Stable Diffusion Locally: Getting Started

January 29, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** You can generate AI images on your own GPU for free — no subscription, no filters, no upload limits. SD 1.5 runs on 4GB VRAM (GTX 1060 and up). SDXL needs 8GB minimum. Flux needs 12GB+. For the fastest start, install Fooocus — download, extract, double-click, and you're generating in under five minutes. For more control, use ComfyUI (power users) or Forge (familiar A1111-style UI with better performance). Start with SDXL and the Juggernaut XL checkpoint.

📚 **More on this topic:** Flux Locally · ComfyUI vs A1111 vs Fooocus · What Can You Run on 8GB VRAM · GPU Buying Guide · Planning Tool

Stable Diffusion is a text-to-image AI model you can run on your own GPU. Type a description, hit generate, and get an image — no cloud service, no subscription, no per-image fee, no content filter telling you what you can and can't create. Everything runs on your machine.

That's the appeal. The barrier used to be complexity: Python environments, command-line installs, YAML configs. In 2026, that barrier is mostly gone. Tools like Fooocus and ComfyUI Desktop have one-click installers. If you have a GPU with 4GB+ VRAM, you can be generating images in under five minutes.

This guide covers what you need, which software to pick, how to generate your first image, and how to get better results once you're running.

Why Run Stable Diffusion Locally

- **Free and unlimited.** No credits, no API costs, no monthly subscription. Generate 10,000 images a day if your GPU can handle it.
- **Private.** Nothing leaves your machine. No prompts logged on someone else's server, no images stored in someone else's cloud.
- **Uncensored.** No corporate content policy deciding what you can generate. You control the model, the prompts, and the output.
- **Customizable.** Install community models, LoRAs, and extensions that add styles, subjects, and capabilities the base model doesn't have.
- **Offline.** After initial setup and model download, everything works without internet.

The tradeoff is hardware. You need a GPU — and the more VRAM it has, the bigger models you can run and the faster images generate. But even a modest card works.

Hardware Requirements

Model	Min VRAM	Comfortable VRAM	Native Resolution	Notes
SD 1.5	4 GB	6 GB	512x512	Runs on GTX 1060, entry-level cards
SDXL	8 GB	12 GB	1024x1024	The current mainstream standard
SD 3.5 Medium	8 GB	12 GB	1024x1024	Better prompt understanding
SD 3.5 Large	12 GB	16 GB	1024x1024	Higher quality, heavier
Flux (NF4)	8-10 GB	12 GB	1024x1024	Best quality, newer
Flux (FP16)	24 GB	24 GB+	1024x1024	Full precision, GPU-hungry

What GPU do you have?

Your GPU	What You Can Run
GTX 1060 / 1660 (6 GB)	SD 1.5 comfortably, SDXL with `--medvram` flag
RTX 3060 Ti / 4060 (8 GB)	SD 1.5, SDXL (tight), Flux NF4 (tight)
RTX 3060 / 4070 (12 GB)	SD 1.5, SDXL comfortably, Flux NF4
RTX 4070 Ti / 3090 (16-24 GB)	Everything, including Flux FP16

Beyond the GPU: 16GB of system RAM minimum (32GB recommended for comfortable multi-model workflows). An SSD matters — models are 2-7GB files, and loading from a spinning disk is painful. Budget 30GB+ of storage for models, LoRAs, and generated images.

Pick Your Software

There are several frontends for Stable Diffusion. Here's how they compare:

Software	Best For	Learning Curve	Model Support	Install
Fooocus	Absolute beginners	Easy	SDXL only	Download + double-click
ComfyUI Desktop	Power users, production	Medium-steep	All models (Flux, SD3.5, video)	Installer app
Forge	A1111 users wanting speed	Easy-medium	SD 1.5, SDXL, Flux	Git clone + batch file
AUTOMATIC1111	Legacy workflows, tutorials	Easy-medium	SD 1.5, SDXL	Git clone + batch file
SD.Next	AMD/Intel GPUs, bleeding-edge	Medium	Everything + video	Git clone + batch file

Fooocus: Just Works

Built by the same developer as ControlNet. Three clicks from download to first image. It auto-enhances your prompts using GPT-2 (similar to Midjourney's hidden preprocessing), includes 276 art style presets, and handles all the technical settings behind the scenes.

The catch: Fooocus entered limited maintenance mode in August 2024 and only supports SDXL. No Flux, no SD 3.5. For beginners who just want to generate images today, it's still the fastest path. When you outgrow it, move to ComfyUI or Forge.

ComfyUI: The Power Tool

ComfyUI uses a node-based interface — you connect processing blocks (load model, set prompt, sample, decode, save) on a visual canvas. Workflows save as JSON files that anyone can reproduce exactly. It's 15% faster than A1111 on the same hardware and has dramatically better VRAM management.

ComfyUI supports every new model first (Flux, SD 3.5, Hunyuan, video generation) and has become the standard for serious image generation work. The learning curve is real — expect 10-20 hours before you're comfortable — but the investment pays off.

ComfyUI Desktop (released January 2025) wraps all of this in a native application with a one-click installer. It bundles ComfyUI Manager for easy extension installation. This is the recommended way to install ComfyUI now.

Forge: A1111 But Faster

A fork of AUTOMATIC1111 by the Fooocus/ControlNet developer. Same familiar UI, but 30-75% faster on GPUs with 6-8GB VRAM. ControlNet, FreeU, and SVD come pre-installed. Supports Flux natively. If you've used A1111 before or are following A1111 tutorials, Forge is the drop-in upgrade.

AUTOMATIC1111: The Original

Still the largest tutorial ecosystem and community. Development has slowed significantly — the master branch currently has a broken dependency on a deleted Stability AI repository, so new installs should use the `dev` branch. Most advanced users have migrated to ComfyUI or Forge, but A1111 remains a solid choice if you want the widest library of existing guides and extensions.

SD.Next: Widest Hardware Support

Fork of A1111 with aggressive feature development. The standout feature is hardware support: NVIDIA CUDA, AMD ROCm (Windows and Linux), Intel Arc/IPEX, DirectML, and OpenVINO. If you're on an AMD GPU or Intel Arc, SD.Next is your best option. It also has native video generation support.

Quick Start with Fooocus

The fastest path from zero to generating images.

Step 1: Check Your GPU

Open a terminal and run:

```
nvidia-smi
```

Look for your GPU name and VRAM. If you see 6GB+ VRAM, you're set. If you don't have an NVIDIA GPU, skip to ComfyUI Desktop or SD.Next.

Step 2: Install Fooocus

Windows:

1. Download the latest release from github.com/lllyasviel/Fooocus/releases (the `.7z` file)
2. Extract with 7-Zip
3. Double-click `run.bat`

Linux:

```
git clone https://github.com/lllyasviel/Fooocus.git
cd Fooocus
python -m venv venv
source venv/bin/activate
pip install -r requirements_versions.txt
python entry_with_update.py
```

First launch downloads the Juggernaut XL model (~6.6 GB). Your browser opens to `127.0.0.1:5000`.

Step 3: Generate Your First Image

Type a description in the prompt box. It can be simple:

```
a cabin in a snowy forest at sunset
```

Click **Generate**. Fooocus expands your prompt automatically, picks good settings, and generates the image. On an RTX 4060, expect about 20-30 seconds per image at 1024x1024.

Step 4: Iterate and Refine

Once you have a first image:

- **Change the style preset** (bottom of the screen) — try "Cinematic," "Anime," "Watercolor" to see how the same prompt looks in different styles
- **Add detail to your prompt** — be specific about lighting, composition, and subject
- **Use the Advanced tab** for more control: adjust aspect ratio, enable the refiner, change the number of images per batch
- **Try image-to-image** — upload an existing image and use a prompt to modify it

---

Quick Start with ComfyUI Desktop

If you want more control from the start, or need Flux/SD 3.5 support:

1. Download the installer from comfy.org/download
2. Run it, select your GPU type, choose an install location
3. ComfyUI opens with a default workflow — load a model, enter a prompt, click **Queue Prompt**
4. Install ComfyUI Manager (bundled) to easily add custom nodes

You'll need to download a model separately. Place `.safetensors` checkpoint files in `ComfyUI/models/checkpoints/`. Start with **Juggernaut XL** from Civitai or Hugging Face.

Understanding Models, LoRAs, and Checkpoints

Term	What It Is	File Size	What It Does
Checkpoint	The full model	2-7 GB	Defines overall style (photorealistic, anime, etc.)
LoRA	Small add-on adapter	10-200 MB	Adds specific concepts — a character, a style, an object
VAE	Image encoder/ decoder	300-800 MB	Affects color vibrancy and detail quality
Embedding	Trained text trigger	<100 KB	Minor style tweaks, commonly used as negative prompts

Checkpoints are the base. You load one checkpoint at a time, and it determines your image style. Think of it like choosing a camera — a photorealistic checkpoint produces photos, an anime checkpoint produces anime.

LoRAs are patches that modify the checkpoint's behavior. Want a specific character? There's a LoRA. Want the style of a particular artist? There's a LoRA. You can stack multiple LoRAs (use weight ~0.6 each when combining to avoid conflicts).

Where to find models:

- **Civitai** — Largest community hub. Thousands of checkpoints, LoRAs, and embeddings with preview images and reviews. Some models require a free account to download.
- **Hugging Face** — Official source for base models (SD 1.5, SDXL, SD 3.5, Flux) from Stability AI and Black Forest Labs.

Recommended Starting Checkpoints

Checkpoint	Base Model	Best For
Juggernaut XL	SDXL	Best all-around SDXL model — photos, architecture, concept art
RealVisXL V5	SDXL	Photorealism specifically
DreamShaper XL	SDXL	Artistic/creative work, digital painting
Pony Diffusion V6	SDXL	Anime and illustration
Realistic Vision	SD 1.5	Photorealistic humans (SD 1.5 ecosystem)

Juggernaut XL, RealVisXL, and Pony Diffusion cover 90% of use cases.

Tips for Better Prompts

Prompt Structure

Stable Diffusion reads prompts left to right and assigns more importance to words that appear earlier. Structure your prompts like this:

1. **Quality tags** — `masterpiece, best quality, highly detailed`
2. **Subject** — what you actually want to see
3. **Setting** — background, environment
4. **Lighting** — time of day, mood
5. **Style** — art style, medium, aesthetic

Example:

```
masterpiece, best quality, a woman reading in a library,
afternoon sunlight through tall windows, warm golden light,
oil painting style, rich colors
```

Negative Prompts

Negative prompts tell the model what to avoid. Most UIs have a separate negative prompt field. A good universal negative prompt:

```
worst quality, low quality, blurry, jpeg artifacts, watermark,
text, signature, extra fingers, mutated hands, poorly drawn face,
deformed, bad anatomy, extra limbs
```

Or use the **EasyNegative** embedding (download from Civitai) — type `easynegative` in the negative prompt field and it handles common issues in a single token.

What Helps

- **Be specific.** "A red-haired woman in a leather jacket, standing on a rainy city street" beats "woman in city."
- **Use camera/lens terms** for photorealism: `DSLR, 85mm lens, f/1.4, bokeh, shallow depth of field`
- **Reference art styles**: `concept art, watercolor, oil painting, Studio Ghibli style, cyberpunk`
- **Weighted terms**: `(rainy night:1.3)` increases emphasis. Keep weights between 0.5 and 1.5 — higher values cause artifacts.

What to Avoid

- **Prompt stuffing.** A wall of quality tags (`beautiful amazing epic detailed ultra realistic photo masterpiece hd best quality trending`) creates noise. Pick 2-3 quality tags and stop.
- **Excessive weighting on SDXL.** Multiple `((( )))` brackets make SDXL outputs look overcooked.
- **CFG scale too high.** 7-9 is the sweet spot. Above 12, images become oversaturated and distorted.

Common Issues and Fixes

Out of VRAM

The most common problem. Your GPU doesn't have enough memory for the model or resolution you're trying to use.

Fixes (in order):

1. **Reduce resolution.** 512x512 uses 75% less VRAM than 1024x1024.

2. **Enable xformers** (`--xformers` flag in A1111/Forge). 20-30% faster and ~50% less VRAM.
3. **Use `--medvram`** (A1111/Forge). Splits model processing to save VRAM — moderate speed penalty.
4. **Use `--lowvram`** as a last resort. Heavy speed penalty (50-70% slower), but lets 4GB cards generate.
5. **Switch to ComfyUI.** Its dynamic model loading handles low-VRAM GPUs better than any A1111-based UI.
6. **Use a smaller model.** Drop from SDXL to SD 1.5, or from Flux FP16 to Flux NF4.

Slow Generation

- **Enable xformers or SDP attention.** If you're running without attention optimization, you're leaving 20-30% performance on the table.
- **Lower sampling steps.** 20-25 steps produces good results in most cases. 50 steps is rarely worth the extra time.
- **Check `nvidia-smi`.** If GPU utilization is low, something else might be using your VRAM. Close Chrome, video players, and game launchers.
- **Use efficient samplers.** DPM++ 2M Karras and Euler a converge fastest.

Black Images

Usually a precision issue on older GPUs (GTX 10-series, 16-series):

1. Try `--no-half-vae` first (keeps only the VAE at full precision)
2. If that fails, try `--upcast-sampling`
3. Last resort: `--no-half` (doubles VRAM usage)

Bad Faces and Hands

Stable Diffusion is notoriously weak at hands. Faces can also look distorted, especially at lower resolutions.

- **ADetailer extension** (A1111/Forge): Automatically detects faces and hands, masks them, and re-inpaints at higher quality. Install from `https://github.com/Bing-su/adetailer`
- **Fooocus Enhance feature**: Built-in face/hand improvement — no extension needed
- **Higher resolution helps.** Faces look much better at 768x768+ than at 512x512.
- **Inpaint manually.** Generate the full image, then use the inpainting tool to regenerate just the face or hands.

Upgrade Path

Faster GPU = faster generations and bigger models.

GPU Tier	VRAM	What It Unlocks for Image Gen	Street Price (Jan 2026)
GTX 1060 / 1660	6 GB	SD 1.5 only	Already own it
RTX 4060 / 3060 Ti	8 GB	SDXL (tight), Flux NF4 (tight)	~$250-300
RTX 3060 / 4070	12 GB	SDXL comfortable, Flux NF4	~$200 (3060 used) / $550 (4070)
RTX 4070 Ti / 5060 Ti	16 GB	SDXL with refiner, SD 3.5 Large, Flux FP8	~$450-500
RTX 3090 / 4090	24 GB	Everything — Flux FP16, multiple models, training	~$700-850 (3090 used)

For most people doing image generation, **12GB is the sweet spot** — it runs SDXL and Flux NF4 without workarounds. A used RTX 3060 12GB (~$200) is the cheapest meaningful entry point. If you're serious and want to run Flux at full precision or train LoRAs, a used RTX 3090 (~$700-850) is the value king.

For the full GPU comparison, see our GPU Buying Guide for Local AI.

The Bottom Line

1. **Check your GPU** with `nvidia-smi`. If you have 6GB+ VRAM, you can generate images today.
2. **Install Fooocus** for the fastest start, or **ComfyUI Desktop** if you want long-term flexibility and Flux support.
3. **Start with the Juggernaut XL checkpoint.** It handles photorealism, concept art, and general use well.
4. **Write specific prompts, use negative prompts, keep CFG at 7-9.** Good prompts matter more than settings.
5. **When you hit VRAM limits**, try `--xformers` and `--medvram` before buying new hardware. When those aren't enough, a used RTX 3060 12GB is $200.

Your GPU is an image generator. Start using it.

Related Guides

- GPU Buying Guide for Local AI
- What Can You Run on 8GB VRAM?
- What Can You Run on 12GB VRAM?
- Used RTX 3090 Buying Guide
- Local AI Planning Tool — VRAM Calculator

Sources: Fooocus GitHub, ComfyUI Official Docs, Forge GitHub, SD.Next GitHub, Stable Diffusion Art Prompt Guide, AIArty GPU Requirements, SDXL System Requirements, Hardware Corner Flux GPU Guide, Tom's Hardware SD Benchmarks

---