Plug-and-Play Diffusion Distillation

Supplementary Material

A. More visualizations on guide model with Latent Diffusion Models

In this section, we provided more visualizations of our methods with stable diffusion v1.5. The figure is shown in Figure 9. This part aims to show that our approach can generate a variety of styles based on the Text prompts. Note that the initial noise in the images generated by the CFG and Full guide model with 16 steps (i.e. first two rows) is identical, but the initial noise for other methods is different.

B. User study

One of the characteristics observed from the injection-based conditioned model (e.g. ControlNet) is that the generated images are more saturated and have higher contrast. Some perceive them as less realistic while others may find them more visually pleasing. We conducted a user study where users were presented with a text prompt along with a pair of images generated from that text prompt (Student Full model 50 steps vs Teacher 50 steps) in a sequential fashion. They were asked to choose the preferred image based on image quality and text-image alignment. In the study, 90 participants collectively assessed a total of 680 unique text-image pairs, resulting in the accumulation of 1.8k votes. The vote distribution indicates that users did not strongly favor the teacher, with 1005 votes (55.65 %) in favor of the teacher and 801 votes (44.35%) in favor of the student.

C. Discussion on model performance with low guidance number

We observe that FID scores of our methods are relatively high when guidance is small (g=2, 4, 6). Due to the formulation of the guidance model, when the guidance value is small, the injection noise is small (as depicted in Figure 7). Therefore, the g=0 corresponds to not using CFG at all, which is known to generate low-quality images. However, when guidance is higher (g=8) our model is comparable to the teacher model.

D. Other Layers in the Feature Maps

In this section, we tried to display all the other feature map injection layers from our guide model. The corresponding figure is shown in Figure 11. Generally, other layers also show that at the beginning of the sampling, the feature map injections are stronger. But there may also be some layers (6th layer, counting from top to bottom) that show an inverse trend.

E. Text prompts

In this section, we will show the precise text prompts of the generated images displayed in the main paper. The text prompt will be ordered from left to right, from top to bottom respectively.

E.1. Text prompts in Figure 2

(a) 3D cartoon style

1. A person on a racing motorcycle making a sharp right turn.
2. A businessman tying a necktie.
3. A plate of french fries and a hamburger and coleslaw.
4. A cat that is looking up while sitting down.
5. Little girl holding a stuffed bunny rabbit toy.
6. This is a bird looking in the direction of a tree.

(b) Watercolor style

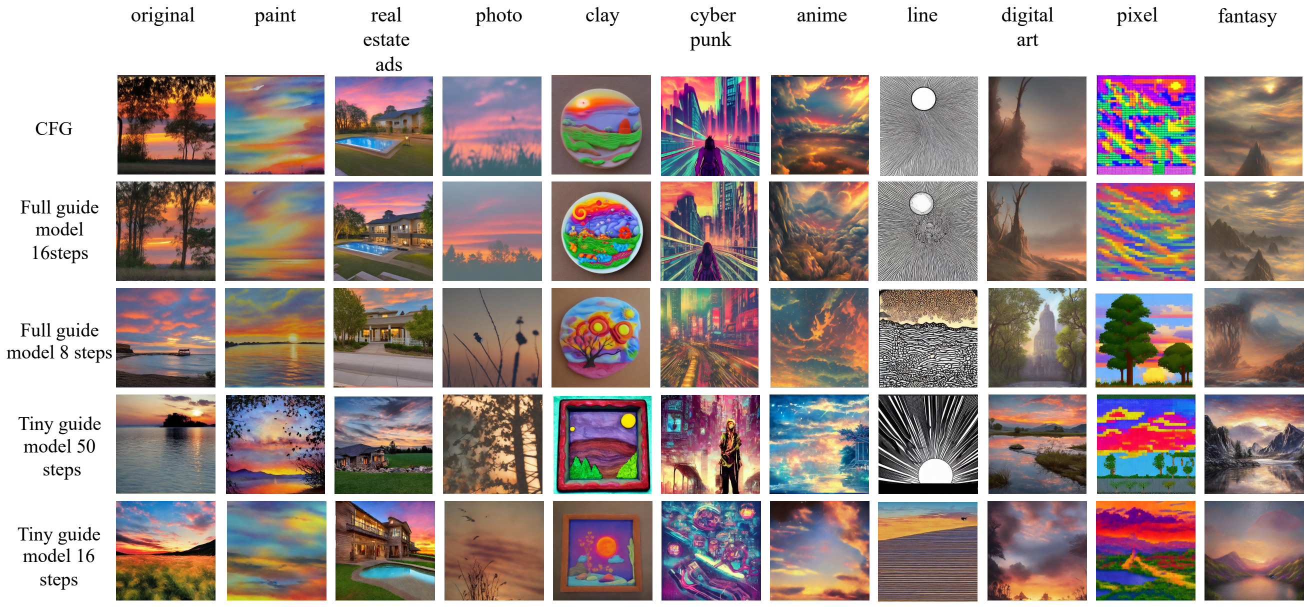
1. A boy walking across a field while flying a kite.
2. An arrangement of yellow flowers with one white flower.
3. There is a cutting board and knife with chopped apples and carrots.
4. A woman walking under one umbrella in the rain.
5. Little girl holding a stuffed bunny rabbit toy.
6. This is a bird looking in the direction of a tree.

(b) Realistic Style

1. A dog is wearing a fluffy hat.
2. A vase holds green leaves and red flowers.
3. A wooden park bench with colorful leaves on the ground.
4. two bears giving each other a nose kiss
5. Little girl holding a stuffed bunny rabbit toy.
6. This is a bird looking in the direction of a tree.

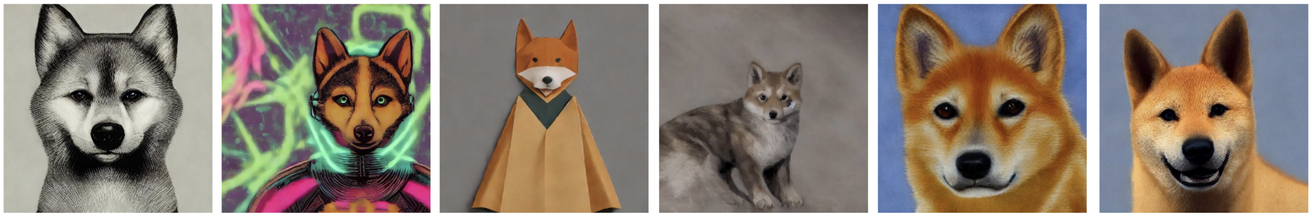
E.2. Text prompts in Figure 6

1. A snow-covered road in rural environment with forest and hills in the swiss alps near schwarzenberg in the canton of lucerne, Switzerland
2. A bowl of soup sitting on a wooden cutting board
3. Jars with different smoothies close-up
4. A person with a short blond hair is looking at the camera
5. Happy female tourist looking sideways while sitting on colorful wooden bridge at sea viewpoint against cloud on blue sky background
6. Satisfied forty years old European woman feels relaxed awakes early enjoys new day wears casual pajama embraces soft blanket rests long during day off



Text prompt: "A beautiful sunset" + style-related descriptions

Figure 9. More visualizations on our methods working with stable diffusion v1.5. Our approach can generate different styles effectively.



Text prompt: "a shibainu" + style related descriptions

Figure 10. More results for tiny guide model 16 steps

E.3. Text prompts in Figure 5

1. A detailed close-up of a cat facing the camera. Its eyes are a striking feature. Vivid and expressive. The fur is meticulously rendered. Showcasing individual strands and the subtle play of light and shadow. Whiskers stand out sharply against a softly blurred background.
2. Long-exposure night photography of a starry sky over a mountain range. with light trails. award winning photography
3. beautiful woman wearing fantastic hand-dyed cotton clothes. embellished beaded feather decorative fringe knots. colorful pigtail. subtropical flowers and plants. symmetrical face. intricate. elegant. highly detailed. 8k. digital painting.
4. b&w photography. model shot. beautiful detailed eyes. professional award winning portrait photography. Zeiss 150mm f/2.8. highly detailed glossy eyes.

E.4. Text prompts in Figure 8

1. elephants standing on top of a grass-covered field.
2. A castle-like building is in the background while the foreground is a green grass lawn, part of which has been mowed.
3. A teddy bear sitting on the grass.
4. A man performs a trick on a running horse in an enclosure
5. The man is riding his motorcycle around the bend.
6. A dog is wearing a fluffy hat.

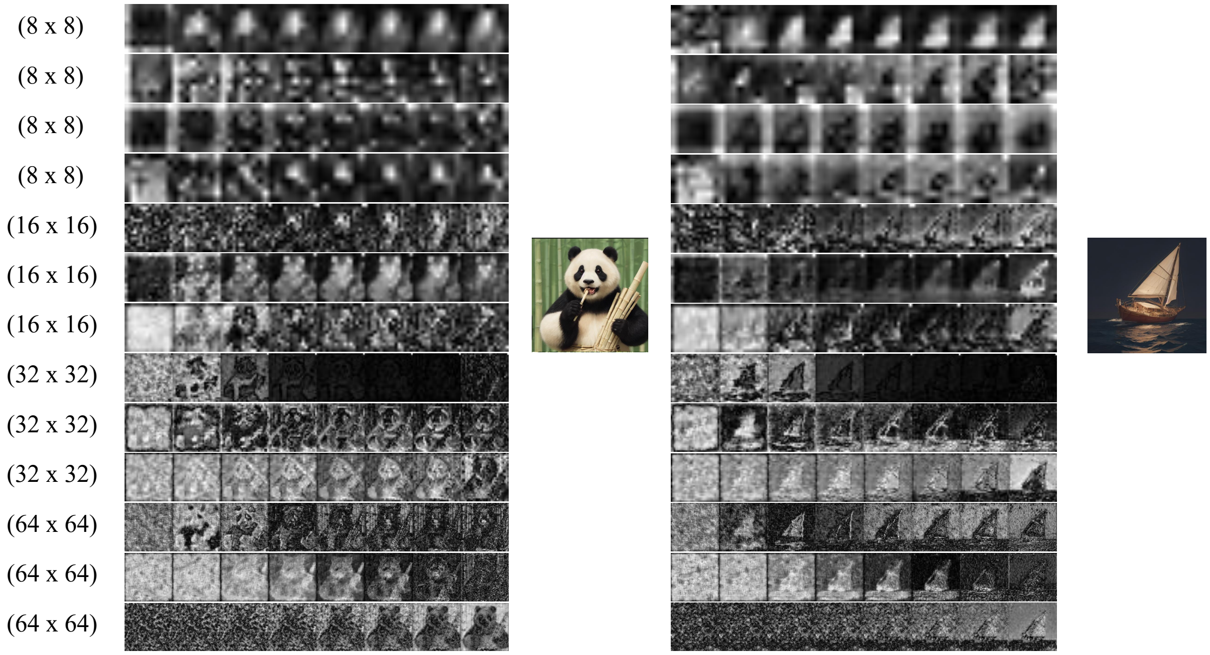


Figure 11. 13 layers of feature map injections across different sampling steps.