RESEARCH · APPLE METAL · BEFORE AND AFTER

Before and After: Every Model, Original Against Exe AI on Apple Silicon

One whole run per bar, from the prompt to the saved file · unmodified stable-diffusion.cpp against the newest Exe AI build that ran the same command

Exe — Exe AI, developer studio

Measured on Apple M2 Max (38-core GPU, 64 GB unified memory)

7.7×
LTX-2.3 1280×704×121, whole run: 2,715.0 s → 350.8 s

111×
LTX-2.3 1280×704×121, video decode alone: 2,278.0 s → 20.6 s

2.2×
median speed-up over 19 whole runs of 10 model families

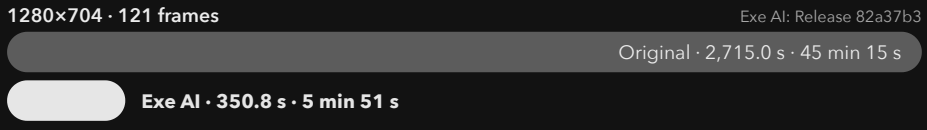
−54 %
all 19 runs together: 3 h 32 min → 1 h 38 min

1 Video models

LTX-2.3 22B

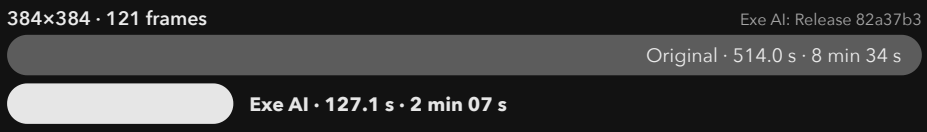
Lightricks · distilled, Q8_0 · 4 steps · video with audio

1280×704 · 121 frames



7.7× faster · −87 % time

384×384 · 121 frames

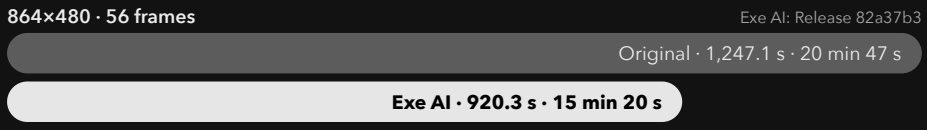


4.0× faster · −75 % time

MiniMax H3

MiniMax · Q8_0 · 20 steps · video with audio

864×480 · 56 frames



1.4× faster · −26 % time

384×384 · 56 frames



1.3× faster · −25 % time

Not yet run in the night-run format; their per-part load is in Report 5, Section 11.3:

Wan 2.2 TI2V 5B
Wan-AI · 1280×704 · 33 frames

Wan 2.1 T2V 1.3B
Wan-AI · 832×480 · 33 frames

HunyuanVideo 1.5
Tencent · 848×480 · 33 frames

LingBot-Video 1.3B
Robbyant · 832×480 · 33 frames

2 Image models

Qwen-Image 2.1

Qwen · F16 · 30 steps · CFG 3

1024²

Exe AI: Release 82a37b3

Original · 1,035.4 s · 17 min 15 s

Exe AI · 479.7 s · 8 min 00 s

2.2× faster · −54 % time

512²

Exe AI: Release 82a37b3

Original · 288.0 s · 4 min 48 s

Exe AI · 159.5 s · 2 min 40 s

1.8× faster · −45 % time

Krea 2 Raw

Krea · F16 · 52 steps · CFG 3.5

1024²

Exe AI: Release 82a37b3

Original · 1,864.9 s · 31 min 05 s

Exe AI · 1,258.6 s · 20 min 59 s

1.5× faster · −33 % time

512²

Exe AI: Release 82a37b3

Original · 489.1 s · 8 min 09 s

Exe AI · 339.2 s · 5 min 39 s

1.4× faster · −31 % time

FLUX.3 Klein Base

Black Forest Labs · Q8_0 · 20 steps · CFG 3.5

1024²

Exe AI: Laboratory · ggml de043708

Original · 465.0 s · 7 min 45 s

Exe AI · 171.0 s · 2 min 51 s

2.7× faster · −63 % time

512²

Exe AI: Laboratory · ggml de043708

Original · 115.5 s · 1 min 55 s

Exe AI · 43.7 s

2.6× faster · −62 % time

Iris 3B

Speridlabs · F16 · 100 steps · CFG 3 · pixel space

1024²

Exe AI: Laboratory · ggml 87d6c0aa

Original · 1,360.1 s · 22 min 40 s

Exe AI · 505.4 s · 8 min 25 s

2.7× faster · −63 % time

512²

Exe AI: Laboratory · ggml 87d6c0aa

Original · 427.2 s · 7 min 07 s

Exe AI · 150.8 s · 2 min 31 s

2.8× faster · −65 % time

MOSS-Image Base

OpenMOSS · F16 · 50 steps · CFG 3

1024²

Exe AI: Laboratory · ggml 87d6c0aa

Original · 741.9 s · 12 min 22 s

Exe AI · 492.8 s · 8 min 13 s

1.5×
faster · −34 % time

512²

Exe AI: Laboratory · ggml 87d6c0aa

Original · 191.6 s · 3 min 12 s

Exe AI · 113.2 s · 1 min 53 s

1.7×
faster · −41 % time

Ming-Image Design

inclusionAI · F16 · 12 steps

1024²

Exe AI: Laboratory · ggml 87d6c0aa

Original · 341.7 s · 5 min 42 s

Exe AI · 166.3 s · 2 min 46 s

2.1×
faster · −51 % time

512²

Exe AI: Laboratory · ggml 87d6c0aa

Original · 167.2 s · 2 min 47 s

Exe AI · 121.6 s · 2 min 02 s

1.4×
faster · −27 % time

SDXL · CyberRealistic XL

Stability AI · F16 · 20 steps · CFG 5

1024²

Exe AI: Laboratory · ggml b100353a

Original · 112.5 s · 1 min 52 s

Exe AI · 37.8 s

3.0×
faster · −66 % time

512²

Exe AI: Laboratory · ggml b100353a

Original · 41.7 s

Exe AI · 11.7 s

3.6×
faster · −72 % time

SD 1.5 · Realistic Vision 6

Stability AI · F16 · 20 steps · CFG 7.5

512²

Exe AI: Laboratory · ggml b100353a

Original · 28.5 s

Exe AI · 7.0 s

4.1×
faster · −76 % time

3 How the bars were measured

Machine	Apple M2 Max, 38-core GPU, 64 GB · macOS 26.5.2 (25F84) · models on the external USB SSD
Command	`nightrun/run_matrix.py`: the same `sd-cli` command for both builds, seed 42, one prompt (`measurements/prompt.txt`), `--flash-attn on`; steps and guidance as on each card
Time	wall time of the whole run, model loading and file writing included (`results.tsv`, column `wall_s`) · one run per bar, each alone on the GPU
Original	stable-diffusion.cpp `master-1000-95ff762`, unmodified, as shipped in Exe AI
Exe AI	the newest build with a whole run of that row, named above each bar: Release `master-1000-82a37b3` (ggml `34e0dc1c`) or the laboratory build `engine/build-edge` with the ggml commit of its log

- LTX-2.3: of the 2,715.0 s of the original, 2,278.0 s are the video decode; Exe AI decodes in 20.6 s, sampling 384.6 s → 282.7 s.
- SDXL and SD 1.5: the original writes a white picture (batched CFG on ggml's GROUP_NORM, Report 4, Section 2); its bar is the time of that white picture. The Exe AI laboratory picture is correct.
- Ming-Image: its text encoder (36 GB F16) reads ≈ 70 s from disk in every run, which keeps the 512² speed-up at 1.4×.
- Not in the bars: the per-step gains of Report 5 for MiniMax H3 (−13 %) and LTX-2.3 (−10 %); the current release `07f8c72` has no whole runs yet.
- Maker marks: the original vendor vectors as used in the Exe AI app, unchanged; Ming-Image carries a letter, as the app holds no mark for inclusionAI. All marks belong to their owners.
- Logs: `measurements/nightrun-…/<family>_<size>_<resolution>_{orig,ours,lab}.log` · bars drawn by `paper/build/build_paper6.py` from `results.tsv`.

EXE AI · DEVELOPER STUDIO