Improving Human Hands in Stable Diffusion

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

1 Introduction

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. Figure 1 showcases some examples of the errors made by SD 1.5.

Figure 1: Example errors made by SD on close-ups. Left: “a photo of a hand"; middle: “A photo of a handshake"; right: “a photo of a fist". These mistakes range from subtle (bottom left of hands has a smudged middle finger; bottom left of handshakes appears to have an extra finger underneath the outer hand’s index finger; top left of fists merges thumb and index finger) to severe.

As mitigations, users frequently rely on negative prompting e.g. ‘too many fingers’ [StabilityAI, 2022] or ‘unnatural hands’ [Kumar, 2023], or use a ControlNet [Zhang et al., 2023] depth map [Kumar, 2023] model to constrain the hands and fingers to natural poses. 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 there is no one widely accepted solution.

Some theorize that hands are simply not primary enough subjects in the training data, while another idea is that they are are highly flexible and detailed, thus requiring more modeling capacity than typical subjects to encode something other than an averaged representation. There is, however, no consensus to the cause, and to our knowledge no survey comparing potential causes and solutions.

In this work, we aim to close that gap. We evaluate some of these 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. The ideal results are thus tested principals that can guide dataset and implementation selection for future SD models, and in the meantime a set of relatively small files (i.e. <1 GB) that can be shared online and loaded alongside

Preprint. Under review.

popular pretrained SD models such as SD v1.5 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.

In some sense, human hands epitomize many kinds of errors made my SD: we also take human and animal limbs, teeth, etc. for granted in real photography and art, while SD struggles. We therefore hope to see some out-of-the-box generalization of any successes on hands to the most similar errors, such as feet and animal hands, as well as to develop a general method that might be applied to the less similar mistakes.

2 Related work

Our experiments investigate pretrained SD models based on Rombach et al. [2022], which moved Ho et al. [2020]'s Denoising Diffusion models into the latent space of a locality-aware pretrained Variation Autoencoder (VAE), and extended them to various kinds of conditional generation. Specifically, we focus primarily on the caption-conditioned version of SD, trained over subsets of the large LAION-5B [Schuhmann et al., 2022] dataset of image, caption pairs, $(x, y) \in \mathcal{D}$, to minimize the squared error between predicted and actual noise: $L_{LDM}(\mathcal{D}; \theta) := \mathbb{E}_{(x,y)\sim\mathcal{U}(\mathcal{D}),\epsilon\sim\mathcal{N}(\mathbf{0},I),t\sim\mathcal{U}(1,T=1000)}\left[\|\epsilon - \epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)\|_2^2\right]$. z_t is a noised version of the image x encoded by the VAE encoder $\mathcal{E}$ such that $z_0 = \mathcal{E}(x)$ and the amount of noise increases with time step t. τ_θ encodes the caption y and is often fixed from a pre-trained model (e.g. Radford et al. [2021]'s CLIP), but more generally may be jointly learned to encode any condition e.g. a class label or a masked image for inpainting. ϵ_θ is a trained feed-forward neural network (NN) 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 $\tau_\theta(y)$.

During sampling, SD models start from the complete noise at step T and rather than taking 1000 denoising steps, in practice, take a user-defined number (often between 20 and 60) of larger extrapolated steps, whose implementation varies based on the specific sampler implementation configured. Furthermore, these steps not only incorporate the predicted conditional noise $\epsilon_\theta\left(z_t, t, \tau_\theta(y)\right)$, but use the notion of Classifier Free Guidance (CFG) [Sanchez et al., 2023] to step in the direction of a weighted difference between the conditional and unconditional predicted noise: $(1 + w)\epsilon_\theta\left(z_t, t, \tau_\theta(y)\right) - w\epsilon_\theta\left(z_t, t, \tau_\theta(\emptyset)\right)$, where w is called the CFG scale and higher values put higher emphasis on the user's prompt while lower values allow the model to be more 'creative'; a common default value is $w \approx 7.5$. Negative prompting exploits this mechanism, instead of stepping away from the prediction conditioned on the negative prompt $\overline{y}$ instead of unconditional prediction: $(1 + w)\epsilon_\theta\left(z_t, t, \tau_\theta(y)\right) - w\epsilon_\theta\left(z_t, t, \tau_\theta(\overline{y})\right)$.

When prompting, negative prompting, and adjusting the CFG scale are not enough, there are a number of methods for further personalizing or conditioning generation. Hu et al. [2021] introduced LoRA as a resource-efficient means of fine-tuning large pretrained language models (LMs) with low-rank updates, but they have since gained popularity for finetuning any kind of large pretrained NN, including SD's ϵ_θ or τ_θ, and, once trained, can easily be shared on the web thanks to their small file size. As demonstrated for LMs, it is often sufficient to fine-tune only a model's attention weights using LoRA, though at the risk of overfitting and catastrophic forgetting [McCloskey and Cohen, 1989] [Kirkpatrick et al., 2017] it is more flexible to also tune SD's U-Net Convolution weights.

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 subject or style, requiring significantly fewer examples (i.e. < 10) than LoRA finetuning. Textual inversion makes a small modification to τ_θ alone, while DreamBooth, often combined with LoRA updates, is closer to vanilla finetuning but with an additional regularizing loss term based on images generated for related subjects by the pretrained model. While these methods are touted for their ability to inject a personal subject into SD's vocabulary, we'll use them in an attempt to narrow it's associations when it comes to human hands.

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.

Figure 2: Example generations of prior works on the same random seed. In some cases they help, but some also significantly change style or ruin the good generation in the bottom right. Prompts: 'a photo of the palm of a hand', 'a photo of a handshake'.

Besides negative prompting, for which we particularly tested the term 'disfigured', customizations on the web to mitigate SD's hands include LoRAs and negative textual inversions. Specifically, we investigate Nerfgun3 [2023]'s 'negative_hand' and 'bad_prompt' negative embeddings, and promqueen [2023]'s 'kHandMix101', Orrreo [2023]'s 'Quickhands', and _Envy_ [2023]'s 'EnvyBetterHands' LoRAs. Some comparisons of their generations to vanilla SD are included in Figure 2. Although these prior works to improve hands exist, there is very little information available on their training data and parameters: they are shared with only the trained weights, example generations that are likely cherry-picked, and usage instructions. In the case of the LoRAs, only by delving into the shared files themselves are we able to discover that all three trained ϵ_θ's attention and covolutional layers as well as τ_θ, respectively using update matrices of rank 32, 128, and a mix of rank 200 in linear and some convolutions and rank 1 in other convolutions. Further, their effectiveness of these solutions varies and they can impact style or ruin otherwise good generations. These are the areas we hope to improve upon in this work.

(a) LoRA finetuning training examples from `Photos(&Drawings)` (top) and `11K-Hands` (bottom) with captions (and retrieval queries). The top right image is included in `Photos&Drawings` but not `Photos`.

(b) Textual inversion and DreamBooth training examples from `Palms` (top) and `Handshakes` (bottom).

Figure 3: Example training images. `Photos&Drawings` is more diverse than the others, but has varying caption quality.

3 Methods

3.1 Datasets

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 the images 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`). The captions are lightly processed to remove some author and source names thought to be uninformative, but are generallly are of varying quality. Some examples are shown in Figure 3a.

For use with Textual Inversion and Dreambooth, two more focused sets of 10 images each without captions 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`). These are demonstrated in Figure 3b.

Also more focused but much larger than `Photos%Drawings`, we utilize Afifi [2019]’s close-ups of of hand palms and the backs of hands (`11K-Hands`). Some examples are included in Figure 3a. It comes with metadata including the person’s age, gender, and skin color, whether the image is of the front or back of the left or right hand, and whether or not it includes nail polish, accessories such as rings, and any ‘irregularities’. This was used to filter images with ‘irregularities’, bringing the dataset size to 10.9K, and generate captions with the grammar: ‘front’|‘back’ ‘of’ ‘fair’|‘medium’|‘dark’ ‘-skinned’ [‘elderly’] ‘male’|‘female’ ‘left’|‘right’ ‘hand’ [‘with’ ‘nail polish’|‘accessories’|‘nail polish and accessories’].

Finally, during manual evaluation we collect some of the particularly poor generations to train a negative embedding using textual inversion similar to some of the prior work. This set includes 260 images, but is not particularly diverse, composed primarily of `11K-Hands`-like generations and samples for the prompts ‘a photo of a handshake’ and ‘a man holding up one hand’.

3.2 Algorithms

We demonstrate the algorithms applied to `Photos` and `Palms`, but application to the others is a simply a substitution of dataset and/or caption.

LoRA is relatively straightforward to apply to any NN architecture, including ϵ_θ and τ_θ. Given a pretrained weight matrix $W_\theta^{(0)} \in \mathbb{R}^{m \times p}$, instead of directly taking stochastic gradient decent steps $W_\theta^{(n)} := W_\theta^{(n-1)} - \gamma \frac{1}{|\mathcal{B}|} \sum_{(x,y) \in \mathcal{B}} \nabla_{W_\theta^{(n-1)}} L_{LDM}(x; \theta)$ for some minibatch $\mathcal{B} \subseteq \mathcal{D}_{\texttt{Photos}}$ and learning rate γ, 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 allow us to share significantly smaller files with those who already have θ. We experiment with both Hugging Face's Mangrulkar et al. [2022] implementation, which only supports tuning ϵ_θ's attention weights, and the Kohya [2023] implementation which also permits tuning ϵ_θ's convolutions and τ_θ. We also utilize Hugging Face's DreamBooth implementation, which incoporates LoRA on τ_θ and ϵ_θ's attention weights for training efficiency, but with a modified loss function: $L_{\text{DreamBooth}} := L_{LDM}(\mathcal{D}_{\texttt{Palms}} \times \{\text{“a photo of [V] hand"}\}) + \lambda L_{LDM}(\mathcal{D}_{\text{generated “a photo of hand”}} \times \{\text{“a photo of hand"}\})$, where the hyperparameter λ adjusts the strength of the "prior preservation loss", with a recommended value $\lambda = 1$.

Textual inversion, on the other hand, does not augment ϵ_θ at all and instead exploits τ_θ's initial embedding of y. Such networks 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 $\arg\min_{v \in \mathbb{R}^d} L_{LDM}(\mathcal{D}_{\texttt{Palms}} \times \mathcal{C}(s^\star); v, \theta)$ via back-propagating gradient descent, where $\mathcal{C}(S^\star)$ is a predefined set of simple captions incorporating $S^\star$, such as 'a photo of a $S^\star$' or 'a rendering of a $S^\star$', and the first layer of τ_θ maps $S^\star$ to v. We utilize Hugging Face's von Platen et al. [2022] implementation.

4 Experiment details

For all our experiments we use the SD 1.5 model, due to its popularity on the web and use in the prior baseline works. Though we experimented with other values, included samples were produced using 30 DDIM [Song et al., 2022] steps and a CFG scale of 7.5, which are common defaults and we found to work as well as higher values.

We train Textual Inversion using the initializations 'hand', 'handshake', and 'hand' respectively on `Palms`, `Handshakes`, and our set of poor generations, for 5000 steps at a constant AdamW [Loshchilov and Hutter, 2019] learning rate of 1e-4. For DreamBooth, we use the recommended default UNet attention LoRA $r = 16, \alpha = 27$, text encoder $r = 16, \alpha = 17$, with 200 generated regularization images, a constant AdamW learning rate 1e-4 with batch size 1 for 800 steps, and try varying the prior preservation loss weight below the default $\lambda = 1$. Following the original work, we use 'sks' as the identifier '[V]'.

For LoRA finetuning, we examine a widge range of hyperparameter settings, including which NN layers to tune, learning rate and schedule, number of epochs, and sampler. On `Photos&Drawings` we hold UNet attention and text encoder both at $r = 32$, $\alpha = 1$, use a batch size of 6, explore including UNet's convolutions at $r = 8$, $\alpha = 1$, and vary AdamW learning rate 1e-4 and 1e-5 at a constant schedule for up to 15 epochs. Using `Photos` again with a batch size 6, we vary text encoder and UNet attention rank between higher values of $r \in [64, 128]$ with $\alpha = r/2$, while keeping the convolutional $r = 8, \alpha = 4$ and training for more epochs, up to 319 for our largest model, at a constant AdamW rate of 5e-4, and try an alternate noise scheduler based on Euler's method instead of DDIM. Finally, on `11K-Hands`, we try tuning only UNet's attention weights, at ranks of 4 or 64, $\alpha = r$, batch size 9, AdamW learning rate 1e-4 on a cosine schedule over 30,000 steps, or roughly 25 epochs.

Figure 4: 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". Bottom: samples after training textual inversion, prompts: “a photo of a <handpalm>", “a photo of a <handshake>".

4.1 Evaluation

For close-up images, both of hands and when testing generalization to related concepts such as feet, we expected qualitative analysis of example generated images to be the most reliable. 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, are 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. For LoRAs trained on `11K-Hands`, we manually rated more than 600 generated images on a three point scale. Generations with an incorrect number of fingers or significant deformities are given a score of zero; those with the correct number of fingers but a number of unlikely artifacts such as thick lines or excessive boniness are given a score of 0.5, and finally only those with five fingers and no artifacts or only minor ones are given a score of one. Figure 8c demonstrates the latter two kinds of samples. Using these manual scores, we are able to perform a quantitative comparison to KID, CLIP scores, and an off-the-shelf hand gesture recognizer score. Furthermore, we fine-tune a Visual Transformer [Dosovitskiy et al., 2021] classifier on the >400 examples which received scores of zero or one, in order to assist with evaluating the utility of reducing the trained LoRA α when sampling and compare it’s scores to these other metrics.

5 Results & Analysis

5.1 Textual inversion

Figure 4 includes baseline samples and samples generated after training on `Palms` and `Handshakes`. 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. It is possible that the initializations are too strong, and training only moves them along a subset of dimensions. It may be worth investigating if we get any better results by initializing with a token with a weak prior such as those used in DreamBooth, or by mixing these strong prior embeddings with a weaker one or the zero vector.

Figure 2 compares our trained negative embedding to the baselines from the web. It seems to have moderate success when it comes to handshakes but less for palms, also, e.g. changing the good generation in the bottom right, and may not be as highly optimized as the baselines. It does seem, to some extent, to move the model away from generating poor palms, but moves it towards generating

Figure 5: DreamBooth results on `Palms` and `Handhshakes`, including the result of recontextualization, composition, and reducing λ.

things that are not palms at all rather than good palms, e.g. top right, middle right, and bottom left. We suspect the learned concept of a bad hand is interleaved with the general concept of a hand and it would be interesting if we could somehow use both positive and negative examples to train an embedding that highlights only these flaws.

5.2 DreamBooth

DreamBooth, on the other hand, we find to work moderately well on `Palms`, but less well on `Handshakes`, as demonstrated in Figure 5. Unfortunately, contrary to the original DreamBooth paper, composing and merging the learned concept i.e. to put it on a person, does not seem to work well, though the model has some success placing it in different contexts, i.e. changing the background. We find that despite the prior for hands being relatively poor, reducing λ does lead to worse generations, even in the absence of recontextualization or composition.

5.3 LoRA

Figure 6 contains samples for experiments on `Photos&Drawings`, while Figure 7 contains samples from the longer trainings with larger ranks. 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 without high quality captions 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 suspect there may be some overfitting to our training set, and there may be some intermediate number of epochs or reduced α at sampling time where we would see improvements for at least the poses and captions in `Photos`.

On the large `11K-Hands`, however, we do see some signficant improvement. Both rank 4 and rank 64 achieve a peak human score of $\approx 60\%$ in comparison to a baseline of 6.3%, though rank 64 reaches this peak in fewer training steps. As depicted in Figure 8b, none of the popular off-the-shelf metrics correlate well with human preference and some are even negatively correlated. Arguably CLIP correlates the best, but increases only very slightly with human preference. Though biased since the classifier was trained on some of the scale 1.0 human scores, its predictions do seem to correlate better with the human scores than the off-the-shelf metrics as shown in Figure 8d. Scaling

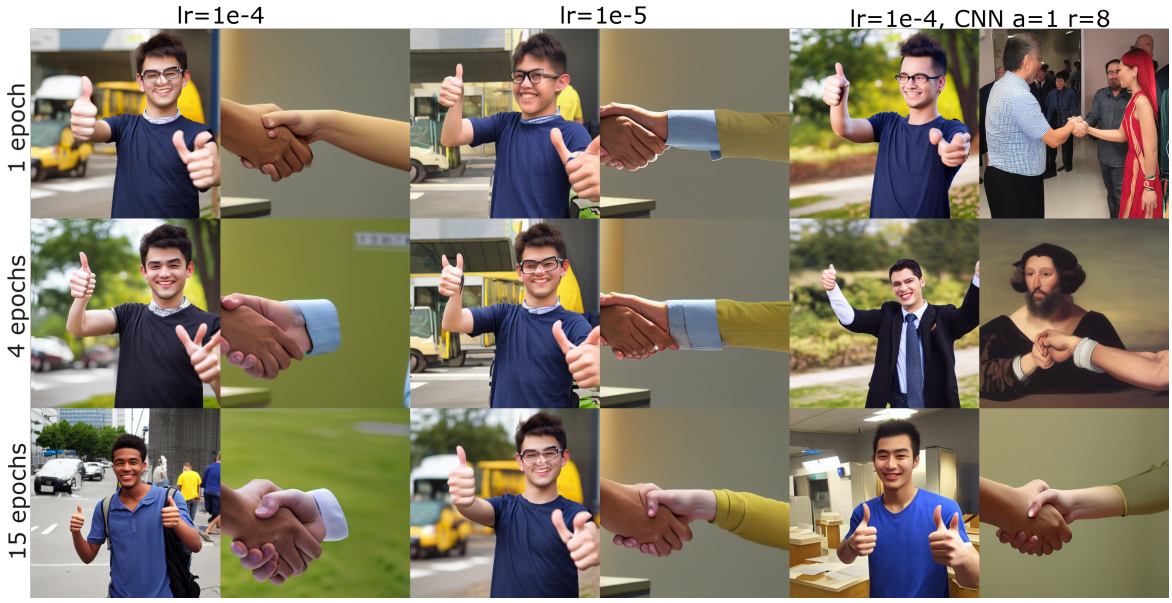


Figure 6: LoRA results on `Photos&Drawings`. For each configuration we include a sample for the prompt “Young Man Gives Thumbs-Up" (which occurred in the training set) and “handshake”.

down the LoRA weights after training does not seem to improve generation quality, except perhaps a little in the case of rank 64 at scale 0.75.

6 Conclusion

Improving human hands in Stable Diffusion is very challenging. Despite examining a wide variety of methods in this work, we are still far from solving the issue entirely. Existing mitigations on the web work just as well if not better than our results, likely involved a lot of effort as well, and it would be beneficial to the community if their methods were documented. Our results on negative textual inversion and DreamBooth demonstrate that it is certainly possible to nudge the model in the right direction with relatively small changes. However, the results on LoRA finetuning show that simply throwing moderate amounts of additional data at the problem is an unlikely solution. While `11K-Hands` is a narrow proof of concept that large amounts of high-quality data brings improvements, the $\approx 60\%$ success rate is still far from ideal.

There are numerous directions for future work. First, we might see how far we can bring improvement in narrow domains like `11K-Hands` with additional modelling power: our experiments on this dataset did not include tuning τ_θ, ϵ_θ's convolutions, or higher rank updates than $r = 64$. They also suggest that with higher quality captions and more data than `Photos&Drawings`, we might be able to bring similarly moderate improvements to a wider domain. Second, negative embeddings work surprisingly well for their size, and this suggests that more clever means of giving the model focused examples of what not to generate, in addition to positive examples, may be fruitful. Third, existing quantitative metrics do not work well for this purpose, and manual qualitative analysis is too labor intensive. While we had some success with a bespoke classifier, this approach does not scale well: any improved means of performing automated evaluation would greatly accelerate experimentation, and their implementations might provide at least insights into the kinds of signals these model need, if not objectives that can be directly incorporated into the optimization. Fourth, we experimented exclusively on SD 1.5, and while we suspect many of our findings will generalize to other models, comparison is warranted. Finally, ControlNet is, in the meantime, the most flexible solution, but it requires far too much user input and experimentation. Our initial attempt to automate the MediaPipe hand landmark ControlNet was not successful, but the idea of automating ControlNet through hand landmark detection or alignment and stitching of depth or canny edge maps is promising.

Figure 7: LoRA results on `Photos` including convolutional layers with $r = 8, \alpha = 4$. For each configuration we include a sample for the prompt “Young Man Gives Thumbs-Up" (which occurred in the training set) and “handshake”.

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`.

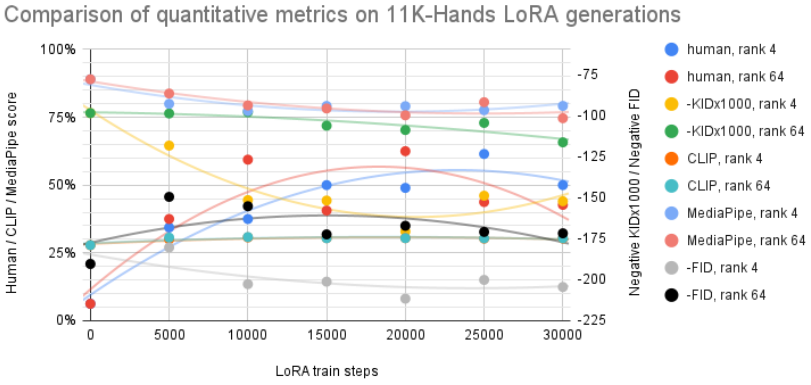
Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models, 2023.

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.

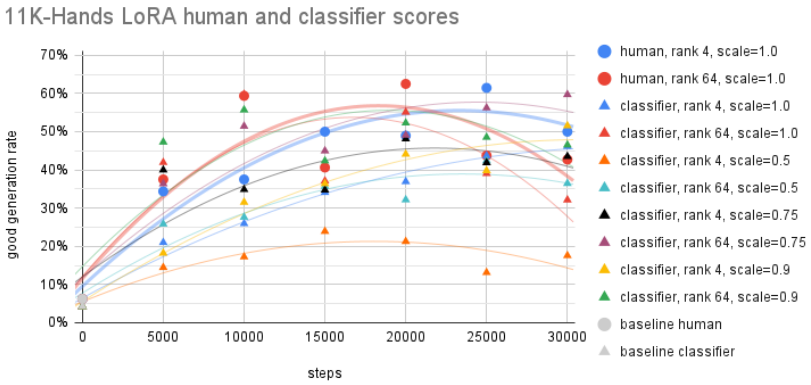
Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

(a) Example high-score (99.8%) detection from MediaPipe on a poor sample.



(b) Comparison of off-the-shelf quantitative metrics and human evaluations, with second-degree polynomial trendlines. FID and KID are negated so that higher values are better like the rest of the metrics, and KID is multiplied by 1000 to put it on a similar scale as FID. Note that CLIP scores on rank 4 are not visible due to its very similar values to those for rank 64.

(c) Example generations after training on `11K-Hands`.



(d) Human and custom classifier scores on samples from LoRA trained on `11K-Hands`, with second degree polynomial trendlines.

Figure 8: Evaluation of LoRAs fine-tuned on `11K-Hands`.

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.

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.

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.

Guillaume Sanchez, Honglu Fan, Alexander Spangher, Elad Levi, Pawan Sasanka Ammanamanchi, and Stella Biderman. Stay on topic with classifier-free guidance, 2023.

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.

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.

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

Orrreo. Quickhand, 2023. URL https://civitai.com/models/44056/quickhand.

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

Mahmoud Afifi. 11k hands: gender recognition and biometric identification using a large dataset of hand images. *Multimedia Tools and Applications*, 2019. doi: 10.1007/s11042-019-7424-8. URL https://doi.org/10.1007/s11042-019-7424-8.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

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.

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

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

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

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.