

Cloud vs. lokale KI-Bildgenerierung



Zwei Wege zum selben Bild – und ein Unterschied in fast allem, was drumherum passiert. 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.	Architektur, Betriebsmodell und Compliance für deinen konkreten Fall.	Vom Prompt bis zur reproduzierbaren Produktionsstrecke, praxisnah.

Bevor du dich mit Prompts, Reglern und Modellen beschäftigst, steht eine Frage im Weg, die du nicht umgehen kannst: Soll das Bild bei einem Anbieter im Netz entstehen oder auf einer Maschine, die dir gehört? Die Antwort klingt nach einer Geschmacksfrage, ist aber eine Architekturentscheidung. Sie bestimmt, wer deine Daten sieht, wer über die Modellversion entscheidet, wer die Kennzeichnung erzeugt und wem du erklären musst, warum ein Bild vor sechs Monaten so aussah, wie es aussah.

Cloud vs. lokale Bildgenerierung ist deshalb keine Glaubensfrage und auch kein Wettkampf mit Sieger. Es ist eine Abwägung über fünf Achsen: Kontrolle, Reproduzierbarkeit, Datenschutz, Kosten und Aufwand. Wer sie ehrlich durchrechnet, kommt oft zu einem Ergebnis, das keine der beiden reinen Varianten ist – sondern eine kontrollierte GPU-Instanz dazwischen. Und wer unter Nachweispflichten arbeitet, merkt schnell: lokal ist dort nicht der Kompromiss, den man eingeht, weil das Budget fehlt. Lokal ist die Kür.

Dieser Beitrag gehört zum Themenfeld [KI-Bildgenerierung – der Überblick](#). Dort findest du die Einordnung aller Bausteine; hier geht es ausschließlich um die Betriebsmodellfrage – mit Zahlen, Lizenzen und Pflichten, Stand September 2026.

FAKTEN · Was sich gerade verschoben hat

Zwei Entwicklungen haben die Rechnung 2026 durcheinandergebracht. Erstens sind Bilder aus der Cloud spürbar billiger geworden: Ein Megapixel aus FLUX.2 [pro] kostet bei Black Forest Labs ab etwa 0,03 US-Dollar, FLUX.2 [klein] ab rund 0,014 US-Dollar, Google Imagen 4 Fast rund 0,02 US-Dollar, die kleinen GPT-Image-Varianten starten unter einem Cent. Zweitens ist lokale Hardware teurer geworden statt günstiger: Die RTX PRO 6000 Blackwell mit 96 GB kam im März 2025 mit einem Listenpreis von 8.565 US-Dollar und stand im September 2026 bei rund 16.000 US-Dollar – knapp 90 Prozent mehr, getrieben von der Knappheit bei GDDR7-Speicher. Auch die RTX 5090 mit 32 GB, offiziell für 1.999 US-Dollar angekündigt, wird real deutlich darüber gehandelt. Die Kostenachse kippt damit klarer zur Cloud – was die anderen Achsen nur umso wichtiger macht.

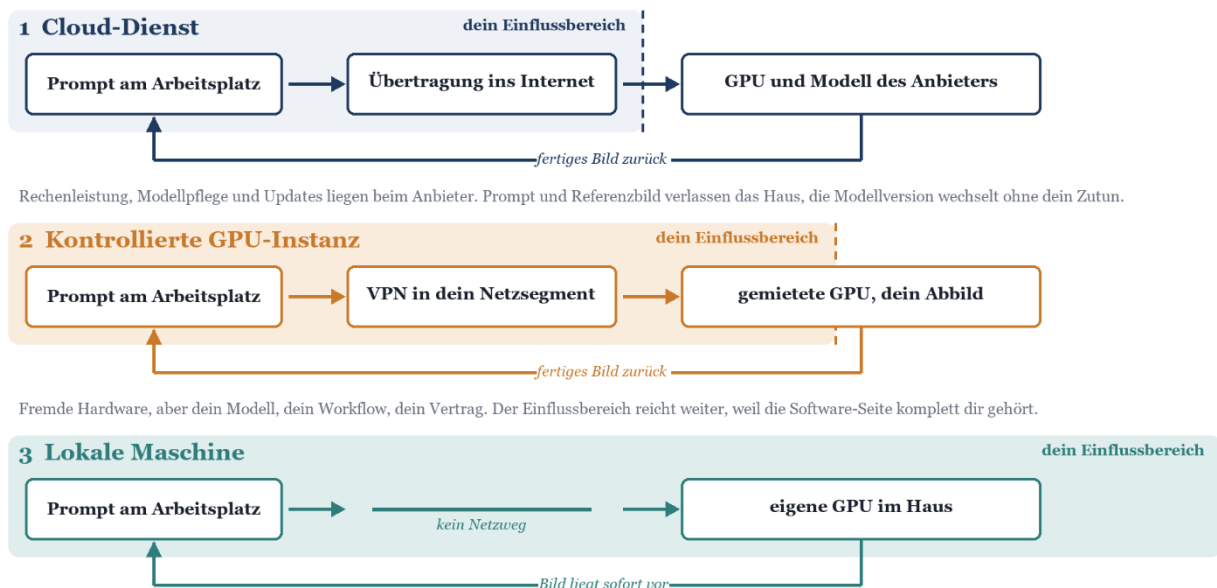
Der eine Unterschied – und die drei Betriebsmodelle, die daraus folgen

Technisch ist die Sache banal: Es geht darum, auf wessen Grafikkarte die Rechnung läuft. Bei einem Cloud-Dienst tippst du deinen Prompt in eine Weboberfläche oder schickst ihn per Schnittstelle los, die GPU des Anbieters rechnet, und du bekommst das Bild zurück. Lokal lädt deine Maschine das Modell in den Videospeicher und rechnet selbst. Prompt, Referenzbilder und Ergebnis bleiben, wo sie sind.

Aus diesem einen Unterschied folgt alles andere. Wer die Rechenleistung stellt, bestimmt die Kostenstruktur. Wer die Daten entgegennimmt, ist datenschutzrechtlich ein Dritter. Wer das Modell betreibt, entscheidet, wann es getauscht wird – und damit, ob dein Bild von letztem Quartal heute noch reproduzierbar ist.

Wo läuft die Rechnung – und wo liegen die Daten?

Die getonte Fläche ist dein Einflussbereich. Alles rechts davon gehört jemand anderem.



Rechenleistung, Modellpflege und Updates liegen beim Anbieter. Prompt und Referenzbild verlassen das Haus, die Modellversion wechselt ohne dein Zutun.

Fremde Hardware, aber dein Modell, dein Workflow, dein Vertrag. Der Einflussbereich reicht weiter, weil die Software-Seite komplett dir gehört.

Kein Empfänger, keine Übermittlung, kein Vertrag über Auftragsverarbeitung. Dafür gehört dir auch das Problem, wenn die Karte stirbt.

Ein technischer Unterschied – wessen Hardware rechnet – erzeugt jede weitere Differenz bei Kosten, Kontrolle und Nachweis.

Skizze 1: Derselbe Prompt, drei Datenwege. Entscheidend ist nicht die Bildqualität am Ende, sondern welche Grenze die Daten unterwegs überschreiten.

Drei Betriebsmodelle, nicht zwei

Die Debatte wird meistens als Entweder-oder geführt, und genau daran scheitert sie. In der Praxis gibt es drei Modelle, und das mittlere ist das, bei dem die meisten Projekte landen:

Cloud-Dienst. Fertige Oberfläche oder Schnittstelle, Abrechnung pro Bild oder pro Monat. Keine Hardware, keine Installation, keine Modellpflege. Du bekommst Qualität ab Werk und gibst dafür Kontrolle ab.

Kontrollierte GPU-Instanz. Du mietest eine Maschine mit Grafikkarte, installierst dein eigenes Abbild darauf – typischerweise ComfyUI mit deinen Modellen und Workflows – und rechnest dort. Die Hardware gehört jemand anderem, die Software-Seite gehört komplett dir. Wie so ein Aufbau aussieht, beschreibt der [Aufbau einer skalierbaren Bild-Maschine mit ComfyUI und Flux](#) im Detail.

Lokale Maschine. Eigene Karte im eigenen Netz. Maximale Kontrolle, maximale Verantwortung. Kein Netzweg, kein Vertrag über Auftragsverarbeitung, keine Ausrede, wenn das Netzteil stirbt.

Die Unterscheidung ist wichtig, weil der mittlere Weg das häufigste Missverständnis auflöst. „Gemietete GPU“ wird oft mit „Cloud“ in einen Topf geworfen, und dann heißt es, der Datenschutzvorteil sei ohnehin weg. Das ist zu grob. Ob eine gemietete Instanz tragfähig ist, hängt an drei Punkten: Steht ein Vertrag über Auftragsverarbeitung mit dem Betreiber? Ist der Standort der Instanz vertraglich festgelegt? Und ist das Logging so konfiguriert, dass der Betreiber deine Prompts nicht mitschreibt? Sind die drei geklärt, ist eine kontrollierte Instanz näher an „lokal“ als an „öffentlicher Dienst“. Sind sie nicht geklärt, ist sie das Schlechteste aus beiden Welten: fremde Hardware ohne Nutzungskomfort.

TIPP · Der Satz, der die Diskussion abkürzt

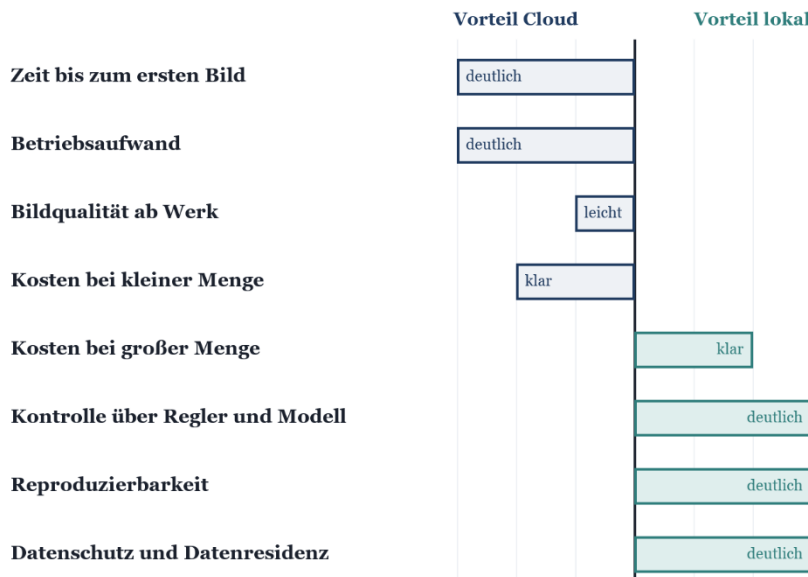
Frag im Projekt nicht „Cloud oder lokal?“, sondern: „Welche Inhalte gehen in den Prompt, und welche in ein Referenzbild?“ In dem Moment, in dem jemand „unser noch nicht angekündigtes Produkt“ oder „ein Foto von einer Kollegin“ sagt, ist die Architekturfrage entschieden – ohne dass ihr über Preise reden musstet.

Die Abwägung Achse für Achse

Je nachdem, was im Projekt hart ist, kippt die Waage. Es gibt keine Gesamtpunktzahl, auf die man sich einigen könnte; es gibt nur die Achse, die im konkreten Fall weh tut. Die folgenden acht sind die, um die es in Beratungsgesprächen tatsächlich geht.

Acht Achsen, kein Pauschalsieger

Ausschlag nach links: Vorteil Cloud. Ausschlag nach rechts: Vorteil lokal.



Die Entscheidung fällt nicht auf der Summe aller Achsen, sondern auf der einen, die im konkreten Projekt hart ist.

Wer unter Nachweispflicht arbeitet, gewichtet Reproduzierbarkeit und Datenresidenz mit dem Faktor drei – dann ist die Sache entschieden.

Skizze 2: Auf vier Achsen liegt die Cloud vorn, auf vier die lokale Lösung. Wer nur die Mehrheit zählt, entscheidet falsch – es kommt auf die Gewichtung an.

Einfachheit und Zeit bis zum ersten Bild

Hier gewinnt die Cloud, und zwar nicht knapp. Anmelden, Prompt eintippen, Bild ansehen – das sind Minuten. Eine lokale Installation kostet dich einen halben bis ganzen Tag, bis ComfyUI läuft, die Modelle im richtigen Ordner liegen, die Treiber passen und du verstanden hast, warum der erste Durchlauf so lange dauert. Das ist kein Argument gegen lokal, sondern eine Zahl, die in die Kalkulation gehört: Deine Arbeitszeit ist der teuerste Posten in der ganzen Rechnung.

Der Aufwand endet auch nicht bei der Installation. Modelle wollen aktualisiert werden, Erweiterungen brechen nach Updates, und irgendwer muss entscheiden, ob der neue Treiber eingespielt wird. Bei einem Cloud-Dienst macht das der Anbieter – was praktisch ist, bis er es tut, ohne dich zu fragen.

Kontrolle und Feintuning

Lokal hast du jeden Regler in der Hand: [Der Seed bei der KI-Bildgenerierung](#), [CFG-Scale: Guidance in KI-Bildgeneratoren](#), [Sampler & Scheduler erklärt](#) und [Sampling-Steps erklärt](#) sind keine versteckten Voreinstellungen, sondern Werte in deinem Workflow. Dazu kommt das, was Cloud-Dienste fast nie anbieten: eigene Zusatzmodelle. Ein [LoRA](#) auf deine Produkte, dein Gesicht, deinen Bildstil trainiert, verändert die Ergebnisqualität für markengebundene Bilder mehr als jeder Promptkniff. Wie das systematisch geht, zeigt der Beitrag zu [markenkonsistenten KI-generierten Bildern](#).

Cloud-Dienste geben dir meist einen Ausschnitt: Seitenverhältnis, manchmal einen Seed, manchmal eine Stilwahl. Dafür liefern sie hohe Qualität, ohne dass du an Reglern drehen musst. Für Entwürfe und Einzelbilder ist das ein fairer Tausch. Für eine Produktionsstrecke, die tausend Bilder in einem definierten Look liefern soll, ist es zu wenig. Und wenn du Randfälle beherrschen willst – warum die [KI den Prompt ignoriert](#) oder warum die [KI keine Hände kann](#) – brauchst du Zugriff auf mehr als ein Eingabefeld.

Reproduzierbarkeit – die Achse, die im Geschäftsumfeld zählt

Das ist der Punkt, an dem die meisten Projekte hängen bleiben, und zwar spät. Reproduzierbarkeit heißt: Du kannst in achtzehn Monaten dasselbe Bild noch einmal erzeugen – weil eine Kampagne verlängert wird, weil ein Motiv in einer anderen Auflösung gebraucht wird, oder weil jemand fragt, wie dieses Bild entstanden ist.

Lokal ist das Buchhaltung: Modelldatei plus Prüfsumme, VAE, Seed, Sampler, Scheduler, Steps, CFG-Wert, Auflösung, Workflow-Version. Alles davon liegt bei dir, alles davon kannst du einfrieren. Die Modellversion ändert sich genau dann, wenn du sie änderst – der Zusammenhang zwischen [Checkpoints, Modellen und VAE](#) und dem Ergebnis bleibt stabil.

Bei einem Cloud-Dienst weißt du oft nicht, welche Modellversion gerade rechnet. Anbieter tauschen Gewichte, verbessern Sicherheitsfilter, ändern Standardparameter – meist ohne Ankündigung und ohne Versionsnummer im Antwortobjekt. Selbst wenn du denselben Seed zurückschicken kannst: Ein anderer Modellstand mit demselben Seed ergibt ein anderes Bild. Das ist keine Unterstellung, sondern die Natur eines gepflegten Dienstes. Wer feste Versionen braucht, muss sie vertraglich zusichern lassen – oder das Modell selbst halten. Wie stark schon kleine Unterschiede in der Ausführung wirken, zeigt [batch size und Reproduzierbarkeit](#).

WARNUNG · Reproduzierbarkeit merkt man erst, wenn sie fehlt

Ein Kunde hatte eine Bildreihe für einen Produktkatalog über einen Cloud-Dienst erzeugt, sechs Monate später kamen vier neue Varianten dazu. Prompt identisch, Seed notiert, Ergebnis: anderer Look, andere Hauttöne, andere Lichtstimmung. Der Anbieter hatte das Modell zwischenzeitlich aktualisiert. Vier Bilder wurden zu einem Neuaufsatz der kompletten Reihe. Die Rechnung für diesen einen Vorfall lag deutlich über dem, was drei Jahre lokale Hardware gekostet hätten.

Datenschutz und Datenresidenz

Der Punkt, der im Unternehmen am schwersten wiegt und beim privaten Ausprobieren am leichtesten untergeht. Jeder Prompt an einen Cloud-Dienst ist eine Übermittlung an einen Dritten. Sobald personenbezogene Daten im Spiel sind – und ein Referenzfoto einer Kollegin ist personenbezogen, ein Gesicht sogar biometrisch relevant – brauchst du eine Rechtsgrundlage, einen Vertrag über Auftragsverarbeitung und Klarheit über den Verarbeitungsort. Bei vertraulichen Inhalten kommt hinzu, dass ein Prompt sehr gesprächig sein kann: „unser neues Gehäusedesign in Mattschwarz mit dem Anschluss an der Seite“ ist eine Produktankündigung, bevor es eine gibt.

Bei den Nutzungsbedingungen lohnt der Blick ins Detail, weil sich die Modelle stark unterscheiden. Für Schnittstellenzugriffe gilt bei OpenAI seit März 2023, dass API-Daten nicht zum Training verwendet werden, sofern du dem nicht ausdrücklich zustimmst; die Standardaufbewahrung liegt bei dreißig Tagen für die Missbrauchserkennung, und eine vollständige Nullaufbewahrung gibt es nur für geeignete Kunden nach Freigabe und nicht für alle Endpunkte. Bei den Endkundendiensten sieht es anders aus: Midjourney zeigt Generierungen in den Tarifen Basic und Standard in der öffentlichen Galerie, der Stealth-Modus ist erst ab Pro verfügbar, und auch dann bleiben die Trainingsrechte des Anbieters an Prompts und Bildern bestehen. Unternehmen mit mehr als einer Million US-Dollar Jahresumsatz brauchen dort ohnehin einen Pro- oder Mega-Tarif für die kommerzielle Nutzung.

WARNUNG · Die öffentliche Galerie ist kein Schönheitsfehler

Wer Kundenprojekte über einen Endkundentarif mit öffentlicher Galerie generiert, publiziert Prompts, Zwischenstände und Motive – nachlesbar, durchsuchbar, mit Zeitstempel. Ich habe in

einem Wettbewerbsumfeld erlebt, dass genau so ein Kampagnenmotiv drei Wochen vor dem Launch bekannt war. Niemand hatte etwas verraten; es stand einfach öffentlich im Werkzeug. Prüfe vor dem ersten Prompt, in welchem Tarif du sitzt und was er mit deinen Eingaben macht.

Lokal stellt sich die Frage nicht. Es gibt keinen Empfänger, keine Übermittlung, keinen Drittlandtransfer. Das ist der stärkste Einzelveil und der Grund, warum in regulierten Branchen die Diskussion meist kurz ist. Bei einer kontrollierten GPU-Instanz ist der Vorteil nicht weg, aber vertraglich abgesichert statt technisch erzwungen – ein Unterschied, den Datenschutzbeauftragte zu Recht sehen wollen.

Der Vergleich in einer Tabelle

Kriterium	Cloud-Dienst	Kontrollierte GPU-Instanz	Lokale Maschine
Zeit bis zum ersten Bild	Minuten	Stunden bis ein Tag	Halber bis ganzer Tag
Einstiegskosten	keine	niedrig, Miete pro Stunde	hoch, Karte und Rechner
Kosten pro Bild	fester Preis, skaliert mit Megapixeln	Mietzeit geteilt durch Bildzahl	nahezu null, Strom und Abschreibung
Regler und Parameter	Ausschnitt	vollständig	vollständig
Eigene LoRAs und Workflows	selten möglich	ja	ja
Modellversion fixierbar	meist nein	ja	ja
Reproduzierbarkeit über Jahre	eingeschränkt	gegeben	gegeben
Prompt verlässt das Haus	ja	ja, in dein Netzsegment	nein
Vertrag über Auftragsverarbeitung	erforderlich	erforderlich	nicht erforderlich
Kennzeichnung nach Artikel 50	meist vom Anbieter beigefügt	in der Pipeline selbst zu erzeugen	in der Pipeline selbst zu erzeugen
Betriebsaufwand laufend	keiner	gering bis mittel	mittel bis hoch
Skalierung nach oben	sofort und unbegrenzt	sofort, kostet Geld	durch Hardware begrenzt
Verfügbarkeit ohne Netz	nein	nein	ja

Tabelle 1: Die drei Betriebsmodelle im direkten Vergleich. Die Zeilen zu Reproduzierbarkeit, Datenweg und Kennzeichnung sind die, an denen Projekte tatsächlich scheitern.

Was es kostet – mit Zahlen, Stand September 2026

Der alte Merksatz „wenig generieren, dann Cloud; viel generieren, dann lokal“ stimmt weiterhin im Prinzip, aber die Schwelle hat sich verschoben, und zwar nach oben. Das liegt an zwei Bewegungen, die in dieselbe Richtung zeigen: Cloud-Bilder sind billiger geworden, lokale Hardware teurer.

Cloud: Abrechnung nach Megapixeln, nicht nach Bildern

Eine Umstellung, die viele Kalkulationen still kaputt macht: Die Abrechnung erfolgt bei FLUX.2 nach Megapixeln der Ausgabe, nicht pauschal pro Bild. Ein Bild mit 1024 mal 1024 Punkten ist etwa ein Megapixel und kostet bei FLUX.2 [pro] rund 0,03 US-Dollar. Verdoppelst du die Kantenlänge, vervierfachst du die Fläche – und den Preis. Bei der Bildbearbeitung zählen Eingabebilder mit, weshalb Editieren teurer startet als Neugenerierung. Wer also

gewohnheitsmäßig in maximaler Auflösung generiert, weil „groß ist besser“, zahlt das Vierfache für Bilder, die am Ende auf eine Webseite mit 1.200 Punkten Breite wandern. Die Grundlagen dazu stehen unter [Auflösung & Seitenverhältnis bei KI-Bildern](#).

Dienst / Modell	Preis ab (1 Megapixel)	Charakter
FLUX.2 [klein] 4B	rund 0,014 US-Dollar	schnell, günstig, für Masse und Iteration
FLUX.2 [klein] 9B	rund 0,015 US-Dollar	etwas mehr Qualität bei ähnlichem Preis
FLUX.2 [pro]	rund 0,03 US-Dollar	Arbeitspferd, Bearbeitung ab etwa 0,045 US-Dollar
FLUX.2 [flex]	rund 0,05 US-Dollar	stabiler Durchsatz, stark bei Text im Bild
FLUX.2 [max]	rund 0,07 US-Dollar	höchste Stufe der Familie
FLUX.1 Kontext [pro]	rund 0,04 US-Dollar	Vorgängergeneration, bildbezogene Bearbeitung
Google Imagen 4 Fast	rund 0,02 US-Dollar	günstige Stufe mit gutem Verhältnis
Google Imagen 4 Standard	rund 0,04 US-Dollar	mittlere Stufe
Google Imagen 4 Ultra	rund 0,06 US-Dollar	obere Stufe
GPT-Image, kleine Varianten	ab unter einem Cent	sehr günstig, für Entwürfe
GPT-Image, Qualitätsstufen	rund 0,10 US-Dollar	Premiumsegment

Tabelle 2: Listenpreise für ein Megapixel, Stand September 2026, gerundet und ohne Mengenrabatte. Stapelverarbeitung über Batch-Schnittstellen halbiert bei mehreren Anbietern den Preis. Dass diese Tabelle in sechs Monaten anders aussieht, ist Teil des Punkts.

Für eine mittlere Marketingabteilung mit, sagen wir, fünfhundert produktiven Bildern im Monat und dem üblichen Faktor fünf an Verwürfen bedeutet das: rund 2.500 Generierungen, bei drei Cent pro Stück unter hundert US-Dollar im Monat. Das ist weniger, als die Diskussion darüber im Meeting kostet. Genau deshalb ist das Kostenargument allein selten der Grund für eine lokale Lösung.

Lokal: Hardware, Strom und der Speicher-Engpass

Die Eintrittshürde für lokales Rechnen ist der Videospeicher. Reicht er nicht, läuft ein Modell gar nicht oder es lagert ständig aus und wird quälend langsam. Die Bandbreite ist dabei größer, als viele denken: Ein schnelles, aktuelles Modell wie Z-Image Turbo läuft in voller Präzision komfortabel auf 16 GB, in einer FP8-Fassung in etwa 8 GB und als quantisierte GGUF-Variante schon in rund 6 GB – mit acht Rechenschritten und fünf bis zehn Sekunden pro Bild auf einer Consumer-Karte mit 16 GB. Am anderen Ende steht FLUX.2 [dev] mit 32 Milliarden Parametern: Die FP8-Fassung belegt rund 32 GB und verlangt damit eine Karte mit 32 GB oder mehr, auf 24 GB läuft sie nur mit Auslagerung; eine Q4-Quantisierung drückt das auf etwa 19 GB und passt damit auf eine 24-GB-Karte. Welches Modell was verlangt, ordnet [Die 15 wichtigsten Modelle zur lokalen Bildgenerierung](#) ein.

Die unangenehme Nachricht von 2026 betrifft die Preise. Durch die Knappheit bei GDDR7-Speicher sind Grafikkarten mit viel Videospeicher deutlich teurer geworden. Die RTX PRO 6000 Blackwell mit 96 GB startete im März 2025 bei 8.565 US-Dollar Listenpreis und wurde im September 2026 mit rund 16.000 US-Dollar geführt – ein Aufschlag von fast 90 Prozent in achtzehn Monaten. Bei der RTX 5090 mit 32 GB liegen Listenpreis und Straßenpreis weit auseinander. Wer eine lokale Strecke plant, rechnet also nicht mehr mit fallenden Hardwarepreisen, sondern kalkuliert den heutigen Preis und kauft, wenn verfügbar.

Der Strom ist dagegen ein Nebenposten, auch wenn er in Diskussionen gerne groß gemacht wird. Bei Gewerbestrompreisen um 25 bis 27 Cent je Kilowattstunde, wie sie 2026 im Bundesdurchschnitt für neue Verträge gelten, und einer Karte, die unter Vollast 450 Watt zieht, kostet ein Bild mit zwanzig Sekunden Rechenzeit weniger als ein Zehntelcent. Selbst zehntausend Bilder im Monat bleiben bei rund sieben Euro Stromkosten. Der Posten, der wirklich zählt, ist die Arbeitszeit.

Kostenposten	Cloud-Dienst	Kontrollierte GPU-Instanz	Lokale Maschine
Anschaffung	—	—	2.000 bis 16.000 Euro je nach Speicher
Laufende Grundkosten	keine oder Abo	rund 0,35 bis 3,30 US-Dollar je Stunde	Abschreibung, Strom, Platz
Verbrauch pro Bild	0,014 bis 0,10 US-Dollar je Megapixel	Mietzeit, bei Spot-Instanzen 50 bis 70 Prozent billiger	unter einem Zehntelcent Strom
Einrichtung einmalig	Minuten	wenige Stunden je Abbild	halber bis ganzer Tag
Pflege laufend	keine	Abbild aktuell halten	Treiber, Modelle, Erweiterungen
Versteckter Posten	Preisänderung des Anbieters	Datenabfluss bei falschem Logging	Kapitalbindung und Alterung

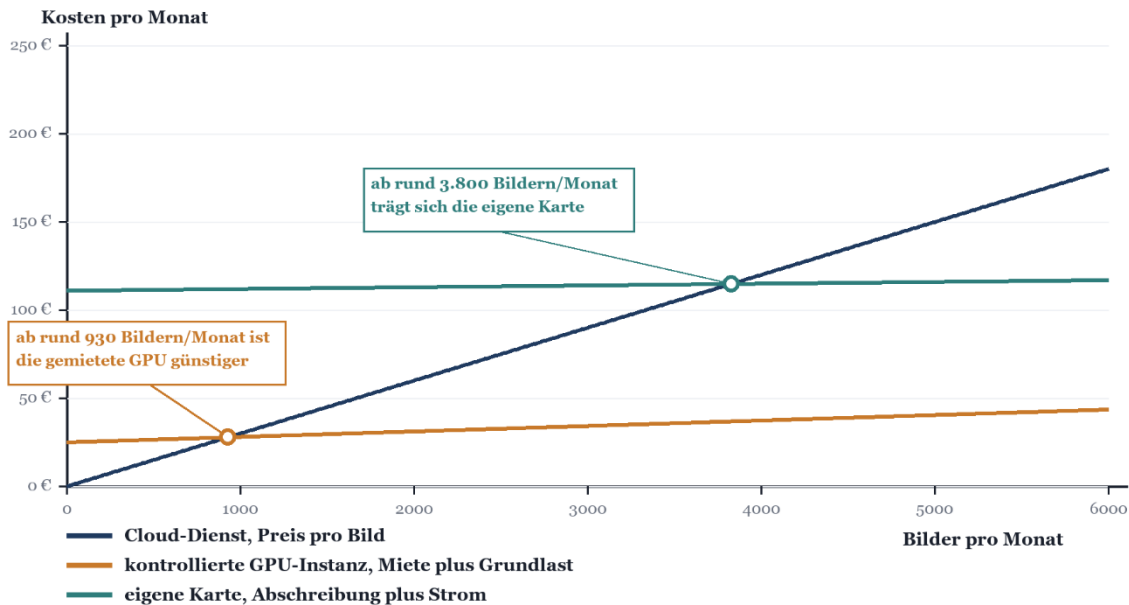
Tabelle 3: Kostenstruktur statt Kostenhöhe. Die letzte Zeile ist die, die in Wirtschaftlichkeitsrechnungen regelmäßig fehlt.

Wo sich was rechnet

Rechnet man das durch, ergeben sich zwei Schwellen statt einer. Bei einem Cloud-Preis von drei Cent pro Bild, einer gemieteten Karte für etwa 55 Cent pro Stunde bei rund 180 Bildern in der Stunde und einer eigenen Karte für 4.000 Euro über drei Jahre liegt die erste Schwelle bei knapp tausend Bildern im Monat – ab dort ist die kontrollierte Instanz billiger als der Dienst. Die zweite liegt bei etwa 3.800 Bildern im Monat, dann trägt sich eigene Hardware gegenüber der Cloud. Bis die eigene Karte auch die Miete unterbietet, braucht es ein Volumen jenseits von zwanzigtausend Bildern im Monat.

Wo sich was rechnet – Modellrechnung, keine Preisliste

Annahmen: 0,03 € pro Bild in der Cloud · gemietete GPU 0,55 €/Stunde bei rund 180 Bildern/Stunde · eigene Karte 4.000 € auf 36 Monate



Nicht eingezeichnet: deine Arbeitszeit. Eine Stunde Einrichtung entspricht in dieser Rechnung rund 5.000 Cloud-Bildern.

Skizze 3: Zwei Schwellen, nicht eine. Die Zahlen sind eine Modellrechnung mit offen gelegten Annahmen – setz deine eigenen Werte ein, die Form der Kurven bleibt gleich.

WICHTIG · Die Kurven sagen nicht, was du tun sollst

Diese Rechnung zeigt nur die reinen Rechenkosten. Sie enthält nicht die Arbeitszeit für Einrichtung und Pflege, nicht den Wert einer eingefrorenen Modellversion und nicht das Risiko eines Datenabflusses. Eine einzige Stunde Beratungszeit entspricht in dieser Modellrechnung rund fünftausend Cloud-Bildern. Wer lokal ausschließlich mit dem Kostenargument begründet, hat in den meisten Fällen die schwächste Begründung gewählt, die zur Verfügung steht.

TIPP · Messen statt schätzen

Bevor du Hardware kaufst, protokolliere vier Wochen lang jede Generierung: Motiv, Auflösung, Anzahl der Verwürfe, verwendetes Modell. Fast immer ist das echte Volumen kleiner als die Schätzung, dafür die Auflösung größer als nötig. Beides verschiebt die Schwelle. Ein Katalog mit Parametern und Ergebnissen – bei mir übernimmt das UB.Dossier – liefert diese Zahlen als Nebenprodukt, ohne dass jemand eine Strichliste führen muss.

Lizenzen, Kennzeichnung und Nachweis

Hier wird es unangenehm, weil sich die Verantwortung genau umdreht. Auf den Achsen Kontrolle und Datenschutz gewinnt lokal. Bei Lizenz und Kennzeichnung liegt die Arbeit plötzlich bei dir.

Was die Lizenz des Modells erlaubt

„Offene Gewichte“ heißt nicht „frei verwendbar“. Der Unterschied, an dem sich in der Praxis alles entscheidet, ist der zwischen den Bildern und dem Modell. Bei FLUX.2 [dev] unter der nicht-kommerziellen Lizenz von Black Forest Labs darfst du die erzeugten Bilder kommerziell nutzen; das Modell selbst kommerziell zu betreiben oder als Bestandteil eines Produkts auszuliefern, verlangt eine bezahlte Lizenz. Dasselbe gilt für FLUX.2 [klein] mit 9 Milliarden Parametern. Wirklich unbeschränkt ist in dieser Familie nur die 4B-Variante von FLUX.2 [klein] unter Apache

2.0. Aus dem Alibaba-Umfeld sind Qwen-Image, ein Modell mit 20 Milliarden Parametern, und Z-Image Turbo ebenfalls unter Apache 2.0 verfügbar – was sie für Fälle interessant macht, in denen die Lizenz das Ausschlusskriterium ist.

Modell	Lizenz	Bilder kommerziell	Modell kommerziell betreiben
FLUX.2 [klein] 4B	Apache 2.0	ja	ja
FLUX.2 [klein] 9B	nicht-kommerzielle Lizenz	ja	nur mit bezahlter Lizenz
FLUX.2 [dev] 32B	nicht-kommerzielle Lizenz	ja	nur mit bezahlter Lizenz
Qwen-Image 20B	Apache 2.0	ja	ja
Z-Image Turbo	Apache 2.0	ja	ja
FLUX.2 [pro] / [flex] / [max]	nur als Dienst	ja, nach Nutzungsbedingungen	nicht vorgesehen, kein Download

Tabelle 4: Lizenzlage der wichtigsten Modelle, Stand September 2026. Die Spalte rechts ist die, die im Einkauf durchfällt, wenn niemand sie geprüft hat.

WARNUNG · Der Klassiker: Agentur nutzt dev-Modell für Kundenprojekte

Eine Agentur richtet eine lokale Strecke mit FLUX.2 [dev] ein und verkauft ihren Kunden „Bildproduktion aus eigener Infrastruktur“. Die Bilder darf sie abrechnen. Den Modellbetrieb als Dienstleistung anzubieten, deckt die nicht-kommerzielle Lizenz nicht. Der Unterschied ist juristisch fein und wirtschaftlich grob. Prüfe vor dem Aufbau, in welcher der beiden Rollen du bist – und wenn es der Modellbetrieb ist, kalkuliere die Lizenz ein oder nimm ein Apache-lizenziertes Modell. Das ist keine Rechtsberatung, sondern der Hinweis, wo genau du hinschauen musst.

Kennzeichnung nach Artikel 50 – und wer sie erzeugt

Seit dem 2. August 2026 gelten die Transparenzpflichten aus Artikel 50 der KI-Verordnung verbindlich. KI-erzeugte oder KI-veränderte Bilder müssen in vielen Fällen erkennbar als solche gekennzeichnet sein, und die Ausgaben generativer Systeme müssen maschinenlesbar als künstlich erzeugt markiert werden. Für Systeme, die vor dem 2. August 2026 in Verkehr gebracht wurden, gibt es für die maschinenlesbare Markierung nach Absatz 2 eine Übergangsfrist bis zum 2. Dezember 2026; die übrigen Transparenzpflichten gelten sofort. Der Bußgeldrahmen reicht bis zu 15 Millionen Euro oder drei Prozent des weltweiten Jahresumsatzes. Was insgesamt seit diesem Datum gilt, ordnet das [Briefing zum EU AI Act](#) ein.

Jetzt kommt der Teil, der in der Cloud-gegen-lokal-Debatte fast immer fehlt. Die technische Markierung ist Aufgabe des Anbieters des KI-Systems, die sichtbare Kennzeichnung im veröffentlichten Inhalt liegt bei dem, der das Bild einsetzt. Große Cloud-Dienste liefern die technische Seite mit: signierte Herkunftsmetadaten nach dem C2PA-Standard, teilweise zusätzlich ein unsichtbares Wasserzeichen wie SynthID, das im Bildinhalt selbst steckt und manche Bearbeitung übersteht. Ein lokal betriebenes Modell mit offenen Gewichten tut das nicht von allein. Die PNG-Datei, die aus deiner eigenen Strecke fällt, enthält keine Content Credentials, solange du sie nicht selbst hineinschreibst.

WICHTIG · Lokal heißt: Kennzeichnung ist dein Arbeitsschritt

Wer selbst betreibt, muss die maschinenlesbare Markierung in die eigene Pipeline einbauen – als Signatur der Herkunftsmetadaten beim Export, nicht als Textzeile im Dateinamen, die beim ersten

Zuschnitt verschwindet. Genau an dieser Stelle nutze ich UB.Forge: Der Export ist ein Schritt im Workflow, der Kennzeichnung und Plausibilitäts-Check erzwingt, bevor eine Datei den Arbeitsbereich verlässt. Ohne so einen Zwang passiert es garantiert, dass ein Bild ohne Markierung in die Veröffentlichung geht – nicht aus Böswilligkeit, sondern weil es Freitag nachmittag war. Woran man KI-Bilder überhaupt erkennt und wie Herkunftsprüfungen funktionieren, steht unter [KI-generierte Bilder erkennen](#).

Nachweis: der Katalog ist die eigentliche Investition

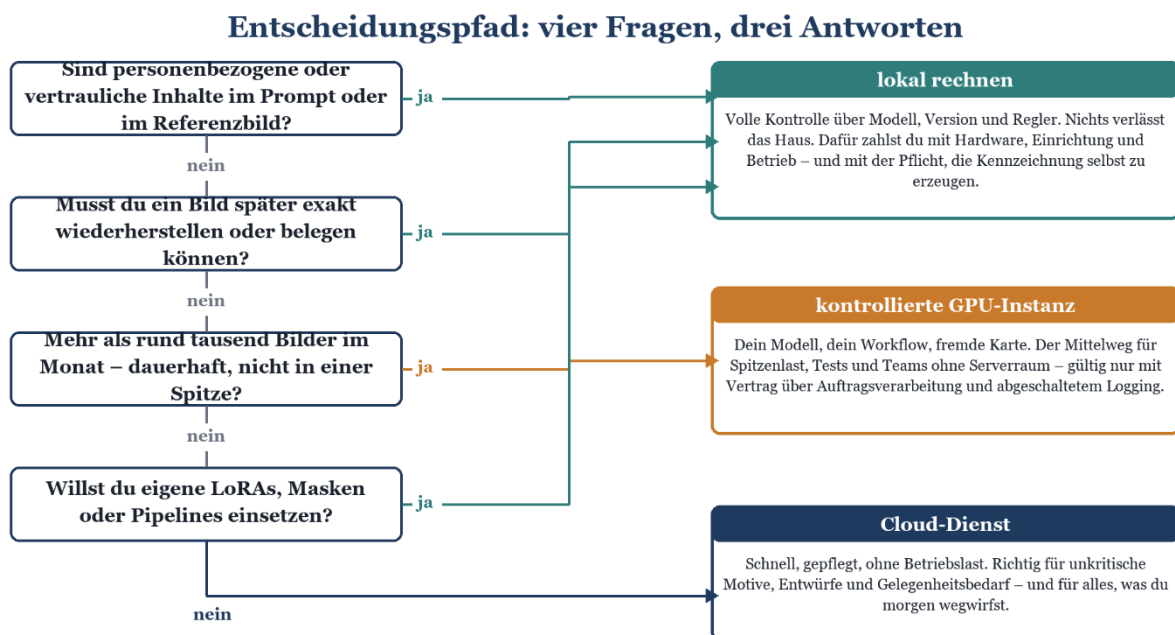
Wenn du die Betriebsmodellfrage sauber lösen willst, ist die Grafikkarte der einfachste Teil. Der schwierige Teil ist die Dokumentation: Welches Bild wurde wann mit welchem Modell, welchem Seed, welchen Reglern und welchem [Prompt und Negativ-Prompt](#) erzeugt, wer hat es freigegeben, wo wird es eingesetzt, und ist es gekennzeichnet? Ohne diese Zuordnung ist lokale Generierung nur ein Ordner voller PNG-Dateien – technisch souverän und organisatorisch wertlos.

Deshalb besteht ein tragfähiger Aufbau aus zwei Teilen: der Rechenstrecke und dem Katalog daneben. Bei mir hält UB.Dossier die Wertekataloge und den Nachweis, UB.Forge steuert Workflow-Änderungen und Produktion. Ob du dafür eigene Werkzeuge nimmst oder auf vorhandene Plattformen aufsetzt, ist eine Frage der Umgebung; wie das mit Microsoft-Mitteln aussieht, beschreibt [Digital Asset Management für KI-Bilder mit SharePoint Online und Power Platform](#). Entscheidend ist, dass der Nachweis nicht am Fleiß einzelner Personen hängt.

Dass ich das nicht nur empfehle, sondern selbst so betreibe, kannst du unter [Dogfooding: Ich setze ein, was ich berate](#) nachlesen – samt der Stellen, an denen es unbequem ist.

Entscheiden: vier Fragen, drei Antworten

In der Beratung stelle ich vier Fragen, und meistens ist die Sache nach der zweiten klar. Die Reihenfolge ist wichtig, weil die harten Kriterien vorne stehen. Kosten kommen zuletzt, nicht zuerst.



Ein einziges Ja bei den ersten beiden Fragen entscheidet die Sache – die Kostenfrage kommt erst danach.

Skizze 4: Der Entscheidungspfad. Ein einziges Ja bei den ersten beiden Fragen beendet die Diskussion – alles andere ist Feinabstimmung.

Die vier Fragen im Klartext

Erstens: Gehen personenbezogene oder vertrauliche Inhalte in den Prompt oder in ein Referenzbild? Wenn ja, ist lokal die naheliegende Antwort und eine kontrollierte Instanz mit ordentlichem Vertrag die zweitbeste. Ein öffentlicher Endkundendienst ist es nicht.

Zweitens: Musst du ein Bild später exakt wiederherstellen oder belegen können?

Kataloge, regulierte Kommunikation, Produktdokumentation, alles mit Freigabeprozess – dann brauchst du eine eingefrorene Modellversion. Die bekommst du bei einem gepflegten Dienst in der Regel nicht.

Drittens: Läuft dauerhaft ein Volumen jenseits von rund tausend Bildern im Monat?

Dann lohnt der Blick auf eine gemietete Karte, auch ohne Datenschutzdruck. Achtung: Es geht um dauerhaftes Volumen, nicht um die eine Kampagne, in der es kurz viel war.

Viertens: Willst du eigene LoRAs, Masken, Inpainting-Strecken oder mehrstufige Workflows?

Dann ist lokal oder kontrollierte Instanz gesetzt, weil die Regler in der Cloud fehlen. Wer nur hübsche Einzelbilder braucht, braucht das alles nicht.

Zweigleisig ist der Normalfall, nicht der Notbehelf

Die meisten funktionierenden Aufbauten sind hybrid, und das ist keine Unentschlossenheit, sondern eine Arbeitsteilung. Entwürfe, Moodboards, Ideenfindung und alles, was morgen im Papierkorb landet, laufen über einen Cloud-Dienst – schnell, billig, ohne Betriebslast. Alles, was veröffentlicht, archiviert, freigegeben oder mit Kundendaten berührt wird, läuft über die kontrollierte Strecke. Die Grenze verläuft nicht zwischen Bildtypen, sondern zwischen Wegwerfbildern und Aktenbildern.

Praktisch heißt das: zwei Eingänge, ein Katalog. Die Bilder aus beiden Wegen landen in derselben Ablage, mit derselben Kennzeichnung und denselben Pflichtfeldern. Wenn du dir das erste gemischte Setup baust, ist das der Teil, den du zuerst festlegen solltest – nicht welche Grafikkarte du kaufst.

Wie lokale Strecken strategisch eingeordnet werden, wenn es nicht um ein Team sondern um ein Unternehmen geht, beschreibt [KI-Bildgenerierung im Unternehmen – Die strategische Alternative zur Cloud](#). Für die Frage, welche Dienste es überhaupt gibt und was sie können, hilft die [Übersicht der wichtigsten KI-Bildgeneratoren](#), und wer das Funktionsprinzip dahinter sauber verstehen will, findet es im [Basiswissen KI-Bildgenerierung: Wie Diffusion funktioniert](#).

TIPP · Der pragmatische Mittelweg für Teams ohne Serverraum

Wenn Datenschutz drückt, aber niemand Hardware betreiben will, ist die kontrollierte GPU-Instanz die Antwort, die in der Debatte regelmäßig übersehen wird. Ein Abbild mit deinen Modellen und Workflows, eine gemietete Karte, abgeschaltetes Logging, fester Standort, Vertrag über Auftragsverarbeitung – und du bekommst Reproduzierbarkeit und Regler ohne Kapitalbindung. Für regulierte Umgebungen ist auch der Weg über eine Unternehmensplattform interessant, weil dort die Modelle im bestehenden Vertragsrahmen liegen; in Microsoft Foundry stehen etwa FLUX.2-Varianten und GPT-Image-Modelle im Katalog. Wie sich FLUX.2 direkt in Office-Werkzeugen nutzen lässt, zeigt das Briefing zu [Black Forest FLUX.2 in PowerPoint und Copilot](#).

Wenn du bei dieser Abwägung nicht allein bleiben willst: Genau das ist der Gegenstand meiner [Beratung zur KI-Bildgenerierung](#) – Betriebsmodell, Architektur, Lizenzlage und Nachweiskonzept in einer Sitzung durchgerechnet, statt in drei Monaten Trial-and-Error. Wenn dein Team die Strecke anschließend selbst fahren soll, ist die [Schulung KI-Bildgenerierung](#) der passende zweite

Schritt. Und wer die Bildgenerierung in eine breitere KI-Strategie einbetten will, findet den Rahmen unter [Claude im Unternehmen](#).

Woran dieser Vergleich bewusst nicht rührt

Zur Ehrlichkeit gehört, was hier nicht steht. Erstens: Die Qualitätsfrage ist nicht mehr trennscharf. Es gibt keine Antwort mehr auf „ist Cloud besser als lokal?“, weil gute Dienste und gute offene Modelle auf einem Niveau liegen, das für die meisten Einsätze mehr als reicht. Wo es Unterschiede gibt, liegen sie in Spezialdisziplinen – Text im Bild, Bildbearbeitung mit Referenz, Gesichtstreue – und wechseln im Halbjahresrhythmus.

Zweitens: Konkrete Modellempfehlungen gehören nicht hierher, sondern in die Modellübersichten, weil sie ein Verfallsdatum haben. Die Betriebsmodellfrage hat keins.

Drittens: Der rechtliche Teil ist hier Orientierung, nicht Beratung. Datenschutzrechtliche Bewertung, Lizenzprüfung und die Frage, ob eine konkrete Verarbeitung zulässig ist, hängen am Einzelfall. Der Hinweis lautet nicht „das wird schon passen“, sondern: Schau bei sensiblen oder kommerziellen Projekten genau hin, und hol im Zweifel jemanden dazu, der es fachlich beurteilen darf.

Fazit

Cloud vs. lokale Bildgenerierung entscheidet sich nicht am Preis pro Bild, auch wenn die Diskussion dort anfängt. Die Kostenachse ist 2026 sogar eindeutiger zur Cloud gekippt als früher: Bilder aus dem Netz sind billiger geworden, Karten mit viel Videospeicher teurer. Wer lokal ausschließlich mit Sparen begründet, argumentiert auf der schwächsten Achse, die zur Verfügung steht.

Die Achsen, die tragen, sind Kontrolle, Reproduzierbarkeit und Datenschutz. Eine eingefrorene Modellversion, die in achtzehn Monaten noch dasselbe Bild liefert. Prompts, die das Haus nicht verlassen. Ein Workflow, den du ändern kannst, ohne auf ein Anbieter-Update zu hoffen. Das ist nichts, was man als Notlösung in Kauf nimmt, weil das Budget für den besseren Dienst fehlt – das ist die anspruchsvollere Variante. Lokal ist die Kür, nicht der Kompromiss.

Gleichzeitig gilt: Der reine Gegensatz ist eine Konstruktion. Zwischen Dienst und eigener Maschine liegt die kontrollierte GPU-Instanz, und sie löst mehr Fälle als beide Extreme. Und über allem steht die Einsicht, dass das schwierige Stück nicht die Hardware ist, sondern der Katalog daneben: Wer nicht sagen kann, wie ein Bild entstanden ist und ob es gekennzeichnet wurde, hat die Betriebsmodellfrage nicht gelöst – nur verschoben.

Wenn du eine einzige Sache mitnimmst: Sortiere deine Bilder in Wegwerfbilder und Aktenbilder. Die erste Sorte darf in die Cloud. Die zweite gehört dorthin, wo du in zwei Jahren noch erklären kannst, was passiert ist.

Häufige Fragen

Was ist der Hauptunterschied zwischen Cloud und lokaler Bildgenerierung?

Wessen Grafikkarte rechnet. Bei einem Cloud-Dienst gehen Prompt und Referenzbilder an den Anbieter, dessen Hardware das Bild erzeugt und zurückschickt. Lokal läuft das Modell auf deiner Maschine, nichts verlässt das Haus. Aus diesem einen Punkt folgt der Rest: Kostenstruktur, Zugriff auf Parameter, Datenschutzlage und die Frage, wer über die Modellversion entscheidet.

Was ist günstiger – Cloud oder lokal?

Bei kleinen und mittleren Mengen die Cloud, deutlicher als noch vor zwei Jahren. Ein Megapixel kostet bei den gängigen Diensten zwischen etwa 0,014 und 0,10 US-Dollar; 2.500 Generierungen im Monat liegen damit oft unter hundert Dollar. Lokale Hardware ist gleichzeitig teurer geworden, weil Speicher knapp ist. In einer Modellrechnung mit drei Cent pro Bild und einer Karte für 4.000 Euro über drei Jahre trägt sich eigene Hardware ab etwa 3.800 Bildern im Monat, eine gemietete Karte schon ab knapp tausend. Rechne aber deine Arbeitszeit mit – eine Beratungsstunde entspricht rund fünftausend Cloud-Bildern.

Ist die lokale Generierung besser für den Datenschutz?

In der Regel ja, und zwar strukturell: Es gibt keinen Empfänger, keine Übermittlung an einen Dritten, keinen Drittlandtransfer und keinen Vertrag über Auftragsverarbeitung, der schiefgehen könnte. Bei Cloud-Diensten musst du dagegen prüfen, was mit Eingaben passiert. Schnittstellenzugriffe sind oft sauberer als Endkundentarife: Bei OpenAI werden API-Daten seit März 2023 nicht zum Training genutzt, sofern man nicht zustimmt, mit dreißig Tagen Standardaufbewahrung für die Missbrauchserkennung. Bei Midjourney landen Generierungen in den Tarifen Basic und Standard in der öffentlichen Galerie, und die Trainingsrechte des Anbieters bleiben auch im Stealth-Modus bestehen.

Brauche ich eine teure Grafikkarte für lokale Generierung?

Das hängt stärker vom Modell ab als vom Anspruch. Schnelle, kompakte Modelle laufen bereits mit 8 bis 16 GB Videospeicher komfortabel, in quantisierten Fassungen teils mit rund 6 GB, und liefern auf einer Consumer-Karte Bilder in fünf bis zehn Sekunden. Große Modelle wie FLUX.2 [dev] mit 32 Milliarden Parametern belegen in FP8 rund 32 GB und wollen entsprechend eine Karte mit 32 GB oder mehr; mit Q4-Quantisierung sinkt das auf etwa 19 GB und passt auf 24 GB. Teuer wird es nicht durch Rechenleistung, sondern durch Videospeicher – und der ist 2026 der Engpass.

Sind Bilder aus Cloud-Diensten reproduzierbar?

Nur eingeschränkt. Selbst wenn du denselben Seed und dieselben Parameter zurückschicken kannst, weißt du meist nicht, welche Modellversion gerade rechnet. Anbieter tauschen Gewichte, ändern Standardwerte und passen Filter an, ohne dass eine Versionsnummer im Ergebnis auftaucht. Für exakte Wiederherstellbarkeit über Monate brauchst du eine fixierte Version – also entweder eine vertragliche Zusicherung oder ein Modell, das du selbst hältst.

Ist eine gemietete GPU dasselbe wie Cloud?

Nein, und diese Verwechslung ist der häufigste Denkfehler in der Debatte. Bei einer gemieteten Instanz läuft dein Abbild mit deinen Modellen und Workflows; du hast alle Regler, kannst Versionen einfrieren und eigene LoRAs einsetzen. Der Unterschied zu lokal ist nicht die Kontrolle über die Software, sondern dass die Hardware jemand anderem gehört. Tragfähig ist das mit Vertrag über Auftragsverarbeitung, festgelegtem Standort und abgeschaltetem Logging. Ohne diese drei ist es wirklich nur Cloud mit Extraarbeit.

Wie ist die Lizenzlage bei lokal betriebenen Modellen?

Differenziert. Bei FLUX.2 [dev] und FLUX.2 [klein] 9B erlaubt die nicht-kommerzielle Lizenz die kommerzielle Verwendung der erzeugten Bilder, nicht aber den kommerziellen Betrieb des Modells selbst – dafür braucht es eine bezahlte Lizenz von Black Forest Labs. Uneingeschränkt nutzbar sind in dieser Familie die 4B-Variante von FLUX.2 [klein] unter Apache 2.0 sowie aus dem Alibaba-Umfeld Qwen-Image und Z-Image Turbo, ebenfalls Apache 2.0. Wenn du Bildproduktion als Dienstleistung verkaufst, ist genau diese Unterscheidung die, die du vorher klären musst.

Muss ich KI-Bilder kennzeichnen – und macht das der Dienst für mich?

Kennzeichnen ja. Seit dem 2. August 2026 gelten die Transparenzpflichten aus Artikel 50 der KI-Verordnung verbindlich; für die maschinenlesbare Markierung nach Absatz 2 gibt es bei Systemen, die vorher in Verkehr gebracht wurden, eine Übergangsfrist bis zum 2. Dezember 2026. Die technische Markierung ist Anbieterpflicht: Große Cloud-Dienste liefern signierte Herkunftsmetadaten nach C2PA, teils zusätzlich ein unsichtbares Wasserzeichen. Ein lokal betriebenes Modell mit offenen Gewichten tut das nicht von allein – dort musst du die Markierung selbst in den Export einbauen. Die sichtbare Kennzeichnung im veröffentlichten Inhalt liegt ohnehin immer bei dir.

Muss ich mich für eines von beiden entscheiden?

Nein, und die meisten funktionierenden Aufbauten tun es nicht. Entwürfe, Moodboards und alles, was morgen weggeworfen wird, laufen über einen Dienst; alles, was veröffentlicht, archiviert oder freigegeben wird, läuft über die kontrollierte Strecke. Die Trennlinie liegt zwischen Wegwerfbildern und Aktenbildern, nicht zwischen Bildarten. Wichtig ist nur, dass beide Wege in denselben Katalog münden – sonst hast du zwei Prozesse und keinen Nachweis.

Liefert die Cloud bessere Bilder als ein lokales Modell?

Nicht mehr pauschal. Auf dem heutigen Stand liefern gute Dienste und gute offene Modelle Ergebnisse auf vergleichbarem Niveau; Unterschiede gibt es in Spezialdisziplinen wie Text im Bild, referenzbasierter Bearbeitung oder Gesichtstreue, und die wechseln schnell. Wer heute noch mit Qualität argumentiert, argumentiert meist mit einem Stand von vor zwei Jahren. Die relevanten Unterschiede liegen im Drumherum.

WEITERLESEN

- › [KI-Bildgenerierung – der Überblick](#) — das gesamte Themenfeld im Zusammenhang
- › [Die 15 wichtigsten Modelle zur lokalen Bildgenerierung](#) — welches Modell welchen Videospeicher braucht
- › [Übersicht der wichtigsten KI-Bildgeneratoren](#) — was die Dienste heute leisten
- › [Aufbau einer skalierbaren Bild-Maschine mit ComfyUI und Flux](#) — wie die kontrollierte Strecke praktisch aussieht
- › [KI-Bildgenerierung im Unternehmen – Die strategische Alternative zur Cloud](#) — die Entscheidung auf Unternehmensebene
- › [Der Seed bei der KI-Bildgenerierung](#) — die Grundlage jeder Reproduzierbarkeit
- › [CFG-Scale: Guidance in KI-Bildgeneratoren](#) — der Regler zwischen Treue und Freiheit
- › [Sampler & Scheduler erklärt](#) — was lokal an Feinsteuerung dazukommt
- › [Was ist ein LoRA?](#) — eigene Stile und Motive ins Modell holen
- › [Digital Asset Management für KI-Bilder mit SharePoint Online und Power Platform](#) — der Katalog, ohne den der Rest nichts wert ist
- › [EU AI Act: Was seit dem 2. August 2026 gilt](#) — Transparenzpflichten im Überblick
- › [Beratung zur KI-Bildgenerierung](#) — Betriebsmodell und Nachweiskonzept gemeinsam festlegen