Supplementary material for the paper "Conditioning Diffusion Models via Attributes and Semantic Masks for Face Generation"

1. Latent P2 Weighting

In this section, we provide a detailed derivation of the P2 weighting and how we introduce it in our method. DMs could be seen as a particular kind of Variational Autoencoder (VAE), which can be trained by optimizing a variational lower bound (VLB), $L_{vlb} = \sum_t L_t$. For each time step t, the loss function could be defined as:

$$L_t = \mathbf{E}_{\mathbf{x},\epsilon}\Big[\frac{\beta_t}{2\alpha_t(1-\bar{\alpha}_t)}||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||^2\Big], \tag{1}$$

where α_t, $\bar{\alpha}_t$ and β_t represent the variance schedule, ϵ is the target Gaussian noise and ϵ_θ is the parametrized U-Net model [2].

When Ho *et al.* proposed DDPM [2], they noticed that by removing the variance schedule-dependant coefficient, they obtained much better results and more stability at training time. Hence, they suggested using the following:

$$L^t_{simple} = \mathbf{E}_{\mathbf{x},\epsilon}\Big[||\epsilon - \epsilon_\theta(\mathbf{x}_t, t)||^2\Big]. \tag{2}$$

By removing the coefficient, the loss function is basically reweighted relative to the timestep term t, as we can see here:

$$L^t_{simple} = \lambda_t L_t, \qquad \lambda_t = \frac{2\alpha_t(1-\bar{\alpha}_t)}{\beta_t} \tag{3}$$

The reason why this kind of reweighting works is explained by Choi *et al.* [1]. They perform a broad analysis across different datasets, architectures and variance schedules in order to understand why the L_{simple} objective improved the perceived quality of the samples. By using perceptual measures like LPIPS [5], they separate the diffusion process into three stages, parametrized on a Signal-to-Noise Ratio (SNR) [3] depending on the variance schedule. These stages define when different levels of detail are lost during the diffusion, or vice-versa when they are generated in the denoising process. In the first stage of denoising, coarse details like color schemes and shapes are generated. Then, in the content stage, more distinguishable features come up. In the final stage, the fine-grained high-frequency details are refined and most of them are not perceivable by the human eyes. They propose a Perception Prioritized (P2) Weighting of DM's Loss function:

$$L^t_{P2} = \lambda'_t L_t, \qquad \lambda'_t = \frac{\lambda_t}{(k + SNR(t))^\gamma} \tag{4}$$

where λ_t is defined in Eq. (3), k is a stabilizing factor to avoid exploding weights for small SNR values, usually set to 1, and γ is an arbitrary exponent to give more or less importance to the reweighting. P2 is a generalization of the L_{simple} re-weighting, defined as follows:

$$Let\ \ \gamma = 0; \qquad \lambda'_t = \frac{\lambda_t}{(k + SNR(t))^\gamma} = \lambda_t; \tag{5}$$

By increasing the value of γ the weights shift towards the coarse and content phases, representing the earlier stages of the denoising process, giving less and less importance

to the loss terms corresponding to fine-grained unperceivable details.
We decided to test P2 in the latent space of LDM since no previous work reports it. Both techniques seem to bring great improvement to DMs and don't show apparent conflicts when combined. We chose to use the proposed γ values for the pixel-space dataset and analyze the experimental results. We then consider the default conditioned LDM loss function:

$$L^t_{LDM} = \mathbf{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\Big[||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(y))||^2\Big] \tag{6}$$

where z_t is the latent representation of the input image obtained by the Encoder $\mathcal{E}$ at diffusion timestep t, τ_θ is the condition encoder model and y is its input, which can be a segmentation mask, an attribute array, a text prompt or anything else. We then updated the objective by adding the P2 weighting term:

$$L^t_{LDM} = \mathbf{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\Big[\frac{||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \tau_\theta(y))||^2}{(k + SNR(t))^\gamma}\Big] \tag{7}$$

where the weight is defined in Eq. (4).

2. Supplementary Qualitative Results

Figures 1, 2 and 3 show additional qualitative results on Attributes-Conditioned Generation, Mask-Conditioned Generation, and Multi-Conditioned Generation respectively. The description of these experiments can be found in their relative section.

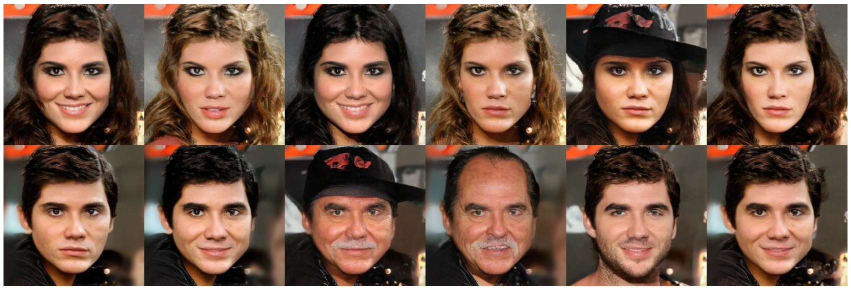

Figure 1: Additional results for Attributes-Conditioned Generation.

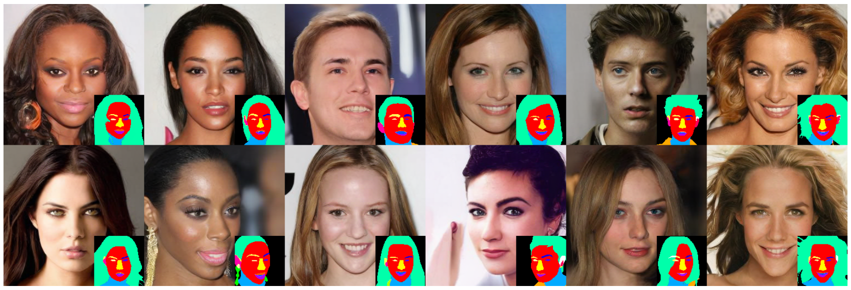

Figure 2: Additional results for Mask-Conditioned Generation.

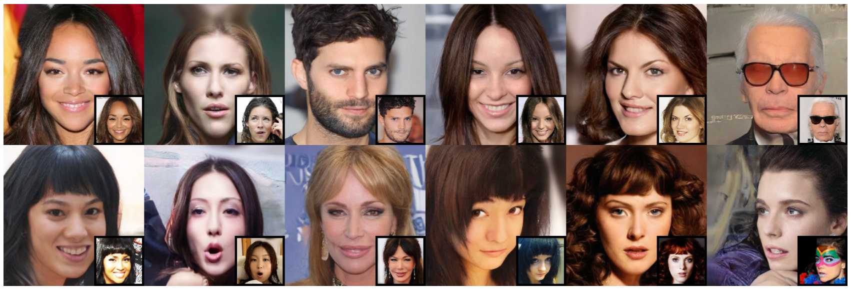

Figure 3: Additional results for Multi-Conditioned Generation.

3. Failure Cases

In this section, we want to report some of our failure cases, represented by non-realistic images. By looking through the generated samples, we noticed our unconditioned model rarely outputs unrealistic samples. Our conditioned models, though, sometimes produce bad samples, as shown in Fig. 4. This is a rare behavior since the faces in Fig. 4 are the only unrealistic results we were able to find among the 5K generated samples obtained using the segmentation masks of the validation set. By generating more samples conditioned on the same masks as the results reported in Fig. 4, we noticed there are two possible behaviors: *(1)* bad samples come from very peculiar masks, which are under-represented in the dataset, hence not reflecting the facial statistics learned by the model (see Fig. 5); *(2)* bad samples don't depend on bad segmentation masks but on a specific combination of mask and noise, where the conditioning strongly collides with the direction the noise is guiding towards, resulting in an unrealistic face. The noise indeed is relevant to the generated images since diffusion models tend to converge to similar latent spaces if they have the same variance schedule, as explained in [4] and shown in Fig. 1.

It is also worth noting that we didn't find any major case of non-faithful images, with respect to attributes and/or masks, through the thousands of generated samples. Our model tends to prioritize the conditioning injection to the image's quality, resulting in faithful but unrealistic generated samples. Quantitatively, this behavior is described by all the results for the conditioned tasks, reported in the main manuscript, where a high-fidelity batch of generated samples brings an increase in FID. Our multi-conditioned scenario fits in this behavior since it performed slightly worse in terms of FID than the single-condition models, but reached high fidelity for both conditionings.

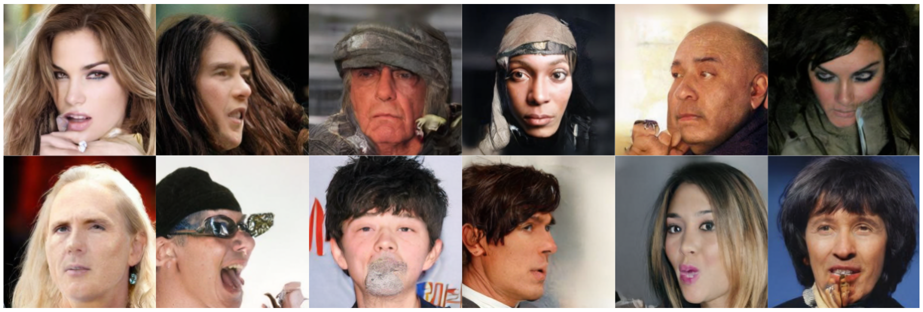

Figure 4: Worst handpicked samples generated by our mask-conditioned model.

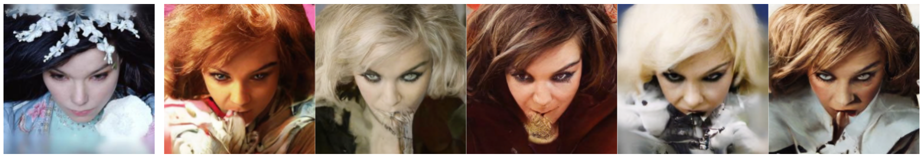

Figure 5: Failure cases generated by our mask-conditioned model on a peculiar mask. On the left, there's the original image from the validation set, while on the right we show our samples generated while conditioning on the reference image's semantic mask.

References

[1] Choi, J., Lee, J., Shin, C., Kim, S., Kim, H., Yoon, S., 2022. Perception prioritized training of diffusion models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 11472–11481.

[2] Ho, J., Jain, A., Abbeel, P., 2020. Denoising diffusion probabilistic models. Advances in Neural Information Processing Systems 33, 6840–6851.

[3] Kingma, D., Salimans, T., Poole, B., Ho, J., 2021. Variational diffusion models. Advances in neural information processing systems 34, 21696–21707.

[4] Wu, C.H., De la Torre, F., 2022. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. arXiv preprint arXiv:2210.05559 .

[5] Zhang, R., Isola, P., Efros, A.A., Shechtman, E., Wang, O., 2018. The unreasonable effectiveness of deep features as a perceptual metric, in: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 586–595.