Improving Latent Diffusion with Perceptual Mask-Aware Loss

Abhinav Gupta Poorvi Hebbar Vibhakar Mohta

Robotics Institute, Carnegie Mellon University
Pittsburgh, PA 15213
{ag6, phebbar, vmohta}@cs.cmu.edu

Abstract

Recent advances in AI-based image generation has made significant strides in text-image synthesis. In particular, diffusion models stand out by combining context using text prompts to generate very realistic and diverse images for text-to-image generation tasks. However, diffusion models struggle to understand and model the spatial and kinematic constraints of the world and therefore perform poorly in depicting complex objects like human faces, body extremities, etc. In this work, we aim to address some of the limitations of diffusion models, in particular, stable diffusion, by optimizing diffusion latents using a mask-aware loss on human faces and body. We hypothesize that conditioning on this loss function will guide the model into focusing on "important" aspects of image generation, like human faces and poses. We believe our work can serve as a foundation for fine-tuning pre-trained diffusion models on more sophisticated loss functions. Our project webpage can be found here and our code is available at Github.

1 Introduction

Generative modeling is a branch of unsupervised learning in which the aim is to learn the probability distribution of input data. Besides potentially helping in tasks like image-to-image translation, image-to-3D, inpainting, pose estimation, music synthesis, etc., the training data collection is reduced due to the unsupervised nature of learning.

Diffusion models(5; 17) particularly stand out as they do not suffer from mode collapse problems associated with Generative Adversarial Networks (GANs) (4; 13) or subpar image quality obtained using Variational Autoencoders (VAEs)(7; 12). Diffusion models typically consist of two Markov chains representing a forward diffusion process of iteratively perturbing input data by adding Gaussian noise and a reverse diffusion process which involves reconstructing the gaussian noise iteratively to an input signal from the original data distribution. Rombach et al. introduced a latent diffusion model, Stable Diffusion (15), which further improved the training and sampling efficiency of diffusion models without degrading the generation quality.

Although stable diffusion is capable of generating a wide variety of visually appealing outputs, it performs poorly in tasks that require fine-grained details, such as depicting human extremities (number of hands, joint positions), human poses, facial expressions, face feature positioning, etc. To alleviate these problems in generative transformer models, Gafni et al.(3) propose an object-aware loss that guides the model into focusing on these fine-grained features, which usually tend to be ignored as they occupy a relatively smaller number of pixels. We aim to guide the diffusion model with such a loss to generate realistic human faces and poses.

2 Prior Work

Conditioning diffusion models for specific tasks or improvements is an active field of research. We categorize the related work as follows:

Text to image generation: Text-to-image generation focuses on generating images from standalone text descriptions. DALL-E (14) trained an auto-regressive transformer on text and image tokens, demonstrating convincing zero-shot capabilities on the MS COCO dataset. Inspired by the high-quality unconditional images generation model, GLIDE (11) used diffusion models conditioned on images and employed guided inference with and without a classifier network to generate high-fidelity images. Many further works like LAFITE (23), employed a pre-trained CLIP (20) model to project text and images to the same latent space, training text-to-image models without text data. Imagen directly diffuses pixels using a pyramid structure, while Stable Diffusion basically is a large-scale implementation of latent diffusion (21) used for text-to-image generation.

Customization and Control of Pre-trained Diffusion Models: Recently, text-to-image methods dominate SOTA image diffusion models, allowing text-guided methods (using CLIP features) more control over a diffusion model. DreamBooth (16), Magic3D (9), Midjourney, Wallace et al. (19), Kumari et al. (8), for example, customize and personalize the contents in the generated results using a small set of images (or single image in the case of Wallace et al.) with the same topics for the object of interest.

Guiding conditional diffusion models: Ho et al. (6) introduced classifier-free guidance, which allows joint training of conditional and unconditional diffusion models. An appropriate guidance scale is used to combine the resulting conditional and unconditional score estimates and attain a trade-off between sample quality and diversity. The above approach reduces the computational burden of training a classifier but requires explicit input conditioning in the form of poses, semantic maps, etc. However, additional input conditioning data may not always be available. For our task of generating high-quality images of the human face and pose from input query, a good prior of human pose or depth map is often not available. This necessitates the need for an effective fine-tuning approach that can guide the diffusion process and help in encapsulating the fine-grained human face or physically feasible poses implicitly. Fine-tuning a pre-trained model is typically achieved by training a noise-aware classifier and backpropagating its gradients. However, due to the lack of availability of noisy data models such an approach usually becomes infeasible.

In our work, we attempt to fine-tune latent diffusion by applying the loss directly in the pixel space after decoding the one-step predictions. This approach bypasses the need to retrain a noisy classifier or segmentation network from scratch.

Adding constraints to generated images: ControlNet (21), introduced a novel architecture where a parallel branch is added to a frozen stable diffusion (15) backbone, which takes as input features or conditions and connects to the backbone through zero convolution layers, guiding the diffusion process while also retaining diversity. ControlNet performs an incredible task of supporting additional input conditions such as pose, depth map etc to control pre-trained large diffusion models. However, as mentioned before we don't have explicit conditioning for our task. Moreover, high-level text embedding and conditioning are successful in generating images with accurate scene representations, but these images often don't capture finer details in human faces. We take inspiration from Make-a-scene (3) where the authors enabled a simple control mechanism by employing domain-specific knowledge over key image regions like face and salient objects by proposing face-aware vector quantization, and face emphasis in the scene space. They backpropagated gradients from a feature-matching loss between ground truth and generated images.

We built upon these two architectures and train the ControlNet branch for a mask-aware loss focusing on poses and faces of human beings in order to better text-to-image generation.

3 Approach

Our broad aim is to control image generation by guiding the diffusion process to minimize a perceptual mask-aware loss function such that it can generate kinematically feasible human poses and realistic

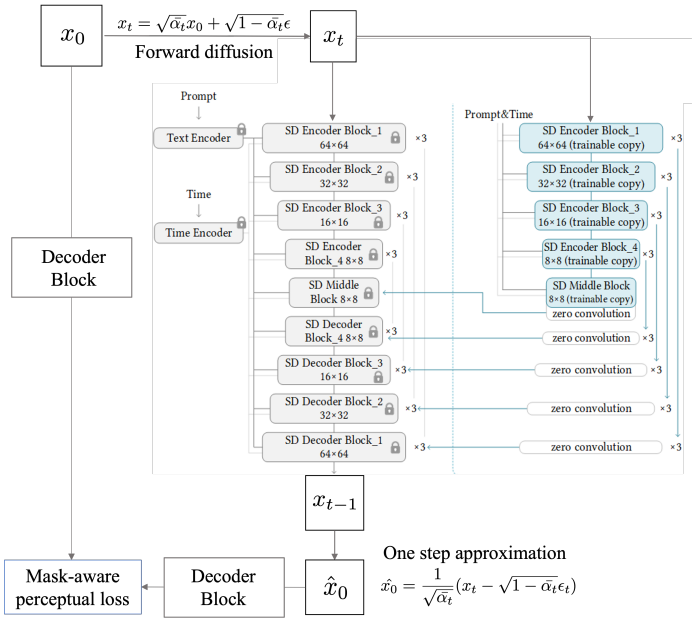

Figure 1: Our Training Pipeline: We freeze the stable diffusion weights and train the parallel control network to minimize the perceptual loss between generated and ground truth image pairs.

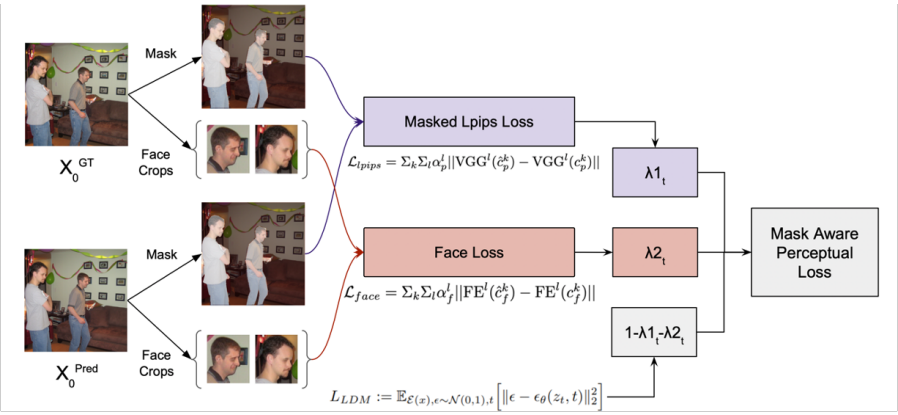

Figure 2: Mask Aware Perceptual Loss Computation

facial expressions. We propose using the ControlNet (21) architecture, and instead of using the parallel network to feed in conditioning, we plan to train it as a second modality that guides the diffusion model into minimizing perceptual loss. This loss function is composed of two parts, inspired by Make-a-scene (3), as shown in figure 2.

1. **Face Aware Loss:** A feature matching loss over the activations of a pre-trained face embedding network (VGGFace 2 (1))

$$\mathcal{L}_{face} = \Sigma_k\Sigma_l\alpha_f^l||FE^l(\hat{c}_f^k) - FE^l(c_f^k)|| \tag{1}$$

where l denotes layers in the face embedding network FE, $\hat{c_f}^k$ and c_f^k are the reconstructed and ground truth face crops of k faces in the image, and α_f^l is a per layer normalization hyperparameter. This face loss helps the model learn low-level features and generate realistic human faces.

2. **Masked LPIPS Loss:** To enforce visual and pose consistency on human images, we additionally use a Learned Perceptual Image Patch Similarity (LPIPS) loss (22) on segmented human masks in the image. LPIPS metric essentially performs a weighted average over

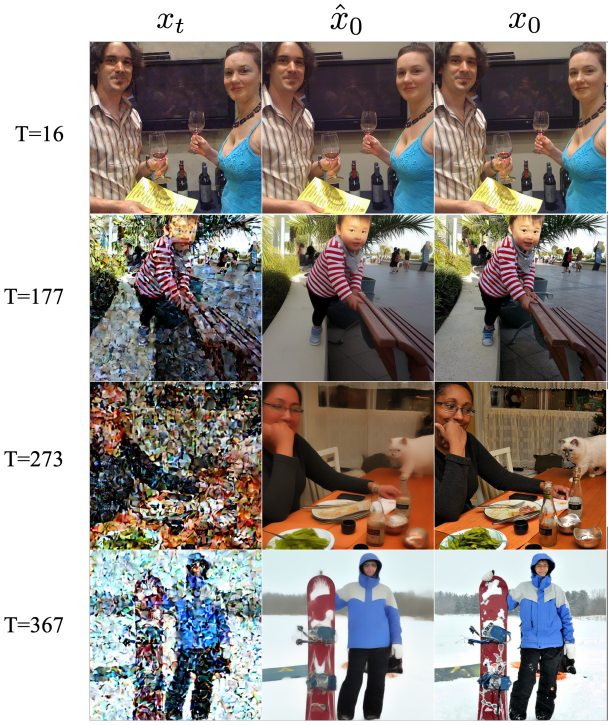

Figure 3: Comparison of one step predicted images with Ground Truth

multiple feature maps obtained from the VGG network. In order to implement masked LPIPS loss, we first upsample feature maps of relevant layers to the original input image size and apply segmented human masks on them. Finally, we compute the LPIPS loss on the masked VGG feature maps. Equation 2 shows LPIPS loss calculation.

$$\mathcal{L}_{lpips} = \Sigma_k \Sigma_l \alpha_p^l ||VGG^l(\hat{c}_p^k) - VGG^l(c_p^k)|| \tag{2}$$

where l denotes layers in the VGG, $\hat{c_p}^k$ and c_p^k are the reconstructed and ground truth image mask crops of k people in the image, and α_p^l is a per layer normalization hyperparameter.

A weighted average of the two loss functions is used to guide the diffusion process and minimize this loss function as represented by 3.

$$\mathcal{L}_{total} = \lambda_1 \mathcal{L}_{lpips} + \lambda_2 \mathcal{L}_{face} + (1 - \lambda_1 - \lambda_2)\mathcal{L}_{LDM} \tag{3}$$

We give a weight λ_1 to this mask perceptual loss, λ_2 to the face loss, and 1-λ_1-λ_2 to the original LDM loss. The model outputs are noisy at any randomly sampled timestep t; hence, the standard implementation of perceptual loss won't work. One approach could be to use DDIM (18) sampling to reproduce the model output at timestep 0 starting from timestep t. However, this would require running DDIM sampling for k steps and storing the computational graph for them amounts to about 1 TB of GPU memory! To circumvent this issue, we instead use a one-step approximation of the model output from a given timestep t, which is given by the following equation 4:

$$\hat{x_0} = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1 - \bar{\alpha}_t}\epsilon_t) \tag{4}$$

As the diffusion model is trained at a randomly sampled timestep, we use the above $\hat{x_0}$, which allows the use of pre-trained networks directly at any sampled timestep by using an efficient invertible diffusion process. To prove that the one-step approximated image is close to the ground truth image, we sample noisy images from time steps and get their one-step prediction. As we can see in 3, for $T < 200$, the one-step approximated image is quite good.

Since our aim is to improve the generation quality of humans, we apply masks on just the humans in the scene and compare the VGG features of the masked ground truth image and the masked one-step

Model	FID	Epochs
Baseline (ControlNet)	63.95	12
Stable Diffusion	62.32	-
Mask LDM ($T = 400$)	63.34	12
Mask LDM ($T = 100$)	61.72	12

Table 1: FID score comparison. $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$ was used for Mask LDMs

predicted image. Similarly, for implementing face loss, we generate bounding boxes for faces in the image and apply perceptual loss on their VGG-Face2 (1) features. Moreover, we directly use the same ground truth mask or bounding box of the image on the $\hat{x_0}$ image since these are quite similar to each other, as shown in 3. This also helps us circumvent the problem of retraining a classifier on noisy segmentation images.

Note that we consider perceptual loss only for time steps less than T and λ_1 and λ_2 are set to 0 for $t > T$. This is illustrated in Fig. 2. As seen in the first rows of 4, 6, 7, 8, the outputs of our model are visually better than our baseline and stable diffusion.

4 Experiments

Dataset: MS COCO dataset subset with images that have a significant amount covered by humans. We have a total of 37470 train images and 1565 validation images.

Metrics: The Frechet Inception Distance (FID) is used to evaluate the quality of generated images with respect to a ground truth dataset. We calculated these scores after every 2 epochs on 1000 validation prompts. It was observed that the FID scores were almost constant throughout training, as seen in Table 1, which may be because the quality of human faces and poses are not captured explicitly. Moreover, we calculated FID on 500 images due to compute limitations; however, recommended settings are at least 10,000 images. The CLIP score is also not extremely relevant for our use case as it measures the similarity between the generated image and text caption, which is not something we wish to improve. We, therefore, believe that human inspection is the best metric for our use case.

Baseline Setup: For our baseline, we re-train the ControlNet branch initialized by pre-trained stable diffusion checkpoint 1.5. Our baseline helps us validate whether simply adding a control network without any loss modification is sufficient for getting better image quality. We train the model for 10 epochs, with an output size of 512x512 and a batch size of 2. Each training epoch takes around 6 hours on an RTX 3090 GPU. As seen in row two of Fig. 4, the output quality has not improved. This motivated the addition of our mask-aware perceptual loss.

Masked LDM: We train our model with the mask-aware perceptual loss as described in Fig. 2. The model was trained for 12 epochs on an RTX 3090 with an output size of 512x512 and a batch size of 1. There are three hyperparameters that were tuned to get desirable results, the loss timestep threshold T, pose loss scale λ_1, and mask loss scale λ_2.

5 Results

Best results were obtained with $\lambda_1 = 0.2$, $\lambda_2 = 0.5$, and T=100 and can be seen in the first rows of Fig. 4, 6, 7, 8. It can be clearly seen that our method outperforms both our baseline and stable diffusion model outputs. It is necessary to highlight here that no prompt engineering was done to obtain the results. A few failure cases of our model are shown in Fig. 9, but here again, it is observed that our model performance is slightly better than the baseline and stable diffusion.

Ablation Studies: We observed that when trained with $T = 400$, the human faces and poses improved, but the generated outputs of the model had extremely smooth features. This can be attributed to the smoothening effect in the one-step predictions at high timesteps. Hence a timestep $T = 100$ was chosen to maintain low-level features in generated output images. As the face pixels cover a small portion of the image, higher values of λ_2 resulted in better face outputs. Tuning λ_1 is again crucial for our

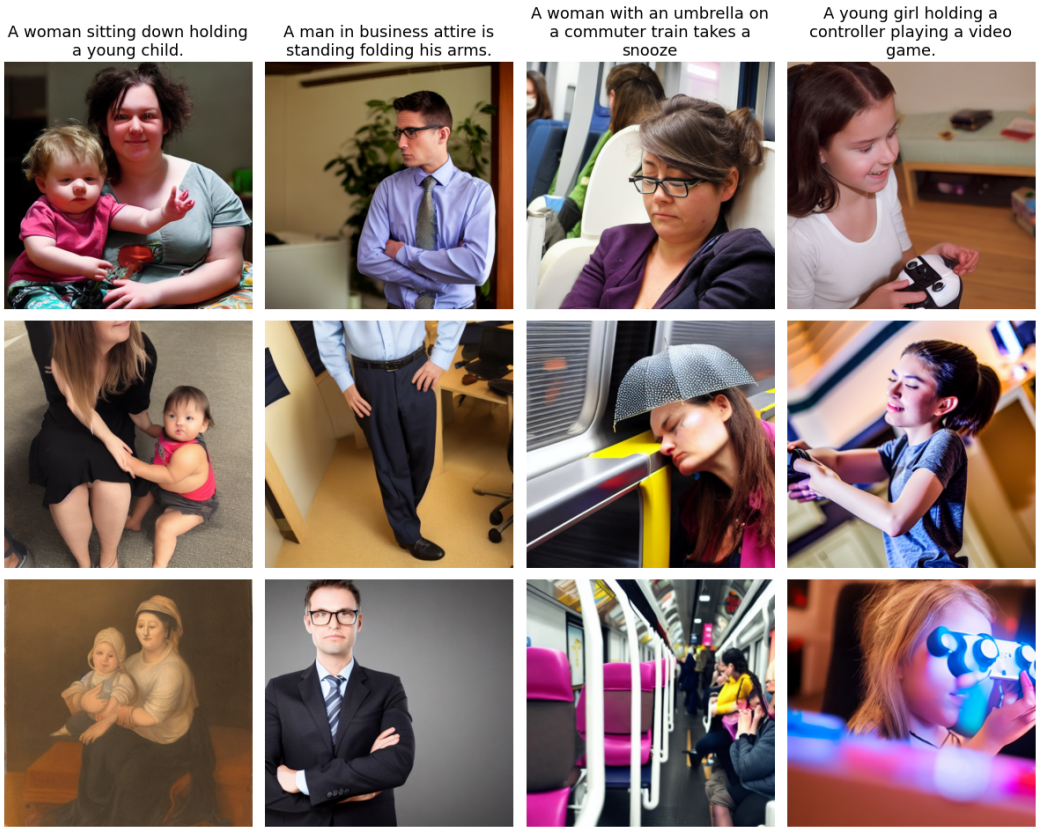

Figure 4: Output comparison, from top row: Masked LDM (ours), Baseline, Stable Diffusion

setup, as it controls how much the model emphasizes the masked humans. It was observed that λ_1 and λ_2 greatly determine output quality, while T controls the smoothness in the output image. Fig 5 shows the sampled model outputs after 12 epochs of training on custom prompts using the same random noise vector. It can be seen by human evaluation that the first row with $\lambda_1 = 0.2, \lambda_2 = 0.5, T = 100$ gave the best results, while the second row with $\lambda_1 = 0.2, \lambda_2 = 0.5, T = 400$ gave the second best results.

6 Conclusion

In this work, we aimed at improving the human image quality generation in latent diffusion models by focusing on the most critical aspects of human perception, such as faces and poses, without explicit input conditioning. We introduced a novel mask-aware loss that encourages the model to learn low-level features of human faces and body poses by minimizing perceptual loss. We directly apply these losses to the pixel space of one-step predictions. This approach helps us circumvent the existing challenges associated with the complexity of classifier guidance which involves retraining an entire noisy segmentation model from scratch. We demonstrate the effectiveness of our method qualitatively by comparing it with Stable Diffusion outputs on a large number of examples. The results show much better image quality and text-image alignment for our model. We believe our work will show a plausible direction in finetuning diffusion models for improving image quality or adding custom concepts.

7 Future Work

The masked LPIPS loss (22) focuses on the entire human body mask; however, fine features like fingers and hands get ignored during training since they occupy a smaller image area. This leads

Figure 5: Ablation study results by changing hyperparameters T, λ_1, λ_2. Images were sampled using the same noise vector on custom prompts. The first row shows the outputs of the best-performing model, and the other rows show the effects of changing specific hyperparameters.

to infeasible hand articulations, as shown in the results above. One approach to mitigate this could be to introduce a keypoints pose loss using keypoint detectors like Open Pose(2) to ensure the kinematic consistency of joints. Recently, Cheng Lu et al. proposed DPM-Solver++ (10), which can sample high-quality images within 15-20 steps in comparison to 50-100 steps of DDIM. This will reduce significant computation time during inference and validation. In our implementation, we used one-step prediction to directly apply losses in pixel space in one step and avoid computation graph explosion; however, Wallace et al. (19) proposed DOODL recently, which is a memory-efficient way of backpropagating gradients by inverting the diffusion process. This will reduce our reliance on one-step approximations for computing loss and will prevent instabilities during training due to the piecewise nature of our loss. Lastly, a better evaluation metric can be used for measuring the quality of human images/poses in the generated image. Current metrics such as FID/KID/IS consider the whole image and a good score can still be achieved despite yielding very visually unappealing images of human faces. Such evaluations can be misleading. Furthermore, CLIP score evaluation can also be performed to quantify the text-image alignment and validate that fine-tuning the model doesn't result in reduced diversity in output images.

References

[1] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. Vggface2: A dataset for recognising faces across pose and age, 2018.

[2] Zhe Cao, Gines Hidalgo, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Openpose: Realtime multi-person 2d pose estimation using part affinity fields, 2019.

[3] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XV*, pages 89–106. Springer, 2022.

[4] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022.

[7] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2013.

[8] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2022.

[9] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation, 2023.

[10] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. Dpm-solver++: Fast solver for guided sampling of diffusion probabilistic models, 2022.

[11] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021.

[12] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2017.

[13] Alec Radford, Luke Metz, and Soumith Chintala. Unsupervised representation learning with deep convolutional generative adversarial networks, 2015.

[14] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[16] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, 2023.

[17] Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015.

[18] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022.

[19] Bram Wallace, Akash Gokul, Stefano Ermon, and Nikhil Naik. End-to-end diffusion latent optimization improves classifier guidance, 2023.

[20] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Yingxia Shao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications, 2022.

[21] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. *arXiv preprint arXiv:2302.05543*, 2023.

[22] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, 2018.

[23] Yufan Zhou, Ruiyi Zhang, Changyou Chen, Chunyuan Li, Chris Tensmeyer, Tong Yu, Jiuxiang Gu, Jinhui Xu, and Tong Sun. Lafite: Towards language-free training for text-to-image generation, 2022.

8 Appendix

Figure 6: Output comparison, from top row: Masked LDM (ours), Baseline, Stable Diffusion

Figure 7: Output comparison, from top row: Masked LDM (ours), Baseline, Stable Diffusion

Figure 8: Output comparison, from top row: Masked LDM (ours), Baseline, Stable Diffusion

Figure 9: Failure Cases, from top row: Masked LDM (ours), Baseline, Stable Diffusion