IP-Adapter: Text Compatible Image Prompt Adapter for Text-to-Image Diffusion Models

paper, code

2/ Using image prompt

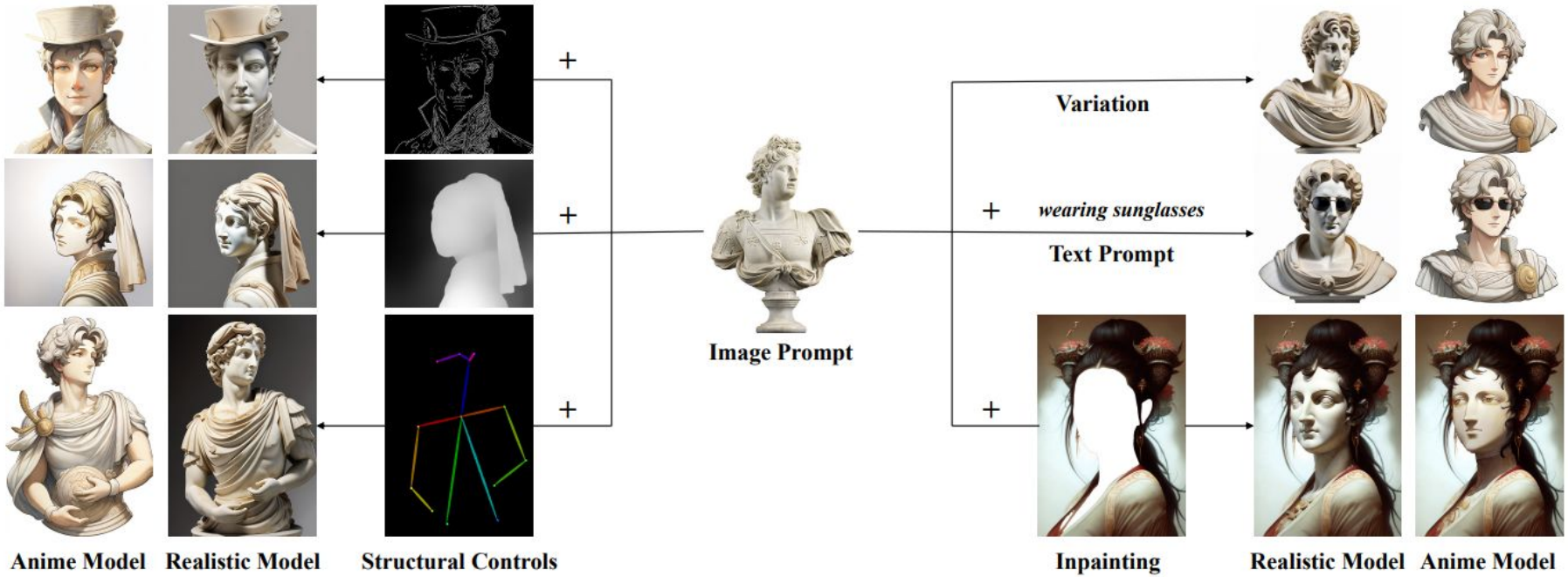
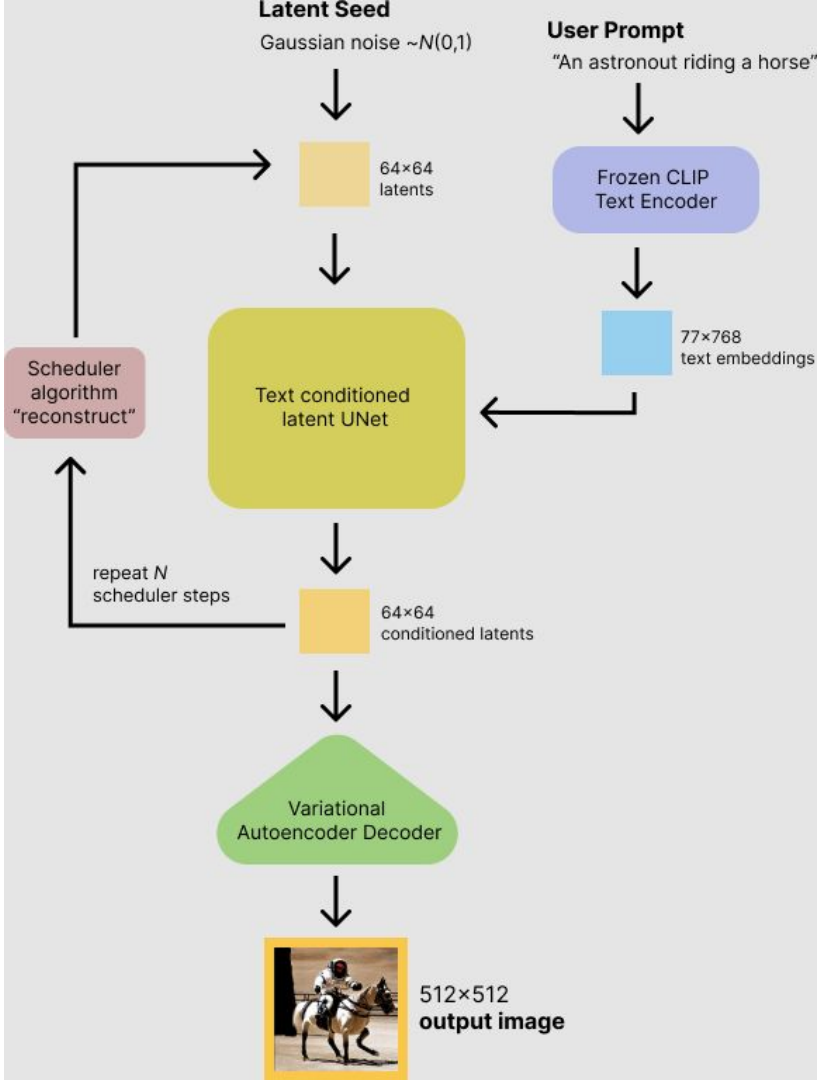


Figure 1: Various image synthesis with our proposed IP-Adapter applied on the pretrained text-to-image diffusion models with different styles. The examples on the right show the results of image variations, multimodal generation, and inpainting with image prompt, while the left examples show the results of controllable generation with image prompt and additional structural conditions.

3/ Stable Diffusion



4/ Proposed approach

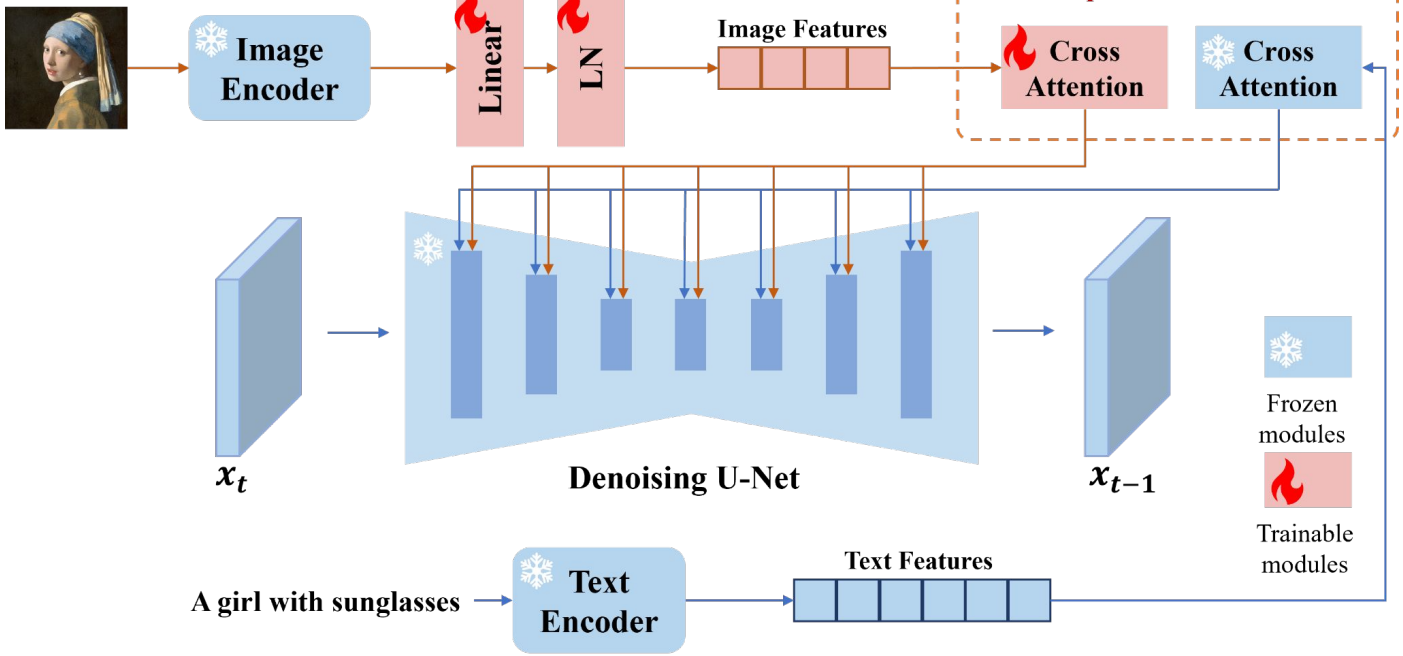


Figure 2: The overall architecture of our proposed IP-Adapter with decoupled cross-attention strategy. Only the newly added modules (in red color) are trained while the pretrained text-to-image model is frozen.

To effectively decompose the global image embedding, we use a small trainable projection network to project the image embedding into a sequence of features with length N (we use N = 4 in this study), the dimension of the image features is the same as the dimension of the text features in the pretrained diffusion model. The projection network we used in this study consists of a linear layer and a Layer Normalization

5/ Decoupled cross-attention

$$\mathbf{Z}' = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V}, \tag{3}$$

where $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K} = \boldsymbol{c}_t\mathbf{W}_k$, $\mathbf{V} = \boldsymbol{c}_t\mathbf{W}_v$ are the query, key, and values matrices of the attention operation respectively, and $\mathbf{W}_q$, $\mathbf{W}_k$, $\mathbf{W}_v$ are the weight matrices of the trainable linear projection layers.

$$\mathbf{Z}'' = \text{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}') = \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}})\mathbf{V}', \tag{4}$$

where, $\mathbf{Q} = \mathbf{Z}\mathbf{W}_q$, $\mathbf{K}' = \boldsymbol{c}_i\mathbf{W}'_k$ and $\mathbf{V}' = \boldsymbol{c}_i\mathbf{W}'_v$ are the query, key, and values matrices from the image features. $\mathbf{W}'_k$ and $\mathbf{W}'_v$ are the corresponding weight matrices. It should be noted that we use the same query for image cross-attention as for text cross-attention. Consequently, we only need add two paramemters $\mathbf{W}'_k$, $\mathbf{W}'_v$ for each cross-attention layer. In order to speed up the convergence, $\mathbf{W}'_k$ and $\mathbf{W}'_v$ are initialized from $\mathbf{W}_k$ and $\mathbf{W}_v$. Then, we simply add the output of image cross-attention to the output of text cross-attention. Hence, the final formulation of the decoupled cross-attention is defined as follows:

$$\mathbf{Z}^{new} = \text{Softmax}(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}})\mathbf{V} + \text{Softmax}(\frac{\mathbf{Q}(\mathbf{K}')^\top}{\sqrt{d}})\mathbf{V}' \tag{5}$$
$$\text{where } \mathbf{Q} = \mathbf{Z}\mathbf{W}_q, \mathbf{K} = \boldsymbol{c}_t\mathbf{W}_k, \mathbf{V} = \boldsymbol{c}_t\mathbf{W}_v, \mathbf{K}' = \boldsymbol{c}_i\mathbf{W}'_k, \mathbf{V}' = \boldsymbol{c}_i\mathbf{W}'_v$$

Sine we freeze the original UNet model, only the $\mathbf{W}'_k$ and $\mathbf{W}'_v$ are trainable in the above decoupled cross-attention.

6/ Training and Inference

During training, we only optimize the IP-Adapter while keeping the parameters of the pretrained diffusion model fixed. The IP-Adapter is also trained on the dataset with image-text pairs[1], using the same training objective as original SD:

$$L_{\text{simple}} = \mathbb{E}_{\boldsymbol{x}_0,\boldsymbol{\epsilon},\boldsymbol{c}_t,\boldsymbol{c}_i,t}\|\boldsymbol{\epsilon} - \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_t, \boldsymbol{c}_i, t)\|^2. \tag{6}$$

We also randomly drop image conditions in the training stage to enable classifier-free guidance in the inference stage:

$$\hat{\boldsymbol{\epsilon}}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_t, \boldsymbol{c}_i, t) = w\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, \boldsymbol{c}_t, \boldsymbol{c}_i, t) + (1-w)\boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t, t) \tag{7}$$

Here, we simply zero out the CLIP image embedding if the image condition is dropped.

As the text cross-attention and image cross-attention are detached, we can also adjust the weight of the image condition in the inference stage:

$$\mathbf{Z}^{new} = \text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) + \lambda \cdot \text{Attention}(\mathbf{Q}, \mathbf{K}', \mathbf{V}') \tag{8}$$

where λ is weight factor, and the model becomes the original text-to-image diffusion model if $\lambda = 0$.

7/ Details

- 10M пар text-image из LAION-2B и COYO-700M;
- SD1.5 в качестве бэйзлайна;
- OpenCLIP ViT-H/14 Image Encoder;
- Совместимость с diffusers;
- 8xV100, 1M steps, batch_size=8 на каждую видеокарту. AdamW, lr=1e-4.
- 22M обучаемых параметров.

8/ Comparison

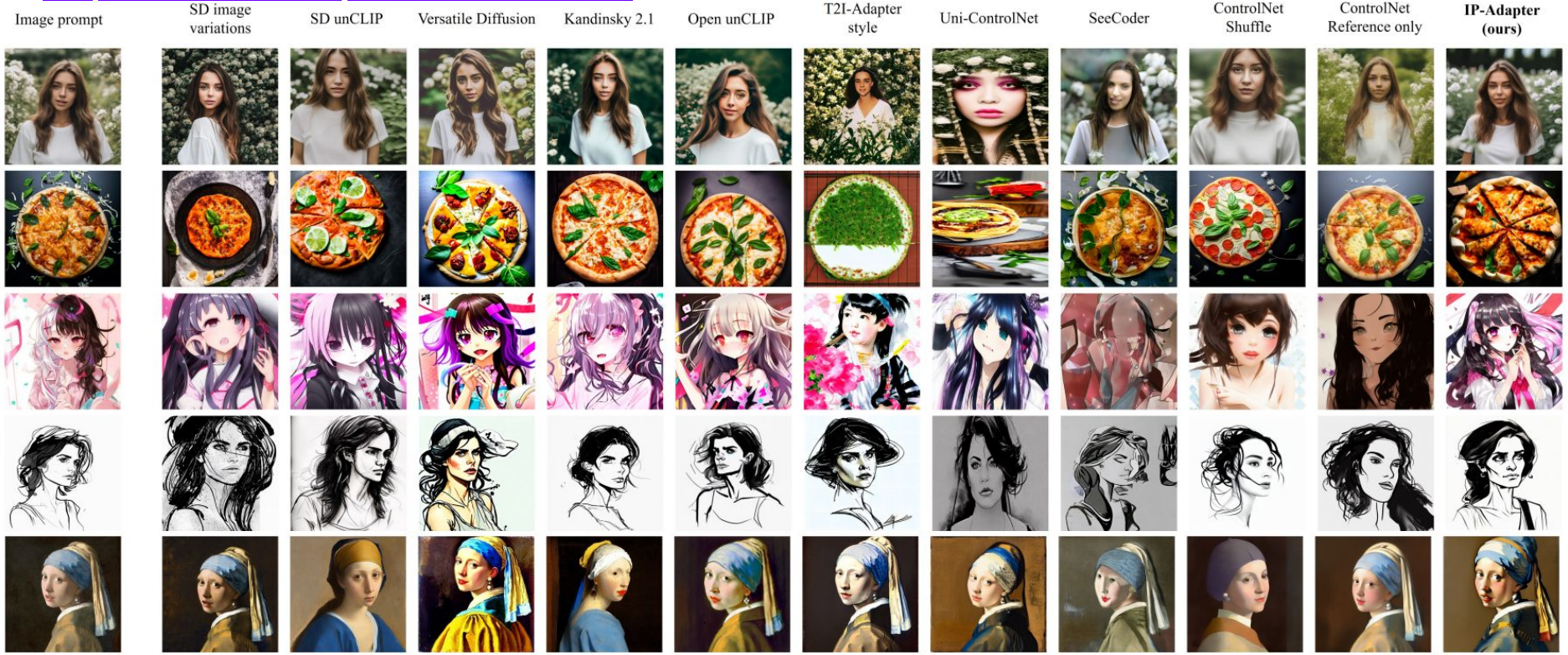


Figure 3: The visual comparison of our proposed IP-Adapter with other methods conditioned on different kinds and styles of images.

9/ Quantitative comparison

Table 1: Quantitative comparison of the proposed IP-Adapter with other methods on COCO validation set. The best results are in **bold**.

Method	Reusable to custom models	Compatible with controllable tools	Multimodal prompts	Trainable parameters	CLIP-T ↑	CLIP-I ↑
Training from scratch						
Open unCLIP	✗	✗	✗	893M	**0.608**	**0.858**
Kandinsky-2-1	✗	✗	✗	1229M	0.599	0.855
Versatile Diffusion	✗	✗	✓	860M	0.587	0.830
Fine-tunining from text-to-image model						
SD Image Variations	✗	✗	✗	860M	0.548	0.760
SD unCLIP	✗	✗	✗	870M	0.584	0.810
Adapters						
Uni-ControlNet (Global Control)	✓	✓	✓	47M	0.506	0.736
T2I-Adapter (Style)	✓	✓	✓	39M	0.485	0.648
ControlNet Shuffle	✓	✓	✓	361M	0.421	0.616
IP-Adapter	✓	✓	✓	22M	**0.588**	**0.828**

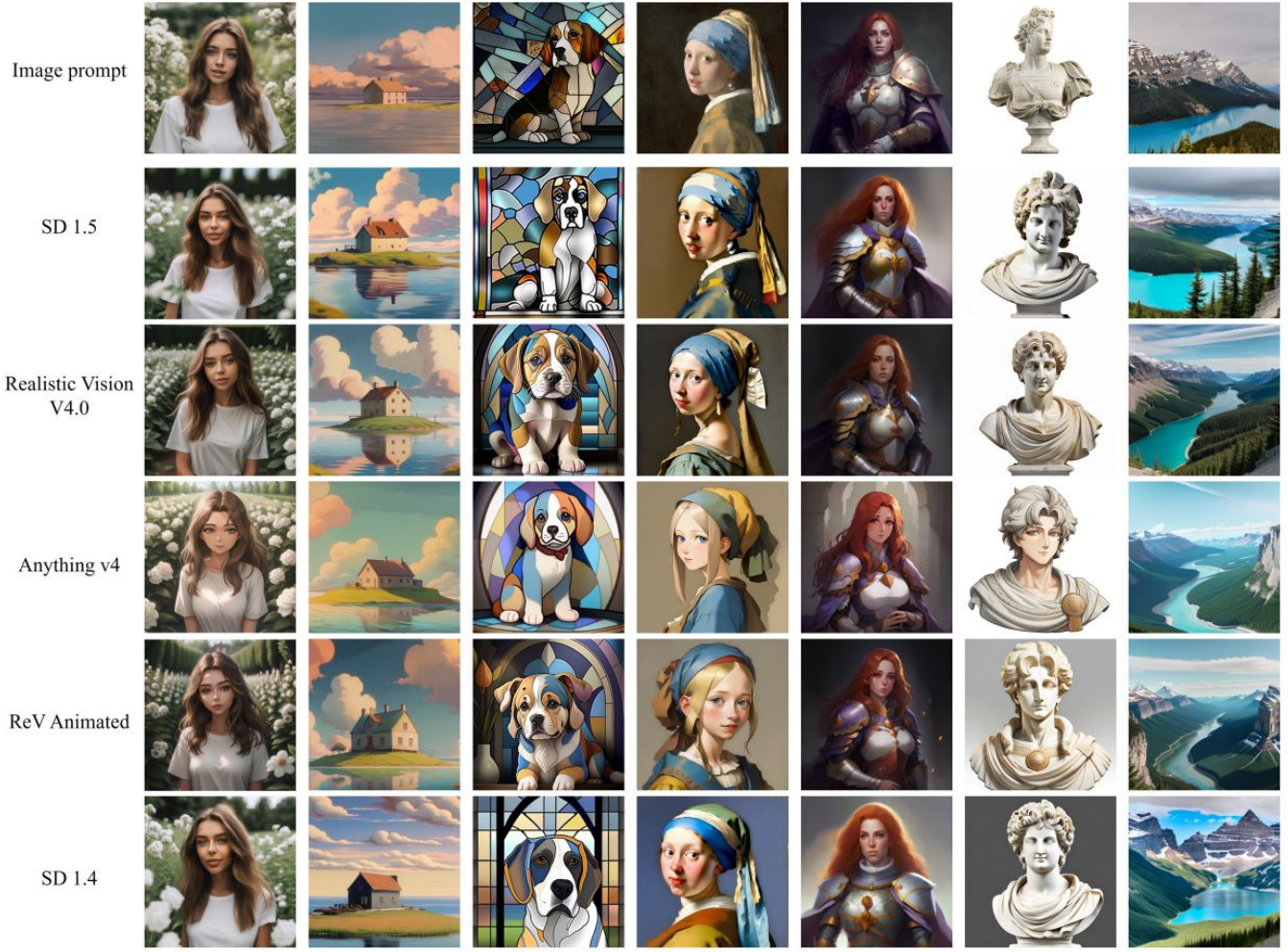


Figure 4: The generated images of different diffusion models with our proposed IP-Adapter. The IP-Adapter is only trained once.

11/ IP Adapter & ControlNet

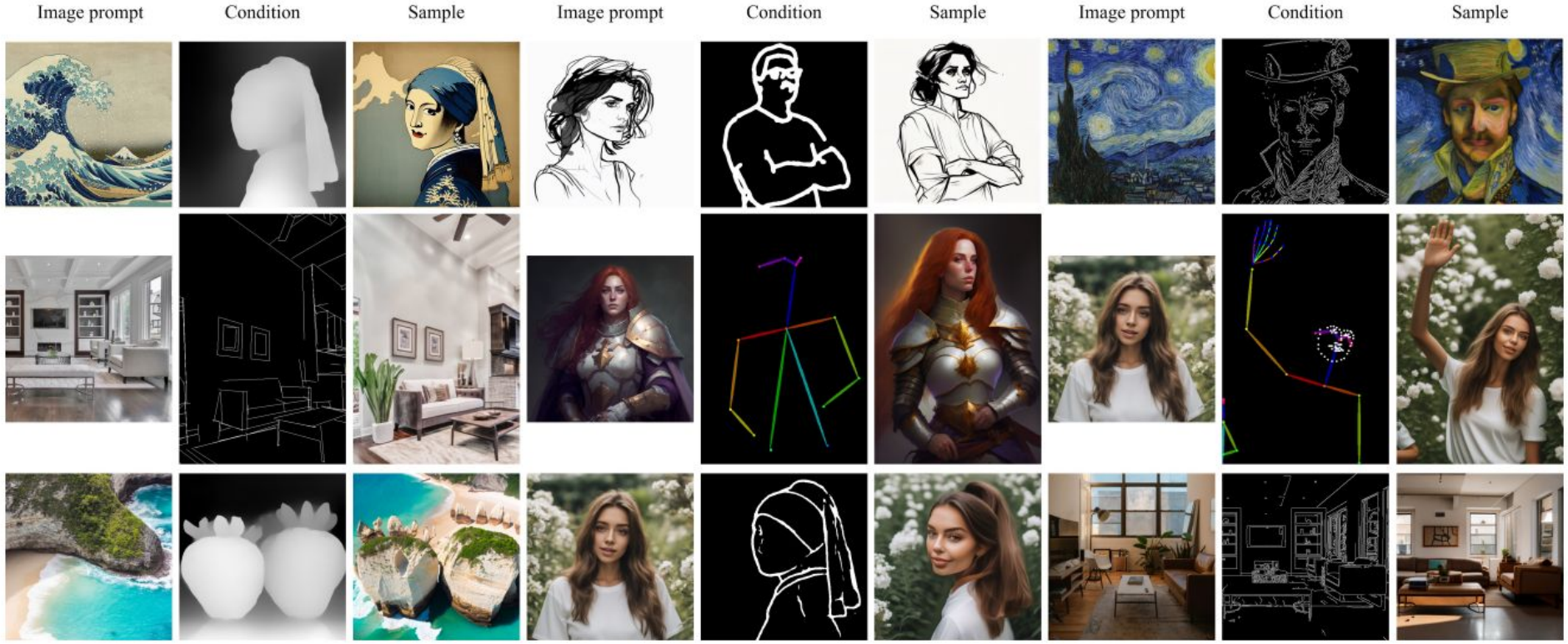


Figure 5: Visualization of generated samples with image prompt and additional structural conditions. Note that we don't need fine-tune the IP-Adapter.

12/ More comparison

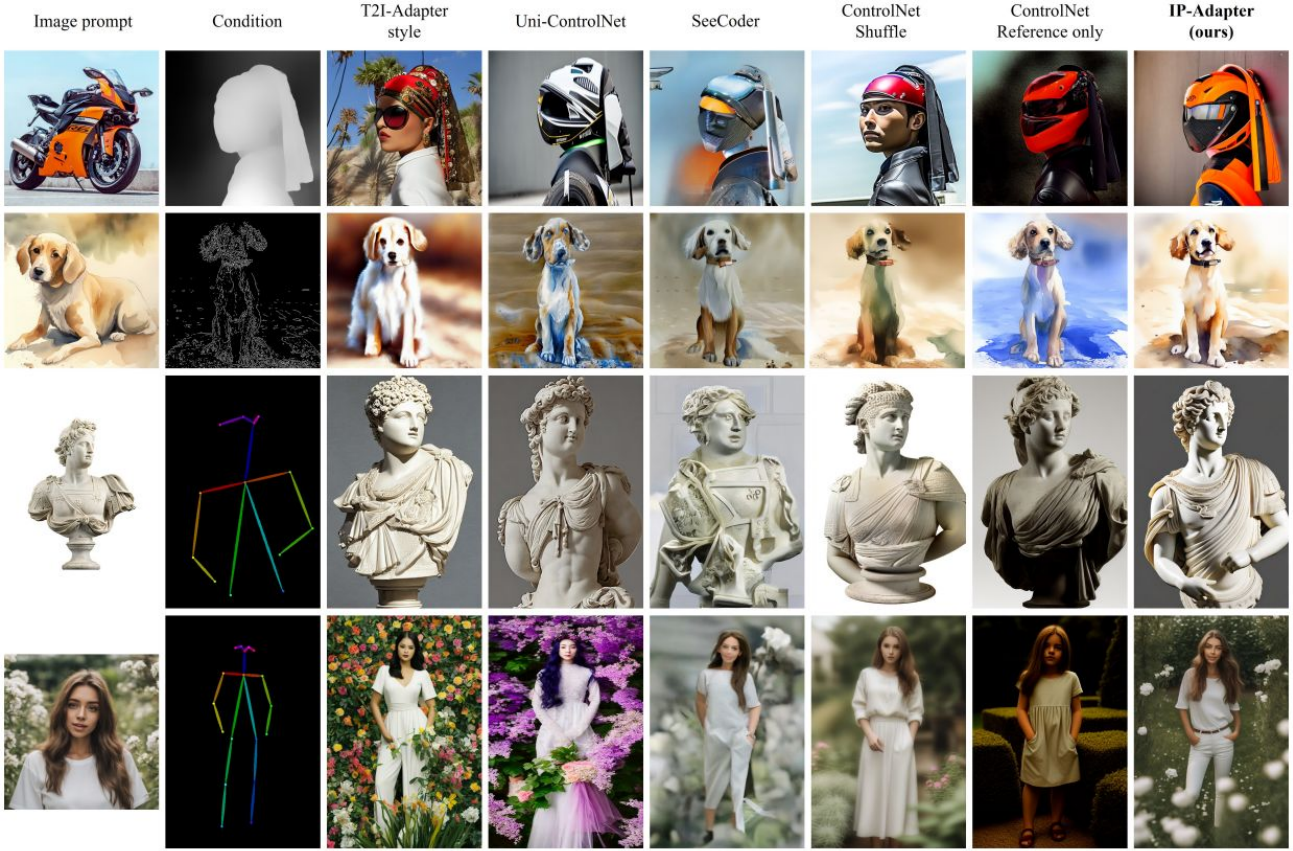


Figure 6: Comparison of our IP-Adapter with other methods on different structural conditions.

13/ Img2img & inpainting modes

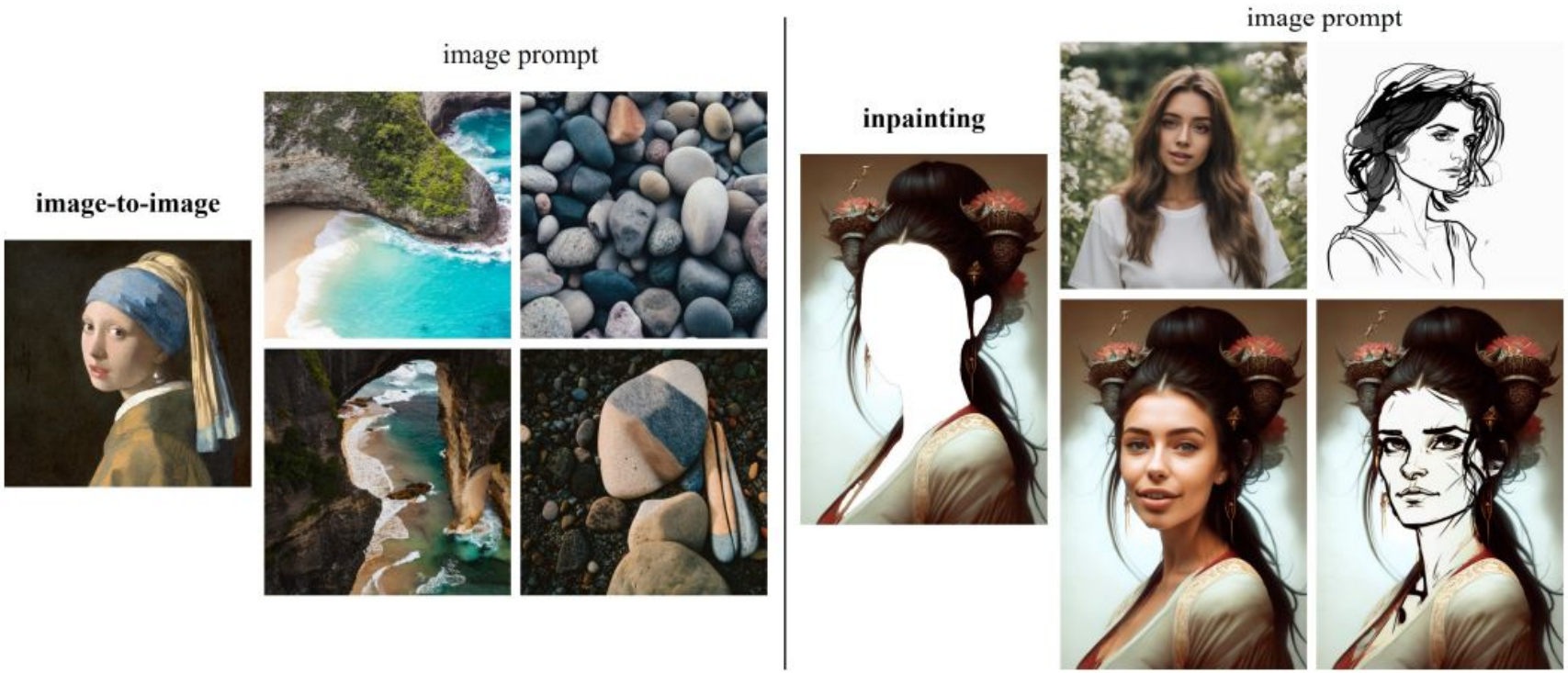


Figure 7: Examples of image-to-image and inpainting with image prompt by our IP-Adapter.

14/ Multimodal prompts

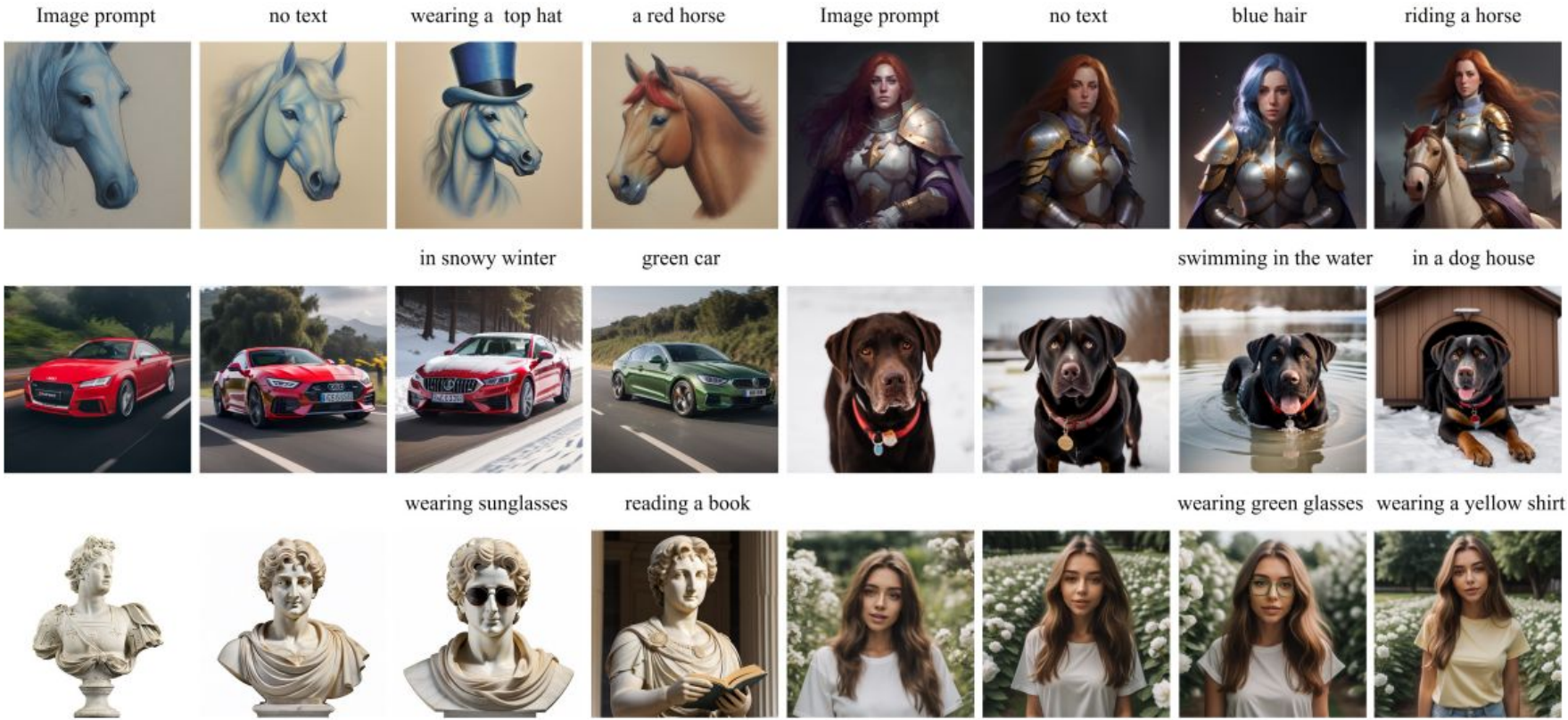


Figure 8: Generated examples of our IP-Adapter with multimodal prompts.

15/ Multimodal prompts

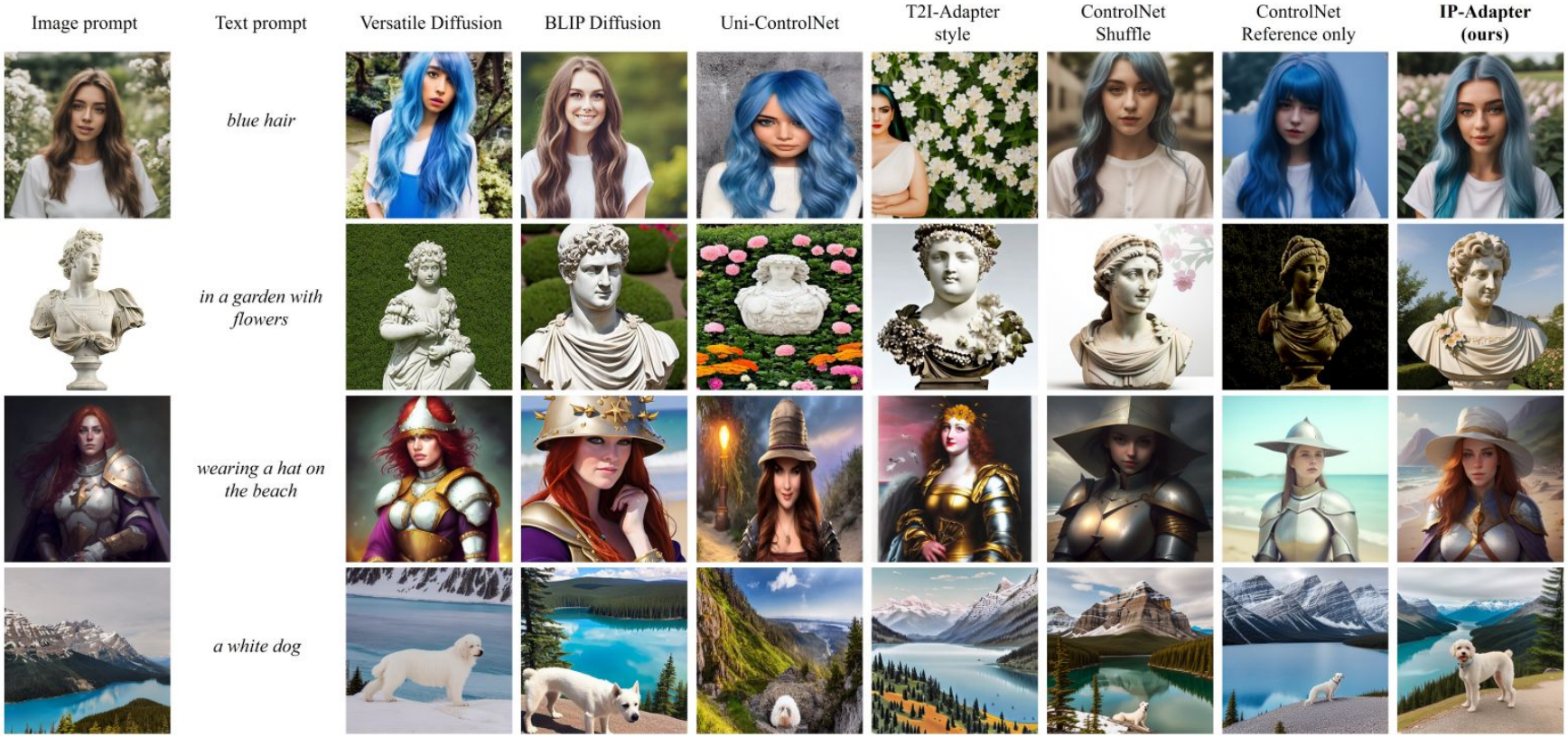


Figure 9: Comparison with multimodal prompts between our IP-Adapter with other methods.

16/ Ablation

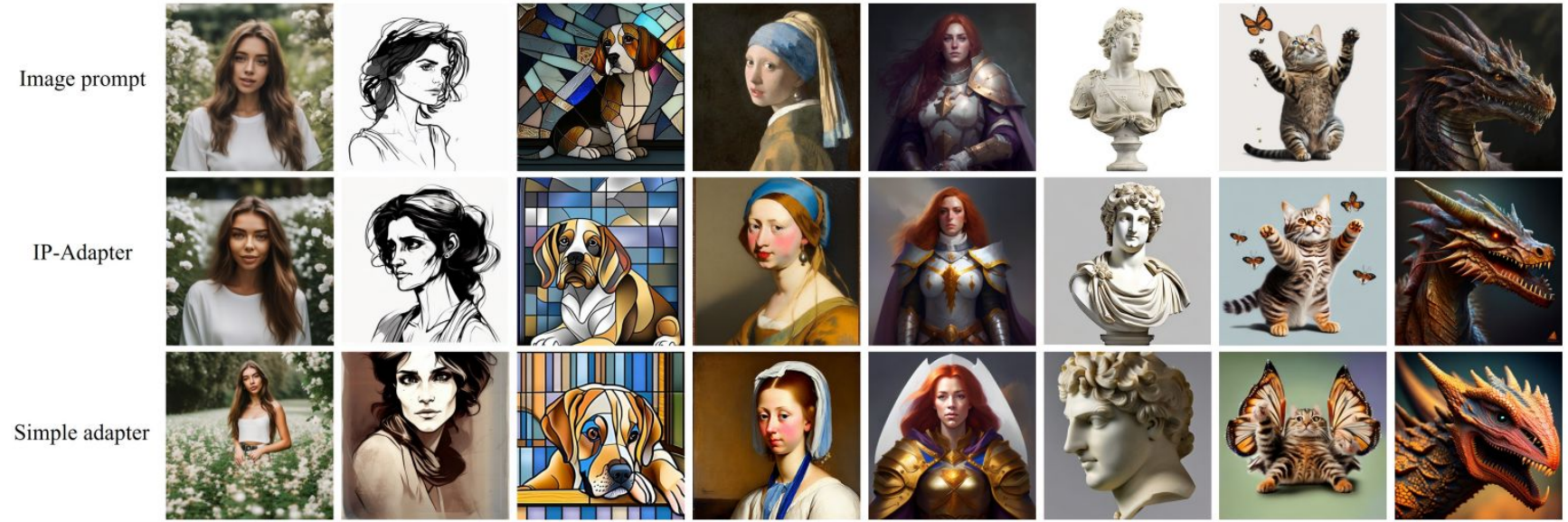


Figure 10: Comparison results of our IP-Adapter with simple adapter. The decoupled cross-attention strategy is not used in the simple adapter.

cross-attention layers. For a fair comparison, we trained both adapters for 200,000 steps with the same configuration. Figure 10 provides comparative examples with the IP-Adapter with decoupled cross-attention and the simple adapter. As we can observe, the IP-Adapter not only can generate higher quality images than the simple adapter, but also can generate more consistent images with image prompts.