Stable Diffusion on Mac: Image Generation with MLX and Draw Things

February 26, 2026 · by Mark Bartlett

Download this guide as PDF

> **Quick Answer:** For most Mac users, Draw Things is the clear winner. It's free on the App Store, optimized for Apple Silicon with Metal FlashAttention, and generates SD 1.5 images in 8-15 seconds on a 16GB M2 Pro. It supports SD 1.5, SDXL, Flux, LoRAs, ControlNet, and inpainting with zero terminal work. ComfyUI works on Mac but runs about 3x slower for the same model and takes an hour to set up. MLX stable diffusion is the fastest native option but requires Python scripting. If image gen is your primary use case and budget matters, a $200 used RTX 3060 PC will beat any Mac -- but if you already own a Mac, Draw Things makes it painless.

More on this topic: Stable Diffusion Locally | Flux Locally | ComfyUI vs A1111 vs Fooocus | Best Local LLMs for Mac | Running LLMs on Mac M-Series

Image generation on Mac works. It's slower than an NVIDIA GPU, and some tools aren't as polished as their Linux/Windows versions, but you can generate real images locally on any Apple Silicon Mac right now. The question is which tool to use, and that depends on whether you want ease, speed, or flexibility.

There are three approaches worth considering: Draw Things (easiest, and honestly the best for most people), MLX stable diffusion (fastest native performance), and ComfyUI (most flexible but slowest on Mac). This guide covers all three with actual speed numbers so you can pick the right one.

Draw Things: the recommended starting point

What it is

Draw Things is a free Mac App Store app built specifically for Apple Silicon. It's not a port of a Linux tool – it's a native Mac app with Metal FlashAttention, Core ML acceleration, and on-demand weight loading that keeps memory usage low. The difference in speed compared to generic PyTorch tools on Mac is immediately obvious.

You install it from the App Store. You pick a model from the built-in downloader. You type a prompt and click generate. No Python, no terminal, no dependency hell.

What it supports

- SD 1.5, SD 2.1, SDXL, SDXL Turbo
- Flux.1 (Schnell and Dev), Flux.2
- LoRA loading and on-device LoRA training
- ControlNet (all major types)
- Inpainting and outpainting
- img2img, upscaling, pose editing
- .safetensors model import (bring your own checkpoints)

I've been using Draw Things as my only image gen tool on Mac for months. Haven't needed anything else.

Speed on Mac

Draw Things uses Metal FlashAttention, which is a custom Metal implementation of the attention mechanism. On M3/M4 chips, version 2.0 of this engine delivers about 20% faster inference than earlier versions. It also runs up to 25% faster than mflux and 94% faster than ggml-based implementations for Flux models (tested on M2 Ultra).

Approximate generation times (20 steps, default settings):

Model	Resolution	M1 base 8GB	M2 Pro 16GB	M3 Pro 18GB	M4 Max 36GB+
SD 1.5	512x512	20-30s	8-15s	6-12s	3-6s
SDXL	1024x1024	Too slow	25-40s	18-30s	8-15s
Flux Schnell	1024x1024	Won't fit	30-50s	20-35s	10-18s
Flux Dev	1024x1024	Won't fit	Very slow	40-60s	15-25s

These times are for the total generation, not per-step. Draw Things is roughly 3x faster than ComfyUI running the same model on the same Mac.

Memory requirements

Draw Things uses on-demand weight loading, which reduces memory overhead by up to 50% compared to tools that load the entire model at once. This is why it can run SD 1.5 on 8GB when ComfyUI can't.

Memory	What works	What doesn't
8GB	SD 1.5 (with 8-bit models), small Flux distilled	SDXL, Flux Schnell/Dev
16GB	SDXL comfortably, Flux Schnell	Flux Dev (loads but swaps)
24GB	Everything including Flux Dev	Flux Dev + large LoRA stacks
36GB+	Everything, comfortable batching	Nothing off-limits

Setup

1. Install from the Mac App Store (free)
2. Open the app, go to the model browser
3. Download SD 1.5 or SDXL (the app suggests compatible models for your hardware)
4. Type a prompt, adjust settings if you want, click Generate

Time from zero to first image: about 5 minutes, most of which is downloading the model.

MLX stable diffusion: fastest native performance

What it is

Apple's MLX framework includes a stable diffusion implementation that runs natively on Apple Silicon's unified memory. It's a Python library, not a GUI app. You write code or run command-line scripts to generate images.

MLX is faster than PyTorch + MPS (Metal Performance Shaders) for the same model because it targets Apple Silicon's unified memory directly instead of going through a generic GPU abstraction. The tradeoff: fewer models supported, and you need to be comfortable with Python.

What it supports

- SD 2.1 and SDXL Turbo (officially)
- img2img
- Quantization (4-bit text encoders, 8-bit UNet) for reduced memory
- Batch generation

The model support is narrower than Draw Things or ComfyUI. SD 1.5 isn't in the official examples (though community forks exist). Flux isn't supported yet. If you need the latest models, MLX isn't the right choice.

When to use it

MLX stable diffusion makes sense if you're:

- Building a Python pipeline that generates images as part of a larger workflow
- Batch-generating hundreds of images and want maximum speed
- Writing scripts that need programmatic control over every parameter
- Comfortable with Python and command-line tools

It doesn't make sense if you want to browse models, experiment with prompts visually, or need Flux/ControlNet/LoRA support.

Usage

```
pip install -r requirements.txt

# Text to image (SDXL Turbo, 4 images)
python txt2image.py "A photo of an astronaut riding a horse on Mars" --n_images 4 --n_rows 2

# Image to image
python image2image.py --strength 0.5 original.png "A lit fireplace"

# Quantized (for 8GB Macs)
python txt2image.py --n_images 4 -q "prompt here"
```

The `-q` flag quantizes text encoders to 4-bit and UNet to 8-bit. This lets SDXL Turbo run on an 8GB Mac Mini without swapping. Without quantization, you need at least 16GB.

ComfyUI on Mac: most flexible, slowest

What it is

ComfyUI is a node-based workflow editor for image generation. You can build things with it that the other tools can't touch: multi-model pipelines, custom samplers, chained refiners, community workflow imports. It's also the slowest option on Mac and the most annoying to install.

The Mac situation

ComfyUI uses PyTorch with Metal Performance Shaders (MPS) for GPU acceleration on Mac. MPS works, but it's 2-4x slower than NVIDIA CUDA for the same operation. Every ComfyUI benchmark you see online with impressive 2-second SDXL generation times? That's an RTX 4090. On Mac, expect 3-5x those numbers.

The ComfyUI Desktop app (beta) supports Apple Silicon, but some features are known to not work. The manual installation route is more reliable.

Installation

```
# Install Python 3.11+ and git via Homebrew
brew install python@3.11 git

# Clone ComfyUI
git clone https://github.com/comfyanonymous/ComfyUI.git
cd ComfyUI

# Create venv and install
python3.11 -m venv venv
source venv/bin/activate
pip install -r requirements.txt

# Run with fp16 (important for Mac speed)
python main.py --force-fp16
```

The `--force-fp16` flag matters. Without it, ComfyUI defaults to fp32 on Mac, which is roughly half the speed and uses twice the memory. I've seen people complain about ComfyUI being unusable on Mac, and half the time it's because they missed this flag.

Speed comparison

On the same M2 Pro 16GB, generating the same image:

Model	Draw Things	ComfyUI (–force-fp16)	MLX
SD 1.5 512x512 (20 steps)	~10s	~30s	~8s
SDXL 1024x1024 (20 steps)	~30s	~90-110s	N/A (limited support)
Flux Schnell 1024x1024 (4 steps)	~35s	~60-80s	N/A

ComfyUI is 3x slower than Draw Things for the same model. That's the Metal FlashAttention vs MPS difference. ComfyUI's PyTorch MPS backend is generic GPU acceleration. Draw Things has hand-tuned Metal shaders for each operation.

When ComfyUI is still worth it

- You need node-based workflows with custom pipelines
- You're importing workflows from the community (ComfyUI has the largest workflow ecosystem)
- You want to chain models: base + refiner + upscaler in one pipeline
- You're already using ComfyUI on another machine and want the same workflow on Mac
- You need custom nodes that only exist in the ComfyUI ecosystem

If none of those apply, use Draw Things. It's faster for every common task.

Known Mac issues

- MPS doesn't support all PyTorch operations. Some custom nodes will fail with cryptic errors.
- Memory reporting in ComfyUI assumes discrete GPU VRAM, not unified memory. The numbers shown in the UI aren't accurate on Mac.
- Some samplers are slower on MPS than others. Euler and DPM++ 2M work well. DDIM can be buggy.
- Flux models need careful memory management on 16GB. Close everything else, use `--force-fp16`, and consider GGUF quantized Flux models to fit.

What to run at each memory tier

Memory	Best tool	Best model	What to expect
8GB	Draw Things	SD 1.5 (8-bit)	Usable for casual generation, 20-30s per image
16GB	Draw Things	SDXL or Flux Schnell	Good quality, 15-40s per image depending on model
24GB	Draw Things	Flux Dev	Everything works, 15-25s for Flux
32GB+	Draw Things or ComfyUI	Flux Dev + LoRAs	Fast enough for iteration, ComfyUI viable for complex workflows

On 8GB, skip SDXL and Flux entirely. They technically load in some tools but swap to disk and generation times balloon to minutes per image. SD 1.5 with Draw Things's 8-bit models is the 8GB sweet spot.

On 16GB, you have real choices. SDXL runs comfortably in Draw Things and produces much better images than SD 1.5. Flux Schnell works too – it's a 4-step model designed for speed, so even at 30-50 seconds per image on Mac, you get fast iterations.

At 32GB+, you're no longer memory-constrained. The speed gap between Mac and NVIDIA still exists, but you can run any model and use ComfyUI for complex workflows without worrying about crashes.

The honest PC comparison

I'd be leaving something out if I didn't say this: if image generation is your primary use case and you don't already own a Mac, a PC with an RTX 3060 12GB ($170 used) will generate images 3-5x faster than an M2 Pro.

Setup	Cost	SD 1.5 512x512	SDXL 1024x1024
M2 Pro 16GB Mac	Already own	~10s	~30s
M4 Max 36GB Mac	$3,000+	~4s	~10s
RTX 3060 12GB PC	~$170 (used GPU)	~3s	~8s
RTX 4090 PC	~$1,600	~1s	~2s

A $170 used GPU matches a $3,000 Mac. CUDA is that much faster for diffusion models.

But if you already have a Mac and don't want a second machine sitting under your desk, Draw Things makes the speed difference tolerable. 10 seconds per SD 1.5 image is fine for iterating on prompts. And newer models like Flux Schnell only need 4 steps, so the per-step speed penalty matters less.

The bottom line

Start with Draw Things. It's free, fast, and takes 5 minutes from install to first image. If you find yourself wanting node-based workflows or custom pipelines, add ComfyUI. If you're writing Python scripts that need image generation, look at MLX.

Most Mac users will never need anything beyond Draw Things. It handles SD 1.5, SDXL, Flux, LoRAs, ControlNet, inpainting, and LoRA training – all through a clean native interface with hand-tuned Metal shaders that actually use your hardware well.

Related guides

- Stable Diffusion Locally: Getting Started
- Flux Locally: Complete Guide
- ComfyUI vs A1111 vs Fooocus
- Best Local LLMs for Mac in 2026
- Running LLMs on Mac M-Series
- AI Art Styles and Workflows

Get notified when we publish new guides.

Subscribe — free, no spam
