A. Limitations

Generation using universal guidance is typically slower than standard conditional generation for several reasons. Empirically, multiple iterations of denoising are required at every noise level t to generate high-quality images with complex guidance functions. However, the time complexity of our algorithm scales linearly with the number of recurrence steps k, which slows down image generation when k is large. Also, as demonstrated in the main paper, backward guidance is required in certain scenarios to help generate images that match the given constraint. Computing backward guidance requires performing minimization with a multi-step gradient descent inner loop. While proper choices of gradient-based optimization algorithms and learning rate schedules significantly speed up the convergence of minimization, the time it takes to compute backward guidance inevitably becomes longer when the guidance function is itself a very-large neural network. Finally, we note that, to get optimal results, sampling hyper-parameters must be chosen individually for each guidance network.

B. CLIP guidance for ImageNet diffusion model

Figure 10. We show that unconditional diffusion models trained on ImageNet can be guided with CLIP to generate high-quality images that match the text prompts, even if these generated images should be *out of distribution*.

CLIP Guidance. We use the same construction of f and ℓ for Stable Diffusion to perform CLIP-guided generation. We use only forward guidance for this experiment. To assess the limit of our universal guidance algorithm, we hand-crafted text prompts such that the matching images are *expected to be out of distribution*. In particular, our text prompts either designate art styles that are far from realistic or designate objects that do not belong to any possible class label of ImageNet. We present the results in Fig. 10, and from the results, we clearly see that our algorithm still successfully guides the generation to produce quality images that also match the text prompts. For all three images, we have $s(t) = w \cdot \sqrt{1-\alpha_t}$, where w is 2, 5 and 2 respectively and k is 10, 5 and 10 respectively.

C. More results on Stable Diffusion

(a) Walker hound, Walker foxhound on snow.

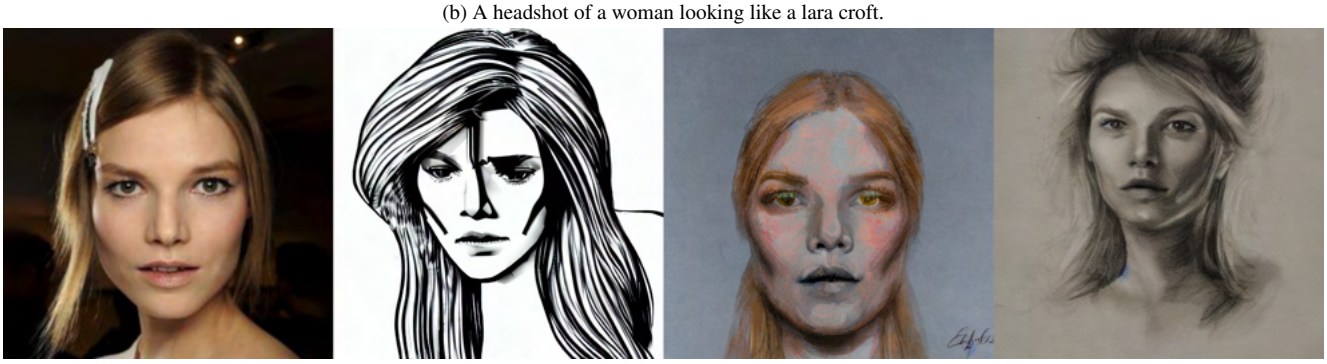
(b) Walker hound, Walker foxhound under water.

(c) Walker hound, Walker foxhound as an oil painting.

Figure 11. More images to show Segmentation guidance. In each subfigure, the first image is the segmentation map used to guide the image generation with its caption as its text prompt.

(a) Headshot of a person with blonde hair with space background.

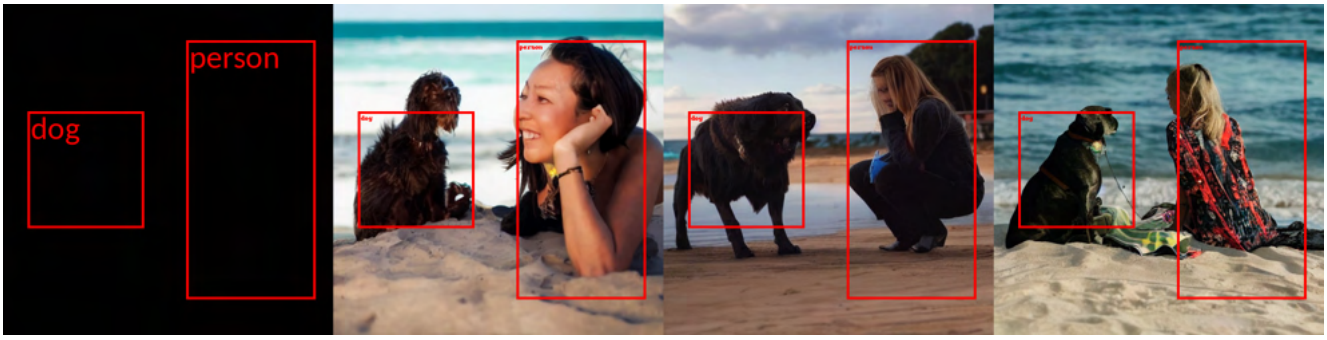
(b) A headshot of a woman looking like a lara croft.



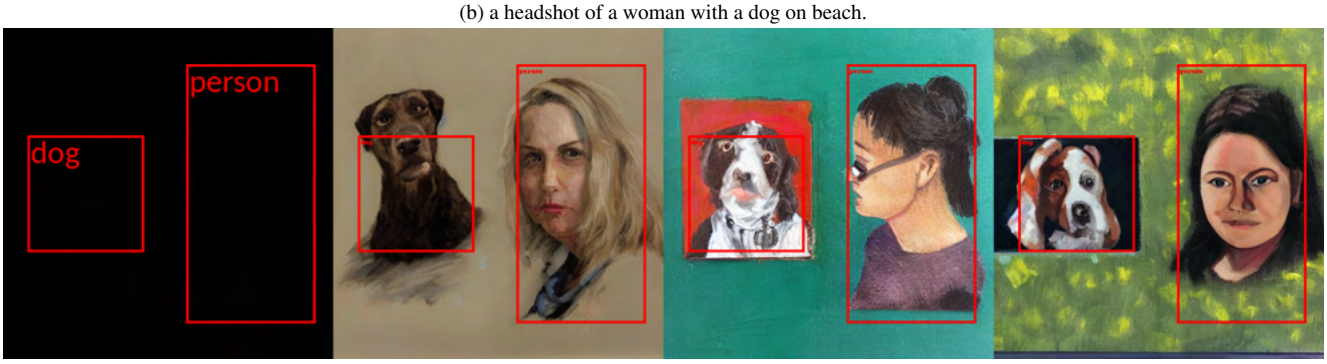
(c) A headshot of a blonde woman as a sketch.

Figure 12. More images to show Face guidance. In each subfigure, the first image is the human identity used to guide the image generation with its caption as its text prompt.

(a) A headshot of a woman with a dog in winter.



(b) a headshot of a woman with a dog on beach.

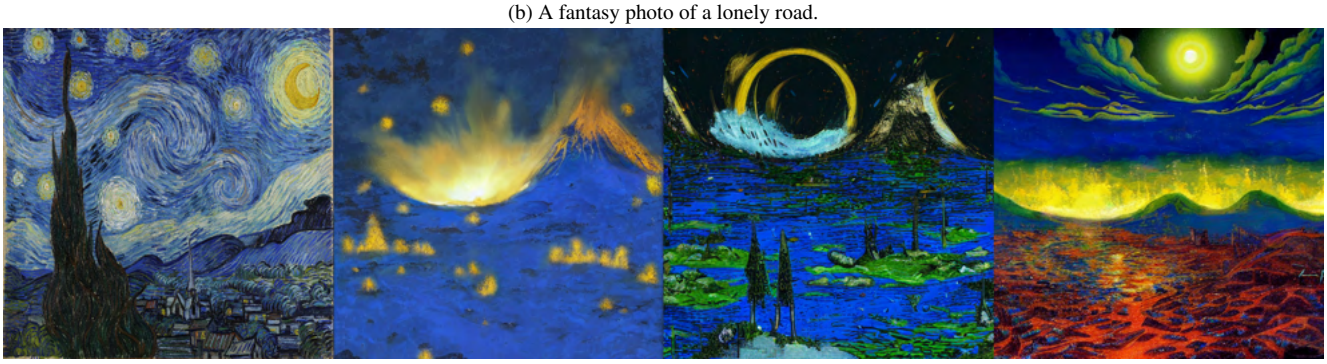


(c) An oil painting of a headshot of a woman with a dog.

Figure 13. More images to show Object Location guidance. In each subfigure, the first image is the object location used to guide the image generation with its caption as its text prompt.

(a) A colorful photo of an Eiffel Tower.

(b) A fantasy photo of a lonely road.



(c) A fantasy photo of volcanoes.

Figure 14. More images to show Style Transfer. In each subfigure, the first image is the styling image used to guide the image generation with its caption as its text prompt.