Multimodal Conversational Interactions for Facial Composites: A Case for Stateful Prompt Orchestration

Rico Städeli and **Roman Leu** and **Jasmin Heierli** and **Max Meisterhans** and **Elena Gavagnin** and **Alexandre de Spindler**
Zurich University of Applied Sciences, Winterthur, Switzerland
{staedric,leurom01}@students.zhaw.ch, {heej,meix,gava,desa}@zhaw.ch

Abstract

The advent of advanced language models has raised expectations for conversational interactions with information systems, but effectively controlling these models to achieve nuanced conversational behavior remains a challenge. This paper introduces PROMISE, a framework that uses state machine modeling to enable dynamic prompt orchestration to facilitate complex interactions. We illustrate PROMISE's application using a multimodal dialogue scenario designed to assist witnesses in generating facial composites of suspects. Our demonstration validates the framework's feasibility and its utility to enable meaningful and complex conversational interactions.

1 Introduction

While the development of increasingly powerful language models (LM) raises expectations for more sophisticated conversational behaviours, directing LM behaviour remains challenging (Hadi et al., 2023). There is therefore a growing need to investigate the feasibility and utility of LM applications.

The capabilities of LMs in extracting structured information from text and generating text from structured information are key enablers for designing and specifying conversational interactions. For example, the extraction of user intents and associated parameters, and the generation of responses from intent-specific, structured data, have long been fundamental in developing conversation flows for assistive conversational agents.

While this dialog management approach is effective for answering questions and executing simple commands, existing platforms such as Google DialogFlow and IBM Watson faced limitations when developing more complex interactions. In DialogFlow, the intertwining of numerous contexts and intents can make it difficult to avoid inadequate intent matches. In Watson, the multitude of logical conditions attached to each step can lead to an extended, often unwieldy tree of conversation possibilities. As the complexity of the dialogs increases, the adaptability of both platforms decreases, making it more difficult to integrate new requirements without great effort.

Consequently, the advent of LMs not only raises the question of how they can be utilized to support designing conversational flows, but also how to harness their advanced capabilities for more complex conversational interactions. Training an LM from scratch to serve a specific purpose is resource-intensive and often impractical for typical development projects. Although fine-tuning can tailor LM responses, it also demands meticulous data preparation, making fast, iterative experimentation difficult. Ultimately, neither approach fully addresses the inherent challenges arising when complex interactions are designed, implemented in variants, and improved iteratively.

Leveraging LMs' zero- and few-shot learning abilities, more efficient approaches for LM control were developed which are commonly referred to as *prompt engineering* (Korzynski et al., 2023; White et al., 2023). While many prompt articulation strategies have been developed (Wei et al., 2022; Fernando et al., 2023; Chu et al., 2023), this alone cannot ensure consistent LM behaviour in complex interactions. Overly detailed prompts that cover the entire interaction may lead to confusion in sequences or levels of partial interactions. Conversely, overly broad prompts risk missing expected responses, may induce erroneous responses, and introduce vulnerabilities (Mozes et al., 2023).

To mitigate the challenges met when using prompts for complex interactions, Wu et al. 2023 proposed a framework (PROMISE) that follows the idea of segmenting complex tasks into sequences of simpler tasks. This was shown to enhance control and predictability while harnessing LM capabilities (Helland et al., 2023). Complex prompts are thus broken down into separate, more specific

prompts, increasing the predictability of LM behaviour while leveraging conversational skills using existing prompting techniques.

PROMISE uses concepts of state machine modeling as a means to orchestrate partial prompts. Conversational behaviour in interaction states, triggers and guards of state transitions, and actions performed when transitioning, can all be implemented with prompts.

In this paper we present a use case for complex and multimodal conversational interactions involving image generation. The idea is to assist witnesses of a crime in recalling and describing the appearance of a person they encountered. Such descriptions are then used to generate visual representations of that person, potentially aiding in the identification of suspects.

In the following Sect. 2, we highlight the challenges of this application domain that necessitate a multimodal approach. We then identify specific requirements for conversational behaviour in Sect. 3. Section 4 details how the PROMISE framework facilitates the realisation of these requirements. Our validation is twofold: Section 5 focuses on the demonstration of the framework-enabled feasibility of the witness assistance application, which serves as a proof of concept. Following this, Sect. 6 shifts focus to the utility of the application, showcasing its effectiveness in facilitating complex conversational interactions. We conclude with final thoughts in Sect. 7.

2 Background

Conversations aimed at eliciting specific information can be considered examples of complex interactions. These usually require an actively managed dialog that dynamically guides the user to obtain the necessary information. An example of this type of interaction is the creation of facial composite images, for which detailed descriptions of a person's appearance must be captured. In such cases, specific characteristics of a person must be recorded during a conversation, whereby the conversation may deviate from the course or contain irrelevant details. As the witness is confronted with resulting images, the description obtained so far may have to be supplemented or corrected.

At present, the collaborative elaboration of facial composites is a highly intricate process involving various technical and psychological aspects. It faces challenges at each stage of translating a text-based description based on human memory into a visual representation of a person (Wells and Hasel, 2007). Although the problem is inherently multimodal and requires the interplay of textual and visual components, current research and efforts to make progress have largely focused on these two elements separately, resulting in unsatisfactory solutions from both visual and linguistic perspectives (Jalal et al., 2023).

From a visual standpoint, the process currently involves either a forensic artist creating a sketch, or an eyewitness iteratively composing a face using specialized software. In both cases, the process is suboptimal as it often leads to a static and unrealistic representation of the suspect (Jalal et al., 2023). Recently, generative deep learning has been introduced, primarily focusing on translating sketches into photographs, predominantly employing generative adversarial networks such as Pix2Pix and CycleGans (Wang et al., 2018a,b; Zhu et al., 2017).

From a linguistic standpoint, the challenge becomes more significant because facial descriptions are frequently affected by noisy information from eyewitnesses. As a matter of fact, sketch-based facial recognition relies primarily on a static textual description given by the witness, which could be inaccurate in the first place and, secondly, does not offer any confidence estimate for each provided feature. In this respect, much linguistic research in the field has concentrated on developing hierarchical analytical methods and frameworks that leverage linguistic theory such as part-of-speech and attribute ontology to effectively extract relevant facial attributes from given descriptions (Karczmarek et al., 2017; Khan and Jalal, 2020).

While both types of approaches address relevant challenges, neither of them seeks to explore how their combination can be used to better assist witnesses in recalling information. In the following, we therefore propose an innovative multimodal conversational interaction that dynamically iterates through linguistic utterances and visual representations to help witnesses recall relevant information.

3 Requirements

We now identify requirements to the conversational behaviour of an assistant supporting witnesses of a crime in recalling and describing the appearance of a person they encountered.

The example conversation in Fig. 1 showcases how descriptive information can be elicited from

a witness. The assistant (light and dark green) uses open-ended questions and empathetic dialogue to create a supportive environment. In a first phase (light green), the witness is guided to provide enough descriptions to generate an initial image, which is then presented (indicated in 1st squared brackets). This initiates a second phase (dark green), where the witness is asked for feedback on the image presented. A new image is then generated based on this feedback, and presented again (indicated in 2nd squared brackets). While this second phase may be repeated, the interaction shall conclude when the witness cannot recall any further information. (light green at the end).

This interaction comprises three phases, which must transition into one another if certain conditions are met. The starting phase needs to conclude when the witness has provided enough descriptive information for an initial image to be generated. The image is generated as part of the transition and presented to the witness to initiate the second phase. This second phase can be left for two reasons. One reason is that the witness provides additional, corrective information, which triggers the generation of a new image. When this new image is presented to the witness, the interaction transitions back to the second phase. The other reason is that the witness no longer suggests any changes. At this point, the interaction transitions to a final phase, in which the witness is thanked and bid goodbye.

To support the generation of an image, transitions need to include an extraction of all descriptive information provided throughout a defined segment of the conversation. For example, the following JSON object should result from such an extraction out of the starting phase of the conversation in Fig. 1.

```
{"Person": {
    "gender": "Female",
    "height": "5'6\"",
    "build": "Medium",
    "demeanor": "Relaxed, cheerful",
    "hair": {
        "color": "Light brown",
        "style": "Long, loose, wavy"
    },
    "eyes": "Light-colored, possibly
       blue or green",
    "facialFeatures": "Friendly face
       with a big smile",
    "distinctiveMarks": "None"
 },
 "Environment": {
    "Location": "City, downtown area",
    "Time": "Day"
}}
```

Accordingly, the following JSON object should be extracted from the second phase.

```
{"hairColor": "caramel brown",
 "eyeColor": "clear blue",
 "earrings": "none"}
```

While the utterances produced by the assistant shown in Fig. 1 demonstrate the benefit of using an LM to generate them, several challenges arise when controlling the LM with a single prompt: Consistently distinguishing different interaction phases, appropriately transitioning between these phases according to specified characteristics of the conversation, comprehensively extracting information from specific conversation segments, all the while promptly following user requests to end the interaction at any time. We are therefore presenting the use of PROMISE as a means of overcoming these challenges while enabling the beneficial use of LMs in the following Sect. 4.

4 Implementation

With PROMISE, the conversational interaction exemplified in Fig. 1 is modelled by a state machine such as in Fig. 2. The state machine consists of an outer state in light gray, an initial node to its left, and a final node to its right. Transitions lead from the initial node to the outer state, and from the outer state to the final node. Transitions depend on triggers and guards, and actions are executed when followed. The outer state contains a sequence of inner states which also starts with an initial node and has its own final node. In between, there are two states shown in green and blue, each specifying the conversational behaviour in the first two interaction phases identified in the previous Sect. 3. The third phase thanking the user and bidding goodbye is specified by the inner final node.

Using PROMISE, states are annotated with a *state prompt* such as "You are having a conversation with a user who ..." for the outer state, "Start by asking open-ended questions about ..." for the green inner state, and "Continue the conversation to refine ..." for the blue inner state. Such state prompts will be used to control the LM while the interaction is in one of these states. Similarly, transitions are also annotated with prompts, for example as indicated by "Details provided", "[]", and "(Extract details, Generate image)" in the case of the transition between the two inner states. These prompts control the LM when analysing the conversation in terms of transition triggers, guards, and

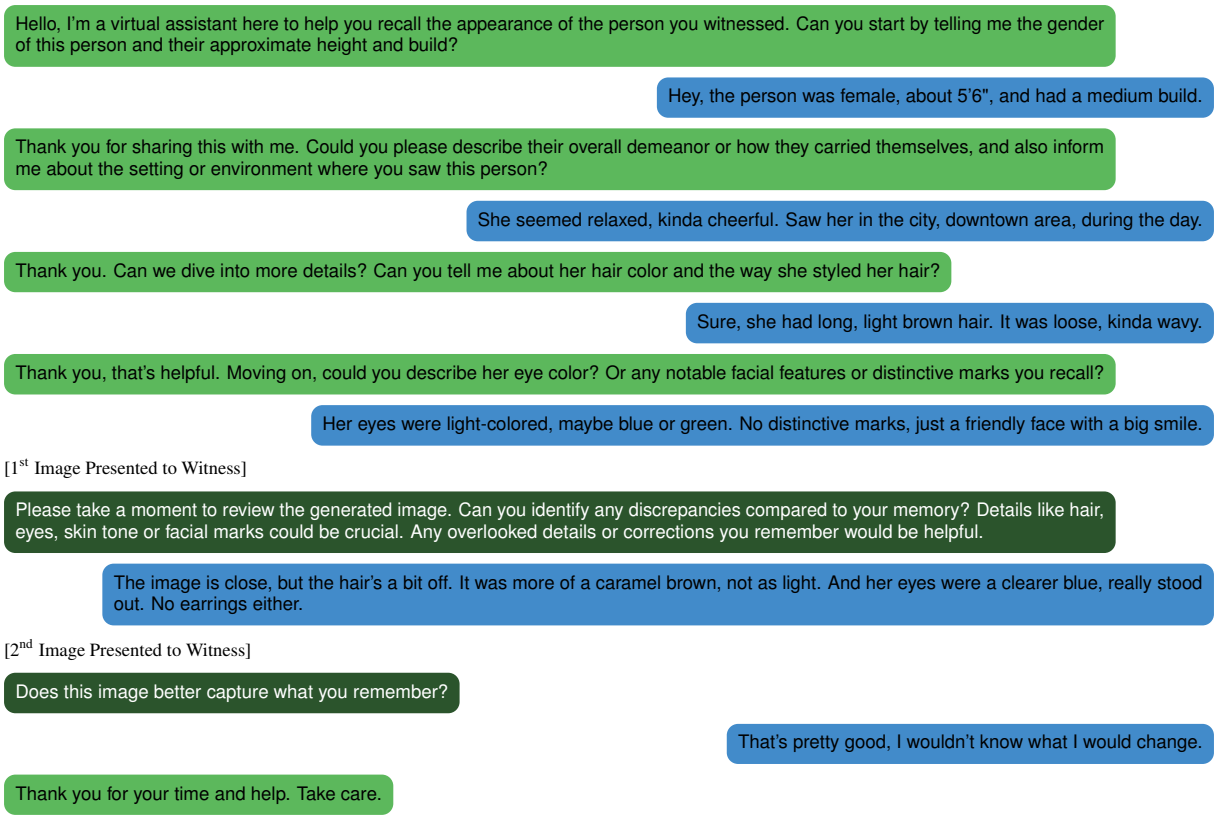

Figure 1: Multi-Phase Elicitation: Assistant (1st: Light Green, 2nd: Dark Green, 3rd: Goodbye) & Witness (Blue)

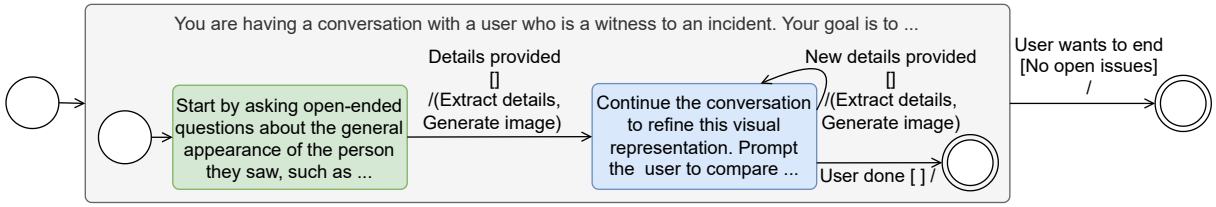

Figure 2: Conversational Interaction Design for Multi-State Elicitation

actions, respectively. As shown next, PROMISE transparently composes more complex prompts by dynamically assembling the prompts associated with states and transitions.

When the interaction is started, the initial node is used to identify the first interaction state. In this case, the initial node points to the outer state, in which the inner initial node points to the green inner state. In this green state, PROMISE will compose a prompt P_c by concatenating the outer state prompt P_{os}

$$P_{os} = \text{"You are having a conversation with ..."}$$

with the inner state prompt P_{is}

$$P_{is} = \text{"Start by asking open-ended ..."}$$

and the *state opening prompt* P_{iso}

$$P_{iso} = \text{"Begin the conversation by ..."}$$

such as to obtain the composed prompt P_c

$$P_c = P_{os} + P_{is} + P_{iso}$$
$$= \text{"You are having ... Start ... Begin the ..."}$$

The state opening prompt, an optional extension of the state prompt, is used when the state is configured to start the conversation, as demonstrated in this example.

The composed prompt P_c is then used to instruct the LM. In the example conversation above, the LM completion returned the utterance "Hello, I'm here to help you ...", which opens the conversation with the witness.

Upon every utterance from the witness, such as the first utterance "Hey, the person was female ...", all outgoing transitions are checked before the LM is used to generate a response to the witness.

First, the list of utterances U_{is}^{t}, which represents the conversation held in this state so far, is extended with the incoming user utterance u_u as follows.

$$\begin{aligned} U_{is}^{t+1} &= U_{is}^{t} + u_u \\ &= [\text{"Hello, ...", "Hey, the person ..."}] \end{aligned}$$

Then, to check a transition, its trigger prompt P_t and guard prompt P_g are used to obtain decisions from the LM about whether the transition should be followed. If it is followed, the action prompt P_a is used to execute the action. The trigger, guard and action prompts are automatically extended with the utterances such as to support decisions based on the conversation so far.

For example, in the case of a transition trigger, the composed prompt

$$\begin{aligned} P_c &= P_t + U_{is}^{t+1} \\ &= \text{"Review the conversation... Determine..."} \\ &+ [\text{"Hello, ...", "Hey, the person ..."}] \end{aligned}$$

is created to let the LM decide whether the conversation so far contains the information required to generate a prompt for the image generation model. While the first witness response mentions a height, build and how the suspect carried themselves, no information about their hair and facial features has been provided so far. Consequently, this transition trigger does not pass, and the interaction stays in the current state.

Multiple decisions may be attached to a single transition. Each decision may contain a prompt for LM-based evaluation, optionally containing placeholders for data injection. Alternatively, decisions may also be specified with code that implements any other evaluation. In our example, a second decision serves as a transition guard, and instructs the LM to decide whether there are no open questions from the witness that should prevent the current interaction from transitioning unexpectedly.

If the interaction stays in the current state, the state prompt and accumulated utterances are included in the newly composed prompt

$$\begin{aligned} P_c &= P_{os} + P_{is} + U_{is}^{t+1} \\ &= \text{"You are having ... Begin the ..."} \\ &+ [\text{"Hello, ...", "Hey, the person ..."}] \end{aligned}$$

which is used to obtain the subsequent response to the witness from the LM. This response is also appended to the state utterances. As seen in the example interaction in Fig.1, the conversation therefore stays in the same state as long as the expected information is incomplete. When all the information is provided, the transition decisions pass, and the conversation transitions to the subsequent state attached to the transition. As shown in Fig. 2, the blue state follows the green state. In this blue state, the LM is controlled as described for the green state, but using the partial prompts associated with the blue state.

Transitions may include multiple actions that contain a prompt or code. In our example, there are two actions. The first is to extract the details provided by the witness. The second action will generate an image based on these details extracted. In both cases, the action is a prompt with which the composed prompt

$$\begin{aligned} P_c &= P_a + U_{is} \\ &= \text{"Review the conversation... Extract..."} \\ &+ [\text{"Hello, ...", "Hey, the person ..."}] \end{aligned}$$

is created and used to instruct the LM such as for extracting details or generating an image. In most cases, the result of an action is an object that is stored in an interaction storage, making it accessible to other states, transition decisions and actions, or surrounding system components.

As opposed to the green state, the blue state has two outgoing transitions. One of them is triggered by the condition "User done", which will be true if the witness has nothing to add to the latest image shown to them. The other transition is triggered by the user providing additional details that can be used to update the image. If this is the case, the additional details will be extracted and used to generate the next image to be shown to the witness as part of this recursive transition.

The outer state simplifies the development of conversational interactions in three ways. First, the outer state prompt P_{os} is transparently prepended to all its inner states. Consequently, developers can avoid redundancies by factoring out common parts of inner state prompts. Second, the outer state may have its own outgoing transitions which enables multilayered interactions. For example, this outer state has a transition triggered by the user wishing to stop the interaction. This trigger decision is tested with each incoming user utterance in all the inner states. As a result, this transition may be triggered at any point in the whole interaction. Third,

the outer state maintains its own list of utterances U_{os} containing all utterances of all its inner states. This enables decisions and actions to be made in the scope of larger conversational contexts.

In summary, PROMISE promotes a separation of concerns when LMs are controlled using prompts. One means of separation results from the support of state-specific prompts and separate, individual prompts for transition decisions and actions. Another means is provided with the ability to factor out recurring prompt parts and reuse them in a common outer state. As a result, it costs less effort to distinguish interaction phases and control their transitions more consistently. The maintenance of state-specific utterances better supports transitions and actions playing out in different interaction segments. Finally, the ability to nest states supports multilayered interactions where different conversation flows are controlled in parallel.

Note that all prompts used in a PROMISE application may feature placeholders in which any text or data may be injected. While this supports the use of prompt engineering for retrieval augmented generation (RAG), a more detailed description of this mechanism is beyond the scope of this paper.

5 Proof-of-Concept Application

This section demonstrates the PROMISE framework's practical applicability by detailing the resulting witness assistant application. Our aim is to showcase the feasibility of bringing nuanced requirements of multithreaded and multimodal conversational flows to reality. The main functionality of the application is outlined in Fig. 3.

A chat frontend used by witnesses is indicated on the left. While the conversation depicted is the one previously shown in Fig. 1, this front-end showcases the idea of iterative questioning, where the witness is repeatedly presented with images generated from the information they provide, and subsequently updated with the additional information they provide when an image is presented. The extractions in JSON format used for the generation of the images are shown on the right.

The conversational interaction was modeled as shown in Fig. 2 and implemented using the PROMISE API. All prompts for the outer and inner states, as well as the transition triggers and actions were generated using ChatGPT with GPT-4 (OpenAI, 2023) using the following meta prompt.

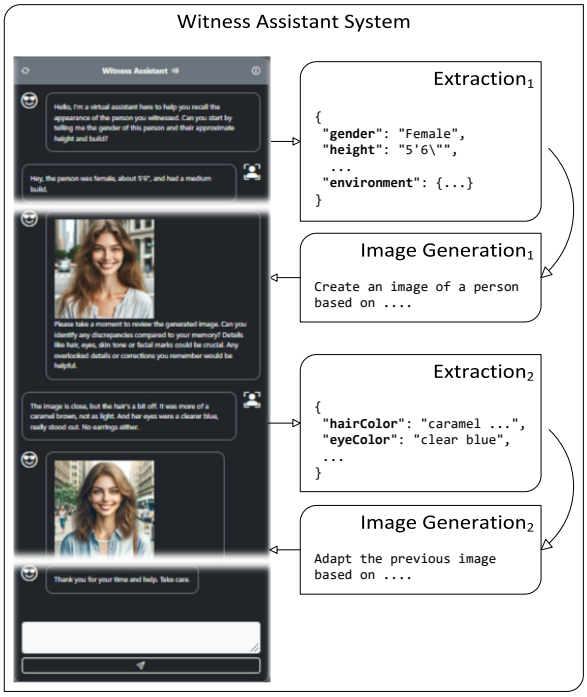

Figure 3: Witness Assistant System

```
Assist me in crafting prompts for
guiding GPT to function as a chatbot.
This chatbot is designed to (...)
```

For example, the meta prompt was then extended with the specific instruction

```
I need a prompt that instructs GPT to
have a conversation with the witness in
order to elicit some key features of
the person they've met. (...)
```

in order to obtain the witness assistant outer state prompt

```
You are having a conversation with a
user who is a witness to an incident.
Your goal is to (...)
```

The fact that this application could be realised with PROMISE demonstrates that the framework effectively supports the development of the required conversational interaction, including the different phases, phase transitions, extractions, image generation and multilayering.

6 Application Validation

Building on the witness assistant application developed using the PROMISE framework, we now turn our attention to the utility of this application. This section is dedicated to evaluating the application's effectiveness in achieving its intended goals, particularly in improving the accuracy and efficiency of information gathering. Through a simulated

interaction, we demonstrate how the application not only operationalizes the framework but also enhances the user experience and outcome quality.

The application was tested by using a fictitious witness simulated using ChatGPT with GPT-4 (OpenAI, 2023), who is asked to describe a person visible in a provided picture. For this purpose, we used the following prompt.

```
Generate responses as if you're a user
interacting with a chatbot (...)
I'll provide the chatbot's messages,
and you'll respond as the user (...)
When asked for descriptions, your
answers should reflect what is visible
about the person in this picture.
[Picture of Person Here]
```

The picture shown in Fig. 4 was obtained from Unsplash[1] and added to the previous prompt at the location indicated with the squared brackets. From

Figure 4: Person to be described by simulated witness

then on, the assistant's utterances such as

```
Hello, I'm a virtual assistant here to
help you recall the appearance of (...)
```

were provided to obtain the utterance of the simulated witness such as

```
Hey, the person was female, about 5'6",
and had a medium build.
```

[1]Guilherme Stecanella on Unsplash

The witness utterances were passed back to the assistant, whose responses were in turn passed on to the simulated witness, and so forth.

The resulting interaction shown in Fig. 1 demonstrates that the assistant behaves according to the combination of the outer state and respective inner state prompts. Furthermore, the different green-coloured utterances, each representing a specific state, adequately triggered state transitions due to the transition decision prompts. Finally, the transition action prompts facilitated the extractions of descriptive information resulting in JSON objects shown in Sect. 3. As can be observed, these JSON objects fully capture all the information provided by the simulated witness.

Based on the JSON object extracted from the first phase of the interaction, the following prompt was used with DALL-E 3 to generate an initial possible image of the person to be identified:

```
Create visual representations based on
witness descriptions. I'll supply you
with JSON objects detailing the
characteristics of an individual, and
your task is to produce images that
match these descriptions.
[JSON Object Here]
```

The image generated using the first JSON object obtained from the first phase of the interaction is shown in Fig. 5.

Figure 5: Image generated after first phase

The generated image was then passed on to the simulated witness, in order to correct or add previously collected attributes and to obtain another JSON object capturing these changes. The following prompt was then used with DALL-E 3 to generate new versions of a previous image based on the

JSON object extracted from subsequent phases of the interaction.

```
The witness was shown an image and
queried about any modifications that
could enhance its similarity to their
recollection. Here is the JSON object
with the witness's suggested
adjustments:
[JSON Object Here]
```

The image generated using the second JSON object obtained from the second phase of the interaction is shown in Fig. 6. A simple quantitative comparison

Figure 6: Image generated after second phase

using the cosine-similarity of the two generated images using Dino2 (Oquab et al., 2024) embeddings confirms the high-level visual matching (cosine = 0.89) and consistent general agreement with the original picture (cosine = 0.75, 0.79).

One limitation we encountered in using image generation models is that it is difficult to generate multiple images in which only prescribed aspects differ. As can be observed when comparing the images in Fig. 5 and 6, despite the JSON object suggesting no more changes than to the hair and eye colors, and not having earrings (which wasn't present in the first image), the second picture shows a collar, a handbag, earrings, and the background was altered. This is a known fact in generative computer vision, and it has its origin in the intrinsic diffusion model at the base.

While the ability to generate such images proves the utility of our application, the validity of the individual images as well as their stepwise adaptation largely depends on the capability of the image generation model. Consequently, any in-depth validation of the differences among the generated images is out of the scope of this paper, as it would rather serve as a validation of the specific image generative model employed instead of validating the witness assistant application or the PROMISE framework.

7 Conclusion

The use of language models (LM) to support conversational interactions is promising but challenging. As the complexity of the expected behaviour grows, so does the prompt specifying the behaviour, which increases the likelihood of misbehaviour. We therefore introduce the notion of stateful prompt orchestration which follows the idea of segmenting complex prompts into smaller ones, which can then be combined dynamically, depending on the state of the interaction.

Given the requirements from an application scenario, we presented the use of the PROMISE framework to design and implement a multimodal conversational interaction. PROMISE supports this by leveraging state machine modelling concepts. This enables developers of interactions to orchestrate prompts, not only to harness the LMs capabilities for open-ended conversations, but also to enable complex conversational interactions including conversation flows, flow transitions, extractions, recursive flows, and multilayered interactions.

With the successful development and simulated use of a proof-of-concept application, we demonstrate that PROMISE effectively supports the design and implementation of useful conversational interactions. The resulting application proves the ability of PROMISE to manage different prompts and effectively orchestrate these prompts to enable complex multimodal interactions beyond what is feasible with single-prompt LM applications.

Our next steps involve augmenting the PROMISE framework with persuasive conversational capabilities by providing the means to dynamically select and apply persuasion strategies. These extensions will enable persuasive conversational interactions that deliver demonstrable benefits in the healthcare sector, for example. As PROMISE is able to respond to defined conversation segments, this will allow the use of different persuasion strategies during the conversation, further enhancing the framework's support for more sophisticated interactions.

References

Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2023. A survey of chain of thought reasoning: Advances, frontiers and future. *arXiv preprint arXiv:2309.15402*.

Chrisantha Fernando, Dylan Banarse, Henryk Michalewski, Simon Osindero, and Tim Rocktäschel. 2023. Promptbreeder: Self-referential self-improvement via prompt evolution. *arXiv preprint arXiv:2309.16797*.

Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. 2023. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*.

Solveig Helland, Elena Gavagnin, and Alexandre de Spindler. 2023. Divide et impera: Multi-transformer architectures for complex nlp-tasks. In *Proceedings of the 8th edition of the Swiss Text Analytics Conference*.

Anand Singh Jalal, Dilip Kumar Sharma, and Bilal Sikander. 2023. Suspect face retrieval using visual and linguistic information. *The Visual Computer*, 39(7):2609–2635.

Paweł Karczmarek, Witold Pedrycz, Adam Kiersztyn, and Przemysław Rutka. 2017. A study in facial features saliency in face recognition: An analytic hierarchy process approach. *Soft Computing - A Fusion of Foundations, Methodologies and Applications*, 21(24):7503–7517.

Mohd. Aamir Khan and Anand Singh Jalal. 2020. A framework for suspect face retrieval using linguistic descriptions. *Expert Systems with Applications*, 141:112925.

Pawel Korzynski, Grzegorz Mazurek, Pamela Krzypkowska, and Artur Kurasinski. 2023. Artificial intelligence prompt engineering as a new digital competence: Analysis of generative ai technologies such as chatgpt. *Entrepreneurial Business and Economics Review*, 11(3):25–37.

Maximilian Mozes, Xuanli He, Bennett Kleinberg, and Lewis D Griffin. 2023. Use of llms for illicit purposes: Threats, prevention measures, and vulnerabilities. *arXiv preprint arXiv:2308.12833*.

OpenAI. 2023. GPT-4 Technical Report.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. DINOv2: Learning Robust Visual Features without Supervision.

Lidan Wang, Vishwanath Sindagi, and Vishal Patel. 2018a. High-Quality Facial Photo-Sketch Synthesis Using Multi-Adversarial Networks. In *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018)*, pages 83–90, Xi'an, China. IEEE Press.

Nannan Wang, Wenjin Zha, Jie Li, and Xinbo Gao. 2018b. Back projection: An effective postprocessing method for GAN-based face sketch synthesis. *Pattern Recognition Letters*, 107:59–65.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems*, 35:24824–24837.

Gary L. Wells and Lisa E. Hasel. 2007. Facial Composite Production by Eyewitnesses. *Current Directions in Psychological Science*, 16(1):6–10.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382*.

Wenyuan Wu, Jasmin Heierli, Max Meisterhans, Adrian Moser, Andri Färber, Mateusz Dolata, Elena Gavagnin, Alexandre de Spindler, and Gerhard Schwabe. 2023. Promise: A framework for model-driven stateful prompt orchestration.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. 2017. Unpaired Image-to-Image Translation Using Cycle-Consistent Adversarial Networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251.