Improving Human Hands in Stable Diffusion

Matthew Hayes
Stanford University
CS236 Project Progress Report
mhayes3@stanford.edu

1 Introduction

Motivation Stable Diffusion (SD) [Rombach et al., 2022] has had immense success in democratizing powerful text-to-image generation models, but can still require careful iteration on the input prompt or other techniques to achieve error-free results. Human hands, in particular, can be notoriously difficult [Podell et al., 2023] [Matthias, 2023] for these models to get right, e.g. producing an unexpected number of fingers. Users frequently rely on negative prompting e.g. 'too many fingers' [StabilityAI, 2022] or 'unnatural hands' [Kumar, 2023], or use ControlNet's [Zhang et al., 2023] depth map [Kumar, 2023] to constrain the hands and fingers in a more specific way. There are also a few examples where users have used textual inversion [Gal et al., 2022] to find a negative embedding or trained a Low-Rank Adapation (LoRA) [Hu et al., 2021], but to our knowledge there has not been any survey comparing these methods or any one widely accepted solution.

Related work Our experiments build upon pretrained SD models based on Rombach et al. [2022]. They are trained over subsets of the large LAION-5B [Schuhmann et al., 2022] dataset of image, caption pairs to minimize the objective $L_{LDM} := \mathbb{E}_{\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right]$, where x is an input image; $\mathcal{E}$ is a pretrained Variation Autoencoder (VAE) encoder; $t \sim Uniform(1, T = 1000)$ is a time step; z_t is a noised version of the embedded image $z_0 = \mathcal{E}(x)$ with the amount of noise increasing with t; y is a condition, for our purposes an image caption, but it might also be e.g. a class label or a masked image for inpainting; τ_θ is an embedding of y, for captions often held fixed from a pretrained model; and ϵ_θ is a trained feed-forward neural network predicting the noise, with an architecture based on Ronneberger et al. [2015]'s U-Net (a deep convolutional NN with skip-connections that progressively downscales and then upscales an image, increasing the number of channels as resolution decreases) with additional blocks for multi-headed cross-attention [Vaswani et al., 2023] with $\tau_\theta(y)$.

Hu et al. [2021] introduced LoRA as a resource-efficient means of fine-tuning large pretrained language models, but they have since gained popularity for finetuning any kind of large pretrained NN, including SD models, and due to their small file size are frequently shared on the web after training to produce particular characters or styles. Gal et al. [2022]'s Textual inversion and Ruiz et al. [2023]'s DreamBooth are popular alternative means to personalize a SD model to produce e.g. a particular person or style, requiring significantly smaller datasets, and even fewer resources than LoRA. Zhang et al. [2023]'s ControlNet is a fourth popular means of manipulating SD's outputs, but rather than optimizing over some personalized dataset, it allows the user to supply additional conditioning to SD as input so that its outputs conform to e.g. a particular human pose, edge map, or depth map. Depth maps, in particular, are commonly reported on the web as means of fixing SD's hands by generating a depth map based on an existing image with hands in the desired pose.

In terms prior work for improving hands specifically, we plan to evaluate popular mitigations on the web such as Nerfgun3 [2023]'s and Euge_ [2023]'s negative text embeddings, and promqueen [2023]'s and _Envy_ [2023]'s LoRAs.

Preprint. Under review.

2 Problem statement

We will evaluate existing approaches to improving SD's synthesized hands and explore new modifications, to find which method or combination of methods brings most improvement with minimal overhead for users. We hope to not only improve user experience for hands specifically, but to gain more general insight into the errors SD makes.

Dataset Images of resolution $\geq 512^2$ were scraped via Google Image Search. With the hope of teaching the model to generate hands in a different poses, compositions, and styles, the largest dataset, for use with LoRA finetuning, is composed of results from a variety of queries, e.g. "hands", "photo of a woman giving thumbs up", "drawing of a man giving a handshake". Up to 20 images along with their alt-text of were scraped for each of 204 queries, duplicates were removed via source URL, and they were then manually filtered to remove those with significant watermarks or low amounts of detail (e.g. clip-art). The result was a set of 618 images of varying resolutions (`Photos+Drawings`), and a subset of 382 images with those whose queries or alt-text mentioned 'drawing' removed (`Photos`).

For use with Textual Inversion and Dreambooth, two sets of 10 images each were collected: a simpler set containing close-ups of a variety of hands showing their palms with fingers spread out against plain backgrounds (`Palms`); and a more challenging set of handshake photos, some with torsos or blurred faces in the background (`Handshakes`).

Expected results The ideal results are: (1) a set of relatively small files (i.e. <1 GB) that can be shared online and loaded alongside popular pretrained SD models such as SD v1.5 and SD v2.1 to produce well-formed human hands by default (i.e. unless prompted for e.g. a six-fingued hand), with little to no additional user input, and without degrading the quality of generations that do not include human hands; (2) a general method that can be applied to similar fine-detail mistakes made by existing SD models, such as human teeth, animal legs etc.; (3) some out-of-the-box generalization of the files in (1) to similar concepts such as human feet and animal hands, without the need to go through the process (2) again.

If the issue is, as some believe [Matthias, 2023], that hands are simply not dominant enough subjects in the training data, we expect finetuning via LoRA to produce the desired results. However, finetuning can lead to catastrophic forgetting [McCloskey and Cohen, 1989] [Kirkpatrick et al., 2017] of information in the pretrained model, degrading the quality of compositions or other subjects. Tuning only a selection portion of the U-Net's weights at specific relative learning rates [Sun et al., 2019] may be necessary to achieve the best of both worlds.

Another idea is that in addition to being relatively minor subjects, they are highly flexible and uniquely detailed. While SD has undoubtedly seen hands in its training set and can produce something resembling a hand when prompted, it likely has not consistently seen similar hands in similar poses and thus associates some mixture of these variants with the term. With the Textual Inversion and Dreambooth methods, we hope to create a single term with a more narrow focus, while a long and detailed prompt might result in a composition, and generally, as shown by Gal et al. [2022] is not as reliable as a trained token. These methods should serve as baselines to produce better hands at least in specific poses.

As reported on the web, we expect Control Net's depth maps to work well for fixing hands, but with some significant user input. After replicating these results, we hope to modify it to produce results of similar quality, but with less user input.

Evaluation For close-up images, both of hands and when testing generalization to related concepts such as feet, we expect qualitative analysis of example generated images to be the most reliable and prioritize it. It's possible existing qualitative metrics, such as Kernel Inception Distance (KID) [Bińkowski et al., 2021] to measure the similarity between scraped and generated images, and CLIP scores [Radford et al., 2021] scores evaluating the similarity between the generated images and prompts such as 'hands' or specific gestures, will be flawed for our purposes in the same way existing SD models are, e.g. rating hands with extra fingers as more 'hand-like' than those with the expected number. However, we plan to evaluate the utility of these metrics on small curated datasets of positive and negative examples, and expect them to be useful for evaluating how well our results generalize to compositions and unrelated prompts. We plan to also evaluate the utility of pre-trained classifiers or object recognizers, and if possible train specialized versions.

3 Technical approach

LoRA is relatively straightfoward to apply to any NN architecture, including ϵ_θ. Given a pretrained weight matrix $W_\theta^{(0)} \in \mathbb{R}^{m \times p}$, instead of directly taking stochastic gradient decent steps $W_\theta^{(n)} := W_\theta^{(n-1)} - \gamma \frac{1}{|\mathcal{B}|} \sum_{x \in \mathcal{B}} \nabla_{W_\theta^{(n-1)}} L(x; \theta)$ for some minibatch $\mathcal{B}$, learning rate γ, and loss function L, we instead hold $W_\theta^{(0)}$ fixed in an augmented network θ' where $W_{\theta'}^{(n)} := W_\theta^{(0)} + \frac{\alpha}{r} A_{\theta'}^{(n)} B_{\theta'}^{(n)}$ and take steps on $A_{\theta'} \in \mathbb{R}^{m \times r}$ and $B_{\theta'} \in \mathbb{R}^{r \times p}$ where α and $r \ll \min(m, p)$ are hyperparameters. If we set $r = \alpha = \min(m, p)$ we recover traditional finetuning, but with smaller values we can often get similar results with much less compute and share significantly smaller files with those who already have θ. We utilize the Kohya [2023] implementation.

Textual inversion, on the other hand, does not augment weights SD's NNs ϵ_θ or τ_θ at all, and instead exploits its text-conditioning input y. Text embedding networks such as τ_θ first segment a sequence of characters into a sequence of tokens from a (typically) fixed vocabulary. In the first layer of τ_θ, each discrete token is mapped to a vector in $\mathbb{R}^d$ learned through back-propagation. It is these continous vectors that are then passed through additional NN layers (typically transformers) to produce an output embedding (used for cross attention in SD's U-Net). Textual inversion adds an additional (user-specified) token $S^\star$, into the vocabulary, and, keeping the vectors for other tokens fixed, learns a vector $v^\star$ for $S^\star$ by optimizing $\arg\min_{v \in \mathbb{R}^d} \mathbb{E}_{x \sim \text{Uniform}(\mathcal{D}), y \sim \text{Uniform}(\mathcal{C}(s^\star))} L_{LDM}(x, y, v; \theta)$ via back-propagating gradient descent, where $\mathcal{D}$ is a small user-specified dataset of images (`Palms` or `Handshakes` for our purposes), $\mathcal{C}(S^\star)$ is a predefined set of simple captions incorporating $S^\star$, such as 'a photo of $S^\star$', and the first layer of τ_θ maps $S^\star$ to v. We utilize the von Platen et al. [2022] implementation.

4 Preliminary results

For all our initial experiments, we use the SD 1.5 model, though we would also like to run using SD 2.1 and even SDXL 1.0 [Podell et al., 2023] if possible to see if the results generalize across the different pretrained models. Though we experimented with other values, included samples were produced using 50 denoising steps and a classifier-free-guidance (CFG) scale of 7.5, which are common defaults and we found to work as well as higher values.

Textual inversion Using the initializations 'hand' and 'handshake' we train Textual Inversion on both `Palms` and `Handshakes`, for 5000 steps at a constant AdamW [Loshchilov and Hutter, 2019] learning rate of 1e-4. Figure 1 includes training examples, baseline samples, and samples generated after training. It's notable that the baseline errors range from subtle (top-center of palms is slightly disfigured and has some bleeding with the background, top left of handshakes appears to have a second hand shadowing the one on the outside) to severe. While the test samples show that some reasonable associations are learned for the new tokens, we don't find them to be more consistent than the baseline prompting, and in the case of `Handshakes`, the small dataset may have actually moved the embedding in a slightly unfavorable direction, e.g. bottom right. We do, however, see accurate hands produced at a slightly higher frequency than the baseline, e.g. bottom right for both datasets, though errors can be just as severe.

LoRA We perform LoRA finetuning on two candidate sets of modules: (1) the U-Net transformer and attention modules and text encoder attention and perceptron modules using `Photos+Drawings`; and (2) the U-Net convolutional upsampling and downsampling layers and residual connections in addition to (1), using `Photos`. We evaluate multiple AdamW learning rates on (1), and examine different numbers of epochs and values for r and α on both. Figure 2 contains samples for experiments on (1) and the in ital result of (2). For (2) we also experiment with an alternative noise scheduler based on Euler's method instead of the default DIMM [Song et al., 2022] and run for a very large number of epochs as we see training loss continuing to reduce. Figure 3 contains samples from the experiments for (2). Across all settings, we do not see consistent quality improvements after finetuning, and hypothesize that similar to the pretraining dataset, our training sets are too diverse to improve even an individual hand pose. After 319 epochs for our largest LoRA, we do observe some reduction in hands appearing deformed, but at the price of overall image quality. At this extreme we

Figure 1: Textual inversion on `Palms` (left) and `Handshakes` (right). Top: baseline samples, prompts: “A photo of the palm of a hand with five fingers spread out", “A photo of a handshake". Middle: example images used for training textual inversion. Bottom: samples after training textual inversion, prompts: “a photo of a <handpalm>", “a photo of a <handshake>".

suspect there may be some overfitting to our training set, and there may be some intermediate number of epochs where we see improvements for at least the poses and captions in `Photos`.

References

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023.

Meg Matthias. Why does AI art screw up hands and fingers?, 2023. URL `https://www.britannica.com/topic/Why-does-AI-art-screw-up-hands-and-fingers-2230501`.

StabilityAI. Stable Diffusion v2.1 and DreamStudio Updates 7-Dec 22, 2022. URL `https://stability.ai/blog/stablediffusion2-1-release7-dec-2022`.

Sujeet Kumar. Ways to Fix Hand in Stable Diffusion, 2023. URL `https://www.pixcores.com/2023/05/fix-hand-in-stable-diffusion`.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H. Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion, 2022.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

Figure 2: LoRA results on `Photos+Drawings` We demonstrate the effect of reducing learning rate and adding the convolutional blocks to the finetuning. For the non-convolutional blocks r and α are held fixed at 32 and 1 respectively. For each configuration we include a sample for the prompt "Young Man Gives Thumbs-Up" (which occrued in the training set) and "handshake".

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.

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.

Nerfgun3. negative_hand negative embedding, 2023. URL `https://civitai.com/models/56519/negativehand-negative-embedding`.

Euge_. badhandv4 - animeillustdiffusion, 2023. URL `https://civitai.com/models/16993/badhandv4-animeillustdiffusion`.

promqueen. k Hand Mix 101, 2023. URL `https://civitai.com/models/47285/k-hand-mix-101`.

Envy. EnvyBetterHands locon, 2023. URL `https://civitai.com/models/47085/envybetterhands-locon`.

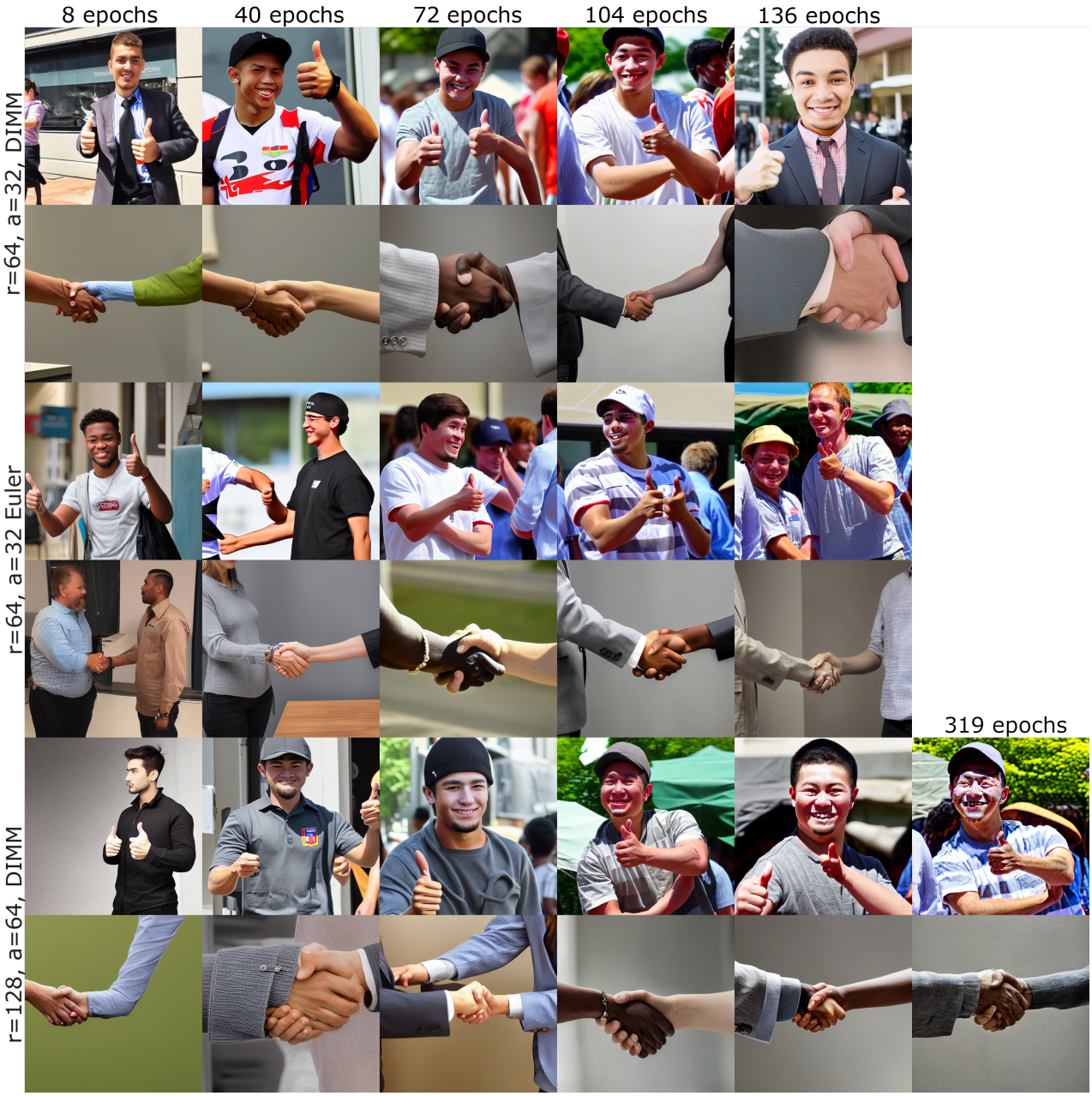

Figure 3: LoRA results on `Photos` including convolutional modules. We use a constant AdamW learning rate of 5e-4 and demonstrate the effect of varying the number epochs and r and α for the non-convolutional blocks. For the convolutional blocks r and α are held fixed at 8 and 4 respectively. For each configuration we include a sample for the prompt “Young Man Gives Thumbs-Up" (which occrued in the training set) and “handshake”.

Michael McCloskey and Neal J. Cohen. Catastrophic interference in connectionist networks: The sequential learning problem. volume 24 of *Psychology of Learning and Motivation*, pages 109–165. Academic Press, 1989. doi: https://doi.org/10.1016/S0079-7421(08)60536-8. URL https://www.sciencedirect.com/science/article/pii/S0079742108605368.

James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, March 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL http://dx.doi.org/10.1073/pnas.1611835114.

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. How to fine-tune bert for text classification?, 2019. URL https://arxiv.org/abs/1905.05583.

Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying mmd gans, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021.

Kohya, 2023. URL https://github.com/Linaqruf/kohya-trainer/.

Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. https://github.com/huggingface/diffusers, 2022.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization, 2019.

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