includes many challenges such as multimodal coreferences and

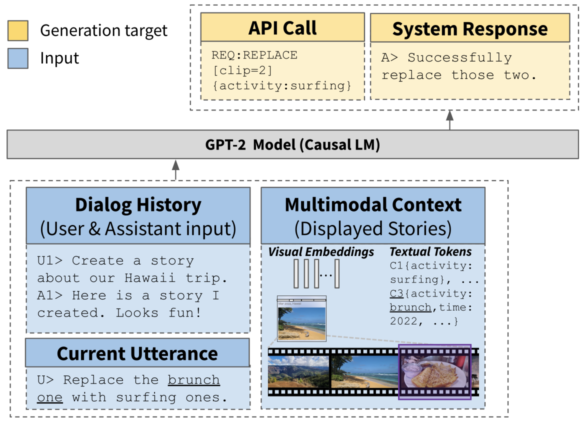



Figure 4: Baseline GPT-2 models for C3. Given the dialog history, multimodal context, and current user utterance, the model predicts the API call at the current turn. As shown, GPT2 (tokens) uses attribute strings to represent memories, while GPT2 (embed) use visual features.

dialog context carryovers. We report the detailed breakdown of the benchmark performances (*e.g.* API prediction, Multimodal Coreference Resolution F1) in Sec. 5

More details on the dataset including the key statistics are provided in Sec. 3.2.

# D Appendix: Ethical Considerations

The data paraphrase task was contracted through an external vendor that specializes in NLP annotations, where annotators are employed as full-time. Annotators were provided with clear instructions including a detailed escalation path ("Report Dialog") for an (unlikely) case where the templated utterance may include sensitive topics.

Please note that the figures used in this paper are from authors' personal media collections, and do not include identifiable faces or sensitive topics.

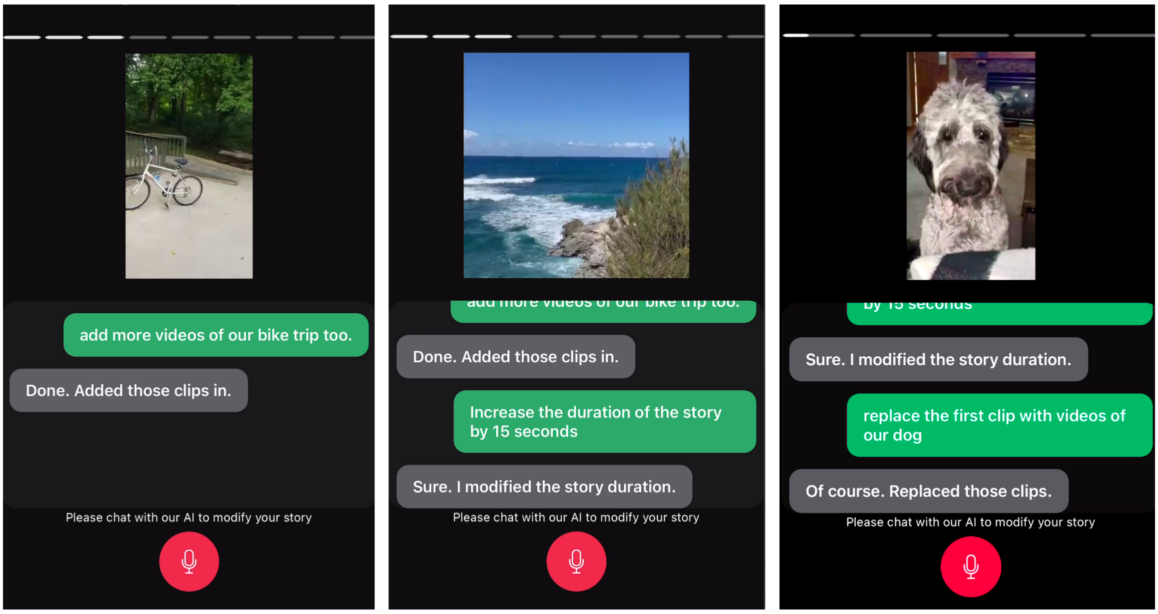



Figure 5: Screenshots of our **mobile demo application**. The dialog model is trained with the C3 dataset, and served on a Python server. A personal media collection was used as a retrieval target set for demonstration purposes.

| | Utterances | API Annotations / Dialog Acts | Story Representation |
|---|---|---|---|
| U | Are you able to build a story comprising Frances and Jacob? | `CREATE_STORY [participant = Frances, Jacob]` | |
| A | Of course! I have a story going that consists of Frances and Jacob. | `ADDED <clip: 3, 35, 50, 17>` | ID: 3 With: Frances, Jacob; ID: 35 With: Frances, Jacob; ID: 50 With: Frances, Jacob; ID: 17 With: Frances, Jacob VIEWING |
| U | What if you search for content catching us at Alki Beach sometime last year? | `REFINE_SEARCH [location = Alki Beach, time = last year]` | |
| A | I was able to grab 2020 content specifically from that same beach. | `ADDED <clip: 34, 8, 32>` | ID: 34 With: Frances, Jacob Time: 2020 Location: Alki Beach; ID: 8 With: Frances, Jacob Time: 2020 VIEWING; ID: 32 With: Frances, Jacob Time: 2020 Location: Alki Beach |
| U | Would you add a minute to the story's length? | `MODIFY_STORY_DURATION [delta_time = 1 min, add = True]` | |
| A | Yes, that's completed. | `MODIFIED:STORY` | |
| U | And scan for media with us skiing and add them. | `ADD_CLIPS [participant = Frances, Jacob & activity = skiing]` | |
| A | I threw in all the skiing clips I came across. | `ADDED <clip: 25, 17, 31>` | ID: 34 With: Frances, Jacob Time: 2020 Location: Alki Beach; ...; ID: 25 With: Frances, Jacob Activity: Skiing; ID: 17 With: Frances, Jacob Activity: Skiing VIEWING; ID: 31 With: Frances, Jacob Activity: Skiing |
| U | Now can you make the first ski trip no longer than 15 seconds? | `CHANGE_CLIP_DURATION <clip: 25> [max_time = 15 sec]` | |
| A | A: I've changed that length of that video for you. | `MODIFIED <clip: 25>` | ID: 34 With: Frances, Jacob Time: 2020 Location: Alki Beach; ...; ID: 25 With: Frances, Jacob Activity: Skiing Duration: Trimmed; ID: 17 With: Frances, Jacob Activity: Skiing VIEWING; ID: 31 With: Frances, Jacob Activity: Skiing |
| U | Finally, replace the one I'm currently viewing with one from the Boston Common. | `REPLACE [location = Boston Common] <clip: 17>` | |
| A | This new one is from the Boston Common, while the one you were looking at is gone. | `REMOVED <clip: 17>` `ADDED <clip: 8>` | ID: 34 With: Frances, Jacob Time: 2020 Location: Alki Beach; ...; ID: 25 With: Frances, Jacob Activity: Skiing Duration: Trimmed; ID: 31 With: Frances, Jacob Activity: Skiing VIEWING; ID: 8 With: Frances, Jacob Location: Boston Common |

Figure 6: **Dataset Example**. Dialog labels include intent, slots, and multimodal coreferences.