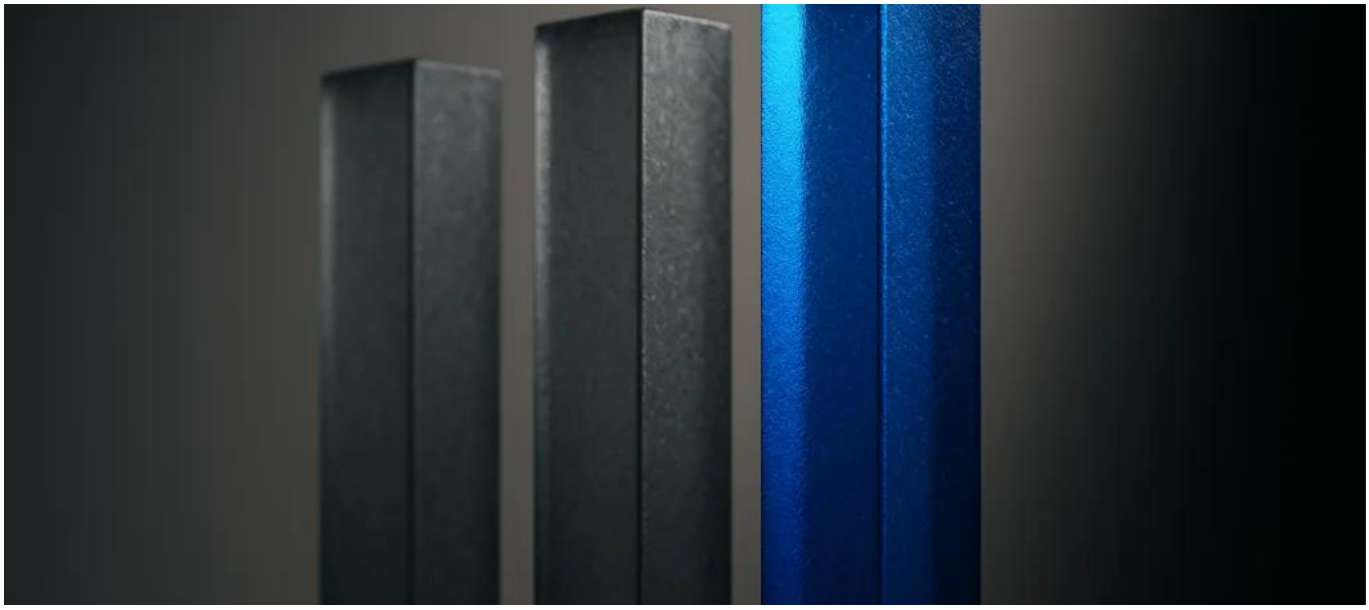

GPT Image 2.5 vs FLUX.2 vs Nano Banana vs Qwen-Image: the honest image model comparison for CTOs

Why run an open-weight image model when the top closed model makes a 1024×1024 image for $0.0059? Most “open vs closed” image-stack arguments I see right now have almost nothing to do with quality scores.

Start with the numbers, because they are unusually clean this time. On the Artificial Analysis text-to-image leaderboard updated 24 September 2026, GPT Image 2.5 Sunburst (max) sits at 1,196 Elo. The best Apache 2.0 open-weight model, FLUX.2 [klein] 9B, scores 940 on the open-weights board. That is a 256-point gap, and the leader’s cheapest quality tier comes to roughly $0.0059 per 1024×1024 image, less than half the $0.014 Black Forest Labs charges for its smallest hosted klein model.

So the cheap option is also the good option, at least on the API. The reasons to look at anything else are licensing, control, and a December deadline.

ATTRIBUTE	GPT IMAGE 2.5	FLUX.2	GEMINI 2.5 FLASH IMAGE	QWEN-IMAGE
Best Elo (text-to-image)	**1,196 (Sunburst max)**	1,025 ([flex])	n/a	886 (open-weights board)
Cost per 1024×1024 image	**$0.0059 low / $0.0132 med / $0.0527 high**	$0.014 (klein 4B) to $0.07 (max)	~$0.039 standard, ~$0.0195 batch	Self-hosted: infra only
Pricing model	Token-based ($30/M image output)	Per image, billed by megapixel	Per image	n/a
Weights available	No	[dev] and [klein] yes, [pro]/[max] no	No	**Yes**
License for commercial use	API terms	klein: Apache 2.0. [dev]: non-commercial, separate BFL license required	API terms	**Apache 2.0**
Tops editing leaderboard	**Yes, 1,181 Elo (Sunburst max)**	n/a	n/a	n/a

A 256-point Elo gap is a weak prior about your workload

Elo on a human-preference board measures which image people pick when shown two. It does not measure whether the model follows a brand style guide, whether it renders your product's specific form factor, or whether it puts legible German text on a packshot. I would treat 1,196 versus 940 as a strong signal about average output quality and a weak one about your specific workload.

Inside the closed tier, the top five are tightly bunched: GPT Image 2.5 Flare (max) at 1,190, GPT Image 2 (high) at 1,171, Grok Imagine Image 2.0 at 1,154, MAI-Image-2.6 at 1,147. Six points separate first from second, forty-nine separate first from fifth. Those margins are small enough that latency, rate limits, region availability and contract terms will decide more than the board does.

The open-weights side is bunched too, and more interestingly. HunyuanImage 3.0 leads the named set at 945 but ships under a non-commercial Tencent Hunyuan license, which

removes it from consideration for anyone shipping a product. Z-Image Turbo posts 942 under Apache 2.0, FLUX.2 [klein] 9B posts 940, also Apache 2.0, and Qwen-Image sits at 886, Apache 2.0. So the practical open-weight ceiling for commercial work is roughly 940 to 942, and the two models at that ceiling are within three points of each other.

A three-point Elo difference between Z-Image Turbo and FLUX.2 [klein] 9B is noise for decision purposes. My take: if you are choosing between them, choose on inference cost, VRAM footprint and ecosystem tooling. I would not defend a ranking claim built on that margin.

The license line runs through FLUX, not between vendors

The most common mistake in this category is treating "FLUX.2" as one thing. The FLUX.2-dev LICENSE.md, updated 9 September 2026, is the FLUX Non-Commercial License: non-commercial and non-production use only. Commercial use requires a separate license from Black Forest Labs, subject to fee, royalty or revenue share.

A team can prototype on [dev], get good results, and discover at production time that the thing they built on requires a negotiated contract with revenue share attached. [klein] under Apache 2.0 does not have that problem. Neither does Qwen-Image. **That distinction matters more than the 86-point Elo spread between klein 9B and Qwen-Image**, because one is a quality difference and the other is a legal cliff.

I wrote about this pattern more generally in the piece on open weights and what they actually grant you. Image models are where it bites hardest, because the output is the product.

Article 50(2) is agnostic about weights

The EU AI Act does not push you toward an open model. Article 50(2) requires that generative image outputs carry an effective, reliable, robust and interoperable machine-readable mark enabling detection as AI-generated. The European Commission's Article 50 transparency FAQ sets the general application date at 2 August 2026, with systems already on the EU market before that date getting until 2 December 2026 to comply with the marking obligation.

The compliance question is therefore about who is responsible for the mark and whether you can verify it survives your pipeline. A hosted API can embed a mark for you, and you inherit whatever it does, including the gap if the provider's implementation is weak or your downstream processing strips it. Self-hosted weights give you nothing for free: you own the marking obligation entirely, and you own the engineering to implement it.

The word "robust" in the text is the one to think about. My take: a mark that survives a re-encode and a crop is a different engineering problem than one that survives only a lossless copy, and the Commission's FAQ language points at the harder version. Unconfirmed: I have not seen a published conformance test that settles where the line sits, so anyone claiming their marking is definitively compliant is ahead of the evidence.

If you are working through the same deadline on the video side, the ranking of video models by what an EU studio can legally ship covers the same obligation with different vendors.

Token pricing will surprise your finance lead

GPT Image 2.5 is priced in tokens rather than images: $5 per million text input, $8 per million image input, $2 per million cached image input, and $30 per million image output. That last number drives everything. The per-image figures of $0.0059, $0.0132 and $0.0527 are derived from those rates for low, medium and high quality at 1024×1024.

Note what that implies. The spread between low and high quality on a single model is nearly 9x, larger than the 5x spread across the entire FLUX.2 hosted ladder, from $0.014 for klein 4B to $0.07 for max. Your quality-tier default matters more for the cost line than your vendor choice does.

Google's Gemini 2.5 Flash Image, the model everyone calls Nano Banana, is reported at roughly $0.039 per image standard and roughly $0.0195 in batch mode. The batch discount is 50%. For any workload that does not need synchronous responses that is a real lever, and it lands Nano Banana between GPT Image 2.5 medium and high on cost. I do not have an Artificial Analysis Elo figure for it in this data set, which is a genuine hole in the comparison: I cannot tell you where it ranks against the models I do have scores for.

One more catch with token pricing. Image input is billed at $8 per million and cached image

input at $2 per million, and editing or iterative refinement workflows feed images back in, so a per-image estimate built only on the $30 output rate will understate an editing pipeline's real cost.

Where each one earns its place

Pick GPT Image 2.5 if you are running a hosted pipeline and cost per image is the binding constraint. It tops the generation board at 1,196 and the separate editing board at 1,181 with Sunburst (max), and at low quality it is the cheapest per-image option here. Those are two different leaderboards, which is worth saying plainly, because generation rank does not automatically transfer to editing rank for any model.

Pick FLUX.2 [klein] 9B under Apache 2.0 if you need weights you can run and relicense freely, and you can accept output quality around 940 Elo. Pick FLUX.2 [pro] or [max] on the API if you want BFL's model family without the license negotiation, and accept $0.03 or $0.07 per image.

Pick Gemini 2.5 Flash Image if your workload batches well and $0.0195 per image against Google's infrastructure suits your existing cloud contracts.

Pick Qwen-Image if Apache 2.0 weights are non-negotiable and 886 Elo clears your bar. It is the lowest-scoring option here, and it is fully commercially usable with no vendor in the loop. That combination has a real audience.

Do not pick FLUX.2 [dev] or HunyuanImage 3.0 for production. Both are non-commercial without a separate arrangement, and building on either creates a cost you have not priced.

My default is GPT Image 2.5 at medium with an Apache 2.0 fallback

My take: for most teams shipping product imagery in the EU, the default is GPT Image 2.5 at medium quality ($0.0132) for production output, with an Apache 2.0 open-weight model held as a fallback path rather than a primary. The reason is dependency rather than quality. A 256-point Elo gap is large, but the ability to keep generating when an API changes terms,

prices or availability is worth building against even if you never exercise it.

The number I would actually manage is the quality tier. Moving a pipeline from high to medium on GPT Image 2.5 cuts per-image cost by roughly 75%, from $0.0527 to $0.0132. No vendor switch in this comparison produces a saving that size.

And the date I would put in the plan is 2 December 2026. Whichever model a team picks, the machine-readable mark has to be verifiable by then for systems already on the EU market, and that is a pipeline engineering task with its own timeline, independent of which of these four names ends up in the config file.

WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting →