


## Supplementary Material

In the supplementary material, we first introduce the details of the user study. Then, we detail the construction pipeline of the proposed CoupleX dataset. Next, we discuss the limitations. Finally, we present more qualitative results from both single-subject personalization and multiple-subject personalization.

### A. User Study Details

Fig. 8 shows the user study interface for evaluating conception personalization (CP-H) and prompt fidelity (PF-H). Each case includes two questions, one on conception personalization and the other on prompt fidelity. Participants rate their responses on a 5-point scale: (1) “Very inconsistent”, (2) “Somewhat inconsistent”, (3) “Fair”, (4) “Quite consistent”, and (5) “Very consistent”.

1.  
Reference image: **berry bowl**  
Prompt for image generation: A **berry bowl** on table, **spoon beside with blueberries inside**.



	Very inconsistent	Somewhat inconsistent	Fair	Quite consistent	Very consistent
Please evaluate the similarity between the berry bowl in the generated image and the one in the reference image.	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Please evaluate how well the content in the generated image matches the provided textual prompt	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Figure 8. User study details.

### B. Dataset Construction Pipeline

In this section, we introduce our dataset construction pipeline with the Large Language Model (LLM) and Vision Language Model (VLM). For the LLM, we use Qwen3-14B, and for the VLM, we use Qwen2.5-VL-32B-InstructIn. In Fig. 9, we present the prompt for data construction.

In Stage 1, LLM generates 50 brief subject prompts based on given subject categories. Each prompt must be realistic, concise, and created using common sense, with no repetition of subjects. The category of the subject is uniformly sampled from the 295 categories selected from the Object365 [23] taxonomy tree. In Stage 2, a short and vivid sentence is written using the given subject by LLM. This sentence should describe a plausible scene or interaction involving the subject, maintaining the exact wording of the

original subject without omission or reordering. Then, we use Flux model to generate images based on the sentence descriptions above. In Stage 3, VLM evaluates the consistency of the subject between two generated images. It assigns a resemblance score from 0 to 9 based on detailed visual comparison, where a score below 2 results in the image being filtered out. To further improve the data quality, we apply manual filtering after VLM filtering. From Fig. 10(b), we generated 72,000 paired images, of which 50,996 remained after filtering with the VLM filter. Due to the efficiency of the data synthesis pipeline, the manually filtered samples account for approximately 10%. In the end, the proposed CoupleX dataset consists of 45,548 high-quality paired images. CoupleX contains 295 categories, which can be grouped into 11 parent categories. The number of images in each parent category is shown in Fig. 10(a).

In Fig. 11, we showcase four representative examples from our proposed CoupleX dataset. CoupleX is enriched with diverse subjects, each accompanied by specific descriptions. For instance, it includes items such as “cooked shrimp” and “sliced hamimelon”. Additionally, our dataset includes common activities of subjects, such as “jellyfish drifting” and “zebras sipping water”. These features are absent in previous synthetic subject-pair datasets. This enhancement in both the variety of subjects and the incorporation of subject behaviors enables a more detailed and applicable range of scenarios for personalization image generation.

### C. Limitation and Discussion

Due to the limitations of the foundation model, the generated paired images cannot fully replicate all real-world scenarios. This constraint subsequently restricts the associative abilities of the proposed Genova. Moreover, our method currently does not extend to tasks involving the generation of three or more objects interacting with each other, as there is a lack of high-quality datasets featuring multiple interacting objects. In future work, we will expand our dataset to include multi-subject paired data and train Genova on it to enhance its application potential for personalization on more subjects.

### D. More Qualitative Results

We show a more qualitative comparison of single-subject personalization in Fig. 12. The baseline methods encounter significant challenges with subject consistency and prompt alignment. For instance, in row 4 of Fig. 12, although TokenVerse successfully places a rose into a vase, the shape of the vase is changed. UNO and Flux-Kontext maintain the target vase better, but the generated image merges the rose and the vase into an unnatural composite. Besides, XVerse and OminiControl exhibit obvious breakdowns in their

## Dataset Construction Pipeline

### *Stage 1: Generation of subject descriptions(LLM)*

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

You will be given a [subject category], based on which you are required to create a brief subject prompt that describes a plausible, real-world entity. The description should focus solely on visual characteristics, and must be grounded in common sense. Repetition across subjects is not allowed.

Example [subject category]: Cat, [subject1]: British Shorthair cat [subject2]: A yellow furry cat...

### *Stage 2: Generation of diptychs(LLM)*

Role: You are an AI expert. Please generate a vivid, concise sentence for text-to-image generation.

Requirements:

1. Use the given subject.
2. Description of a sentence about a subject that can be active in a certain scene or interact with another object.
3. Include specific background or setting details, but not too complex.
4. The sentence must include all words in the subject exactly as provided, with no omissions or reordering.

### *Stage 3: Image Filtering(VLM)*

Role: You are an AI expert tasked with objectively assessing the consistency of subjects across two images. You will analyze two images. Describe each image and determine if the subject from the first is present in the second.

Step 1: Thoroughly inspect the most prominent subject in both images. Deconstruct it into key evaluative components; however, you do not need to include these in your output.

Step 2: Conduct a detailed comparison of each identified component, noting all differences. These details do not need to be listed in your output.

Step 3: Based on your comprehensive analysis, provide a single integer score from 0 to 9 that reflects the overall similarity of the subject between the two images.

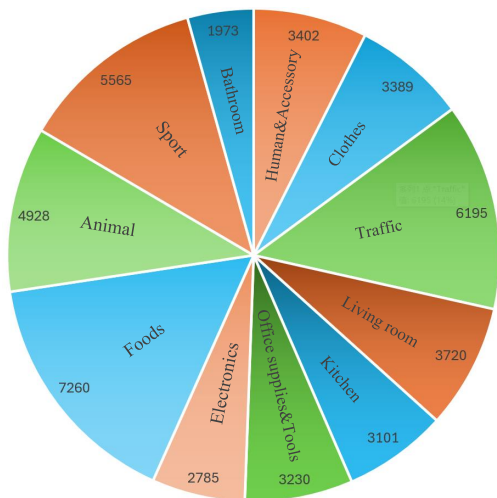
Output Format: Only output a single integer from 0 to 9, and nothing else.

**The image with a score less than 2 will be filtered out.**

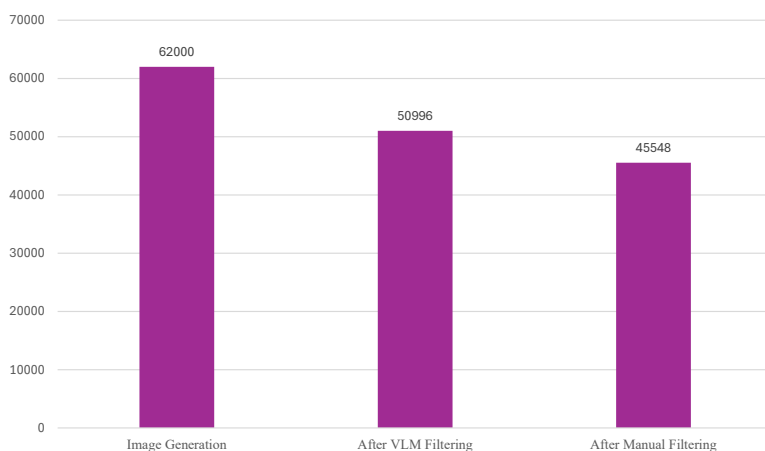
Figure 9. The dataset construction pipeline with LLM/VLM.

output. Similarly, in row 3, while TokenVerse, XVerse, and UNO do generate images of a “foggy park”, the shape of the street lamp is noticeably altered. Neither Flux-Kontext nor OminiControl accurately reflects the “foggy” aspect specified in the prompt. In contrast, our method achieved outstanding performance in maintaining high fidelity and semantic consistency, demonstrating its superiority against the challenges faced by the baseline models.

For multi-subject image generation, we present additional results generated by Genova in Fig. 13. These examples demonstrate Genova’s capability to preserve detailed features of subjects, such as the teddy bear’s goggles in row 3 and the “Genova” chocolate brand in row 4. Furthermore, it adeptly combines the given two objects according to the prompt, as seen with the giraffe wearing a red scarf in row 2, and the cherries submerged in a water glass in row 4. These results highlight the model’s proficiency in maintaining subject consistency while creatively integrating distinct subjects as specified in the prompts.



(a) Data category distribution



(b) Number of filtered paired images

Figure 10. (a) Data category distribution and (b) the number of generated images after VLM and manual filtering steps.

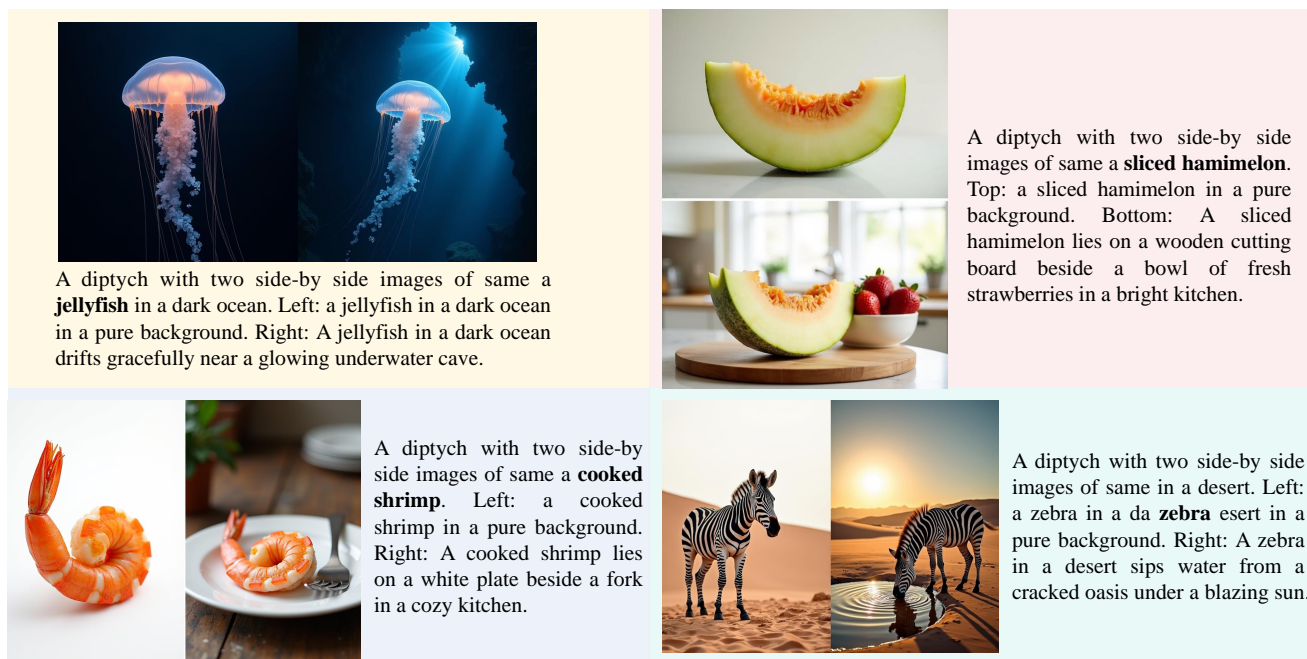


Figure 11. Examples from our proposed CoupleX dataset. The bold texted text represents the subject.















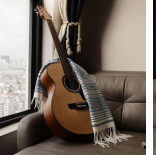






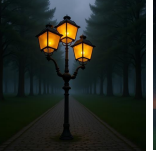
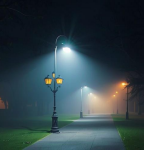





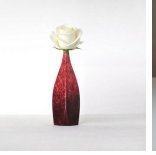









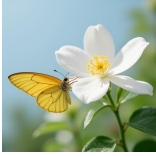


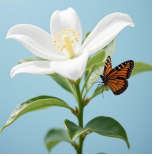
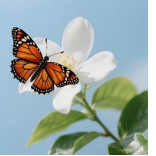






















Reference Image	Prompt	Genova(Ours)	TokenVerse	XVerse	UNO	Flux-Kontext	OminiControl
	A <b>boat</b> floating in the air, <b>carried</b> by a cluster of vibrant balloons against a sunset backdrop.						
	A scarf drapes over the <b>guitar</b> neck.						
	A <b>street lamp</b> illuminating a foggy park pathway.						
	A single white rose rests in the vase.						
	The <b>yellow alarm clock</b> rests beside a folded blanket on the bed.						
	A <b>butterfly</b> rests on a white lily.						
	A <b>bird</b> perches next to the <b>wolf plushie</b> on the rock.						
	A car halts in front of the <b>stop sign</b> .						
	A small leaf falls onto the <b>candle</b> sitting on the table.						

Figure 12. More qualitative comparison of subject-driven image generation. In the prompt, the bolded text represents the subject, the purple text indicates another object with which the target subject interacts, and the blue text represents the action.

Reference Images	Prompt	Output Images	Reference Images	Prompt	Output Images
 	a <b>backpack</b> and a <b>stuffed animal</b> on the beach.		 	a <b>cat</b> <b>lying</b> on a <b>hat</b> , floating on a river	
 	a <b>duck</b> <b>resting</b> comfortably on the edge of a <b>boat</b> .		 	a <b>giraffe</b> <b>wearing</b> a <b>red scarf</b>	
 	a <b>small bird</b> <b>in</b> a <b>nest</b> on a branch.		 	a <b>teddy bear</b> <b>on</b> flamingo float, bright pool	
 	a <b>cherry</b> <b>in</b> a glass of water.		 	a <b>chocolate sign</b> <b>places</b> on top of a cake.	

Figure 13. More results of multi-subject image generation. In the prompt, the purple text indicates the subjects, and the blue text represents the interaction.