# 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] Electrical and Computer Engineering Department, UC San Diego, USA
Email: haz035@ucsd.edu, zcheustc@gmail.com

WACV 2026 TUCSON, AZ 3/6 – 3/10

## Introduction

**Conversational image generation is non-Markovian**: many user requests depend on visual states or bindings introduced several turns earlier.

**Existing benchmarks and models exploit a Markov shortcut**, leading to systematic failures in multi-round editing and identity consistency [1, 2].

**We address this with non-Markov datasets and a history-conditioned generative framework**

- Non-Markov multi-round data:
  - Rollback-style editing dataset
  - Name-based personalization dataset
- token-level caching
- high-fidelity detokenization
- staged personalization training

## Non-Markov Multi-round Image Generation

We represent a conversation up to turn $t$ as an interleaved history

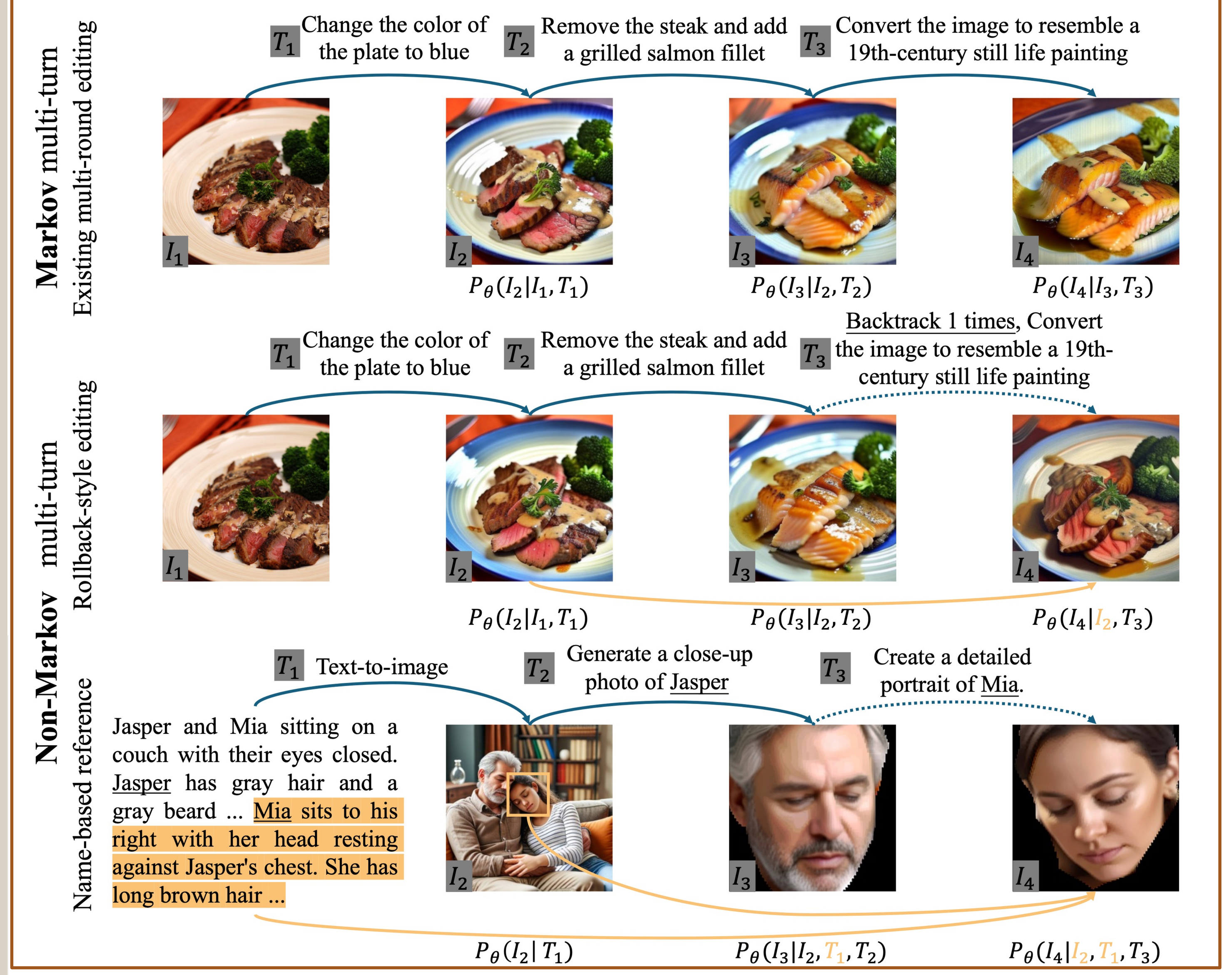
$$H_t = \{(T_1, I_1), (T_2, I_2), \ldots, (T_t, I_t)\}$$

where $T_i$ is the user instruction at round $i$ and $I_i$ is the model-generated image at that round. Given $H_t$ and a new instruction $T_{i+1}$, model generates next image:

$$I_{t+1} \sim p_\theta(I \mid T_{t+1}, H_t)$$

Many existing multi-turn editing benchmarks [1, 2] are *effectively Markov*

$$p_\theta(I \mid T_{t+1}, H_t) \approx p_\theta(I \mid T_{t+1}, I_t)$$




## Non-Markov Dataset Construction

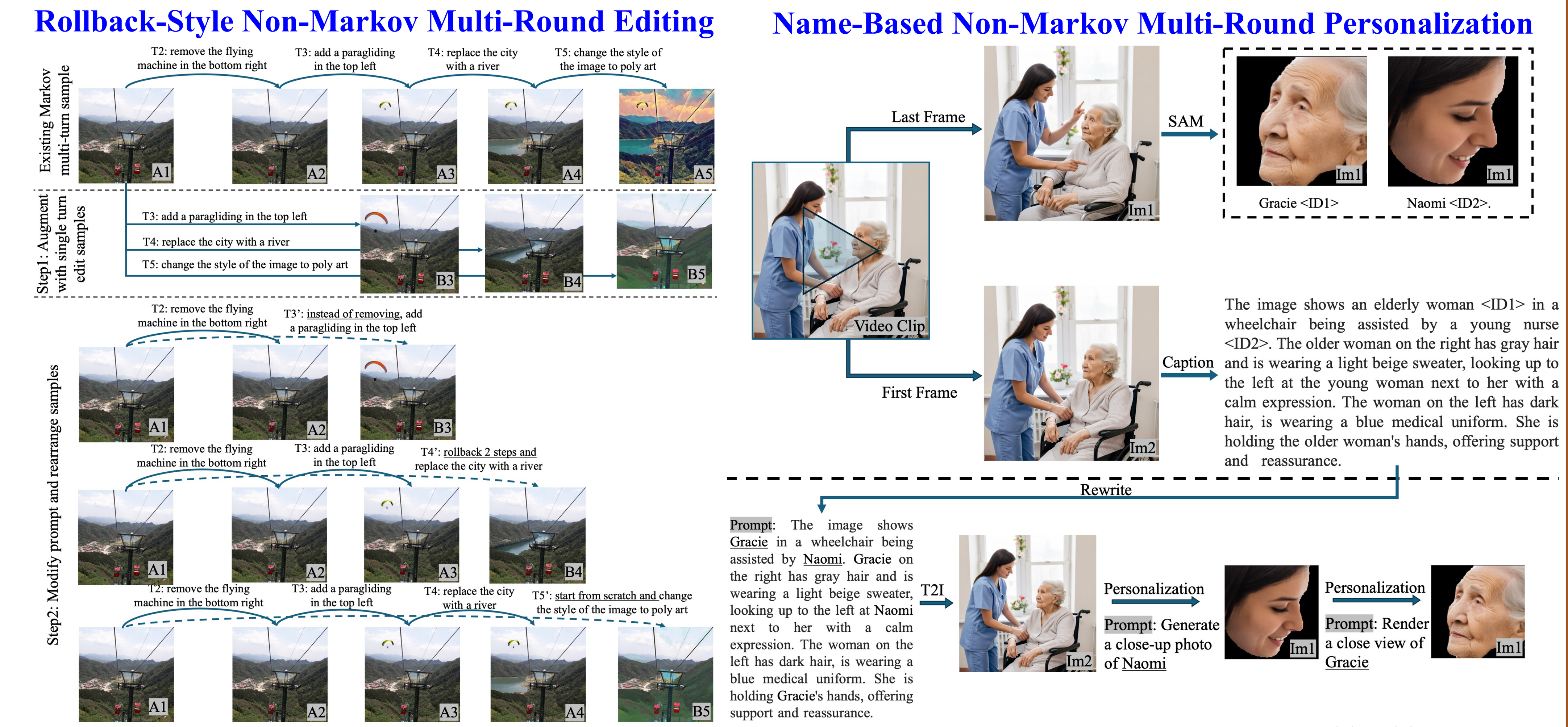
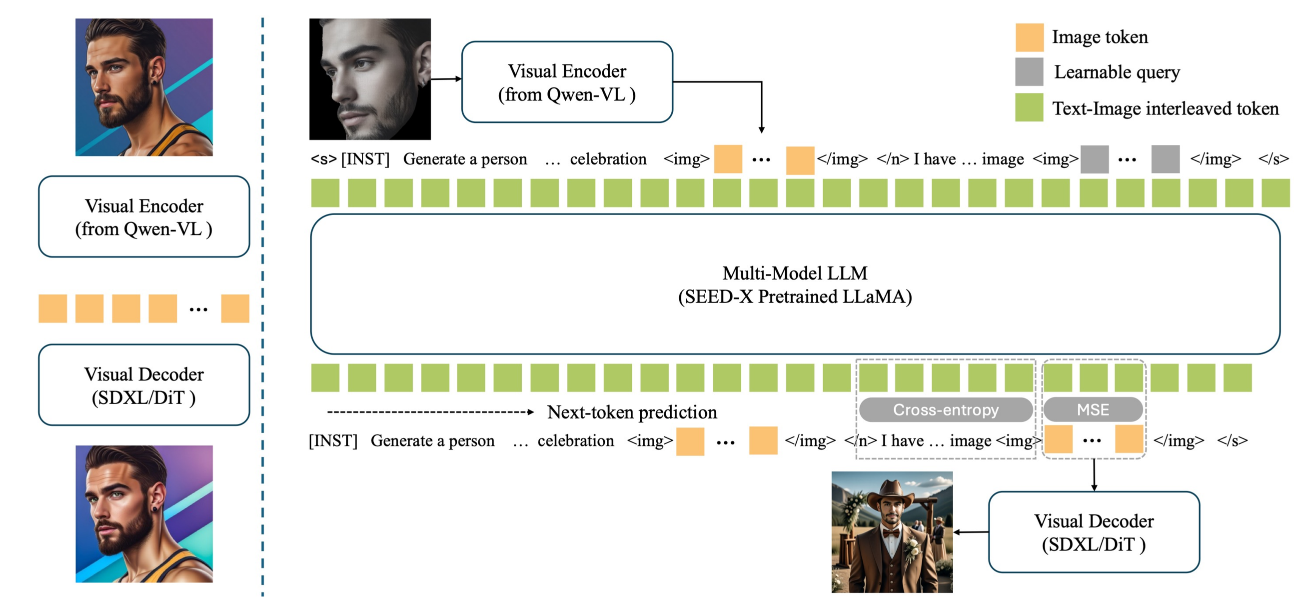



Image-level: $p_\theta(I \mid T_{t+1}, H_t) \approx p_\theta(I \mid T_{t+1}, I_{base})$ token-level: $p_\theta(I_3 \mid T_3, H_2) \approx p_\theta(I_3 \mid T_3, T_1^{(e)}, I_1^{(e)})$

## MLLM and Enabling Components



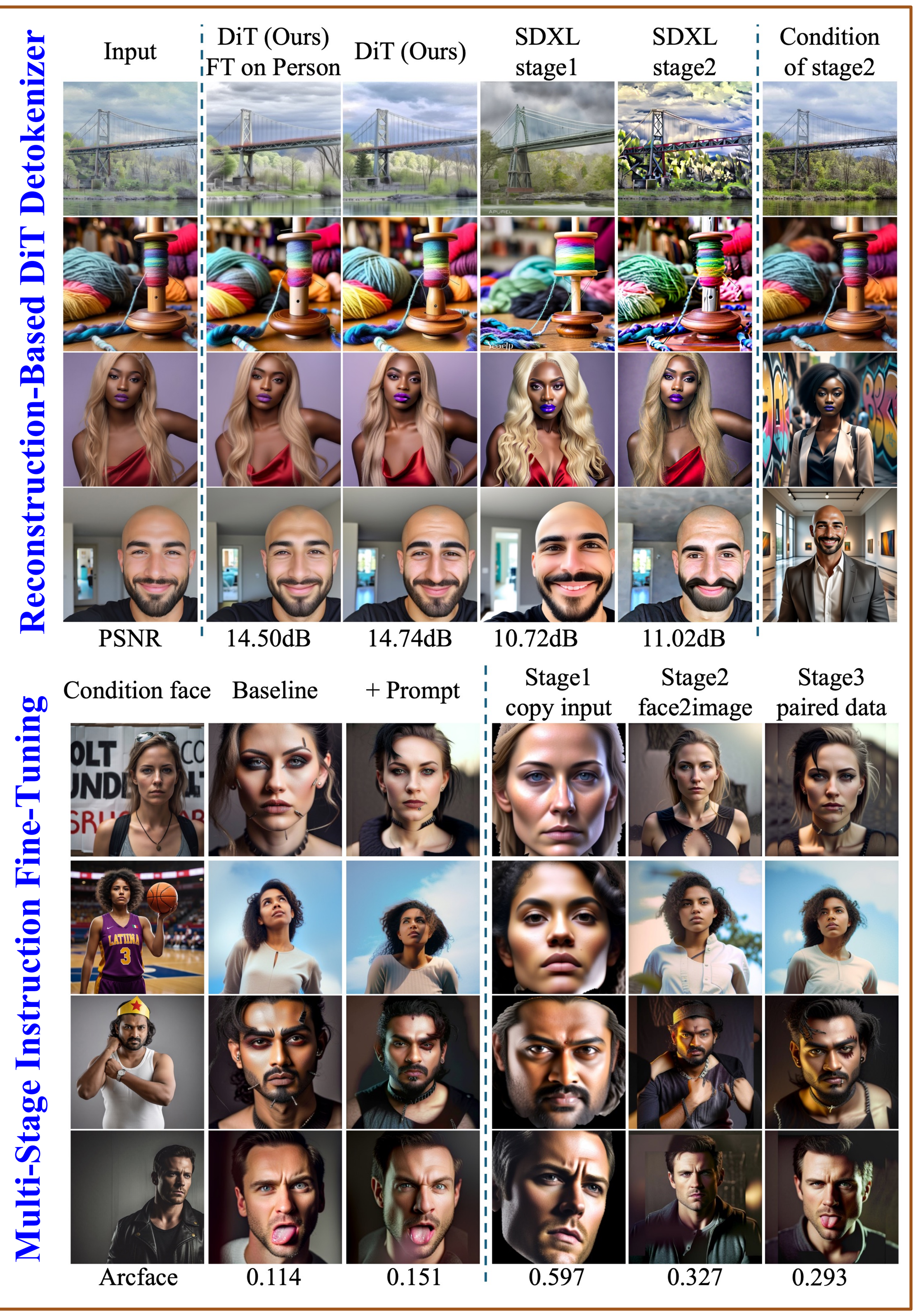

**SEED-X [1] Framework**

- It treats images as a visual language via a tokenizer [3] – detokenizer [4] interface.
- However, vanilla framework suffers from reconstruction drift and identity degradation.

**Token-Level Caching** $H_t = \{(T_1, \hat{I}_1), (T_2, \hat{I}_2), \ldots, (T_t, \hat{I}_t)\}$ not $\Phi(I_t)$

At round $t$, the model produces image tokens $\hat{I}_t$ (before detokenization)

We cache $\hat{I}_t$ and reuse it as the history representation for later rounds




Multi-Stage Instruction Fine-Tuning:

- Stage1: Who it is
- Stage2: Where it is
- Stage3: Who it remains

## Non-Markov Multi-Round Experimental Results

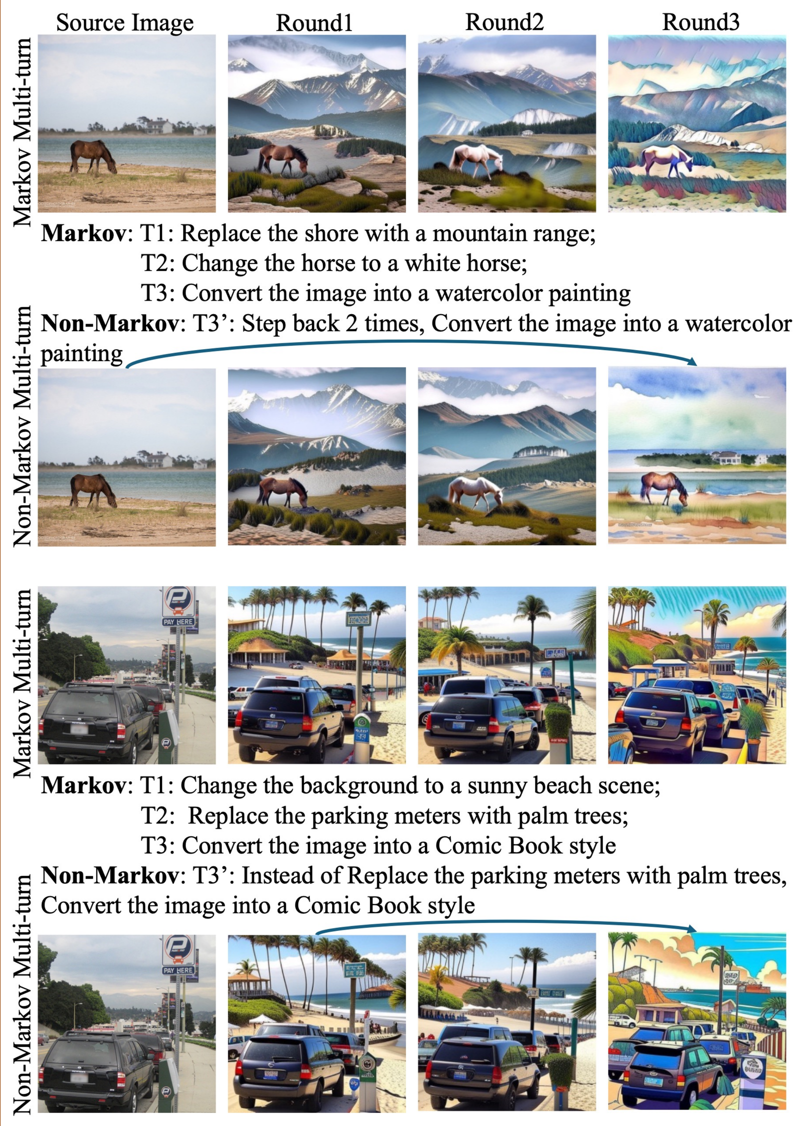
(Left: Multi-turn editing results) Finetuned on our dataset, MLLM supports both Markov and non-Markov editing ✅

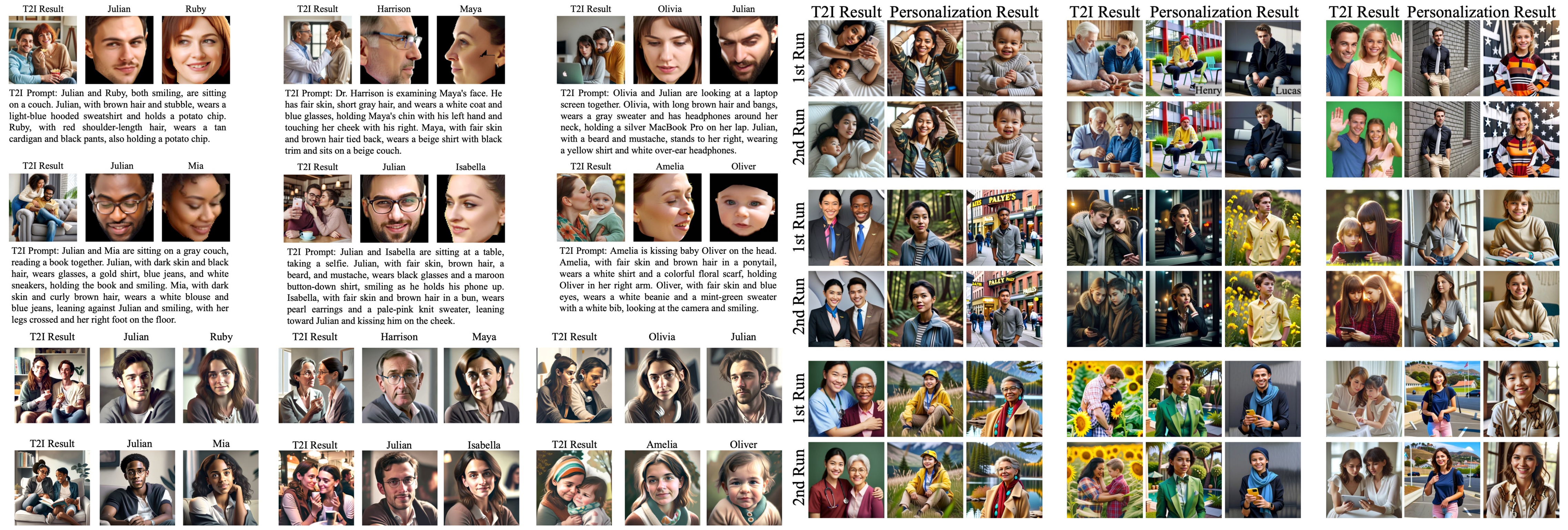
(Middle: Multi-turn personalization results, ours vs baseline) → Ours: Correctly generates two distinct individuals and preserves name–identity bindings across rounds. ✅ SEED-X: Fails at two-person generation and defaults to text-to-image behavior, ignoring prior visual context.

(Right: Further explorations on full-body multi-turn personalization.)

**Full-body personalization**: Fine-tuning enables multi-round full-body generation, though face fidelity is lower than in close-up portraits. This is due to noisier, diffusion-synthesized supervision and a smaller dataset.

**Non-Markov evidence**: When early rounds leave attributes implicit (e.g., age), the model infers a plausible identity and consistently preserves it across later rounds—even with identical text prompts. ✅




## Conclusions

**In summary**, conversational image generation requires explicit history conditioning to resolve rollback edits and identity references across multiple turns. Our non-Markov datasets and history-aware modeling pipeline can improve multi-round editing reliability and personalization consistency.

**Future Directions:**

- Standardized non-Markov conversational benchmarks for image generation.
- More diverse and complex non-Markov multi-round datasets.
- Memory mechanisms for long-horizon dialogue and visual state retrieval.
- Improved image token- and detokenization for stable long-term generation.

## References

1. Y. Ge, S. Zhao, J. Zhu, Y. Ge, K. Yi, L. Song, C. Li, X. Ding, and Y. Shan, "SEED-X: Multimodal models with unified multi-granularity comprehension and generation," arXiv preprint arXiv:2404.14396, 2024.
2. K. Zhang, L. Mo, W. Chen, H. Sun, and Y. Su, "Magicbrush: A manually annotated dataset for instruction-guided image editing," in NIPS, 2023.
3. J. Bai, S. Bai, S. Yang, S. Wang, S. Tan, P. Wang, J. Lin, C. Zhou et al., "Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond," 2023.
4. D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. M¨ uller, J. Penna, and R. Rombach, "SDXL: Improving latent diffusion models for high-resolution image synthesis," arXiv preprint arXiv:2307.01952, 2023.

## Prompt for full-body personalization

- Round1: 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: 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.
- Round3: 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.

**Please also refer to our extended arXiv version:**