Supplementary Materials for “More Control for Free! Image Synthesis with Semantic Diffusion Guidance”

Xihui Liu[1,4*] Dong Huk Park[1] Samaneh Azadi[1] Gong Zhang[2,3]
Arman Chopikyan[2] Yuxiao Hu[2] Humphrey Shi[2,3] Anna Rohrbach[1] Trevor Darrell[1]
[1]UC Berkeley [2]Picsart AI Research (PAIR) [3]University of Oregon [4]The University of Hong Kong

1. Implementation Details

1.1. Dataset Licenses

The images in LSUN dataset [4] (`https://www.yf.io/p/lsun`) are obtained by Google Image Search with Creative Commons licenses. The images of FFHQ dataset [1] (`https://github.com/NVlabs/ffhq-dataset`) are published in Flickr by their respective authors under either Creative Commons BY 2.0, Creative Commons BY-NC 2.0, Public Domain Mark 1.0, Public Domain CC0 1.0, or U.S. Government Works license. The FFHQ dataset itself (including JSON metadata, download script, and documentation) is made available under Creative Commons BY-NC-SA 4.0 license by NVIDIA Corporation.

1.2. Model and Training Details

We adopt diffusion models pretrained on LSUN dataset from `https://github.com/openai/guided-diffusion` and diffusion models pretrained on FFHQ dataset from `https://github.com/jychoi118/ilvr_adm` for our SDG experiments. We finetune CLIP ResNet 50× 16 model [3] from `https://github.com/openai/CLIP` on the noised images with initial learning rate 10^{-4} and weight decay 10^{-3}, with a batch size of 256. On FFHQ dataset, the learning rate decays by a factor of 0.1 every 3,000 iterations, and the model is trained for 14,000 iterations. On LSUN cat, LSUN horse, and LSUN bedroom datasets, the learning rate decays by a factor of 0.1 every 30,000 iterations, and the model is trained for 100,000 iterations. When synthesizing images with our SDG, the scaling factor is a hyperparameter that we manually adjust for each guidance. Code will be released upon acceptance.

1.3. Text Instructions for Evaluating Text-to-Image Synthesis

In Section 4.2 of the main paper, we evaluate text-to-image synthesis with 400 text instructions for FFHQ face images. The text instructions are defined based on compositions of gender and face attributes from CelebA-Attributes [2]. We provide detailed information on how to generate the 400 text instructions. The sentence is in the structure of “A photo of {a / a chubby / a smiling / an attractive} {man / woman / girl / boy} with {*attribute*}”, where *attribute* is one of the following: bags under eyes, big lips, big nose, black hair, blond hair, brown hair, red hair, bushy eyebrows, double chin, eyeglasses, high cheekbones, slightly open mouth, narrow eyes, oval face, pale skin, pointy nose, rosy cheeks, a hat, short hair, straight hair, curly hair. Additionally, we define 9 attributes specifically for women: arched eyebrows, bangs hair, heavy sunglasses, wavy hair, earrings, lipstick, necklace, long hair, bob-style hair; 5 attributes specifically for men: goatee, mustache, no beard, sideburns, necktie; and 1 attribute “gray hair” for both men and women. The compositions of the above descriptions, genders, and attributes result in 400 unique text instructions.

2. Qualitative Results

We show more examples of images generated by our SDG model with image guidance (Figure 1) and language guidance (Figure 2). Our model is able to generate diverse images that semantically match the guidance signal. In Figure 3 we illustrate multimodal guidance, where both image and language guidance are provided. The model is able to generate images that are semantically consistent with both guidance signals. In the last three rows of Figure 3, we demonstrate that the model is able to take out-of-domain images as guidance, and generate real images according to the semantic guidance.

Our method aims at generating diverse images according to the semantic guidance, which is different from the goal of image editing. Specifically, our approach generates di-

[*] This work was done when Xihui Liu was a postdoc in UC Berkeley.

verse images, while image editing aims at generating a single image that preserves most information (e.g., face identity) from the input image.

3. Effect of Different Scale Factors

As explained in Section 3.1 and Algorithm 1 of the main paper, there is a user-controllable scaling factor s that controls the diversity and semantic consistency of the synthesized images. We illustrate the effect of the scaling factor in Figure 4, where each row shows images synthesized by a fixed scaling factor, and each column shares the same random seed for sampling during the generation process. As the scaling factor becomes larger, the consistency between generated images and the guidance signal becomes better, while the diversity of generated images becomes lower.

4. Limitations and Failure Cases

Figure 5 shows a failure case of our SDG model. When facing novel compositions of concepts that are not common in the training set, the model might have a difficulty disentangling different concepts. As shown in the example, given the language instruction "A photo of a bedroom with yellow curtains", the model generates images of bedrooms with yellow bed or yellow walls. Better vision-language embedding models, language-guided masking or attention schemes, and compositional image synthesis models may hopefully address this problem.

We observe that some images from the LSUN training set have watermarks, so the model learns the bias from the training set and generates images with similar text and watermark patterns.

5. Social Impact

Our Semantic Diffusion Guidance (SDG) model can help designers and artists design arts or generate images as desired in a controllable way. The image guidance and language guidance provides a simple and intuitive interface for users. However, image synthesis models have as much potential for misuse in applications as they have for beneficial applications. We should be aware of the potential negative social impact if image synthesis is used for generating fake images, fake videos, or fake news to mislead people. Especially, we should be cautious about using the image synthesis models for face synthesis. In addition, the CLIP [3] model used in our language diffusion guidance is pretrained on large-scale image-caption pairs from the web, which might encodes undesired biases that could propagate to our image synthesis process. Researchers should be aware of the potential biases encoded by the pretrained CLIP model.

References

[1] Rinon Gal, Or Patashnik, Haggai Maron, Gal Chechik, and Daniel Cohen-Or. Stylegan-nada: Clip-guided domain adaptation of image generators. *arXiv preprint arXiv:2108.00946*, 2021.

[2] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. *arXiv preprint arXiv:2103.00020*, 2021.

[4] Fisher Yu, Ari Seff, Yinda Zhang, Shuran Song, Thomas Funkhouser, and Jianxiong Xiao. Lsun: Construction of a large-scale image dataset using deep learning with humans in the loop. *arXiv preprint arXiv:1506.03365*, 2015.

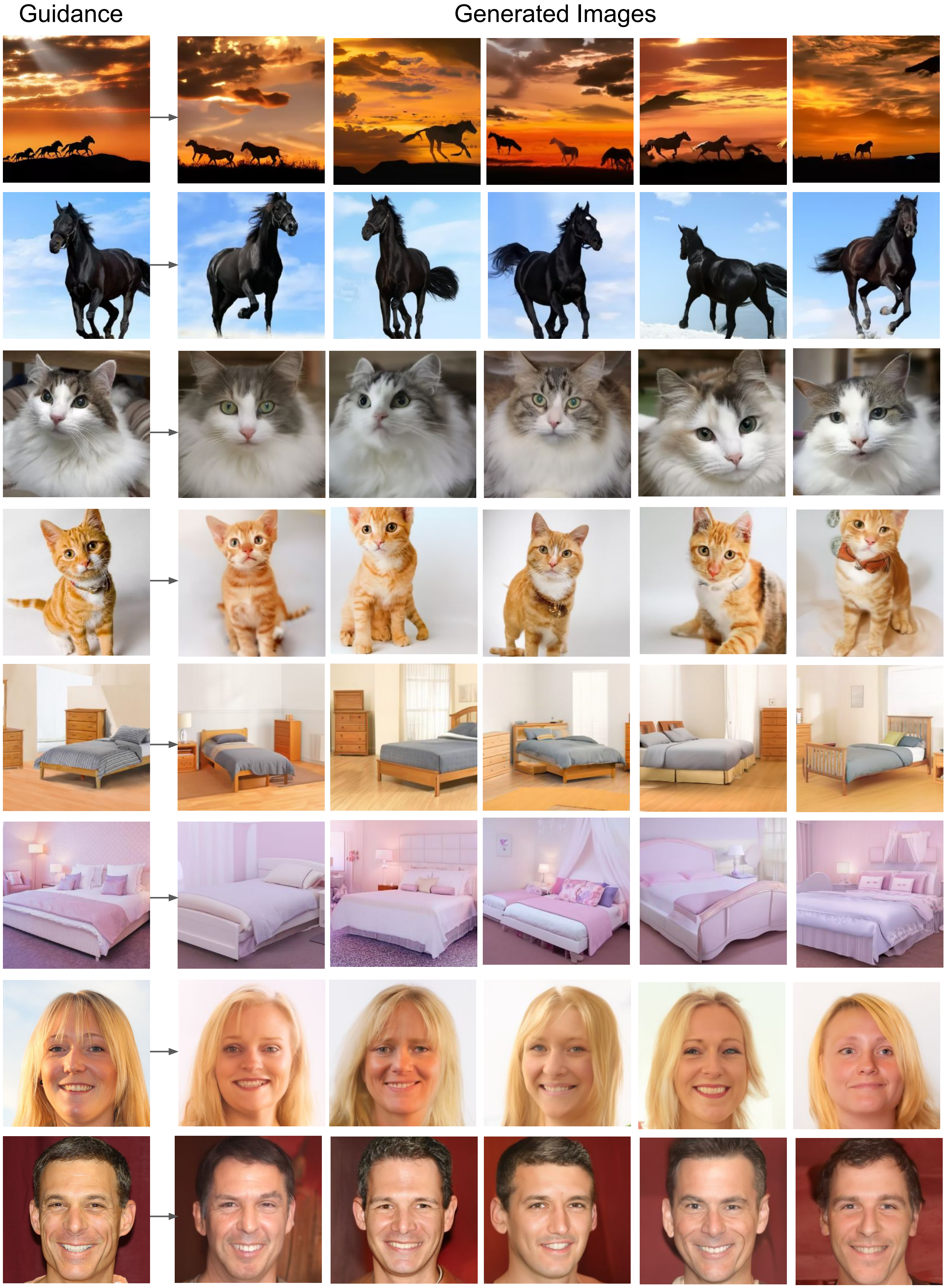

Figure 1: Image synthesis results with image guidance on LSUN and FFHQ datasets. Given a guidance image, the model is able to generate semantically similar images with different pose, layout, and structure.

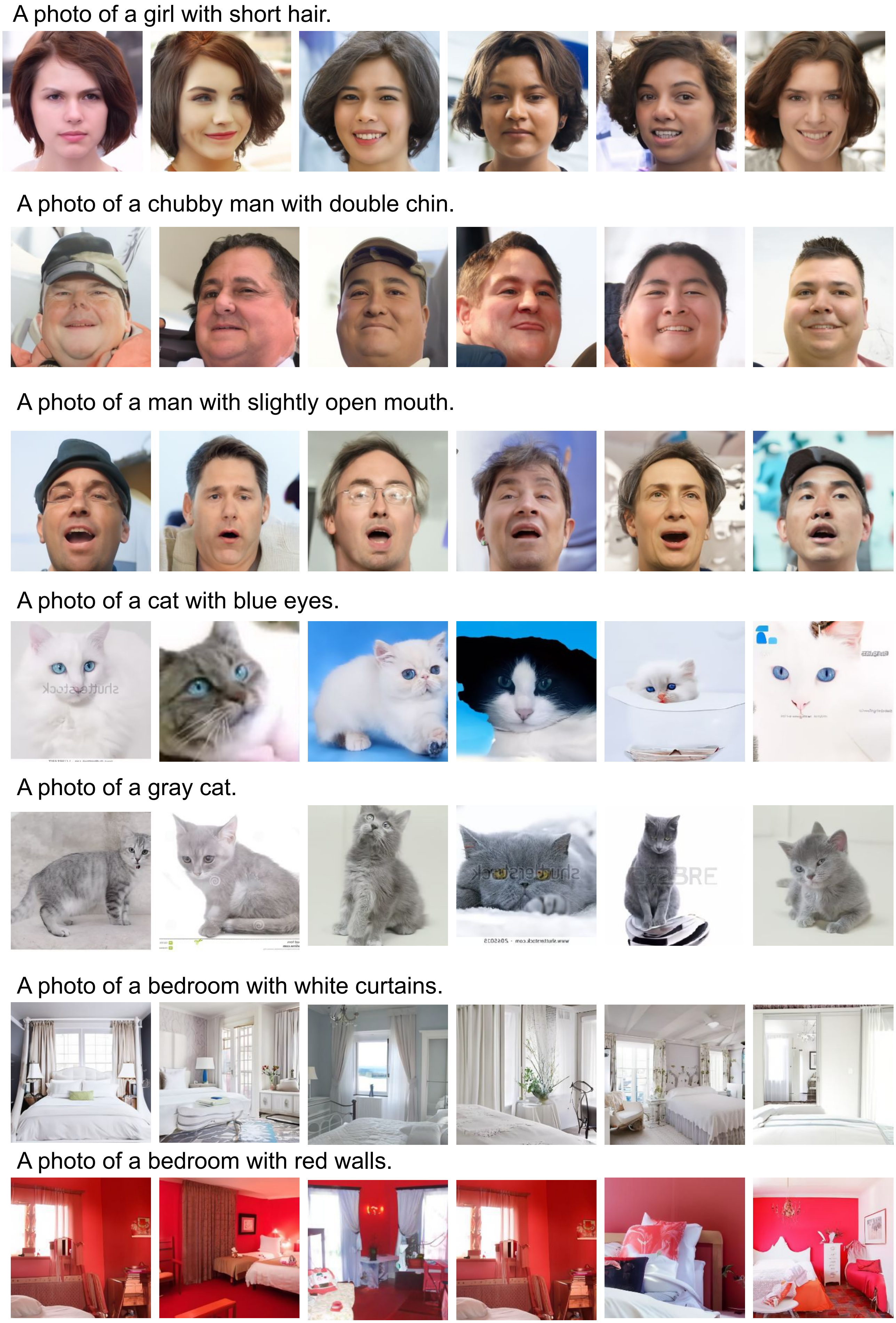
A photo of a girl with short hair.
A photo of a chubby man with double chin.
A photo of a man with slightly open mouth.
A photo of a cat with blue eyes.
A photo of a gray cat.
A photo of a bedroom with white curtains.
A photo of a bedroom with red walls.

Figure 2: Image synthesis results with language guidance on LSUN and FFHQ datasets.

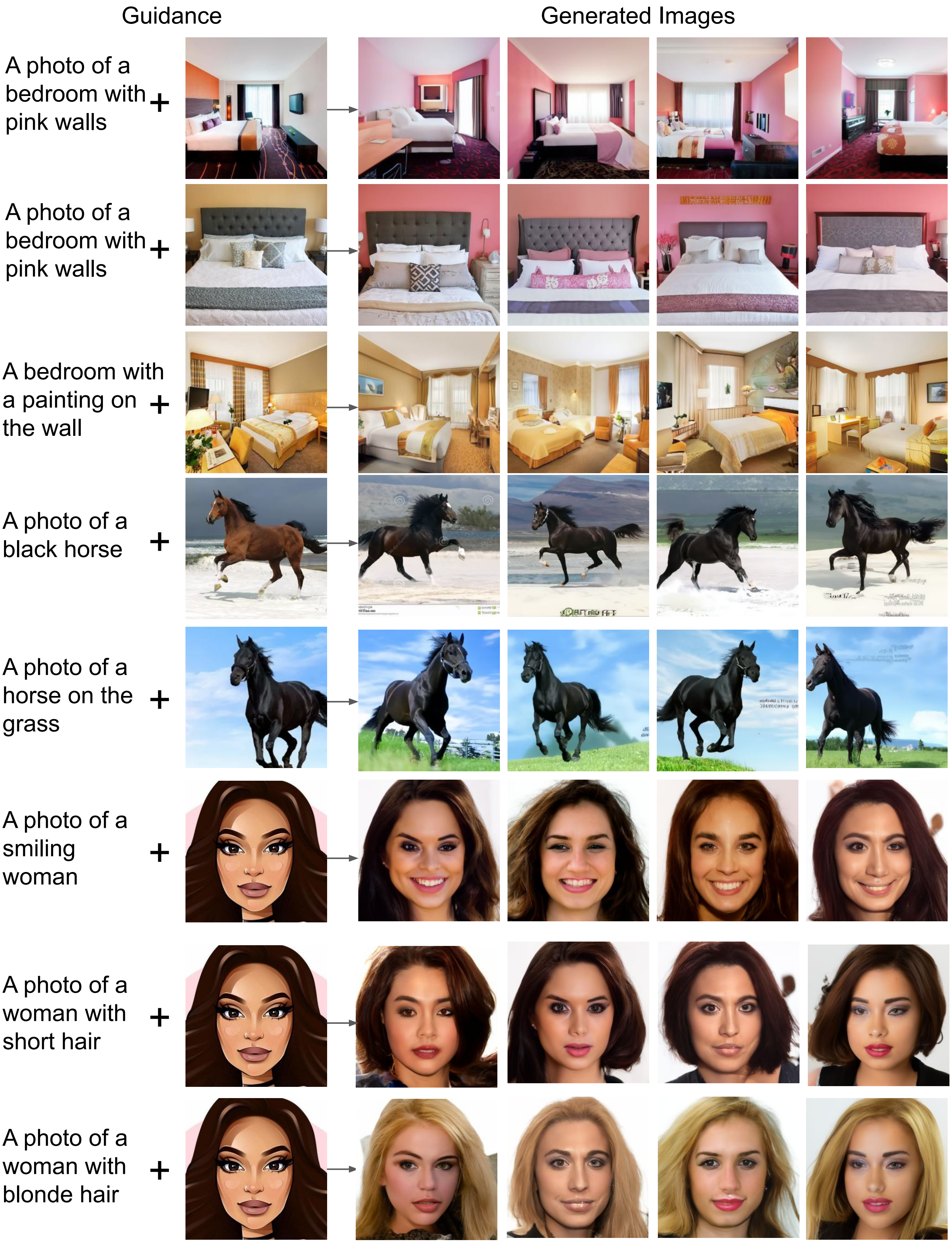
Guidance
Generated Images
A photo of a bedroom with pink walls +
A photo of a bedroom with pink walls +
A bedroom with a painting on the wall +
A photo of a black horse +
A photo of a horse on the grass +
A photo of a smiling woman +
A photo of a woman with short hair +
A photo of a woman with blonde hair +

Figure 3: Image synthesis results with both image and language guidance on LSUN and FFHQ datasets.

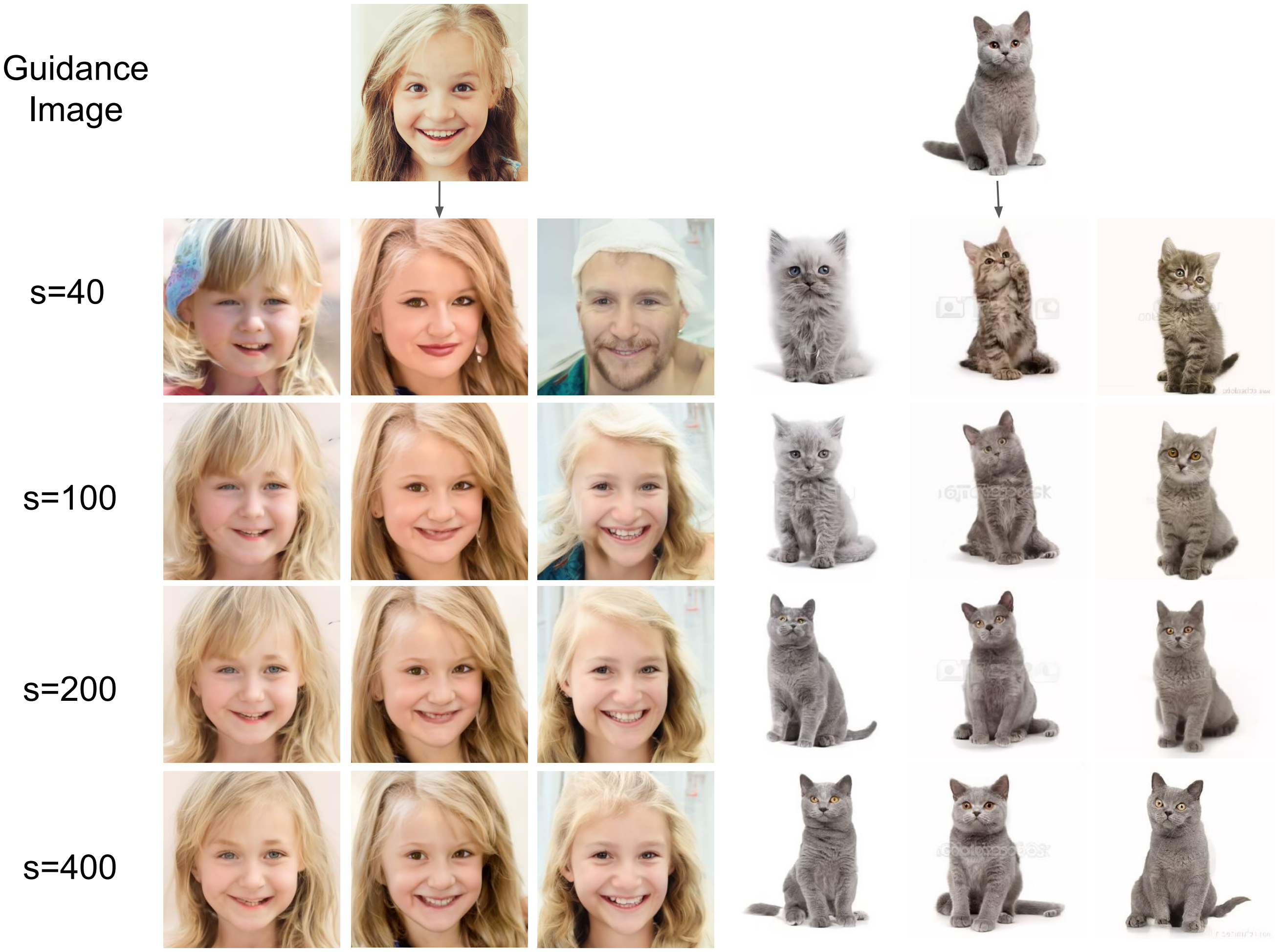

Figure 4: Image synthesis results with different scaling factors. s denotes the value of the scaling factor. Larger scaling factors result in lower diversity and more consistency with the guidance. The generated images in the same column share the same random seed for sampling.

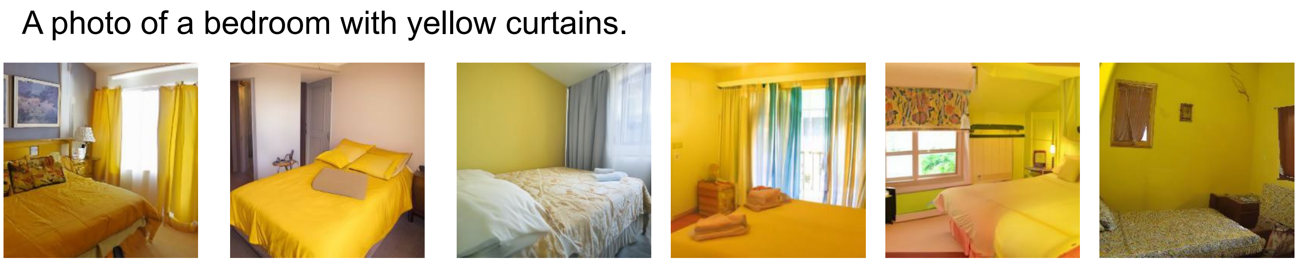

Figure 5: Failure cases of our SDG model.