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

Haochen Zhang[1], Animesh Sinha, Felix Juefei-Xu, Haoyu Ma, Kunpeng Li,
Zhipeng Fan, Xiaoliang Dai, Tingbo Hou, Peizhao Zhang, Zecheng He

[1]UC San Diego

haz035@ucsd.edu, zcheustc@gmail.com

Abstract

Recent advancements in diffusion models have significantly enhanced personalized image generation, enabling high-fidelity synthesis of human-subject-specific images. However, existing approaches are constrained by the inherent limitations of diffusion models, which lack conversational capabilities, and operate in a single-round setting, restricting user interaction. In this work, we propose a novel framework that integrates multi-modal large language models (MLLMs) for multi-round conversational personalization. To achieve this, we identified a performance bottleneck in the detokenizer of current MLLMs, which struggles to reconstruct fine-grained facial identity details. Thus, we enhance the detokenizer with a personalization-enhaced Diffusion Transformer (DiT). We also introduce a multi-stage instruction fine-tuning strategy to balance face preservation and prompt alignment effectively. To support multi-round generation, we implement a chat-history caching mechanism and construct the first multi-round personalization dataset from video clips. Experimental results demonstrate that our approach achieves state-of-the-art performance among MLLM-based personalization methods. To the best of our knowledge, this is the first work to enable conversational personalization, unlocking new capabilities for MLLMs in personalized image generation.

1. Introduction

Recent advancements in diffusion models [14, 29, 32] have revolutionized image generation, demonstrating remarkable capabilities in semantic understanding and high-fidelity synthesis. Trained on large-scale datasets of image-text pairs, these models excel in generating diverse, photorealistic images from textual prompts and have been widely adopted for tasks such as image editing [2, 5], inpainting [45, 49], style transfer [42, 47], and controllable generation [23, 51]. Among these applications, subject-driven image generation [7, 40] has gained significant attention, aiming to synthesize images that faithfully preserve specific subject identities or attributes from reference images. In particular, personalization—the generation of diverse images of a given human face that align with textual prompts while maintaining identity consistency—has emerged as a key challenge.

Early approaches to personalization in diffusion models, such as Textual Inversion [10] and DreamBooth [33], achieved personalization through diffusion model fine-tuning. More recent works [13, 18, 40] have focused on extracting visual embeddings from reference images and injecting them into the diffusion process, enabling subject-driven generation without requiring per-user fine-tuning. While these methods demonstrate strong performance, they operate in a single-round setting, where a reference image and textual prompt are processed once to generate an output. This limitation arises from the inherent nature of diffusion models, which lack conversational capabilities to support multi-round generation through conversation.

In contrast, our work unlocks the potential of multi-round conversational personalization through natural conversation. We first developed a framework that leverages multi-modal LLMs (MLLMs) to enable conversational sessions, akin to current text-based LLMs [39, 54], empowering users to personalize generated outputs in multi-round chatting. However, we discovered that vanilla MLLMs, which are trained on general-purpose data, fall short in preserving subject identity in personalized image generation. Our investigation revealed that this shortcoming is largely due to the detokenizer's limited ability to reconstruct fine-grained details as reference images. To address this, we enhanced the detokenizer with a more powerful Diffusion Transformer (DiT) [27], specifically fine-tuned on human images, thereby significantly improving identity preservation in the generated images.

Furthermore, we extend our framework to the more challenging task of multi-round personalization—a scenario where diffusion models falter due to their inability to maintain contexts. To overcome this limitation, we harness the conversational strengths of MLLMs by integrating a chat-history caching mechanism. This proven strategy retains context across iterations, ensuring that each new input is enriched by the accumulated conversation history. We also construct the first multi-round personalization dataset, featuring an initial text-to-image generation round followed by multiple rounds of name-based personalization. This dataset enables MLLMs to reason from both text and image chat histories, demonstrating their ability to generate personalized outputs consistently across conversations.

To the best of our knowledge, this is the first work to enable MLLMs for personalized image generation and the first to demonstrate their unique strength in multi-round personalized image generation via conversation. Our main contributions are summarized as follows:

- We proposed a new framework using MLLM for multi-round conversational personalized image generation.
- We identified a detokenizer bottleneck in reconstructing fine-grained image details in MLLMs and improved using a personalization-enhanced DiT.
- We introduce a multi-stage instruction fine-tuning strategy for better identity and editability tradeoff.
- We proposed a new methodology creating multi-round personalization data from video clips and constructed the first name-based multi-round personalization dataset.
- Experiments show that our approach achieves state-of-the-art performance in MLLMs based personalization, as validated through human assessment, and proves new capability in multiround conversational personalization.

2. Related Work

2.1. Image Generation

Diffusion models, starting from DDPM [14], have significantly progressed in text-to-image generation, with models like Stable Diffusion [29, 32], DALL-E [31], and Imagen [4] excelling in producing visuals from textual prompts through iterative denoising. To improving visual-conditioned generation, enhancements like ControlNet [51] and T2I-Adapter [23] have been integrated into these models, supporting applications in image editing [2, 5], composition [40], subject-driven generation [7], and so on.

In this paper, we focus on human-centric subject-driven generation, referred to as personalization. Early personalization approaches, such as Textual Inversion [10] and DreamBooth [33], fine-tune diffusion models on identity-specific tokens but often suffer from limited generalizability. Recent methods aim to balance personalization and efficiency by integrating visual and textual features, thereby reducing the need for user-specific fine-tuning. ELITE [43] maps vision features into the text-embedding space using both local and global transformations. PhotoMaker [18] fuses vision and text tokens using cross-attention. IP-Adapter [48] leverages face classification embeddings and a CLIP vision encoder to enhance identity preservation. InstantID [40] incorporates ControlNet [51] to enable precise control over pose and facial expression. However, these methods operate in a single-round setting, constrained by the inherent limitations of diffusion models, which lack conversational capabilities.

2.2. Multimodal Large Language Models

Research in LLMs [25, 26, 39] has surged after ChatGPT-3.5 [24] which demonstrated its advancements in generating human-like text. Subsequently, notable efforts have extended them beyond text to incorporate images, thus forming VLLMs [6, 21, 28, 56]. These models have exhibited remarkable capabilities in vision-language understanding tasks. For instance, LLaVA [20] proposes a language-image instruction dataset and integrates visual perception into LLMs through a MLP vision-language connector.

Recent developments have introduced MLLMs [12, 35, 38] generating both text and images, in which a key component is the visual encoder-decoder pair, tokenizing images as a new language. For example, Chameleon [37] employs image tokenizer to generate discrete image tokens for LLM modeling. The EMU series [35, 36] and SEED-X [12] have leveraged CLIP [3, 34] features for continuous-space modeling. Concurrent advancement EMU3 [41] have reported an enhanced text-to-video generation performance by improving the tokenizer and data quality. Additionally, innovative works like TransFusion [55] and Show-O [44] have unified diffusion and autoregressive methods into a single model. Building on this foundation, our paper explores new capacities of MLLMs in single round and conversational multi-round personalization.

2.3. Multi-round Image Generation Datasets

To develop a MLLM capable of generating images interactively based on chat history, multi-round instruction fine-tuning datasets for image generation are essential. Unfortunately, popular multi-round datasets, such as LLaVA [20], SVIT [53], and LLaVAR [52], feature text outputs. With modifications, some text-image interleaved datasets, like MMC4 [57], VIST [16], and LeafInstruct[46], can be converted into multi-round text-image interleaved datasets. However, the images in these datasets often lack consistency, limiting their effectiveness in conditional image generation, such as personalization.

Existing datasets closely aligned with our requirements are some multi-turn editing datasets like MagicBrush [50] and SEED-Data-Edit [12], which provide a source image and a series of editing prompts. Each target image corresponding to a prompt is an edited version based on the

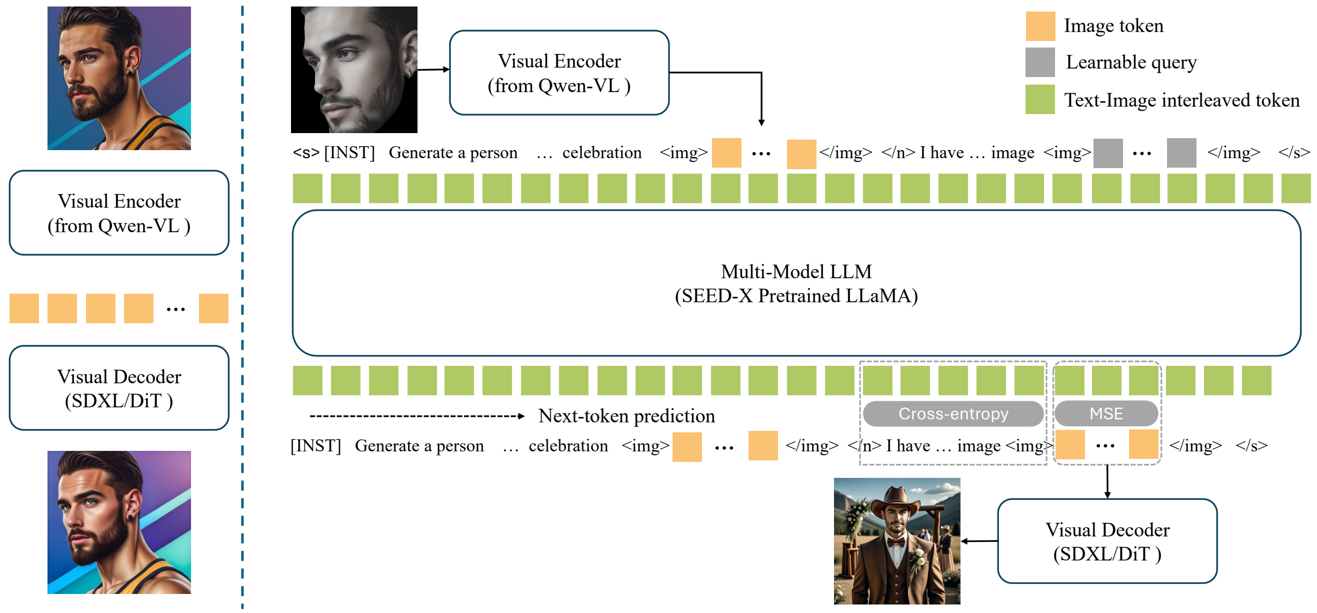

Visual Encoder
(from Qwen-VL)
Visual Decoder
(SDXL/DiT)
Image token
Learnable query
Text-Image interleaved token
Visual Encoder
(from Qwen-VL)
<s> [INST] Generate a person … celebration <img> ··· </img> </n> I have … image <img> ··· </img> </s>
Multi-Model LLM
(SEED-X Pretrained LLaMA)
Cross-entropy
MSE
Next-token prediction
[INST] Generate a person … celebration <img> ··· </img> </n> I have … image <img> ··· </img> </s>
Visual Decoder
(SDXL/DiT)

Figure 1. Overview of our framework and data pipeline. Prompt is from example in Figure 9.

previous target, following a Markov property. This property downgrades multi-round image generation to multiple single rounds as detailed in Seciton 3.3.1. Thus, it is necessary to construct new datasets for real multi-round image generation. Specifically, we created a name-based multi-round dataset from video clips for multi-turn personalization, which requires fine-grained chat hisotry analysis. The details will be further elaborated in Section 3.3.2.

3. Method

Our proposed method build upon mainstream MLLMs, where images are treated as a new language, enabling text-image interleaved pretraining. As illustrated in Figure 1, we adopt SEED-X [12] as the backbone, integrating a pre-trained and frozen image encoder from Qwen-VL [3], to extract visual features. The image inputs are processed into 64 image tokens via average pooling, while text inputs are tokenized using a text tokenizer. These tokens are then fed into the LLaMA [39] model for next-token prediction, where both text and image tokens are used for understanding and reasoning. During training, text token predictions are optimized using cross-entropy loss, while image token predictions use regression loss. In the inference stage, the predicted text tokens are detokenized to textual outputs, and the predicted image tokens are passed through the image decoder, SDXL [29], to reconstruct the final image output. Please also refer to their original paper [12] for details.

To minimize unnecessary complexity, our approach focuses solely on the instruction fine-tuning stage, bypassing the computationally expensive pretraining step. We make only essential modifications, such as upgrading the SDXL detokenizer to DiT detailed in the next section.

3.1. DiT based Visual Detokenization

As outlined above, the quality of the output images is influenced by two key components: the reasoning module, LLaMA, and the visual SDXL detokenizer. In this section, we address the bottleneck in the SDXL detokenizer, while improvements to the LLaMA training are discussed in Section 3.2. Specifically, SEED-X [12] has two stages for finetuning SDXL to serve as a detokenizer. The initial stage [11] utilizes a reconstruction approach, where the decoder is provided with the CLIP feature of the input to reconstruct the input itself. In contrast, the second stage compensates for the loss of detailed information in the encoder by concatenating a conditional image to the noise map of the diffusion model, finetuning it on editing data. Figure 2 illustrates the detokenizer decoding results.

Figure 2 reveals two critical observations: 1) the stage1 reconstruction-based detokenizers struggle to preserve content accurately, especially for human faces; 2) the editing-finetuned detokenizer demonstrates improved performance only in scenarios where the conditional image and input share an editing relationship, whereas cannot be generalized to other context, such as face preservation. Moreover, when finetuned with editing data, the SDXL decoder architecture was modified to fit the task by introducing extra layers further limited the generalizability. This leads to generate artifacts if out-of-distribution condition images are provided. These observations underscore the first challenge addressed in this paper: enhancing the performance of detokenizers

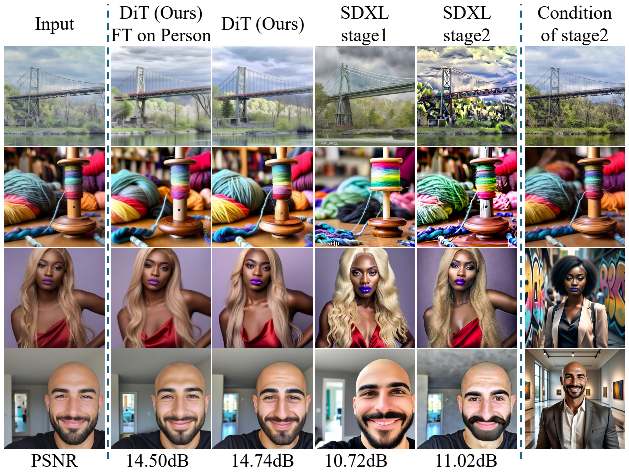


Figure 2. Detokenizer performance comparison. SDXL stage1 detokenizer struggles with detail preservation. Stage2 uses a condition image to keep shapes but introduces artifacts and fails to maintain faces. Our DiT detokenizer preserves details without additional conditions and reconstructs faces well after tuning on human images. PSNR analysis also shows superior reconstruction quality of our DiT over SDXL detokenizer

that rely solely on reconstruction objectives, thereby improving their reconstruction ability and generalizability.

Our approach to addressing this challenge involves the adoption of the DiT detokenizer. To avoid re-pretraining LLaMA component, we retain to employ the same pretrained Qwen-VL image encoder as the tokenizer, which is kept frozen during training. We also employ a 1D average pooling to downsample tokens from 256 to 64[1]. We then adopt a vanilla structure [27] DiT and fine-tuned it to reconstruct natural images from image tokens, initialized with a text-to-image model [30]. This process is analogous to the stage1 fine-tuning of the SDXL detokenizer in SEED-X.

As depicted in Figure 2, some artifacts occurred when using SDXL stage2 detokenizer. Compared to SDXL stage1 detokenizer, there was a significant improvement in content preservation for both natural scenes and human subjects. However, slight alterations in facial features were observed, which could potentially affect personalization performance. To mitigate this, we further finetuned the DiT on human images, but still on the reconstruction task. This additional finetuning enhanced face preservation. The PSNR evaluation on a subset of COCO2014[19] images further demonstrates a significant improvement in reconstruction quality with our DiT compared to the baseline SDXL detokenizer. Despite our DiT detokenizer did not achieve perfect reconstruction as some VQGAN based detokenizers such as SBER-MoVQGAN [22], partly because of the number of image tokens, it is good enough to serve our purpose.

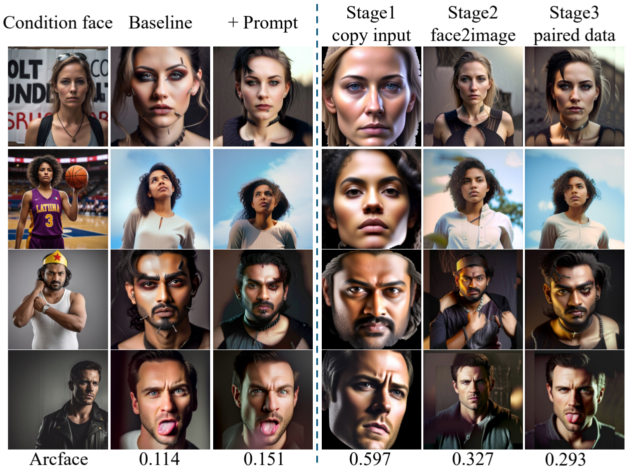


Figure 3. Personalization results w.r.t. improvements. Prompt instructions enhance face preservation. Stage1 yields identical face generation; Stage2 maintains face similarity with editability beyond faces but struggles with complex prompts; Stage3 achieves the best balance between face identity and editability.

3.2. MLLM for Single Round Personalization

After addressing the bottleneck in detokenizer, in this section, we move to LLaMA component enhancements to realize single turn personalized image generation with MLLMs.

As the baseline, we first finetune MLLM on personalization data [13] using a prompt template "`<s> [INST] Generate the image shows {caption} <img> {source embedding} </img> [/INST] \n I have generated an image. <img> {target embedding} </img> </s>`", where `\n` splits input prompt and model response. Please refer to Figure 9 for a specific example. Results are illustrated in Figure 3 where we observed that, despite following the prompts accurately, the generated images often retained characteristics such as race, gender and hairstyle, but featured completely different faces. This discrepancy highlighted the need for refinements in LLM component.

To address this challenge, we refined our training approach by focusing on two key aspects: prompt design and training strategy. Leveraging the fact that MLLM possesses the ability to comprehend human language, we incorporated dense textual context to preserve identity attributes during finetuning. Specifically, we augmented the prompt with instructions such as "Please keep the face identical" and included confirmatory responses like "I keep the face unchanged" before generating the desired image tokens. This approach enabled us to improve facial consistency while still allowing for creative freedom in image generation. Considering that ArcFace [8] score indicates a slight improvement, from 0.114 to 0.151, we regarded this as a trick and not explicitly listed in our contributions.

[1] Refer to Suppl. Material 7 for detokenizer results without pooling.

To ensure a better personalization performance, we proposed a multi-stage fine-tuning strategy to transition MLLM from merely replicating faces to achieving editable personalization: 1) Initially, we trained the model to output images resembling the input, preserving high identity fidelity even trained on non-human data. Although this stage never complied with prompt instructions, it served as a good starting point for identity and editability trade-off. 2) Subsequently, we trained the model to predict the original images with the condition of cropped and masked human faces. This stage taught MLLM to replicate face areas while adhering to the prompt in the others. Models trained after this stage have exhibited good prompt following ability while keeping the face identical. However, as shown in Figure 3, the model fails to follow complex prompts which require face changes, such as 'goth makeup' or 'sticking out their tongue' detailed in Suppl. Material 6. 3) Thus, our final stage involved fine-tuning MLLM with paired images of same people, where faces from one image were used to predict the other whole image, supplemented by regularization using data from the second stage. This stage resulted in a model that optimally balanced identity preservation with prompt alignment.

3.3. Conversational Multi-Round Generation

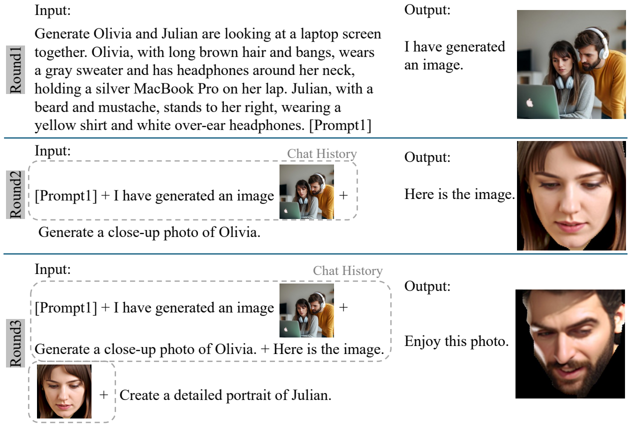
This section explores a novel task—conversational image generation—which requires the analysis of text-image interleaved chat history, a crucial capability for multimodal conversations. To demonstrate this, we develop an instruction fine-tuning dataset in context of personalization, enabling MLLMs to effectively handle multi-round text-image interleaved contextual dependencies.

3.3.1. Conversational Generation Definition

Before introducing our proposed dataset, we first clarify key terminologies: single round, multiple single rounds, and multi-round. In the context of MLLMs, given a text input X_t and a visual input X_v, the **single round** text-image interleaved response X_a can be modeled as $p(X_a|X_t, X_v)$. Consider existing multi-round editing datasets such as MagicBrush [50] and SEED-Data-Edit [12], both of them adhere to the Markov property: the edited image X_a^i depends solely on the current instruction X_t^i and the previously generated result X_a^{i-1}, while ignoring longer history. This property simplifies multi-round image generation, reducing it to a **multiple single turn** process, *i.e.* $\{p(X_a^i|X_t^i, X_a^{i-1})\}_{i=1}^N$. Different from existing multiple single turn setting, our goal is to enable MLLMs to generate images via conversation, akin to text-based LLMs [39, 54]. To achieve this, we cache the conversation history and incorporate it as additional input alongside the current user inputs. We formalize conversational **multi-round** image generation as:

$$p(X_a^i|X_t^i, X_v^i, \{X_t^k, X_v^k, X_a^k\}_{k=1}^{i-1})$$

where past interactions, both input and output, serve as contextual information for the current generation step.



Generate Olivia and Julian are looking at a laptop screen together. Olivia, with long brown hair and bangs, wears a gray sweater and has headphones around her neck, holding a silver MacBook Pro on her lap. Julian, with a beard and mustache, stands to her right, wearing a yellow shirt and white over-ear headphones. [Prompt1]

I have generated an image.

[Prompt1] + I have generated an image +

Generate a close-up photo of Olivia.

Here is the image.

[Prompt1] + I have generated an image +

Generate a close-up photo of Olivia. + Here is the image.

+ Create a detailed portrait of Julian.

Enjoy this photo.

Figure 4. Example of multi-round personalization inference. The 1st round is T2I generation. The 2nd and 3rd round of personalization leverages both the textual input and visual output from round one for person description and appearance, respectively.

Consider a simplified two-round case, denoted as $I_1 \xrightarrow{T_1} I_2 \xrightarrow{T_2} I_3$. The process unfolds as follows: In the first round, the initial input is $[I_1, T_1]$, producing the output $p(I_2|T_1, I_1)$. In the second round, although the user provides only T_2 as input, the MLLM processes the full context $[\{I_1, T_1, I_2\}, T_2]$. Consequently, the second-round output is given by $p(I_3|T_2, I_1, T_1, I_2)$, incorporating information from the entire chat history alongside the current input. Considering some special cases: 1) If I_3 primarily depends on I_1, the generation can be simplified to $p(I_3|T_2, I_1)$, where the chat history integrates essential information I_1. 2) If I_3 relies on I_2 only, the process reduces to multiple single round generations, formulated as $p(I_3|T_2, I_2)$. This shows that existing multi-turn datasets [12, 50] composed of independent single turns are simply a special case of the broader non-Markov multi-turn dataset we study.

3.3.2. Name-based Multi-Round Personalization Dataset

With these definitions established, it becomes obvious that conversational multi-round image generation remains largely unexplored in existing literature, with no suitable multi-round instruction fine-tuning datasets available. To address this gap, we take the first step toward this unexplored domain by introducing a name-based multi-round personalization dataset.

In general, we target on a conversational multi-round personalization illustrated in Figure 4. The process begins with an initial text-to-image (T2I) generation round, followed by two rounds of personalization. In the first round, individuals Julian and Olivia are generated based on detailed textual descriptions and assigned names. In the second round, only Olivia's name is referenced without additional appearance description, requiring MLLMs to reason across chat histories. A similar level of in-context reasoning is also required in the third round of personalization.

It is important to notice that this multi-round design is intentionally complex to enable fine-grained chat history analysis. Unlike a straightforward multi-round design that combines personalization with incremental editing refinements, our task necessitates precise **token-level** retrieval and association of past information. For example, in Figure 4, when given only the name 'Olivia' in the current prompt, the MLLM must first identify her corresponding textual description from the first round input—such as 'long brown hair' and 'headphones around her neck'—and then use this information to locate her appearance in the T2I-generated results. Notably, only a partial textual description pertains to Olivia, and only a subset of image tokens represents her face. Completing this more difficult task demonstrates the MLLM's capability for chat history analysis and highlights its potential for complex conversational image generation.

To achieve the Figure 4 capability, we propose to develop the name-based multi-round personalization dataset derived from video clips featuring two individuals. As depicted in Figure 5, we establish paired images via utilizing the first and last frames of these clips, each containing two subjects. For the captions of the first frame, we adopt the captioning strategy outlined in [13]. Subsequently using LLaMA-3 [9], we assign appropriate names to each individual and modify the captions accordingly. Concurrently, we employ ArcFace [8] and SAM [17] to detect and segment each person's face in the last frame. This setup allows us to construct a multi-round personalization sample as illustrated at the bottom of Figure 5, realizing conversational personalization. Ideally with masked faces, a further refinement could be generating synthetic full-body images using diffusion model-based personalization methods [13, 48]. We discuss more in Suppl. Material 8 and visualize some examples in Figure 8.

4. Experimental Results

4.1. Implement Details

Dataset details. For single round personalization, we use the same training dataset proposed in [13]. For multi-round personalization, our proposed dataset is derived from approximately 170,000 videos, each featuring two subjects. After filtering out videos where the subjects were too far apart to fit within a 512x512 square box, around 150,000 videos remained. We utilized LLaMA [9] to generate captions for the frames and subsequently rewrote these captions. By filtering out captions that did not contain exactly two names, we obtained 92,471 samples for model training. Additionally, for the full-body personalizaion experiment described in Section 8, we applied ArcFace [8] scores to

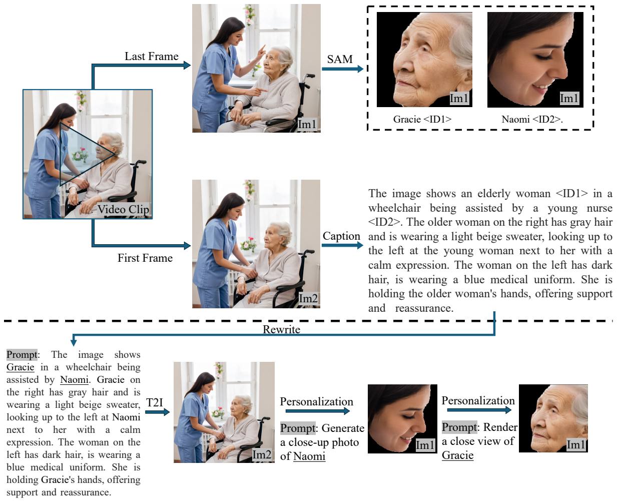

Last Frame
SAM
Im1
Im1
Im1
Gracie <ID1>
Naomi <ID2>.
Video Clip
First Frame
Im2
Caption
The image shows an elderly woman <ID1> in a wheelchair being assisted by a young nurse <ID2>. The older woman on the right has gray hair and is wearing a light beige sweater, looking up to the left at the young woman next to her with a calm expression. The woman on the left has dark hair, is wearing a blue medical uniform. She is holding the older woman's hands, offering support and reassurance.
Rewrite
Prompt: The image shows Gracie in a wheelchair being assisted by Naomi. Gracie on the right has gray hair and is wearing a light beige sweater, looking up to the left at Naomi next to her with a calm expression. The woman on the left has dark hair, is wearing a blue medical uniform. She is holding Gracie's hands, offering support and reassurance.
T2I
Im2
Personalization
Prompt: Generate a close-up photo of Naomi
Im1
Personalization
Prompt: Render a close view of Gracie
Im1

Figure 5. Illustration of name-based multi-round personalization data derived from video. We caption one frame and assign each person a name for the first T2I round. We segment the faces in the other frame as the ground truth for personalization rounds.

filter diffusion model-based personalization results, resulting in 24,793 training samples. During inference, all single round quantitative evaluations are conducted on a subset (∼400) of the dataset from [13], as no widely accepted public benchmark currently exists for personalization tasks.

Implementation. Details on MLLM finetuning are elaborated in Suppl. Material Section 9.

4.2. Single Round Personalization

To demonstrate the effectiveness of our proposed DiT detokenizer and multi-stage fine-tuning strategy, we conduct experiments comparing our approach with state-of-the-art MLLMs, SEED-X and EMU2. All baseline results are obtained using the official SEED-X and EMU2 models and implementations to ensure a fair comparison.

Qualitative results. Figure 6 shows the visual comparison among SEED-X, EMU2 and our models. Please refer to the Suppl. Material Section 6 for the prompts used in this figure. As SEED-X does not provide a instruction finetuned model on personalization task, we here use the pretrained version in comparison. As can be observed, SEED-X can hardly preserve human face identity no matter what detokenizer is used, operating similarly to a text-to-image model. If focusing on three examples on the left, EMU2 can keep human identity in some extent. However on the right side of examples where the condition images and the prompts have conflicts, for example the typical characters of "pirate captain" and "USA president look" are usually male, EMU2 struggles to keep the identity in condition image and generate the wrong gender. In contrast, our model shows the best identity preservation ability while keeping

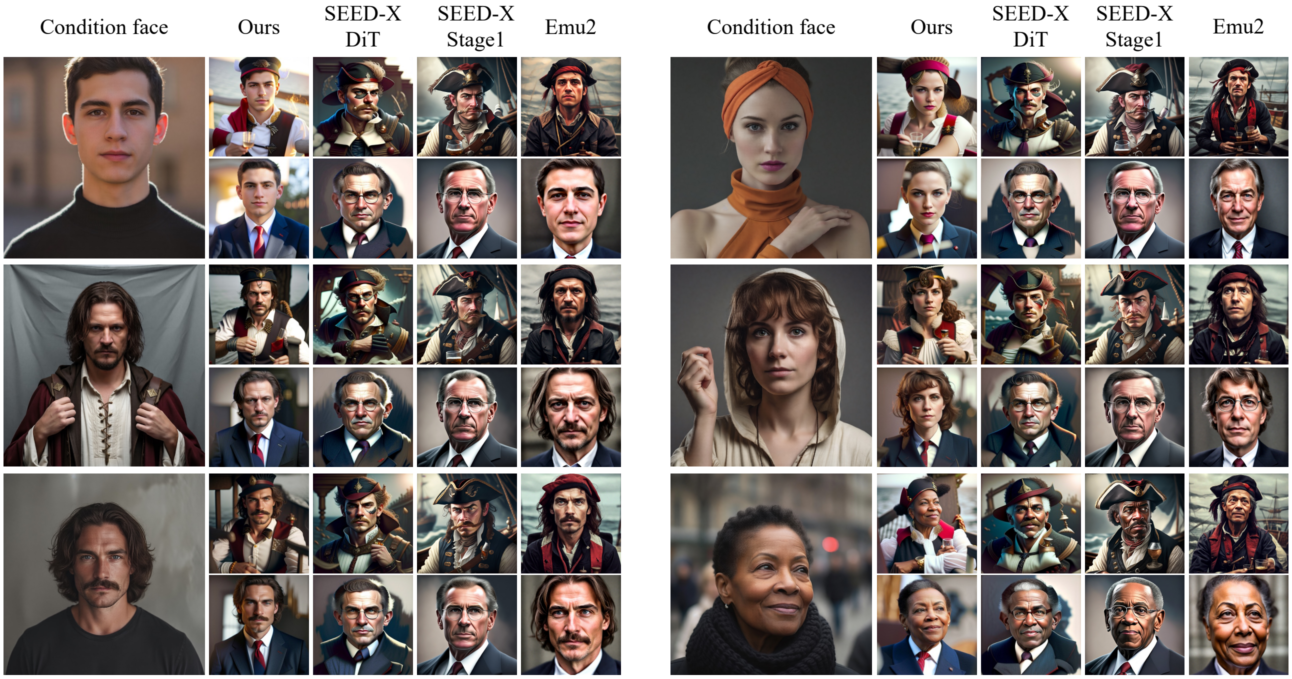


Figure 6. Performance comparison on single-turn personalization task. Please refer to the Suppl. Material 6 for the prompts. SEED-X and EMU2 struggle to perserve human faces especially when the condition images and the prompts have conflicts. Our model shows the best tradeoff between face perseveration and prompt alignment. More visual inspection can be found in Figure 11.

Table 1. User study result. Our model shows significantly better image quality and face identity. For prompt alignment, while the baseline model, SEED-X, functions like a text-to-image model with strong prompt adherence, our model still outperforms it.

Measurement	Ours wins	Loses	Tie
Quality	73.75%	3.75%	22.5%
Alignment	36.25%	15%	48.75%
Face ID	71.25%	2.5%	26.25%

aligned with given prompts. Considering that our model is improved based on SEED-X instead of EMU2, the improvements is significant. Additional examples from our model are shown in Figure 11.

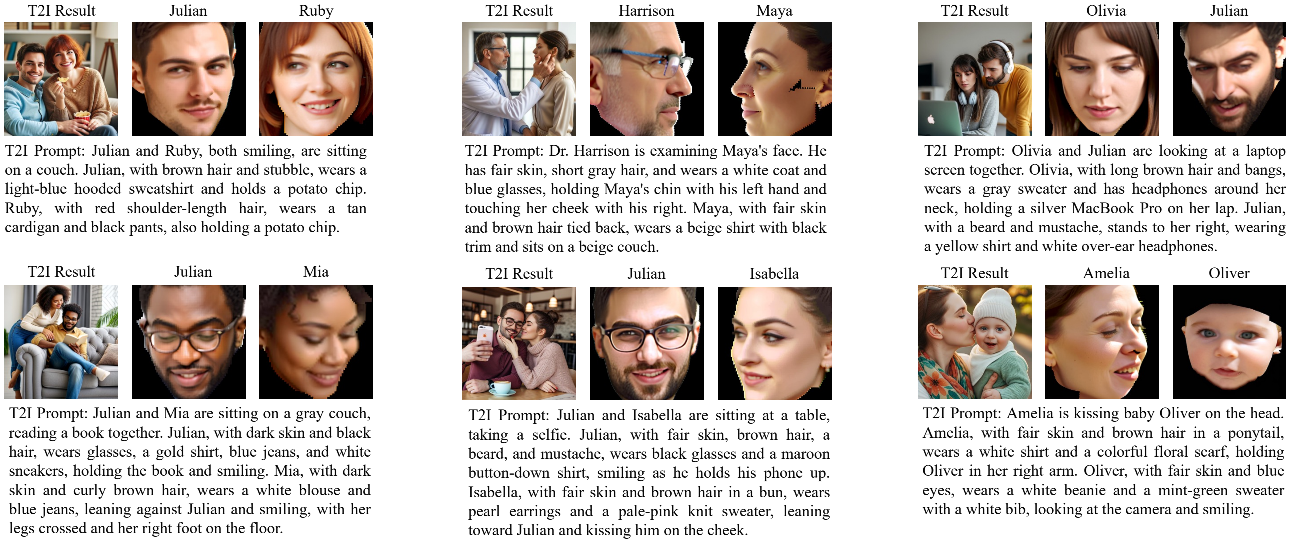
Quantitative results. Our qualitative evaluation further demonstrates the superior performance of our model compared to the baseline SEED-X. During the model development phase, we use ArcFace [8] and CLIP score [33] as preliminary evaluation metrics to assess fine-tuned models. We observe a significant improvement in **ArcFace** (0.094 → 0.293) and comparable **CLIP score** (28.36 → 28.59) for single round personalization. Once a satisfactory model is achieved, we conduct human evaluations across three key aspects: image quality, prompt alignment, and face preservation. As observed in Table 1, our model consistently outperforms the baseline SEED-X if not tie. Among these aspects, SEED-X demonstrates strong performance in prompt alignment, as it primarily functions as a text-to-image model with high adherence to textual prompts while often neglecting visual conditions. This limitation is evident in the fourth column of Figure 6, where the three "USA president look" results appear similar despite differing input face conditions.

It is important to note that user studies provide the most accurate yet costly evaluation. Discrepancies between human evaluation and automated metric scores highlight inherent limitations in automated evaluation metrics. The substantial performance gap between our model and SEED-X underscores a significant improvement. To the best of our knowledge, this is the first work to demonstrate the capability of MLLMs in personalization tasks.

4.3. Multi-Round Personalization

Due to the novelty of the multi-round personalization task, there is limited prior work for direct comparison. Thus, we evaluate the effectiveness mainly through visual inspection.

The multi-round personalization results are presented in Figure 7. This task involves three rounds: The first round is a text-to-image generation, where we provide a detailed description of two individuals with corresponding names. In the subsequent two rounds, the MLLM is tasked with generating the face of each named individual. This requires the model to reason not only from the round one input to iden-



T2I Prompt: Julian and Ruby, both smiling, are sitting on a couch. Julian, with brown hair and stubble, wears a light-blue hooded sweatshirt and holds a potato chip. Ruby, with red shoulder-length hair, wears a tan cardigan and black pants, also holding a potato chip.

T2I Prompt: Dr. Harrison is examining Maya's face. He has fair skin, short gray hair, and wears a white coat and blue glasses, holding Maya's chin with his left hand and touching her cheek with his right. Maya, with fair skin and brown hair tied back, wears a beige shirt with black trim and sits on a beige couch.

T2I Prompt: Olivia and Julian are looking at a laptop screen together. Olivia, with long brown hair and bangs, wears a gray sweater and has headphones around her neck, holding a silver MacBook Pro on her lap. Julian, with a beard and mustache, stands to her right, wearing a yellow shirt and white over-ear headphones.

T2I Prompt: Julian and Mia are sitting on a gray couch, reading a book together. Julian, with dark skin and black hair, wears glasses, a gold shirt, blue jeans, and white sneakers, holding the book and smiling. Mia, with dark skin and curly brown hair, wears a white blouse and blue jeans, leaning against Julian and smiling, with her legs crossed and her right foot on the floor.

T2I Prompt: Julian and Isabella are sitting at a table, taking a selfie. Julian, with fair skin, brown hair, a beard, and mustache, wears black glasses and a maroon button-down shirt, smiling as he holds his phone up. Isabella, with fair skin and brown hair in a bun, wears pearl earrings and a pale-pink knit sweater, leaning toward Julian and kissing him on the cheek.

T2I Prompt: Amelia is kissing baby Oliver on the head. Amelia, with fair skin and brown hair in a ponytail, wears a white shirt and a colorful floral scarf, holding Oliver in her right arm. Oliver, with fair skin and blue eyes, wears a white beanie and a mint-green sweater with a white bib, looking at the camera and smiling.

Figure 7. Examples of multi-turn personalization results. The inference involves 3 rounds: Round 1 takes T2I prompt as input to generate an image; Round 2, 3 use prompt "Generate a close-up photo of {name}" to generate faces. Our model generates images of the two individuals in the 1st round as well as faces of the correct individuals in personalization rounds.

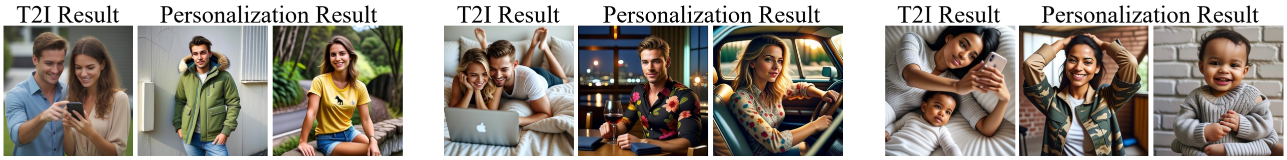


Figure 8. Examples of full-body multi-turn personalization results. More examples are presented in Suppl. Material Figure 14.

tify the person but also from the round one output in the chat history to determine the person's appearance. As observed, our model effectively generates images of the two individuals in the first round, despite minor mismatches, such as "red shoulder-length hair." More importantly, in the second and third rounds, our model generates faces of the correct individuals as appeared in the first round output. Additional examples are provided in the Suppl. Material Figure 13. These results demonstrate the capabilities of MLLMs in generating personalized images during conversations with users, inspiring more developments of AI assistants generate both text and images interactively.

We also extend the name-based multi-round personalization task to full-body cases, detailed in Suppl. Material Section 8. We conduct experiments and present example results in Figure 8. Replacing 'close-up photo' with detailed prompts, our model effectively generate full-body personalized images. We observe that the face preservation in this setting might not as good as that in Figure 7. This outcome mainly attributes to our training data: Instead of using real images, the full-body personalization ground truth is synthesized via diffusion models [13], whose results may not always perfectly preserve the conditional faces, introducing noise into the training samples. In this regard, this result matches our expectation and we leave further face refinements as our future research. This paper focuses on exploring conversational image generation, emphasizing the ability of MLLMs to reason over chat history during image generation. This capability has been well demonstrated in both Figure 7, Figure 8 and even failure cases in Figure 15.

5. Conclusion

This paper introduces a novel framework for multi-round conversational personalized image generation using MLLMs. It integrates DiT as an improved detokenizer and employs a multi-stage fine-tuning strategy for enhanced face preservation. Additionally, it leverages MLLMs' conversational strengths through a chat-history caching mechanism and the first name-based multi-round personalization dataset. Experimental results confirm MLLMs' ability to handle multi-turn personalization by analyzing chat history.

This work marks an initial step toward conversational image generation, with challenges remaining for future research. Key areas for improvement include enhancing the detokenizer's content preservation for long conversations and developing more comprehensive multi-turn instruction fine-tuning datasets beyond name-based personalization.

References

[1] Sören Auer, Dante AC Barone, Cassiano Bartz, Eduardo G Cortes, Mohamad Yaser Jaradeh, Oliver Karras, Manolis Koubarakis, Dmitry Mouromtsev, Dmitrii Pliukhin, Daniil Radyush, et al. The SciQA scientific question answering benchmark for scholarly knowledge. *Scientific Reports*, 13 (1):7240, 2023. 6

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *CVPR*, pages 18208–18218, 2022. 1, 2

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, et al. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond. 2023. 2, 3

[4] Jason Baldridge, Jakob Bauer, Mukul Bhutani, Nicole Brichtova, Andrew Bunner, Kelvin Chan, Yichang Chen, Sander Dieleman, Yuqing Du, Zach Eaton-Rosen, et al. Imagen 3. *arXiv preprint arXiv:2408.07009*, 2024. 2

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023. 1, 2

[6] Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023. 2

[7] Wenhu Chen, Hexiang Hu, Yandong Li, Nataniel Ruiz, Xuhui Jia, Ming-Wei Chang, and William W Cohen. Subject-driven text-to-image generation via apprenticeship learning. *NIPS*, 36, 2024. 1, 2

[8] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *CVPR*, pages 4690–4699, 2019. 4, 6, 7

[9] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 6

[10] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1, 2

[11] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, et al. Making llama see and draw with seed tokenizer. *arXiv preprint arXiv:2310.01218*, 2023. 3

[12] Yuying Ge, Sijie Zhao, Jinguo Zhu, Yixiao Ge, Kun Yi, Lin Song, Chen Li, Xiaohan Ding, and Ying Shan. SEED-X: Multimodal models with unified multi-granularity comprehension and generation. *arXiv preprint arXiv:2404.14396*, 2024. 2, 3, 5

[13] Zecheng He, Bo Sun, Felix Juefei-Xu, Haoyu Ma, Ankit Ramchandani, Vincent Cheung, Siddharth Shah, Anmol Kalia, Harihar Subramanyam, Alireza Zareian, et al. Imagine yourself: Tuning-free personalized image generation. *arXiv preprint arXiv:2409.13346*, 2024. 1, 4, 6, 8, 3, 5

[14] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NIPS*, 33:6840–6851, 2020. 1, 2

[15] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, et al. LoRA: Low-rank adaptation of large language models. In *ICLR*, 2022. 5

[16] Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. Visual storytelling. In *Proceedings of the 2016 conference of the North American chapter of the association for computational linguistics: Human language technologies*, pages 1233–1239, 2016. 2

[17] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. In *ICCV*, pages 4015–4026, 2023. 6

[18] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. In *CVPR*, pages 8640–8650, 2024. 1, 2

[19] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO: Common objects in context. In *ECCV*, pages 740–755, 2014. 4

[20] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023. 2

[21] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024. 2

[22] Anastasia Maltseva, Arseniy Shakhmatov, Andrey Kuznetsov, and Denis Dimitrov. SBER-MoVQGAN. https://github.com/ai-forever/MoVQGAN, 2023. 4

[23] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2I-Adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *AAAI*, pages 4296–4304, 2024. 1, 2

[24] OpenAI. Chatgpt-3.5: Optimizing language models for dialogue. https://openai.com/blog/chatgpt, 2022. 2

[25] OpenAI. Chatgpt. https://chat.openai.com, 2023. 2

[26] OpenAI. GPT-4 technical report. *ArXiv*, abs/2303.08774, 2023. 2

[27] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pages 4195–4205, 2023. 1, 4

[28] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023. 2

[29] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 1, 2, 3

[30] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, et al. Movie Gen: A cast of media foundation models. *arXiv preprint arXiv:2410.13720*, 2024. 4, 5

[31] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2

[32] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 1, 2

[33] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. DreamBooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*, pages 22500–22510, 2023. 1, 2, 7

[34] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. 2

[35] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners. In *CVPR*, pages 14398–14409, 2024. 2

[36] Quan Sun, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multimodality. *ICLR*, 2024. 2

[37] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024. 2

[38] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2

[39] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. LLAMA 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023. 1, 2, 3, 5

[40] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, and Anthony Chen. InstantID: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 1, 2

[41] Xinlong Wang, Xiaosong Zhang, Zhengxiong Luo, Quan Sun, Yufeng Cui, Jinsheng Wang, Fan Zhang, Yueze Wang, Zhen Li, Qiying Yu, et al. Emu3: Next-token prediction is all you need. *arXiv preprint arXiv:2409.18869*, 2024. 2

[42] Zhizhong Wang, Lei Zhao, and Wei Xing. Stylediffusion: Controllable disentangled style transfer via diffusion models. In *ICCV*, pages 7677–7689, 2023. 1

[43] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: Encoding visual concepts into textual embeddings for customized text-to-image generation. In *ICCV*, pages 15943–15953, 2023. 2

[44] Jinheng Xie, Weijia Mao, Zechen Bai, David Junhao Zhang, Weihao Wang, Kevin Qinghong Lin, Yuchao Gu, Zhijie Chen, Zhenheng Yang, and Mike Zheng Shou. Show-O: One single transformer to unify multimodal understanding and generation. *arXiv preprint arXiv:2408.12528*, 2024. 2

[45] Dejia Xu, Xingqian Xu, Wenyan Cong, Humphrey Shi, and Zhangyang Wang. Reference-based painterly inpainting via diffusion: Crossing the wild reference domain gap. *arXiv preprint arXiv:2307.10584*, 2023. 1

[46] Zhiyang Xu, Minqian Liu, Ying Shen, Joy Rimchala, Jiaxin Zhang, Qifan Wang, Yu Cheng, and Lifu Huang. Lateralization lora: Interleaved instruction tuning with modality-specialized adaptations. *arXiv preprint arXiv:2407.03604*, 2024. 2

[47] Serin Yang, Hyunmin Hwang, and Jong Chul Ye. Zero-shot contrastive loss for text-guided diffusion image style transfer. In *CVPR*, pages 22873–22882, 2023. 1

[48] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 6

[49] Tao Yu, Runseng Feng, Ruoyu Feng, Jinming Liu, Xin Jin, Wenjun Zeng, and Zhibo Chen. Inpaint anything: Segment anything meets image inpainting. *arXiv preprint arXiv:2304.06790*, 2023. 1

[50] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *NIPS*, 2023. 2, 5

[51] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023. 1, 2

[52] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun. LLaVaR: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*, 2023. 2

[53] Bo Zhao, Boya Wu, Muyang He, and Tiejun Huang. SVIT: Scaling up visual instruction tuning. *arXiv preprint arXiv:2307.04087*, 2023. 2

[54] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric. P Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena, 2023. 1, 5

[55] Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. Transfusion: Predict the next token and diffuse images with one multi-modal model. *arXiv preprint arXiv:2408.11039*, 2024. 2

[56] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

[57] Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *NIPS*, 36, 2024. 2

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:

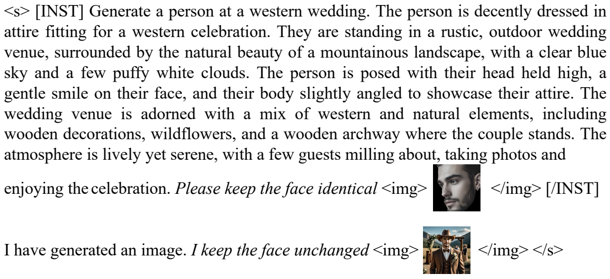
1&3) “A person with goth makeup, with their face visible and distinguishable. The person has a pale complexion, with dark eyeliner and mascara accentuating their eyes. Their lips are painted a deep red, and their eyebrows are plucked and drawn on to create a sharp, angular shape. A silver stud pierces 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.”



<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.

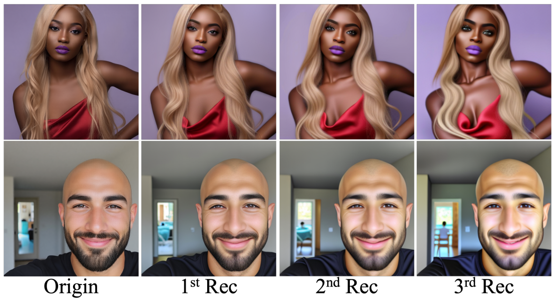


Figure 10. Illustration of accumulated error when encoding and decoding an image several times. This result suggests to caching CLIP features in chat history instead of images when performing multi-round inference.

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.”

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

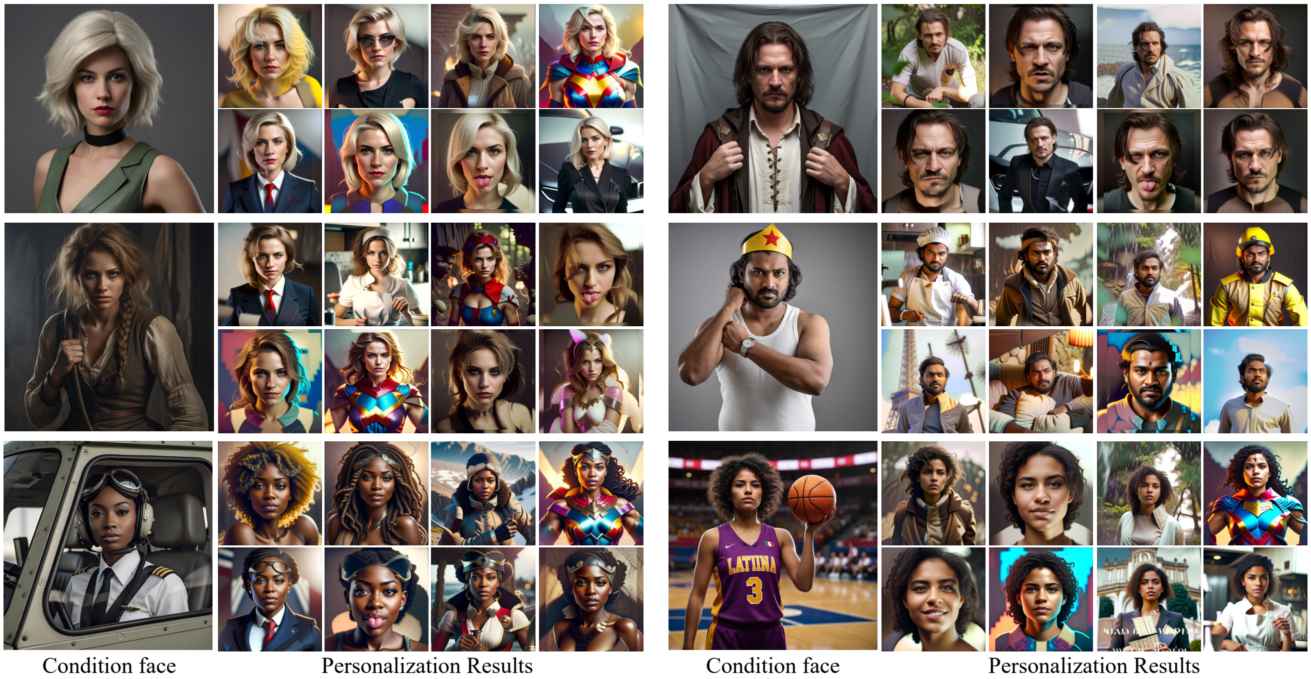


Figure 11. Single-turn personalization visual examples of our MLLM.

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

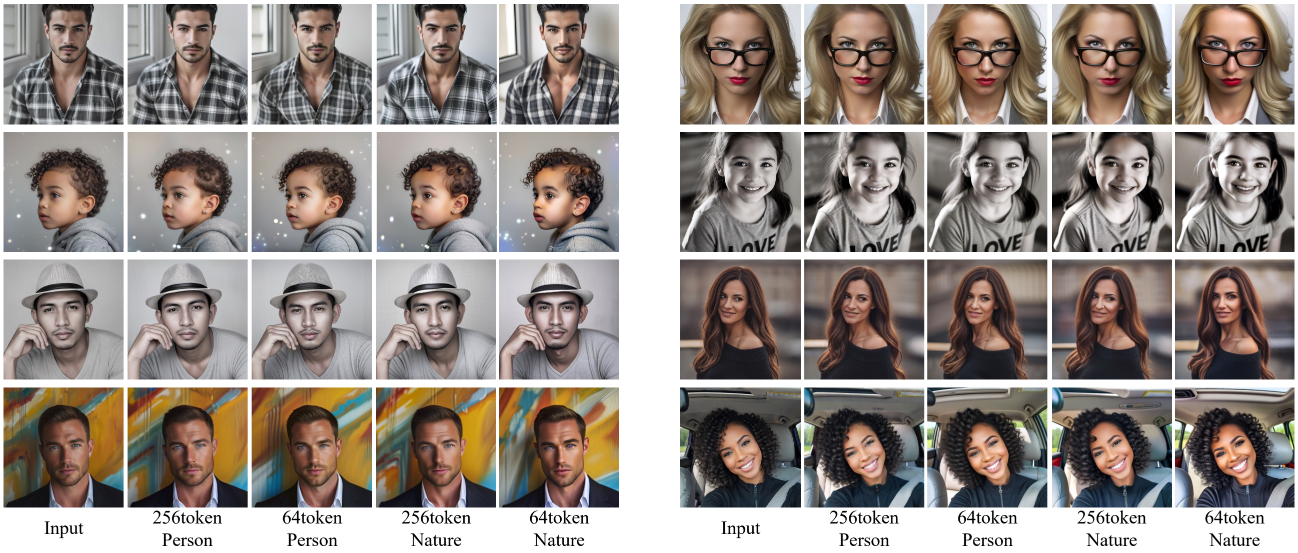


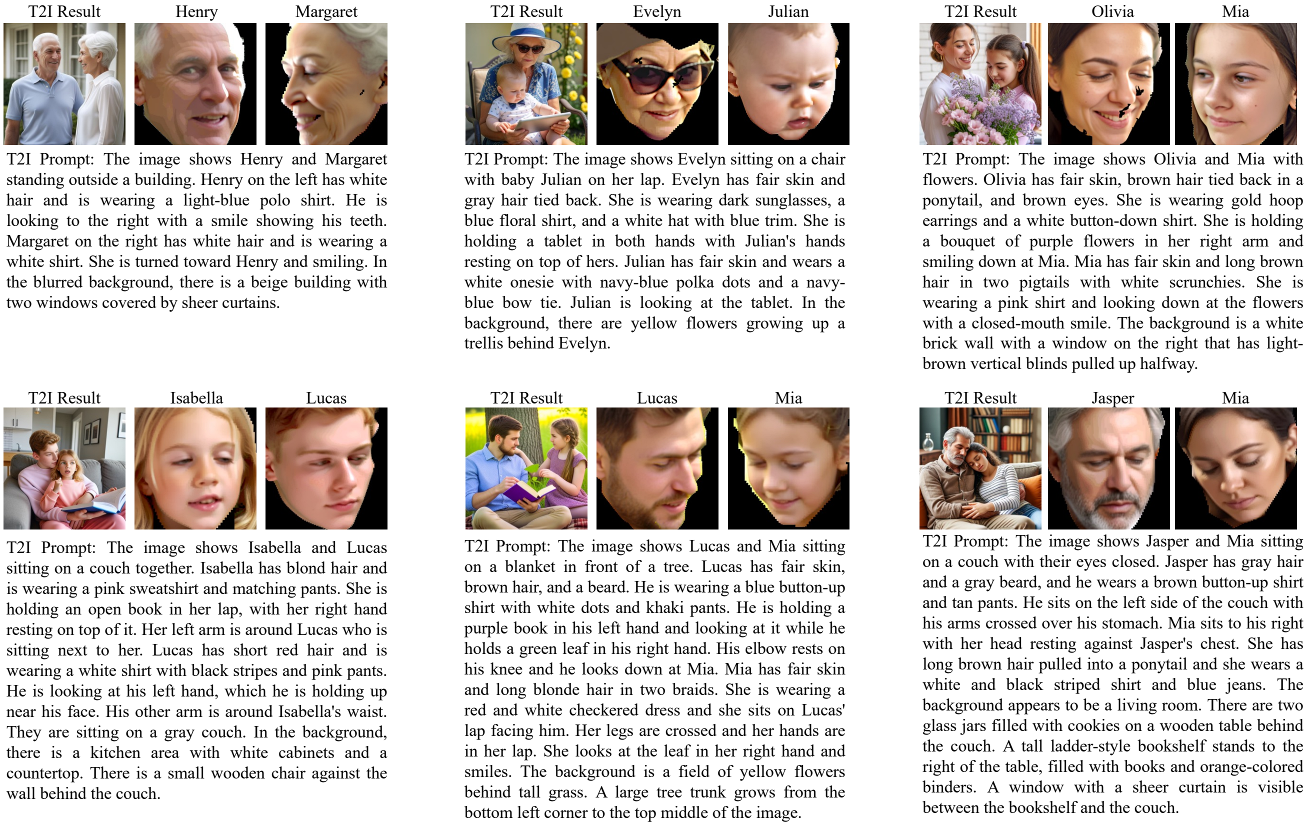
Figure 12. Comparison of DiT Detokenizer Results. "256 token" indicates the absence of 1D pooling after Qwen-VL. Notably, 1) the 256 token configuration demonstrates superior face reconstruction capabilities, but its integration into our pipeline would necessitate re-pretraining the LLaMA component. 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."



T2I Prompt: The image shows Henry and Margaret standing outside a building. Henry on the left has white hair and is wearing a light-blue polo shirt. He is looking to the right with a smile showing his teeth. Margaret on the right has white hair and is wearing a white shirt. She is turned toward Henry and smiling. In the blurred background, there is a beige building with two windows covered by sheer curtains.

T2I Prompt: The image shows Evelyn sitting on a chair with baby Julian on her lap. Evelyn has fair skin and gray hair tied back. She is wearing dark sunglasses, a blue floral shirt, and a white hat with blue trim. She is holding a tablet in both hands with Julian's hands resting on top of hers. Julian has fair skin and wears a white onesie with navy-blue polka dots and a navy-blue bow tie. Julian is looking at the tablet. In the background, there are yellow flowers growing up a trellis behind Evelyn.

T2I Prompt: The image shows Olivia and Mia with flowers. Olivia has fair skin, brown hair tied back in a ponytail, and brown eyes. She is wearing gold hoop earrings and a white button-down shirt. She is holding a bouquet of purple flowers in her right arm and smiling down at Mia. Mia has fair skin and long brown hair in two pigtails with white scrunchies. She is wearing a pink shirt and looking down at the flowers with a closed-mouth smile. The background is a white brick wall with a window on the right that has light-brown vertical blinds pulled up halfway.

T2I Prompt: The image shows Isabella and Lucas sitting on a couch together. Isabella has blond hair and is wearing a pink sweatshirt and matching pants. She is holding an open book in her lap, with her right hand resting on top of it. Her left arm is around Lucas who is sitting next to her. Lucas has short red hair and is wearing a white shirt with black stripes and pink pants. He is looking at his left hand, which he is holding up near his face. His other arm is around Isabella's waist. They are sitting on a gray couch. In the background, there is a kitchen area with white cabinets and a countertop. There is a small wooden chair against the wall behind the couch.

T2I Prompt: The image shows Lucas and Mia sitting on a blanket in front of a tree. Lucas has fair skin, brown hair, and a beard. He is wearing a blue button-up shirt with white dots and khaki pants. He is holding a purple book in his left hand and looking at it while he holds a green leaf in his right hand. His elbow rests on his knee and he looks down at Mia. Mia has fair skin and long blonde hair in two braids. She is wearing a red and white checkered dress and she sits on Lucas' lap facing him. Her legs are crossed and her hands are in her lap. She looks at the leaf in her right hand and smiles. The background is a field of yellow flowers behind tall grass. A large tree trunk grows from the bottom left corner to the top middle of the image.

T2I Prompt: The image shows Jasper and Mia sitting on a couch with their eyes closed. Jasper has gray hair and a gray beard, and he wears a brown button-up shirt and tan pants. He sits on the left side of the couch with his arms crossed over his stomach. Mia sits to his right with her head resting against Jasper's chest. She has long brown hair pulled into a ponytail and she wears a white and black striped shirt and blue jeans. The background appears to be a living room. There are two glass jars filled with cookies on a wooden table behind the couch. A tall ladder-style bookshelf stands to the right of the table, filled with books and orange-colored binders. A window with a sheer curtain is visible between the bookshelf and the couch.

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

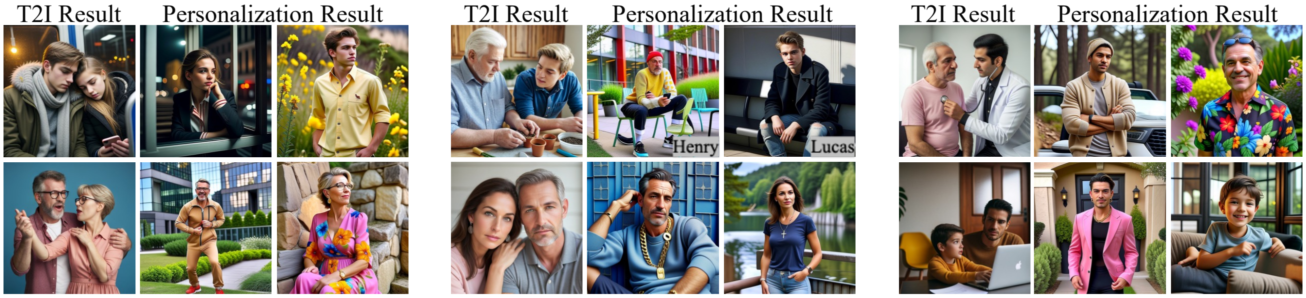


Figure 14. More examples of full-body multi-turn personalization results.

- **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.

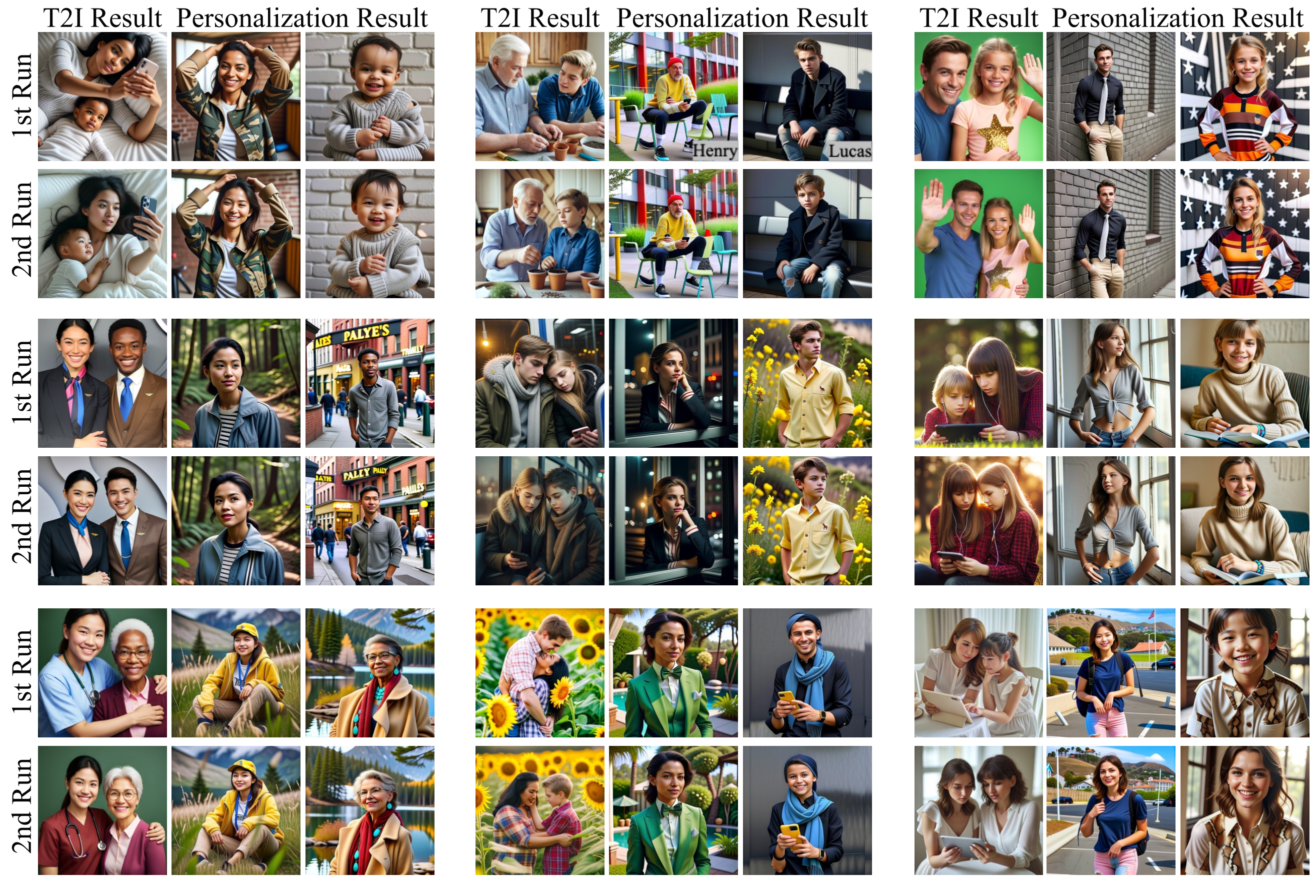


Figure 15. Examples of multi-turn personalization results. As can be observed, 1) This model can follow the prompt well, but struggles to maintain consistent face identity. Please see Section 8 last paragraph for discussion. 2) When inferring twice with the same prompt, distinct 1st-turn T2I results 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

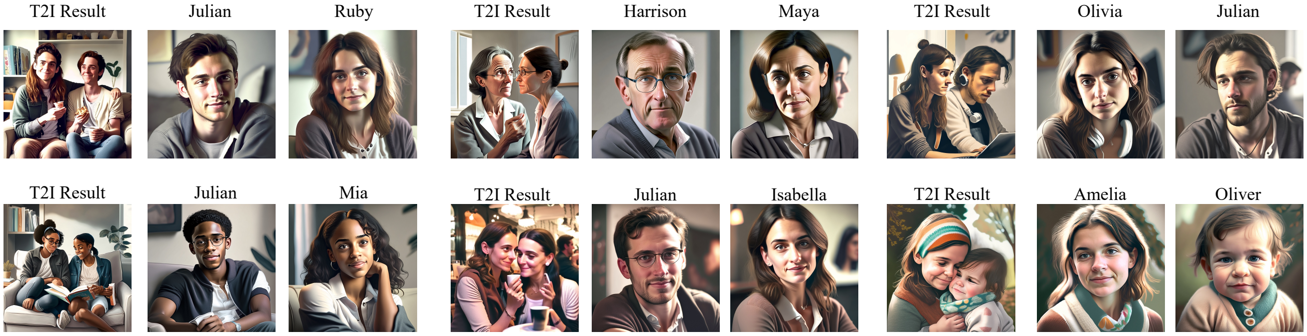


Figure 16. Multi-turn personalization results of SEED-X without fine-tuning on our proposed multi-turn dataset. These results can be directly compared to Figure 7. As can be observed, firstly, this baseline model has difficuties in generate a reasonable two-person images in the first round; Then in personalization 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.