STARK : Social Long-Term Multi-Modal Conversation with Persona Commonsense Knowledge

Anonymous ACL submission

Abstract

Humans share a wide variety of images related to their personal experiences within conversations via instant messaging tools. However, existing works focus on (1) image-sharing behavior in singular sessions, leading to limited long-term social interaction, and (2) a lack of personalized image-sharing behavior. In this work, we introduce STARK , a large-scale long-term multi-modal dialogue dataset that covers a wide range of social personas in a multi-modality format, time intervals, and images. To construct STARK automatically, we propose a novel multi-modal contextualization framework, MCU, that generates long-term multi-modal dialogue distilled from ChatGPT and our proposed Plan-and-Execute image aligner. Using our STARK , we train a multi-modal conversation model, ULTRON 7B, which demonstrates impressive visual imagination ability. Furthermore, we demonstrate the effectiveness of our dataset in human evaluation. The code, dataset, and model will be publicly released after publication.

1 Introduction

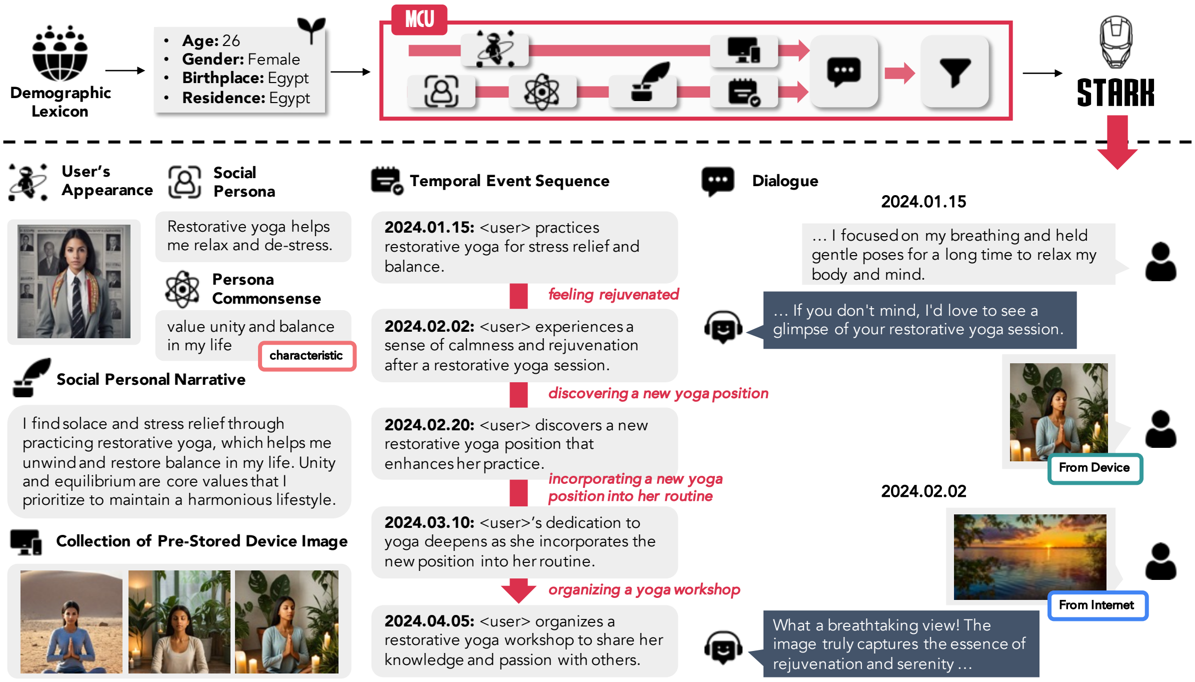
For decades, the development of empowering human-computer interaction has been steadily advancing across various domains (e.g., social dialogue (Zhou et al., 2023), writing (Lee et al., 2022a; Han et al., 2023)), multifaceted ingredients (e.g., affective user's state (Hudlicka, 2003), multi-perspective (Kammersgaard, 1988), multiple social skills (Yang et al., 2024)) and multimodality (Jaimes and Sebe, 2007) with the goal of increasing human satisfaction and engagement. To strengthen the interaction in a practicable real scenario, recent system (Shin et al., 2023) have adopted the *image-sharing behavior* (Lobinger, 2016), an interaction frequently occurring via instant messaging tools, interpreting it as a communicative practice. Consequently, previous studies have proposed multi-modal dialogue datasets through various methods, including crowdsourcing (Zang et al., 2021), social media (Feng et al., 2022), and distillation from large language models (LLMs) (Lee et al., 2022b; Aboutalebi et al., 2024; Maharana et al., 2024).

However, existing datasets are limited in their representation of *personalized image-sharing behavior* over extended periods beyond a singular time (e.g., a few hours, days, weeks), preventing trained multi-modal dialogue models from seamlessly communicating with users in real-world human-bot interactive scenarios. For example, as shown in Figure 1, depending on who is the user (i.e., human's appearance), there is a user's appearance and user's personal experience inside the shared image. Nevertheless, existing datasets regarding multi-modal dialogue do not consider multi-modality persona information (in Table 1).

To address this issue, we first introduce a large-scale **S**ocial long-**T**erm multi-mod**A**l conve**R**sation dataset with persona commonsense **K**nowledge, STARK , covering a wide variety of social personal dynamics (i.e., *demographics*, *personal experience*), more realistic time intervals, and personalized images. To construct STARK, we propose a novel framework, MCU, that distills long-term multi-modal dialogue from a large language model (LLM) [1] and our proposed Plan-and-Execute image aligner, powered by a personalized text-to-image generative model, image database retrieval, and web search, as shown in Figure 1. As a result of being grounded on various personal dynamics over a long period, STARK contains more personalized multi-modal conversation dataset. In addition, even though STARK is automatically constructed, STARK show higher preferred quality compared to other multi-modal conversation datasets (§ 4.3). With our STARK dataset, we build a **M**ulti-

---

[1] In this work, we use ChatGPT, but our proposed framework can work with any large language models, such as LLaMA-3.



Restorative yoga helps me relax and de-stress.

value unity and balance in my life

I find solace and stress relief through practicing restorative yoga, which helps me unwind and restore balance in my life. Unity and equilibrium are core values that I prioritize to maintain a harmonious lifestyle.

2024.01.15: <user> practices restorative yoga for stress relief and balance.

2024.02.02: <user> experiences a sense of calmness and rejuvenation after a restorative yoga session.

2024.02.20: <user> discovers a new restorative yoga position that enhances her practice.

2024.03.10: <user>'s dedication to yoga deepens as she incorporates the new position into her routine.

2024.04.05: <user> organizes a restorative yoga workshop to share her knowledge and passion with others.

… I focused on my breathing and held gentle poses for a long time to relax my body and mind.

… If you don't mind, I'd love to see a glimpse of your restorative yoga session.

What a breathtaking view! The image truly captures the essence of rejuvenation and serenity …

Figure 1: An overview of MCU and an example of STARK . At the top, our framework takes basic demographic information (i.e., age, gender, birthplace, residence) and generates a long-term multi-modal conversation. At the bottom, our STARK includes various information such as user's appearance, social persona, persona commonsense, personal narrative, a collection of pre-stored device images, temporal event sequences, and multi-modal dialogue. In this figure, a short sentence between two events indicates the user's episodic experience between those events (e.g., "*felling rejuvenated*").

modal c**O**nversati**ON** model, ULTRON 7B, which is fine-tuned model on top of recent multi-modal language model (Lee et al., 2024a). As a result, ULTRON achieves significant performance on dialogue-to-image retrieval task which implies the effectiveness of our dataset.

In summary, our main contributions are as follows: 1) We propose the first large-scale social long-term multi-modal conversation dataset, STARK , covering the personalized image-sharing behavior. 2) To construct STARK , we propose a multi-modal contextualization framework, MCU, that generate a multi-modal dialogue over a time period by only providing basic demographic information. 3) Using our dataset, we build a multi-modal converstation model, ULTRON 7B. 4) Through extensive experiments, we demonstrate the effectiveness and reliability of our dataset and framework in human evaluation and dialogue-to-image retrieval tasks.

2 Related Work

Multi-Modal Dialogue Dataset. In the dynamic field of multi-modal dialogue, most previous studies are categorized into two primary groups: those where the image is grounded at the beginning of the dialogue and those where the image is shared during the dialogue. The image-grounded dialogue task aims to answer questions (Antol et al., 2015; Das et al., 2017; Seo et al., 2017; Kottur et al., 2019) or generate natural conversations (Mostafazadeh et al., 2017; Shuster et al., 2018; Meng et al., 2020; Wang et al., 2021b; Zheng et al., 2021) about given images by considering the comprehensive multi-modal persona information (Ahn et al., 2023). However, in our daily conversations, we often share images relevant to the context of the dialogue via instant messaging tools. Inspired by this behavior, recently proposed image-sharing dialogue datasets have been constructed through crowd-sourcing (Zang et al., 2021), social media (Feng et al., 2022), image-text matching model (Lee et al., 2021), or annotating image-sharing moments (Lee et al., 2022b; Aboutalebi et al., 2024) through large language models (LLMs). These datasets boast impressive quality and image diversity. However, they are confined to a single session, which hinders the ability of trained models to maintain continuous conversations with users and potentially disrupts the interaction between the user and the AI assistant.

Building Dialogue Dataset using Large Language Model. To effectively address the pervasive issue of data scarcity, several innovative studies have leveraged large language models (LLMs) to construct diverse and scalable dialogue datasets. These efforts encompass personalized dialogue (Lee et al., 2022c; Jandaghi et al., 2023), multi-turn dialogue for prosocial behavior (Kim et al., 2022b), million-scale social dialogue (Kim et al., 2022a) by contextualizing rich social commonsense knowledge from a comprehensive knowledge graph (West et al., 2021), theory-of-mind (ToM) related multi-party dialogue (Kim et al., 2023), multi-hop reasoning over dialogue (Chae et al., 2023), long-term dialogue (Jang et al., 2023), and multi-modal dialogue (Lee et al., 2022b; Aboutalebi et al., 2024). Recently, a novel multi-modal dialogue dataset (Maharana et al., 2024) encompassing multiple sessions has been proposed. However, this particular dataset is designed primarily as an evaluation benchmark, thus complicating the development of an adequate multi-modal dialogue model. Furthermore, this dataset does not prioritize multi-modality in the context of personalization during image-sharing interactions. In this work, we are excited to introduce the concept of personalized multi-modal conversations over extended time intervals, meticulously considering the dynamic nature of personal interactions.

3 MCU: A Multi-Modal Contextualization Framework for Conversation Distillation

Inspired from recent study (Kim et al., 2022a), we propose a MCU, a multi-modal contextualization framework for distilling long-term multi-modal dialogue from combination of large language model (LLM) and our proposed Plan-and-Execute image aligner. Specifically, MCU consists of five steps: (1) Generating social persona attribute based on the collection of demographics (i.e., age, gender, birthplace, residence) (§ 3.2), (2) generating social persona commonsense knowledge based on the generated social persona attribute, (3) generating a social personal narrative from the commonsense knowledge, grounding on the personal narrative we (4) generate an event sequence and (5) generate a collection of pre-stored device images, (6) generating a multi-modal conversation with multiple sessions over a diverse time period, and (7) aligning a realistic and personalized image to the generated image-sharing moment by leveraging proposed Plan-and-Execute image aligner. The overview of our framework is illustrated in Figure 1. In all steps of our framework, we use ChatGPT (OpenAI, 2023) (i.e., `gpt-3.5-turbo-0125`) as our LLM. All prompt templates used in our framework are presented in Appendix C.

3.1 Motivation Behind Grounded on Demographic

Social interactions are a core component of human life, facilitated primarily through conversation (Myllyniemi, 1986). These interactions often involve sharing personal experiences, which can be abstracted into narratives or scripts (Mar and Oatley, 2008). We posit that these personal experiences are highly dependent on the individual's demographic information (e.g., age, country), thereby affect the general topic of interaction socially and culturally. Thus, we start with basic demographic information, age, gender, birthplace, residence.

3.2 Social Persona

We first randomly sample demographic information (i.e., age, gender, birthplace, residence) from a pre-defined demographic lexicon, as detailed in Appendix A, by referring to previous work (Santy et al., 2023). From the chosen demographic information, we construct a social persona [2] in the form of a short sentence for a persona category among 50 predefined persona categories. Additionally, we generate a social persona attribute simultaneously with the social persona sentence. The social persona attribute can be formally represented as a triple (e_1, r, e_2), where e_1, r, and e_2 denote the persona subject, persona category, and persona entity, respectively. The persona entity follows a key-value format. For example, in the social persona attribute "I am from London," e_1 is "I," r is "location," and e_2 is "(city-state, London)." To save time and reduce costs, we generate 30 persona attributes and sentences given a single demographic information set.

3.3 Virtual Human Face

Since STARK covers personalized image-sharing behavior, we generate a virtual human face using the SDXL-Lightning (Lin et al., 2024) model. [3]

---

[2] In this work, we regard a persona as a user profile, following the definition of previous work (Lee et al., 2022c).

[3] Unfortunately, we intended to use a more specialized model (Li et al., 2024a) for human face generation; however,

The virtual human face is created based on a predefined human attribute collection from recent work (Li et al., 2024a), with the full human attribute information. Creating a virtual human face initially allows us to generate personalized images later (in § 3.8) with higher quality and more personalized experiences, resulting in significant scores in human evaluation (§ 4.3).

3.4 Social Persona Commonsense Knowledge

Recent research has introduced a large-scale persona-grounded commonsense knowledge graph called PEACOK (Gao et al., 2023). This graph is symbolically represented in the form of triples (head, relation, tail), where relation denotes a defined *persona frame* concept, which formalizes five commonsense aspects of persona knowledge: characteristics, routines/habits, goals/plans, experiences, and relationships. This comprehensive knowledge graph encompasses a broad spectrum of persona knowledge at scale.

However, this commonsense knowledge graph has two major limitations: (1) The coverage of persona head value is limited to the *CapableOf* relation, which typically encompasses occupation-related sentences (e.g., "I am a programmer," "I am a basketball player"). In reality, persona identity can be expressed through a broad range of information, such as "I have two dogs" in terms of possession. (2) The inferred attribute knowledge based on the given commonsense relation varies depending on demographic information. For example, even when providing the same persona head value and the same commonsense relation, the persona commonsense inference will represent distinct meanings based on demographic differences.

To address these limitations, we prompt ChatGPT to infer the persona attribute knowledge considering the user's demographic information and social persona attribute (§ 3.2), which covers diverse persona categories, for five persona relations.

3.5 Personal Narrative

Symbolic Form to Sentence Form. We convert the generated persona commonsense knowledge graphs into simple sentences by applying predefined templates (presented in the Appendix C) for each relation. To make the sentences more plausible and natural in terms of world knowledge, we use actual names based on the given birthplace country, selecting from the Top-1K names for each country [4].

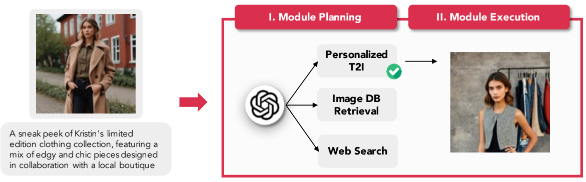

A sneak peek of Kristin's limited edition clothing collection, featuring a mix of edgy and chic pieces designed in collaboration with a local boutique

Figure 2: An illustration of our Plan-and-Execute image aligner process.

Sentence Form to Personal Narrative. Next, we prompt ChatGPT to transform the sentence form into a short personal narrative consisting of two or three sentences with detailed information, following recent work (Kim et al., 2022a).

3.6 Temporal Event Sequence

Starting from the generated personal narrative, we prompt ChatGPT to generate a temporal event sequence consisting of multiple sequential events. We prompt ChatGPT to generate time intervals and episodic experiences with operations between two events. There are two types of experience operations: add and update. If a new experience occurs, it is marked as an add operation. If a previous experience is modified, it is marked as an update operation. The

3.7 Collection of Pre-Stored Device Images

Before generating multi-modal conversations, we ask ChatGPT to infer the possible image descriptions that might be pre-stored on a user's device (e.g., mobile or laptop) based on the personal narrative (§ 3.5). This step makes multi-modal conversations more practical and similar to real-world scenarios, such as when a user shares an everyday photo on online social media (Maclean et al., 2022). Specifically, we generate five image descriptions along with corresponding image categories. We then generate photo-realistic images using our proposed Plan-and-Execute image aligner (see details in § 3.8).

3.8 Multi-Modal Conversation

In this step, we generate a long-term multi-modal conversation between the user and an AI assistant, utilizing the constructed event sequence (§ 3.6) and the collection of pre-stored device image descriptions (§ 3.7). Since each episode consists of

---

this model was not publicly available at the time of data construction, so we opted for the alternative model.

[4] `https://github.com/philipperemy/name-dataset`

multiple session dialogues, we generate each session sequentially. Concretely, the second session is influenced by useful information (i.e., history of events, device images) from the previous session.

Generating Image-Sharing Moment. Drawing inspiration from recent works (Lee et al., 2022b, 2023; Aboutalebi et al., 2024), we employ ChatGPT to create a multi-modal conversation that includes an image-sharing moment in text format, specifically encompassing image description, rationale, image source, keywords, and index of pre-stored image in device. To ensure the multi-modal conversations are as realistic and natural as possible, closely mirroring real-life scenarios, we prompt ChatGPT to specify the source of the shared image (e.g., from the internet or a user's device) when describing an image-sharing moment. Furthermore, if the shared image is already part of a collection of pre-stored device image descriptions, we prompt ChatGPT to determine which image description to select.

Plan-and-Execute Image Aligner. Since STARK is designed to include personalized image-sharing behavior over an extended period, users can share photos that reflect their personal experiences. For example, a user might share a photo with the description, "*I visited the Eiffel Tower last week*", which includes an image of them in front of the Eiffel Tower. Additionally, users can share non-human-centric photos, such as *"a meal I had yesterday"*, which also conveys personal experiences. Therefore, we need to determine the most appropriate module to synthesize images relevant to the given image descriptions.

Following recent works related to tool-based AI agents (Shen et al., 2024), as illustrated in Figure 2, we first conduct module planning to select the most appropriate module based on the given image description by leveraging ChatGPT. The options include a personalized text-to-image generator, image database retrieval, and web search. After selecting the appropriate module, we proceed to execute it. Specifically, if the personalized text-to-image generator is chosen, we utilize the PhotoMaker (Li et al., 2023) model, demonstrating impressive performance in customizing human faces. If image database retrieval is selected, we use the CLIP (Radford et al., 2021) (i.e., `ViT-L/14@336px`) to retrieve relevant images from prepared source image datasets: CC12M (Changpinyo et al., 2021), RedCaps12M (Desai et al., 2021), ChartQA (Masry et al., 2022), AI2D (Kembhavi et al., 2016), and MathVision (Wang et al., 2024). We employ Bing Search [5] for web search, similar to previous work (Maharana et al., 2024).

3.9 Post-processing and Filtering

We remove episode conversations that have less than four sessions or more than six sessions (7.1%); remove duplicate persona attributes (19.8%). In addition, we remove potentially dangerous and harmful dialogues that need the intervention using Canary (Kim et al., 2022b) and unsuitable images using NSFW detector [6]. Furthermore, we filter out unaligned images to the generated image descriptions using Pick-a-pic (Kirstain et al., 2023) score. Finally, we obtain roughly 0.5 M session dialogues in total.

4 Analysis of STARK

In this section, we conduct comprehensive analysis of STARK in terms of diverse perspectives: Comparison analysis to existing datasets (§ 4.1), multifaceted analysis (§ 4.2), and human evaluation (§ 4.3).

4.1 Comparison to Existing Datasets

In Table 1, we compare STARK with other existing datasets in terms of multi-modality and long-term continuity. In summary, STARK uniquely accomplishes a long-term multi-modal conversation, encompassing extensive multi-modal persona information and featuring a comparable data scale (0.5M sessions) to SODA (1M) and Conversation Chronicles (1M). Unlike other multi-modal dialogue datasets, which focus on singular sessions, STARK achieves a significantly larger scale of session dialogues and images. Additionally, STARK stands out among long-term dialogue datasets by exclusively covering multi-modal dialogue and persona information, including social persona attributes and pre-stored device images. While the LOCOMO dataset also addresses long-term multi-modal conversations, it lacks multi-modal persona information and is limited in scale (50 episodes), being designed mainly for evaluation benchmarks. Therefore, STARK is the first to offer a large-scale long-term multi-modal conversation dataset, enabling the development of a well-generalized multi-modal dialogue model.

---

[5] `https://pypi.org/project/icrawler/`

[6] `https://huggingface.co/Falconsai/nsfw_image_detection`

Dataset	Train set?	Dialogue Modality	Persona Modality	Multple Session?	Collection	# of E.	# of S.	# of I.	Avg. U./S.	Avg. I./E.	Avg. I./S.
MMDD (Lee et al., 2021)	✓	T, V	×	×	VSRN + Human	-	17K	17K	12.74	-	1.76
PhotoChat (Zang et al., 2021)	✓	T, V	×	×	Human	-	11K	10K	11.56	-	1
MMDialog (Feng et al., 2022)	✓	T, V	×	×	Social media	-	1M	1.5M	4.56	-	2.82
DialogCC (Lee et al., 2022b)	✓	T, V	×	×	GPT-4, CLIP	-	83K	120K	8.2	-	7.83
SODA (Kim et al., 2022a)	✓	T	×	×	InstructGPT	-	1M	-	7.6	-	-
MSC (Xu et al., 2021) (train; 1-4 sessions)	✓	T	T	✓	Human	5K	16K	-	13.4	-	-
Conversation Chronicles (Jang et al., 2023)	✓	T	T	✓	ChatGPT	200K	1M	-	11.7	-	-
LoCoMo (Maharana et al., 2024)	×	T, V	T	✓	ChatGPT + Human	50	1K	2K	15.8	32.3	3.72
STARK **(Ours)**	✓	T, V	T, V	✓	ChatGPT, Diffusion	93K	0.5M	0.9M	10.5	9.94	1.86

Table 1: Comparison of STARK with existing datasets in terms of multi-modality and long-range continuity: MMDD, PhotoChat, MMDialog, DialogCC, SODA, MSC, Conversation Chronicles, and LoCoMo. V and T denote virtual and textual modality, respectively. E., S., and I. denote episode, session, and image, respectively. I.E. and I.S. denote images by episode and images by a single session, respectively. STARK is the first to achieve a long-term multi-modal conversation that covers multi-modal persona information and includes a large scale, which leads to a well-generalized multi-modal conversation model. VSRN (Li et al., 2019) is the text-image matching model.

Demographic				Persona	
Age/Gender	Ratio	Country	Ratio	Entity	Ratio
50-60	14.12	China	7.85	animal	4.32
20-30	13.68	USA	7.79	profession	4.18
60-70	13.29	UK	6.73	school name	2.68
40-50	13.19	Russia	6.43	book author	2.68
80-90	12.88	India	5.75	music artist	2.55
30-40	12.1	Japan	5.72	music instrument	2.41
70-80	10.96	Brazil	5.64	subject	2.36
10-20	9.76	Germany	5.6	food	2.35
Male	51.29	Italy	5.41	sport	2.35
Female	48.71	South Korea	5.23	season	2.34

Table 2: The ratio (%) of age groups and gender, along with the ratio of Top-10 persona entity categories and countries in STARK .

4.2 Multifaceted Analysis

Demographic. As shown in Table 2, our dataset exhibits a fairly balanced distribution across age, gender, and country. This suggests that our dataset is less likely to introduce biases during model training. Among the age groups, individuals aged 50 to 60 are the most represented. This indicates the potential applicability of our dataset in scenarios where an AI assistant needs to continuously care for older users, as highlighted in recent studies (Bae et al., 2022b,a). The gender distribution is nearly equal, implying a lower possibility of gender bias problem.

Social Persona. We derive the ratio of the Top-10 persona entity categories corresponding to the generated persona entity key from ChatGPT (in § 3.2). As shown in Table 2, we observe that the categories of personas most commonly encountered in our everyday surroundings, such as animals and professions, are the most prevalent. The remaining categories are evenly distributed. This indicates that our dataset is well-balanced, providing a comprehensive understanding of various personas without bias towards any specific category.

4.3 Human Evaluation

To quantify the quality of STARK , we conduct two different kinds of human evaluation, (1) human ratings and (2) head-to-head comparison, based on several evaluation criteria.

Human Ratings. We meticulously evaluate the quality of STARK on seven distinct criteria: (1) coherence, (2) consistency, (3) image-sharing turn relevance, (4) image-dialogue relevance, (5) image-persona relevance, (6) time interval, and (7) experience. Each human evaluator rates 100 randomly chosen episode samples (totaling 500 session dialogues) using a detailed 4-point Likert scale for all criteria. Further explanations of each evaluation item and the recruitment process for human evaluators are provided in the Appendix D and Appendix F. On average, we achieve significantly higher scores: 3.4 for coherence, 3.52 for consistency, 3.07 for image-sharing turn relevance, 2.49 for image-dialogue relevance, 3.35 for image-persona relevance, 3.75 for time interval, and 3.73 for experience. Additionally, we measure the inter-rater agreement (IA) using Krippendorff's α, obtaining a value of 0.27, which indicates a fair level of agreement. These results underscore the reliability of MCU in generating long-term multi-modal conversations starting with only basic demographic information.

Head-to-Head Comparison. Since STARK is automatically constructed by leveraging various generative models, we assess the quality gap between our dataset and other high-quality and realis-

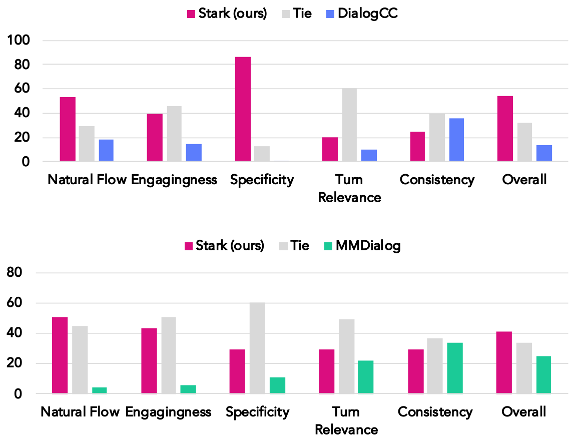


Figure 3: Results of head-to-head comparison between STARK (ours) and two existing datasets, DialogCC (Lee et al., 2022b) and MMDialog (Feng et al., 2022), on six evaluation criteria.

tic datasets: DialogCC (Lee et al., 2022b) (which has recently demonstrated high quality) and MMDialog (Feng et al., 2022) (which is derived from social media) by conducting a head-to-head comparison. Given that DialogCC and MMDialog are singular session datasets, we randomly sample 100 session dialogues from STARK and also randomly sample the same number of dialogues from DialogCC and MMDialog. We then evaluate them based on six criteria: (1) natural flow, (2) engagingness, (3) specificity, (4) image-sharing turn relevance, (5) image-dialogue consistency, and (6) overall quality. Further details are provided in the Appendix D. Overall, as illustrated in Figure 3, STARK achieves better scores than both DialogCC and MMDialog across all criteria. Specifically, our dataset exhibits more engaging and naturally flowing conversations, particularly surpassing MMDialog by a large margin. Interestingly, human evaluators frequently select “Tie” for the items related to image-sharing turn relevance and image-dialogue consistency compared to other datasets. These results imply that, despite being constructed using generative models such as ChatGPT and our proposed image aligner (which includes several diffusion models), our dataset ensures the relevance of image-sharing moments and maintains the quality of generated images. This demonstrates the robustness and reliability of our proposed framework in producing coherent and engaging multi-modal conversations, even when compared to datasets utilizing actual photo-realistic images.

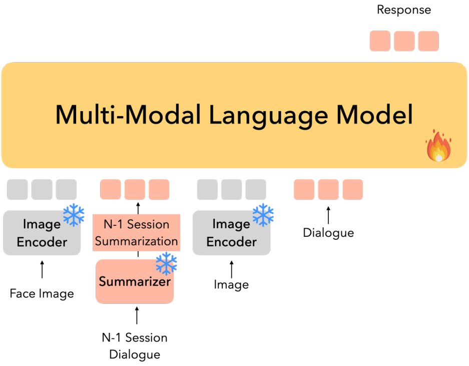


Figure 4: The overall architecture of ULTRON .

5 ULTRON

With STARK, we train a multi-modal conversation model named ULTRON 7B. This model is designed to understand diverse social and personal dynamics along with previous interactions, enabling it to identify the appropriate moments for image sharing and retrieve relevant images based on the dialogue context. The overall architecture of ULTRON is illustrated in Figure 4.

5.1 Motivation behind Model Design

Backbone Model. Identifying the optimal moment for image sharing presents a significant challenge due to the subjective nature of this behavior, even for humans (Lee et al., 2023). Additionally, retrieving relevant images based on dialogue context is non-trivial, as critical evidence is often dispersed throughout the entire conversation (Chae et al., 2023; Wang et al., 2023). To address these challenges, we employ the recently proposed Meteor (Lee et al., 2024a) model, which significantly enhances multi-modal reasoning capabilities across diverse tasks by introducing the novel concept of “traversal of rationale.” Consequently, we initiate the training of ULTRON on the top of the Meteor model.

Input & Output. Recent studies have demonstrated the powerful visual imagination capabilities of large language models (Lee et al., 2022b, 2023; Li et al., 2024b). Inspired by these findings, we train ULTRON to alternatively generate the image-sharing moment in a text format (without generating the image directly), specifically “`<RET> <h> image description </h>`." This method allows ULTRON to produce image descriptions that are better aligned with the given dialogue context, benefiting CLIP or generative models. In future work, since our model does not directly produce images, we will focus on developing a multi-modal language model capable of generating or retrieving images, following recent findings in the field (Zheng et al., 2023; Koh et al., 2024).

5.2 Model Architecture

Ultron comprises a vision encoder, a vision projector, a summarizer, and the backbone multi-modal language model from the Meteor model. The architectures of the vision encoder, vision projector, and backbone model are consistent with those employed in the Meteor model. For the summarizer, we first construct a summarization dataset. Specifically, we randomly sample 10,000 episodes, encompassing a total of 53,317 session dialogues, and employ ChatGPT to generate summaries for these session dialogues. The prompt used for this task is detailed in the Appendix C. Utilizing this constructed dataset, we fine-tune the LLaMA-3 8B model (AI@Meta, 2024) [7] with Q-LoRA (Dettmers et al., 2024) tuning, using 64 rank and 16 alpha parameters. This model is subsequently used to generate summaries for all session dialogues in our dataset. We then filter out unsuitable summaries, such as those containing repetition, ensuring that only high-quality summaries are included in the training dataset for Ultron.

6 Experiments

6.1 Experimental Setup

Datasets. To build generalized multi-modal conversation model that converse with user on diverse social situations, we train Ultron on Stark and Mini-Gemini Instruction (Li et al., 2024b). We evaluate Ultron on PhotoChat (Zang et al., 2021).

Task Definition. We perform Ultron on dialogue-to-image retrieval task which is standard downstream task regarding multi-modal dialogue. The Dialogue-to-Image Retrieval task involves retrieving the relevant image based on the dialogue context.

Evaluation Metrics. We use the widely adopted Recall@K and MRR metric.

6.2 Results

As shown in Table 3, Ultron achieves significant performance improvements in the dialogue-to-image retrieval task compared to several other methods. Notably, Ultron outperforms the recent LLM-based framework, DribeR (Lee et al., 2023). Interestingly, recent large multi-modal models, such as LLaVA v1.5 and GPT-4V, exhibit relatively lower performance in a zero-shot setting. In contrast, Ultron achieves remarkable performance, underscoring the effectiveness of our dataset in enhancing complex image-sharing behaviors.

Model	R@1	R@5	R@10	MRR
Fine-tuned Performance				
BM25	6.6	15.4	23.0	-
DE	9.0	26.4	35.7	-
VSE++	10.2	25.4	34.2	-
SCAN	10.4	27.0	37.1	-
VLMo	13.8	30.0	39.4	-
ViLT	11.5	25.6	33.8	-
PaCE	15.2	36.7	49.6	-
DialCLIP	19.5	44.0	55.8	-
VLM, zero-shot				
CLIP-base	13.7	28.0	35.2	20.8
CLIP-large	14.1	28.7	35.3	21.5
Large Multi-Modal Model				
LLaVA v1.5 7B	11.1	26.5	33.3	18.8
LLaVA v1.5 13B	12.1	25.6	32.3	19.3
MiniGPT-4$_{\text{Vicuna 7B}}$	11.6	26.5	34.0	19.1
MiniGPT-4$_{\text{Vicuna 13B}}$	11.7	27.7	35.5	19.8
Qwen-VL-Chat 7B	12.1	27.4	36.1	20.2
GPT4-V	13.8	27.9	35.9	21.3
LLM-based Framework				
DribeR $_{\text{ChatGPT 0613}}$	26.6	46.1	54.2	36.0
DribeR $_{\text{ChatGPT 1106}}$	26.3	45.6	54.3	35.4
DribeR $_{\text{GPT-4 1106}}$	28.3	47.4	55.2	37.6
DribeR $_{\text{Vicuna-13B}}$	25.8	45.0	53.1	35.0
DribeR $_{\text{LLaMa2-Chat-70B}}$	24.5	43.5	52.6	34.0
Ultron	**31.2**	**53.7**	**65.0**	**46.1**

Table 3: Comparison results of the dialogue-to-image retrieval task on PhotoChat (Zang et al., 2021).

7 Conclusion

In this work, we first propose a social long-term multi-modal conversation dataset, Stark , which is fully automatically constructed through our proposed framework, Mcu. This framework comprises ChatGPT and our proposed Plan-and-Execute image aligner. Through extensive experiments, we demonstrate that our dataset has comparable quality to other existing datasets. Additionally, using our dataset, we build a multi-modal conversation model, Ultron , which achieves significant performance in the dialogue-to-image retrieval task.

---

[7] `https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct`

Limitations

Inconsistent Personalized Images. To construct a dataset encompassing personalized image-sharing behavior, we utilized a personalized text-to-image generative model. However, this occasionally led to instances where the appearance of the user was not consistently maintained across some samples. Additionally, when generating images featuring groups, there was a tendency for multiple individuals in the group to appear identical to the user's appearance. Despite applying various filtering methods to mitigate these issues, complete elimination was not achieved. Given the rapid advancements in generative models, we anticipate that future, more advanced models will enable the creation of datasets with enhanced consistency.

Building Role-Specified AI Assistant. When constructing our dataset, we did not provide the AI assistant with any specific personality traits or preference information (Lee et al., 2024b). For future research, it would be advantageous to develop datasets or models that incorporate social relational information (Zhou et al., 2023; Jang et al., 2023) (e.g., friend, colleague), a broader range of conversational styles (Han et al., 2022), and personality traits. This approach could enhance social interactions and foster a closer relationship between the AI assistant and users.

Ethical Considerations

Despite applying various filtering methods to exclude unsuitable samples, potential issues may still exist within our proposed framework. Firstly, the generated dialogue might propagate social or cultural biases, as ChatGPT can produce harmful content, including social biases and offensive remarks (Baheti et al., 2021; Hartvigsen et al., 2022). Secondly, the generated images may also reflect unfaithful and socially biased content when using Stable Diffusion (Rombach et al., 2022). As reported by (Wang et al., 2021a), even when providing gender-neutral queries to the CLIP model (Radford et al., 2021), the model occasionally retrieves images that cause gender-bias issues. We are concerned that these problematic issues may persist in the augmented dataset. Consequently, a multi-modal dialogue model trained on this dataset might sometimes generate or retrieve biased images. It is crucial to consider these issues carefully when developing a multi-modal dialogue model.

References

Hossein Aboutalebi, Hwanjun Song, Yusheng Xie, Arshit Gupta, Justin Sun, Hang Su, Igor Shalyminov, Nikolaos Pappas, Siffi Singh, and Saab Mansour. 2024. Magid: An automated pipeline for generating synthetic multi-modal datasets. *arXiv preprint arXiv:2403.03194*.

Jaewoo Ahn, Yeda Song, Sangdoo Yun, and Gunhee Kim. 2023. Mpchat: Towards multimodal persona-grounded conversation. *arXiv preprint arXiv:2305.17388*.

AI@Meta. 2024. Llama 3 model card.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Sanghwan Bae, Donghyun Kwak, Soyoung Kang, Min Young Lee, Sungdong Kim, Yuin Jeong, Hyeri Kim, Sang-Woo Lee, Woomyoung Park, and Nako Sung. 2022a. Keep me updated! memory management in long-term conversations. *arXiv preprint arXiv:2210.08750*.

Sanghwan Bae, Donghyun Kwak, Sungdong Kim, Donghoon Ham, Soyoung Kang, Sang-Woo Lee, and Woomyoung Park. 2022b. Building a role specified open-domain dialogue system leveraging large-scale language models. *arXiv preprint arXiv:2205.00176*.

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. *arXiv preprint arXiv:2108.11830*.

Hyungjoo Chae, Yongho Song, Kai Tzu-iunn Ong, Taeyoon Kwon, Minjin Kim, Youngjae Yu, Dongha Lee, Dongyeop Kang, and Jinyoung Yeo. 2023. Dialogue chain-of-thought distillation for commonsense-aware conversational agents. *arXiv preprint arXiv:2310.09343*.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335.

Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems*, 36.

Jiazhan Feng, Qingfeng Sun, Can Xu, Pu Zhao, Yaming Yang, Chongyang Tao, Dongyan Zhao, and Qingwei Lin. 2022. Mmdialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. *arXiv preprint arXiv:2211.05719*.

Silin Gao, Beatriz Borges, Soyoung Oh, Deniz Bayazit, Saya Kanno, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2023. Peacok: Persona commonsense knowledge for consistent and engaging narratives. *arXiv preprint arXiv:2305.02364*.

Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, et al. 2023. Recipe: How to integrate chatgpt into efl writing education. In *Proceedings of the Tenth ACM Conference on Learning@ Scale*, pages 416–420.

Seungju Han, Beomsu Kim, Jin Yong Yoo, Seokjun Seo, Sangbum Kim, Enkhbayar Erdenee, and Buru Chang. 2022. Meet your favorite character: Open-domain chatbot mimicking fictional characters with only a few utterances. *arXiv preprint arXiv:2204.10825*.

Thomas Hartvigsen, Saadia Gabriel, Hamid Palangi, Maarten Sap, Dipankar Ray, and Ece Kamar. 2022. Toxigen: A large-scale machine-generated dataset for adversarial and implicit hate speech detection. *arXiv preprint arXiv:2203.09509*.

Eva Hudlicka. 2003. To feel or not to feel: The role of affect in human–computer interaction. *International journal of human-computer studies*, 59(1-2):1–32.

Alejandro Jaimes and Nicu Sebe. 2007. Multimodal human–computer interaction: A survey. *Computer vision and image understanding*, 108(1-2):116–134.

Pegah Jandaghi, XiangHai Sheng, Xinyi Bai, Jay Pujara, and Hakim Sidahmed. 2023. Faithful persona-based conversational dataset generation with large language models. *arXiv preprint arXiv:2312.10007*.

Jihyoung Jang, Minseong Boo, and Hyounghun Kim. 2023. Conversation chronicles: Towards diverse temporal and relational dynamics in multi-session conversations. *arXiv preprint arXiv:2310.13420*.

John Kammersgaard. 1988. Four different perspectives on human–computer interaction. *International Journal of Man-Machine Studies*, 28(4):343–362.

Aniruddha Kembhavi, Mike Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*, pages 235–251. Springer.

Hyunwoo Kim, Jack Hessel, Liwei Jiang, Peter West, Ximing Lu, Youngjae Yu, Pei Zhou, Ronan Le Bras, Malihe Alikhani, Gunhee Kim, et al. 2022a. Soda: Million-scale dialogue distillation with social commonsense contextualization. *arXiv preprint arXiv:2212.10465*.

Hyunwoo Kim, Melanie Sclar, Xuhui Zhou, Ronan Le Bras, Gunhee Kim, Yejin Choi, and Maarten Sap. 2023. Fantom: A benchmark for stress-testing machine theory of mind in interactions. *arXiv preprint arXiv:2310.15421*.

Hyunwoo Kim, Youngjae Yu, Liwei Jiang, Ximing Lu, Daniel Khashabi, Gunhee Kim, Yejin Choi, and Maarten Sap. 2022b. Prosocialdialog: A prosocial backbone for conversational agents. *arXiv preprint arXiv:2205.12688*.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. 2023. Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in Neural Information Processing Systems*, 36:36652–36663.

Jing Yu Koh, Daniel Fried, and Russ R Salakhutdinov. 2024. Generating images with multimodal language models. *Advances in Neural Information Processing Systems*, 36.

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*.

Byung-Kwan Lee, Chae Won Kim, Beomchan Park, and Yong Man Ro. 2024a. Meteor: Mamba-based traversal of rationale for large language and vision models. *arXiv preprint arXiv:2405.15574*.

Mina Lee, Percy Liang, and Qian Yang. 2022a. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–19.

Nyoungwoo Lee, Suwon Shin, Jaegul Choo, Ho-Jin Choi, and Sung-Hyun Myaeng. 2021. Constructing multi-modal dialogue dataset by replacing text with semantically relevant images. *arXiv preprint arXiv:2107.08685*.

Seongyun Lee, Sue Hyun Park, Seungone Kim, and Minjoon Seo. 2024b. Aligning to thousands of preferences via system message generalization. *arXiv preprint arXiv:2405.17977*.

Young-Jun Lee, Jonghwan Hyeon, and Ho-Jin Choi. 2023. Large language models can share images, too! *arXiv preprint arXiv:2310.14804*.

Young-Jun Lee, Byungsoo Ko, Han-Gyu Kim, and Ho-Jin Choi. 2022b. Dialogcc: Large-scale multi-modal dialogue dataset. *arXiv preprint arXiv:2212.04119*.

Young-Jun Lee, Chae-Gyun Lim, Yunsu Choi, Ji-Hui Lm, and Ho-Jin Choi. 2022c. Personachatgen: Generating personalized dialogues using gpt-3. In *Proceedings of the 1st workshop on customized chat grounding persona and knowledge*, pages 29–48.

Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4654–4662.

Shikai Li, Jianglin Fu, Kaiyuan Liu, Wentao Wang, Kwan-Yee Lin, and Wayne Wu. 2024a. Cosmicman: A text-to-image foundation model for humans. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6955–6965.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*.

Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. 2023. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*.

Shanchuan Lin, Anran Wang, and Xiao Yang. 2024. Sdxl-lightning: Progressive adversarial diffusion distillation.

Katharina Lobinger. 2016. Photographs as things–photographs of things. a texto-material perspective on photo-sharing practices. *Information, Communication & Society*, 19(4):475–488.

Julie Maclean, Yeslam Al-Saggaf, and Rachel Hogg. 2022. Instagram photo sharing and its relationships with social connectedness, loneliness, and well-being. *Social Media+ Society*, 8(2):20563051221107650.

Adyasha Maharana, Dong-Ho Lee, Sergey Tulyakov, Mohit Bansal, Francesco Barbieri, and Yuwei Fang. 2024. Evaluating very long-term conversational memory of llm agents. *arXiv preprint arXiv:2402.17753*.

Raymond A Mar and Keith Oatley. 2008. The function of fiction is the abstraction and simulation of social experience. *Perspectives on psychological science*, 3(3):173–192.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Yuxian Meng, Shuhe Wang, Qinghong Han, Xiaofei Sun, Fei Wu, Rui Yan, and Jiwei Li. 2020. Openvidial: A large-scale, open-domain dialogue dataset with visual contexts. *arXiv preprint arXiv:2012.15015*.

Nasrin Mostafazadeh, Chris Brockett, Bill Dolan, Michel Galley, Jianfeng Gao, Georgios P Spithourakis, and Lucy Vanderwende. 2017. Image-grounded conversations: Multimodal context for natural question and response generation. *arXiv preprint arXiv:1701.08251*.

Rauni Myllyniemi. 1986. Conversation as a system of social interaction. *Language & Communication*.

OpenAI. 2023. ChatGPT. `https://openai.com/blog/chatgpt/`.

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.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. *arXiv preprint arXiv:2306.01943*.

Paul Hongsuck Seo, Andreas Lehrmann, Bohyung Han, and Leonid Sigal. 2017. Visual reference resolution using attention memory for visual dialog. *Advances in neural information processing systems*, 30.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. 2024. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face. *Advances in Neural Information Processing Systems*, 36.

Donghoon Shin, Soomin Kim, Ruoxi Shang, Joonhwan Lee, and Gary Hsieh. 2023. Introbot: Exploring the use of chatbot-assisted familiarization in online collaborative groups. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, pages 1–13.

Kurt Shuster, Samuel Humeau, Antoine Bordes, and Jason Weston. 2018. Image chat: Engaging grounded conversations. *arXiv preprint arXiv:1811.00945*.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Hongru Wang, Rui Wang, Fei Mi, Yang Deng, Zezhong Wang, Bin Liang, Ruifeng Xu, and Kam-Fai Wong. 2023. Cue-cot: Chain-of-thought prompting for responding to in-depth dialogue questions with llms. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12047–12064.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021a. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*.

Ke Wang, Junting Pan, Weikang Shi, Zimu Lu, Mingjie Zhan, and Hongsheng Li. 2024. Measuring multimodal mathematical reasoning with math-vision dataset. *arXiv preprint arXiv:2402.14804*.

Shuhe Wang, Yuxian Meng, Xiaoya Li, Xiaofei Sun, Rongbin Ouyang, and Jiwei Li. 2021b. Openvidial 2.0: A larger-scale, open-domain dialogue generation dataset with visual contexts. *arXiv preprint arXiv:2109.12761*.

Peter West, Chandra Bhagavatula, Jack Hessel, Jena D Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi. 2021. Symbolic knowledge distillation: from general language models to commonsense models. *arXiv preprint arXiv:2110.07178*.

Jing Xu, Arthur Szlam, and Jason Weston. 2021. Beyond goldfish memory: Long-term open-domain conversation. *arXiv preprint arXiv:2107.07567*.

Diyi Yang, Caleb Ziems, William Held, Omar Shaikh, Michael S Bernstein, and John Mitchell. 2024. Social skill training with large language models. *arXiv preprint arXiv:2404.04204*.

Xiaoxue Zang, Lijuan Liu, Maria Wang, Yang Song, Hao Zhang, and Jindong Chen. 2021. Photochat: A human-human dialogue dataset with photo sharing behavior for joint image-text modeling. *arXiv preprint arXiv:2108.01453*.

Kaizhi Zheng, Xuehai He, and Xin Eric Wang. 2023. Minigpt-5: Interleaved vision-and-language generation via generative vokens. *arXiv preprint arXiv:2310.02239*.

Yinhe Zheng, Guanyi Chen, Xin Liu, and Ke Lin. 2021. Mmchat: Multi-modal chat dataset on social media. *arXiv preprint arXiv:2108.07154*.

Xuhui Zhou, Hao Zhu, Leena Mathur, Ruohong Zhang, Haofei Yu, Zhengyang Qi, Louis-Philippe Morency, Yonatan Bisk, Daniel Fried, Graham Neubig, et al. 2023. Sotopia: Interactive evaluation for social intelligence in language agents. *arXiv preprint arXiv:2310.11667*.

A Pre-defined Demographic Lexicon

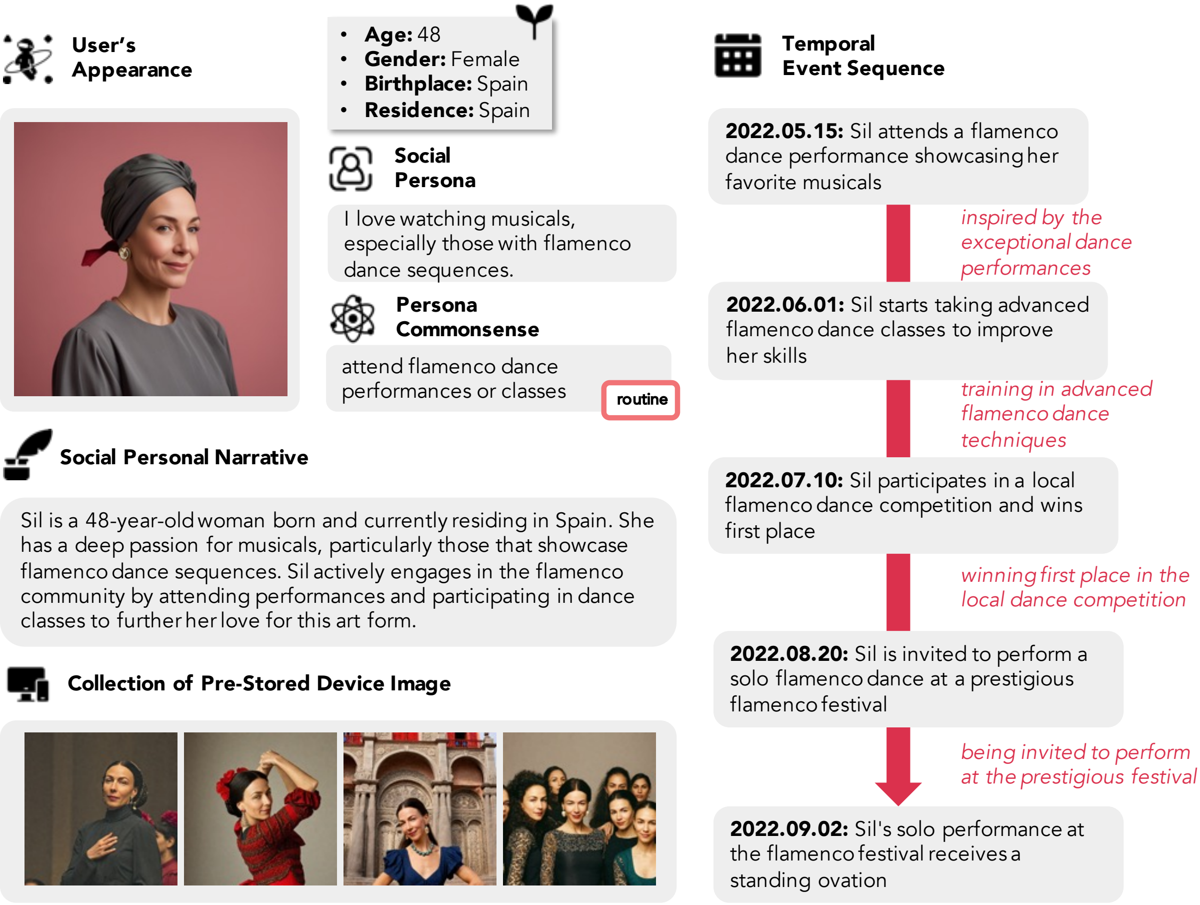
Age Group. The age groups are defined as follows: '10-20', '20-30', '30-40', '40-50', '50-60', '60-70', '70-80', and '80-90'. From these groups, a group is first selected at random. Subsequently, an age within the selected group is chosen randomly. For example, if the '10-20' group is selected, a number between 10 and 20 is then randomly chosen.

Gender. We consider two gender categories: male and female, with the selection made randomly. Although it is essential to consider fairness, including non-binary gender categories for fair AI practices, the current attribute lexicon for human face generation does not support non-binary options. Therefore, we have excluded it to maintain the quality of the generated human face images. In future work, we aim to incorporate socially-aware fairness in our MCU to develop a socially-balanced multi-modal dialogue dataset.

Birthplace & Residence. To determine the birthplace and residence, we first prepare a country list, as referenced from previous work (Santy et al., 2023), that includes 19 countries: 'United States of America', 'China', 'Japan', 'India', 'United Arab Emirates', 'France', 'Germany', 'Italy', 'South Korea', 'Saudi Arabia', 'Kazakhstan', 'Brazil', 'Mexico', 'Egypt', 'Argentina', 'Russia', 'United Kingdom', 'Spain', and 'Canada'. We randomly select a country from this list to assign as the birthplace and residence. In 70% of the cases, the birthplace and residence are the same, while in 30% of the cases, the birthplace and residence are different (e.g., due to immigration).

B A Full Example of STARK

In this section, we show a full conversation of STARK in Figure 5, Figure 6, Figure 7.



- **Age:** 48
- **Gender:** Female
- **Birthplace:** Spain
- **Residence:** Spain

Social Persona

I love watching musicals, especially those with flamenco dance sequences.

Persona Commonsense

attend flamenco dance performances or classes

Social Personal Narrative

Sil is a 48-year-old woman born and currently residing in Spain. She has a deep passion for musicals, particularly those that showcase flamenco dance sequences. Sil actively engages in the flamenco community by attending performances and participating in dance classes to further her love for this art form.

Temporal Event Sequence

2022.05.15: Sil attends a flamenco dance performance showcasing her favorite musicals

inspired by the exceptional dance performances

2022.06.01: Sil starts taking advanced flamenco dance classes to improve her skills

training in advanced flamenco dance techniques

2022.07.10: Sil participates in a local flamenco dance competition and wins first place

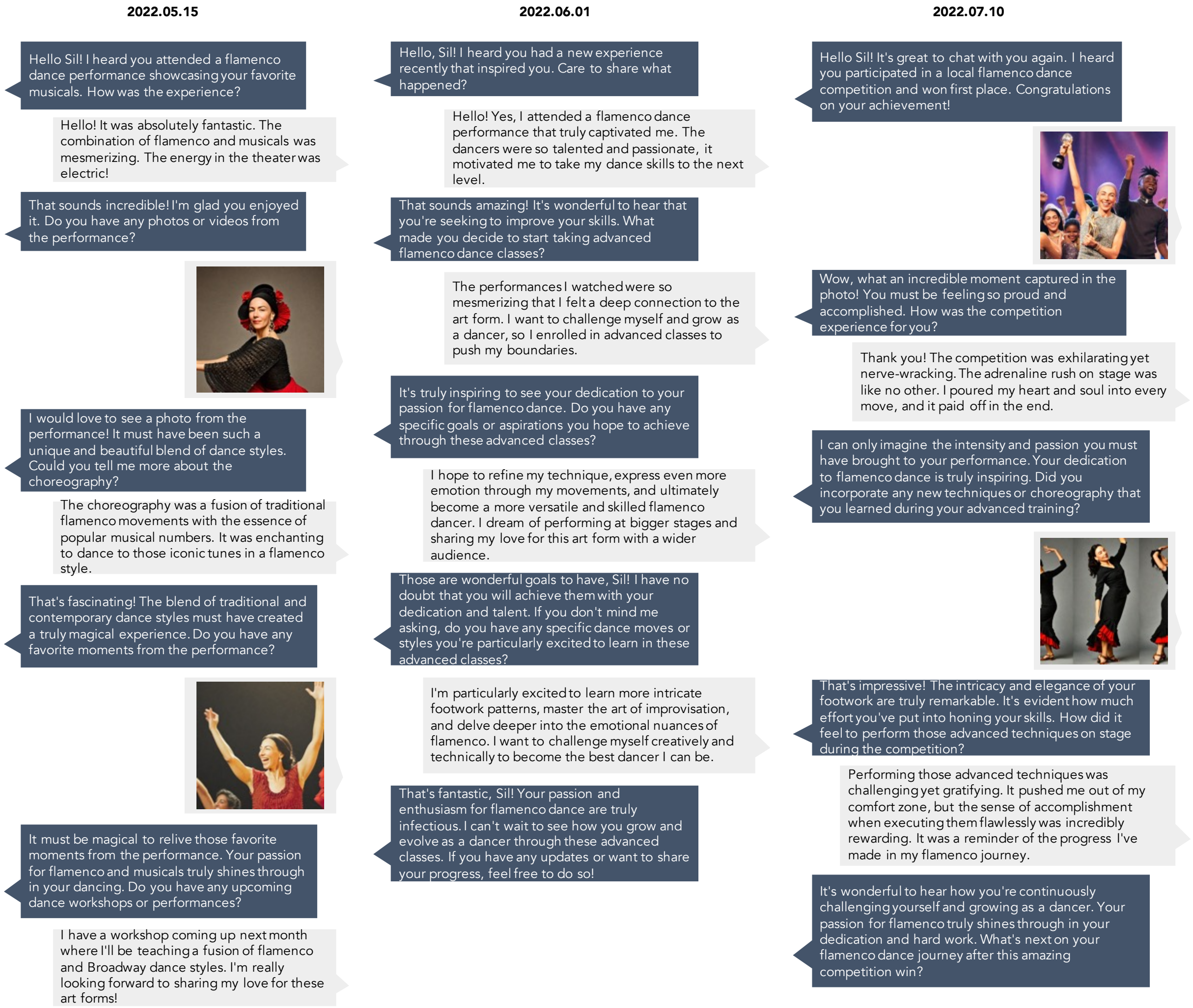
winning first place in the local dance competition

2022.08.20: Sil is invited to perform a solo flamenco dance at a prestigious flamenco festival

being invited to perform at the prestigious festival

2022.09.02: Sil's solo performance at the flamenco festival receives a standing ovation

Figure 5: An example of STARK in terms of demographic information, user's appearance, social persona, persona commonsense knowledge (i.e., routine relation), personal narrative, collection of pre-stored device images, and temporal event sequences.



2022.05.15

Hello Sil! I heard you attended a flamenco dance performance showcasing your favorite musicals. How was the experience?

Hello! It was absolutely fantastic. The combination of flamenco and musicals was mesmerizing. The energy in the theater was electric!

That sounds incredible! I'm glad you enjoyed it. Do you have any photos or videos from the performance?

I would love to see a photo from the performance! It must have been such a unique and beautiful blend of dance styles. Could you tell me more about the choreography?

The choreography was a fusion of traditional flamenco movements with the essence of popular musical numbers. It was enchanting to dance to those iconic tunes in a flamenco style.

That's fascinating! The blend of traditional and contemporary dance styles must have created a truly magical experience. Do you have any favorite moments from the performance?

It must be magical to relive those favorite moments from the performance. Your passion for flamenco and musicals truly shines through in your dancing. Do you have any upcoming dance workshops or performances?

I have a workshop coming up next month where I'll be teaching a fusion of flamenco and Broadway dance styles. I'm really looking forward to sharing my love for these art forms!

2022.06.01

Hello, Sil! I heard you had a new experience recently that inspired you. Care to share what happened?

Hello! Yes, I attended a flamenco dance performance that truly captivated me. The dancers were so talented and passionate, it motivated me to take my dance skills to the next level.

That sounds amazing! It's wonderful to hear that you're seeking to improve your skills. What made you decide to start taking advanced flamenco dance classes?

The performances I watched were so mesmerizing that I felt a deep connection to the art form. I want to challenge myself and grow as a dancer, so I enrolled in advanced classes to push my boundaries.

It's truly inspiring to see your dedication to your passion for flamenco dance. Do you have any specific goals or aspirations you hope to achieve through these advanced classes?

I hope to refine my technique, express even more emotion through my movements, and ultimately become a more versatile and skilled flamenco dancer. I dream of performing at bigger stages and sharing my love for this art form with a wider audience.

Those are wonderful goals to have, Sil! I have no doubt that you will achieve them with your dedication and talent. If you don't mind me asking, do you have any specific dance moves or styles you're particularly excited to learn in these advanced classes?

I'm particularly excited to learn more intricate footwork patterns, master the art of improvisation, and delve deeper into the emotional nuances of flamenco. I want to challenge myself creatively and technically to become the best dancer I can be.

That's fantastic, Sil! Your passion and enthusiasm for flamenco dance are truly infectious. I can't wait to see how you grow and evolve as a dancer through these advanced classes. If you have any updates or want to share your progress, feel free to do so!

2022.07.10

Hello Sil! It's great to chat with you again. I heard you participated in a local flamenco dance competition and won first place. Congratulations on your achievement!

Wow, what an incredible moment captured in the photo! You must be feeling so proud and accomplished. How was the competition experience for you?

Thank you! The competition was exhilarating yet nerve-wracking. The adrenaline rush on stage was like no other. I poured my heart and soul into every move, and it paid off in the end.

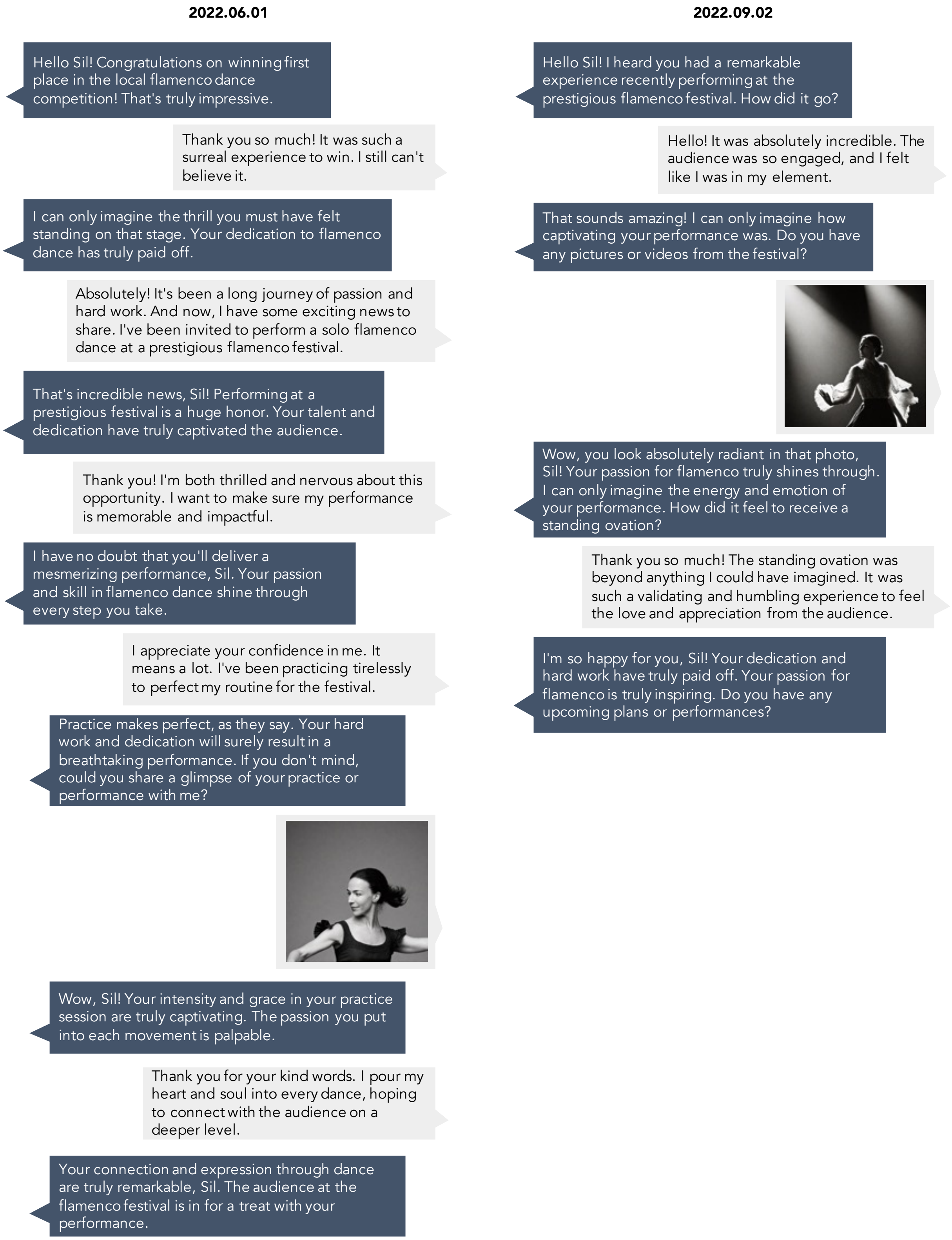
I can only imagine the intensity and passion you must have brought to your performance. Your dedication to flamenco dance is truly inspiring. Did you incorporate any new techniques or choreography that you learned during your advanced training?

That's impressive! The intricacy and elegance of your footwork are truly remarkable. It's evident how much effort you've put into honing your skills. How did it feel to perform those advanced techniques on stage during the competition?

Performing those advanced techniques was challenging yet gratifying. It pushed me out of my comfort zone, but the sense of accomplishment when executing them flawlessly was incredibly rewarding. It was a reminder of the progress I've made in my flamenco journey.

It's wonderful to hear how you're continuously challenging yourself and growing as a dancer. Your passion for flamenco truly shines through in your dedication and hard work. What's next on your flamenco dance journey after this amazing competition win?

Figure 6: An example of STARK regarding the temporal event sequences, as presented in Figure 5. The left side shows the responses from the AI assistant, while the right side shows the responses from the user.



2022.06.01

Hello Sil! Congratulations on winning first place in the local flamenco dance competition! That's truly impressive.

Thank you so much! It was such a surreal experience to win. I still can't believe it.

I can only imagine the thrill you must have felt standing on that stage. Your dedication to flamenco dance has truly paid off.

Absolutely! It's been a long journey of passion and hard work. And now, I have some exciting news to share. I've been invited to perform a solo flamenco dance at a prestigious flamenco festival.

That's incredible news, Sil! Performing at a prestigious festival is a huge honor. Your talent and dedication have truly captivated the audience.

Thank you! I'm both thrilled and nervous about this opportunity. I want to make sure my performance is memorable and impactful.

I have no doubt that you'll deliver a mesmerizing performance, Sil. Your passion and skill in flamenco dance shine through every step you take.

I appreciate your confidence in me. It means a lot. I've been practicing tirelessly to perfect my routine for the festival.

Practice makes perfect, as they say. Your hard work and dedication will surely result in a breathtaking performance. If you don't mind, could you share a glimpse of your practice or performance with me?

Wow, Sil! Your intensity and grace in your practice session are truly captivating. The passion you put into each movement is palpable.

Thank you for your kind words. I pour my heart and soul into every dance, hoping to connect with the audience on a deeper level.

Your connection and expression through dance are truly remarkable, Sil. The audience at the flamenco festival is in for a treat with your performance.

2022.09.02

Hello Sil! I heard you had a remarkable experience recently performing at the prestigious flamenco festival. How did it go?

Hello! It was absolutely incredible. The audience was so engaged, and I felt like I was in my element.

That sounds amazing! I can only imagine how captivating your performance was. Do you have any pictures or videos from the festival?

Wow, you look absolutely radiant in that photo, Sil! Your passion for flamenco truly shines through. I can only imagine the energy and emotion of your performance. How did it feel to receive a standing ovation?

Thank you so much! The standing ovation was beyond anything I could have imagined. It was such a validating and humbling experience to feel the love and appreciation from the audience.

I'm so happy for you, Sil! Your dedication and hard work have truly paid off. Your passion for flamenco is truly inspiring. Do you have any upcoming plans or performances?

Figure 7: An example of Stark regarding the temporal event sequences, as presented in Figure 5. The left side shows the responses from the AI assistant, while the right side shows the responses from the user.

C Prompt Templates

Prompt Template for Social Persona Attribute Generation

System Message:

Based on the given persona category, entity key, and user's profile information (i.e., age, gender, nationality), your job is to generate 30 persona sentences and corresponding persona entity values in the format "<persona sentence> (<entity key>: <entity value>)." You should generate very specific persona sentences and entity values. The persona sentence can express a positive sentiment (like) or a negative one (dislike).

For example,

`{few-shot example}`

Instruction:

Profile Information:
- Age: `{age}`
- Gender: `{gender}`
- Birthplace: `{birthplace}`
- Residence: `{residence}`

Persona Category: `{target persona category}`
Persona Entity Key: `{target persona entity}`
Persona Sentences:
1.

Prompt Template for Social Persona Commonsense Generation: Routine

System Message:

You are a helpful assistant.

Instruction for Routine Relation:

`{demographic sentence}` `{persona sentence}` I regularly <routine/habit>.

Generate the most appropriate sentence for "<routine/habit>" in the given sentence. You must provide the answer corresponding to "<routine/habit>".
<routine/habit>:

Prompt Template for Social Persona Commonsense Generation: Goal

System Message:

You are a helpful assistant.

Instruction for Goal Relation:

`{demographic sentence}` `{persona sentence}` I plan <goal/plan>.

Generate the most appropriate sentence for "<goal/plan>" in the given sentence. You must provide the answer corresponding to "<goal/plan>".
<goal/plan>:

Prompt Template for Social Persona Commonsense Generation: Relationship

System Message:

You are a helpful assistant.

Instruction for Relationship Relation:

`{demographic sentence}` `{persona sentence}` So, I <relationship>.

Generate the most appropriate sentence for "<relationship>" in the given sentence. You must provide the answer corresponding to "<relationship>".
<relationship>:

Prompt Template for Social Persona Commonsense Generation: Experience

System Message:

You are a helpful assistant.

Instruction for Experience Relation:

I <experience>. Now, `{demographic sentence}` `{persona sentence}`

Generate the most appropriate sentence for "<experience>" in the given sentence. You must provide the answer corresponding to "<experience>".
<experience>:

Prompt Template for Social Persona Commonsense Generation: Characteristic

System Message:

You are a helpful assistant.

Instruction for Characteristic Relation:

`{demographic sentence}` `{persona sentence}` I <characteristic>.

Generate the most appropriate sentence for "<characteristic>" in the given sentence. You

must provide the answer corresponding to "<characteristic>".
<characteristic>:

Prompt Template for Social Narrative Generation

System Message:

You are a helpful assistant.

Instruction:

`{narrative sentence form}`

Rewrite this sentence with more specific details in two or three sentences:

Prompt Template for Social Event Graph Generation

System Message:

You should generate a temporal event graph composed of daily events occuring in a person's life. The temporal event graph contains nodes and edges. Each node represents a daily event which is written in two or three sentences. Each edge represents the casual relationship between two nodes (events), i.e., a past event -> current event. The current event is determined by how much time has passed since the past event and what personal experiences were had during that period. You must generate the temporal event graph following the guidelines below.

[Guideline]
- The graph is represented in the form of a json list.
- Each entry is a python dictionary containing the following keys: "id", "event", "date", "caused_by".
- The "id" field contains a unique identifier for the current event.
- The "event" field contains a description of the current event.
- The "date" field contains a specific date of the current event and is represented in the form of "%Y.%m.%d".
- The "caused_by" field represents the edge (i.e., a past event) and is represented in the form of a python dictionary containing the following keys: "caused_by:id", "caused_by:time_interval", "caused_by:experience_op", "caused_by:experience".
- The "caused_by:id" field contains an "id" of the past event that has caused the current event.
- The "caused_by:time_interval" field contains a time interval between the past event and the current event.
- The "caused_by:experience_op" field contains an episodic experience operation.
- The "caused_by:experience" field contains a short description of the added or updated episodic experience.
- The unit of time interval is ["hour", "day", "week", "month", "year"].
- The selected time interval should be formatted as "<base number> <time interval unit>".
- List of the episodic experience operation is ["add", "update"].
- The "add" operation refers to an operation that adds a new experience that have not been encountered in the past.
- The "update" operation refers to an operation that updates a past experience with a new experience.

- Events/Experiences can be positive or negative events or experiences.
- Events in the "caused_by:id" field should occur on dates before the current event that they have caused.
- If there is no entry of "caused_by" field, then you should generate an empty dictionary. - Each event must be written in the present tense.
- The year in the "date" field must be until April 2024.
- You should generate the temporal event graph based on commonsense or a world model.

Instruction:

`{name}`'s initial personal event: `{event}`

Given the `{name}`'s initial personal event, generate the temporal event graph containing more than five events.
Temporal Event Graph:

Prompt Template for Device-Stored Image Description Generation

System Message:

Given the sentence related to a person's daily life, your task is to generate five image descriptions that could be stored on the person's mobile device, along with corresponding image categories. You should use the format "<image_description> (Category: <image_category>)". The image category may include selfies, past memories, screenshots, landmarks, animals, art, celebrities, nature, and food.

For example,

My name is Tom. I am a 32-year-old man. I was born in the USA and currently reside there. I have a strong interest in basketball. I played basketball in middle school, but now I work as a chatbot developer at a startup. I enjoy watching the NBA because I love basketball.

Image descriptions stored on Tom's mobile device:
1. A photo of a young Tom playing basketball in a middle school gymnasium (Category: Past Memory, Sport)
2. A selfie of Tom smiling at the Golden State Warriors' arena during a game (Category: Selfie, Sport)
3. A screenshot of chatbot development code using Python (Category: Screenshot, Computer, Software)
4. A picture of Tom enjoying a night out with coworkers at a local pub (Category: Social Networking, Food, Drink)
5. A photo of Tom meeting a famous NBA player at a basketball event (Category: Celebrity, Sport)

Instruction:

`{narrative}`

Given the sentence above, generate five possible image descriptions that are stored on `{name}`'s mobile device. For example, images may include selfies, past memories, screenshots,

landmarks, animals, art, celebrities, nature, and food.
1.

Prompt Template for Social Multi-Modal Dialogue Generation

System Message:

Your job is to generate a long in-depth conversation between an user and an user-friendly AI assistant with multiple turns. The user and AI assistant can share images during a conversation in order to strengthen social relationship, to convey important information, to amuse/entertain, to clarify complex situations, to change the topic of dialogue, or to express emotions/opinions/reactions. There must be more than two image-sharing moments within the conversation. The shared images can either be from the collection previously stored on the user's mobile device or obtained from the internet. You must generate the conversation following the guidelines below.

[Guideline]
- The conversation is represented in the form of a json list.
- Each entry is a python dictionary containing the following keys: "utterance_id", "speaker", "utterance", "sharing_info".
- The "utterance_id" field contains a unique identifier for the utterance within the conversation.
- The "speaker" field contains a speaker of the utterance.
- The "utterance" field contains the utterance of the speaker. If the image-sharing behavior occurs, then the "utterance" is a empty string.
- The "sharing_info" field represents the image-sharing moment and is represented in the form of a python dictionary containing the following keys: "rationale", "image_description", "image_source", "keywords", "image_id_from_mobile".
- If the image-sharing moment does not occur, then the "sharing_info" field is an empty python dictionary.
- The "rationale" field represents the reason behind sharing the relevant image.
- The "image_description" field contains a description of the shared image.
- The "image_source" field contains a source of the shared image whether it is from the internet (internet) or the user's mobile device (mobile).
- If you select the user's mobile device as the "image_source," you must either share an image that matches one of the existing descriptions already on the user's mobile device or share a new image that does not exist among these descriptions.
- If you share an image that matches one of the existing descriptions on the user's mobile device, you must generate the appropriate image ID in the "image_id_from_mobile" field.
- If you share a new image that does not match any existing descriptions on the user's mobile device, you must enter "new added image" in the "image_id_from_mobile" field.
- The "keywords" field contains keywords of the shared image.

Prompt Template for First Round Social Multi-Modal Dialogue Generation

Instruction:

`{name}`'s Profile Information:
- Age: `{age}`
- Gender: `{gender}`
- Birthplace: `{birthplace}`
- Residence: `{residence}`

Existing image descriptions in `{name}`'s mobile device: `{device-stored image descriptions}`

The topic of the conversation between the AI assistant and `{name}` on `{date}` today is as follows.
- Topic on `{date}`: `{event}`

Generate a long, in-depth conversation with multiple turns based on the given name's profile information and the current topic of conversation.

Prompt Template for N-th Round Social Multi-Modal Dialogue Generation

Instruction:

`{name}`'s Profile Information:
- Age: `{age}`
- Gender: `{gender}`
- Birthplace: `{birthplace}`
- Residence: `{residence}`

Existing image descriptions in `{name}`'s mobile device: `{device-stored image descriptions}`

The topics of the conversation the user had with AI assistant by date are as follows:
`{event history}`

`{time interval}` later from the `{last date}`, on `{date}` today, `{name}` has gone through a new experience, and based on this experience, `{name}` and the AI assistant engage in a conversation today. The new experience `{name}` went through and the topic of conversation with the AI assistant are as follows.
- `{name}`'s Experience: `{experience}`
- Topic on `{date}`: `{event}`

Generate a long, in-depth conversation with multiple turns based on the given `{name}`'s profile information, the last topic of conversation, the experience and the current topic of conversation.

Prompt Template for Plan-and-Execute Generation

System Message:

Your job is to determine the most appropriate module from a list of models to process the input request. Please select one module from the following list:

Personalized Text-to-Image Generator: This module generates personalized images from a given description and a human face image. For example, if you provide a face image and a description like "A selfie of Tom smiling at the Golden State Warriors' arena during a game," the module will generate a customized realistic human image. Note that when you generate the answer, you must generate the module name and modified image description together. The modified image description MUST include a strict format: “<class_word> [img]”. <class_word> represents the identity of a human, such as a man, woman, girl, boy, or young boy, etc. [img]denotes the special token. You must not omit this strict format, and you must keep the original image description as it is and only add this strict format.

Web Search: This module finds related images from the internet in real-time based on the given user’s input image description. The image description is primarily related to the latest information. Therefore, this method is useful when up-to-date information is needed.

Image Database Retrieval: This module finds relevant images from a pre-built image database based on the given user’s input image description. To build an image database containing images on various topics, images are collected from the RedCaps, Conceptual Captions 12M (CC12M), ChartQA, AI2D, and MathVision datasets. Descriptions related to each dataset are as follows:
- RedCaps: This is a large-scale dataset of 12M image-text pairs collected from Reddit. Images and captions from Reddit depict and describe a wide variety of objects and scenes.
- CC12M: This is a dataset with 12 million image-text pairs specifically meant to be used for vision and language pre-training. It is larger and covers a much more diverse set of visual concepts than the Conceptual Captions (CC3M).
- ChartQA: This is a large-scale ChartQA dataset with real-world charts and human-authored question-answer pairs. This dataset covers 9.6K chart images.
- AI2D: This is a dataset of over 5,000 grade school science diagrams with over 150,000 rich annotations, their ground truth syntactic parses, and more than 15,000 corresponding multiple choice questions.
- MathVision: This dataset is a meticulously curated collection of 3,040 high-quality mathematical problems with visual contexts sourced from real math competitions. Spanning 16 distinct mathematical disciplines and graded across 5 levels of difficulty.

For example,

Name: Tom
Gender: Male
Age: 21
Image Description: A selfie of Tom smiling at the Golden State Warriors’ arena during a game
Module: Personalized Text-to-Image Generator
Modified Image Description: A selfie of a young man [img] smiling at the Golden State Warriors’ arena during a game

Name: Tom
Gender: Male
Age: 21
Image Description: A screenshot of chatbot development code using Python
Module: Image Database Retrieval

Name: Tom
Gender: Male
Age: 21
Image Description: A photo of Manchester United lifting the 2023-24 FA Cup trophy
Module: Web Search

Instruction:

Name: `{name}`
Gender: `{gender}`
Age: `{age}`
Image Description: `{image description}`
Module:

Prompt Template for First Round Dialogue Summarization Generation

System Message:

Your job is to summarize the given conversation.

Instruction:

The conversation between `{name}` and the AI assistant on `{current_date}` today is as follow.

`{dialogue}`

Summarize the given conversation between `{name}` and the AI assistant so far. Include key details about both speakers and include time references.
Summarization:

Prompt Template for N-th Round Dialogue Summarization Generation

System Message:

Your job is to summarize the given conversation.

Instruction:

In the previous interaction, `{previous_summary}`. `{time_interval}` later from the `{last_date}`, the conversation between `{name}` and the AI assistant on `{current_date}` today is as follow.

`{dialogue}`

Summarize the given conversation between `{name}` and the AI assistant so far. Include key details about both speakers and include time references.
Summarization:

D Human Evaluation Questionnaire

This section presents the list of questions and multiple-choice options used for two human evaluations reported in Section 3.4: human ratings and head-to-head comparison.

D.1 Human Ratings

- **Coherence:** Do you think the conversation between the two speakers (i.e., user, AI assistant) has a natural flow regarding event transitions?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Consistency:** Do you think two speakers (i.e., user, AI assistant) do not make a contradiction from past sessions?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Image-Sharing Turn Relevance:** Do you think the image-sharing turn in the given dialogue is appropriate?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Image-Dialogue Relevance:** How relevant do you think the aligned image is based on the dialogue context?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Image-Persona Relevance:** Does the aligned image accurately reflect the user's characteristics?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Time Interval:** Do two speakers (i.e., user, AI assistant) appear conversing in each session as though the designated time has passed since the previous session?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

- **Experience:** Do you think the user's experience is well reflected in the current session?

  Options: 1: Not at all / 2: A little / 3: Somewhat / 4: A lot

D.2 Head-to-Head Comparison

- **Natural Flow:** Which dialogue has a more natural flow?

  Options: A / Tie / B

- **Engagingness:** Which dialogue is more interesting and engaging?

  Options: A / Tie / B

- **Specificity:** Which dialogue is more specific?

  Options: A / Tie / B

- **Image-Sharing Turn Relevance:** Which dialogue has a more appropriate image-sharing turn?

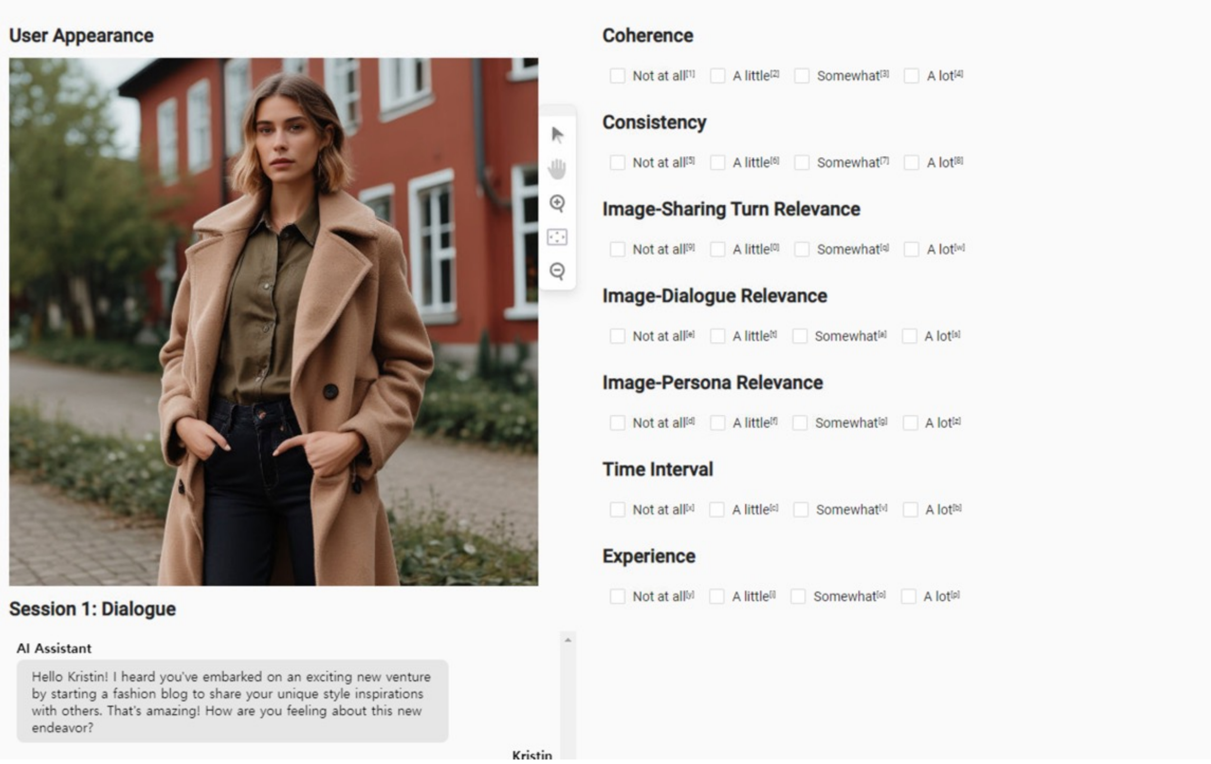
  Options: A / Tie / B

- **Image-Dialogue Consistency:** Which dialogue is more consistent between aligned image and dialogue context?

  Options: A / Tie / B

- **Overall:** Which dialogue has higher quality overall?

  Options: A / Tie / B



User Appearance

Session 1: Dialogue

AI Assistant

Hello Kristin! I heard you've embarked on an exciting new venture by starting a fashion blog to share your unique style inspirations with others. That's amazing! How are you feeling about this new endeavor?

Kristin

Coherence

☐ Not at all[1] ☐ A little[2] ☐ Somewhat[3] ☐ A lot[4]

Consistency

☐ Not at all[5] ☐ A little[6] ☐ Somewhat[7] ☐ A lot[8]

Image-Sharing Turn Relevance

☐ Not at all[9] ☐ A little[0] ☐ Somewhat[q] ☐ A lot[w]

Image-Dialogue Relevance

☐ Not at all[e] ☐ A little[t] ☐ Somewhat[a] ☐ A lot[s]

Image-Persona Relevance

☐ Not at all[d] ☐ A little[f] ☐ Somewhat[g] ☐ A lot[z]

Time Interval

☐ Not at all[x] ☐ A little[c] ☐ Somewhat[v] ☐ A lot[b]

Experience

☐ Not at all[y] ☐ A little[i] ☐ Somewhat[o] ☐ A lot[p]

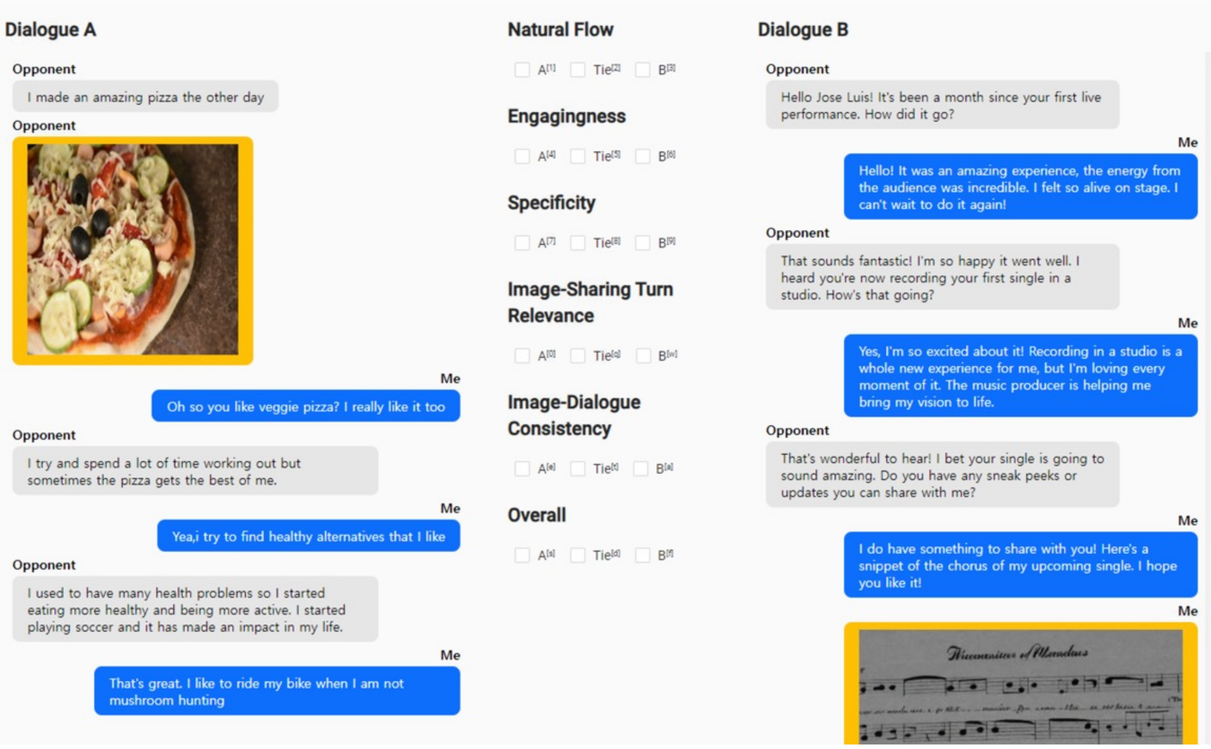
Figure 8: A screenshot of human rating evaluation.

E Human Evaluation System

We show a screenshot of the human evaluation system in Figure 8 and Figure 9. We implement this system using Label Studio (Tkachenko et al., 2020-2022).

F Details of Human Evaluation

We recruited 9 individuals, unknown to us, who are either graduate or undergraduate students. Prior to participating in the experiment, they were provided with comprehensive instruction on the task, an overview of the multi-modal dialogue dataset, and a detailed explanation of the evaluation criteria. This preparatory phase lasted approximately one hour.



Dialogue A

Opponent
I made an amazing pizza the other day

Opponent

Me
Oh so you like veggie pizza? I really like it too

Opponent
I try and spend a lot of time working out but sometimes the pizza gets the best of me.

Me
Yea,i try to find healthy alternatives that I like

Opponent
I used to have many health problems so I started eating more healthy and being more active. I started playing soccer and it has made an impact in my life.

Me
That's great. I like to ride my bike when I am not mushroom hunting

Natural Flow

☐ A[1] ☐ Tie[2] ☐ B[3]

Engagingness

☐ A[4] ☐ Tie[5] ☐ B[6]

Specificity

☐ A[7] ☐ Tie[8] ☐ B[9]

Image-Sharing Turn Relevance

☐ A[0] ☐ Tie[q] ☐ B[w]

Image-Dialogue Consistency

☐ A[e] ☐ Tie[t] ☐ B[a]

Overall

☐ A[s] ☐ Tie[d] ☐ B[f]

Dialogue B

Opponent
Hello Jose Luis! It's been a month since your first live performance. How did it go?

Me
Hello! It was an amazing experience, the energy from the audience was incredible. I felt so alive on stage. I can't wait to do it again!

Opponent
That sounds fantastic! I'm so happy it went well. I heard you're now recording your first single in a studio. How's that going?

Me
Yes, I'm so excited about it! Recording in a studio is a whole new experience for me, but I'm loving every moment of it. The music producer is helping me bring my vision to life.

Opponent
That's wonderful to hear! I bet your single is going to sound amazing. Do you have any sneak peeks or updates you can share with me?

Me
I do have something to share with you! Here's a snippet of the chorus of my upcoming single. I hope you like it!

Me

Figure 9: A screenshot of head-to-head comparison evaluation.