r/StableDiffusion • 3 yr. ago
terrariyum

Advanced advice for model training / fine-tuning and captioning

Tutorial | Guide

This is advice for those who already understand the basics of training a checkpoint model and want to up their game. I'll be giving very specific pointers and explaining the reason behind them using real examples from my own models (which I'll also shamelessly plug). My experience is specific to checkpoints, but may also be true for LORA.

Summary

- **Training images**
  - Originals should be very large, denoised, then sized down
  - Minimum 10 per concept, but much more is much better
  - Maximize visual diversity and minimize visual repetition (except any object being trained)
- **Style captioning**
  - The MORE description, the better (opposite of objects)
  - Order captions from most to least prominent concept
  - You DON'T need to caption a style keyword (opposite of objects)
  - The specific word choice matters
- **Object captioning**
  - The LESS description, the better (opposite of styles)
  - Order captions from most to least prominent concept (if more than one)
  - You DO need to caption an object keyword (opposite of styles)
  - The specific word choice matters
- **Learning rate**
  - Probably 5e-7 is best, but it's slowwwww

The basic rules of training images

I've seen vast improvements by increasing the number of images and quality in my training set. Specifically, the improvements were: more reliably generating images that match trained concepts, images that more reliably combined concepts, images that are more realistic, diverse, and detailed, and images that didn't look exactly like the trainers (over-fitting). But why is that? This is what you need to know:

1. Any and every large and small visual detail of the training images will appear in the model.
2. Anything visual detail that's repeated in multiple training images will be massively amplified.
3. If base-SD can already generate a style/object that's similar to training concepts, then fewer trainer images will be needed for those concepts.

How many training images to use

So when I use the term "concept" in this post, I mean the word or words in your caption file that match a specific visual element in your trainers. For example, my Emotion-Puppeteer model contains several concepts: one for each different eye and mouth expression. One such concept is "seething eyes". That's the caption I used in each image that contained a face with eyes the look angry with the brows scrunched together in a >:(shape. Several trainers shared that concept even though the faces were different people and the mouths paired with the "seething eyes" were sometimes different (e.g. frowning or sneering).

So how many images do you need? Some of the eye and mouth concepts only needed 10 training images to reliably reproduce the matching visual element in the output. But "seething eyes" took 20 images. Meanwhile, I have 20 trainers with "winking eyes", and that output is still unreliable. In a future model, I'll try again with 40 "winking eye" trainers. I suspect it's harder to train because it's less common in the LAION dataset used to train SD. Also keep in mind, that the more trainers per concept the less over-fitting and the more diversity of the output. Some amateurs are training models with literally thousands of images.

In my Huggingface, I list exactly how many images I used for each concept used to train Emotion Puppeteer so that you can see how those difference cause bias.

How to select trainer images

This may seem obvious - just pick images that match the desired style/object right? Nope! Consider trainer rules #1 and #2. If your trainers are a bit blurry or contain artifacts, those will be amplified in the resulting model. That's why it's import, for every single training image to:

- **Start with images that are no smaller than 1,000****^{2} before resizing.**
- **Level-balance, color-balance, and denoise before resizing.**

Note that the 1000^{2} size is the minimum for a typical 512^{2} model. For a 768^{2} model, the minimum is 1,500^{2} images. If you don't follow the above, your model will be biased towards lacking contrast, having color-bias, having noise, and having low detail. The reason you need to start with higher-res images is that you need to denoise them. Even with high-quality denoising software, some of the fine detail besides the noise will be unavoidably lost. But if you start large, then any detail loss will be hidden when you scale down (e.g. to 512^{2}). Also, if you're using images found only, they will typically be compressed or artificially upscaled. So only the largest images will have enough detail. You can judge the quality difference yourself by starting with two different sized images, denoising both, then scaling both down to a matching 512^{2}.

The reverse of trainer rule #1 is also true: anything that's NOT in the trainers won't appear in the model. That including fine detail. For example, My Emotion-Puppeteer model generates closeups of faces. In an earlier version of the model, all output lacked detail because I didn't start with high-res images. In the latest model I started with hi-res trainers, and even when scaled to 512^{2}, you can see skin pores and fine wrinkles in the trainers. While nothing is guaranteed, these details can show up in the output of the latest model.

If you can't find larger training images, then at least upscale before resizing to the training size. Start with a denoised image, then quadruple its size using upscaling software (e.g. the "extras" tab within Auto1111). Finally, scale it down to to train size. That at least will make all of the edges clean and sharp, remove artifacts, and smooth

- **Avoid visual repetition as much as possible except for the thing you want to reproduce.**

Remember trainer rule #2. Here's an example. For my Emotion-Puppeteer model, I needed to images of the many eye and mouth positions I wanted to train. But it's hard to find high-quality images of some facial expressions. So for one of the mouth positions (aka concepts), I found several photos of the same celebrity making that expression. Out of all the trainers I found for that mouth concept, I ended up with about ~10% that were photos of that celebrity. In my latest model, when that mouth keyword is used in a prompt, the face looks recognizably like that celebrity, I'd guess, about a 3rd of the time. The 10% of that celebrity has been amplified by about 3x.

This amplification effect isn't only limited to the things that you explicitly describe in the captions. Literally anything that's visually similar across images, anywhere in those images will be trained and amplified.

Here another example: The reason for that was that, in an earlier version of Emotion-Puppeteer, I had cropped all of my trainer photos at the neck. So the model struggled to generate output that was zoomed-out and cropped at the waist. To get around that limitation, I tried an experiment. I found one photo that was cropped at the waist, and then I used my model with inpainting to generate new images of various different faces. I then added those new images to my training set and trained a 2nd model.

Those generate images only made up about ~15% of the training set that I used to train the 2nd model, but the background was the same for each, and it happened to be a wall covered in flowers. Note that none of my captions contained "flowers". Nevertheless the result was that most of the images generated by that 2nd model contained flowers! Flowers in the background, random flowers next to random objects, flowers in people's hair, and even flowers in the fabric print on clothing. The ~15% of uncaptioned flowers made the whole model obsessed with flowers!

- **Visually diverse trainers are critical for style and object matters**

This is similar to the advice to avoid visual repetition, but it's worth calling out. For a style model, the more diverse and numerous the objects in the trainers, the more examples of objects in that style the model has to learn from. Therefore, the model is better able to extract the style from those example objects and transfer it to objects that aren't in the trainers. Ideally, your style trainers will have examples from inside, outside, closeup, long-shot, day, night, people, objects, etc.

Meanwhile, for an object model, you want the trainers to show the object being trained as many different angles and lighting conditions as possible. For an object model, the more diverse and numerous the "styles" (e.g. lighting conditions) in the trainers, the more examples of styles of that object the model has to learn from. Therefore, the model is better able to extract the object from those example styles and transfer onto it styles that aren't in the trainers. The ideal object trainer set will show the object from many angles (e.g. 10), repeating all that set of angles in several lighting conditions (e.g. 10x10), and using a different background in every single trainer (e.g. 100 different backgrounds). That prevents the backgrounds from appearing unprompted in the output.

- **Some concepts are hard to train, and some concepts probably can't be trained**

This is trainer rule #3, and mostly you'll discover this through experimentation. Mostly. But if the base SD model struggles with something, you know that'll be harder to train. Hands are the most obvious example. People have tried to train a model that just does hands using hundreds of images. That hasn't been successful because the base SD 1.5 model doesn't understand hands at all. Similarly SD 2.1 doesn't understand anatomy in general, and people haven't been able to train anatomy back in. The base or starting point for the fine-tuning is just too low.

In my own experience with Emotion-Puppeteer, so far I haven't been able to train the concept of a the lip-biting expression. Maybe I could if I had a 100 trainers. The "winking eyes" concept is merely unreliable. But I actually had to remove the lip-biting trainer images entirely from the model and retrain because including that concept resulted in hideously deformed mouths even when caption keyword wasn't used in the prompt. I even tried switching the caption from "lip-biting" mouth to "flirting mouth", but it didn't help.

Here's another example: I tried to train 4 concepts using ~50 images for each: a.) head turned straight towards the camera and eyes looking into the camera, b.) head turned straight towards the camera but eyes looking away from it, c.) head turned to a three-quarter angle but eyes looking into the camera, and d.) head turned away and eyes looking away. While a, b, and d, worked, c failed to train, even with 50 images. So in the latest model, I only used concepts a and d. For the ~100 images of 3/4 head turn, whether eyes looking to camera or not, I captioned them all as "looking away". For the ~50 images of head facing forward but eyes looking away, I didn't caption anything, and for the other ~50, I captioned "looking straight". This resulted in looking into camera and 3/4 head turn both becoming more reliable.

The basic rules of captioning

You've probably heard by now that captions are the best way to train, which is true. But I haven't found any good advice about how to caption, what to caption, what words to use, and why. I already made one post about how to caption a style, based what I learned from my Technicolor-Diffusion model. Since that post, I've learned more. This is what you need to know:

1. The specific words that you use in the captions are the same specific words you'll need to use in the prompts.
2. Describe concepts in training images that you want to reproduce, and don't describe concepts that you don't want to reproduce.
3. Like imagery, words that are repeated will be amplified.
4. Like prompting, words at the start of the caption carry more weight.
5. For each caption word you used, the corresponding visual elements from your trainers will be blended with the visual elements that the SD base model already associates with that word.

How to caption ~style~ models

- **The MORE description the better.**

An ideal style model will reproduce the style no matter what subject you reference in the prompt. The greater the visual diversity or subject matter of the images, the better SD is able to guess what that visual style will look like on subjects that it hasn't seen in that style. Makes sense, right? So why are more word descriptions better? Because it's also the case that the greater the linguistic diversity of the captions, the better SD is able to relate those words to the adjacent words it already knows, and the better it will apply the visual style to those adjacent concepts that aren't in the captions. Therefore, you should describe in detail every part of every object in the image, the positions and orientations of those objects and parts of objects, and whether they're in the foreground or background. Also describe more abstract concepts such as the lighting conditions, emotions, beautiful/ugly, etc.

Consider captioning rule #1. In my earlier post about training Technicolor-Diffusion, I showed an example where using one of the full and exact captions as the prompt reproduced that training image nearly exactly. And I "woman", then you can only reliably change "woman" in the output image. But if you captioned "blonde woman", then you can reliably change "blonde" (e.g. to redhead) while keeping woman. You can't over-describe, as long as you don't describe anything that's NOT in the image.

- **Describe the image in order from most to least prominent concept (usually biggest to smallest part of image).**

Consider captioning rule #4. Let's say that you have an illustration of a man sitting in a chair by a pool. You could - and should - caption a hundred things about that image from the man's clothing and hairstyle, to the pair of sunglasses in his shirt-pocket, down to the tiny glint of sunlight off the water in the pool in the distance. But if you asked an average person what the image contained, they'd say something like "a man sitting in a chair by a pool" because those are both the biggest parts of the image and the most obvious concepts.

Captioning rule #4 says that, just as words at the start of the prompt are most likely to be generated in the image, words at the start of the caption are most likely to be learned from the trainer image. You hope your style model will reproduce that style even in glint of light in the distance. But that detail is hard to learn because it's so small in pixel size and because "glint" as a concept isn't as obvious. Again, you can't over describe so long as you order your captions by concept prominence. Those words and concepts at the end of the caption are just less likely to be learned.

- **You don't need to caption a style keyword - e.g. "in blob style"**

The traditional advice has been to include "blob style" at the front of every caption - where "blob" is any random keyword that will be used in the prompt to invoke the style. But, again, that just means that you're now required to put "blob style" into every prompt in order to maximize the output of that style. Meanwhile, your blob model output is always going to be at least a bit "blobby", so your fine-tuned style model is already ruined as a completely generic model, and that's the whole point. Why would anyone use your "blob style" model if they don't want blobby images? It's easy enough to switch models. So it's better to just leave "blob style" out of your captions.

The reason for the traditional advice is captioning rule #3. By repeating the word "style", you ensure that the training ends up amplifying the elements of style in the images. But the issue is that "style" is too generic to work well. It can mean artistic, fashionable, or a type of something (e.g. "style of thermos"). So SD doesn't know what part of the images to map the concept of style. In my experience, putting it in doesn't make the model more effective.

- **Use words with the right level of specificity: common but not too generic.**

This is a hard to understand idea that's related to captioning rule #5. SD will take each word in your captions and match it with a concept that it recognizes in your trainers. It can do that because it already has visual associations with that word. It will then blend the visual information from in your trainers with its existing visual associations. If your caption words are too generic, that will cause lack of style transfer, because there are too many existing visual associations. Here's an example. Let's say that one of your trainer images for your style model happens to contain an visual of a brandy snifter. If you caption that as "a container", the base SD model knows a million examples of container that come in vastly different sizes and shapes. So they style of the brandy snifter becomes diluted.

object model that that exact special snifter specifically, you would want caption like that. Essentially, this tells SD, "the snifter you see in the image is unique from other snifters - it's a specialblob." That way when you prompt "specialblob", the output will be that exact snifter from the training image rather than some generic snifter. But for a style model, you don't care about the snifter itself but rather the style (e.g. swirly brush strokes) of the snifter.

Rather than "container" or "snifter", a good middle-ground of specificity might be "glassware". That's a more common word, yet all glassware all somewhat similar - at least semi-transparent and liquid holding. This middle-ground allows SD to match the snifter with a smaller pool of similar images, so swirliness of your trainer image is less diluted. I only have limited anecdotal evidence for this advice, and it's very subjective. But I think using simple common words is a good strategy.

- **You may or may not want to caption things that are true of ALL the training images**

Here the rules conflict, and I don't have solid advice. Captioning rule #3 is that words repetitions will be amplified. So if All of the trainers are "paintings with "swirly brush strokes", then theoretically including those words in the captions will make the training pay attention to those concepts in the training images and amplify them. But trainer rule #2 is that visual repetitions will be amplified even if you don't caption them. So the swirliness is gauranteed to be learned anyway. Also, captioning rule #1 is that if you do include "swirly brush strokes" in the caption for every image, then you'll also need to include those words in the prompt to make the model generate that style most effectively. That's just a pain and needlessly eats up prompt tokens.

This likely depends on how generic these concepts are. Every training image could be captioned as "an image". But that's certainly useless since an image could literally look like anything. In this example, where every image is a painting, you could also use the caption "painting" for every trainer. But that's probably also too generic. Again, relating to rule #5, the captioned visual concepts get blended with existing SD's existing visual concepts for that word, so that's blending with the millions of styles of "painting" in LAION. "Swirly brush strokes" might be specific enough. Best to experiment.

How to caption ~object~ models

You can find proof for most of this advice in my other post that shows an apples to apples comparison of object captioning methods.

- **DO use keywords - e.g. "a blob person". (opposite from style models)**

Let's say that you're training yourself. You need a special keyword (aka "blob") to indicate that you are a special instance of a generic object, i.e. "person". Yes, you are a special "blob person"! Every training image's caption could be nothing more than "blob person". That way, the prompt "blob person" will generate someone who looks like you, while the prompt "person" will still generate diverse people.

However, you might want to pair the special keyword with multiple generic objects. For example, if you're training yourself, you may want to use "blob face" for closeups and "blob person" or "blob woman" for long-shots. SD is sometimes bad at understanding that a closeup photo of an object is the same object as a long-shot photo of that object. It's also pretty bad at understand the term "closeup" in general.

- **The LESS description the better. (opposite from style models)** captioning rule #1 and its opposite. For every caption word you use, the corresponding detail of the training images will be regenerated when you use that word in the prompt. For an object, you don't want that. For example, let's say a trainer has a window in the background. If you caption "window", then it's more likely that if you put "window" into the prompt, it'll generate that specific window (over-fitting) rather than many different windows.

Similarly, you don't want to caption "a beautiful old black blob woman", even when all of those adjectives are true. Remember caption rule #3. Since that caption will be repeated for every trainer, you're teaching the model that every "beautiful old black woman" looks exactly like you. And that concept will bleed into the component concepts. So even "old black woman" will look like you, and probably even "old black man"! So use as few words as possible, e.g. "blob woman".

There are cases were you do need to use more than just "blob person". For example, when the photos of you have some major difference, such as a two different hairstyles. In that case, SD will unsuccessfully try to average those differences in the output, creating a blurry hairstyle. To fix that, expand the captions as little as needed, such as to "blob person, short hair" and "blob person, long hair". That also allows you to use "short" and "long" in the prompts to generate those hairstyles separately. Another example is if you're in various different positions. In that case, for example, you might caption, "blob person, short hair, standing" and "blob person, short hair, sitting."

SD already understands concepts such as "from above" and "from below", so you don't need to caption the angle of the photo for SD to be able to regenerate those angles. But if you want to reliably get that exact angle, then you should caption it, and you'll need several trainer images from that same angle.

- **For multiple concepts, describe the image in order from most to least prominent concept. (same as for style models)**

Read the same advice for style models above for the full explanation. This is less important for an object model because the captions are so much shorter - maybe as short as "blob person". But if you're adding hair style to the caption, for example, then the order you want is "blob person, short hair" since "person" is more prominent and bigger in the trainer image than "hair".

In my Emotion-Puppeteer model, I captioned each images as "X face, Y eyes, Z mouth". The reason for "X face" is that I wanted to differentiate between "plain" and "cute" faces. Face is first because it's a bigger and broader concept that eyes and mouths. The reason for "Y eyes" and "Z mouth" is that I wanted to be able to "puppeteer" the mouth and eyes separately. Also, it wouldn't have worked to caption, "angry face" or "angry emotion" because an angry person may be frowning, pouting, gnashing their teeth. SD would have averaged those very different trainers together into a blurry or grotesque mess. After face, eyes, and mouths, I also included the even less prominent concepts of "closeup" and "looking straight". All of those levers were successfully trained.

- **Use words with the right level of specificity: common but not too generic. (same as for style models)**

Read the same advice for style models above for the full explanation. This is a bit tricky. If you are a woman, you could theoretically caption yourself as "blob image", "blob person", "blob woman", "blob doctor", or "blob homo sapiens". As described above, "image" is way too generic. "Doctor" is too specific, unless your images are all of you in scrubs and you want the model to always generate you in scrubs. "Homo sapiens" is too uncommon, and your likeness may get blended (captioning rule #5) with other homo sapiens images that are hairy and naked. "Woman" or "person" are probably the right middle-ground.

and not your mouth - i.e. "smizing", and it's also possible to smile with your mouth and not your eyes - i.e. a "fake smile". So in an earlier version of my model, I used the caption "smiling eyes". This didn't work well because the base SD model has such a strong association of the word "smile" with mouths. So whenever I prompted "smiling eyes, frowning mouth", it generated smiling mouths.

To fix this in the latest model, I changed the caption to "pleasing eyes", which is a very specific and uncommon word combination. Since the LAION database probably has few instances of "pleasing eyes", it acts like a keyword. It ends up being the same as if I had used a unique keyword such as "blob eyes". So now when you prompt "pleasing eyes", the model gives you eyes similar to my training images, and you can puppeteer those kind of eyes separately from the mouths.

Learning rate

The slower the better, if you can stand it. My Emotion-Puppeteer model was trained for the first third of its steps at 1.5e -6, then sped up to 1.0e -6 for the final two-thirds. I saved checkpoints at several stages and published the model with that generates all of the eye and mouth keywords the most reliably. However, that published model is "over-trained" and needs CFG of 5 or else the output looks fried. I had the same problem with my Technicolor-Diffusion model: the style didn't become reliable until the model was "over-trained".

The solution is either an even slower learning rate or even more training images. Either way, that means a longer training time. Everydream2 defaults to 1.5e -6, which is deffo too fast. Dreambooth used to default to 1.0e -6 (not sure now). Probably 5e -7 (aka half the speed of 1.0e -6) would be best. But damn, that's slow. I didn't have the patience. Some day I'll try it.

The best training software

- **As of Feb 2023, Everydream2 is the best checkpoint training software.**

Note that I'm not affiliated with it in any way. I've tried several different options, and here's why I make this claim: Everydream2 is definitely the fastest and probably the easiest. You can use training images with several different aspect ratios, which isn't possible in most other software. Lastly, it's easy to set up on Runpod if you don't have an expensive GPU. Everydream2 doesn't use prior-preservation or a classifier image set. That's no longer necessary to prevent over-fitting, and that saves you time.

Of course, this could all be obsolete soon given how quickly as things keep advancing!

If you have any experience that contradicts this advice, please let me know!

Archived post. New comments cannot be posted and votes cannot be cast.

245 113 Share

Freonr2 • 3y ago • Edited 3y ago

Great writeup! Definitely not a ton of great info out there for people doing large projects that extend beyond your basic "here's my face, 'dreambooth' it" type stuff.

ED2 author here, a few notes:

- Shouldn't need to worry too much about downsizing your images prior to training, they're resized on the fly (bicubic, which should be best general case resize), and crop jitter feature needs them to be slightly larger than your target training size (i.e. if training at 512, you ideally want like 520x520 bare minimum, but 2000x2000 is fine too, I personally recommend 1.5+ megapixel just to allow yourself headroom to train at higher res in the future as tech improves). You can feed in 4K images if you want, shouldn't have any appreciable impact on performance as the data loader is multithreaded and preloads stuff on CPU. Having 4k+ images shouldn't hurt anything but your disk space. You may kick yourself in the future if you resize everything to 512x512 or 768x768 or whatever. Crop jitter is also a quality improvement and it needs "buffer" in the training image size to slice off a few edge pixels to shift the image around every epoch. Here's a video that talks about crop jitter and a bit about resolution and aspects, etc: https://www.youtube.com/watch?v=0xswM8QYFD0
- You might consider toying with conditional dropout especially to "force" a style into the model, but high values can start to cause weird behavior. Its a way to help make a style take over the whole model. Conditional dropout is a fairly powerful tool. I might suggest if you want to completely take over the model with style using 0.10-0.15. Higher values will cause bleeding, especially at lower CFG scale at inference.

> Order captions from most to least prominent concept

Definitely, character names should be up front, and if you have 2+ characters better to just list their names instead of trying to cram outfit information in as well, and instead use the solo images to details outfits and such, and keep your 2+ character images to <15%, maybe even <10%, but you can train SD to paint 2 characters at once if you give it enough data and examples. 3+ is still very elusive, probably needs inference tricks, inpainting, maybe some controlnet stuff would help now.

You mention starting at 1.5e-6 then going to 1e-6, makes sense, make sure you use the chaining feature. You can setup a few copies of train.json (or look at chain0.json, chain1.json etc) with different settings and run them from a batch file in order. `"resume_ckpt": "findlast"` will resume from the last training sessions. There's an example `chain.bat` (can rename to .sh for linux) in the repo and chain0.json, chain1.json, chain2.json that shows how you can chain them together. Only the first chain0 would use "resume_ckpt": "sd_v1-5_vae" or whatever base model, then the rest use "findlast" to resume in order. This means you can tweak any setting and walk away to let something run overnight and have it change settings as it goes. I feel chaining is a bit underutilized in the community.

For training smaller dreambooth type models, I've found it useful to actually copy your training images, one with a full caption, the other with just the person's name. Ex. "joe smith" and "joe smith in a blue cardigan sitting at a desk". Most useful when you are just doing a face/person with like 20-40 images.

42 ···

4 ···

⊕ **Angelotheshredder** • 3y ago

⊕ 6 more replies

ProGamerGov • 3y ago

There's also a recent discovery from the end of January about how to improve finetuning of images bright and dark images: https://www.crosslabs.org/blog/diffusion-with-offset-noise

Stable Diffusion generally defaults to 50% gray (half way between black and white) regardless of training data, while the training fix allows you to reach the full range of white to black.

⊖ 5 ···

Freonr2 • 3y ago

Just added this to EveryDream2 the other day.

https://huggingface.co/panopstor/ff7r-stable-diffusion/blob/main/zero_freq_test_biggs.webp

It definitely helps, even just set to 5% (0.05). Improves contrast and color saturation to be more true to the originals.

Suggest using it.

New arg is "zero_frequency_noise_ratio" and use a ratio number just like conditional dropout. i.e. "0.05" would be 5%.

5 ···

⊕ 3 more replies

⊕ 2 more replies

gxcells • 3y ago

Why do you need to denoise the original images? Wouldn't it make them to "flat" and smooth especially for photography?

5 ···

⊕ 1 more reply

hsadg • 3y ago

Awesome write up! Saved for when I inevitably try it myself :)

4 ···

Skip to main content

Log In

any recommendations on denoising software?

4 ...

Distinct-Quit6909 • 3y ago

The top noise reduction software is ai driven and there's no reason I can think that pix2pix wont get the job done with a bit of encouragement, I'm going to attempt it with pix2pix "remove noise" .

4 ...

1 more reply

terrariyum OP • 3y ago

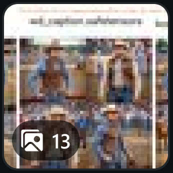
I use Affinity Photo's build in denoise, which works well and is very customizable. clipdrop.co is free, but a pain for bulk. The upscale in Auto111 extras also denoises, but I find it to be too extreme hard to customize and hard for bulk.

4 ...

1 more reply

r/StableDiffusion • 1 yr. ago

Compared Effect Of Image Captioning For SDXL Fine-tuning / DreamBooth Training for a Single Person, 10.3 GB VRAM via OneTrainer, WD14 vs Kosmos-2 vs Ohwx Man, More Info In Comments



19 upvotes · 21 comments

r/Frat • 4 yr. ago

Drop your best insta caption

317 upvotes · 69 comments

r/StableDiffusion • 3 yr. ago

Consistency - Definitive Guide to Having Multiple Faces of the Same Character. the brazilian guide uahuahuaha guide in the comments (elegant, highly detailed, digital paintin...Oops... for...

255 upvotes · 42 comments

r/StableDiffusion • 1 yr. ago

InternVL-State of the Art Captioning Model

Skip to main content

Log In

r/funny • 12 yr. ago

Inappropriate Captions

1.3K upvotes · 234 comments

r/StableDiffusion • 1 yr. ago

FLUX is smarter than you! - and other surprising findings on making the model your own

653 upvotes · 148 comments

r/StableDiffusion • 2 yr. ago

What is the best model for caption images with all the details that exist in the image and also that is uncensored?

9 upvotes · 8 comments

r/StableDiffusion • 3 yr. ago

CharTurner - A work in progress resource for character artists.

200 upvotes · 100 comments

r/StableDiffusion • 1 yr. ago

Lora - image captioning best practices

13 upvotes · 11 comments

r/StableDiffusion • 8 mo. ago

Is there a video captioning (scene description) model?

5 upvotes · 2 comments

r/onlyfansadvice • 2 yr. ago

Blanking on captions, need advice on good formulas to reliably come up with captions

10 upvotes · 22 comments

r/StableDiffusion • 3 yr. ago

What does overtraining look like? An Experiment

8 upvotes · 8 comments

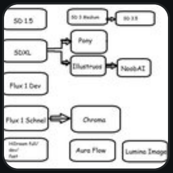
r/AskWomen • 9 yr. ago

Skip to main content

Log In

r/StableDiffusion • 4 mo. ago

Do I get the relations between models right?



543 upvotes · 98 comments

r/StableDiffusion • 9 mo. ago

JoyCaption: Free, Open, Uncensored VLM (Progress Update)

301 upvotes · 46 comments

r/StableDiffusion • 4 mo. ago

What the best model for character consistency right now?

1 upvote · 6 comments

r/StableDiffusion • 1 yr. ago

JoyCaption: Free, Open, Uncensored VLM (Alpha One release)

457 upvotes · 131 comments

r/StableDiffusion • 1 mo. ago

How to create consistent body and face?

10 comments

r/StableDiffusion • 3 yr. ago

I'm going insane trying to train large datasets for poses, any input would be greatly appreciated I've been stuck for days

17 upvotes · 12 comments

r/StableDiffusion • 3 yr. ago

Discussion on training face embeddings using textual inversion

5 upvotes · 16 comments

r/StableDiffusion • 1 mo. ago

Has multi-subject/character consistency been solved? How do people achieve it?

4 upvotes · 7 comments

r/StableDiffusion • 13 days ago

Log In

r/StableDiffusion • 2 mo. ago

As someone who is already able to do 3d modelling, texturing, animation all on my own, is there any new ai software that i can make use of to speed up my workflow or improve the quality of my outputs?

9 upvotes · 11 comments

r/StableDiffusion • 3 mo. ago

is there a model that can relight an image?

8 upvotes · 9 comments

r/modhelp • 8 mo. ago

Adding captions when posting multiple photos

2 upvotes · 14 comments

LANGUAGES

Français

Português

TOP POSTS

Reddit

reReddit: Top posts of February 17, 2023

Reddit

reReddit: Top posts of February 2023

Reddit

reReddit: Top posts of 2023

Reddit Rules Privacy Policy User Agreement Accessibility