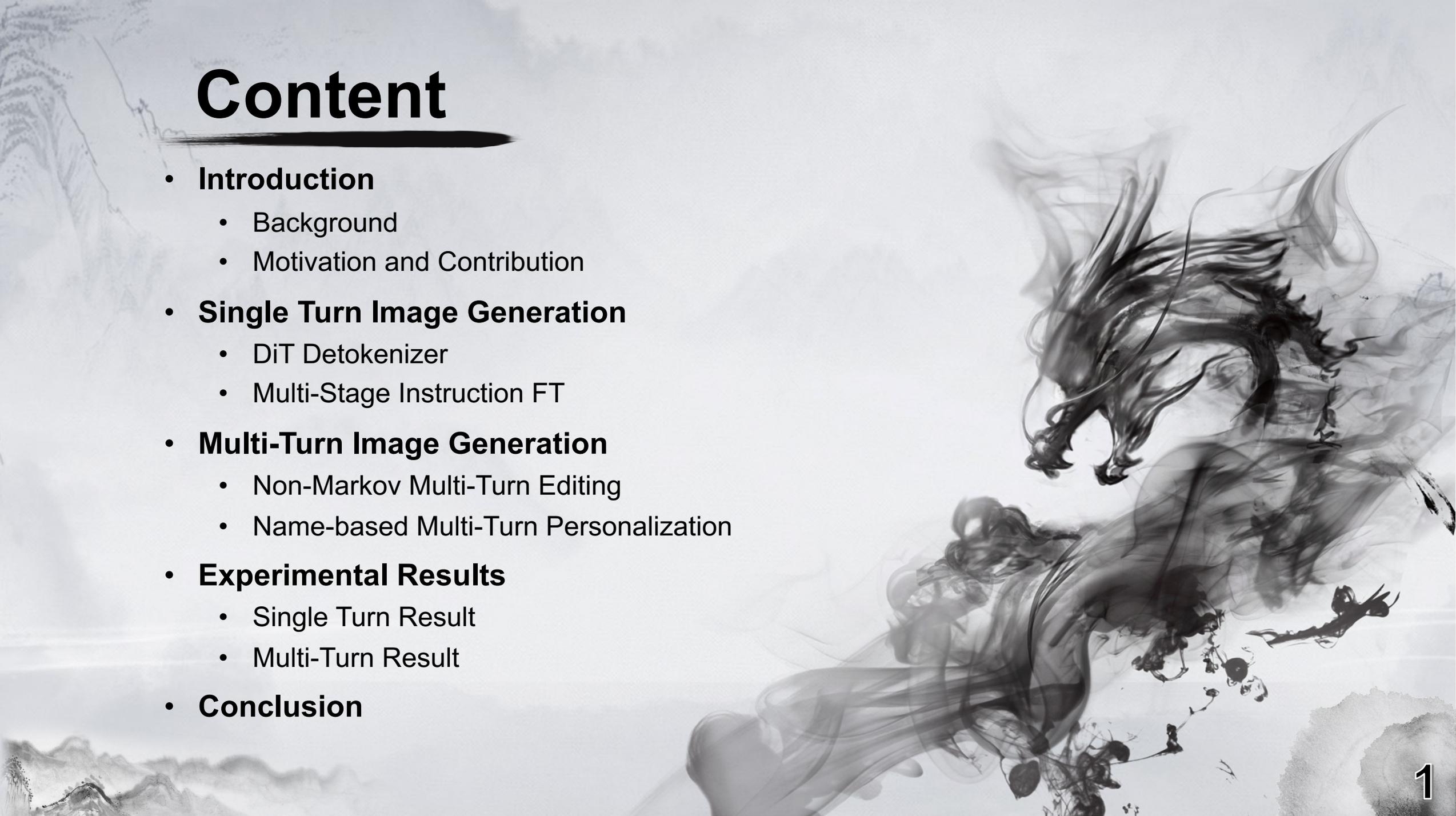




Conversational Image Generation: Towards Multi-Round Image Generation with Multi-Modal LLM

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

Content

The background of the slide features a traditional Chinese ink wash painting. It depicts a person riding a horse through a misty, mountainous landscape. The style is expressive, with fluid lines and varying shades of grey and black ink. The horse is in a dynamic, galloping pose, and the rider is leaning forward. The background shows soft, misty mountains and a hazy sky.

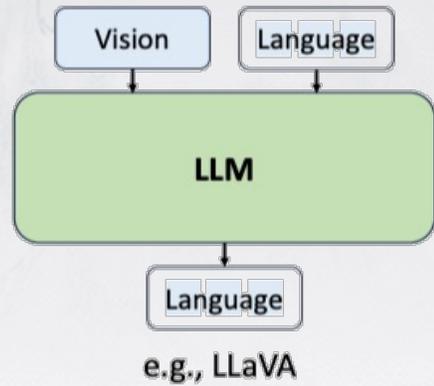
- **Introduction**
 - Background
 - Motivation and Contribution
- **Single Turn Image Generation**
 - DiT Detokenizer
 - Multi-Stage Instruction FT
- **Multi-Turn Image Generation**
 - Non-Markov Multi-Turn Editing
 - Name-based Multi-Turn Personalization
- **Experimental Results**
 - Single Turn Result
 - Multi-Turn Result
- **Conclusion**

1

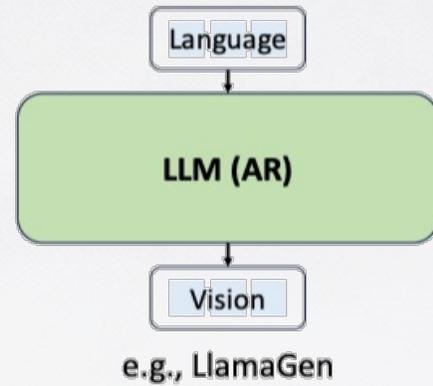
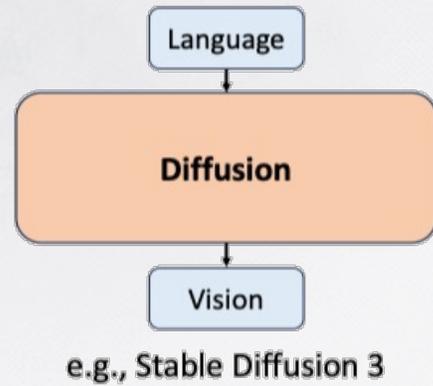
Background LLM and Diffusion model



(a) Understanding Only

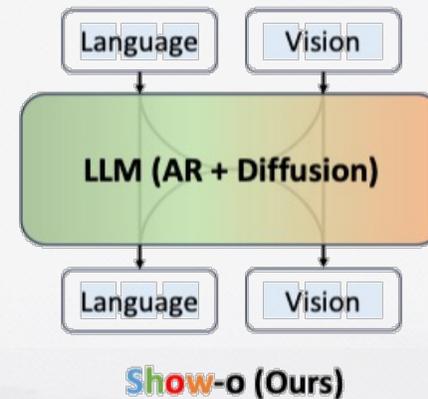
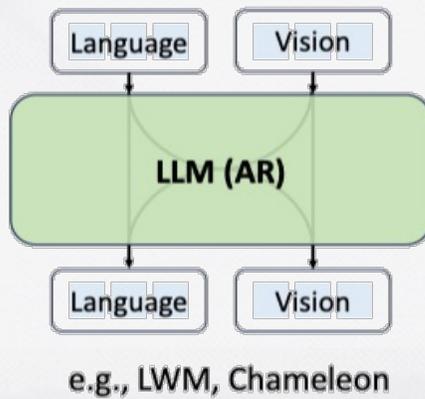
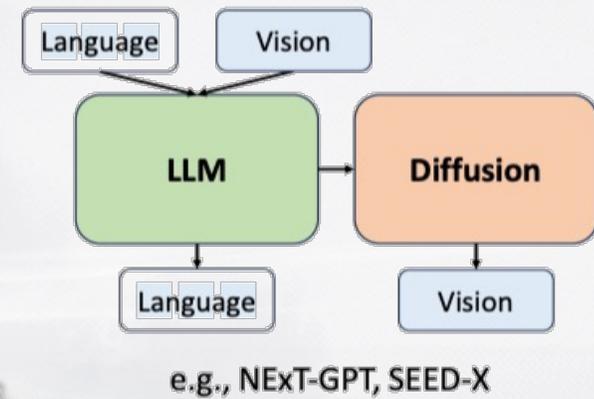


(b) Generation Only



Continuous
Discrete
AR: Autoregressive

(c) Unified Model (Understanding & Generation)



- Text-based LLM: Chatgpt
- Vision LLM: LLaVA
- Multimodal LLM: SEED-X

- Diffusion model can also take vision as input, such as
 - ControlNet
 - T2I-Adapter

1

Background Capability of Diffusion model

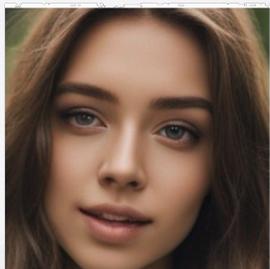


Image Editing



Realistic Model

Subject-Driven Generation
(Personalization)



+

*a beautiful girl
wearing casual
shirt in a garden*



Anime Model

1

Background

Capability of MLLM

Text-to-image
Editing
Personalization?
Multi-turn?



I want to build a two-story cabin in the woods, with many commanding windows. Can you help me draw a picture to show a visualization of it?

Absolutely, I'd be happy to help you visualize your two-story cabin in the woods. Here's a picture.



I love reading books. Can you add a bookshelf in the bedroom?

Absolutely, here is a picture.



Please explain your design.

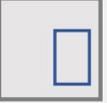
My design is a two-story cabin nestled in the heart of a dense forest. The cabin is constructed from **rustic wood**, giving it a warm and inviting appearance. It features a **wrap-around porch** on the ground floor, perfect for enjoying the fresh air and the sounds of nature. The upper floor is adorned with **multiple windows**, offering breathtaking views of the surrounding trees and wildlife. The roof is **covered in shingles**, providing a sturdy and durable protection against the elements. The overall design of the cabin is simple yet elegant, blending seamlessly with its natural surroundings.

Show me the location of the bookshelf.



SEED-X

SEED-X: Multimodal Models with Unified Multi-granularity Comprehension and Generation
Generative Multimodal Models are In-Context Learners

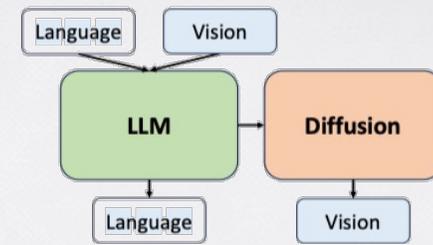
cute baby panda			Make it wearing a very cool sunglasses	
<a panda>			on the beach	
<a panda>			and <a dog>	
			on the beach	
An oil painting of two animals			, in the style of Van Gogh	

EMU-2

1

Background

Single-turn MLLM $p(x_a^i | X_v^i, X_t^i)$



Editing template

```
prompt = [image, "wearing a red hat on the beach."]
```



Grounded image gen template

```
prompt = [  
  "<grounding>",  
  "An oil painting of three dogs,",  
  "<phrase>the first dog</phrase>",  
  "<object>",  
  dog1_mask,  
  "</object>",  
  dog1,  
  "<phrase>the second dog</phrase>",  
  "<object>",  
  dog2_mask,  
  "</object>",  
  dog2,  
  "<phrase>the third dog</phrase>",  
  "<object>",  
  dog3_mask,  
  "</object>",  
  dog3,  
]
```

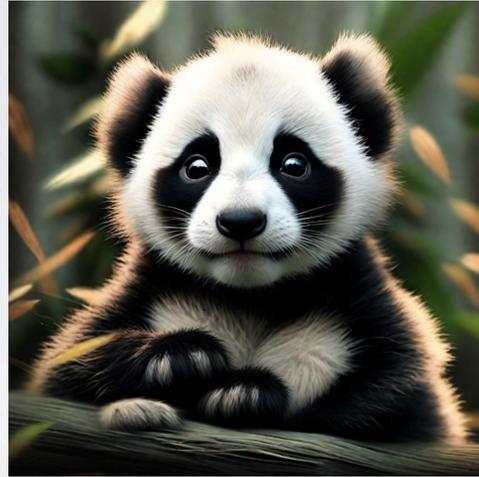


1

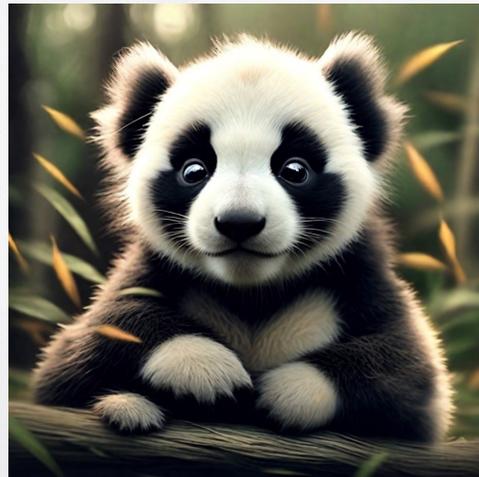
Background

Multi-turn MLLM

["cute baby panda"]



["cute baby panda"]



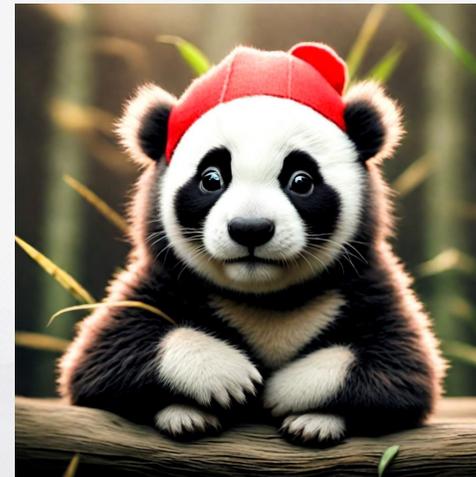
Multi-turn

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

['make', , 'wearing a red hat']



['cute baby panda', , 'make it wearing a red hat.']



1

Background Multi-turn MLLM

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



["<A panda> [bbox, ] on the beach"]

X_v^3 / X_a^2

X_t^3



x_a^3

Multiple Single-turn

X_t^1

X_a^1

['cute baby panda',



, 'make it wearing a red hat.',



X_a^2

X_t^3

, 'make it on the beach.']

X_t^2



x_a^3

Multi-turn

1

Background Multi-turn MLLM

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



["<A panda> [bbox, ] and <A dog> [bbox, ] on the beach"]



Multiple Single-turn

X_t^1 ['cute baby panda',  X_a^1], 'make it wearing a red hat.',  X_t^2 X_a^2 , 'make it on the beach.',  X_t^3 X_a^3 ,
'Add <A dog> [bbox, ] on it.'] X_v^4 X_a^4

Multi-turn



x_a^4

1

Background Multi-turn MLLM

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



["An oil painting of two animals  in the style of Van Gogh."]



Multiple Single-turn

X_t^1

['cute baby panda',



X_t^2

, 'make it wearing a red hat.',



X_a^2

X_t^3

, 'make it on the beach.',



X_a^3

Multi-turn



'Add <A dog> [bbox,

X_t^4



X_v^4



X_a^4

, 'make it an oil painting in the style of Van Gogh.']

X_t^5

x_a^5

1

Motivation & Contribution



Current model

- Multi-turn image generation ❌
- Multiple single-turn image generation ✅

LLM is known for its ability of reasoning

- Reasoning between text element (phrase) and image element (patch)
 - Image editing
 - Personalization
- Reasoning from chat history for condition selection
 - Non-Markov multi-turn edit
 - Name-based multi-turn personalization

Contributions

Improve the SDXL detokenizer with DiT structure

Multi-stage instruction finetuning

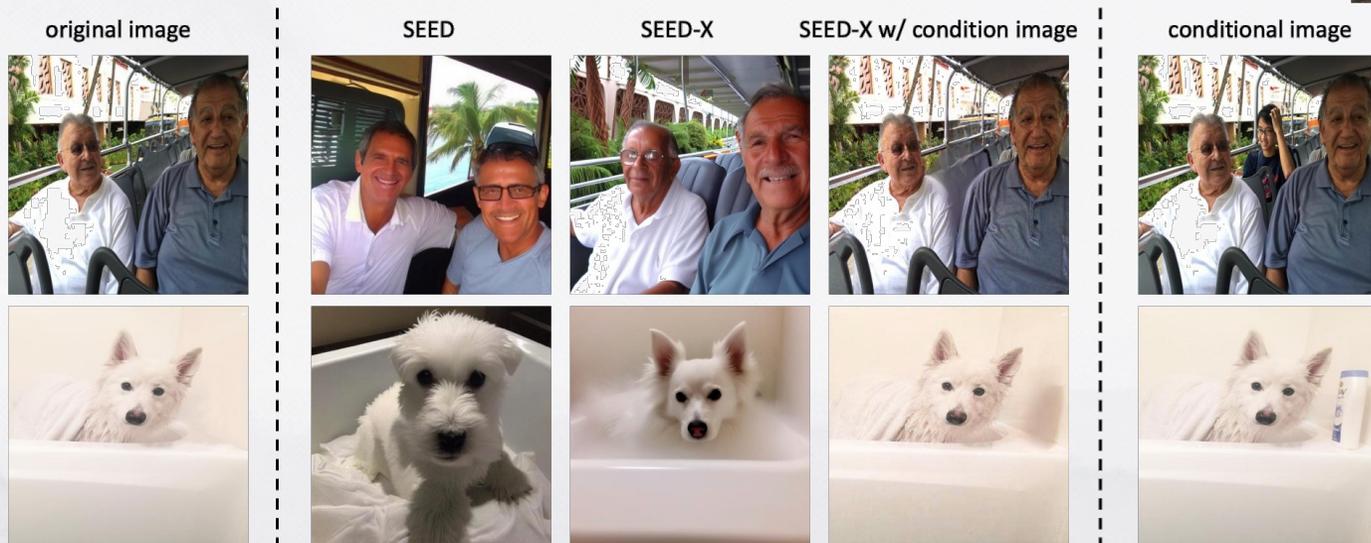
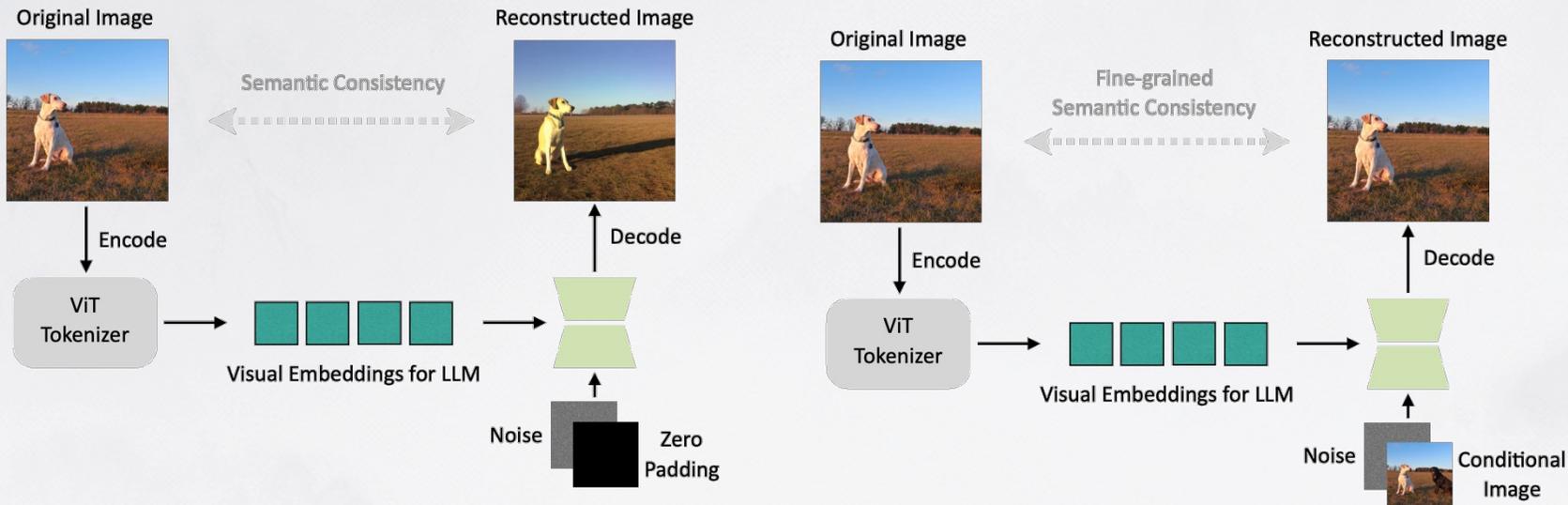
Datasets and training pipeline

② Single Turn Image Generation

- SEED-X backbone
- DiT Detokenizer
- Multi-Stage Instruction FT

2

Preliminaries Seed-X detokenizer



1

Preliminaries

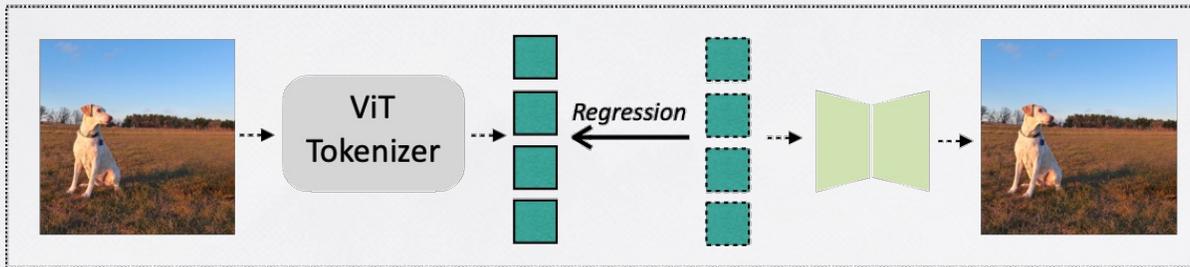
Seed-X LLaMA



Training sample:

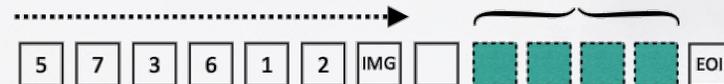
...  Lucky likes playing
in the park  ...

Training

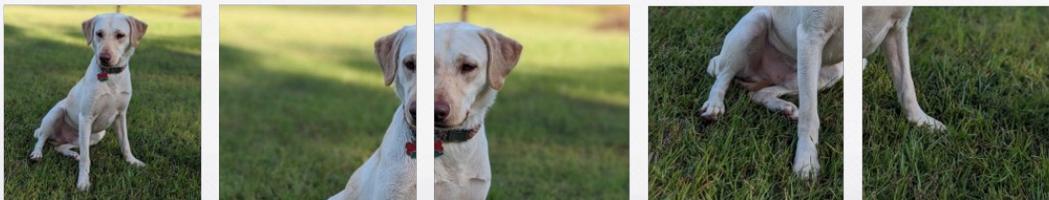


Inference

Next-word prediction



SEED-X: Large Multimodal Model



Lucky likes playing in the park

Learnable queries

Image gridding to support arbitrary sizes and aspect ratios

1

Preliminaries Emu2



Encoder

Encoder



Decoder



<s> [IMG]   ...  [/IMG] An emu egg that will hatch into a baby emu [IMG]   ...  [/IMG]</s>

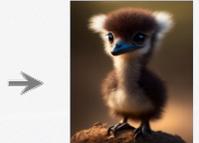


Generative Multimodal Model

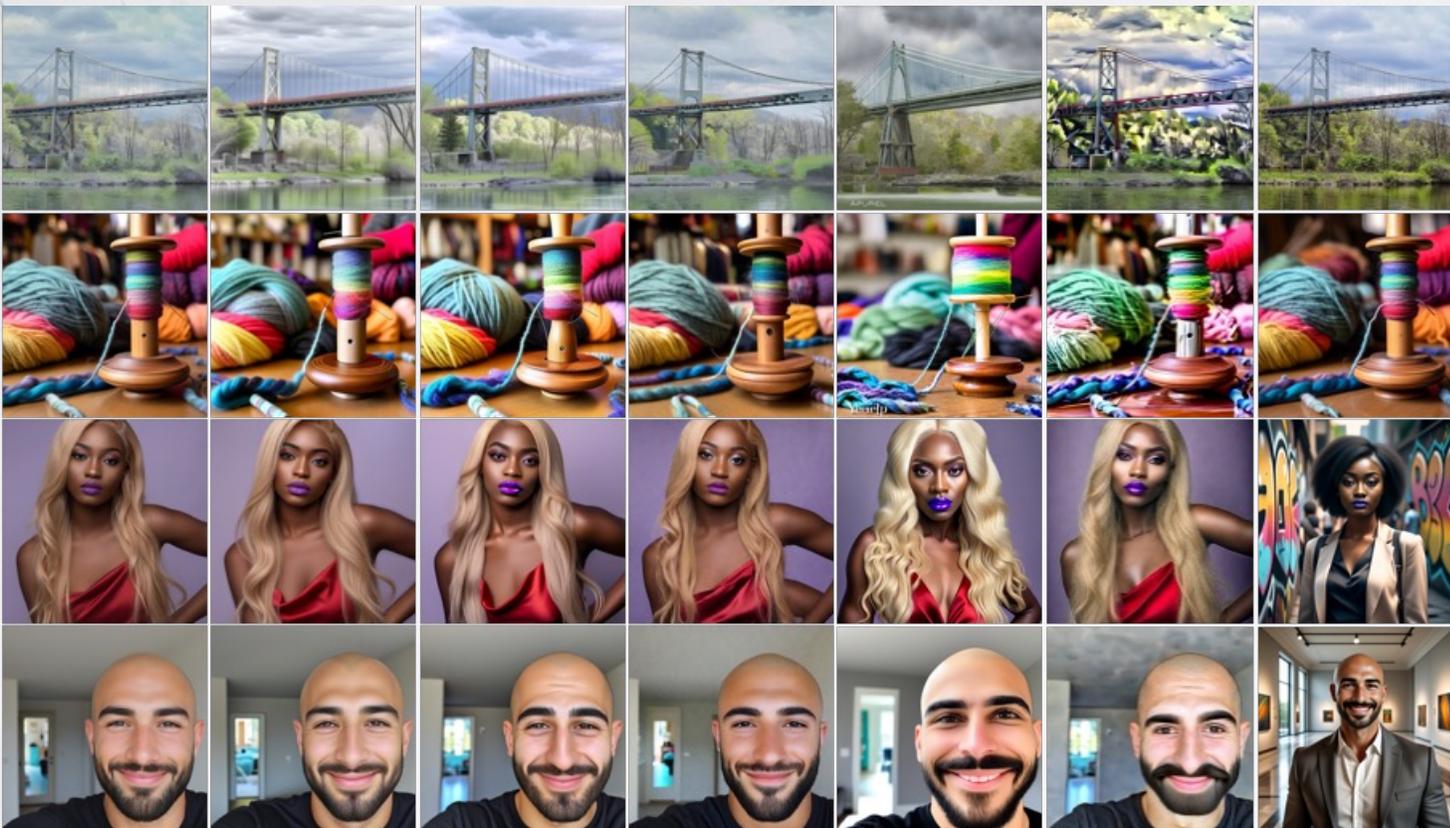


↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ ↓ Classification Regression ↓
[IMG]   ...  An emu egg that will hatch into a baby emu [IMG]   ...  </s>

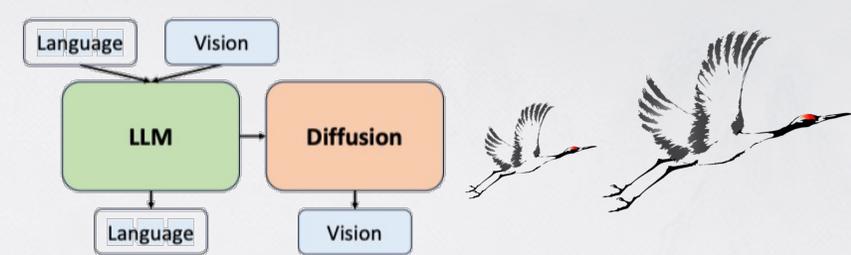
Decoder



② DiT detokenizer

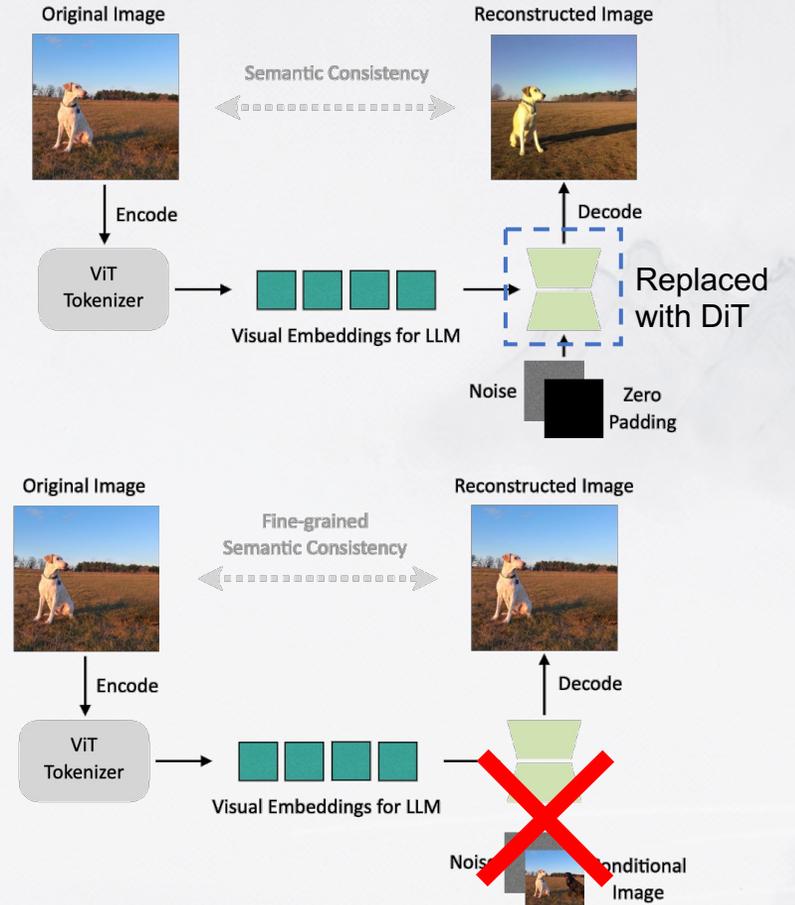
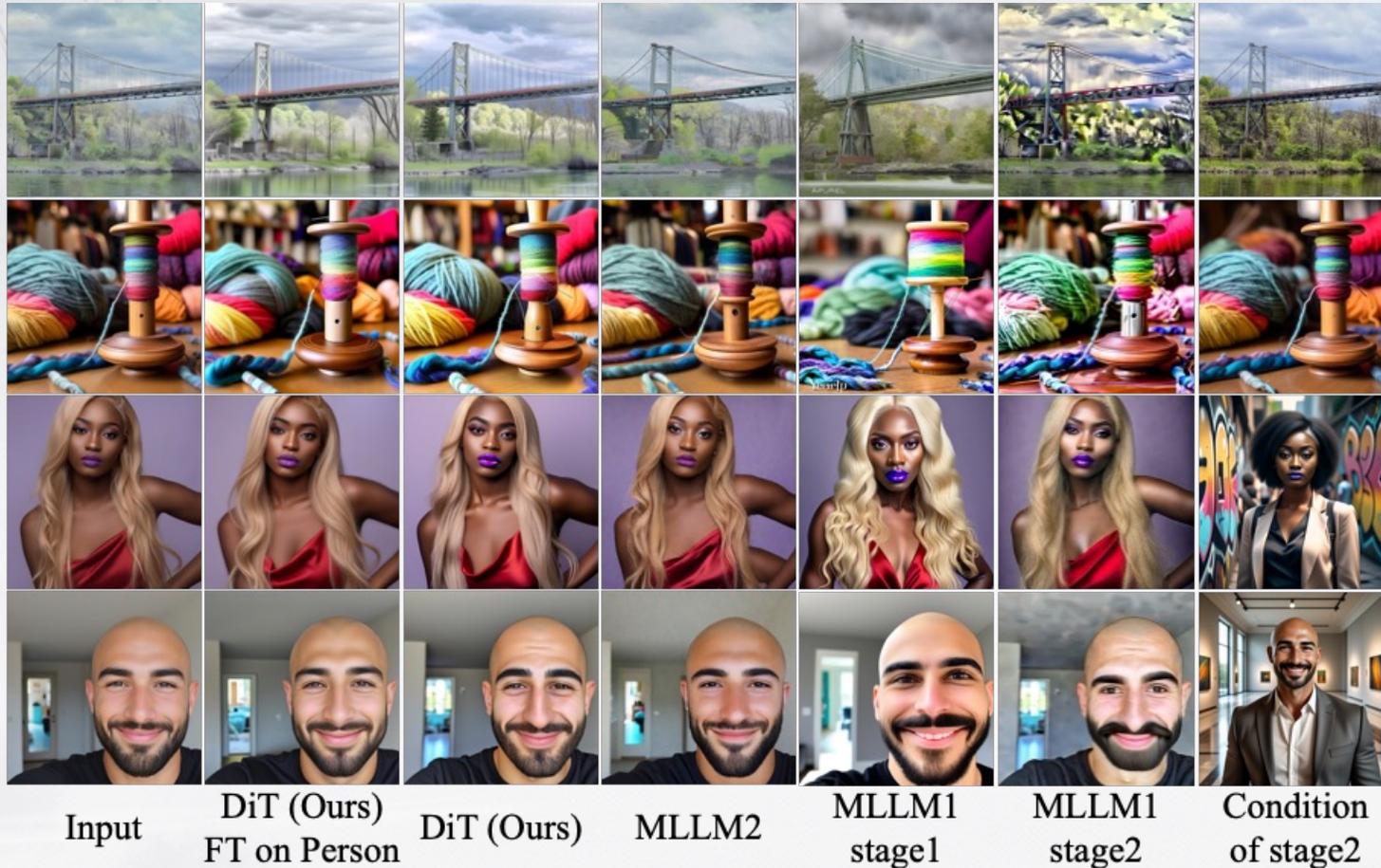


Input DiT (Ours) FT on Person DiT (Ours) MLLM2 MLLM1 stage1 MLLM1 stage2 Condition of stage2



- The stage1 reconstruction-based detokenizers struggle to preserve content accurately, especially for human faces;
- The editing-finetuned detokenizer demonstrates improved performance only in scenarios where the conditional image and input share an editing relationship.
- The stage2 detokenizer also introduce some artifacts if OOD condition images are provided

② DiT detokenizer



Compared to MLLM1 stage1 detokenizer,

- There was a significant improvement in content preservation without additional conditions
- After finetuning on human images enhanced face preservation.

2

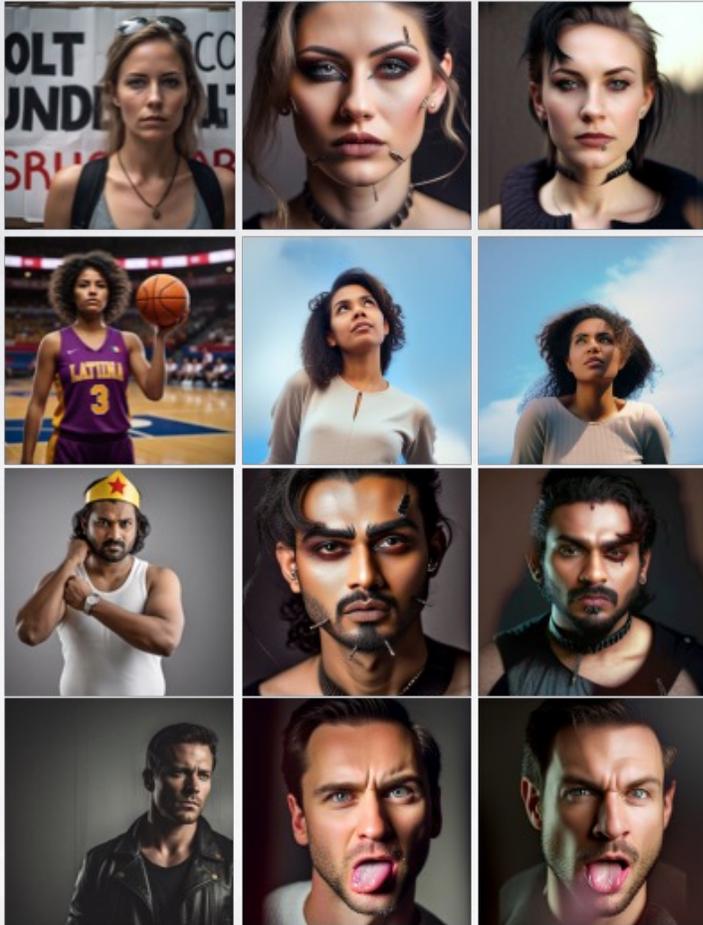
DiT detokenizer



Input	Ours	MLLM1 DiT	MLLM1 Stage2	MLLM1 Stage1	MLLM2
Add the word 'mocha' to the side of the closest cup.					
Remove the word "Stop" from the red sign					
Change "A's" to "B's".					
Add a green bowl on the branch					
Add a fence behind the largest tree.					
Change the British flag into an American flag					

Input	Ours	MLLM1 DiT	MLLM1 Stage2	MLLM1 Stage1	MLLM2
Alter the background so its inside a mall					
Change the image so it looks like the area is rusting away.					
Make background in outer space.					
Change the style of the image into a Monet painting.					
Change the image to a 1969 Scooby-Doo cartoon art style.					
Change the style to that of a children's book.					

② Prompt for Personalization



Condition face Baseline + Prompt

Input:

<s> [INST] Generate {prompt}.



 [/INST]

Output:

I have generated an image.

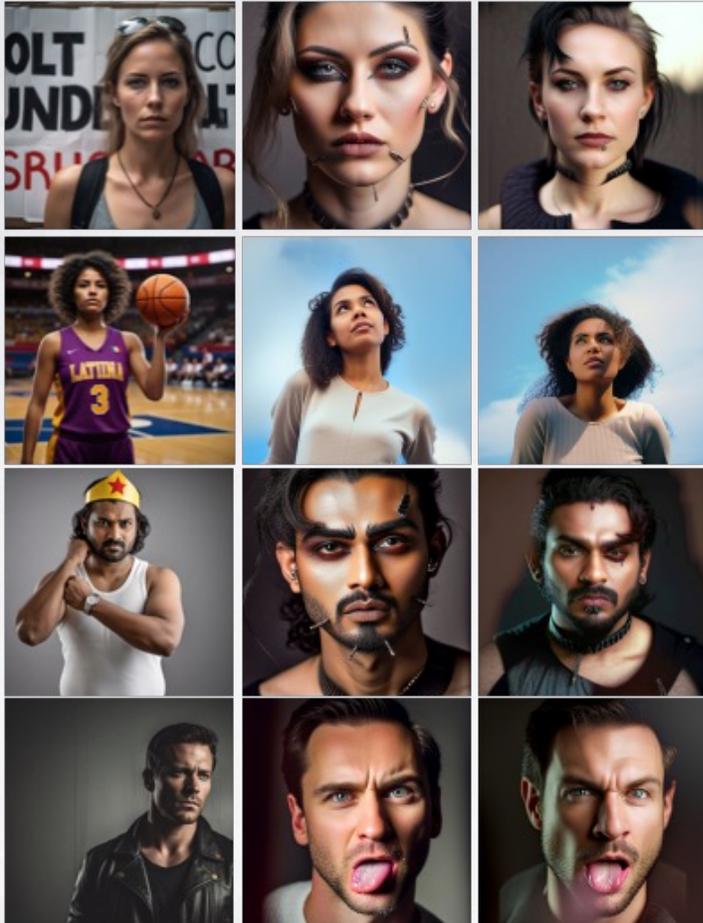


 </s>

{prompt}:

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

② Prompt for Personalization



Condition face Baseline + Prompt

Input:

<s> [INST] Generate {prompt}. Please keep the face identical



 [/INST]

Output:

I have generated an image. I keep the face unchanged

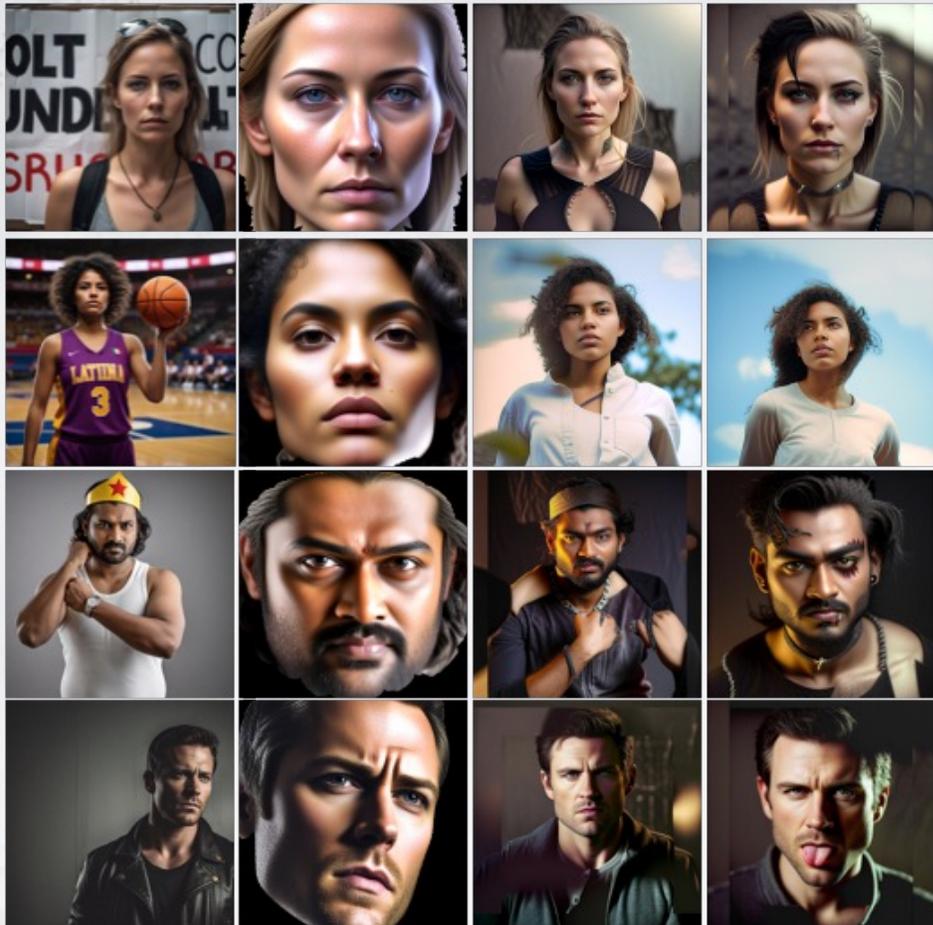
  </s>

{prompt}

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

Multi-stage FT



Tradeoff between *face preservation* and *prompt alignment*

Multi-stage	input	output	Face ID	Alignment
stage1	cropped face	Input itself	best	worst
stage2	cropped face	Image the face cropped from	Good	Worse in face area
stage3	cropped face	Another image with same face	OK	Good

{prompt}

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.

Condition face

Stage1
copy input

Stage2
face2image

Stage3
paired data

2

Personalization result



Condition face

Personalization Results

Condition face

Personalization Results

③ Multi-Turn Image Generation

- Non-Markov Multi-Turn Editing
- Name-based Multi-Turn Personalization

3

Multi-turn Dataset



Existing multi-turn dataset

- Text only: LLaVA, SVIT
- Text-image Interleaved: MMC4, **VIST**
- Multi-turn Editing: MagicBrush, SEED-Data-Edit

1 Story ≠ 5 Captions

	1	2	3	4	5
					
Desc-in-Isolation	A black frisbee is sitting on top of a roof.	A man playing soccer outside of a white house with a red door.	The boy is throwing a soccer ball by the red door.	A soccer ball is over a roof by a frisbee in a rain gutter.	Two balls and a frisbee are on top of a roof.
Desc-in-Sequence	A roof top with a black frisbee laying on the top of the edge of it.	A man is standing in the grass in front of the house kicking a soccer ball.	A man is in the front of the house throwing a soccer ball up.	A blue and white soccer ball and black Frisbee are on the edge of the roof top.	Two soccer balls and a Frisbee are sitting on top of the roof top.
Story-in-Sequence	A discus got stuck up on the roof.	Why not try getting it down with a soccer ball?	Up the soccer ball goes.	It didn't work so we tried a volley ball.	Now the discus, soccer ball, and volleyball are all stuck on the roof.

Visual instruction tuning
SVIT: Scaling up visual instruction tuning
Multimodal c4: An open, billion-scale corpus of images interleaved with text.

3

Multi-turn Dataset



(a) Source Image (b) Target Image in Turn 1 (c) Target Image in Turn 2 (d) Target Image in Turn 3

Multi-turn Editing

Visual storytelling.

Magicbrush: A manually annotated dataset for instruction guided image editing.

SEED-X: Multimodal models with unified multi-granularity comprehension and generation.

Existing multi-turn dataset

- Text only: LLaVA, SVIT
- Text-image Interleaved: MMC4, VIST
- Multi-turn Editing: **MagicBrush**, SEED-Data-Edit

Markov multi-turn edit

Current result only depends on last edit results

Markov multi-turn edit

$$p(x_a^i | X_a^{i-1}, X_t^i)$$

Single turn edit

$$p(x_a^i | X_v^i, X_t^i)$$

Multi-turn edit

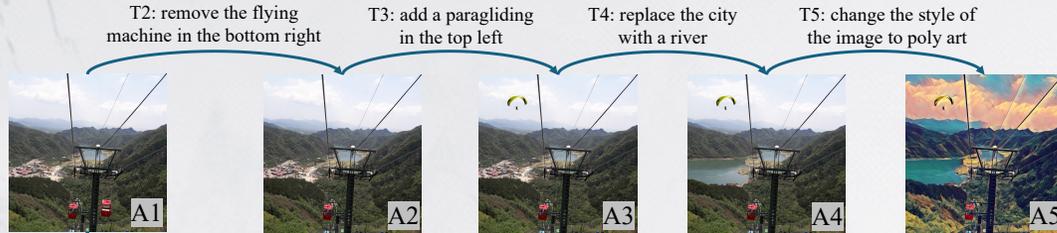
$$p(x_a^i | X_v^i, X_t^i, \{X_v^k, X_t^k, X_a^k\}_{k=1}^{i-1})$$

3

Non-Markov Editing dataset



Existing Markov multi-turn sample



T2: remove the flying machine in the bottom right
 T3: add a paragliding in the top left
 T4: replace the city with a river
 T5: change the style of the image to poly art

$$A_1 \xrightarrow{T_2} A_2 \xrightarrow{T_3} A_3 \xrightarrow{T_4} A_4 \xrightarrow{T_5} A_5$$

Step1: Augment with single turn edit samples



T3: add a paragliding in the top left
 T4: replace the city with a river
 T5: change the style of the image to poly art

$$\{A_1 \xrightarrow{T_i} B_i\}_{i=3}^5$$

Step2: Modify prompt and rearrange samples



T2: remove the flying machine in the bottom right
 T3': instead of removing, add a paragliding in the top left

T2: remove the flying machine in the bottom right
 T3: add a paragliding in the top left
 T4': rollback 2 steps and replace the city with a river

T2: remove the flying machine in the bottom right
 T3: add a paragliding in the top left
 T4: replace the city with a river
 T5': start from scratch and change the style of the image to poly art

$$A_1 \xrightarrow{T_2} A_2 \xrightarrow{T'_3} B_3$$

A_2 depends on A_1 (Markov)

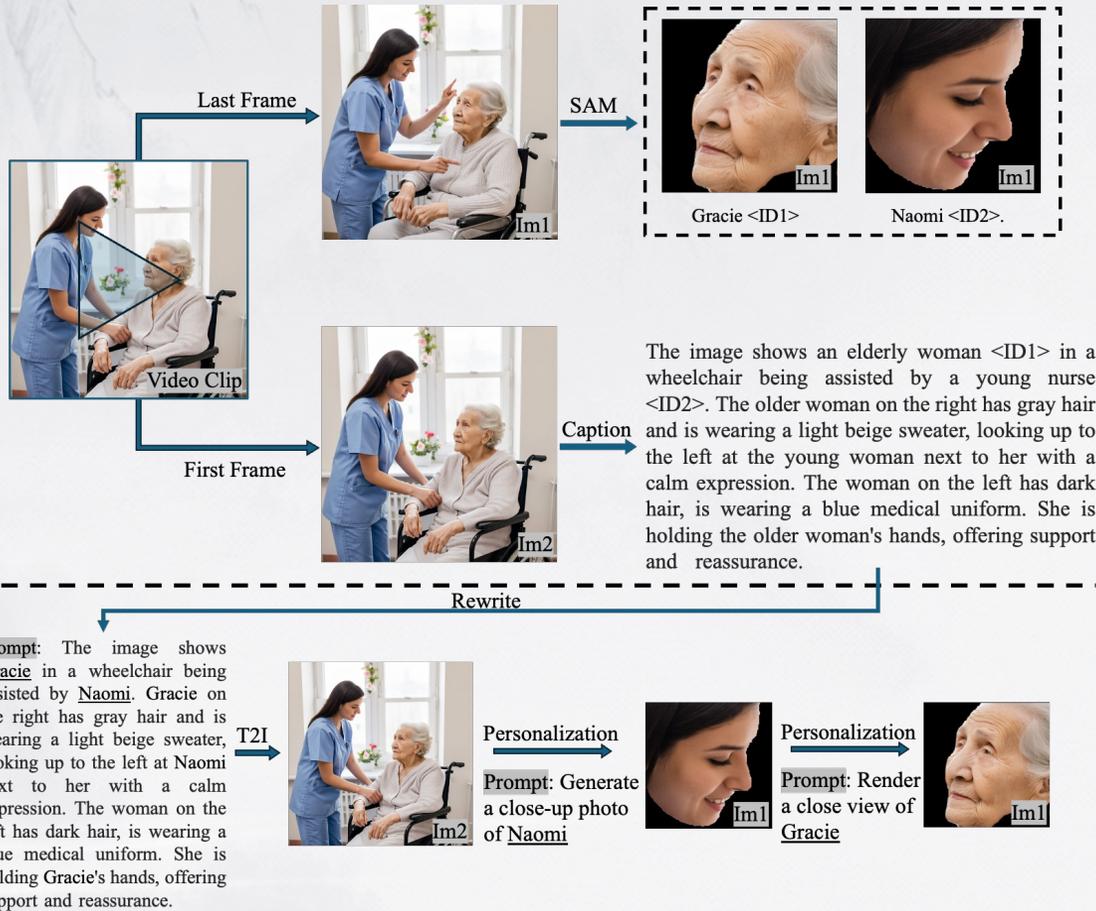
B_3 depends on A_1 instead of A_2 (non – Markov)



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

3

Name-based personalization dataset



Round1

Input: 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]

Output: I have generated an image.

Round2

Input: [Prompt1] + I have generated an image + Generate a close-up photo of Olivia.

Output: Here is the image.

Round3

Input: [Prompt1] + I have generated an image + Here is the image. + Create a detailed portrait of Julian.

Output: Enjoy this photo.

④ Experiment Results

- Single Turn Result
- Multi-Turn Result



4

Single Turn Editing Result



Add the word 'mocha' to the side of the closest cup.



Remove the word "STOP" from the red sign



Change "A's" to "B's".



Add a green bowl on the branch



Add a fence behind the largest tree.



Change the British flag into an American flag



Alter the background so its inside a mall



Change the image so it looks like the area is rusting away.



Make background in outer space.



Change the style of the image into a Monet painting.



Change the image to a 1969 Scooby-Doo cartoon art style.



Change the style to that of a children's book.

4

Single Turn Editing Result



Add the word "EXIT" over the window



Get rid of the largest animal



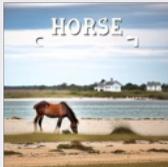
Change the background to a college campus



Recreate the image as a minimalism painting



Change the color of the plate to blue



Add the word "HORSE" to the sky



Add a fence behind the largest tree



Change the background to a construction zone



Recreate the image as an Edgar Degas painting



Alter the car color from red to blue



Add the word "Nutrition" to the top



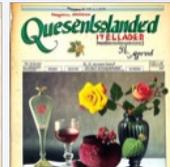
Change the bear to a large dog



Change the background to a corn field in the summer



Change the image to a 1969 Scooby-Doo cartoon art style



Change the color of the roses to yellow



Add USA to the front of the plane



Change the bench into an abandoned couch



Make the background on a cruise ship



Change the image to a Monet art style



Change the bird to all white



Add "BANANAS 39 cent" over the bananas



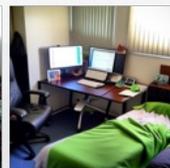
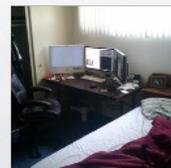
Remove the yellow motorbike from the photo



Make the picture seem like it was taken during the fall



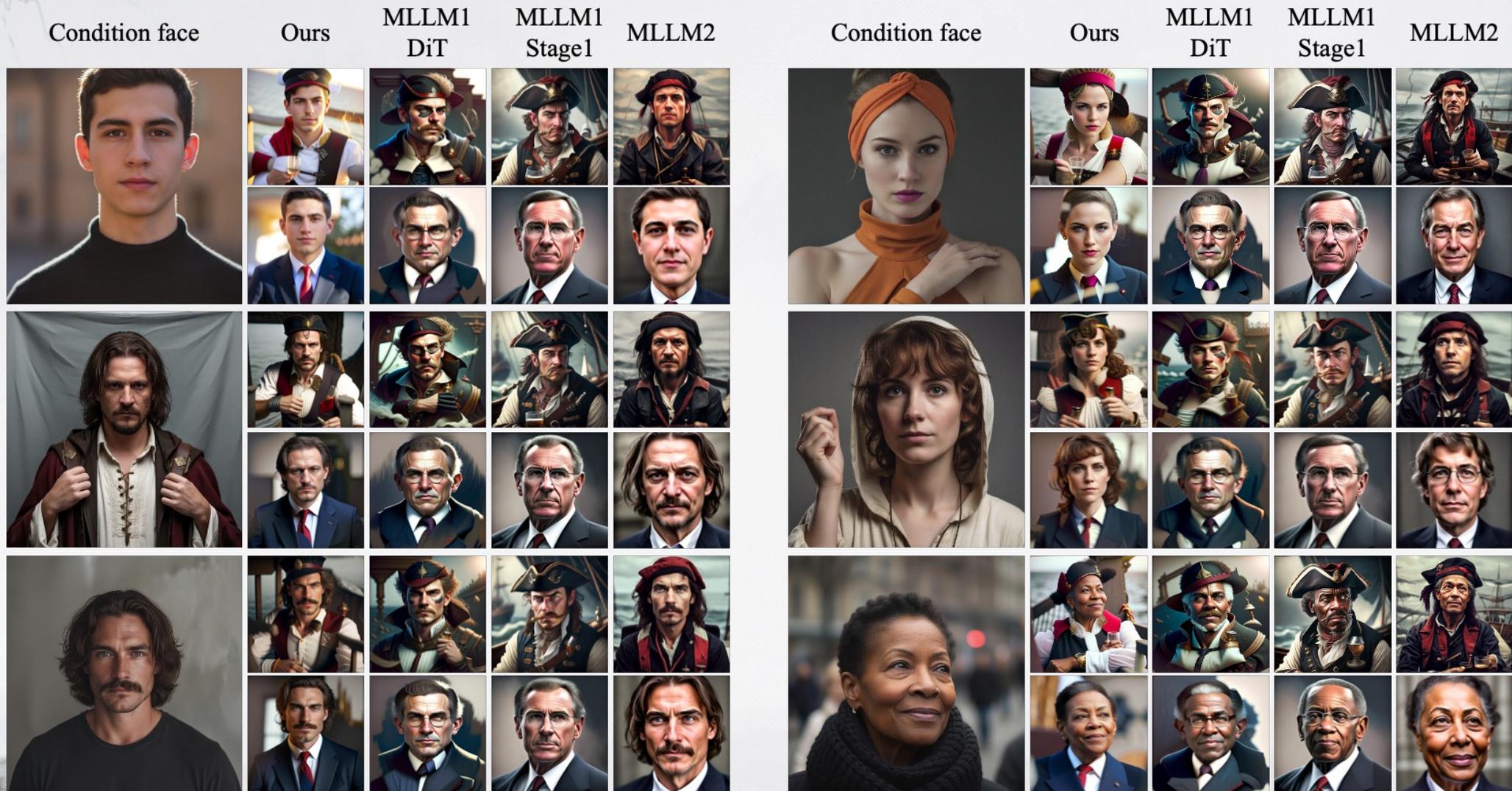
Change the image to a Salvador Dali art work



Change the blanket to bright green

4

Single Turn Personalization Result



4

User study



Single turn editing

Measurement	Ours wins	Ours loses	Tie
Image Quality	58%	2%	40%
Prompt Alignment	39%	9%	52%
Content Preservation	69%	2%	29%

Single turn personalization

Measurement	Ours wins	Ours loses	Tie
Image Quality	73.75%	3.75%	22.5%
Prompt Alignment	36.25%	15%	48.75%
Face Preservation	71.25%	2.5%	26.25%

4

Multi-turn Editing Result



Markov Multi-turn

Source Image



Round1



Round2



Round3



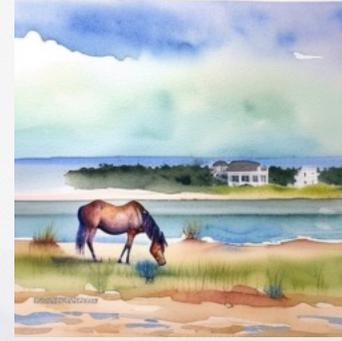
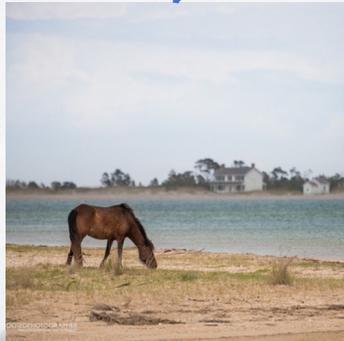
Markov: T1: Replace the shore with a mountain range;

T2: Change the horse to a white horse;

T3: Convert the image into a watercolor painting

Non-Markov Multi-turn

Non-Markov: T3': Step back 2 times, Convert the image into a watercolor painting



4

Multi-turn Editing Result



Markov Multi-turn



Markov: T1: Change the background to a sunny beach scene;
T2: Replace the parking meters with palm trees;
T3: Convert the image into a Comic Book style

Non-Markov Multi-turn

Non-Markov: T3': Instead of Replace the parking meters with palm trees, Convert the image into a Comic Book style



4

Multi-turn Personalization Result



T2I Result



Julian



Ruby



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.

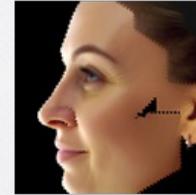
T2I Result



Harrison



Maya

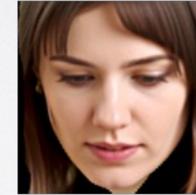


T2I Prompt: Dr. Harrison is examining Maya's face. He has fair skin, short gray hair, and wears a white coat and blue glasses, holding Maya's chin with his left hand and touching her cheek with his right. Maya, with fair skin and brown hair tied back, wears a beige shirt with black trim and sits on a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

T2I Result



Olivia



Julian



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.

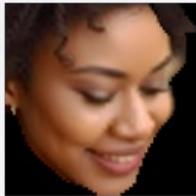
T2I Result



Julian



Mia



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.

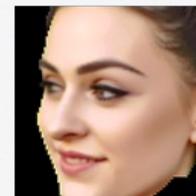
T2I Result



Julian

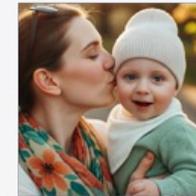


Isabella



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 Result



Amelia



Oliver



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.

⑤ Conclusion



5 Conclusion



- Demonstrated MLLMs' personalization capabilities and potential for reasoning-based image generation.
- Adopted DiT as a powerful detokenizer and proposed a multi-stage fine-tuning strategy for face preservation.
- Built two multi-turn editing and personalization datasets
- Experiments showed MLLMs effectively handle single-turn and multi-turn image generation by analyzing chat history.

Future Work

- Enhance the detokenizer to preserve content, especially for multi-turn editing.
- Develop more comprehensive multi-turn instruction fine-tuning datasets.

Thank you for listening



4

Multi-turn Editing Result



Markov



Markov: T1: Change the color of the plate to blue; T2: Remove the steak and add a grilled salmon fillet; T3: Convert the image to resemble a 19th-century still life painting

Non-Markov: T3': Backtrack 1 times, Convert the image to resemble a 19th-century still life painting

Non-Markov



Markov: T1: Convert this into a 80s anime; T2: Remove the lake and replace it with a futuristic cityscape; T3: Change the giraffe into a huge spider; T4: Incorporate a dramatic sunset sky with purple hues

Non-Markov: T4': Recall 1 actions, Incorporate a dramatic sunset sky with purple hues



4

Multi-turn Personalization Result



	T2I Result	Personalization Result	T2I Result	Personalization Result	T2I Result	Personalization Result
1st Run						
2nd Run						
1st Run						
2nd Run						

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

标注

字体使用

中文 微软雅黑
英文 Arail

行距

正文 1.2

图片出处

freepik.com

声明

本网站所提供的任何信息内容（包括但不限于 PPT 模板、Word 文档、Excel 图表、图片素材等）均受《中华人民共和国著作权法》、《信息网络传播权保护条例》及其他适用的法律法规的保护，未经权利人书面明确授权，信息内容的任何部分(包括图片或图表)不得被全部或部分的复制、传播、销售，否则将承担法律责任。

作者

OfficePLUS