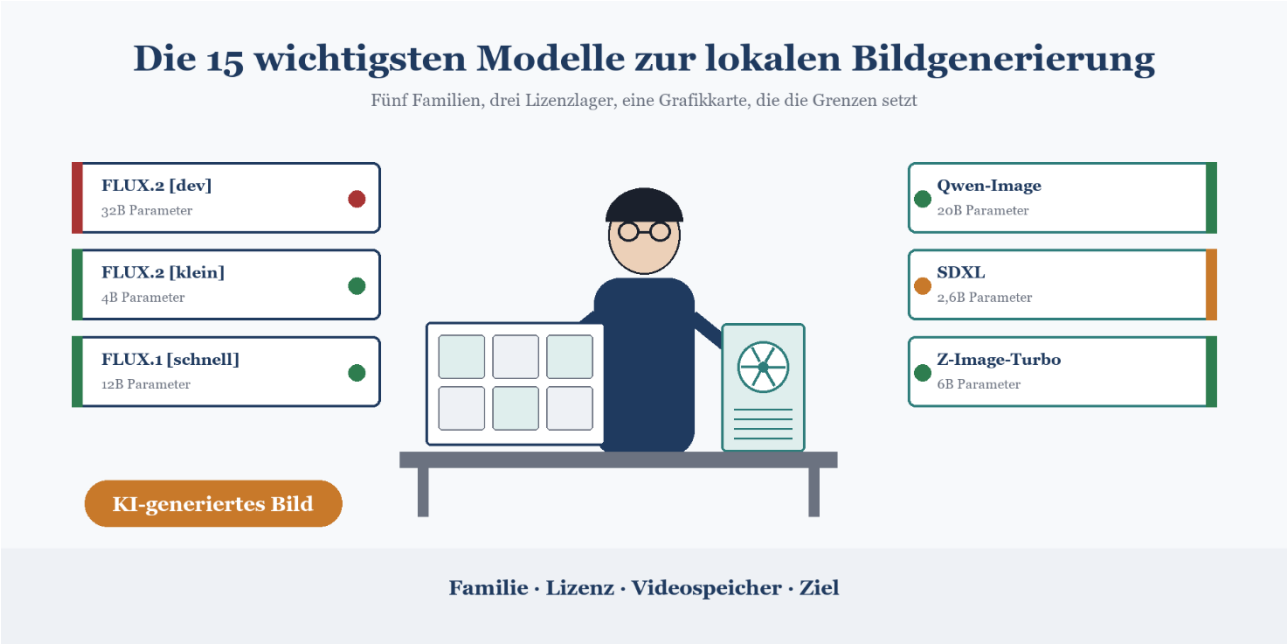


Die 15 wichtigsten Modelle zur lokalen Bildgenerierung



Ein Arbeitsplatz, eine Grafikkarte, ein Regal voller Modelle – und die Frage, welches davon du überhaupt benutzen darfst. Dieses Beitragsbild ist KI-generiert und entsprechend gekennzeichnet.

WISSEN <u>KI-Bildgenerierung – der Überblick</u>	BERATUNG <u>Beratung zur KI-Bildgenerierung</u>	SCHULUNG <u>Schulung KI-Bildgenerierung</u>
Das gesamte Themenfeld von Diffusion bis Governance an einem Platz.	Modellauswahl, Lizenzlage und Betriebsmodell für deinen konkreten Fall.	Vom ersten Prompt bis zur reproduzierbaren Produktionsstrecke.

Es gibt einen Moment, den jeder erlebt, der zum ersten Mal lokal Bilder erzeugen will. Du hast die Software installiert, du hast eine Grafikkarte, die dich mehr gekostet hat als dein erstes Auto, und dann stehst du vor einem Verzeichnis, das leer ist. Denn ohne Modell macht die schönste Oberfläche nichts. Also gehst du auf eine der großen Plattformen, tippst „realistic“ ins Suchfeld und bekommst eine sechsstellige Trefferzahl zurück. SDXL, SD 1.5, SD 3.5, Flux, Pony, Illustrious, NoobAI, Juggernaut, dazu Kürzel wie fp8, Q4_K_M, GGUF, Turbo, Lightning, klein, schnell, dev. Willkommen im Dschungel.

Die gute Nachricht zuerst: Du musst keine zehntausend Modelle kennen. Du musst fünf Familien kennen. Hinter dem Namenswirrwarr stehen wenige Architekturen, und fast alles, was du herunterladen kannst, ist ein Ableger von einer davon. Wenn du die Familien verstehst, ordnest du jedes neue Modell in wenigen Sekunden ein: Wie groß ist es, wie viel Videospeicher braucht es, welche Zusatzwerkzeuge passen dazu, und – der Teil, der in Unternehmen am meisten schiefgeht – darfst du die Bilder überhaupt verkaufen?

Genau um diese vier Fragen geht es hier. Fünfzehn Modelle, nach Familien sortiert, jedes mit Stärken, Schwächen, Lizenzlage und ehrlichem Speicherbedarf. Keine Punkteskala, keine Rangliste mit Gewinner, weil eine Rangliste in diesem Feld nach sechs Wochen Altpapier ist. Stattdessen die

Struktur, die bleibt – und die konkreten Namen, die heute die Arbeit machen. Wenn du wissen willst, wie das Ganze technisch überhaupt funktioniert, findest du das im großen Überblick [KI-Bildgenerierung – der Überblick](#). Hier geht es nur um Modelle.

Ein Wort zur Erwartungshaltung: Dieser Bereich bewegt sich schnell, aber nicht so schnell, wie die Marketingabteilungen behaupten. Die Familie, die du heute wählst, wird dich ein bis zwei Jahre begleiten. Der Ableger, den du innerhalb der Familie nimmst, wechselt vielleicht alle drei Monate. Baue also deine Prozesse auf Familien und deine Bilder auf Ablegern – nicht umgekehrt.

Fünf Familien statt hundert Namen

Ein Modell zur Bildgenerierung ist eine Datei mit mehreren Gigabyte, in der die gelernten Gewichte eines neuronalen Netzes liegen. Die Datei heißt in der Praxis Checkpoint, und was genau darin steckt – Diffusionsmodell, Textencoder, VAE – erklärt der Beitrag [Checkpoints, Modelle und VAE](#). Für die Einordnung reicht eine Ebene darüber: Jeder Checkpoint gehört zu einer Architektur, und die Architektur bestimmt alles Wichtige.

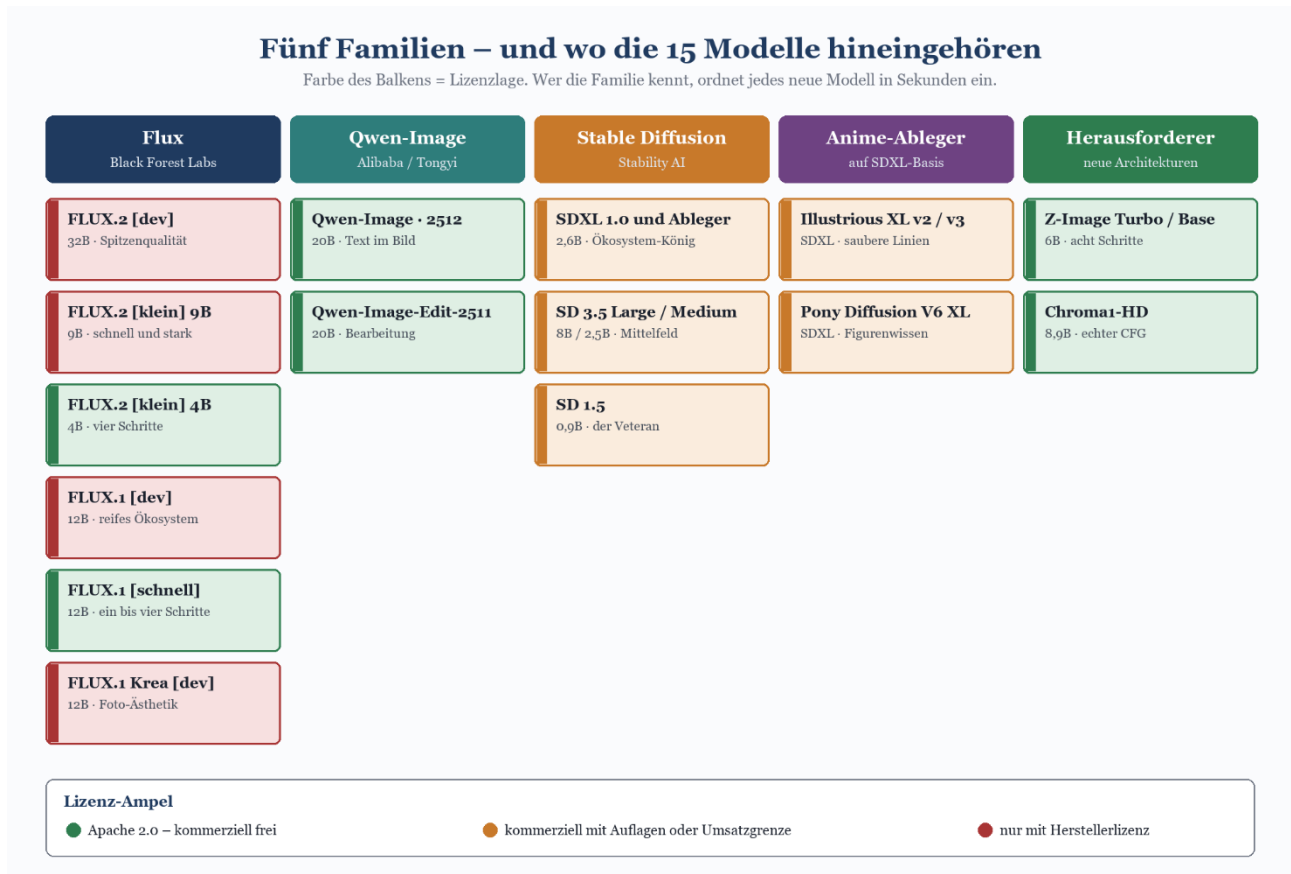
Warum die Familie mehr sagt als der Name

Die Architektur entscheidet, welche Zusatzpakete passen. Ein LoRA, das für SDXL trainiert wurde, funktioniert nicht mit Flux und nicht mit Qwen-Image – nicht schlecht, sondern überhaupt nicht. Es lädt einfach nicht. Dasselbe gilt für Steuerwerkzeuge, für Inpainting-Modelle, für Auflösungs-Tricks. Wer das nicht weiß, verbringt einen Abend damit, Fehlermeldungen zu lesen, die alle dasselbe sagen: falsche Familie. Was ein LoRA ist und warum es an die Basis gebunden bleibt, steht unter [Was ist ein LoRA?](#).

Die Architektur entscheidet zweitens, wie du Prompts schreibst. SDXL und seine Ableger reagieren auf Stichwortlisten mit Kommas, und bei den Anime-Modellen sind das sogar feste Schlagwörter aus einem Katalog. Flux und Qwen-Image wollen ganze Sätze, am liebsten eine kleine Bildbeschreibung in Prosa. Wer seine SDXL-Prompts eins zu eins auf Flux überträgt, bekommt matschige Ergebnisse und schimpft aufs Modell. Warum das Ergebnis manchmal trotzdem nicht zum Prompt passt, behandelt [Warum die KI den Prompt ignoriert](#); die Mechanik von Prompt und Negativ-Prompt erklärt [Prompt und Negativ-Prompt in der KI-Bildgenerierung](#).

Und die Architektur entscheidet drittens über die Lizenz. Das klingt seltsam, ist aber so: Fast alle Ableger erben die Lizenzbedingungen ihrer Basis. Wenn die Basis kommerzielle Nutzung nur mit gekaufter Lizenz erlaubt, dann gilt das für jeden Fine-Tune darauf – auch wenn der Fine-Tune auf der Plattform fröhlich als „free to use“ markiert ist. Diese eine Erkenntnis hat mir in Projekten mehr unangenehme Gespräche erspart als jede Qualitätsdiskussion.

Die Landkarte



Skizze 1: Die fünf Familien mit ihren Mitgliedern. Die Farbe des Seitenbalkens zeigt die Lizenzlage – grün heißt Apache 2.0 und damit kommerziell frei, gelb heißt Auflagen, rot heißt: nur mit Herstellerlizenz.

Zwei Dinge fallen an dieser Landkarte sofort auf. Erstens: Die Familie mit der besten Bildqualität ist gleichzeitig die mit der schlechtesten Lizenzlage. Zweitens: Die beiden Familien mit der saubersten Lizenz – Qwen-Image und die Herausforderer – sind die jüngsten und haben das kleinste Ökosystem. Das ist kein Zufall, sondern das zentrale Spannungsfeld der lokalen Bildgenerierung im Jahr 2026. Du bekommst nicht gleichzeitig Spitzenqualität, freie Lizenz und ein reifes Ökosystem. Du bekommst zwei von drei, und welche zwei, hängt davon ab, was du vorhast.

FAKTEN · Was sich in zwei Jahren verschoben hat

Drei Entwicklungen haben die Landkarte umgepflügt. Erstens ist Stability AI nicht mehr das Zentrum: Stable Diffusion 3.5 ist seit Ende 2024 die aktuelle Generation des Hauses, ein Nachfolger ist nicht erschienen, und das Unternehmen hat sich 2026 auf Audio und Markenwerkzeuge verlagert. Zweitens kommt die interessanteste Konkurrenz aus China und liegt unter Apache 2.0 – Qwen-Image und Z-Image sind vollständig frei nutzbar, auch kommerziell. Drittens hat Black Forest Labs mit FLUX.2 die Qualitätsspitze weiter ausgebaut, den Speicherbedarf aber gleich mit: Das Spitzenmodell hat 32 Milliarden Parameter und will in voller Genauigkeit etwa 64 Gigabyte Videospeicher.

Die 15 Modelle im Einzelnen

Ab hier wird es konkret. Die Reihenfolge folgt den Familien, nicht einer Wertung. Alle Speicherangaben sind Richtwerte für ein Bild um 1024 Pixel Kantenlänge und beziehen sich auf den Videospeicher der Grafikkarte, nicht auf den Arbeitsspeicher des Rechners.

Flux: die Qualitätsspitze mit dem Lizenzhaken

Black Forest Labs wurde von Leuten gegründet, die vorher an Stable Diffusion gearbeitet haben, und man merkt es. Die Flux-Modelle sind technisch eine andere Baustelle als SDXL: statt eines klassischen U-Net ein Transformer-Aufbau mit rektifizierten Flüssen, und im Ergebnis eine Prompt-Treue, die SDXL nie erreicht hat. Flux setzt Beziehungen zwischen Objekten um, zählt halbwegs zuverlässig, schreibt lesbare Schrift ins Bild und liefert Hände, die meistens fünf Finger haben – warum das überhaupt ein Problem war, erklärt [Warum KI keine Hände kann](#).

Die aktuelle Generation ist FLUX.2. Das Spitzenmodell FLUX.2 [dev] hat 32 Milliarden Parameter, beherrscht Text im Bild souverän und kann mehrere Referenzbilder gleichzeitig auswerten – für markenkonsistente Serien ist das der praktisch wichtigste Fortschritt. Wie so eine Markenkonsistenz konkret aufgesetzt wird, zeigt [Markenkonsistente KI-generierte Bilder](#). Der Preis ist doppelt: Speicher und Lizenz. In voller Genauigkeit reden wir über etwa 64 Gigabyte Videospeicher, in FP8 über rund 32, und erst als stark quantisierte GGUF-Variante passt das Modell mit knapp 19 Gigabyte auf eine 24-GB-Karte. Und kommerziell nutzen darfst du es nur mit einer gekauften Lizenz von Black Forest Labs oder über die Hersteller-API.

Die Antwort auf den Speicherhunger heißt FLUX.2 [klein]. Das ist keine abgespeckte Notlösung, sondern eine destillierte Variante, die Bild und Bearbeitung in einem Modell vereint und mit sehr wenigen Schritten auskommt. Es gibt sie in zwei Größen, und hier lohnt genaues Hinsehen: Die 4-Milliarden-Variante steht unter Apache 2.0 und ist damit kommerziell frei; die 9-Milliarden-Variante ist besser, steht aber wieder unter der nicht-kommerziellen Lizenz. Wer eine Karte mit 12 bis 16 Gigabyte hat und legal Geld verdienen will, landet mit ziemlicher Sicherheit bei FLUX.2 [klein] 4B. Dass FLUX.2 inzwischen auch in Microsoft-Werkzeugen auftaucht, ist ein eigenes Thema – nachzulesen unter [Black Forest FLUX.2 in PowerPoint und Copilot](#).

Die Vorgängergeneration FLUX.1 ist nicht tot, im Gegenteil. Sie hat das größere Ökosystem, weil zweieinhalb Jahre lang jeder darauf trainiert hat: tausende LoRAs, ausgereifte Steuerwerkzeuge, erprobte Arbeitsabläufe. FLUX.1 [dev] mit 12 Milliarden Parametern ist nach wie vor das Arbeitspferd vieler Produktionsstrecken, FLUX.1 [schnell] die Apache-2.0-Variante für ein bis vier Schritte, und FLUX.1 Krea [dev] der Spezialist gegen den überzuckerten Standardlook, gemeinsam mit Krea AI gebaut. Nur eben: alles außer [schnell] ist nicht-kommerziell.

Modell	Größe	VRAM (quantisiert bis voll)	Stärke	Schwäche	Lizenz
1. FLUX.2 [dev]	32B	19 – 64 GB	Beste Prompt-Treue, Text im Bild, mehrere Referenzbilder	Speicherhunger, langsam, große Downloads	nicht kommerziell
2. FLUX.2 [klein] 9B	9B	12 – 24 GB	Nahe an [dev], erzeugt und bearbeitet in einem Modell	Lizenz blockiert kommerziellen Einsatz	nicht kommerziell
3. FLUX.2 [klein] 4B	4B	8 – 13 GB	Vier Schritte, unter einer Sekunde auf guter Karte, frei nutzbar	Weniger Detailtiefe als die großen Geschwister	Apache 2.0
4. FLUX.1 [dev]	12B	7 – 24 GB	Reifstes Flux-Ökosystem, sehr viele LoRAs und Werkzeuge	Kein klassischer Negativ-Prompt, Lizenz	nicht kommerziell
5. FLUX.1 [schnell]	12B	7 – 24 GB	Apache 2.0, ein bis vier Schritte, gute Basis für Fine-Tunes	Flacher im Detail, wenig Stilbandbreite	Apache 2.0

6. FLUX.1 Krea [dev]	12B	7 – 24 GB	Bricht den typischen KI-Glanz, glaubwürdige Fotoästhetik	Eigenwilliger Stil, nicht neutral	nicht kommerziell
----------------------	-----	-----------	--	-----------------------------------	-------------------

Die Flux-Familie. VRAM-Spanne von stark quantisiert bis volle Genauigkeit; die Lizenzspalte entscheidet in Unternehmen oft vor der Qualität.

WARNUNG · „dev“ ist ein Entwicklungsname, keine Freigabe

Die Bezeichnung [dev] bei Flux steht nicht für „darfst du frei benutzen“, sondern für die Entwicklungs- und Forschungsvariante. FLUX.2 [dev], FLUX.1 [dev], FLUX.1 Krea [dev] und FLUX.2 [klein] 9B stehen unter einer nicht-kommerziellen Lizenz. Du darfst damit lernen, testen, trainieren und private Bilder erzeugen. Sobald ein Bild in einer Broschüre, auf einer Unternehmensseite oder in einer Kundenpräsentation landet, brauchst du eine gekaufte Lizenz von Black Forest Labs – gestaffelt von der einzelnen Domain bis zum Enterprise-Vertrag. Ich habe mehr als einmal erlebt, dass eine komplette Bildstrecke nach der Freigabe neu erzeugt werden musste, weil jemand „das kostenlose Flux“ genommen hatte.

Qwen-Image: der Text-Champion unter Apache 2.0

Wenn ich in Projekten eine einzige Empfehlung aussprechen müsste, ohne die Lizenzlage vorher zu kennen, wäre es Qwen-Image. Das Modell aus dem Qwen-Team von Alibaba hat 20 Milliarden Parameter, steht unter Apache 2.0 – also ohne Umsatzgrenze, ohne Nennungspflicht, ohne Rückfrage – und ist bei einer Disziplin schlicht das beste offene Modell: Schrift im Bild. Nicht „lesbar mit Mühe“, sondern Plakate, Verpackungen, Beschilderung, Diagrammbeschriftungen, und das in lateinischer wie chinesischer Schrift.

Die Familie wird gepflegt. Auf das Ursprungsmodell folgte eine überarbeitete Fassung mit glaubwürdigeren Menschen und feineren Naturtexturen, und daneben steht der eigentliche Schatz: Qwen-Image-Edit. Die Bearbeitungsvariante nimmt ein vorhandenes Bild und eine Anweisung in natürlicher Sprache – „mach den Himmel bedeckt, aber lass die Person unverändert“ – und setzt das um, ohne dass du Masken malen musst. Die aktuelle Fassung hält Figuren über mehrere Bearbeitungsschritte hinweg konsistent und nimmt bis zu drei Referenzbilder gleichzeitig an, etwa um zwei Personen in ein gemeinsames Bild zu setzen.

Der Haken: 20 Milliarden Parameter wollen Platz. In voller Genauigkeit sind das um 40 Gigabyte, in FP8 rund 16, als quantisierte GGUF-Variante etwa 14, und im brutal komprimierten Fall passt es mit acht Gigabyte auf eine kleine Karte – dann aber mit sichtbaren Abstrichen. Wer mit Qwen-Image ernsthaft arbeiten will, sollte 16 Gigabyte Videospeicher als Untergrenze einplanen. Und es gibt eine zweite Generation, die deutlich schlanker gebaut ist und nativ in 2K rechnet – die läuft allerdings bisher nur über die Programmierschnittstelle des Herstellers, offene Gewichte gibt es dafür nicht. Für lokal heißt das: erste Generation, und die ist gut genug.

Modell	Größe	VRAM (quantisiert bis voll)	Stärke	Schwäche	Lizenz
7. Qwen-Image (inkl. überarbeiteter Fassung)	20B	14 – 40 GB	Bestes offenes Modell für Text im Bild, mehrsprachig, sehr ordentliche Fotos	Groß und langsam, kleineres LoRA-Ökosystem als SDXL	Apache 2.0
8. Qwen-Image-Edit	20B	14 – 40 GB	Bearbeitung per Anweisung ohne Maske,	Gleiche Speicherklasse	Apache

(aktuelle Fassung)			stabile Figuren, bis zu drei Referenzbilder	wie das Basismodell	2.0
--------------------	--	--	---	---------------------	-----

Die Qwen-Image-Familie. Zwei Modelle, eine Lizenz, und die klarste Antwort auf die Frage nach Schrift im Bild.

TIPP · Text im Bild ist ein Modellproblem, kein Prompt-Problem

Wenn ein Modell keine Schrift kann, hilft kein besserer Prompt und kein höherer Guidance-Wert – du bekommst nur sauberer gerenderten Buchstabensalat. Bei SDXL und seinen Ablegern ist Text im Bild strukturell schwach; das ist keine Einstellungssache. Für alles mit Beschriftung nimmst du Qwen-Image, FLUX.2 oder Z-Image. Bei längeren Textblöcken bleibt auch dort die Regel: kurze Zeichenfolgen in Anführungszeichen im Prompt, und alles, was wirklich korrekt sein muss, setzt du hinterher im Layoutprogramm darüber.

Die Stable-Diffusion-Linie: das Ökosystem, das nicht stirbt

Man könnte die Stable-Diffusion-Linie für erledigt halten. Technisch ist sie überholt, der Hersteller hat den Fokus verschoben, und jede Qualitätsmessung fällt gegen Flux und Qwen-Image aus. Trotzdem ist SDXL bis heute die Familie mit den meisten Downloads, und das hat einen guten Grund: Das Werkzeug-Ökosystem ist unerreicht. Wenn du eine Aufgabe hast, die über „schönes Bild aus Text“ hinausgeht – Posen vorgeben, Kanten übernehmen, Tiefenkarten nutzen, Regionen getrennt prompten, Gesichter austauschen, Bereiche nachzeichnen – dann gibt es das Werkzeug dafür für SDXL, und zwar in ausgereifter Form. Für die neuen Familien gibt es davon vielleicht die Hälfte.

Dazu kommt die Genügsamkeit. SDXL läuft mit acht Gigabyte Videospeicher anständig und mit sechs notfalls auch. Das Basismodell benutzt kaum jemand direkt; man nimmt einen Fine-Tune, den jemand auf ein Ziel weitertrainiert hat. Juggernaut XL ist seit Jahren der Allrounder für Fotorealismus, RealVis XL der Spezialist für Haut, Gesichter und natürliches Licht, DreamShaper XL der vielseitige Kandidat für alles Stilisierte. Diese drei sind keine exotische Wahl, sondern der Konsens einer sehr großen Anwenderschaft – und genau das ist ihr Wert: Für jedes Problem findest du innerhalb von zehn Minuten eine Anleitung.

Stable Diffusion 3.5 ist der unglückliche Teil der Geschichte. Technisch ist es ein Fortschritt gegenüber SDXL, vor allem bei Prompt-Treue und Schrift. Praktisch sitzt es in einem ungemütlichen Mittelfeld: mehr Speicherbedarf als SDXL, weniger Qualität als Flux oder Qwen-Image, ein Ökosystem, das nie richtig gewachsen ist. Und die Lizenz ist die einzige in dieser Liste mit einer Umsatzgrenze – kostenlos kommerziell nur bis eine Million US-Dollar Jahresumsatz, darüber braucht es einen Vertrag. In der Praxis überspringen die meisten diese Generation. Es gibt sie in zwei Größen, Large mit acht und Medium mit zweieinhalb Milliarden Parametern.

Und SD 1.5? Das Modell ist von 2022, es hat 0,9 Milliarden Parameter, es passt in vier Gigabyte Videospeicher, und es ist qualitativ chancenlos. Aber es ist immer noch die Antwort auf genau eine Frage: Was mache ich mit einem alten Notebook oder einer Karte mit vier Gigabyte? Dafür, und für schnelle Kompositionsskizzen, hat es weiter seinen Platz. Ansonsten: Finger weg, die Zeit ist vorbei.

Modell	Größe	VRAM (quantisiert bis voll)	Stärke	Schwäche	Lizenz
9. SDXL 1.0 samt Fine-Tunes	2,6B	6 – 10 GB	Größtes Werkzeug- und LoRA-Ökosystem, genügsam, schnell	Prompt-Treue deutlich unter Flux, Text im Bild praktisch	Open RAIL+ +-M

				unbrauchbar	
10. Stable Diffusion 3.5 (Large / Medium)	8B / 2,5B	8 – 18 GB	Besser als SDXL bei Prompt-Treue und Schrift	Unbequemes Mittelfeld, dünnes Ökosystem, Umsatzgrenze in der Lizenz	Community License
11. Stable Diffusion 1.5	0,9B	4 – 6 GB	Läuft auf fast jeder Hardware, riesige Altbestände an LoRAs	Qualität nicht mehr konkurrenzfähig, 512-Pixel-Basis	Open RAIL-M

Die Stable-Diffusion-Linie. Technisch nicht mehr vorn, aber bei Werkzeugvielfalt und Genügsamkeit weiter konkurrenzlos.

Fine-Tune auf SDXL	Wofür	Was ihn auszeichnet
Juggernaut XL	Fotorealismus allgemein	Der gutmütige Allrounder über Menschen, Produkte und Szenen hinweg – der häufigste Einstieg
RealVis XL	Menschen und Natur	Haut, Gesichter, natürliches Licht; ruhiger und weniger plakativ als die meisten Alternativen
DreamShaper XL	Konzeptkunst, Fantasy	Hang zum Stilisierten, gut für Bilder, die nicht dokumentarisch aussehen müssen

Die drei Fotorealismus-Ableger, die sich als Arbeitspferde etabliert haben. Alle drei erben die SDXL-Lizenz und legen eigene Bedingungen darauf – Modellkarte lesen.

Anime und Illustration: eine eigene Welt

Dieser Teil des Feldes funktioniert nach eigenen Regeln, und wer aus der Fotoecke kommt, unterschätzt ihn regelmäßig. Anime- und Illustrationsmodelle werden nicht mit Sätzen geprompted, sondern mit Schlagwörtern aus festen Katalogen, oft ergänzt um Qualitätsmarker, die zum Modell gehören und nirgends sonst funktionieren. Dafür bekommst du eine Kontrolle über Figuren, Posen und Stile, die kein allgemeines Fotomodell bietet.

Illustrious XL ist der Referenzpunkt geworden. Gezielt für hochwertige Illustration gebaut, mit saubereren Linien, stabileren Farben und besserer Anatomie als die älteren Anime-Modelle, inzwischen in mehreren Generationen, von denen die neueren nativ in höheren Auflösungen rechnen. Darauf aufgebaut ist NoobAI-XL, ein Ableger mit sehr breitem Training über Millionen von Bildern, der deutlich mehr Stilbandbreite und Figurenwissen mitbringt. Die Mehrheit der aktuellen Anime-Checkpoints auf den Plattformen stammt aus dieser Linie.

Pony Diffusion V6 XL ist der andere Pol – trotz des Namens kein Pony-Modell, sondern ein Charakterspezialist mit enzyklopädischem Figurenwissen und einem eigenen, punktebasierten Qualitätssystem im Prompt. Eine der größten LoRA-Sammlungen überhaupt hängt daran. Der Nachfolger V7 ist erschienen, allerdings auf einer anderen Architektur und mit erkennbaren Kinderkrankheiten bei Gesichtern und Schrift; eine Korrekturfassung wird gearbeitet. In der Praxis arbeitet die Mehrheit weiter mit V6. Wer hier kommerziell unterwegs ist, muss besonders genau lesen: Bei den Pony-Modellen gibt es Zusatzbedingungen, die den Einsatz in bezahlten Generierdiensten einschränken.

Modell	Basis	VRAM	Stärke	Schwäche
12. Illustrious XL (v2 / v3)	SDXL	8 – 12 GB	Saubere Linien, stabile Farben, gute Anatomie; große aktive Ableger-Familie inklusive NoobAI-XL	Lizenzlage je Version unterschiedlich, Schlagwort-Prompts nötig

13. Pony Diffusion V6 XL	SDXL	8 – 12 GB	Unerreichtes Figurenwissen, riesige LoRA-Sammlung, eigenes Qualitätssystem im Prompt	Eigenwillige Prompt-Konventionen, Zusatzbedingungen gegen bezahlte Generierdienste
--------------------------	------	-----------	--	--

Die beiden Pole der Anime-Welt. Beide sitzen auf SDXL, damit passt das ganze SDXL-Werkzeugregal – aber nicht die LoRAs der jeweils anderen Linie.

Die Herausforderer: Z-Image und Chroma

Zwei Modelle fallen in keine der etablierten Familien und sind trotzdem – oder gerade deswegen – die interessantesten Neuzugänge. Z-Image aus dem Tongyi-Labor von Alibaba ist mit sechs Milliarden Parametern erstaunlich klein und liefert Fotorealismus, der sich mit deutlich größeren Modellen messen lässt. Die destillierte Turbo-Fassung braucht acht Schritte und ist damit auf einer vernünftigen Karte in wenigen Sekunden fertig; in voller Genauigkeit will sie 14 bis 16 Gigabyte, in FP8 etwa acht, quantisiert sechs. Sie beherrscht zweisprachige Beschriftungen und folgt Anweisungen erfreulich brav. Daneben gibt es die nicht-destillierte Basis-Fassung für Leute, die eigene Fine-Tunes bauen wollen – die braucht dann wieder 30 bis 50 Schritte. Alles unter Apache 2.0.

Chroma1-HD ist der Sonderfall mit Geschichte. Es ist ein vollständig neu trainierter Ableger auf Basis von FLUX.1 [schnell], knapp neun Milliarden Parameter, unter Apache 2.0 – und das ist der Punkt: Damit existiert eine Flux-nahe Architektur, die du ohne Lizenzgespräch kommerziell nutzen und weitertrainieren darfst. Obendrein hat Chroma zurückgebaut, was bei Flux fehlt: echte klassifikatorfreie Guidance und damit einen funktionierenden Negativ-Prompt. Wer aus der SDXL-Welt kommt und den Negativ-Prompt vermisst, findet hier den Umstieg ohne Entzug. Wie dieser Regler wirkt, erklärt [CFG-Scale: Guidance in KI-Bildgeneratoren](#). Als Basismodell ist Chroma absichtlich neutral gehalten, also nicht auf einen Look getrimmt – das macht es zu einem hervorragenden Ausgangspunkt für eigene Fine-Tunes und zu einem etwas spröden Modell für den schnellen Griff.

Modell	Größe	VRAM (quantisiert bis voll)	Stärke	Schwäche	Lizenz
14. Z-Image Turbo (und Basis-Fassung)	6B	6 – 16 GB	Sehr gutes Verhältnis von Qualität zu Größe, acht Schritte, zweisprachiger Text	Junges Ökosystem, wenige LoRAs und Steuerwerkzeuge	Apache 2.0
15. Chroma1-HD	8,9B	6 – 18 GB	Flux-nahe Architektur unter Apache 2.0, echter Negativ-Prompt, neutrale Trainingsbasis	Kein fertiger Look, braucht mehr Prompt-Arbeit	Apache 2.0

Die beiden Herausforderer. Beide unter Apache 2.0, beide auf mittleren Karten lauffähig – und beide mit deutlich kleinerem Ökosystem als die Platzhirsche.

Außer Reichweite, aber erwähnenswert

Es gibt offene Modelle, die du zwar herunterladen darfst, aber zu Hause nicht sinnvoll betreiben kannst. Das prominenteste ist HunyuanImage 3.0 von Tencent: 80 Milliarden Parameter in einer Mixture-of-Experts-Architektur, von denen pro Durchlauf etwa 13 Milliarden aktiv sind. Es vereint Erzeugen, Bearbeiten und Bildverstehen in einem einzigen Modell und liefert entsprechend beeindruckende Ergebnisse. Die Herstellerempfehlung lautet acht Karten mit je 80 Gigabyte. Quantisiert lässt es sich mit 24 Gigabyte irgendwie überreden, aber der Spaß hält nicht lange. Die

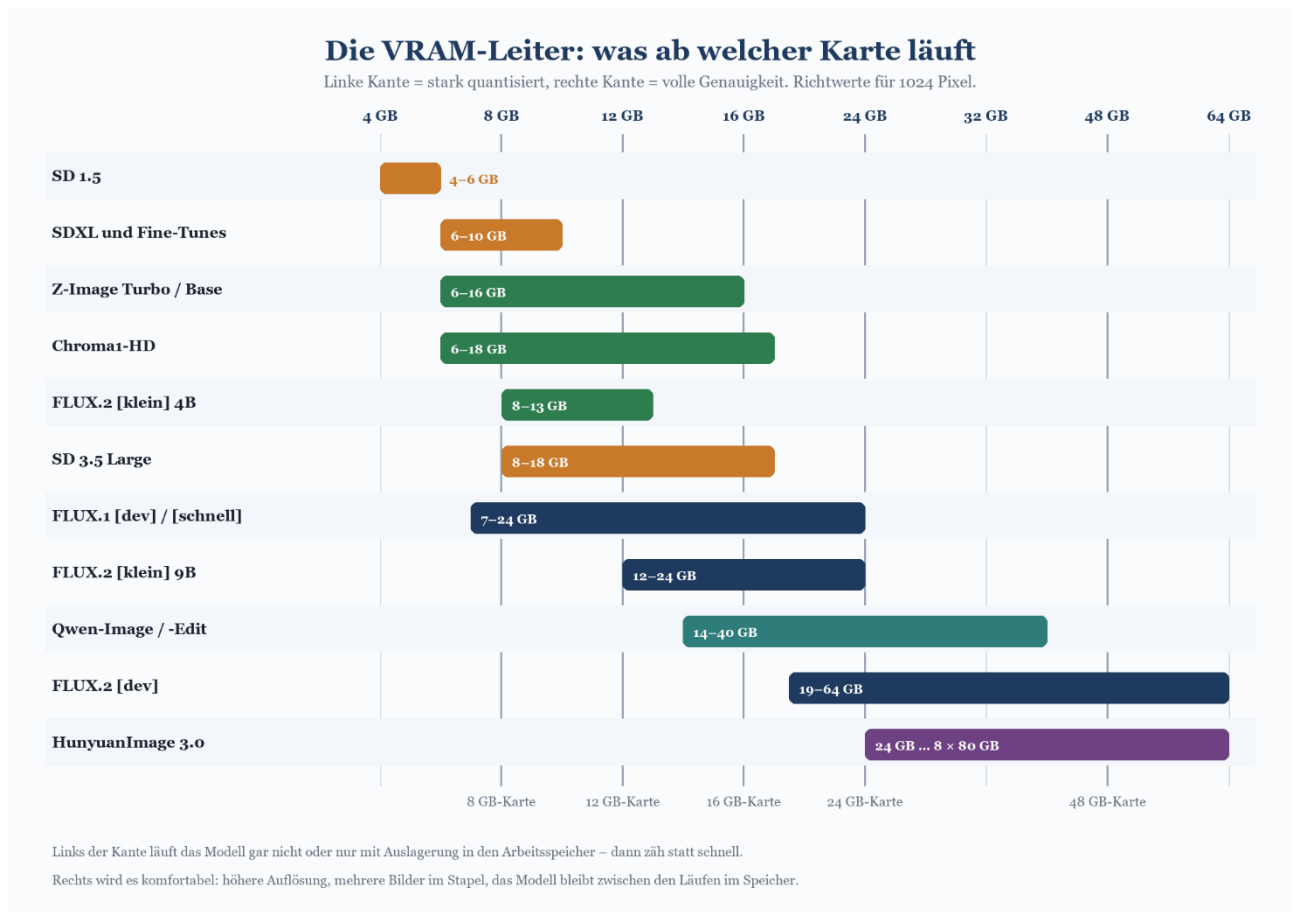
Lizenz ist übrigens die freundlichste dieser Größenklasse: kommerzielle Nutzung ist erlaubt, eine Zusatzvereinbarung braucht erst, wer mehr als hundert Millionen monatlich aktive Nutzer bedient. Das dürfte die meisten nicht betreffen.

Der Punkt an diesem Beispiel: Offene Gewichte bedeuten nicht lokal lauffähig. Die Lücke zwischen „das Modell ist frei“ und „ich kann es betreiben“ wird größer, nicht kleiner. Wenn du an diese Grenze stößt, ist die ehrliche Antwort oft ein gemischtes Betriebsmodell – die Abwägung dazu steht in [Cloud vs. lokale KI-Bildgenerierung](#), und was die geschlossenen Dienste heute leisten, in [Übersicht der wichtigsten KI-Bildgeneratoren](#).

Was auf deiner Hardware wirklich läuft

Der wichtigste Filter bei der Modellwahl ist nicht die Qualität, sondern der Videospeicher deiner Grafikkarte. Reicht er nicht, passiert eines von zwei Dingen: Die Software bricht mit einer Speicherfehlermeldung ab, oder sie lagert Teile des Modells in den Arbeitsspeicher aus und wird dabei so langsam, dass du aus einem Achtzehn-Sekunden-Bild ein Vier-Minuten-Bild machst. Beides kannst du in fünf Minuten selbst herausfinden, aber du kannst es auch vorher wissen.

Die VRAM-Leiter



Skizze 2: Speicherbedarf der Modelle von stark quantisiert (linke Kante) bis volle Genauigkeit (rechte Kante). Die grauen Linien markieren die üblichen Kartengrößen. Wo dein Balken beginnt, wird es möglich – wo er endet, wird es komfortabel.

Die Leiter liest sich in beide Richtungen. Von links: Mit acht Gigabyte bist du nicht raus, aber du lebst von Quantisierung und Geduld. Mit zwölf läuft das meiste, mit sechzehn wird es entspannt, mit vierundzwanzig hast du freie Wahl inklusive FLUX.2 [dev] in komprimierter Form. Von rechts: Wenn du nicht quantisieren willst – weil du trainieren willst, weil du maximale Qualität brauchst

oder weil du keine Lust auf Fehlersuche in exotischen Dateiformaten hast – verschiebt sich alles um eine Klasse nach oben.

Karte	Was komfortabel läuft	Was mit Kompromissen geht	Was du bleiben lassen solltest
4 – 6 GB	SD 1.5 und seine Ableger	SDXL in kleiner Auflösung, Z-Image stark quantisiert	Alles ab 12 Milliarden Parametern
8 GB	SDXL samt Fine-Tunes, Anime-Modelle, Z-Image quantisiert	FLUX.1 und FLUX.2 [klein] 4B als GGUF, Chroma quantisiert	Qwen-Image in brauchbarer Qualität, FLUX.2 [dev]
12 GB	SDXL-Welt, FLUX.1 in FP8, FLUX.2 [klein] 4B, Z-Image, Chroma	Qwen-Image als GGUF, SD 3.5 Large, FLUX.2 [klein] 9B quantisiert	FLUX.2 [dev]
16 GB	Alles Vorige plus Qwen-Image und Qwen-Image-Edit als GGUF, Z-Image in voller Genauigkeit	FLUX.2 [klein] 9B, erste LoRA-Trainings auf SDXL	FLUX.2 [dev] in FP8
24 GB	Praktisch die ganze Liste, FLUX.1 in voller Genauigkeit	FLUX.2 [dev] als GGUF Q4, LoRA-Training auf Flux	HunyuanImage 3.0 im Alltagsbetrieb
48 GB und mehr	FLUX.2 [dev] in FP8, Qwen-Image in voller Genauigkeit, Trainingsläufe	FLUX.2 [dev] in voller Genauigkeit über zwei Karten	Nichts aus dieser Liste mehr

Grobe Zuordnung von Kartengröße zu Modellklasse. Richtwerte für Einzelbilder um 1024 Pixel – höhere Auflösungen und größere Stapel verschieben alles nach oben.

Quantisierung: der Trick, der fast immer funktioniert

Quantisierung heißt, die Gewichte des Modells mit weniger Bits pro Zahl zu speichern. Aus 16 Bit werden acht, oder vier, manchmal zwei. Das Modell wird kleiner, passt in weniger Speicher, rechnet oft sogar schneller – und verliert ein wenig Präzision. Die praktische Erfahrung: Der Sprung von voller Genauigkeit auf FP8 ist in Blindtests kaum zu sehen. Der Sprung auf vier Bit ist sichtbar, aber meist akzeptabel, vor allem bei Feinheiten wie Haaransätzen, kleinen Gesichtern und – ausgerechnet – Schrift. Unter vier Bit wird es hässlich.

In der Praxis triffst du auf zwei Formate. FP8 ist die einfache Variante, die direkt von den üblichen Oberflächen unterstützt wird und moderne Karten voraussetzt. GGUF ist das flexiblere Format mit vielen Stufen von Q8 bis Q2, braucht aber eine Zusatzkomponente in der Oberfläche. Wenn du auf einer 8- oder 12-GB-Karte arbeitest, wird GGUF dein Standardformat, und Q4_K_M ist der Kompromiss, mit dem die meisten glücklich werden.

TIPP · Quantisierung zuerst am Textencoder sparen

Bei den großen Modellen steckt ein erheblicher Teil des Speicherbedarfs nicht im Bildmodell, sondern im Textencoder – bei Flux und Qwen-Image sind das eigene Sprachmodelle. Wenn dir zwei Gigabyte fehlen, lohnt der erste Versuch dort: Textencoder in FP8, Bildmodell in voller Genauigkeit. Das kostet weniger Bildqualität als der umgekehrte Weg, weil der Encoder nur den Prompt in Zahlen übersetzt und nicht jedes Pixel malt. Wie du das in einer produktiven Strecke verdrahtest, zeigt [Aufbau einer skalierbaren Bild-Maschine mit ComfyUI und Flux](#).

Was der Videospeicher nicht löst

Drei Dinge werden bei der Hardwareplanung regelmäßig vergessen. Erstens der Arbeitsspeicher: Ein 32-Milliarden-Parameter-Modell wird beim Laden durch den Hauptspeicher geschleust, und

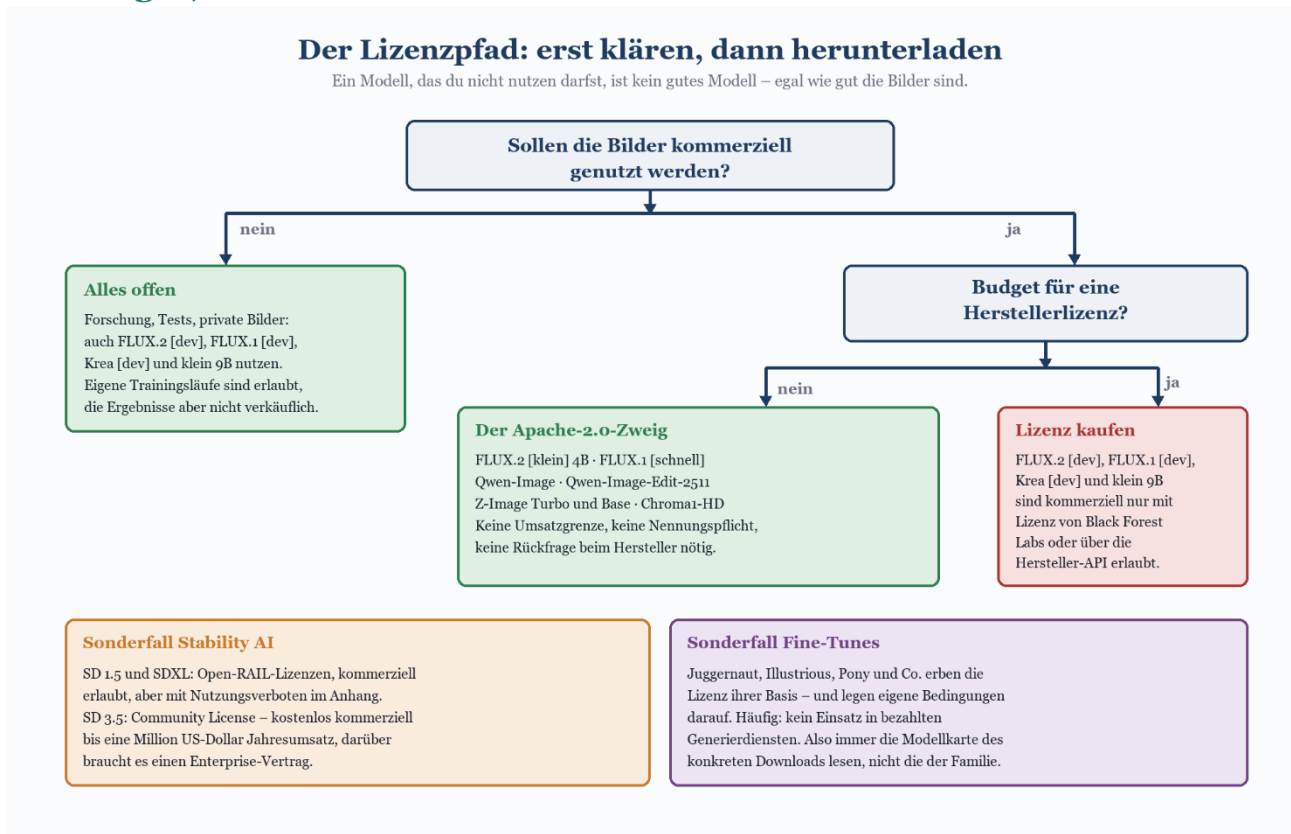
wer mit 16 Gigabyte antritt, erlebt eine Überraschung. 32 Gigabyte sind die vernünftige Untergrenze für die großen Modelle, 64 wenn du auslagern willst. Zweitens der Plattenplatz: Ein vollständiger FLUX.2-Satz aus Modell, Textencoder und VAE belegt schnell 60 Gigabyte, und du wirst nicht bei einem Modell bleiben. Rechne mit einer halben Terabyte-SSD, nur für Modelle.

Drittens die Auflösung. Jedes Modell hat eine Auflösung, für die es trainiert wurde – bei SD 1.5 512 Pixel, bei SDXL und den meisten neueren rund 1024, bei den jüngsten Modellen teils darüber. Wenn du weit darüber hinausgehst, steigt der Speicherbedarf quadratisch, und die Bildqualität sinkt oft trotzdem, weil das Modell Motive doppelt. Der saubere Weg ist erzeugen auf Modellauflösung und hinterher hochskalieren; die Details dazu in [Auflösung & Seitenverhältnis bei KI-Bildern](#). Und wer mehrere Bilder gleichzeitig rechnen lässt, sollte wissen, dass die Stapelgröße den Speicherbedarf mitskaliert und unter Umständen die Reproduzierbarkeit beeinflusst – nachzulesen in [batch size und Reproduzierbarkeit](#).

Die Lizenzfrage kostet Geld

Jetzt der Teil, den fast alle Modellvergleiche im Netz überspringen, weil er weder Bilder noch Benchmarks hergibt. Er entscheidet aber darüber, ob deine Arbeit verwertbar ist. Ein Modell, dessen Ergebnisse du nicht veröffentlichen darfst, ist für ein Unternehmen kein gutes Modell – auch wenn es die schönsten Bilder macht.

Drei Lager, klar unterscheidbar



Skizze 3: Der Entscheidungspfad. Zwei Fragen genügen, um in einem der Lager zu landen – und die beiden Sonderfälle darunter sind genau die, an denen es in Projekten hängt.

Lager eins ist Apache 2.0. Diese Lizenz erlaubt kommerzielle Nutzung ohne Umsatzgrenze, ohne Namensnennung und ohne Rückfrage. Sie ist der Grund, warum ich in Unternehmensprojekten fast immer bei Qwen-Image, Z-Image, Chroma1-HD, FLUX.1 [schnell] oder FLUX.2 [klein] 4B lande. Nicht weil diese Modelle die besten Bilder machen, sondern weil sie die einzigen sind, bei denen die Rechtsabteilung keine Rückfragen hat.

Lager zwei sind die Lizenzen mit Auflagen. Die Open-RAIL-Familie bei Stable Diffusion erlaubt kommerzielle Nutzung, hängt aber einen Katalog verbotener Verwendungen an, den du einhalten musst und den fast niemand liest. Die Community License bei Stable Diffusion 3.5 zieht eine Umsatzgrenze bei einer Million US-Dollar pro Jahr; darüber brauchst du einen Vertrag mit Stability AI. Die Tencent-Lizenz bei HunyuanImage zieht die Grenze bei hundert Millionen monatlich aktiven Nutzern. Das sind handhabbare Bedingungen, aber sie sind Bedingungen.

Lager drei ist nicht-kommerziell. Dazu gehören FLUX.2 [dev], FLUX.2 [klein] 9B, FLUX.1 [dev] und FLUX.1 Krea [dev]. Du darfst die Gewichte herunterladen, damit arbeiten, Fine-Tunes trainieren und die Ergebnisse privat nutzen. Für jede kommerzielle Verwendung brauchst du eine Lizenz von Black Forest Labs – gestaffelt von der Nutzung innerhalb einer Domain über Produkt- und Agenturmodelle bis zum Enterprise-Vertrag – oder du nutzt die Modelle über die Hersteller-API, wo die Rechte mitgeliefert werden.

FAKTEN • Lizenzlage der 15 Modelle auf einen Blick

Die folgende Tabelle ist der Teil dieses Artikels, den du dir ausdrucken und an den Monitor hängen solltest. Sie beantwortet die einzige Frage, die vor dem Download zählt: Darf das, was hinten herauskommt, Geld verdienen? Stand September 2026 – Lizenzen können sich ändern, die Modellkarte des konkreten Downloads ist immer die verbindliche Quelle.

Modell	Lizenz	Kommerziell nutzbar?	Worauf du achten musst
FLUX.2 [dev]	FLUX.2 [dev] Non-Commercial	Nur mit gekaufter Lizenz	Lizenz staffelt nach Einsatzart; API-Nutzung bringt die Rechte mit
FLUX.2 [klein] 9B	Non-Commercial	Nur mit gekaufter Lizenz	Wird oft mit der 4B-Variante verwechselt – Dateiname prüfen
FLUX.2 [klein] 4B	Apache 2.0	Ja, ohne Auflagen	Die einzige FLUX.2-Variante, die kommerziell frei ist
FLUX.1 [dev]	FLUX.1 [dev] Non-Commercial	Nur mit gekaufter Lizenz	Gilt auch für alle Fine-Tunes und Merges darauf
FLUX.1 [schnell]	Apache 2.0	Ja, ohne Auflagen	Basis vieler freier Ableger, unter anderem Chroma
FLUX.1 Krea [dev]	FLUX.1 [dev] Non-Commercial	Nur mit gekaufter Lizenz	Dieselbe Lizenz wie [dev], trotz eigenem Namen
Qwen-Image und Folgefassungen	Apache 2.0	Ja, ohne Auflagen	Auch Fine-Tunes bleiben frei, sofern der Autor nichts draufsetzt
Qwen-Image-Edit	Apache 2.0	Ja, ohne Auflagen	Bei Bearbeitung fremder Bilder zählt das Urheberrecht am Ausgangsbild
SDXL 1.0	CreativeML Open RAIL++-M	Ja, mit Nutzungsverboten	Verbotskatalog im Lizenzanhang gilt weiter, auch für Ableger
Stable Diffusion 3.5	Stability Community License	Ja, bis 1 Mio. USD Jahresumsatz	Darüber Enterprise-Vertrag; Umsatz des Unternehmens zählt, nicht des Projekts
Stable Diffusion 1.5	CreativeML Open	Ja, mit	Herkunft der Gewichte bei alten

	RAIL-M	Nutzungsverboten	Kopien oft unklar
Illustrious XL	je Version unterschiedlich	Meist ja, prüfen	Frühe Versionen unter einer eigenen Lizenz, spätere SDXL-nah
Pony Diffusion V6 XL	SDXL-Basis plus Zusatzbedingungen	Eingeschränkt	Zusatzbedingungen richten sich gegen bezahlte Generierdienste
Z-Image Turbo und Basis	Apache 2.0	Ja, ohne Auflagen	Bearbeitungsvariante der Familie ist noch nicht veröffentlicht
Chroma1-HD	Apache 2.0	Ja, ohne Auflagen	Bewusst ohne Inhaltsfilter – die Verantwortung liegt komplett bei dir

Lizenzlage der fünfzehn Modelle, Stand September 2026. Diese Tabelle ist keine Rechtsberatung, sondern eine Landkarte, wo du genauer hinschauen musst.

Fine-Tunes erben – und legen nach

Der häufigste Lizenzfehler in der Praxis ist nicht, dass jemand die Lizenz eines Basismodells missversteht. Es ist, dass jemand sie gar nicht anschaut, weil er einen Fine-Tune heruntergeladen hat und die Plattform ein grünes Häkchen zeigt. Ein Fine-Tune erbt die Bedingungen seiner Basis. Ein auf FLUX.1 [dev] trainiertes Fotomodell ist nicht kommerziell nutzbar, egal was in der Modellbeschreibung steht. Und umgekehrt kann ein Autor auf eine freie Basis eigene Bedingungen setzen – etwa das Verbot, das Modell in einem bezahlten Bildgenerierdienst anzubieten.

Praktisch heißt das: zwei Lizenzen prüfen, immer. Die der Basis und die des konkreten Downloads. Und dann aufschreiben, denn in zwei Jahren erinnert sich niemand mehr, welcher Checkpoint hinter welcher Bildstrecke steckte. Genau dafür pflege ich in meinen Projekten einen Modellkatalog: pro Checkpoint der Dateiname, die Prüfsumme, die Basis, die Lizenz, das Datum und die Freigabe. In UB.Dossier sind das schlicht Datensätze mit Nachweis – wenn sich eine Lizenz ändert, ändert sich ein Wert, und man sieht danach, welche Bilder betroffen sind. Ohne so einen Katalog ist die Frage „durften wir das“ nach einem Jahr nicht mehr beantwortbar, und die Antwort „wir glauben schon“ überzeugt in einer Prüfung niemanden.

WARNUNG · Die Lizenz ist nur die halbe Prüfung

Was ein Modell darf, regelt die Lizenz. Was es gelernt hat, regelt sie nicht. Ein Fine-Tune, der gezielt auf geschützte Marken, Figuren aus laufenden Serien oder reale Personen trainiert wurde, ist rechtlich heikel, auch wenn er unter Apache 2.0 steht – denn Marken-, Urheber- und Persönlichkeitsrecht interessieren sich nicht für Modelllizenzen. Wenn ein Modell auf Nachfrage erkennbar eine bestimmte Schauspielerin oder ein bestimmtes Comic-Universum ausgibt, weißt du genug. Für Unternehmensbilder gilt die einfache Regel: neutrale Basis, eigener Stil per LoRA aus eigenem Material. Wie das aufgesetzt wird, besprechen wir in der [Beratung zur KI-Bildgenerierung](#).

Kennzeichnung ist keine Modellfrage – aber deine Pflicht

WICHTIG · Transparenzpflichten gelten seit dem 2. August 2026

Kein Modell in dieser Liste erzeugt von sich aus eine rechtssichere Kennzeichnung. Die Transparenzpflichten des EU AI Act für synthetische Inhalte liegen bei dir, nicht beim Modellhersteller – und sie gelten unabhängig davon, ob das Bild lokal oder in der Cloud entstanden ist. Praktisch brauchst du zwei Dinge: eine sichtbare Kennzeichnung am Bild und maschinenlesbare Metadaten in der Datei. Was konkret gilt, steht in [EU AI Act: Was seit dem 2. August 2026 gilt](#); welche Spuren KI-Bilder überhaupt hinterlassen, in [KI-generierte Bilder erkennen](#).

Der Zusammenhang zur Modellwahl ist indirekt, aber real: Wenn du lokal arbeitest, bist du für die Metadaten selbst verantwortlich, denn die großen Dienste schreiben ihre Herkunftsinformationen automatisch mit, eine lokale Installation nicht. Der Modellname, die Version, der Seed, die Schrittzahl und die Guidance gehören in die Datei – nicht aus Ordnungsliebe, sondern weil du sonst bei einer Rückfrage nicht sagen kannst, womit ein Bild entstanden ist. Wie man diesen Nachweis strukturiert ablegt, behandelt [Digital Asset Management für KI-Bilder mit SharePoint Online und Power Platform](#).

Modell nach Ziel wählen, nicht nach Rangliste

„Was ist das beste Modell?“ ist die Frage, die ich am häufigsten höre, und sie hat keine Antwort. Sie ist so gestellt wie „was ist das beste Werkzeug?“ – die Antwort hängt davon ab, ob du eine Schraube drehen oder eine Wand aufstemmen willst. Was es gibt, ist ein passendes Modell pro Ziel, und die Zuordnung ist stabiler als jede Bestenliste.

Die Ziel-Matrix

Nicht das beste Modell suchen, sondern das passende		
Jedes Ziel hat eine natürliche erste Wahl – und einen Ersatz, wenn der Videospeicher nicht reicht.		
Fotorealistische Menschen und Produkte ERSTE WAHL FLUX.2 [dev] oder FLUX.2 [klein] WENN NICHT, DANN Z-Image Turbo · Juggernaut XL auf SDXL-Basis	Lesbarer Text im Bild ERSTE WAHL Qwen-Image bzw. Qwen-Image-2512 WENN NICHT, DANN Z-Image für kurze Beschriftungen · FLUX.2	Vorhandenes Bild gezielt ändern ERSTE WAHL Qwen-Image-Edit-2511 WENN NICHT, DANN FLUX.2 [klein] erzeugt und bearbeitet in einem Modell
Anime, Manga, Illustration ERSTE WAHL Illustrious XL v2/v3 oder NoobAI-XL WENN NICHT, DANN Pony Diffusion V6 XL für breites Figurenwissen	Viele Varianten in kurzer Zeit ERSTE WAHL Z-Image Turbo mit acht Schritten WENN NICHT, DANN FLUX.2 [klein] 4B mit vier · FLUX.1 [schnell]	Eigenes Fine-Tune oder LoRA trainieren ERSTE WAHL SDXL – das größte Werkzeug-Ökosystem WENN NICHT, DANN Chroma1-HD als neutrale Apache-Basis · FLUX.1 [dev]

Skizze 4: Sechs typische Ziele, je eine erste Wahl und eine Alternative für den Fall, dass Videospeicher oder Lizenz dagegen sprechen.

Wenn du das willst	Nimm das	Weil
Fotorealistische Menschen, Produkte, Szenen	FLUX.2 [klein] 4B, bei Lizenz und Speicher FLUX.2 [dev]	Beste Prompt-Treue, saubere Anatomie, glaubwürdiges Licht
Dasselbe, aber auf einer 8-GB-Karte	Juggernaut XL oder RealVis XL auf SDXL	Genügsam, erprobt, unzählige Anleitungen und Hilfsmittel
Lesbaren Text im Bild	Qwen-Image	Das einzige offene Modell, bei dem Beschriftungen verlässlich stimmen
Ein vorhandenes Bild gezielt ändern	Qwen-Image-Edit	Anweisung in Prosa statt Maskenmalerei, stabile Figuren über Schritte
Anime, Manga, Illustration	Illustrious XL oder NoobAI-XL	Saubere Linien, Schlagwortsteuerung, riesige Stil- und Figurenbibliothek

Viele Varianten in kurzer Zeit	Z-Image Turbo	Acht Schritte, kleines Modell, gute Qualität – ideal für Vorauswahl
Einen eigenen Stil antrainieren	SDXL, alternativ Chroma1-HD	SDXL: reifste Trainingswerkzeuge. Chroma: neutrale Basis unter Apache 2.0
Maximale Rechtssicherheit im Unternehmen	Qwen-Image, Z-Image, Chroma1-HD, FLUX.1 [schnell]	Apache 2.0 ohne Umsatzgrenze und ohne Nennungspflicht

Ziel, Modell, Begründung. Diese Zuordnung ist über Modellgenerationen hinweg stabiler als jede Qualitätsrangliste.

Der Blindtest, der zwei Stunden spart

Egal wie gut ein Vergleich geschrieben ist – welches Modell zu deinen Motiven, deiner Schreibweise und deinem Geschmack passt, weiß nur dein Bildschirm. Der Test dafür ist schnell und erschreckend aussagekräftig: drei Kandidaten, derselbe Prompt, derselbe Seed, dieselbe Schrittzahl, dieselbe Auflösung. Zehn Motive, die deine echte Arbeit abbilden, nicht zehn Drachen im Sonnenuntergang. Dann die Bilder anonymisiert nebeneinander legen und auswählen, ohne zu wissen, welches von welchem Modell ist.

Warum der Seed dabei nicht verhandelbar ist, erklärt [Der Seed bei der KI-Bildgenerierung](#). Ohne festen Seed vergleichst du Zufall mit Zufall. Ebenso wichtig: Nicht jedes Modell verträgt dieselben Einstellungen. Destillierte Modelle wie FLUX.2 [klein], FLUX.1 [schnell] oder Z-Image Turbo arbeiten mit wenigen Schritten und sehr niedriger Guidance; wer sie mit den SDXL-Werten von 30 Schritten und CFG 7 füttert, bekommt verbrannte Bilder und hält das Modell für schlecht. Die Grundlagen dazu in [Sampling-Steps erklärt](#) und [Sampler & Scheduler erklärt](#).

TIPP · Prompts übersetzen, nicht kopieren

Beim Familienwechsel ist der Prompt die häufigste Fehlerquelle. SDXL-Prompts sind Stichwortlisten, Flux- und Qwen-Prompts sind Beschreibungen in ganzen Sätzen. Nimm deinen besten SDXL-Prompt und lass ihn in eine Bildbeschreibung umformulieren, bevor du urteilst – dafür ist ein Sprachmodell das naheliegende Werkzeug, siehe [Claude im Unternehmen](#). Der Qualitätsunterschied zwischen kopiertem und übersetztem Prompt ist größer als der zwischen zwei Modellgenerationen.

Wie viele Modelle du wirklich brauchst

Drei. In fast jedem Projekt, das ich begleitet habe, reichen drei Modelle: ein Arbeitspferd für den Alltag, ein schnelles für Vorauswahl und Varianten, ein Spezialist für den einen Fall, der sonst nicht geht – meist Text im Bild oder Bearbeitung. Wer zwölf Checkpoints im Verzeichnis hat, hat nicht mehr Möglichkeiten, sondern nur mehr Unsicherheit darüber, womit welches Bild entstand.

Die praktische Kombination, mit der ich am häufigsten arbeite: Qwen-Image als Arbeitspferd, weil Lizenz und Textfähigkeit stimmen. Z-Image Turbo für die schnelle Vorauswahl, weil acht Schritte acht Schritte sind. Und FLUX.2 [klein] 4B für das letzte Stück Prompt-Treue, wenn eine Szene kompliziert wird. Alle drei unter Apache 2.0, alle drei auf einer 16-GB-Karte lauffähig, keine Lizenzgespräche. Wenn dann noch ein SDXL-Ableger für die Fälle dabei ist, in denen ein exotisches Steuerwerkzeug gebraucht wird, ist der Werkzeugkasten komplett.

Ein Modellwechsel in einer laufenden Produktionsstrecke ist übrigens kein Dateiumbenennen. Andere Familie heißt anderer Textencoder, andere VAE, andere Schrittzahl, andere Guidance, andere LoRAs. Das ist eine Änderung am Arbeitsablauf, nicht an einem Wert – in UB.Forge wird so ein Wechsel als Workflow-Änderung geführt, mit Plausibilitäts-Check, ob Auflösung, Schritte und Speicher zur neuen Familie überhaupt passen. Diese Trennung klingt pedantisch, verhindert aber

genau den Abend, an dem niemand mehr weiß, warum die Bilder seit Dienstag anders aussehen. Dass ich das selbst so betreibe und nicht nur empfehle, steht in [Dogfooding: Ich setze ein, was ich berate](#).

Was diese Liste nicht leisten kann

Ehrlichkeit gehört dazu. Diese Liste ist eine Momentaufnahme, und zwei Sorten von Aussagen darin haben unterschiedliche Halbwertszeiten. Die Familien, ihre Rollen und die Lizenzlager sind stabil – daran wird sich in den nächsten Jahren wenig ändern. Die konkreten Versionsnummern, die Speicherwerte und die Antwort auf „welches ist gerade das beste“ können in sechs Wochen anders lauten. Genau deshalb steht hier keine Punktetabelle.

Zweitens kann diese Liste nicht die Qualitätsfrage für deinen Anwendungsfall beantworten. Benchmarks messen Prompt-Treue an Testsätzen, die mit deinen Motiven nichts zu tun haben. Und drittens ersetzt sie keine Rechtsprüfung. Sie zeigt dir, wo du hinschauen musst – nicht, was dein Justiziar dazu sagt. Für die strategische Ebene, also die Frage, ob und wie ein Unternehmen überhaupt lokal produziert, ist [KI-Bildgenerierung im Unternehmen – Die strategische Alternative zur Cloud](#) der passende Einstieg; die technischen Grundlagen der Verfahren stehen in [Basiswissen KI-Bildgenerierung: Wie Diffusion funktioniert](#).

Fazit

Fünfzehn Modelle, fünf Familien, drei Lizenzlager – und eine Reihenfolge, in der du die Fragen stellen solltest. Erstens Lizenz: Dürfen die Bilder Geld verdienen? Damit fällt die halbe Liste weg, und zwar ausgerechnet die qualitativ stärkste Hälfte. Zweitens Videospeicher: Was läuft überhaupt auf deiner Karte, ohne dass du für jedes Bild Kaffee holen gehst? Damit fällt der Rest auf eine Handvoll zusammen. Drittens Ziel: Fotorealismus, Text, Bearbeitung, Illustration, Tempo oder eigenes Training? Damit steht dein Modell.

Wenn du nur eine Empfehlung mitnehmen willst: Fang mit Qwen-Image an, wenn du 16 Gigabyte Videospeicher hast, und mit einem bewährten SDXL-Ableger, wenn du acht hast. Beide lassen dich kommerziell arbeiten, beide sind gut dokumentiert, und beide bringen dich weit genug, um herauszufinden, was dir tatsächlich fehlt. Flux nimmst du dazu, wenn du entweder bei privaten Projekten bist oder bereit, eine Lizenz zu kaufen – und dann in der [klein]-Variante mit 4 Milliarden Parametern, die ohnehin kostenlos kommerziell nutzbar ist.

Und nimm die drei Zahlen ernst, die über allem stehen: drei Modelle im Verzeichnis, zwei Lizenzen pro Download geprüft, ein Katalog, in dem steht, was womit entstanden ist. Das ist weniger spannend als eine neue Modellgeneration, aber es ist der Unterschied zwischen einer Bildproduktion und einem Hobby mit Nebenwirkungen.

Häufige Fragen

Welches lokale Modell ist das beste?

Es gibt kein bestes Modell, nur ein passendes. Für maximale Prompt-Treue und Text im Bild steht FLUX.2 [dev] an der Spitze, ist aber kommerziell lizenzpflichtig und speicherhungrig. Für die Kombination aus guter Qualität, freier Lizenz und Textfähigkeit ist Qwen-Image die vernünftigste Wahl. Für das größte Werkzeug-Ökosystem und den geringsten Speicherbedarf bleibt es SDXL. Und für Tempo bei brauchbarer Qualität Z-Image Turbo. Die Frage nach dem besten Modell ist in der Praxis immer eine Frage nach Lizenz, Speicher und Ziel.

Welches Modell nehme ich als Einstieg?

Mit 16 Gigabyte Videospeicher oder mehr: Qwen-Image. Es steht unter Apache 2.0, du läufst in keine Lizenzfalle, und es kann als einziges offenes Modell verlässlich Schrift. Mit 8 bis 12 Gigabyte:

ein bewährter SDXL-Ableger wie Juggernaut XL, weil du für jedes Problem innerhalb von Minuten eine Anleitung findest. Mit weniger: FLUX.2 [klein] 4B als GGUF oder Z-Image Turbo quantisiert. Nimm genau ein Modell und bleib zwei Wochen dabei, bevor du sammelst.

Was ist der Unterschied zwischen SDXL, Flux und Qwen-Image?

Drei Architekturen mit drei Profilen. SDXL ist das genügsame Arbeitstier mit 2,6 Milliarden Parametern, läuft ab 8 Gigabyte, hat das größte Ökosystem und bei Prompt-Treue wie Schrift die größten Schwächen. Flux ist die Qualitätsspitze, technisch ein Transformer-Aufbau, reicht von 4 bis 32 Milliarden Parametern und hat bei den starken Varianten das Lizenzproblem. Qwen-Image ist mit 20 Milliarden Parametern der Textspezialist unter Apache 2.0 und die pragmatische Mitte. LoRAs und Steuerwerkzeuge sind zwischen den drei Familien nicht austauschbar.

Warum gibt es so viele Modelle mit ähnlichen Namen?

Weil die meisten Downloads keine eigenen Modelle sind, sondern Fine-Tunes und Merges – Ableger eines Basismodells, die jemand auf ein Ziel weitertrainiert oder mit anderen verrechnet hat. Juggernaut XL, RealVis XL, DreamShaper XL, Illustrrious, NoobAI und die Mehrzahl der Anime-Checkpoints sitzen alle auf SDXL. Deshalb ist die Basis die wichtigere Information als der Name. Was genau in so einer Datei steckt, erklärt der Beitrag zu Checkpoints, Modellen und VAE.

Darf ich mit diesen Modellen kommerziell arbeiten?

Das hängt am konkreten Modell und ist die Frage, die vor dem Download geklärt sein muss. Ohne Auflagen kommerziell nutzbar sind Qwen-Image und Qwen-Image-Edit, Z-Image, Chroma1-HD, FLUX.1 [schnell] und FLUX.2 [klein] 4B – alle unter Apache 2.0. Mit Auflagen nutzbar sind die Stable-Diffusion-Modelle, bei SD 3.5 mit einer Umsatzgrenze von einer Million US-Dollar. Nicht ohne gekaufte Lizenz nutzbar sind FLUX.2 [dev], FLUX.2 [klein] 9B, FLUX.1 [dev] und FLUX.1 Krea [dev]. Fine-Tunes erben die Bedingungen ihrer Basis, also immer beide Lizenzen prüfen.

Brauche ich für Flux eine teure Grafikkarte?

Für FLUX.2 [dev] ja – unter 24 Gigabyte Videospeicher wird es auch quantisiert unangenehm, in voller Genauigkeit reden wir über 64 Gigabyte. Für FLUX.1 reichen 12 Gigabyte in FP8 gut aus, quantisiert geht es ab etwa 7. Und FLUX.2 [klein] 4B läuft mit rund 13 Gigabyte in voller Genauigkeit, quantisiert ab 8. Die Aussage „Flux braucht eine teure Karte“ stimmt also nur noch für die Spitze der Familie.

Lohnt sich Stable Diffusion 1.5 heute noch?

Nur in einem Fall: wenn deine Karte weniger als 6 Gigabyte Videospeicher hat. Dann ist SD 1.5 die einzige Familie, die noch vernünftig läuft, und für schnelle Kompositionsskizzen taugt sie weiter. Qualitativ ist sie nicht konkurrenzfähig, die 512-Pixel-Basis merkt man jedem Bild an, und selbst auf kleinen Karten sind stark quantisierte neuere Modelle inzwischen oft die besseren Kandidaten.

Ist Stable Diffusion 3.5 eine Empfehlung?

Meistens nicht. Es ist besser als SDXL bei Prompt-Treue und Schrift, braucht aber mehr Speicher, liegt qualitativ hinter Flux und Qwen-Image, hat ein dünnes Ökosystem und ist das einzige Modell in dieser Liste mit einer Umsatzgrenze in der Lizenz. Wer schon eine gute SDXL-Strecke hat, gewinnt wenig; wer neu anfängt, greift besser direkt zu Qwen-Image oder FLUX.2 [klein].

Welches Modell kann Text im Bild?

Qwen-Image ist hier das offene Referenzmodell und schreibt auch längere Beschriftungen zuverlässig, in lateinischer und chinesischer Schrift. FLUX.2 ist gut, Z-Image für kurze Zeilen brauchbar. SDXL und seine Ableger können es strukturell nicht, und daran ändert kein Prompt

etwas. Für alles, was wirklich korrekt sein muss – Preise, Namen, Rechtstexte – gilt trotzdem: Fläche generieren, Text hinterher im Layoutprogramm setzen.

Kann ich LoRAs zwischen den Familien tauschen?

Nein. Ein LoRA ist an die Architektur gebunden, auf der es trainiert wurde. SDXL-LoRAs laufen nicht mit Flux, Flux-LoRAs nicht mit Qwen-Image, und innerhalb der Anime-Welt sind Pony- und Illustrious-LoRAs ebenfalls nicht beliebig austauschbar, obwohl beide auf SDXL sitzen – die Prompt-Konventionen unterscheiden sich zu stark. Wenn du eine Familie wechselst, wechselst du deinen ganzen Zubehörbestand mit.

Wie erkenne ich, ob ein Modell überhaupt auf meine Karte passt?

Grobe Rechnung: Milliarden Parameter mal zwei ergibt den Speicherbedarf in Gigabyte bei voller Genauigkeit, mal eins bei FP8, mal 0,6 bei einer Vier-Bit-Quantisierung. Dazu ein bis zwei Gigabyte für Textencoder, VAE und Arbeitsspeicher der Bildberechnung. Ein 20-Milliarden-Modell will also rund 40 Gigabyte voll, 20 in FP8 und 14 quantisiert. Das sind Richtwerte, aber sie liegen erfahrungsgemäß näher an der Wahrheit als die Angaben in Ankündigungen.

Sollte ich lokal arbeiten oder einen Dienst nutzen?

Das ist eine Architekturentscheidung und keine Modellfrage. Lokal gewinnst du Kontrolle über Modellversion, Reproduzierbarkeit und Datenfluss, zahlst dafür mit Hardware und Betriebsaufwand. Ein Dienst nimmt dir den Betrieb ab, entscheidet dafür selbst, wann sich das Modell unter deinen Bildern ändert. In der Praxis läuft es häufig gemischt: Wegwerfbilder über einen Dienst, alles was veröffentlicht oder archiviert wird über die eigene Strecke.

WEITERLESEN

- › [KI-Bildgenerierung – der Überblick](#) – das gesamte Themenfeld im Zusammenhang
- › [Checkpoints, Modelle und VAE](#) – was in so einer Modelldatei eigentlich steckt
- › [Was ist ein LoRA?](#) – eigene Stile und Motive ins Modell holen
- › [Prompt und Negativ-Prompt in der KI-Bildgenerierung](#) – warum Flux und SDXL verschiedene Sprachen sprechen
- › [Sampling-Steps erklärt](#) – warum destillierte Modelle mit vier Schritten auskommen
- › [Sampler & Scheduler erklärt](#) – die Einstellungen, die je Familie anders wollen
- › [CFG-Scale: Guidance in KI-Bildgeneratoren](#) – der Regler, den Flux nicht kennt
- › [Der Seed bei der KI-Bildgenerierung](#) – die Grundlage jedes fairen Modellvergleichs
- › [Auflösung & Seitenverhältnis bei KI-Bildern](#) – warum jede Familie ihre Lieblingsgröße hat
- › [Cloud vs. lokale KI-Bildgenerierung](#) – die Entscheidung vor der Modellwahl
- › [Aufbau einer skalierbaren Bild-Maschine mit ComfyUI und Flux](#) – wie die Produktionsstrecke praktisch aussieht
- › [Digital Asset Management für KI-Bilder mit SharePoint Online und Power Platform](#) – Modellkatalog und Nachweis an einem Ort
- › [EU AI Act: Was seit dem 2. August 2026 gilt](#) – Kennzeichnungspflichten im Überblick
- › [Schulung KI-Bildgenerierung](#) – Modellwahl und Prompt-Handwerk am eigenen Material üben
- › [Beratung zur KI-Bildgenerierung](#) – Lizenzlage, Hardware und Betriebsmodell gemeinsam klären