FAQ

How do I setup the leaked NovelAI model?

0. Download the leak:
   `magnet:?xt=urn:btih:5bde442da86265b670a3e5ea3163afad2c6f8ecc&dn=novelaileak&tr=udp%3A%2F%2Ftracker.opentrackr.org%3A1337%2Fannonce&tr=udp%3A%2F%2F9.rarbg.com%3A2810%2Fannounce&tr=udp%3A%2F%2Ftracker.openbittorrent.com%3A6969%2Fannounce&tr=http%3A%2F%2Ftracker.openbittorrent.com%3A80%2Fannounce&tr=udp%3A%2F%2Fopentracker.i2p.rocks%3A6969%2Fannounce`
1. Move one of the model files (model.ckpt) from stableckpt to webui/models/Stable-diffusion
    - Feel free to rename it
2. Move stableckpt/animevae.pt to webui/models/Stable-diffusion and rename it to be the same name as your NovelAI model but with .vae.pt
    - Note: optional, it may be worth experimenting with/without vae as sometimes non-vae results can be better
3. Move stableckpt/modules/modules/*.pt (anime.pt, anime_2.pt, etc) to webui/models/hypernetworks
    - Create the directory if it does not exist
4. Launch the webui and the model should appear with the others in the settings

Preface prompt with: Masterpiece, best quality,
NovelAI's default negative prompt:
lowres, bad anatomy, bad hands, text, error, missing fingers, extra digit, fewer digits, cropped, worst quality, low quality, normal quality, jpeg artifacts, signature, watermark, username, blurry
CFG scale of 12
CLIP skip 2
Eta noise seed delta 31337

Leak pt2 (not required, including it here just for those that are interested):
https://pastebin.com/6wX7Bx7w

NAIFU: NovelAI model + backend + frontend

From an anon (>>>/g/89097704):

> Runs on Windows/Linux on Nvidia with 8GB RAM. Entirely offline after installing python dependencies.
> Web frontend is really good, with danbooru tag suggestions, past generation history, and mobile support.
> `magnet:?xt=urn:btih:4a4b483d4a5840b6e1fee6b0ca1582c979434e4d&dn=naifu&tr=udp%3a%2f%2ftracker.opentrackr.org%3a1337%2fannounce`

Looks like it comes with some of the same models/hypernetworks from the first leak, so if you already downloaded that you can save yourself some time.

See the README.txt in the torrent for setup/usage instructions.
Extract it like this: https://imgur.com/a/gvUCiCy.gif

Using a different model in NAIFU

1. Copy the directory of the one you want from stableckpt (the first leak) into the models folder alongside animefull-final-pruned

    - you can skip moving files around and just do the next step, but make sure you use the absolute path to the model's directory
2. Edit run.bat or run.sh and change MODEL_PATH to point to the directory that contains the model you want
3. Restart NAIFU to load the new model

You can also run non-NovelAI models as long as you edit the run script and point to a directory that contains a model named "model.ckpt", and a "config.yaml" file. I don't know what changes you might need in the yaml file but just copying one from the other NovelAI models seemed to work fine.

xformers (increase your it/s) (MORE CARDS SUPPORTED)

> Make sure you `git pull` to the latest webui version first

> Pascal, Turing and Ampere are now supported automatically just by using --xformers as of Oct 10th.

Using xformers will affect your generated images somewhat.

If you are running an Pascal, Turing and Ampere (1000, 2000, 3000 series) card
Add `--xformers` to COMMANDLINE_ARGS in webui-user.bat and that's all you have to do.

If you are running an older card you need to build xformers yourself and force the webui to accept it.

> Some anons have reported significantly worse performance with xformers on 700 and 900 series cards, consider this before you proceed

Add `--force-enable-xformers` to COMMANDLINE_ARGS in webui-user.bat, then either use a prebuilt xformers or build it yourself

Prebuilt xformers

1. Download the .whl for your GPU from https://mega.nz/folder/f1UAyaLL#50Sq07s18kC3Tn095LZ8zQ
    - GTX 900 series = 5.2
    - You can double check with https://developer.nvidia.com/cuda-gpus under CUDA-Enabled GeForce and TITAN Products
2. Place it in the webui folder
3. Open up cmd prompt or bash in that folder
4. `venv\Scripts\activate.bat` or `source ./venv/bin/activate`
5. `pip install xformers-0.0.14.dev0-cp310-cp310-win_amd64.whl`

If the .whl isn't available for your GPU architecture you will need to build it yourself, or obtain the .whl from someone else who has built it for your architecture.

Building xformers

> If the path to your webui folder at all long (like 40-50 characters idk exact number), you will run into major problems. See the **Common problems** section below

Make sure your **Python version is 3.10 or later**
`python --version`
Install **CUDA Toolkit 11.3** https://developer.nvidia.com/cuda-11.3.0-download-archive
Install **Build Tools for Visual Studio 2022** https://visualstudio.microsoft.com/downloads/?q=build+tools#build-tools-for-visual-studio-2022
(You only need Desktop development with C++)

Open up cmd prompt/bash

Confirm nvcc is available
`nvcc --version`

Go to the webui directory
`cd C:\path\to\SD\stable-diffusion-webui`

Download xformers repo
`cd repositories`
`git clone https://github.com/facebookresearch/xformers.git`
`cd xformers`

`git submodule update --init --recursive`

Create venv and activate
`python -m venv venv`
Depending on where you're running from (cmd prompt, powershell, bash), run either
`venv\Scripts\activate.bat` or `source ./venv/bin/activate`

To avoid issues with getting the CPU version, install pyTorch seperately
`pip install torch torchvision --extra-index-url https://download.pytorch.org/whl/cu113`

Install the rest of the dependencies
`pip install -r requirements.txt`
`pip install wheel`
`pip install ninja`

Force enable CUDA to be built with MS Build Tools 2022
cmd prompt: `set NVCC_FLAGS=-allow-unsupported-compiler`
bash: `export NVCC_FLAGS=-allow-unsupported-compiler`
For the next part you need to set TORCH_CUDA_ARCH_LIST so it uses your architecture. Grab your GPU arch from these lists:
https://developer.nvidia.com/cuda-gpus (consumer GPUs will be under CUDA-Enabled GeForce and TITAN Products)
For example, if your GPU is a GTX 1070, based on that list the architecture is 6.1 and the command would be:
`set TORCH_CUDA_ARCH_LIST=6.1`

Set it for your architecture:
cmd prompt: `set TORCH_CUDA_ARCH_LIST=<YOUR ARCH>`
bash: `export TORCH_CUDA_ARCH_LIST=<YOUR ARCH>`

Build xformers
`python setup.py build`
This may take a long time.

Build the .whl
`python setup.py bdist_wheel`

Copy the resulting .whl file to the webui folder
`copy dist\xformers*.whl ..\..\`

Activate the webui venv
`cd ..\..`
`venv\Scripts\activate.bat` or `source ./venv/bin/activate`

`dir xformers*` or `ls xformers*`
Copy the full name of the .whl
And install it
`pip install <FULL .whl FILENAME>`

Then you should be good to go.

Common problems

> Filename too long
> or
> fatal error C1083: Cannot open compiler generated file: '': Invalid argument
> error: command 'C:\Program Files\NVIDIA GPU Computing Toolkit\CUDA\v11.3\bin\nvcc.exe' failed with exit code 4294967295

Windows problem with path lengths.

Move your stable-diffusion-webui folder higher up and/or shorten its name.
For example, either of these should be fine unless you have a really long username:
C:\stable-diffusion-webui\
C:\Users\<name>\sd-webui\

> RuntimeError: CUDA error: no kernel image is available for execution on the device

More GPU architectures are now automatically supported, try running a clean install and just use the flag --xformers

If you move this then you will have to delete the venv directory inside (run webui-user.bat again to recreate it), or deal with moving it manually

> ./venv/bin/activate: No such file or directory

If you are using bash on Windows, source from Scripts.
`source ./venv/Scripts/activate`

> If you encounter some error about torch not being built with your cuda version blah blah, then try:

`pip install setuptools==49.6.0`

What does ()/[]/{} or (word:number) mean?

() adds emphasis to a term, [] decreases emphasis, both by a factor of 1.1. You can either stack ()/[] for increasing/decreasing emphasis or use the new syntax which takes a number directly - it looks like this:
(word:1.1) == (word)
(word:1.21) == ((word))
(word:0.91) == [word]
To use literal ()/[] in your prompt, escape them with \
See https://github.com/AUTOMATIC1111/stable-diffusion-webui/wiki/Features for full details and additional features.

{word} is for NovelAI's official service only. It is similar to (word) but the emphasis is only increased by a factor of 1.05. If you are using the leaked models in the webui you shouldn't be using this syntax.

float16 vs. float32?

float32 for older gpus or if you want 100% precision. The outputs of both should be nearly identical, the main difference is size and the gpus that support it.

What does model1+model2 (WD+yiffy/50% SD + 50% WD) mean?

Refers to merged models, see the "Checkpoint Merger" tab in the webui.

What is the best model?

It depends the type of stuff you want to generate. Generally just grab the latest model/highest epoch of the type you want in https://rentry.org/sdmodels

AND syntax

For now, see https://energy-based-model.github.io/Compositional-Visual-Generation-with-Composable-Diffusion-Models/