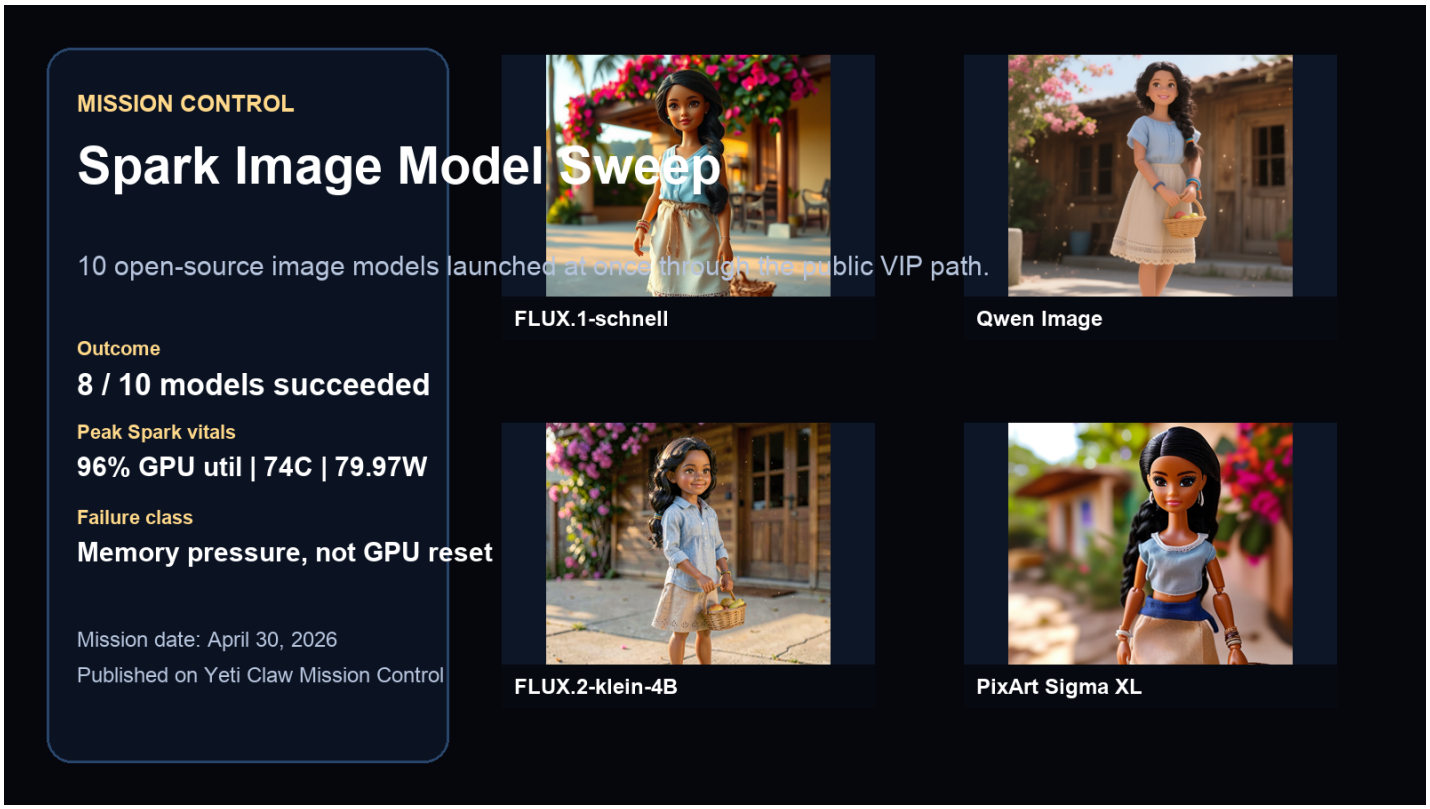


Committee Review | Spark Image Model Sweep

Mission Control committee report for the 10-model concurrent open-source image generation sweep run through the public VIP path on April 30, 2026.

Run window. Started 2026-05-01 03:09:12 UTC | **Path.** chat.neonflux.co via 192.168.12.163 | **Prompt class.** realism / character generation.

Executive summary

Outcome. Eight of ten models completed successfully during a full 10-way concurrent burst. The two failures were operationally distinct: **Qwen Image Edit** failed because the request lacked a required source image, while **HunyuanDiT v1.2** failed inside a memory-pressure window. Spark stayed online, stayed reachable, and did not emit off-the-bus or full-chip reset signatures during the run.

- Fastest successful render: FLUX.1-schnell in 217.61 seconds.
- Slowest successful render: FLUX.2-klein-base-4B in 834.71 seconds.
- Peak Spark envelope: 96% GPU utilization, 74C GPU temperature, 79.97W power draw.
- No GPU off-the-bus, full-chip reset, or reset-required signature was observed in the benchmark summary window.

Test configuration

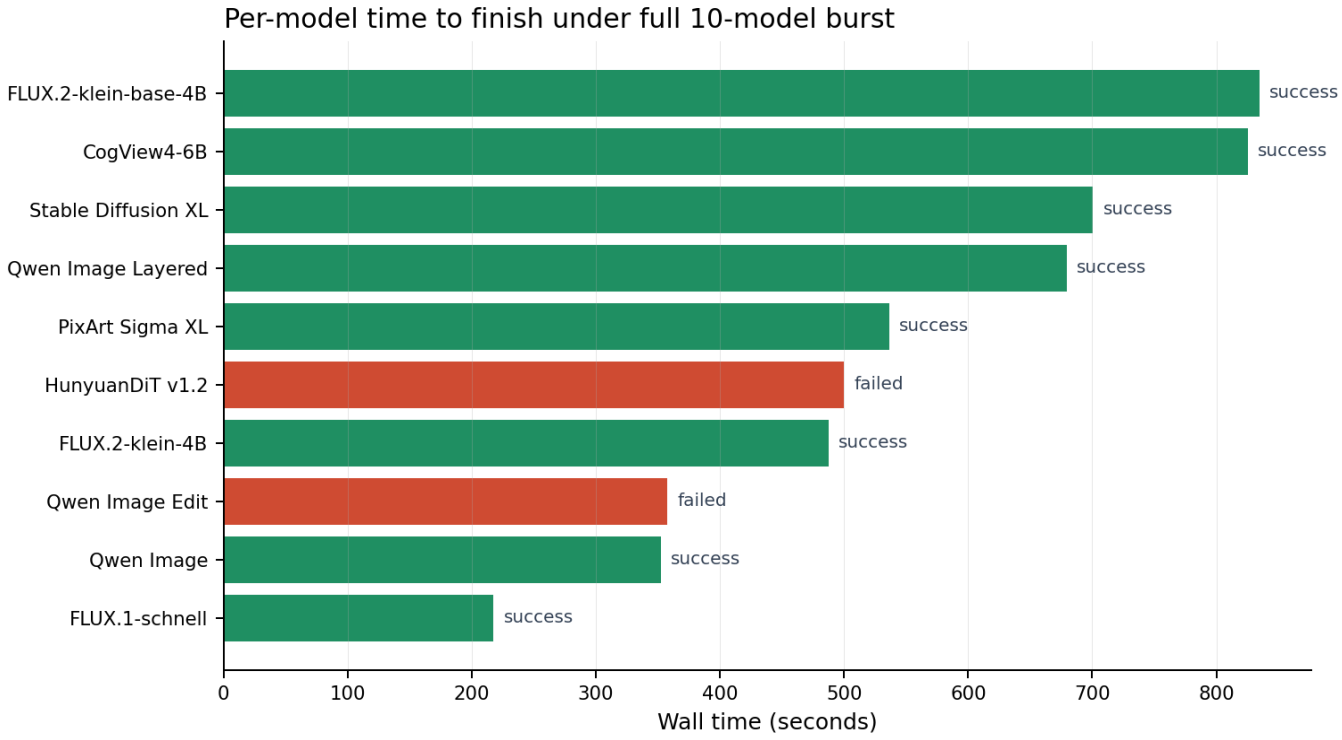
Field	Value
Path tested	http://192.168.12.163/api/chat with Host: chat.neonflux.co
Launch style	10 image models fired at once
Prompt	Salvadoran realism prompt

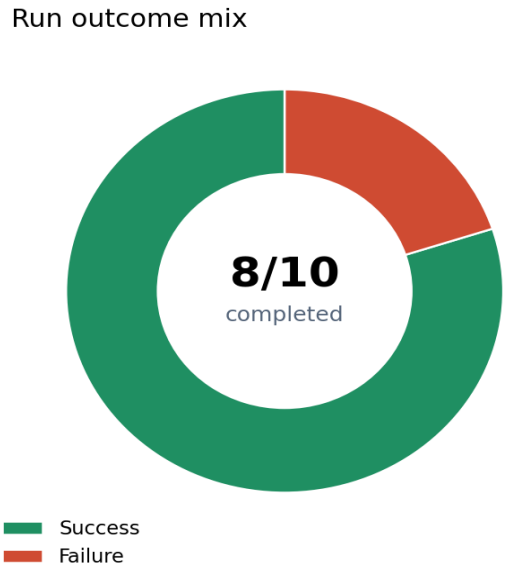
Spark samples	176
HTTP outcome mix	8 x 200 OK, 2 x 502 Bad Gateway

Model	Status	Wall s	API ms	HTTP	Note
FLUX.1-schnell	success	217.61	217512	200	Fastest successful render.
Qwen Image	success	352.14	352052	200	Base Qwen image model completed cleanly.
Qwen Image Edit	failed	357.45	n/a	502	Requires source image upload.
FLUX.2-klein-4B	success	487.59	487464	200	Completed cleanly.
HunyuanDiT v1.2	failed	499.98	n/a	502	Failure window aligned with Spark GPU memory pressure.
PixArt Sigma XL	success	536.57	536458	200	Completed cleanly.
Qwen Image Layered	success	679.47	679373	200	Completed cleanly.
Stable Diffusion XL	success	700.62	700492	200	Completed cleanly.
CogView4-6B	success	825.30	825199	200	Completed cleanly.
FLUX.2-klein-base-4B	success	834.71	834599	200	Slowest successful render.

Performance curves

The timing chart shows the true completion spread across the model set. The result is not a flat cluster: the fastest model finished in a little over three and a half minutes, while the slowest successful completion crossed thirteen minutes. That spread matters when deciding whether the public surface should allow free-form all-model bursts or should shape traffic into tiers.

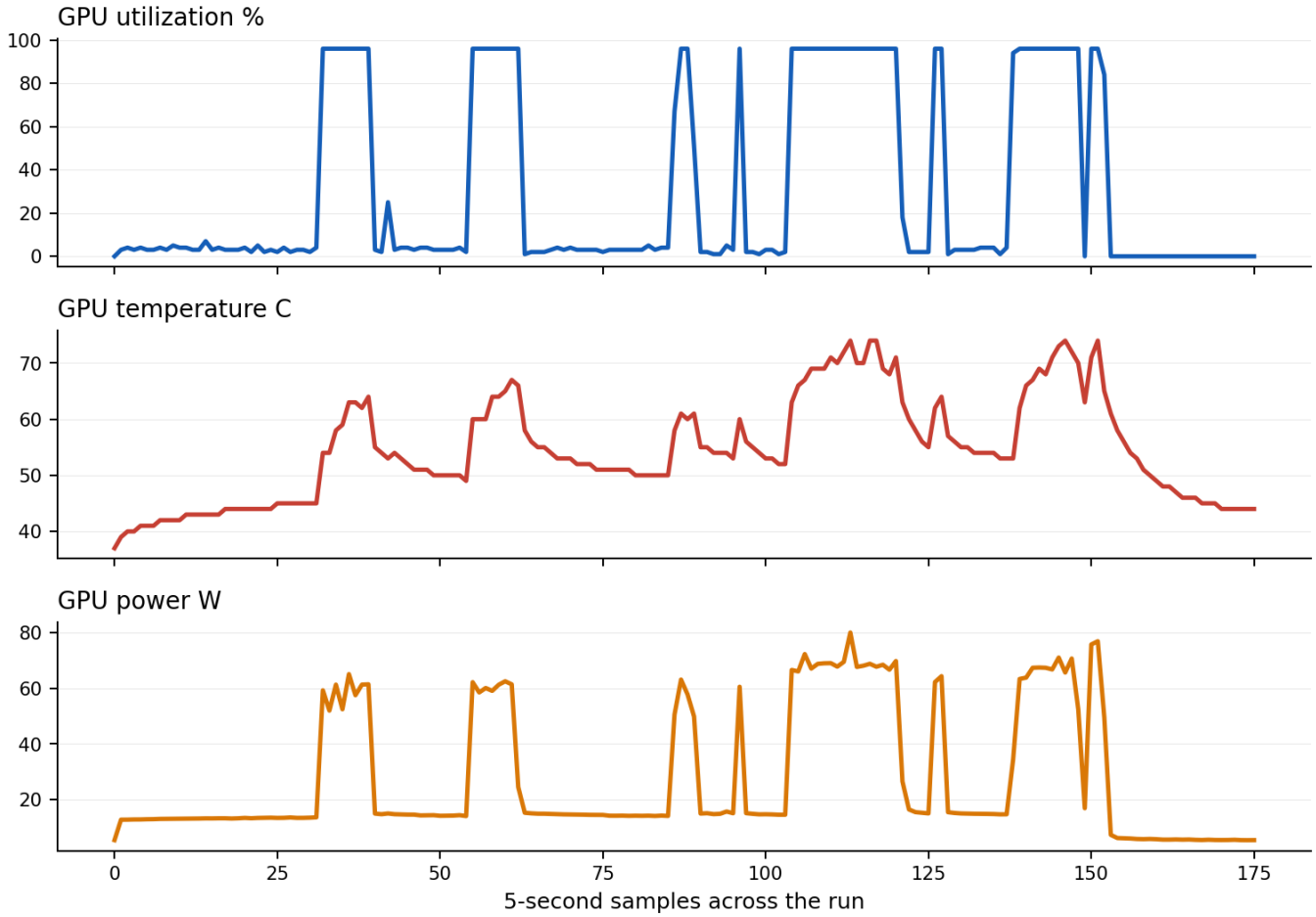




Environmental telemetry

Spark behaved like a loaded but healthy image factory. GPU utilization rose into the mid-90s, temperature peaked at 74C, and power climbed to nearly 80W before dropping back to idle once the queue drained. The important operational fact is that memory pressure showed up before thermal instability or transport loss.

Spark environmental envelope during the concurrent image sweep



Output gallery

The contact sheet below shows the eight successful outputs captured from the public path. These images were decoded directly from the response payloads saved during the run, not regenerated afterward.

Spark Image Load Test | Successful Out

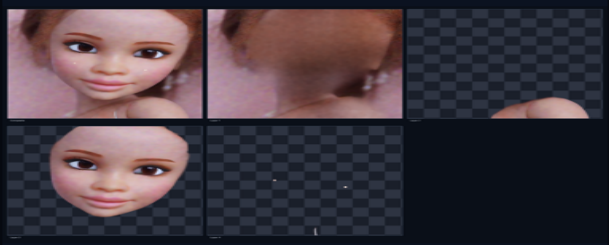
Eight models completed under a 10-model concurrent launch through the public VIP

FLUX.1-schnell
217.61s | HTTP 200

Qwen Image
352.14s | HTTP 200

FLUX.2-klein-4B
487.59s | HTTP 200

PixArt Sigma XL
536.57s | HTTP 200



Qwen Image Layered
679.47s | HTTP 200

Stable Diffusion XL
700.62s | HTTP 200

CogView4-6B
825.30s | HTTP 200

FLUX.2-klein-base-4B
834.71s | HTTP 200

Operational opinion

What the test proves. Spark can survive a 10-model image burst without transport collapse or GPU reset, but it does not mean every model mix is equally safe. The failing Hunyuan window is the signal to respect memory-heavy models as a separate tier.

- Good public-facing default lane: SDXL, PixArt, FLUX.2-klein-4B, and Qwen Image one at a time or in a shaped queue.
- Keep edit-style pipelines behind capability checks because they require an input image and fail noisily without one.
- Treat Hunyuan-class or similarly memory-hungry pipelines as premium jobs that should not be admitted blindly into a full free-for-all burst.
- The absence of bus-reset signatures is a strong positive result. The bottleneck here was memory pressure, not host instability.