

Less-to-More Generalization: Unlocking More Controllability by In-Context Generation

Supplementary Material

F. Experiments Setting

F.1. Implementation Details.

To self-evolution our base DiT-based T2I model, we firstly take the FLUX.1 dev [20] as the pretrained model. We train the model with a learning rate of 10^{-5} and a total batch size of 16. For the progressive cross-modal alignment, we first train the model using single-subject pair-data for 5,000 steps. Then, we continue training on multi-subject pair-data for another 5,000 steps. Specifically, we generated 230k and 15k data pairs for these two stages respectively using the in-context data generation method mentioned above. We trained the model using a LoRA [14] rank of 512 throughout the training process.

F.2. Evaluation Metrics.

Following previous works, we use standard automatic metrics to evaluate both subject similarity and text fidelity. Specifically, we employ cosine similarity measures between generated images and reference images within CLIP [30] and DINO [25] spaces, referred to as CLIP-I and DINO scores, respectively, to assess subject similarity. Additionally, we calculate the cosine similarity between the prompt and the image CLIP embeddings (CLIP-T) to evaluate text fidelity. For single-subject driven generation, we measure all methods on DreamBench [36] for fairness. For multi-subject driven generation, we follow previous studies [17, 22] that involve 30 different combinations of two subjects from DreamBench, including combinations of non-live and live objects. For each combination, we generate 6 images per prompt using 25 text prompts from DreamBench, resulting in 4,500 image groups for all subjects.

G. In-Context Data Generation Pipeline

In this section, we give a detailed description of our in-context data generation pipeline. We first build a taxonomy tree in Sec. G.1 to obtain various subject instances and scenes. Then we generate subject-consistent image-pair data with the in-context ability of pretrained Text-to-Image (T2I) model and utilize Chain-of-Thought (CoT) [45] to filter the synthesized data in Sec. G.2. Finally, for multi-subject data, we train a Subject-to-Image (S2I) model to generate subject-consistent reference image instead of the cropped one to avoid the copy-paste issue in Sec. G.3.

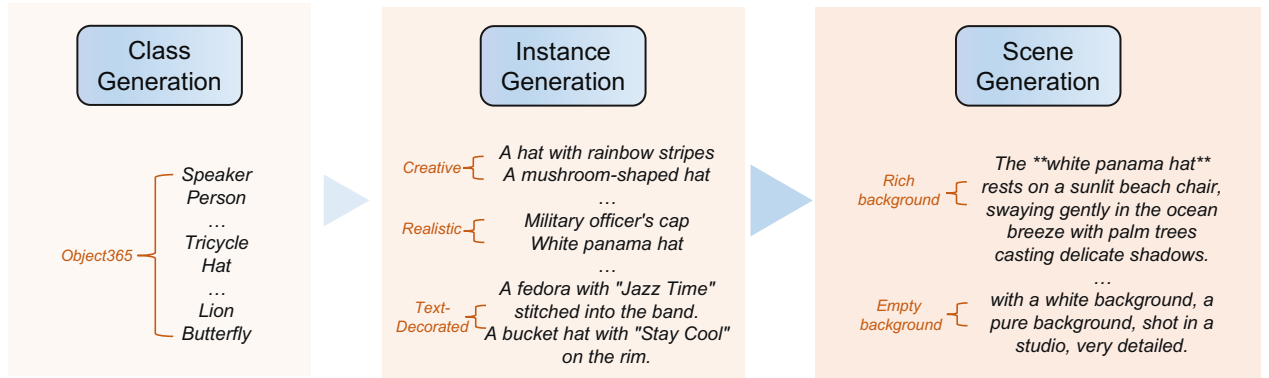



Figure 10. Illustration of the taxonomy tree.

G.1. Taxonomy Tree Generation

To ensure the diversity of the generated dataset, we first construct a taxonomy tree that includes common categories of people and objects, as shown in Fig. 10. Specifically, we use the 365 general classes from Object365 [39] as the basis for our taxonomy tree. To obtain more diverse categories, we employ Large Language Model (LLM) to generate various subject instances and diverse scenes. The instructions in Fig. 12 make LLM generate subject instances according to the given asset category in creative, realistic, and text-decorated ways. In addition, we instruct LLM to generate scene descriptions according



```

Template_1
A diptych with two side-by-side images of same <\subject1>.
Left: <\subject1> in <\scene1>.
Right: <\subject1> together with <\subject2> in <\scene2>.

Template_2
A diptych with two side-by-side images of same <\subject1>.
Top: <\subject1> in <\scene1>.
Bottom: <\subject1> together with <\subject2> in <\scene2>.

```

Figure 11. Diptych text template for generating subject-consistent image-pair with FLUX.1[20].

to the given subject with system prompt in Fig. 13. Following the steps above, we build a taxonomy tree and get plenty of diverse subjects and scene descriptions.

G.2. Single-Subject In-Context Data Generation

In-context ability of T2I model has been proved in OminiControl [41] and we also utilize it to generate subject-consistent image-pair data as the basic part of our final in-context synthesized data in Sec. G.2.1. Moreover, we filter out low quality data with bad subject consistency according to the similarity from DINOv2 [25] and Vision-Language Model (VLM) in Sec. G.2.2.

G.2.1. Subject-Consistent Image-Pair Generation

Combining the constructed taxonomy tree with the predefined diptych text template in Fig. 11, we utilize the inherent in-context ability of FLUX.1 [20], one of the state-of-the-art T2I model, to generate subject-consistent image-pair. Since FLUX.1 [20] has multi-resolution generation ability, we directly produce three different high-resolution (*i.e.*, 1024×1024 , 1024×768 , 768×1024) image-pairs, with great balance of quality and efficiency.

G.2.2. Subject-Consistent Image-Pair Filter

Though FLUX.1 [20] shows great in-context generation ability, synthesized image-pairs suffer several issues, especially subject inconsistency and missing subjects. We highlight that the high quality of synthesized data can notably accelerate the convergence and improve the subject consistency. To efficiently filter synthesized data, we first split the diptych image-pair into reference image I_{ref}^1 and target image I_{tgt} with Hough Transform. According to the template in Fig. 11, both I_{ref}^1 and I_{tgt} contain the same *subject1* while I_{tgt} has another *subject2*. To ensure I_{ref}^1 and I_{tgt} have consistent *subject1*, we then calculate cosine similarity with DINOv2 [25] and set a threshold to filter out image-pairs with significantly low consistency.

However, since the reference image I_{ref}^1 and the target image I_{tgt} have different scene settings, the *subject1* in the image-pairs may not be spatially aligned, resulting in incorrect cosine similarity with feature from DINOv2 [25]. We further employ VLM to provide a fine-grained score list evaluating various aspects adaptively, *i.e.* appearance, details, and attributes. We only keep the data with highest VLM score, which indicating the highest quality and subject consistency in the synthesized data. Specifically, inspired by [4], we utilize CoT [45] for better discrimination of the *subject1* in I_{ref}^1 and I_{tgt} , as shown in Fig. 14. To demonstrate the effectiveness of the CoT filter, we sample data from different VLM score intervals in Fig. 15. Image-pairs with low score suffer severe subject inconsistency while those with highest score (*i.e.* score is 4) show highly consistent subject in the reference image and the target image. We also count the amount of data in each score interval as shown in Fig. 17, indicating that around 35.43% data would be remained with the VLM CoT filter. Also, there are seldom of data with extremely low VLM score after DINOv2 filter, showing its effectiveness.

G.3. Multi-Subject In-Context Data Generation

Following the above pipeline, we construct single-subject in-context data containing subject-consistent image-pairs ($I_{\text{ref}}^1, I_{\text{tgt}}$). Both the reference image and the target image have *subject1* while only the target image contain *subject2*. Since I_{tgt} has multi-subject, the simplest way to build multi-subject in-context data is utilizing open-vocabulary detector (OVD) to identify and crop the *subject2* in I_{tgt} as the second reference image \tilde{I}_{ref}^2 . However, we find that the cropped \tilde{I}_{ref}^2 would make severe copy-paste issue. To alleviate the issue, a S2I model is trained with the single-subject in-context data and then used to generate new reference image I_{ref}^2 , which has the consistent subject with the cropped \tilde{I}_{ref}^2 but different scenes. Thus we have high quality multi-subject in-context data with subject-consistent image-pairs ($I_{\text{ref}}^1, I_{\text{ref}}^2, I_{\text{tgt}}$) after very similar filter pipeline for

Role:
Please be very creative and generate 50 brief subject prompts for text-to-image generation.

Follow these rules:

1. You will be given an "asset category", you need to create an asset(brief subject prompt) based on the "asset category".
2. These descriptions can refer only to appearance descriptions/or to certain brands. e.g. "Elon Musk in pajamas", "a tiger in a black hat", "A Mercedes sports car", "A blonde", "A door red on the left and green on the right".
3. Do not repeat each asset, you need to use your imagination and common sense of life to create.
3. No more than 12 words.

Example1

[asset category]: Book
[asset1]: A book with a green cover
[asset2]: comic book
[asset3]: math book
[asset4]: An open book
[asset5]: Roten books
[asset6]: A book made of candy
[asset7]: The book with "love and power" on the cover
[asset8]: A book with a blue key on it
[asset9]: Triangular book
...
(Up to [asset50])
[asset category]:

(a) System prompt of LLM used to generate subject instances in creative type.

Role:
Please be very careful and generate 50 brief subject prompts for text-to-image generation.

Follow these rules:

1. You will be given an "asset category", you need to create an asset(brief subject prompt) based on the "asset category".
2. These descriptions can refer only to appearance descriptions/or to certain brands. But it has to be something that can exist in the real world. e.g. "Elon Musk in pajamas", "a white tiger", "A Mercedes sports car", "A blonde", "A rotten wooden door".
3. Do not repeat each asset, you need to use common sense of life to create.
3. No more than 12 words.

Example1

[asset category]: Book
[asset1]: A book with a green cover
[asset2]: comic book
[asset3]: math book
[asset4]: An open book
[asset5]: Roten books
[asset6]: The book with "Harry Potter and the Sorcerer's Stone" on the cover
...
(Up to [asset50])
[asset category]:

(b) System prompt of LLM used to generate subject instances in realistic type.

Role:
Please be very careful and generate 50 brief subject prompts for text-to-image generation.

Follow these rules:

1. You will be given an "asset category", you need to create an asset(brief subject prompt) based on the "asset category".
2. Please add a text description in the brief subject. The text can appear anywhere.
2. These descriptions can refer only to appearance descriptions/or to certain brands. e.g. "Elon Musk in his pajamas with the words 'beat it'", "a white tiger holding a sign that says 'go'", "A Mercedes sports car with '101' written on it", "A blonde".
3. Do not repeat each asset, you need to use common sense of life to create.
3. No more than 12 words.

Example1

[asset category]: Person
[asset1]: A surfer with "Catch the Waves" on a surfboard.
[asset2]: A guitarist with "Rock On" on a t-shirt.
[asset3]: A dancing ballerina with "Grace" written on her tutu.
[asset4]: A chef wearing a hat that says "Cook Master"
...
(Up to [asset50])
[asset category]:

(c) System prompt of LLM used to generate subject instances in text-decorated type.

Figure 12. System prompt of LLM used to generate subject instances.

Role:
Please be very creative and generate 50 brief subject prompts for text-to-image generation.

Follow these rules:

1. Given a brief subject prompt of an asset, you need to generate 8 detailed **Scene Description** for the asset.
2. Each **Scene Description** should be a detailed description, which describes the background area you imagine for an identical extracted asset, under different environments/camera views/lighting conditions, etc (please be very very creative here).
3. Each **Scene Description** should be one line and be as short and precise as possible, do not exceed 77 tokens, Be very creative!

Example1

[asset]: Scientist with exploding beakers
[SceneDescription1]: The scientist with exploding beakers stands in a futuristic laboratory with holographic equations swirling around them.
[SceneDescription2]: Amidst the chaos of a stormy outdoor field lab, the scientist with exploding beakers conducts dramatic experiments as lightning crashes overhead.
[SceneDescription3]: In an ancient alchemist's den filled with dusty tomes, the scientist with exploding beakers looks surprised as colorful liquid bursts forth.
[SceneDescription4]: The scientist with exploding beakers is immersed in a vibrant neon-lit urban laboratory, surrounded by robotic assistants.
[SceneDescription5]: A desert makeshift tent serves as the lab where the scientist with exploding beakers creates a plume of shimmering dust.
[SceneDescription6]: On an alien planet bathed in ethereal light, the scientist with exploding beakers observes bioluminescent reactions in awe.
[SceneDescription7]: In a steampunk inspired workshop, the scientist with exploding beakers wears goggles and smiles amidst gears and steam as an experiment erupts.
[SceneDescription8]: The scientist with exploding beakers stands on a floating platform in the clouds, conducting experiments as colorful bursts light up the sky.
[asset]:

Figure 13. System prompt of LLM used to generate scene descriptions.

the synthesized I_{ref}^2 . There are some case randomly sample from our final data in Fig. 16. Interestingly, we find that a small part of image-pairs have more than 2 reference images, due to the randomness of T2I generation and OVD, empowering generalization for more-subject generation.

H. Analysis on LoRA Rank

In this section, we further conduct an ablation study on the LoRA rank. Since training parameters are strongly related to the final performance, we scale the rank from 4 to 512. As shown in the Fig. 18, increasing the rank gradually brings sustained gains, but when the rank reaches 128, the performance improvement slows down. Finally, considering both performance and resource consumption, we set UNO to a rank of 512.

I. More Qualitative Results

I.1. Qualitative Results on Multi-Subject Driven Generation

We show a more qualitative comparison of multi-subject driven generation in Fig. 19. It is clear that UNO generate images with best multi-subject consistency, following edit instructions to the subject and background.

Role
You are an expert AI assistant specializing in the objective evaluation of the consistency of subjects in two images.

Input Format
You will receive two images. You need to describe two pictures and determine whether the subject in the first picture is in the second picture.

(a) System prompt of the filter VLM.

Step 1:
Briefly describe these two images, as well as the most prominent subject that exists. Think carefully about which parts of the subject you need to break down in order to make an objective and thoroughly evaluation. Don't make evaluations at this step.

Important Notes

- Focus solely on the subject itself.
- If there is text on the subject, each text itself should be considered as an important separate part.
- Ignore the difference of subject's background, environment, position, size, etc.
- Ignore the difference of subject's actions, poses, expressions, viewpoints, lightning, etc.

Output Format

[subject]: [subject in IMG1, e.g., a man in a white shirt and black pants]
[caption1]: [IMG1 caption, e.g., a man in a white shirt and black pants]
[caption2]: [IMG2 caption, e.g., a man in a white shirt and black pants holds a blue cup, butterflies and flowers swirled around him]
[break down]: [Break down the evaluation parts of the subject in IMG1]

(b) Prompt for the first round CoT of the filter VLM.

Step 2:
For each part you have identified, compare this aspect of the subject in the two images and describe the differences in **extreme extreme extreme extreme extreme detail**. You need to be meticulous and precise, noting every tiny detail.

Important Notes

- Provide quantitative differences whenever possible. For example, "The subject's chest in the first image has 3 blue circular lights, while the subject's chest in the second image has only one blue light and it is not circular."
- Ignore differences in the subject's background, environment, position, size, etc.
- Ignore differences in the subject's actions, poses, expressions, viewpoints, additional accessories, etc.
- Ignore the extra accessory of the subject in the second image, such as hat, glasses, etc.
- Consider that when the subject has a large perspective change, the part may not appear in the new perspective, and no judgment is needed at this time. For example, if the subject in the first image is the back of the sofa, and the subject in the second image is the front of the sofa, determine the similarity of the two sofas based on your association ability.

(c) Prompt for the second round CoT of the filter VLM.

Step 3:
Based on the differences analyzed in Step 2, assign a specific integer score to each part. More and larger differences result in a lower score. The score ranges from 0 to 4:

- Very Poor (0): No resemblance. This subject part in the second image has no relation to the part in the first image.
- Poor (1): Minimal resemblance. This subject part in the second image has significant differences from the part in the first image.
- Fair (2): Moderate resemblance. This subject part in the second image has modest differences from the part in the first image.
- Good (3): Strong resemblance. This subject part in the second image has minor but noticeable differences from the part in the first image.
- Excellent (4): Near-identical. This subject part in the second image is virtually indistinguishable from the part in the first image.

Output Format

[Part 1]: [Part 1 Score]
[Part 2]: [Part 2 Score]
[Part 3]: [Part 3 Score]
[Part N]: [Part N Score]
You must adhere to the output format strictly. Each part name and its score must be separated by a colon and a space.

(d) Prompt for the third round CoT of the filter VLM.

Figure 14. CoT prompt of the filter VLM.

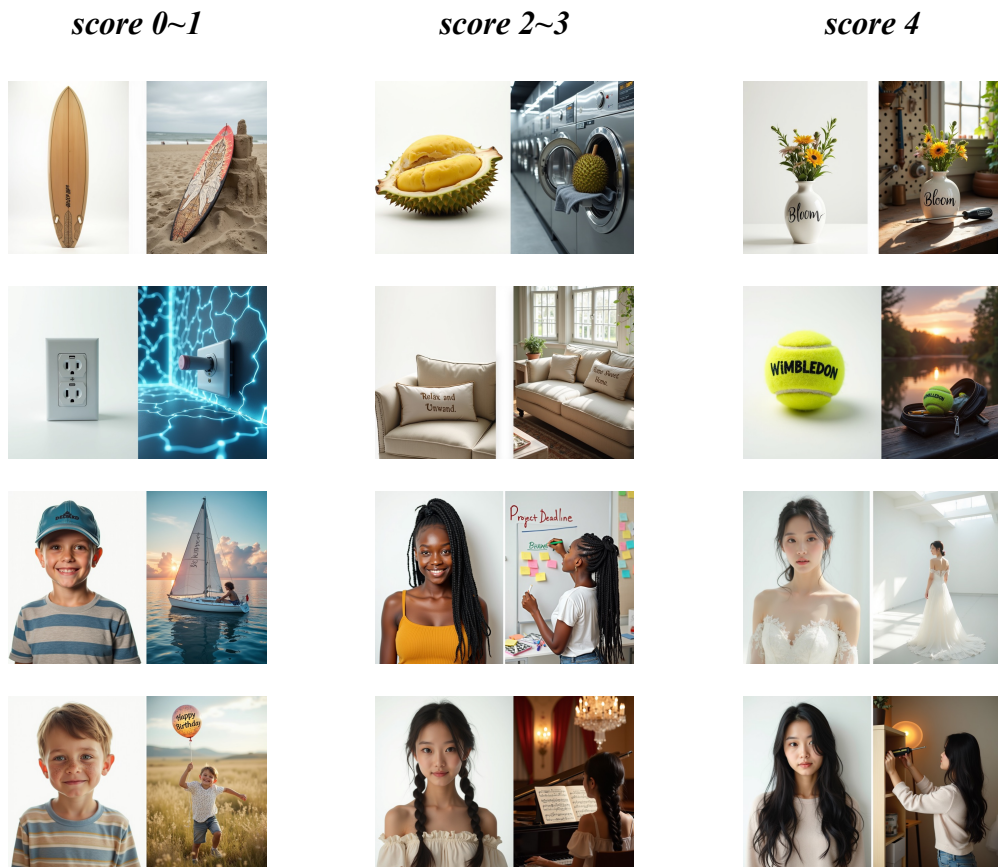


Figure 15. Sampled data from different VLM score intervals.



Figure 16. Sampled data from our final multi-subject in-context data.

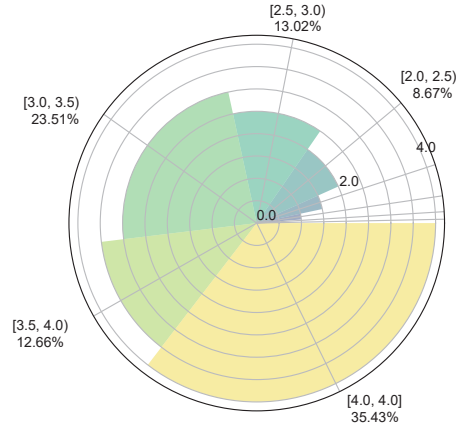


Figure 17. Amount of data in each VLM score interval.

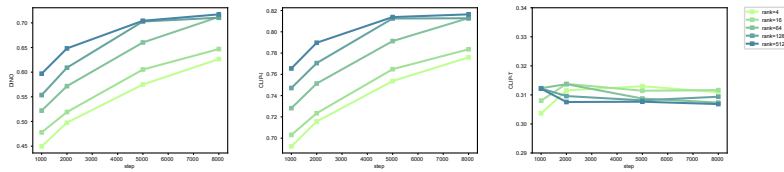


Figure 18. Analysis of model performance under different LoRA ranks.

I.2. Application Scenarios

We evaluated our UNO model across diverse multi-image conditional scenarios, such as identity preservation, virtual try-on, and stylized generation. We found that UNO demonstrated exceptional generalization capabilities, even with minimal

Scenarios	Prompt
One2One	"A clock on the beach is under a red sun umbrella" "A doll holds a 'UNO' sign under the rainbow on the grass"
Two2One	"The figurine is in the crystal ball" "The boy and girl are walking in the street"
Many2One	"A penguin doll, a car and a pillow are scattered on the bed" "A boy in a red hat wear a sunglasses"
Stylized Generation	"Ghibli style, a woman" "Ghibli style, a man"
Virtual Try-on	"A man wears the black hoodie and pants" "A girl wears the blue dress in the snow"
Product Design	"The logo and words 'Let us unlock!' are printed on the clothes" "The logo is printed on the cup"
Identity-preservation	"The figurine is in the crystal ball" "A penguin doll, a car and a pillow are scattered on the bed"
Story Generation	"A boy in green is in the arcade" "A man strolls down a bustling city street under moonlight" "The man and a boy in green clothes are standing among the flowers by the lake" "The man met a boy dressed in green at the foot of the tower"

Table 7. Text prompts used in Fig. 1.

exposure to such data during training.

Multi-subject Driven Generation: we have showcased additional results from our UNO model in Fig. 20. Beyond effectively handling multi-subject scenarios, UNO excels in complex applications like logo design and the integration of virtual and real elements, demonstrating its strong generalization capabilities. **Virtual Try-on:** as shown in Fig. 21, UNO performs exceptionally well in virtual try-on scenarios, despite the absence of specialized training on such datasets. This demonstrates that UNO has learned to understand relationships between objects rather than simply performing copy-paste operations. It also suggests that UNO could provide novel optimization strategies for virtual try-on applications, a promising direction we leave to further exploration. **Identity Preservation:** another notable observation is that UNO performs well in both pure ID scenarios and ID-subject combinations in Fig. 22. This flexibility reduces reliance on additional ID plugins, fostering open-source community development. We attribute this capability to our systematic training data construction. As mentioned in Sec. G.1, our taxonomy tree covers extensive human-object combinations, enabling this versatile performance. **Stylized Generation:** as depicted in Fig. 23, UNO has inherited stylization ability inherited from the original DiT model, despite the lack of specific paired data in our training set. This stems from our training approach, which smoothly transitions from T2I to S2I, allowing the model to evolve multi-condition control while maintaining strong semantic alignment.

J. Limitation and Discussion

Although we have established an automated data curation framework, this paper primarily focuses on subject-driven generation. Our dataset currently contains limited editing and stylization data. While UNO is a unified and customizable framework with sufficient generalization capabilities, the types of synthetic data may somewhat restrict its abilities. In the future, we plan to expand our data types to further unlock UNO’s potential and cover a broader range of tasks.

Reference Images	Prompt	UNO (Ours)	OmniGen	MS-Diffusion	MIP-Adapter	SSR-Encoder
	a toy and a toy on top of a white rug					
	a purple sneaker and a <i>purple</i> toy					
	a boot and a stuffed animal with a city in the background					
	a cartoon and a toy with a wheat field in the background					
	a dog in a <i>purple</i> wizard outfit, next to it is a can					
	a dog in a <i>police</i> outfit, next to it is a glasses					
	a dog in a <i>firefighter</i> outfit, next to it is a teapot					
	a cat wearing <i>pink</i> glasses, next to it is a cat					

Figure 19. More comparison with different methods on multi-subject driven generation. We *italicize* the subject-related editing part of the prompts.



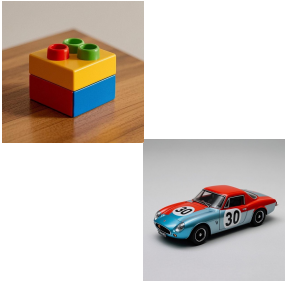








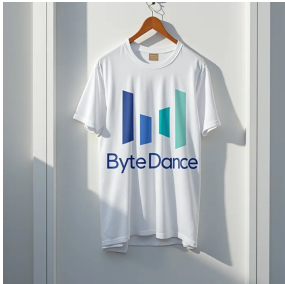




<i>Input</i>	<i>Output</i>	<i>Input</i>	<i>Output</i>
			
“A vase is next to the case, in the desert”		“A colorful block on top of the toy car”	
			
“A black cat wearing a black hat is riding a little yellow duck, in the forest”		“A little bear carries a blue bag”	
			
“The girl wears a clothes with the red logo”		“The logo is printed on the clothes”	
			
“”		“”	

Figure 20. More multi-subject generation results from our UNO model.










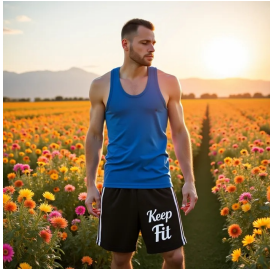






<i>Input</i>	<i>Output</i>		
	in the flowers	near the sea	in the street
			
A pretty woman wears a flower petal dress,			
			
A girl is wearing a white blouse and a blue skirt,			
			
A man wears a blue tank top and black shorts,			
			
A woman wears the dress and holds a bag,			

Figure 21. More virtual try-on results from our UNO model.

<i>Input</i>	<i>Output</i>			
				<p>1. The girl stands before a large painting in an art gallery. Her dark eyes reflect curiosity and appreciation as she studies the brushstrokes up close. She wears a tailored black blazer over a simple white blouse, exuding an air of sophistication. Around her, other patrons whisper in admiration, but she remains lost in her own artistic contemplation. 2. The girl poses in a botanical garden. She wears a floral dress, holding a bouquet, blending natural and artistic beauty. 3. The girl navigates busy city streets. She wears a trench coat and trousers, exploring the urban landscape.</p>
 				<p>1. The woman sits on a park bench, a brown leather handbag by her side. She wears a floral dress, enjoying the sunshine. The bag rests on her lap, a small umbrella tucked inside. She occasionally rummages through it for snacks. 2. The woman walks at dusk, a brown velvet evening bag in hand. 3. The woman is sitting in a cafe, with a brown bag on the chair.</p>
				<p>1. The man was lecturing on the podium, and the blackboard was full of mathematical formulas. 2. The man stands in a futuristic laboratory, wearing a sleek silver jumpsuit. He adjusts the settings on a bizarre machine, its glowing panels and wires humming with energy. 3. The man plays chess.</p>
 				<p>1. The man is sitting in a cafe, with a violin on the chair. 2. The man wears a suit and played the violin on the stage. 3. This man has fairy ears and plays the violin in the forest.</p>

Figure 22. More identity preservation results from our UNO model.

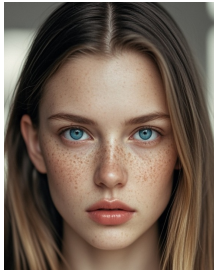







<i>Input</i>		<i>Output</i>		
		Ghibli style,	Comic style,	3D cartoon style,
				
		a woman		
				
		a man		

Figure 23. More stylized generation results from our UNO model.