Run a *model* shoot-out for your studio

Pick the model for a recurring job with a fair, repeatable shoot-out: same input, same prompt, a scorecard, and a recorded winner and runner-up.

PLAYBOOK

IN SHORT

- A dated scorecard that names your team's model for one recurring job, with a runner-up, cost and wait.
- Change only the model: same prompt, same reference, same aspect ratio, each at its highest quality setting.
- Four runs, one per model, then a second seed for the top two; started in parallel, they finish in the time of the slowest.

You need ✓ Any plan with credits ✓ One real brief from your team's work ✓ The reference images that brief uses

SEE IT MADE

▷ Nine Models, One Brief

A scorecard that names the model your team uses for one recurring job, with a runner-up, the cost and wait of each, and the inputs to rerun it when a new model ships. For team leads and leads of craft.

As of October 2026

The models, settings and screens in this article are the ones Scenario offered in October 2026. New models and updates arrive often, so check the app for newer options before you choose.



Same prompt, same reference. Top left: GPT Image 2.5 Sunburst. Top right: GPT Image 2.5 Flare.
Bottom left: Nano Banana 2.1. Bottom right: Seedream 5.0 Pro.

The brief

Field	Brief
Job	Product in a location scene from a studio reference
Input	The studio launch image of the Aurelle Solenne, an invented electric grand tourer in liquid satin champagne
Test	The car on a gravel overlook on a mountain pass at sunrise, three-quarter front, 3:2
Candidates	Four image models that take a reference, picked from the platform ranking for image to image at quality priority
Fixed	Prompt, reference, aspect ratio, one output each, quality high
Score	Wordmark, light bar, wheels and calipers, paint vs reference, light and mood vs brief, artifacts, price per accepted image, wait
Decide	Winner and runner-up, recorded with the date

The job tested is the one the Product lifestyle scenes playbook does all day: put a branded product from a reference into a new scene.



The reference: the approved studio image.

Approach

1	2	3	4	5
Shortlist Platform ranking, image to image, quality	**Test prompt** Product keep-list plus test scene	**Run four models** Sunburst, Flare, Nano Banana 2.1, Seedream 5.0 Pro	**Line up** Same-size grid and front crops	**Score and record** Winner, runner-up, cost, wait, date

A fair shoot-out changes one thing: the model (Compare models on your brief). The prompt stays word for word, with the same reference, the same aspect ratio and the highest quality setting each model offers.

Use a real brief from your team: a model that wins on a generic landscape can lose on your paint color or your wordmark. Shortlist with data, then decide with eyes:

1. **Shortlist** three or four candidates from the capability ranking. In this run the platform's recommendation for image to image at quality priority ranked GPT Image 2.5 Sunburst, GPT Image 2.5 Flare, GPT Image 2, Meta Muse Image Edit and Seedream 5.0 Pro; GPT Image 2 was left out as the older sibling of the first two, and Nano Banana 2.1 was added as a strong model from another family. Use **Choose a Model** with the **Image to Image** capability and **Quality: Highest First**, or ask Scenario Genie (Choose a model).

2. **Run** each once with identical inputs. One take per model is enough to eliminate; run a second seed for the top two when they are close.
3. **Score** on the criteria the brief cares about, in order. A wordmark error eliminates regardless of beauty.
4. **Record** the winner and runner-up with cost, wait and date, and keep the inputs to rerun later.

Steps

1 **Write the test prompt once.** It is the product keep-list plus the test scene:

Prompt

```
The exact car from the reference image, with its body shape and
proportions, liquid satin champagne paint, one continuous thin LED light
bar across the full width of the nose, the spaced AURELLE wordmark in
chrome capitals centered under the light bar, wide dark lower intake with
fine horizontal slats, dark twisted multi-spoke turbine wheels with gold
brake calipers, and black glass roof unchanged, parked on a gravel
overlook on a high mountain pass at sunrise. Snow-capped peaks and a
winding road behind, low warm sun from the left raking along the side of
the car, cold blue shadows, thin mist in the valley below. Three-quarter
front view at headlight height, the light bar lit and the AURELLE
wordmark facing the camera and sharp. Soft contact shadow under the tires
on the gravel. 3:2 landscape. No other cars, no people, no other text or
logos. Style: automotive launch photography, controlled reflections,
perfect panel gaps, no other brands.
```

2 **Run model 1. Create** > **Images**, **GPT Image 2.5 Sunburst**, the studio image in **Reference Images**, **Dimensions** 1536x1024, **Quality** `high`, **Image Count** 1.

3 **Run models 2 to 4** (Run a same-input shoot-out) with the same prompt and reference: **GPT Image 2.5 Flare** at the same settings; **Nano Banana 2.1** with aspect ratio `3:2` and resolution `2K`; **Seedream 5.0 Pro** at 2048x1360. Where a model offers only aspect presets, pick the matching ratio.

4 **Line them up.** Put the four results in one grid at the same size, and crop the front of each (light bar, wordmark, intake) at the same scale into one image.



Top left: Sunburst. Top right: Flare. Bottom left: Nano Banana 2.1. Bottom right: Seedream 5.0 Pro. All four spelled AURELLE and kept the light bar; the decision moves to paint and light.

5 **Score each criterion** and fill the scorecard below. Look at the reference while scoring the product.

6 **Run a second seed for the two closest.** Sunburst and Flare were close, so each ran once more with the same inputs. Both held the car again; the ranking did not change.

7 **Record the decision** in the team's model notes and set the winner as the default in the team App or workflow for this job.

With the Scenario MCP: call `recommend` with the capability (here image to image) and quality priority to get the shortlist, `model_schema_get` for each candidate's parameters, then four `model_run` calls with `wait: false` and one `jobs_wait`.

An agent can write the scorecard skeleton (cost, wait, output size) from the job results; the visual scores stay human.

Check the result

Check before you accept

Accept the comparison when: every model received the same prompt and reference; outputs are compared at the same size; the wordmark and the product were checked at 1:1 against the reference; cost and wait come from the actual jobs. **Reject the comparison when:** prompts were tuned per model; one model ran at a lower quality setting; the winner was picked from thumbnails; the brief was a showcase prompt instead of real work.

The scorecard from this run (1 = poor, 5 = excellent):

Criterion	Sunburst	Flare	Nano Banana 2.1	Seedream 5.0 Pro
Wordmark spelled correctly	yes	yes	yes	yes
Light bar, turbine wheels, gold calipers	held	held	held	held
Paint vs reference (liquid satin champagne)	5	5	3 (cooler, silvery)	3 (greyer)
Light and mood vs brief (sunrise, raking light, mist)	5	4	3 (flat light, car smaller in frame)	3 (cool, overcast light)
Artifacts	none seen	none seen	none seen	none seen
Price per accepted image, read from the quote	same as Flare and Nano Banana 2.1	same	same	highest of the four
Wait	about 40 s	about 27 s	about 35 s	about 126 s

- **Winner: GPT Image 2.5 Sunburst**, for the light that best matched the brief: low sun raking along the flank, mist in the valley, controlled reflections on the paint.
- **Runner-up: GPT Image 2.5 Flare**, which held the product on both seeds at about two thirds of the wait and is the better pick for drafts and volume.
- **Nano Banana 2.1** had the largest native output (2528 px wide at 2K) but turned the champagne cooler and silvery, which matters for a paint color.
- **Seedream 5.0 Pro** held the car's shape but read as an overcast morning rather than sunrise, and was the slowest and most expensive of the four.

One run per model only screens the candidates: the second seed for the top two is what made the winner a decision rather than a lucky take.



Winner: Sunburst. Champagne paint and raking sunrise light closest to the brief.



Nano Banana 2.1: wordmark right, paint turned cooler and silvery, light flat.



Sunburst, second seed: the product held again.



Flare, second seed: the product held again; close to Sunburst.

Check yourself

2 questions

1 **Same prompt and studio reference, two models. The brief asks for sunrise light raking along the car and mist in the valley, and the paint must match the reference. Both wordmarks read correctly. Which take wins?**



Take A



Take B

A Take A: the wordmark is correct and its model gave the largest native resolution

B **Take B: its paint and its raking sunrise light match the reference and the brief**

C Neither stands out: both wordmarks pass, so pick the faster model

Answer: B. Once the wordmark passes, the decision moves to paint and light. Take A's paint turned cooler and silvery under flat light, which matters for a car sold on its color.

2 **A colleague ran the shoot-out but rewrote the prompt for each model to get its best result. What do you do with the scorecard?**

- **A** **Reject the comparison and rerun with one prompt, word for word, for every model**
- B Accept it: each model was shown at its best, which is fairer
- C Accept it, as long as each model also ran at its highest quality setting

Answer: A. Prompts tuned per model are in Reject the comparison when. A fair shoot-out changes one thing, the model.

LEVEL 3 · A SYSTEM

Make it repeatable

- **Keep a test pack per job** (Keep sources as data). The prompt, the reference images, the settings and the scorecard template in one collection. Rerunning the shoot-out when a model ships is then a 10 minute task. See Keep up with new models.
- **Build the shoot-out as a workflow.** In the **Node Editor**: an **Image** input (`Reference`), a **Text** input (`Test prompt`), and four **Image Generator** nodes, one per candidate, all fed by the same two inputs. Duplicate it per job; change only the models.
- **Write the decision where people work.** Put "Product scenes: GPT Image 2.5 Sunburst; drafts: Flare" in the **Project Instructions** so Scenario Agents follow it, and set it as the model in the team's App. See Instructions and memory.
- **Restrict when needed.** If the decision must be enforced, admins can limit the models available to the organization. See Control model access.
- **Date every decision.** Models change monthly; a scorecard without a date is an opinion.

Runs and time

Step	Model	Runs or takes	Kept	Wait
Candidate 1	GPT Image 2.5 Sunburst, high	1	1 (winner)	about 40 s
Candidate 2	GPT Image 2.5 Flare, high	1	1 (runner-up)	about 27 s
Candidate 3	Nano Banana 2.1, 2K	1	0	about 35 s
Candidate 4	Seedream 5.0 Pro, 2048x1360	1	0	about 126 s

Step	Model	Runs or takes	Kept	Wait
Second seed, top two	Sunburst and Flare, high	2	2	about 40 s and 27 s
Grid, crops, scorecard	local	1	1	about 10 min of review

Started in parallel, the four candidates finish in the time of the slowest one. Each candidate costs one run; read each price on the **Generate** button, or from a dry-run quote through the Scenario MCP, and copy it into the scorecard.

Variations

- **Video:** same method with a fixed start frame and motion prompt; score drift, motion direction and artifacts frame by frame.
- **3D:** fixed input image; score topology, texture and silhouette in a viewer, plus export compatibility.
- **Audio:** fixed script or music brief; score with a transcript, a listen on headphones and the duration.
- **Upscalers:** fixed low-res input; score text, faces and invented detail at 1:1.
- **Cost-first teams:** add a "cost per accepted image" row: cost per run divided by the share of runs you would keep.

If it breaks

- **Every model looks good.** The test is too easy. Use the hardest real brief: small text, hands, a product with a precise color or fine detail.
- **The results flip on a second run.** The models are close on this job. Pick on cost or speed, and note both as acceptable.
- **A model needs different settings to work.** Give each its own best quality setting, but never a different prompt; record the settings in the scorecard.
- **The team ignores the decision.** Put the winner in the App, the workflow and the project instructions, not only in a document.
- **A new model appears the week after.** Rerun the test pack with the new model against the current winner only.

For Team lead, Artist Page 25 of 26 in Use-case playbooks

Verified in Scenario on Oct 16, 2026.