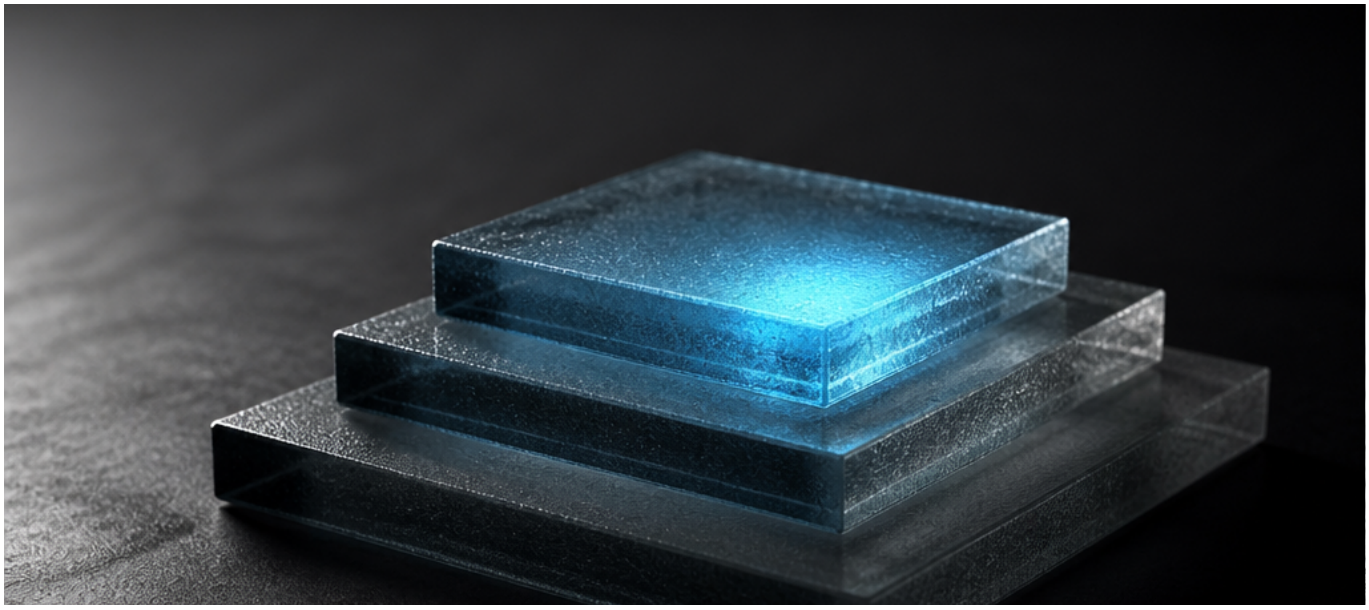


FLUX.2 klein vs MiniMax H3-Image vs Qwen-Image-2.1-Turbo: which open-weight image model can a Swiss CTO actually ship?

Picture the procurement review: three open-weight image models on the table, weights downloaded, benchmarks green. Legal reads the licenses and crosses off two of them before lunch.

Exactly one tier in this comparison carries a license you can put into a commercial product without a separate negotiation: FLUX.2 [klein] 4B and 4B Base, under Apache 2.0, at 8.4 GB and 9.2 GB of VRAM respectively per the [klein] model page. Everything above that line, including klein's own 9B tiers at 19.6 GB and 21.7 GB, is non-commercial, region-restricted, or research-only. That is the whole decision, and it has almost nothing to do with image quality.

The klein 4B ships and the klein 9B does not

Black Forest Labs released FLUX.2 [klein] on August 27, 2026, per its [release notes](https://docs.us.bfl.ai/release-notes), in four principal local variants. The [model page](https://bfl.ai/models/flux-2-klein) splits them cleanly: 4B and 4B Base are Apache 2.0, 9B and 9B Base ship under the FLUX Non-Commercial License. The Base variants are the undistilled foundation models and the non-Base ones are distilled for faster inference, which [the official repository](https://github.com/black-forest-labs/flux2) states directly. So the 4B Base at 9.2 GB is the undistilled, fully commercial-licensed option, and 4B at 8.4 GB is the quick one.

One wrinkle worth knowing before you size hardware: BFL's own documentation is inconsistent about klein 4B's VRAM. The product page says 8.4 GB, while a BFL blog figure puts it nearer 13 GB. I would plan against the higher number. A 16 GB card covers either reading with headroom, and a 12 GB card is a gamble on which document is current.

The bigger sibling, FLUX.2 [dev], is 32B parameters, released November 25, 2025, and wants roughly 80 GB of VRAM, which means H100-class hardware. Its [model card](https://huggingface.co/black-forest-labs/FLUX.2-dev) is unambiguous that commercial deployment and production use require a separate license or the commercial API. You can rent an H100 for an afternoon and still not be allowed to serve a single customer image from it.

MiniMax H3-Image is a 33.1B-parameter image checkpoint, [documented](https://openmodelweights.com/models/minimaxai-minimax-h3-image/) with an open weights release in November 2026. It ships under the MiniMax-H3-Image Community License Agreement, which is not OSI-approved. The license grants rights in a territory defined as "worldwide, excluding the European Union, the United Kingdom, the Republic of Korea, and the United States". [Coverage of the license](https://timesof.ai/2026/11/minimax-h3-image-open-weights-4k-license-excludes-us-eu-uk-korea) notes that use of both the model and its outputs outside the granted territory is unauthorized, with a separate authorization threshold referenced above roughly $1M annual revenue or for hosted services, and excluded regions invited to seek formal authorization.

Read that twice if you are in Basel, London, Seoul, or San Francisco. **The restriction is on use, and it extends to the images the model produces**, not merely to redistribution of the weights.

Qwen-Image-2.1-Turbo released October 9, 2026 and is, by the numbers available, the strongest of the three: 1,227 on LMArena at #22 overall as of November 17, 2026, the best-rated open image model in [that roundup](https://www.csuite.so). Hosted inference runs around $0.0141 per image. The license is described as research, which makes the LMArena position a benchmark fact rather than a procurement fact.

ATTRIBUTE	FLUX.2 [KLEIN] 4B	MINIMAX H3-IMAGE	QWEN-IMAGE-2.1-TURBO
License	**Apache 2.0**	MiniMax Community License (not OSI-approved)	Research license
Usable commercially in the EU	**Yes**	No (EU, UK, US, Korea excluded)	No
Parameters	4B	**33.1B**	n/a
VRAM to self-host	**8.4 GB (product page; ~13 GB per BFL blog)**	n/a	n/a
Public quality rating	n/a	n/a	**1,227 LMArena, #22 overall (Nov 17, 2026)**
Marginal cost per image	**$0 licensing, self-hosted**	n/a	$0.0141 hosted
Release date	Aug 27, 2026	Nov 2026	Oct 9, 2026

The empty cells matter as much as the filled ones. For MiniMax H3-Image I have a parameter count and a license territory, with no published VRAM figure or quality rating in the research in front of me. For Qwen-Image-2.1-Turbo I have a rating and a hosted price, with no self-hosting footprint. Anyone presenting you a clean three-way quality ranking across these models is filling gaps I cannot fill.

Why an EU company has to care about a license territory at all

Because the license is contract law, and it does not care that the weights were a free download. If your pipeline sits in Zurich and you generate marketing images with a model whose license grant explicitly excludes the European Union, you have no grant. That is the absence of permission, not a compliance nuance you remediate later with a policy document, and in the MiniMax case it extends to the outputs, which means the images themselves inherit the problem.

The same logic, less dramatically, applies to the non-commercial tiers. FLUX.2 [dev] and klein 9B are perfectly legal to run on a workstation for evaluation, concept work, or internal research. The moment output touches a customer-facing surface, you need BFL's commercial license or their API. That is a legitimate commercial model and BFL states it plainly. It just means the 80 GB machine you costed is a research expense rather than a production one.

My take: license tiering is now the primary product decision in open-weight image models, ahead of quality. Three vendors independently shipped capable weights behind restrictive terms in the same quarter, and the pattern I read is deliberate: give the ecosystem the artifact, keep the revenue. I expect the commercially clean tier to stay one size below the flagship, because that gap is the business model. Unconfirmed, but I would budget as if it holds through 2027.

What klein 4B costs you in VRAM, and what it buys

An 8.4 GB model fits on consumer hardware. A 16 GB card handles it comfortably even if the 13 GB figure turns out to be the accurate one. That changes the shape of a self-hosted pipeline: no H100 reservation, no 80 GB allocation, no per-image API meter. Against Qwen's $0.0141 per hosted image, klein 4B's marginal licensing cost is zero, and you pay in GPU time you already own or rent cheaply.

What you give up is unmeasured. There is no published quality comparison between klein 4B and the 9B tiers, and none between klein 4B and Qwen-Image-2.1-Turbo. A 4B distilled model against a 33.1B checkpoint is an unequal fight on paper. Whether the gap matters depends on your output: product photography variants and marketing backgrounds are a different bar from editorial illustration.

This is why I would structure evaluation around the licensed tier rather than around the

leaderboard. Run klein 4B Base (the undistilled one, 9.2 GB) against your actual prompts and your actual acceptance criteria. If it clears the bar, the license question is already answered and you ship. If it does not, you are choosing between BFL's commercial license for the bigger tiers and a hosted API, which is a pricing negotiation rather than an architecture problem. The GPU and node-risk side of this sits in the [self-hosted ComfyUI pipeline guide](https://www.arturmarkus.com/setting-up-a-self-hosted-comfyui-image-pipeline-a-ctos-guide-to-licences-gpu-costs-and-node-risk/), and the conclusion there holds here: the licence column decides the architecture, not the benchmark column.

One more model worth a line, because it keeps appearing in comparisons and should not. Bria Fibo 2.0, an 8B model, has roughly 25.5 GB of checkpoint components per [an October 23, 2026 write-up](https://www.freakyjolly.com), with no published official VRAM requirement and no confirmed license. A model with no confirmed license is a bookmark, not a candidate.

Pick klein 4B Base, then verify the VRAM figure on the day

My take: for a European company building a self-hosted image pipeline, FLUX.2 [klein] 4B Base is the only serious option in this set, and the reason is Apache 2.0 rather than anything about its output. **It is the single tier across three model families where you can download, deploy, serve customers, and never talk to a vendor's legal team.** MiniMax H3-Image is unusable in the EU by the plain text of its own license. Qwen-Image-2.1-Turbo is the best-rated open model here and is research-licensed, so its 1,227 LMArena score is interesting and commercially inert.

The verification I would not skip: confirm the current klein 4B VRAM figure directly against whatever BFL publishes on the day you size hardware, because the 8.4 GB versus 13 GB discrepancy sits inside BFL's own documentation and I cannot resolve it from the outside. And read the Apache 2.0 grant on the specific tier you download rather than the family page, because the 4B/9B split in this family is the whole point, and a mislabelled checkpoint is an expensive mistake to discover in production. I made a similar argument about licence-first selection when comparing [the hosted and open image models earlier this year](https://www.arturmarkus.com/gpt-image-2-5-vs-flux-2-vs-nano-banana-vs-qwen-image-the-honest-image-model-comparison-for-ctos/); the models changed, the

ordering of criteria did not.

Three model families, four license regimes, one tier you can ship. Size the 16 GB card.



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 →