Conversational Image Generation: Towards Multi-Round Personalized Generation with Multi-Modal Language Models

Supplementary Material

6. Personalization Prompts

The full prompts used in Figure 3 are:

1&3) "A person with goth makeup, with their face visible and distinguishable. The person has a pale complexion, with dark eyeliner and mascara accentuating their eyes. Their lips are painted a deep red, and their eyebrows are plucked and drawn on to create a sharp, angular shape. A silver stud pierces their left eyebrow, and a choker made of black leather adorns their neck. The person's hair is black and styled in a messy, spiky fashion, adding to their goth aesthetic."

2) "A person looking up into the sky, with their face visible and distinguishable. The person is standing with their feet shoulder-width apart, their eyes squinting slightly as they gaze upwards. They are wearing a light-colored shirt and jeans, and their hair is blowing gently in the wind. The sky above is a brilliant blue, with only a few wispy clouds scattered across it. The person's expression is one of wonder and awe, as if they are marveling at the vastness of the sky."

4) "A person sticking out their tongue, with their face visible and distinguishable. The person's eyes are wide open, and their eyebrows are raised, creating a comical expression. Their tongue is bright pink and slightly curled, adding to the playful appearance. The person's face is positioned close to the camera, emphasizing the tongue-sticking-out gesture. The image is cropped closely around the person's face, focusing attention on the tongue and facial expression."

The prompts used in Figure 6 are

1) "A person dressed as a pirate captain, with their face visible and distinguishable, standing at the helm of a ship navigating through the rough waters of the North Sea. The captain is wearing a white shirt with billowy sleeves, a red vest, and a black tricorn hat adorned with a golden chain and a feather. A whiskey tumbler glass is held tightly in their hand, with a hint of whiskey remaining at the bottom. The captain's facial expression is one of determination and focus, with a hint of ruggedness and weathered skin, suggesting a seasoned sailor."

2) "A person with a USA president look, with their face visible and distinguishable. They are wearing a navy blue suit with a white shirt and a red tie. A pair of glasses perches on the end of their nose, and a hint of a smile plays on their lips. Their hair is neatly combed and gray, suggesting a sense of wisdom and experience. The person exudes an air of confidence and authority, as if they are about to deliver an important speech or address the nation. The focus is on the person's face and upper body, with a blurred background that emphasizes their presence and leadership."

<s> [INST] Generate a person at a western wedding. The person is decently dressed in attire fitting for a western celebration. They are standing in a rustic, outdoor wedding venue, surrounded by the natural beauty of a mountainous landscape, with a clear blue sky and a few puffy white clouds. The person is posed with their head held high, a gentle smile on their face, and their body slightly angled to showcase their attire. The wedding venue is adorned with a mix of western and natural elements, including wooden decorations, wildflowers, and a wooden archway where the couple stands. The atmosphere is lively yet serene, with a few guests milling about, taking photos and enjoying the celebration. *Please keep the face identical* <img>  </img> [/INST]

I have generated an image. *I keep the face unchanged* <img>  </img> </s>

Figure 9. Examples of single turn personalization training prompt template. The images in prompt represent the 64 image tokens, *i.e.* CLIP features.

Figure 10. Illustration of accumulated error when encoding and decoding an image several times. This result suggests to caching CLIP features in chat history instead of images when performing multi-round inference.

The full prompts used in Figure 7 are:

1) "The image shows Julian and Ruby sitting on a couch together, smiling at the camera. Julian has fair skin with brown hair and stubble. He is wearing a light-blue hooded sweatshirt and holding a potato chip in his right hand. Ruby has fair skin with red shoulder-length hair. She is wearing a tan cardigan and black pants. She is holding a potato chip in her left hand. The background appears to be a living room. There are white shelves on the left with books and decorations on them. On the right, there is a tall floor lamp with a yellow lampshade and a white bookcase behind it."

2) "The image shows Dr. Harrison examining Maya's face. Dr. Harrison has fair skin and short gray hair, is wearing a white coat and blue glasses, and is holding Maya's

Figure 11. Single-turn personalization visual examples of our MLLM.

chin with his left hand while he looks at her face with his right eye closed and his right hand touching her cheek. Maya has fair skin and brown hair tied back, is wearing a beige shirt with black trim and is sitting on a beige couch. The background is a room with white walls and black-framed windows."

3) "The image shows Olivia and Julian looking at a laptop screen together. Olivia, with long brown hair and bangs, looks down at a silver MacBook Pro that she holds on her lap. She is wearing a gray sweater and has headphones around her neck. Julian, with a beard and mustache, stands to her right, also looking down at the laptop screen. He is wearing a yellow shirt and white over-ear headphones. The background is a blurred room with white walls and a window on the left."

4) "The image shows Julian and Mia sitting on a gray couch, reading a book together. Julian has dark skin and black hair, and he's wearing glasses, a gold shirt, blue jeans, and white sneakers. He's holding an open book with both hands, looking down at it and smiling. Mia has dark skin and curly brown hair. She's wearing a white blouse and blue jeans, and she's leaning against Julian, looking down at the book and smiling. Her legs are crossed, and her right foot is resting on the floor. The couch is light-gray with two tufted seats and two matching throw pillows. Behind them, there's a tall, dark-gray metal bookshelf with a woven basket on top. A green plant peeks out from behind the basket. In front of the couch, there's a window with a white sheer curtain covering it."

5) "The image shows Julian and Isabella sitting at a table in a restaurant, taking a selfie. Julian has fair skin and brown hair with a beard and mustache. He wears black glasses and a maroon button-down shirt. He sits on the left side of the table and smiles as he holds his phone up to take a selfie. Isabella has fair skin and brown hair pulled into a bun. She wears pearl earrings and a pale-pink knit sweater. She leans toward Julian and kisses him on the cheek while holding her hand under her chin. In front of them on the table are two white coffee cups and saucers. The background is blurred and appears to be a restaurant or cafe. There are hanging lights above the couple and more tables set with dishes and glassware behind them."

6) "The image shows Amelia kissing baby Oliver on the head. Amelia has fair skin and brown hair tied back in a ponytail. She is wearing a white shirt and a colorful scarf with orange, green, blue, black and yellow flowers on it. She is holding Oliver in her right arm who is looking at the camera and smiling. Oliver has fair skin and blue eyes. He is wearing a white beanie and a mint-green sweater with a white bib underneath. In the background there are trees and a path on the right side."

7. Detokenizer Discussion Continued

As discussed in Section 3.1, even with our proposed DiT-based detokenizer, perfect reconstruction remains elusive. In this section, we explore how the number of image token matters. We trained an additional DiT detokenizer on 256

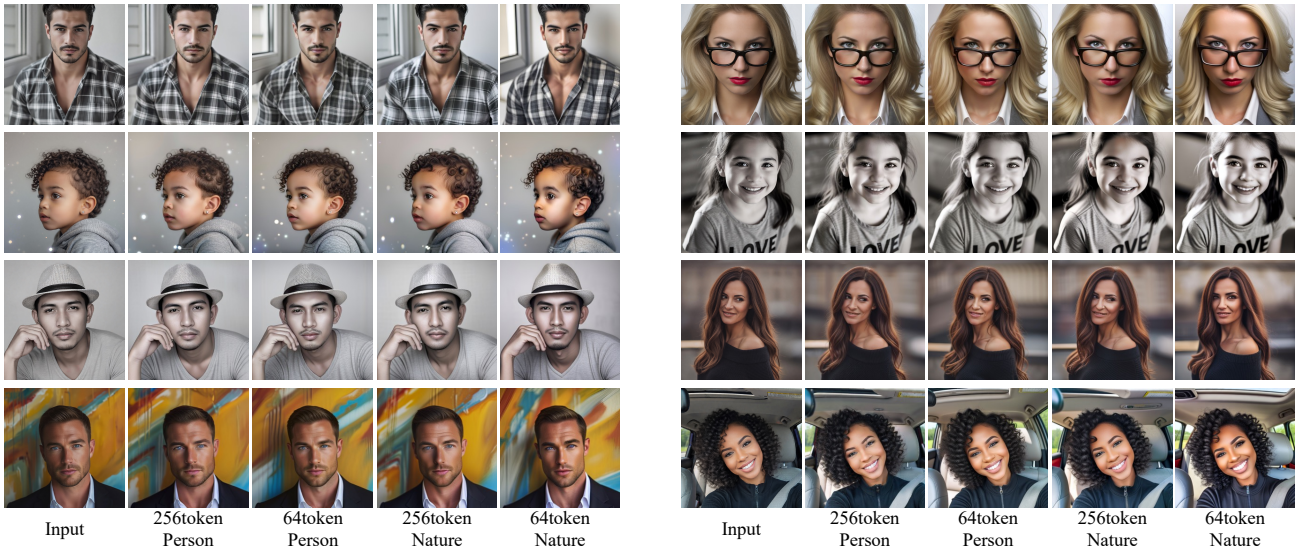

Figure 12. Comparison of DiT Detokenizer Results. "256 token" indicates the absence of 1D pooling after Qwen-VL. Notably, 1) the 256 token configuration demonstrates superior face reconstruction capabilities, but its integration into our pipeline would necessitate re-pretraining the LLaMA component. 2) Fine-tuning on human images consistently enhances the face preservation ability of both detokenizers.

Qwen-VL features without pooling, using the same strategy. The results in Figure 12 indicate that the DiT detokenizer exhibits improved content preservation when provided with 256 image tokens. However, this approach is not feasible for our experiments, as altering the number of tokens in the tokenizer necessitates pretraining the LLaMA component again on text-and-image interleaved data. Since our primary objective is to demonstrate the conversational multi-round image generation capability, we focus on instruction fine-tuning based on the SEED-X pre-trained model. This result also shows that fine-tuning on human images can enhance the face preservation ability even with 256 tokens.

8. Name-based Multi-turn Personalization Continued

In Section 3.3.2, we described our method for constructing a name-based multi-round personalization dataset, which involves generating a segmented face (close-up photo) for the second and third personalization rounds. In this section, we present results from our further exploration. Using the segmented faces shown in Figure 5 and a pool of personalization prompts, we employed a diffusion-based personalization model [13] to generate full-body images as ground truths for personalization rounds. This process resulted in another multi-round personalization dataset, exemplified as follows:

- **Round1 text-to-image generation with two-person prompt including their names**, *e.g.* "Generate a image shows *Henry* and *Lucas* sitting at a table together. Henry has white hair and a white beard, and he is wearing a blue-and-white checkered button-up shirt. He is looking down at his hands as he holds two small pots with brown dirt in them. There is a hand protruding from the bottom left corner of the image holding a handful of seeds that are spilling out into the pots. Lucas is on the right side of the table. He has blond hair and he is wearing a navy-blue button-up shirt. He is looking down at the table with a neutral expression. There are gardening tools on the table in front of him. The background shows a kitchen with light-brown wood panel walls. There is a white sink on the left edge of the image. Above it, there is a white countertop with a white faucet. On the back wall, there is a white electrical outlet with a white switch above it. There is a white cabinet underneath the countertop on the right side of the image".
- **Round2 name-based personalization with full-body prompt**, *e.g.* "*Henry* is sitting on a light-green metal folding chair at an outdoor cafe. They wear a red beanie, a yellow long-sleeve shirt with a white checkered pattern, black pants, white socks, and black and white Nike shoes. Their left leg is bent upward and their right leg is stretched out behind them. They hold a phone in their left hand and look at the camera with a neutral expression. In front of them are two light-blue chairs and one light-yellow chair. The background is a gray sidewalk with green grass growing between it and a building. On the other side of the sidewalk is a glass wall with tall red spikes protruding from the ground."

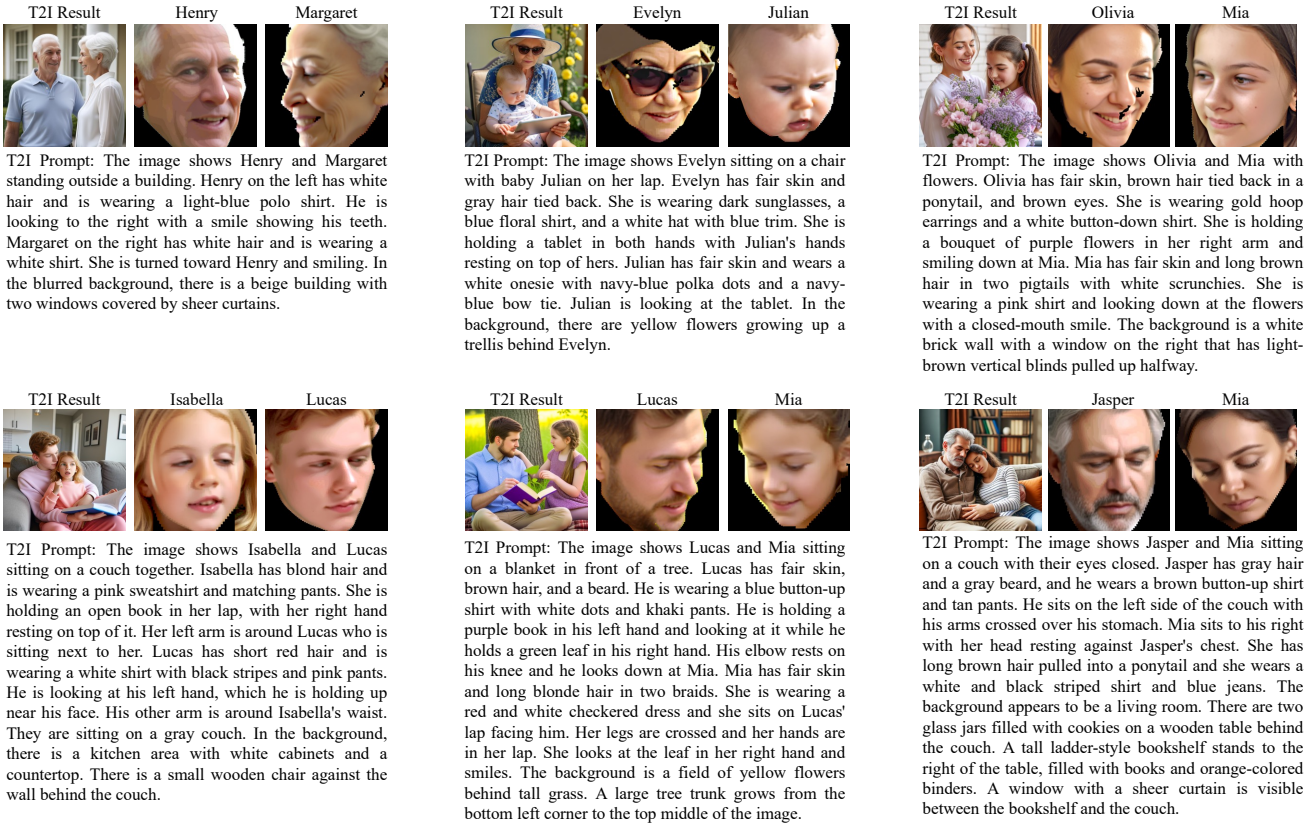

Figure 13. More examples of multi-turn personalization results.

Figure 14. More examples of full-body multi-turn personalization results.

- **Round3 name-based personalization with full-body prompt**, *e.g.* "*Lucas* is sitting on a black bench. They are wearing a black peacoat over a black shirt and blue jeans with holes in the knees. They look at the camera with a neutral expression. The background is a white wall with a shadow falling on the left side."

The above example is actually part of our evaluation set, corresponding to the second image in Figure 14, and the training set follows the same structure.

Fine-tuned with this dataset, our MLLM generates multi-turn results, as exemplified in Figure 15, where we perform inference twice with the same text prompts. Since the T2I prompt does not specify whether Lucas is a boy or a teenager, we produce two different images as the 1st-round output. Notably, the personalization round generates full-body images with subjects of similar age and race; for instance, if a teenager is generated in the 1st round, the personalization result in 3rd round is also a teenager. This behavior strongly indicates that our MLLM can reason from both the 1st-round input and output to generate a contextually appropriate personalization result.

Additionally, our model effectively follows the full-body prompt, although the faces may vary. This outcome might result from the synthetic ground truth in our training dataset.

Figure 15. Examples of multi-turn personalization results. As can be observed, 1) This model can follow the prompt well, but struggles to maintain consistent face identity. Please see Section 8 last paragraph for discussion. 2) When inferring twice with the same prompt, distinct 1st-turn T2I results are generated. Subsequently, the 2nd- and 3rd-turn personalization results are different as well. For instance, if the model initially generates an image containing a boy rather than a teenager, the subsequent personalization results will also depict a boy. This behavior is a strong evidence that our model can generating image based on reasoning from text-image interleaved chat histories.

Since the full-body personalization ground truth is generated by diffusion models [13], the ground truth may not always match the condition image even after filtering, introducing noise into the training set. Consequently, this task presents plenty of work for future research. This paper focuses on the chat history analyzing capability of MLLM in image generation, and Figure 15 effectively demonstrates this capability, especially compared with Figure 16.

9. Implement Details

In our approach to DiT-based detokenizer training, we integrate an MLP adapter atop DiT to adjust the dimension of Qwen-VL image encoder, ensuring compatibility with its input dimension. We employ the same dataset and methodology as outlined in [30] to fine-tune DiT for detokenization purposes. A critical configuration involves using a constant learning rate of 10^{-5} with an effective batch size of 1024, as smaller batch sizes result in model non-convergence. The DiT was fine-tuned 180,000 iterations on nature images and another 96,000 iterations on human images.

For LLaMA fine-tuning, we generally adhere to the default settings of SEED-X, incorporating necessary modifications. The LLM model is initialized with a pretrained LlamaForCausalLM and trained with LoRA [15] strategy. We utilize the AdamW optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, and $\epsilon = 10^{-6}$. By default, training is configured for 60,000 iterations with a weight decay of 0.05 and a maximum gradient norm of 1.0, employing mixed precision training with 'bf16'. All models are trained on a single node with 8 GPUs, using gradient accumulation to adjust the effective batch size.

For single-turn personalization, we set the LoRA rank and α to 1280 across all three stages. In the first stage, LLaMA is trained to output the input directly, using the SEED-X pretrained version with a constant learning rate of 10^{-5} and an effective batch size of 1024 for 6,000 iterations. The second stage involves inputting a cropped face

Figure 16. Multi-turn personalization results of SEED-X without fine-tuning on our proposed multi-turn dataset. These results can be directly compared to Figure 7. As can be observed, firstly, this baseline model has difficuties in generate a reasonable two-person images in the first round; Then in personalization rounds, this model functioned similarly to a T2I model by identifying text near the name and generating a new face using its T2I capability, entirely disregarding the first-round output conditions.

and caption to predict the entire image, with a learning rate of 10^{-6} and an effective batch size of 512 for 30,000 iterations. In the final stage, paired data is introduced with a learning rate of 10^{-7} and an effective batch size of 1024 for 24,000 iterations, maintaining a fixed 1:2 ratio between stage 2 data and paired stage3 data.

For multi-turn personalization, we use a LoRA rank and α of 1280, with learning rates of 10^{-4}, 10^{-5}, and 10^{-6} for 28,000-, 50,000-, and 12,000- iteration training, respectively. Due to time constraints, the effective batch size is limited to 512. Ideally, multi-stage training should be employed, but due to time limitations, all datasets are mixed for training. In addition to the multi-turn dataset constructed in Section 3.3.2, we use single-turn stage 2 and stage 3 data as augmentation. Stage 2 prompts are used to predict corresponding full images for T2I tasks (Agmnt1), and full images from stage 2 and stage 3 data are used to predict cropped faces for personalization (Agmnt2). SciQA [1] is used as a regularization to maintain reasoning ability. Consequently, the dataset mix ratio is multi-turn: Agmnt1: Agmnt2: SciQA = 6:2:3:1.