Unified Customized Generation by Disentangled Reward Modeling

Supplementary Material

6. Appendix

6.1. Experiments Setting

6.1.1. Implementation Details.

We begin with FLUX.1 dev [15] and the SigLIP [43] pretrained model. We train on triplets $\{I^c_{ref}, I^s_{ref}, I_{tgt}\}$ for $21,000$ steps at batch size 64, learning rate $8e-5$, resolution 1024 and reward steps $S = 18,000$. LoRA [10] rank 128 is used throughout.

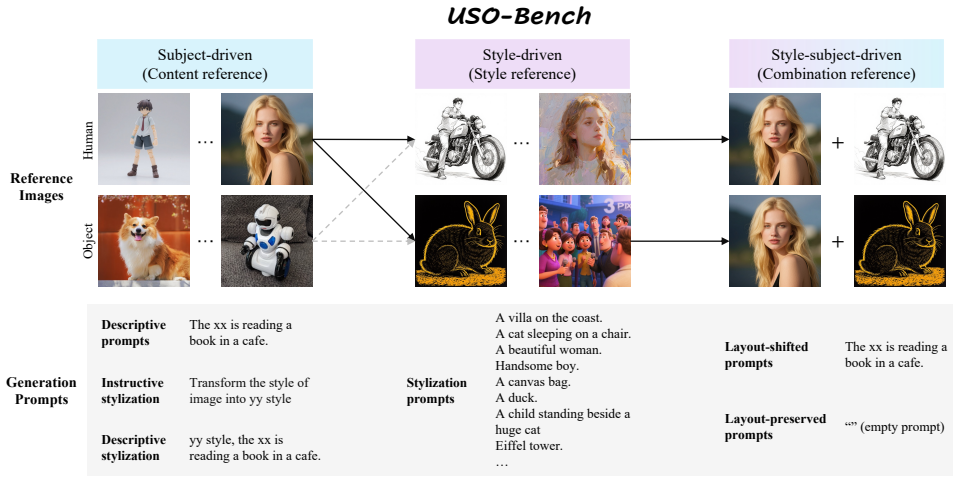

Figure 8. Examples of USO-Bench.

6.1.2. Details of USO-Bench.

USO-Bench is built to evaluate subject-driven, style-driven, and joint style-subject-driven generation. As shown in Figure 8, each subject-driven sample uses three prompt types: descriptive, instructive-stylization, and descriptive-stylization. By pairing these prompts with style-reference images from style-driven tasks, we obtain style-subject-driven samples via their Cartesian product. The resulting prompts are further split into layout-shifted and layout-preserved variants.

6.2. Auxiliary Style Reward

We present detail algorithm in Algorithm 1.

6.3. Additional Experiments

6.3.1. User Study.

We further conduct an online user-study questionnaire to compare state-of-the-art subject-driven and style-driven methods. Questionnaires were distributed to both domain experts and non-experts, who ranked the best results for each task. (1) *Subject-driven tasks* were evaluated on text fidelity, visual appeal, subject consistency, and overall quality. (2) *Style-driven tasks* were judged on text fidelity, visual appeal, style similarity, and overall quality. As shown in Figure 9, our USO achieves top performance on both tasks, validating the effectiveness of our cross-task co-disentanglement and showcasing its capability to deliver state-of-the-art results.

6.3.2. Quantitative Evaluation on DreamBench [28].

To further assess USO, we evaluate it on DreamBench [28] in addition to USO-Bench. Following UNO [37], we generate six images per prompt, yielding 4,500 image groups across all subjects. As shown in Table 4, USO achieves the highest CLIP-I and DINO scores, and with a CLIP-T score of 0.316, it trails the top result (0.318) by only a narrow margin. These results demonstrate USO's superior subject consistency among state-of-the-art methods.

Algorithm 1 Auxiliary Style Reward (ASR) with Flow Matching

Require: Customization model $\texttt{net}$ with pretrained parameters θ; pretrain loss $\mathcal{L}_{\text{Pre}}$; reward loss $\mathcal{L}_{\text{ASR}}$; reward model $\mathcal{M}_{\text{RM}}$; balancing coefficient λ; noise-schedule steps T; ASR fine-tuning interval $[t_s, t_e]$; dataset $\mathcal{D} = \{(y, I_0, I^c_{\text{ref}}, I^s_{\text{ref}})\}$, y is prompt, I_0 is target image and I^c_{ref}, I^s_{ref} are reference content and style images (Section 3.1)

for $(y, I_0, I^c_{\text{ref}}, I^s_{\text{ref}}) \in \mathcal{D}$ **do**
  $\mathcal{L}_{\text{Pre}} \leftarrow \texttt{net}_\theta(y, I_0, I^c_{\text{ref}}, I^s_{\text{ref}})$ // calculate pretrain loss with Equation (4)
  $t \sim \mathcal{U}(t_s, t_e)$ // pick a random time step in $[t_s, t_e]$
  $x_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  for $\tau = T, \ldots, t+1$ **do**
    $\hat{v}_\tau \leftarrow \text{no-grad}(\texttt{net}_\theta(y, x_\tau, I^c_{\text{ref}}, I^s_{\text{ref}}))$
    $x_{\tau-1} \leftarrow x_\tau - \hat{v}_\tau \Delta t$ // reverse-step update
  end for
  $\hat{v}_t \leftarrow \texttt{net}_\theta(y, x_t, I^c_{\text{ref}}, I^s_{\text{ref}})$
  $\hat{I}_0 \leftarrow \text{decode}(x_t - \hat{v}_t \Delta t)$ // predict original image
  $\mathcal{L}_{\text{ASR}} \leftarrow -\mathcal{M}_{\text{RM}}(\hat{I}_0, I^s_{\text{ref}})$ // calculate ASR loss with negative reward with Equation (3)
  $\mathcal{L} \leftarrow \mathcal{L}_{\text{Pre}} + \lambda \mathcal{L}_{\text{ASR}}$
  $\theta \leftarrow \theta - \eta \nabla_\theta \mathcal{L}$ // update model parameters via gradient descent (η is learning rate)
end for

Method	DINO ↑	CLIP-I ↑	CLIP-T ↑
Oracle(reference images)	0.774	0.885	-
Textual Inversion [7]	0.569	0.780	0.255
DreamBooth [28]	0.668	0.803	0.305
BLIP-Diffusion [18]	0.670	0.805	0.302
ELITE [34]	0.647	0.772	0.296
Re-Imagen [1]	0.600	0.740	0.270
BootPIG[25]	0.674	0.797	0.311
SSR-Encoder[44]	0.612	0.821	0.308
RealCustom++ [12, 20]	0.702	0.794	**0.318**
OmniGen [39]	0.693	0.801	0.315
OminiControl [31]	0.684	0.799	0.312
FLUX.1 IP-Adapter	0.582	0.820	0.288
UNO [37]	0.760	0.835	0.304
USO (Ours)	**0.800**	**0.838**	0.316

Table 4. Quantitative results for single-subject driven generation on Dreambench [28]. We highlight the best and second-best values for each metric.

6.3.3. Additional Ablation Experiments

Effect of Disentangled Encoder (DE). We provide a visual comparison of using a single encoder versus separate encoders for the two conditions. As shown in Figure 10, the "cheetah" reverts to a photorealistic appearance, while the man's identity suffers a marked loss, further demonstrating the effectiveness of our disentanglement training.

Effect of Hierarchical Projector. To demonstrate the effectiveness of the Hierarchical Projector, we freeze all other parameters and fine-tune only this module to create a stylized variant that enables the pretrained T2I model to accept style-reference images as conditional input. This allows us to isolate its contribution. As shown in Table 5, the hierarchical projector achieves the highest CSD and a top CLIP-T score, confirming its key role in style-alignment training.

6.4. More Results.

We present additional qualitative results from USO:

- From Figures 11 to 14, USO demonstrates the ability to extract task-relevant content features while maintaining subject consistency across diverse textual prompts—capabilities that prior work typically treats as isolated tasks (e.g., subject-driven generation, instruction-based stylized editing, and identity preservation).

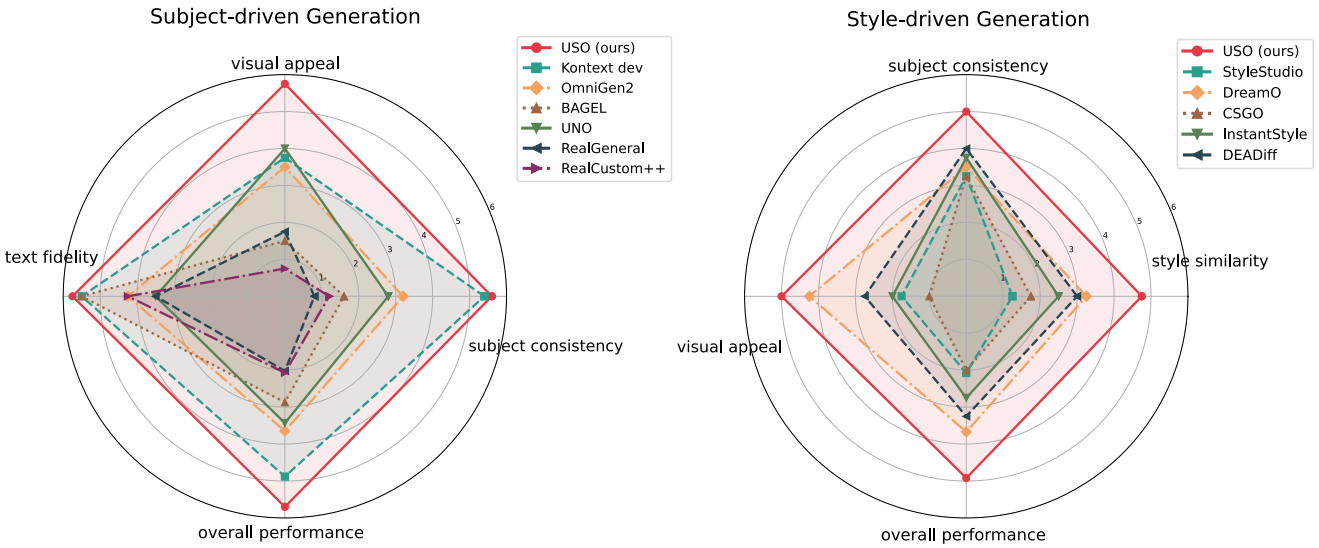

Figure 9. Radar charts of user evaluation of methods for subject-driven and style-driven generation on different dimensions.

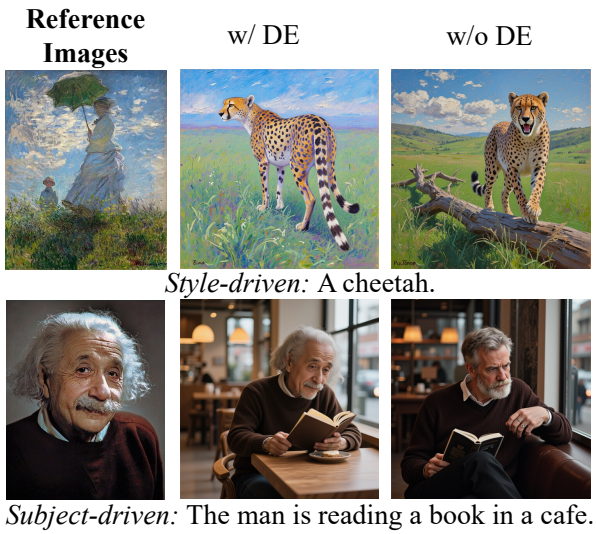

Figure 10. Ablation study of disentangled encoder. Zoom in for details.

Table 5. Ablation study of different projector in USO.

Model	CSD↑	CLIP-T↑
resampler (depth=1)	<u>0.336</u>	0.279
resampler, unfreeze siglip	0.155	**0.288**
mlp (depth=1)	0.277	<u>0.284</u>
mlp, unfreeze siglip	0.179	**0.288**
hierarchical projector	**0.402**	<u>0.284</u>

- In Figures 15 and 16, USO exhibits high stylistic fidelity, capturing both fine-grained characteristics (e.g., brushwork and material textures) and abstract artistic styles—far beyond simple color transfer.

(1) A dog in the jungle. ***(2)*** A dog on a cobblestone street. ***(3)*** A dog on the beach. ***(4)*** A dog on top of a wooden floor. ***(5)*** A dog on top of fabric. ***(6)*** Transform the style of image into Studio Ghibli anime style. ***(7)*** Oil painting style, a dog with a house in the background. ***(8)*** Pixel style, a dog on a cobblestone street. ***(9)*** Retro comic style, a dog in the jungle.

Figure 11. More results on subject-driven generation.

- In Figures 17 and 18, USO freely combines arbitrary subjects with arbitrary styles, supporting both layout-preserving and layout-shifting generations.

(1) This man is surfing, with the waves behind him chasing after him. ***(2)*** Handsome man is playing Piano. ***(3)*** This man in suit was playing the violin on the stage when a beam of light shone upon him. ***(4)*** This man is holding a cat in the garden. ***(5)*** This man stood on the moon and made a "yeah" sign, with a miniature of the Earth behind him. ***(6)*** The boy is reading a book in the coffe. ***(7)*** This man is playing football on the playground under the setting sun. ***(8)*** Handsome man in the city. ***(9)*** This man is in the water, with fish circling around him.

Figure 12. More results on subject-driven generation.

(1) The man is reading a book in a cafe. ***(2)*** The man carried a backpack with a kitten inside. ***(3)*** A man in a silver sequin jacket dances in a club, strobe lights bouncing off his coat like. ***(4)*** A man fixes a bike at dusk, wrench shining in orange twilight. ***(5)*** This man was walking on the street at night, with the blurry neon lights behind him reading "USO". ***(6)*** Sketch style, the man is walking with a dog, on the path in the park. ***(7)*** Pixel style, the man in flower shops carefully match bouquets, conveying beautiful emotions and blessings with flowers. ***(8)*** Lego building block wind, the man is reading a book in a cafe. ***(9)*** Studio Ghibli anime style, The man gave an impassioned speech on the podium.

Figure 13. More results on identity-driven generation.

(1) The woman crouched down in the garden and carefully trimmed the flower branches. ***(2)*** The woman is reading a book in a cafe. ***(3)*** A woman in a black leather jacket at night, streetlights streaking past like gold lines, her jacket collar flipping to catch cool blue neon. ***(4)*** A woman is mixing paint in a sunny art studio. ***(5)*** This woman writes on the blackboard, side view, the blackboard blurs "USO inspires creativity". ***(6)*** Retro comic style, the woman is walking in a retro alley, with the sky drizzling and the raindrops clearly visible. ***(7)*** Pixel style, the woman crouched down in the garden and carefully trimmed the flower branches. ***(8)*** 3D Cartoon Style, the woman rides a deer in the forest. ***(9)*** Studio Ghibli anime style, the woman gave an impassioned speech on the podium.

Figure 14. More results on identity-driven generation.

Figure 15. More results on style-driven generation.

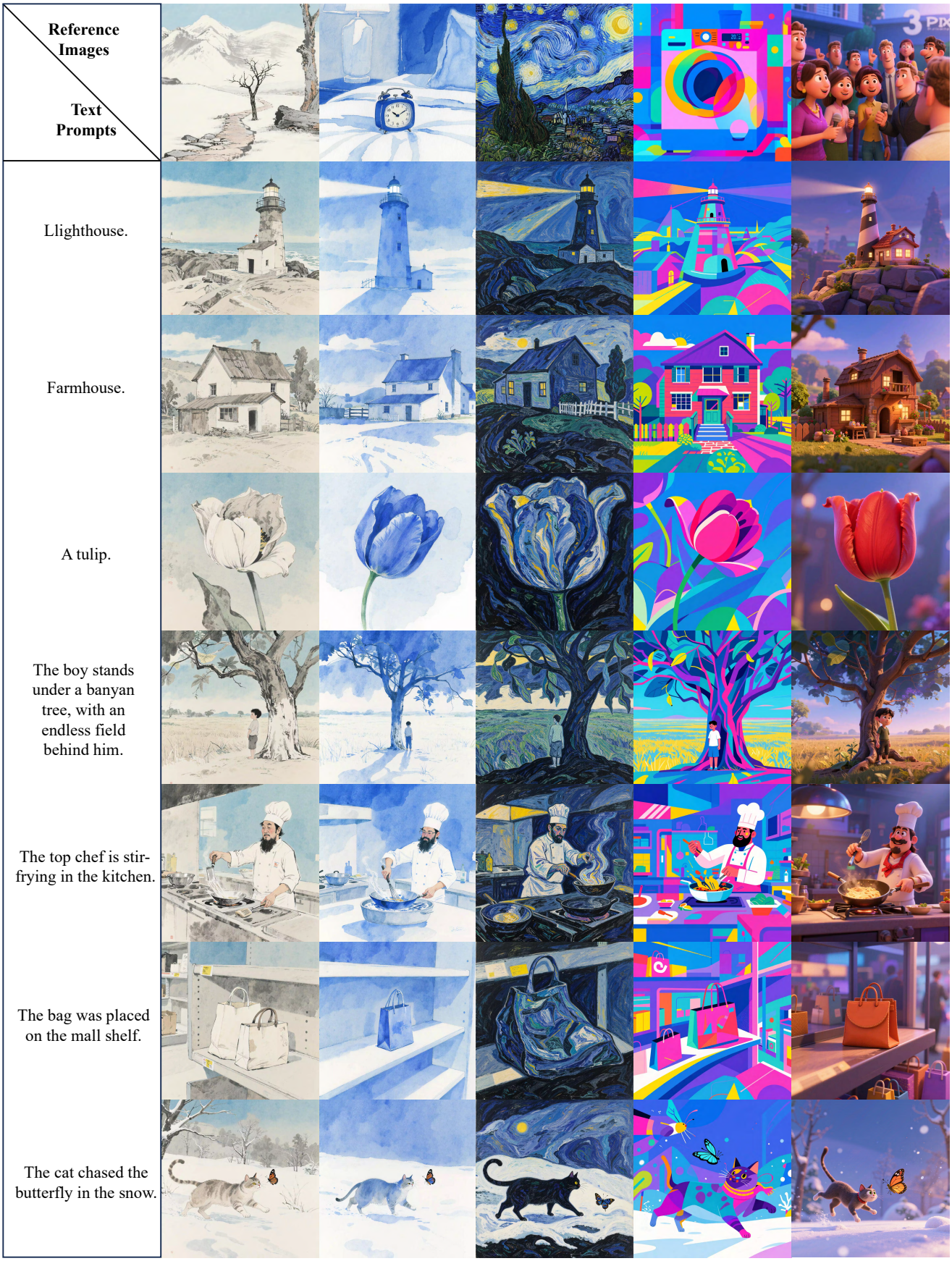

Figure 16. More results on style-driven generation.

Figure 17. More results on style-subject-driven generation. We set prompt to empty for layout-preserved generation.

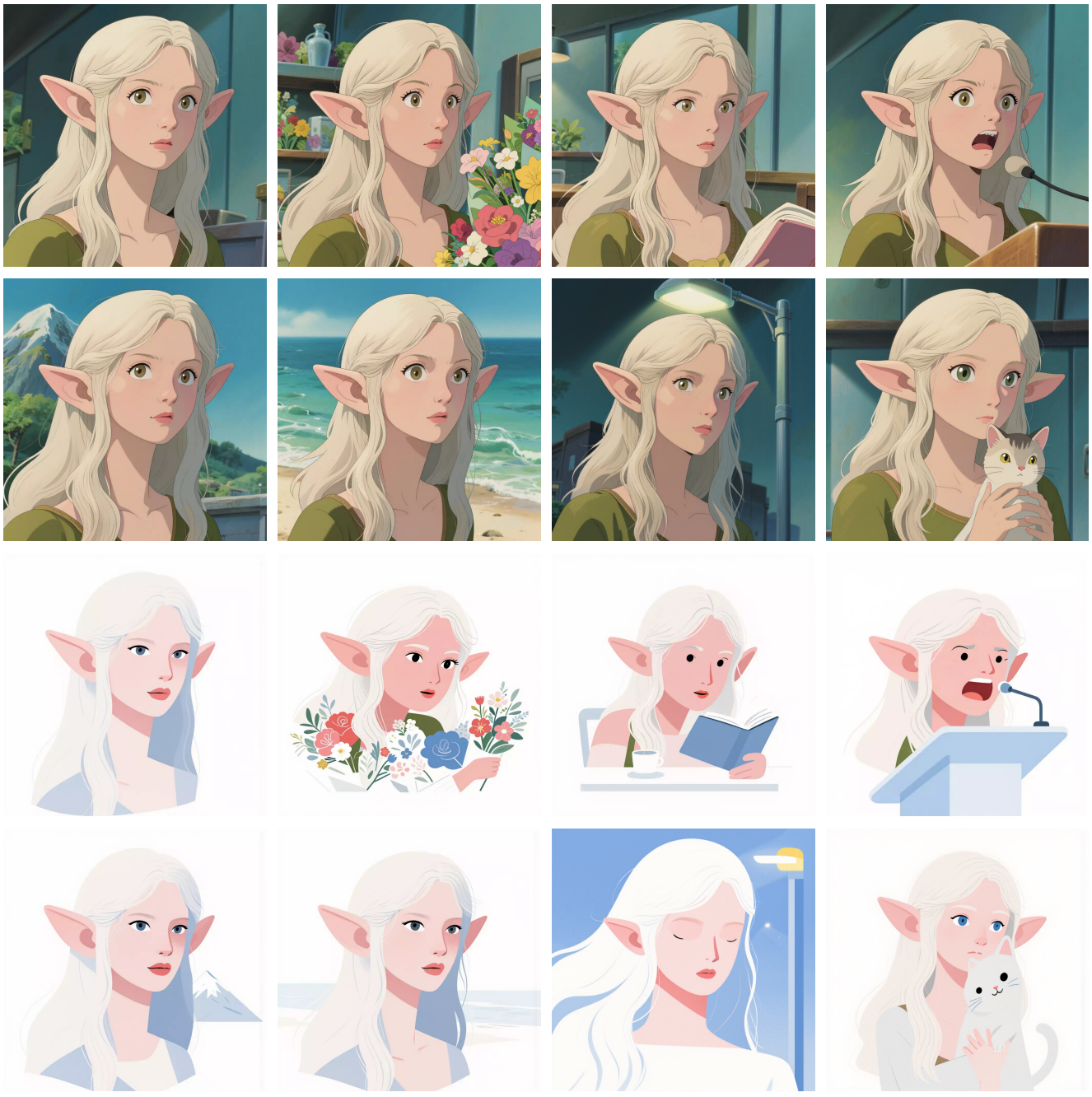

Figure 18. More results on style-subject-driven generation. USO supports any subject combined with any style in any scenario.