

18 TESTS · 9 IMAGE MODELS / MAKING OF

Nine models, one brief.

A comparison of nine image models on Scenario. Every model got the same prompt and the same input, one image each at default settings, and every credit and second on the page comes from the job records.

9
image models

162
results on the page

18
tests

2,941
credits, every job



09 · The suitcase lighthouse, one image per model

One agent, one log.

Claude Opus 5.5 in Claude Code ran the whole comparison through the Scenario MCP. It chose the models, priced the run, sent each test's prompt and input to all nine models, waited for the jobs and logged every job id, asset id and cost. The grids, the notes and the page were built from that log. Four source images feed the tests that start from a picture.



Gouache style sheet · tests 01–03



Badger character · tests 05–08



Portrait · test 11



Courtyard · tests 16–18

The lineup

Nine current models that can both generate and edit, found with the MCP's recommend and search tools, read with `model_schema_get` and priced with a dry run before anything ran. Two models edit with a sister model: Meta Muse Image Edit and MAI Image 2.5 Edit. The planned ninth model, Ideogram 4.5, returned an access error for the team, so Grok Imagine Image 2.0 took its place.

The rules

One output per model, each model's default quality, no prompt rewriting. The same prompt and the same input went to all nine. Only failed jobs were rerun: two Grok jobs timed out on the server side and were retried successfully, and Scenario refunded the failures.

Every prompt has something to check.

Each of the 18 tests has one fixed prompt with details that can be verified by eye: exact text, counts, sides, layouts and materials. Game art comes first, then imagination, a portrait, a relight, the teacup test, a risograph, a painted sign, a watchmaker and three photo edits.



01· A new character in the sheet's gouache



12· Five cups, the third upside down



14· Exact spelling, with an em dash

Checkable, not generic

The teacup prompt asks for exactly five mismatched cups in one row, smallest to largest, the middle one upside down, a green chess knight in the largest, one paper crane on the left and a line of turmeric running corner to corner. Sunburst, Flare and Grok got every clause right. The sign must read “MUTTON DUMPLINGS — OPEN UNTIL THE WOLVES COME”; most models turned the em dash into a hyphen.

Consistency, not prompt-following

The character tests say “keep everything else identical to the reference” without listing the badger's monocle, earring, brass arm, scarf, apron or tools, so each model has to read them from the image. Results are judged against the reference, not the prompt.

Game art comes first.

Tests 01 to 08 cover the assets a game team asks for: a new character in a borrowed style, an isolated building, a nine-icon inventory set, a seamless texture, a new pose, a turnaround sheet, an expression sheet and App Store art. The notes below are a first visual read, not a scored evaluation.



04 · Frozen koi pond, a tile per model



07 · Six emotions, 3 columns × 2 rows



08 · Three lines of exact text

What held

All nine returned exactly nine icons in a 3×3 grid, and every model spelled “FORGE & LANTERN”, “PLAY FREE” and “Hammer. Glow. Repeat.” correctly. On the expression sheet, all nine returned six portraits in 3 columns and 2 rows once the layout was spelled out.

What broke

In the new pose, Gemini and FLUX dropped the monocle, MAI gave the badger two mechanical arms and Muse duplicated the scarf. FLUX's fourth turnaround figure was a second back view instead of a three-quarter view. Sunburst replaced the R in LANTERN with a lantern icon: clever, but a misspelled logo.

Then people, light and edits.

The general tests start from text or from a photograph. The edits keep the source image and change one thing: the light, one added object, two removed objects or the whole medium. They show how much of the source each model keeps.



11 · Daylight portrait relit for night

The relight

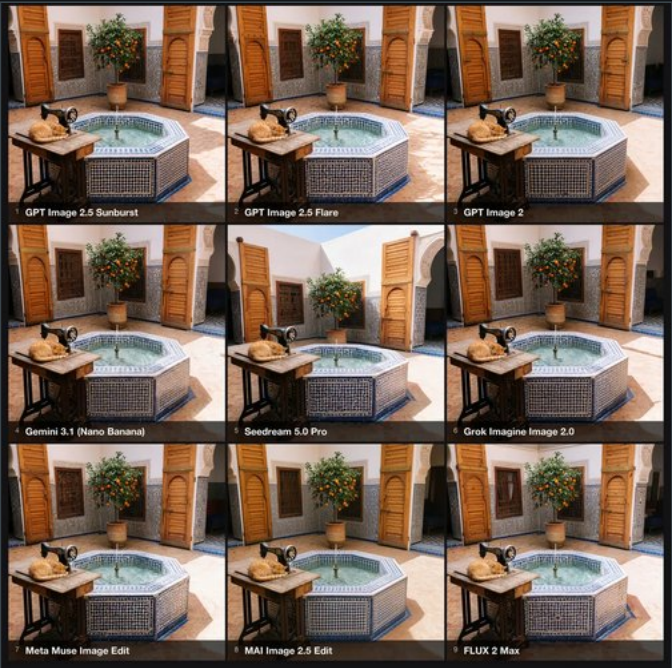
The prompt asks for night, with blue light from the left and pink from the right, and the same woman and pose. Every model kept her identity and pose. FLUX gave the most dramatic blue and pink split and the GPT Image 2.5 pair got the balance right. MAI left the window bright white, so it never really became night.



16 · A flamingo added, with its reflection

Add, remove, restyle

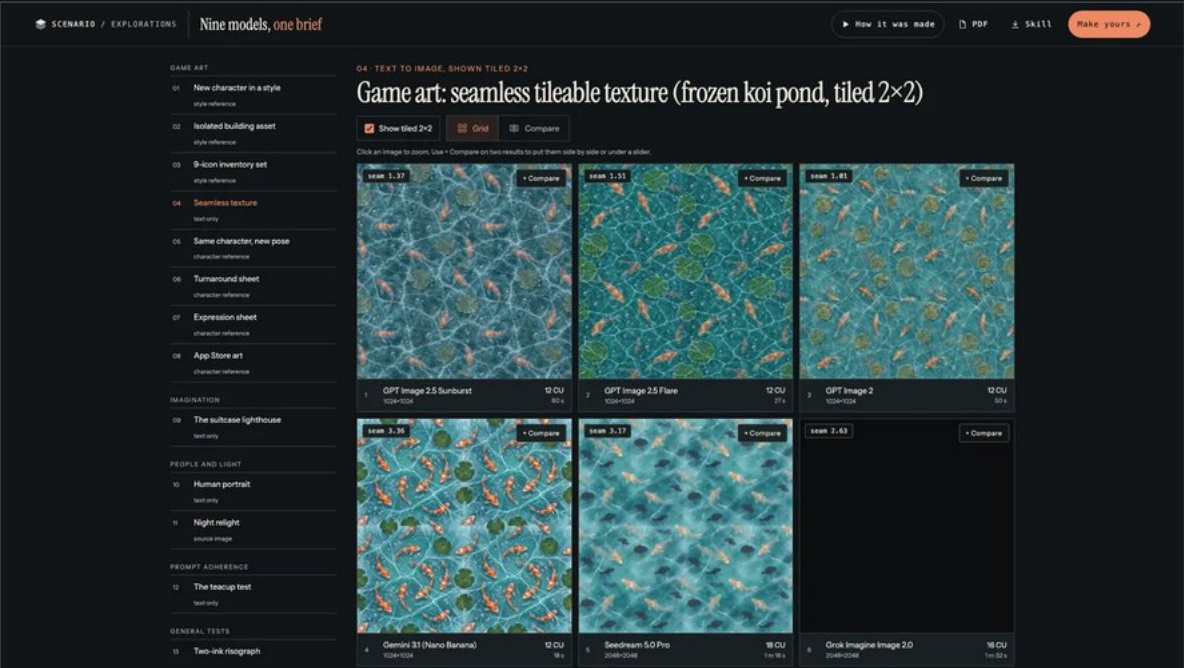
All nine added the flamingo; Grok drew its reflection on the fountain's wall, and FLUX's ripples were the most convincing. Every model removed the bicycle and the laundry lines, though Seedream re-framed the courtyard. All nine switched to a yellow duotone and kept the composition.



17 · The bicycle and laundry removed

Credits and seconds from the job records.

Credits are what Scenario charged for each job. Time runs from the moment Scenario started processing a job to the finished result, read from each job's status history with job_get, so time spent waiting in the agent's own batch queue is left out. The texture test gets one more number: a seam score.



04 on the page · tiles repeated 2×2, seam score on each

The seam score

seam_check.py divides the mean colour jump across the wrap-around edges by the mean jump between neighbouring pixels inside the tile. 1.0 means the wrap is as smooth as the texture itself. GPT Image 2 scored 1.01, effectively seamless with cracks and fish crossing the edges. FLUX 2 Max scored 4.58, with a square frame around every tile.

EVERY MODEL, ALL 18 TESTS

MODEL	CREDITS / IMAGE	MEDIAN TIME
GPT Image 2.5 Sunburst	12.1	34 s
GPT Image 2.5 Flare	12.1	20 s
GPT Image 2	12.1	32 s
Gemini 3.1 (Nano Banana)	12.6	23 s
Seedream 5.0 Pro	18.0	1 m 55 s
Grok Imagine Image 2.0	16.6	32 s
Meta Muse Image	2.0	24 s
MAI Image 2.5	11.2	32 s
FLUX 2 Max	18.7	24 s

Cheapest and slowest

Meta Muse Image cost 2 credits an image, FLUX 2 Max 18.7 on average. Seedream 5.0 Pro had the slowest median time, 1 m 55 s, and GPT Image 2.5 Flare the fastest, 20 s. The page's ledger sorts all nine by credits per image, credits over the 18 tests, median time and range.

Review, then redo what was weak.

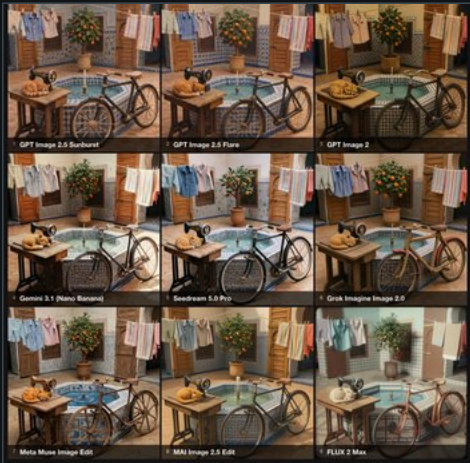
After the first pass, five tests were run again with a better brief, and their first attempts were archived instead of mixed into the page: 45 images in all. The texture became a frozen koi pond, the relight moved to a portrait, the restyle got a stronger style, and the two multi-panel sheets moved to wide canvases.



06 first · a square canvas squeezed four figures



06 redo · 16:9, exact order



18 first · stop-motion, too subtle



18 redo · yellow duotone

What did not work

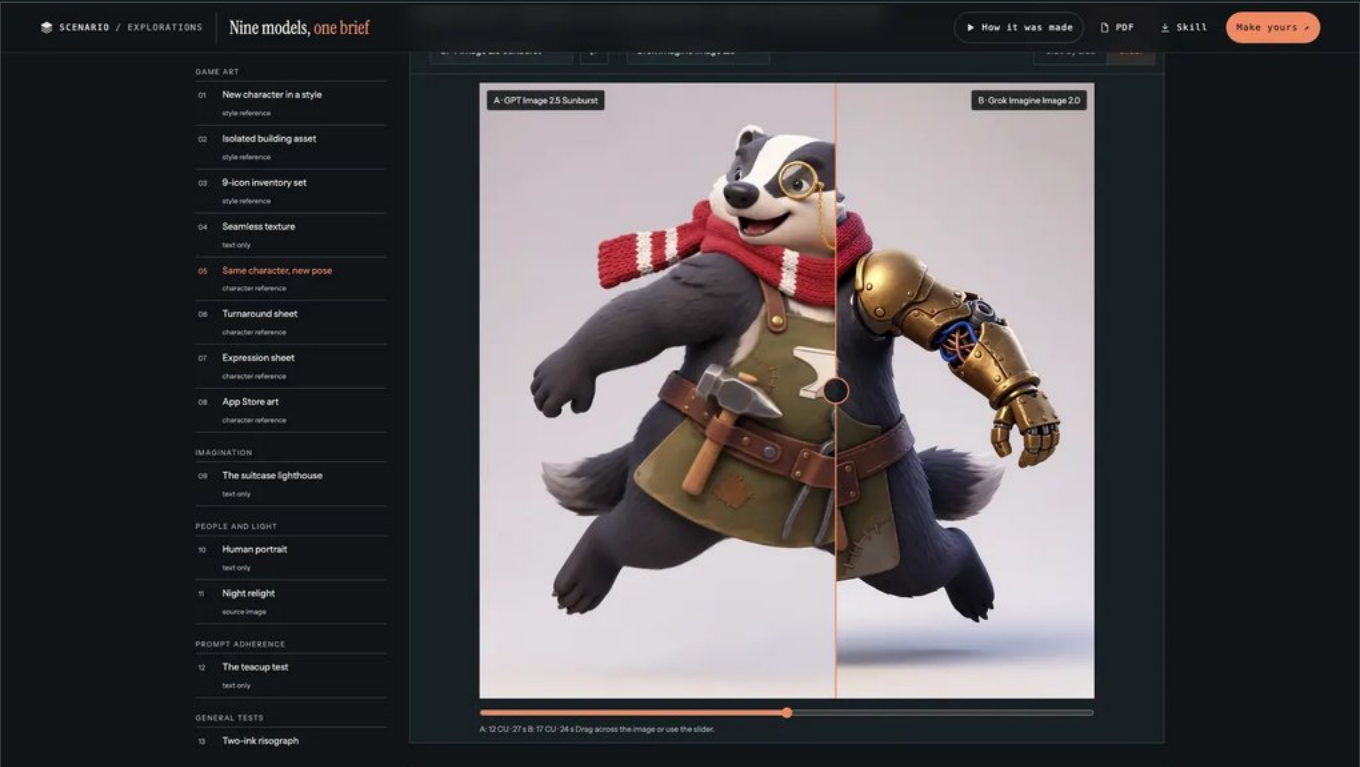
Square canvases squeezed the four-figure turnaround and gave the expression sheet different row counts, and the first grid script cropped Grok's wide 3552 × 1152 turnaround to a square. The stop-motion restyle was too subtle: three models barely changed the photo. Because Grok's edits keep the input's shape, every model got the same reference already placed on the 16:9 or 3:2 canvas.

The tip

For any multi-panel sheet, give the canvas shape and spell out the exact layout: rows, columns and order. With the wide canvas and the explicit layout, all nine models delivered four figures in one row with nothing squashed or cropped.

Judge it for yourself.

The page shows each test's 3×3 grid, its prompt, its source image and what we saw. Zoom into any result up to 100%, step through the nine with the arrow keys, or pick two and compare them side by side or under a slider. The ledger at the bottom totals credits and time for every model.



05 on the page · two results under the slider

How it was built

build_page.py reads config.json, the results table and each job's status history, writes web-sized WebP images and one data file, and the page renders every grid, prompt and ledger row from it. The narrated film in How it was made uses the same job records.

The receipt

162 results on the page, 18 tests × 9 models	2,076 CU
45 archived first attempts, 5 redone tests	593 CU
17 narration clips for the film	272 CU
Total, 224 finished Scenario jobs	2,941 CU

From the production's Scenario job records: the results table, the two archived results tables and the film's narration records. The two Grok jobs that timed out were refunded and are not counted.