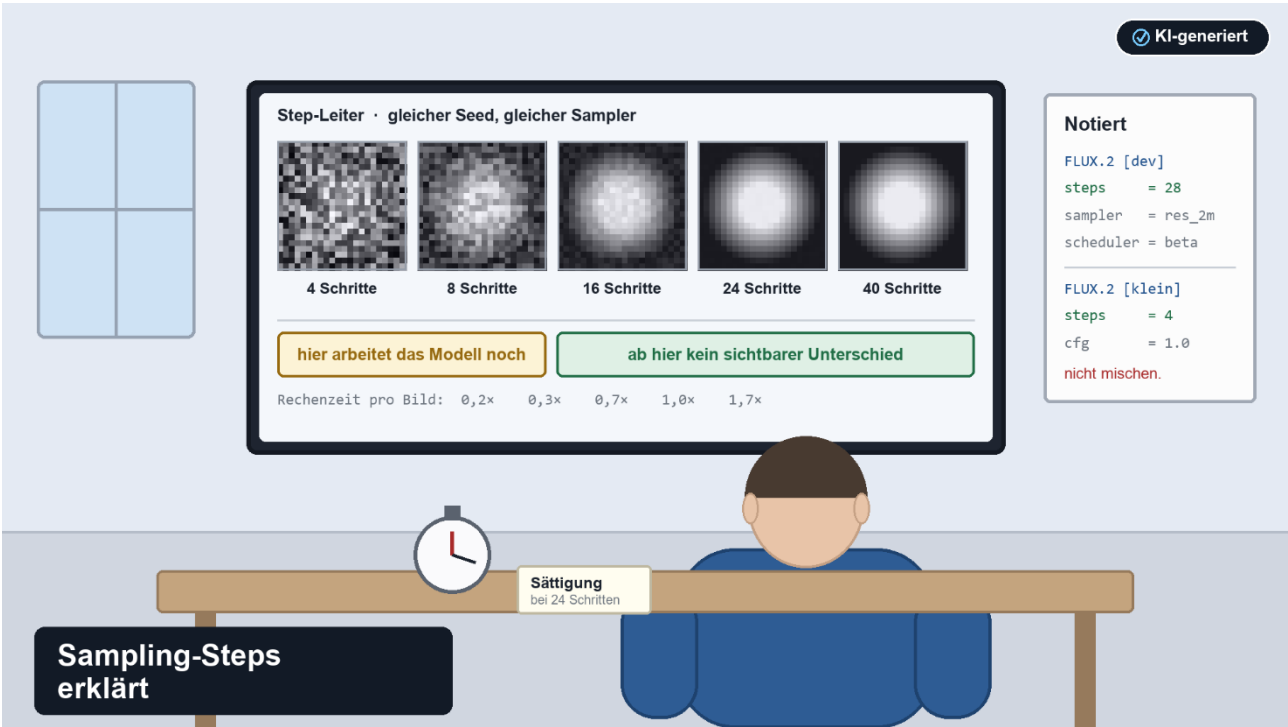


Sampling-Steps erklärt



Dieselbe Szene in fünf Schrittzahlen: Irgendwo zwischen der dritten und vierten Kachel hört der Gewinn auf, die Rechenzeit aber nicht. Bild selbst erzeugt und als KI-generiert gekennzeichnet.

Wissen	Beratung	Schulung
Alle Grundlagen zur KI-Bildgenerierung an einer Stelle – vom Diffusionsprinzip bis zur Governance.	Vom Einzelversuch zur Bildproduktion, die Termine hält und Rechenzeit nicht verschenkt.	Dein Team lernt, welcher Regler welches Problem löst – und welcher eben nicht.
KI-Bildgenerierung	KI-Bildgenerierung Beratung	KI-Bildgenerierung Schulung

Der Regler heißt „Steps“, „Sampling-Schritte“ oder „Sampling steps“, steht meist ganz oben in der Oberfläche und lädt zu genau einer Fehlannahme ein: mehr ist besser. Das ist verständlich. Bei Auflösung stimmt es ungefähr, bei Bitrate stimmt es, bei Arbeitsspeicher sowieso. Bei Sampling-Schritten stimmt es für die ersten paar Dutzend Prozent und danach überhaupt nicht mehr.

Ich habe das bei einem Industriedienstleister in Reinform gesehen. Dort lief eine Bildstrecke für einen Produktkatalog, 400 Motive, Abgabe in vier Tagen. Jemand hatte in der Vorlage die Schrittzahl auf 150 gesetzt, weil das Ergebnis „in Druckqualität“ gebraucht wurde – eine Logik, die aus der Welt der Bildauflösung stammt und in dieser Welt nichts zu suchen hat. Die Rechenknechte liefen drei Nächte. Als wir hinterher zwei Bilder nebeneinander gelegt haben, eines mit 150 Schritten und eines mit 26, brauchte es die Lupe, um den Unterschied überhaupt zu finden. Er war da. Er war nur nicht besser, sondern anders. Die Sonderschicht hatte rund fünfmal so viel Rechenzeit gekostet wie nötig.

Dieser Beitrag klärt, was bei einem Schritt tatsächlich passiert, warum ab einem bestimmten Punkt nichts Sichtbares mehr dazukommt, wie du diesen Punkt für dein Modell selbst misst statt ihn aus einem Forenbeitrag abzuschreiben – und warum manche Modelle ein fertiges Bild in vier Schritten liefern, wo andere dreißig brauchen. Wenn dir das Fundament dahinter fehlt, ist [Basiswissen KI-](#)

[Bildgenerierung: Wie Diffusion funktioniert](#) der passende Einstieg; die ganze Reglerlandschaft steht im Überblick unter [KI-Bildgenerierung](#).

Was du aus diesem Beitrag mitnimmst

Ein Sampling-Schritt ist eine Stützstelle auf dem Weg vom Rauschen zum Bild, keine zusätzliche Schicht Politur. Mehr Schritte heißt: dem gleichen Weg genauer folgen.

Die Schrittzahl legt gemeinsam mit dem Scheduler den kompletten Rauschplan fest. Änderst du sie, bekommst du nicht dasselbe Bild mit mehr Details, sondern ein anderes Bild. Das ist der Grund, warum Vergleiche so oft in die Irre führen.

Der Sättigungspunkt liegt bei nicht destillierten Modellen typischerweise zwischen 20 und 30 Schritten. Darüber sind die Änderungen kleiner als die Wahrnehmungsschwelle – die Rechenzeit steigt trotzdem linear weiter.

Destillierte Modelle – FLUX.1 [schnell], FLUX.2 [klein], die Turbo-, Lightning- und Hyper-Varianten, Z-Image Turbo, Qwen-Image-Lightning – haben den Sprungabstand im Training eingebrannt. Sie liefern bei 4 bis 8 Schritten ihr Bestes und werden darüber schlechter, nicht besser.

Steps reparieren keinen Prompt, keine Hände, keine Auflösung und kein falsch gewähltes Modell. Wer dort landet, dreht am falschen Regler.

Was bei einem Schritt tatsächlich passiert

Die verbreitete Erklärung lautet: Das Modell startet mit Rauschen und entfernt in jedem Schritt ein bisschen davon, bis das Bild da ist. Das ist nicht falsch, aber es erzeugt genau die Vorstellung, die später zu Fehlbedienung führt – nämlich die, dass Schritte aufeinander aufbauen wie Schichten Lack. Tun sie nicht.

Der Pfad liegt fest, die Schritte sind Stützstellen

Ein Diffusions- oder Flow-Matching-Modell beschreibt mathematisch einen Weg: von einem vollständig verrauschten Ausgangszustand hin zu einem rauschfreien Bild. Dieser Weg ist durch das Modell, den Prompt und den Startzustand eindeutig bestimmt. Er ist kontinuierlich – er hat keine Schritte. Die Schritte kommen erst dadurch ins Spiel, dass ein Rechner diesen Weg nicht in unendlicher Feinheit ablaufen kann. Er nähert ihn in geraden Teilstücken an.

Genau das leistet der Sampler: Er ist das Verfahren, mit dem die Teilstücke gelegt werden, und manche Verfahren treffen den Pfad mit wenigen Stützstellen deutlich besser als andere. Welches Verfahren du wählst, ist eine eigene Entscheidung – nachzulesen unter [Sampler & Scheduler erklärt](#). Die Schrittzahl beantwortet nur die zweite Hälfte der Frage: wie viele Stützstellen das Verfahren bekommt.

Daraus folgt die erste praktische Konsequenz. Wenige Schritte bedeuten große, gerade Sprünge, die den gekrümmten Pfad an den Ecken abschneiden. Das Ergebnis landet neben dem eigentlichen Ziel: unscharf, flau, manchmal mit merkwürdiger Farbgebung. Viele Schritte bedeuten kurze Teilstücke, die sich eng an den Pfad schmiegen. Ab einer gewissen Feinheit liegt das Ergebnis so nah am mathematisch korrekten Ziel, dass weitere Verfeinerung nur noch Stellen hinter dem Komma verschiebt.

Ein Schritt ist eine Stützstelle, keine Schicht Politur

Der Weg vom Rauschen zum Bild liegt fest. Die Schrittzahl entscheidet, wie genau ihm gerechnet wird.



Skizze 1: Oben derselbe Pfad mit vier und mit 24 Stützstellen. Unten der Grund, warum ein Schritt mehr kein Feinschliff ist: Zwei Schrittzahlen erzeugen zwei völlig verschiedene Rauschpläne, die sich nur an Anfang und Ende treffen.

Warum 21 Schritte nicht „20 plus einer“ sind

Das ist der Punkt, an dem die meisten Vergleichsversuche scheitern. Die Schrittzahl steht nicht allein. Zusammen mit dem Scheduler – dem Teil der Einstellung, der die Rauschmengen über den Verlauf verteilt – erzeugt sie einen Plan: eine Liste von Rauschwerten, die von 100 Prozent bis auf null führt. Diese Liste wird für jede Schrittzahl komplett neu berechnet. Bei 20 Schritten liegen die Stützstellen an anderen Stellen als bei 21. Nicht an denselben plus einer weiteren. An anderen.

Das bedeutet: Ein Lauf mit 20 Schritten und ein Lauf mit 24 Schritten – gleicher Seed, gleicher Prompt, gleiches Modell – erzeugen nicht dasselbe Bild in zwei Schärfegraden, sondern zwei verwandte, aber unterschiedliche Bilder. Gesichtszüge können sich leicht verschieben, ein Faltenwurf anders fallen, ein Hintergrundobjekt seine Form ändern. Wer das nicht weiß, hält den ersten Unterschied für einen Qualitätsgewinn und dreht weiter.

Der häufigste Denkfehler beim Vergleichen

Ein Bild mit höherer Schrittzahl ist keine verbesserte Version des Bildes mit niedrigerer Schrittzahl. Es ist ein neues Bild aus demselben Ausgangsrauschen, auf einem feiner abgetasteten Weg gerechnet. Deshalb ist „sieht anders aus“ kein Beleg für „ist besser“. Die Frage beim Vergleichen lautet nicht, ob sich etwas verändert hat – es verändert sich immer etwas –, sondern ob das spätere Bild handwerklich sauberer ist: schärfere Kanten, sauberere Materialien, weniger Matsch in dunklen Flächen.

Ab dem Sättigungspunkt verändert sich auch das nicht mehr. Dann gleichen sich die Bilder bis auf Pixelrauschen, das kein Mensch in der Druckvorstufe je bemerken wird.

Wo Guidance dazwischenfunkt

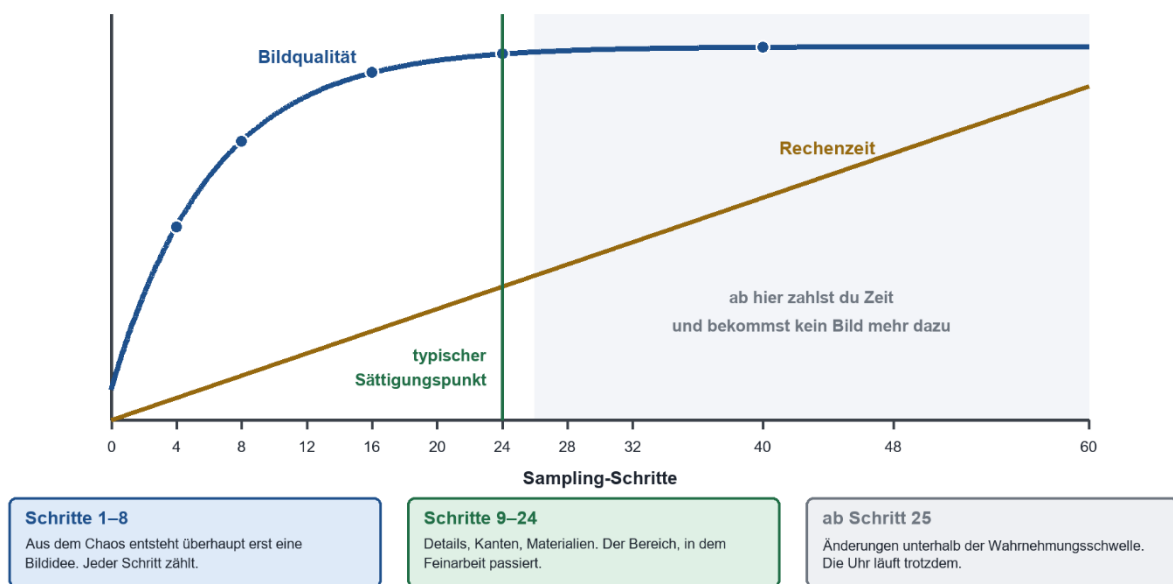
Eine Nebenwirkung, die in der Praxis regelmäßig für Verwirrung sorgt: Bei hoher Guidance – dem Regler, der die Prompt-Treue erzwingt – wird der Pfad steiler und unruhiger. Das Verfahren braucht dann mehr Stützstellen für dasselbe Ergebnis, und typische Überzeichnungsartefakte wie ausgefressene Farben werden mit steigender Schrittzahl nicht schwächer, sondern sauberer herausgerechnet. Wer also verbrannte Farben mit mehr Schritten bekämpft, verstärkt sie. Der richtige Regler steht in [CFG-Scale: Guidance in KI-Bildgeneratoren](#).

Der Sättigungspunkt – ab wann mehr Schritte nichts mehr bringen

Jedes Näherungsverfahren hat einen Fehler, und dieser Fehler schrumpft mit kleineren Teilstücken. Das Entscheidende ist, wie schnell er schrumpft: nicht linear, sondern steil abfallend. Die ersten Schritte holen den größten Teil des Fehlers weg, die mittleren einen erkennbaren Rest, die späten fast nichts mehr. Irgendwann ist der verbleibende Fehler kleiner als das, was ein Bildschirm darstellen und ein Auge unterscheiden kann. Das ist der Sättigungspunkt.

Der Sättigungspunkt

Der Gewinn pro Schritt fällt exponentiell. Die Rechenzeit steigt linear. Irgendwo kreuzen sich die Interessen.



Skizze 2: Die Qualität steigt steil und flacht ab, die Rechenzeit steigt geradlinig weiter. Der interessante Bereich ist der Knick – rechts davon kaufst du nur noch Wartezeit.

Warum höherwertige Verfahren früher sättigen

Der Sättigungspunkt ist keine Eigenschaft des Modells allein. Ein einfaches Verfahren wie Euler legt gerade Teilstücke und braucht entsprechend viele davon. Verfahren höherer Ordnung – die DPM++-Familie, UniPC, die neueren Mehrschritt-Varianten – schätzen die Krümmung des Pfades mit ab und treffen ihn mit weniger Stützstellen. Das ist der Grund, warum dieselbe Modelldatei mit dem einen Sampler bei 30 Schritten sättigt und mit dem anderen schon bei 18. Wer die Schrittzahl senken will, wechselt oft besser zuerst das Verfahren; welches wofür taugt, steht in [Sampler & Scheduler erklärt](#).

Was „mehr Schritte machen es schlechter“ wirklich heißt

Diese Behauptung geistert durch jedes Forum, und sie stimmt – aber aus drei verschiedenen Gründen, die man auseinanderhalten sollte, weil jeder eine andere Konsequenz hat.

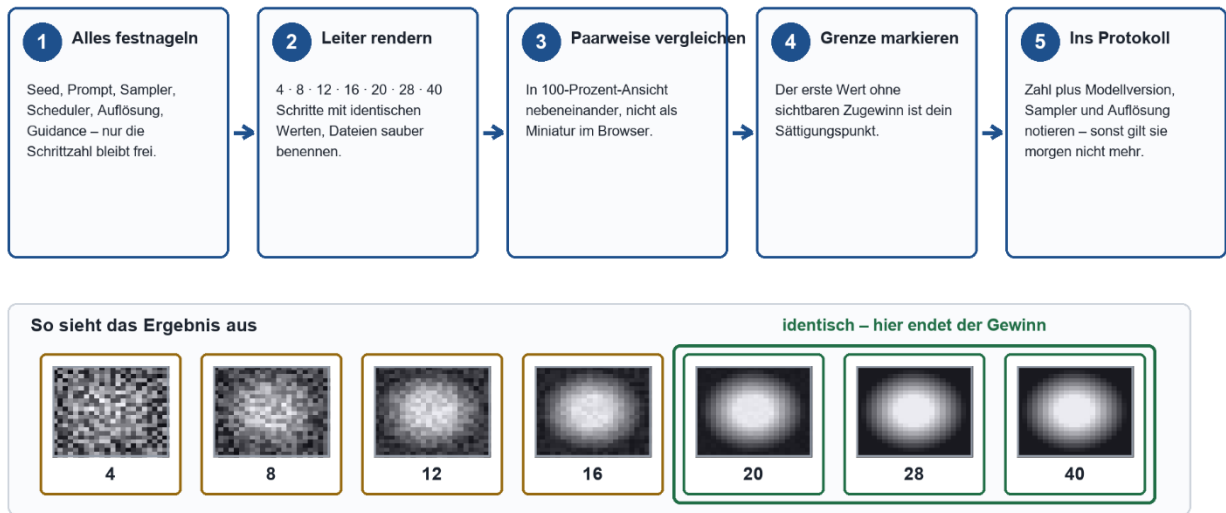
Beobachtung	Ursache	Was zu tun ist
Das Bild wird mit jedem Erhöhen wieder anders, es beruhigt sich nie	Ein Sampler aus der ancestral- oder SDE-Familie mischt in jedem Schritt frisches Rauschen zu. Solche Verfahren konvergieren grundsätzlich nicht.	Für Vergleichsläufe und für die Produktion ein deterministisches Verfahren wählen. Ancestral-Varianten sind zum Stöbern gut, zum Messen nicht.
Farben brennen aus, Kanten bekommen Halos, Haut wirkt wie Plastik	Zu hohe Guidance. Mehr Schritte rechnen die Überzeichnung nur präziser heraus.	Guidance senken, nicht die Schrittzahl erhöhen.
Ab einer bestimmten Zahl wird alles flau, matschig oder bekommt Farbstiche	Ein destilliertes Modell wird jenseits seines trainierten Sprungabstands betrieben.	Zurück auf den Sollwert des Modells – meist 4 bis 8 Schritte. Mehr ist dort kein Feintuning, sondern ein Bedienfehler.
Kein sichtbarer Unterschied mehr, nur längere Laufzeit	Der Sättigungspunkt ist erreicht. Alles Weitere ist Rundungsrauschen.	Freuen und den Wert festschreiben. Das ist der Normalfall und das Ziel.

Den eigenen Sättigungspunkt messen, nicht raten

Jede Zahl, die du irgendwo liest – auch die in diesem Beitrag – ist ein Startwert. Der tatsächliche Sättigungspunkt hängt an Modellversion, Sampler, Scheduler, Auflösung, Guidance und am Motiv. Ein Porträt mit weichem Hintergrund sättigt früher als eine Werkhalle voller Kanten und Beschriftungen. Die gute Nachricht: Die Messung dauert zwanzig Minuten und gilt danach für alle Läufe mit diesem Aufbau. Halte dabei den Seed fest, sonst misst du den Zufall statt der Schrittzahl – warum das so ist, steht in [Der Seed bei der KI-Bildgenerierung](#).

Die Step-Leiter: den eigenen Sättigungspunkt messen statt raten

Einmal pro Modell, zwanzig Minuten Aufwand, danach steht die Zahl für alle Läufe.



Skizze 3: Die Step-Leiter in fünf Handgriffen. Einmal pro Modell und Auflösung durchgeführt, danach steht die Zahl – und man hört auf, sie bei jedem Auftrag neu zu erfinden.

Step-Leiter: das Testrezept

Nimm ein Motiv, das für deine Bildstrecke typisch ist – nicht das hübscheste, sondern das

anspruchsvollste. Feine Strukturen, Text auf einem Schild, viele Kanten.

Friere alles ein außer der Schrittzahl: Seed, Prompt, Negativ-Prompt, Sampler, Scheduler, Auflösung, Guidance, Modellversion. Ein einziger veränderlicher Wert, sonst ist die Messung wertlos.

Rendere die Leiter 4 – 8 – 12 – 16 – 20 – 28 – 40 und benenne die Dateien nach der Schrittzahl.

Vergleiche paarweise in 100-Prozent-Ansicht, nicht als Miniatur. Der Unterschied zwischen 16 und 20 Schritten ist in einer Vorschau von 300 Pixeln Breite grundsätzlich unsichtbar.

Der erste Wert, ab dem du beim Umschalten nichts Handwerkliches mehr gewinnst, ist dein Sättigungspunkt. Nimm ihn plus einen Sicherheitsaufschlag von zwei bis vier Schritten – und notiere ihn zusammen mit Modell, Sampler und Auflösung, denn ohne diesen Kontext gilt er nicht.

Wenn du solche Läufe stapelweise durchziehst, lohnt sich der Blick auf die Stapelgröße: Ein Testlauf über sieben Schrittzahlen ist der klassische Fall, in dem man versehentlich sieben verschiedene Seeds vergleicht. Wie Stapel und Seed zusammenhängen, steht in [batch size und Reproduzierbarkeit](#). Und wenn die gefundene Zahl später im Produktions-Workflow geändert wird, prüft bei mir UB.Forge beim Speichern, ob sie zum Modelltyp passt – 40 Schritte auf einem destillierten Modell sind kein mutiges Feintuning, sondern ein Plausibilitätsfehler, der vor dem Rendern auffallen sollte und nicht danach.

Warum destillierte Modelle mit vier Schritten auskommen

Jetzt der interessante Teil. Wenn ein ordentliches Bild zwanzig bis dreißig Stützstellen braucht – wie liefert dann ein Modell wie FLUX.2 [klein] in unter einer Sekunde ein fertiges Bild in vier Schritten? Die Antwort ist nicht, dass die Schritte größer geworden wären. Die Antwort ist, dass die Arbeit an eine andere Stelle verlagert wurde.

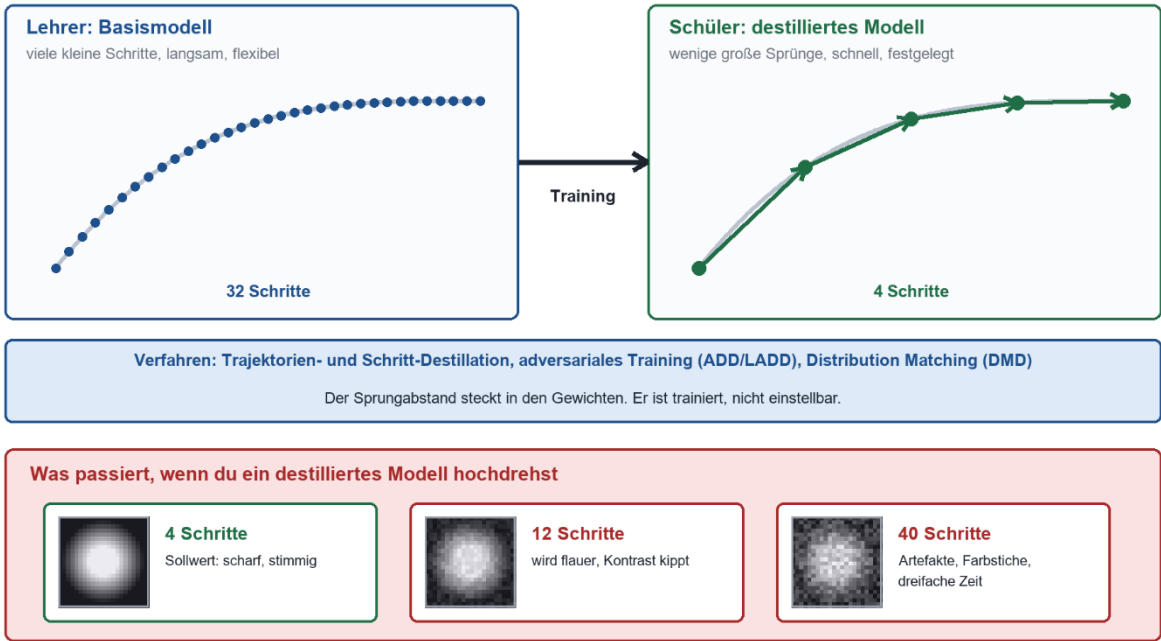
Destillation: der Umweg über das Training

Bei der Destillation dient ein großes, langsames Modell als Lehrer. Es rechnet den Bildpfad mit vielen kleinen Schritten aus, so wie es das immer tut. Ein zweites Modell – der Schüler – wird darauf trainiert, dasselbe Ziel mit wenigen, sehr großen Sprüngen zu treffen. Der Schüler lernt also nicht, Rauschen zu entfernen; er lernt, den Endpunkt eines ganzen Pfadabschnitts direkt vorherzusagen. Der Sprungabstand ist damit kein Regler mehr, sondern eine Eigenschaft der Gewichte.

Die technischen Verfahren dahinter sind unterschiedlich und haben sich in den letzten Jahren deutlich weiterentwickelt. Bei den frühen Turbo-Modellen von Stability kam Adversarial Diffusion Distillation zum Einsatz: Ein zweites Netz lernt parallel, echte von erzeugten Bildern zu unterscheiden, und zwingt den Schüler damit zu schärferen Ergebnissen. FLUX.1 [schnell] nutzt dasselbe Prinzip im Latentraum – latent adversarial diffusion distillation – und wurde dabei nicht vom offenen Dev-Modell abgeleitet, sondern vom nicht öffentlichen Pro-Modell. Die neueren Vertreter wie Z-Image Turbo arbeiten mit Verteilungsabgleich statt mit einem Gegenspieler, in diesem Fall mit einer entkoppelten Variante der Distribution Matching Distillation.

Warum vier Schritte reichen können

Destillation verlegt die Arbeit vom Erzeugen ins Training: Das Schülermodell lernt die großen Sprünge.



Skizze 4: Der Lehrer rechnet den Pfad in vielen kleinen Schritten, der Schüler lernt, dasselbe Ziel in vier Sprüngen zu treffen. Und unten das, was passiert, wenn man den Sprungabstand nachträglich ignoriert.

Guidance-Destillation ist etwas anderes als Schritt-Destillation

Diese Unterscheidung ist wichtig, weil sie im Alltag ständig durcheinandergeht. Es gibt zwei Dinge, die man aus einem Modell herausdestillieren kann, und sie haben unterschiedliche Folgen für die Schrittzahl.

	Guidance-Destillation	Schritt-Destillation
Was eingespart wird	Der zweite Rechendurchgang pro Schritt, den klassische Guidance braucht	Die Anzahl der Schritte selbst
Wirkung auf die Schrittzahl	keine – das Modell braucht weiterhin 20 bis 30 Schritte	drastisch – 1 bis 8 Schritte statt 20 bis 30
Der Guidance-Regler	wird zu einem eingebakkenen Wert, der das Ergebnis stilistisch steuert	muss in der Regel auf 1 bleiben, sonst kippt das Bild
Typische Vertreter	FLUX.1 [dev], FLUX.2 [dev]	FLUX.1 [schnell], FLUX.2 [klein], Turbo-, Lightning- und Hyper-Varianten, Z-Image Turbo, Qwen-Image-Lightning

Die praktische Merkregel: Ein guidance-destilliertes Modell verhält sich bei der Schrittzahl wie ein normales Modell. Ein schritt-destilliertes Modell hat einen Sollwert, und dieser Sollwert ist eine Vorgabe, kein Vorschlag.

Destillierte Modelle vertragen kein Hochdrehen

Wer bei FLUX.2 [klein] von den vorgesehenen vier Schritten auf zwanzig geht, bekommt kein besseres Bild, sondern ein flaueres – und zahlt dafür das Fünffache an Rechenzeit. Dasselbe gilt für Turbo-, Lightning- und Hyper-Varianten.

Ebenso wenig verträgt diese Modellklasse die gewohnte Guidance. Der Wert gehört bei den meisten schritt-destillierten Modellen auf 1,0 – bei manchen bis 1,5. Höhere Werte erzeugen Überzeichnung,

die dann fälschlich der Schrittzahl angelastet wird.

Der umgekehrte Fehler ist genauso teuer: ein nicht destilliertes Modell mit vier Schritten zu fahren, weil der Wert aus dem letzten Projekt noch in der Vorlage stand. Das Ergebnis ist verlässlich Matsch – und wird dann gern dem Modell angelastet statt der Einstellung.

Der Preis der wenigen Schritte

Ehrlichkeitshalber: Destillation ist kein kostenloses Mittagessen, auch wenn der Abstand in den letzten beiden Jahren erheblich geschrumpft ist. Schritt-destillierte Modelle zeigen typischerweise weniger Variantenvielfalt bei gleichem Prompt – sie haben gelernt, zuverlässig auf einen Zielpunkt zu springen, und das äußert sich in einer gewissen Gleichförmigkeit über verschiedene Seeds hinweg. Bei sehr langen oder widersprüchlichen Prompts verlieren sie eher Details als das große Ausgangsmodell, und feine Materialstrukturen geraten auf großen Flächen schneller glatt. Für Entwurfsschleifen, Varianten-Suche und alles Interaktive ist das völlig gleichgültig. Für das finale Motiv, das in 3.000 Katalogen landet, kann es der Unterschied sein. Welches Modell was leistet, ordnet [Die 15 wichtigsten Modelle zur lokalen Bildgenerierung](#) ein.

Und ein Punkt, der in Unternehmen mehr zählt als jede Bildqualitätsdebatte: Die Lizenzlage unterscheidet sich zwischen den Varianten derselben Familie erheblich. FLUX.1 [schnell] steht unter Apache 2.0 und darf kommerziell frei eingesetzt werden, FLUX.1 [dev] und FLUX.2 [dev] nicht. Bei FLUX.2 [klein] gilt Apache 2.0 für die 4-Milliarden-Variante, während die größere 9-Milliarden-Variante unter einer nicht-kommerziellen Lizenz steht. Z-Image Turbo und Qwen-Image stehen unter Apache 2.0. Die Turbo-Modelle von Stability laufen unter der Stability AI Community License, die kommerzielle Nutzung bis eine Million US-Dollar Jahresumsatz erlaubt und darüber eine Unternehmenslizenz verlangt. Wer das schnellste Modell wählt, ohne diese Zeile zu prüfen, hat unter Umständen das teuerste gewählt.

Faustregeln je Modellfamilie

Die folgenden Werte sind Startpunkte für die eigene Messung, keine Naturkonstanten. Sie gelten für Text-zu-Bild bei üblichen Auflösungen und mit einem Sampler höherer Ordnung. Der Stand ist September 2026.

Nicht destillierte und guidance-destillierte Modelle

Modellfamilie	Schritte	Guidance	Anmerkung
Stable Diffusion 1.5	20–30	5–8	Klassisches Diffusionsmodell, gutmütig, sättigt spät bei einfachen Samplern
SDXL 1.0 (Base)	25–35	5–8	Der optionale Refiner übernimmt die letzten Prozent, nicht die Schrittzahl
Stable Diffusion 3.5 Large	28–40	3,5–5	Flow Matching; reagiert empfindlicher auf den Scheduler als auf die Schrittzahl
FLUX.1 [dev], 12 Mrd. Parameter	20–30 (Standardwert 28)	3,5 (eingebacken)	Guidance-destilliert, aber nicht schritt-destilliert; nicht-kommerzielle Lizenz
FLUX.2 [dev], 32 Mrd. Parameter	20–50	etwa 4	Seit November 2025; volle Kontrolle über alle Werte, hoher Speicherbedarf
FLUX.2 [flex] (API)	4–50 frei wählbar	frei wählbar	Der einzige große Cloud-Zugang, der die Schrittzahl offenlegt
Qwen-Image / Qwen-Image 2.0 (7 Mrd.)	20–40	3–4	Version 2.0 seit Februar 2026, Apache 2.0, starke Typografie bis 2K

Schritt-destillierte Modelle

Modellfamilie	Schritte	Guidance	Anmerkung
SD Turbo / SDXL Turbo	1–4	1,0	Adversarial Diffusion Distillation; Stability AI Community License
SDXL Lightning, Hyper-SD, LCM	1–8	1–2	Auch als LoRA auf vorhandene Modelle aufsetzbar, dann gilt der Sollwert des LoRA
SD 3.5 Large Turbo	4	1,0	Destillierte Fassung des großen Modells
FLUX.1 [schnell]	1–4	1,0	Latent adversarial distillation, abgeleitet vom Pro-Modell; Apache 2.0
FLUX.2 [klein] 4B / 9B	4	1,0–1,5	Seit Januar 2026; 4B unter Apache 2.0, 9B nicht-kommerziell. Es gibt auch eine nicht destillierte Base-Variante mit freier Schrittzahl
Z-Image Turbo (6 Mrd.)	8–10	1,0	Seit November 2025, Apache 2.0, Decoupled-DMD; unter 8 Schritten kippt es
Qwen-Image-Lightning	4–8	1,0	Destillat der Qwen-Familie, rund zehnfache Beschleunigung

Die zwei Zahlen, die man sich merken kann

Nicht destilliert: irgendwo zwischen 20 und 30 anfangen, dann einmal messen. Über 40 zu gehen ist fast nie die Lösung eines Bildproblems.

Schritt-destilliert: den Sollwert des Modells nehmen – meist 4, bei Z-Image Turbo 8 – und ihn nicht anfassen. Die Guidance bleibt bei 1.

Alles dazwischen ist eine Frage der Messung, nicht der Meinung.

Auflösung, zweite Durchgänge und Bild-zu-Bild

Drei Sonderfälle, in denen die Faustregeln nicht direkt greifen. Erstens die Auflösung: Flow-Matching-Modelle verschieben den Rauschplan abhängig von der Bildgröße – in den Oberflächen heißt dieser Wert meist „Shift“. Große Formate brauchen deshalb tendenziell ein paar Schritte mehr als das Quadrat in Standardgröße, und der gemessene Sättigungspunkt gilt streng genommen nur für die Auflösung, bei der du ihn gemessen hast. Was Auflösung sonst noch anrichtet, steht in [Auflösung & Seitenverhältnis bei KI-Bildern](#).

Zweitens die zweistufigen Abläufe, bei denen ein kleines Bild erzeugt, hochskaliert und ein zweites Mal durch den Sampler geschickt wird. Der zweite Durchgang startet nicht bei 100 Prozent Rauschen, sondern irgendwo in der Mitte des Plans. Dort reichen regelmäßig zehn bis fünfzehn Schritte für ein Ergebnis, für das die erste Stufe fünfundzwanzig gebraucht hätte – und der Gesamtaufwand sinkt trotzdem, weil die erste Stufe in kleiner Auflösung läuft.

Drittens Bild-zu-Bild, wo die Denoising-Stärke mitspielt – und zwar in den Werkzeugen unterschiedlich. In der WebUI-Familie wirkt sie direkt auf die tatsächlich gerechneten Schritte: 20 eingestellte Schritte bei einer Stärke von 0,5 ergeben rund zehn gerechnete. In ComfyUI rechnet der KSampler die eingestellten 20 Schritte tatsächlich, verkürzt aber den Rauschplan entsprechend. Wer Werte aus der einen Welt in die andere überträgt, wundert sich über die doppelte Laufzeit oder das halbfertige Bild.

Wenn der Regler gar nicht da ist

In den großen Cloud-Generatoren suchst du die Schrittzahl vergeblich. Die Bildmodelle der Gemini- und GPT-Familien nehmen Auflösung und Qualitätsstufe entgegen, keine Sampling-Schritte. Midjourney kennt einen Qualitätsparameter, der intern an der Rechenzeit dreht, aber

keine Schrittzahl im hier beschriebenen Sinn – und keine offene Schnittstelle. Bei Black Forest Labs ist FLUX.2 [flex] die Ausnahme: Dort ist die Schrittzahl von 4 bis 50 frei wählbar, während die Pro- und Max-Varianten sie intern setzen.

Das ist kein Nachteil, es ist eine Arbeitsteilung: Der Anbieter optimiert den Wert für sein Modell und rechnet dir das Ergebnis ab. Es ist aber ein handfester Unterschied für alle, die Kosten und Bildwirkung selbst steuern wollen – die Abwägung steht in [Cloud vs. lokale KI-Bildgenerierung](#). Wie sich ein solches Modell in die Office-Werkzeugkette einfügt, zeigt [Black Forest FLUX.2 in PowerPoint und Copilot](#).

Was Steps nicht können

Der teuerste Fehler beim Thema Schrittzahl ist nicht ein falscher Wert. Es ist der Versuch, mit diesem Regler Probleme zu lösen, die an ganz anderer Stelle sitzen. Drei davon kommen in jedem Projekt vor.

Kein Ersatz für einen brauchbaren Prompt

Wenn das Modell dein Motiv nicht trifft, weil die Beschreibung widersprüchlich ist, weil ein Begriff im Training nicht vorkam oder weil zwei Anforderungen sich gegenseitig ausschließen, dann liefern mehr Schritte exakt dasselbe Problem, nur präziser ausgerechnet. Das Thema gehört zu [Prompt und Negativ-Prompt in der KI-Bildgenerierung](#) und – wenn es hartnäckiger ist – zu [Warum die KI den Prompt ignoriert](#).

Dasselbe gilt für die klassischen anatomischen Aussetzer. Eine sechsfingrige Hand ist kein halbfertiges Bild, sondern ein sauber ausgerechneter Fehler. Sie wird bei 60 Schritten nicht zur fünffingrigen, sie wird zur gestochen scharfen sechsfingrigen. Warum das so ist, steht in [Warum KI keine Hände kann](#).

Kein Ersatz für Auflösung

Schritte steuern, wie fein der Entrauschungsweg abgetastet wird – nicht, wie viele Bildpunkte am Ende herauskommen. Ein Bild mit 1024 mal 1024 Punkten hat bei 80 Schritten exakt dieselbe Größe wie bei 20. Wer größere Dateien braucht, arbeitet mit Auflösung und Hochskalierung, nicht mit der Schrittzahl; siehe [Auflösung & Seitenverhältnis bei KI-Bildern](#). Auch der zweite große Hebel für Bildcharakter liegt woanders: bei der Wahl des Modells und gegebenenfalls eines Feinabstimmungs-Zusatzes, nachzulesen in [Checkpoints, Modelle und VAE](#) und [Was ist ein LoRA?](#).

Kein Posten, den man beim Rechnen vergessen darf

Die Schrittzahl geht praktisch linear in die Rechenzeit ein – und damit in Stromkosten, Kartenbelegung und, bei Cloud-Nutzung mit Schrittregler, direkt in die Rechnung. Das ist bei fünf Bildern egal und bei einer Bildstrecke nicht mehr.

Schritte	Relative Rechenzeit	800 Bilder, relativ	Sichtbarer Unterschied zu 24 Schritten
4	0,17×	rund ein Sechstel	deutlich: unfertig, flau (außer bei destillierten Modellen)
8	0,33×	ein Drittel	sichtbar: Details fehlen, Kanten weich
16	0,67×	zwei Drittel	erkennbar bei feinen Strukturen, sonst kaum
24	1,00×	Bezugsgröße	–
40	1,67×	zwei Drittel mehr	praktisch keiner
80	3,33×	mehr als das Dreifache	keiner, der eine Sonderschicht rechtfertigt

Genau hier liegt der wirtschaftliche Kern des Themas: Zwischen 24 und 80 Schritten liegt der Faktor drei in der Rechenzeit und nichts Sichtbares im Bild. Bei 800 Motiven ist das der Unterschied zwischen einem Nachmittag und einem Wochenende. Wer eine Bildproduktion aufsetzt, die Termine halten soll, kalkuliert die Schrittzahl deshalb wie eine Stückzeit in der Fertigung – dazu passt [Aufbau einer skalierbaren Bild-Maschine mit ComfyUI und Flux](#); beim Aufbau im Unternehmen ist [KI-Bildgenerierung Beratung](#) der richtige Einstieg.

Und die Zahl gehört dokumentiert. Nicht aus Bürokratiefreude, sondern weil sie ohne Kontext wertlos ist: „28 Schritte“ sagt nichts, solange nicht dabeisteht, mit welchem Modell, welchem Sampler und welcher Auflösung. Ich führe diese Werte in UB.Dossier mit, zusammen mit Modellversion und Lizenzstand – im selben Katalog, aus dem später der Nachweis gezogen wird, wenn jemand wissen will, wie ein Motiv entstanden ist. Dass diese Frage inzwischen regelmäßig gestellt wird, hat regulatorische Gründe; die Pflichtenlage beschreibt [EU AI Act: Was seit dem 2. August 2026 gilt](#). Wie so ein Katalog technisch aussieht, steht in [Digital Asset Management für KI-Bilder mit SharePoint Online und Power Platform](#).

Häufige Fragen zu Sampling-Schritten

Wie viele Sampling-Schritte sind ideal?

Bei nicht destillierten Modellen sind 20 bis 30 Schritte ein guter Startwert; darunter wirkt das Bild oft unfertig, darüber gewinnst du kaum noch etwas. Schritt-destillierte Modelle wie FLUX.1 [schnell], FLUX.2 [klein] oder die Turbo- und Lightning-Varianten arbeiten mit 4 bis 8 Schritten und sollen dort auch bleiben. Den genauen Wert für deinen Aufbau findest du mit einer Step-Leiter in zwanzig Minuten – und dann gilt er belastbar statt gefühlt.

Werden Bilder mit mehr Schritten immer besser?

Nein. Der Gewinn pro Schritt fällt steil ab, und ab dem Sättigungspunkt liegen die Unterschiede unterhalb dessen, was ein Auge auf einem Bildschirm trennen kann. Was sich weiterhin ändert, ist das Bild selbst – weil jede Schrittzahl einen eigenen Rauschplan erzeugt. Diese Veränderung mit einer Verbesserung zu verwechseln, ist der häufigste Fehler beim Vergleichen.

Warum sieht mein Bild bei 25 Schritten völlig anders aus als bei 20?

Weil die Schrittzahl zusammen mit dem Scheduler den gesamten Rauschplan festlegt. Bei 25 Schritten liegen die Stützstellen an anderen Stellen des Weges als bei 20 – nicht an denselben plus fünf weiteren. Gleiches Ausgangsrauschen, anderer Rechenweg, anderes Ergebnis. Nur Anfang und Ende des Plans sind identisch.

Warum brauchen Turbo- und Schnell-Modelle nur wenige Schritte?

Sie wurden destilliert: Ein langsames Lehrmodell rechnet den Bildpfad vor, ein Schülermodell lernt, dasselbe Ziel in wenigen großen Sprüngen zu treffen. Der Sprungabstand steckt danach in den Gewichten und ist nicht mehr einstellbar. Die frühen Turbo-Modelle nutzten dafür ein gegnerisches Training, FLUX.1 [schnell] dieselbe Idee im Latentraum, neuere Modelle wie Z-Image Turbo arbeiten mit Verteilungsabgleich.

Was bringt es, die Schrittzahl bei einem destillierten Modell zu erhöhen?

Nichts Gutes. Das Modell wird außerhalb seines trainierten Sprungabstands betrieben; typisch sind flauer Kontrast, glatt gebügelte Strukturen und Farbstiche – bei mehrfacher Rechenzeit. Wer mehr Detailtiefe braucht, wechselt das Modell, nicht die Schrittzahl. Umgekehrt gilt: Auch die Guidance gehört bei diesen Modellen auf 1 und nicht auf den Wert aus dem letzten Projekt.

Wie viele Schritte brauche ich bei Flux?

Das hängt an der Variante, und die Unterschiede sind groß. FLUX.1 [dev] arbeitet mit 20 bis 30 Schritten, Standardwert 28. FLUX.2 [dev] liegt im Bereich 20 bis 50. FLUX.2 [flex] lässt 4 bis 50 zu. FLUX.1 [schnell] und FLUX.2 [klein] sind schritt-destilliert und liefern bei 4 Schritten ihr Bestes. Die Dev-Modelle sind zwar guidance-destilliert, aber eben nicht schritt-destilliert – deshalb gilt für sie die normale Faustregel.

Machen mehr Schritte das Bild hochauflösender?

Nein. Die Schrittzahl bestimmt nur, wie fein der Entrauschungsweg abgetastet wird. Die Bildgröße ist eine eigene Einstellung, und für echtes Vergrößern gibt es eigene Verfahren. Ein Bild wird durch 80 Schritte keinen einzigen Bildpunkt größer – siehe [Auflösung & Seitenverhältnis bei KI-Bildern](#).

Warum wird mein Bild bei jedem Erhöhen wieder anders, ohne sich zu beruhigen?

Dann läuft vermutlich ein Sampler aus der ancestral- oder SDE-Familie. Diese Verfahren mischen in jedem Schritt frisches Rauschen zu und konvergieren deshalb grundsätzlich nicht: Auch bei 200 Schritten kommt kein stabiles Ergebnis heraus, sondern immer wieder ein anderes. Für Messreihen und für die Produktion gehört ein deterministisches Verfahren in die Vorlage.

Hängen Schritte und Sampler zusammen?

Eng. Verfahren höherer Ordnung schätzen die Krümmung des Pfades mit und treffen ihn mit weniger Stützstellen; einfache Verfahren brauchen mehr davon. Derselbe Prompt kann mit dem einen Sampler bei 18 Schritten fertig sein und mit dem anderen erst bei 30. Wenn dir die Rechenzeit zu hoch ist, lohnt der Wechsel des Verfahrens oft mehr als das Herunterdrehen der Schrittzahl.

Gilt die Schrittzahl bei Bild-zu-Bild genauso?

Nicht direkt, weil dort die Denoising-Stärke mitspielt – und die Werkzeuge rechnen unterschiedlich. In der WebUI-Familie ergibt sich die tatsächliche Schrittzahl aus eingestellten Schritten mal Stärke; 20 Schritte bei Stärke 0,5 sind rund zehn gerechnete. In ComfyUI werden die eingestellten Schritte tatsächlich gerechnet, allerdings über einen verkürzten Rauschplan. Wer Werte zwischen den Werkzeugen überträgt, sollte das wissen.

Wie finde ich den Sättigungspunkt für mein Modell?

Mit einer Step-Leiter: alles festnageln außer der Schrittzahl, dann 4, 8, 12, 16, 20, 28 und 40 Schritte rendern und paarweise in voller Größe vergleichen. Der erste Wert ohne handwerklichen Zugewinn ist dein Sättigungspunkt; nimm zwei bis vier Schritte Sicherheitsaufschlag. Wichtig ist der feste Seed, sonst misst du den Zufall – warum, steht in [Der Seed bei der KI-Bildgenerierung](#).

Warum bieten Cloud-Dienste keinen Schrittreger?

Weil sie die Einstellung für ihr Modell selbst optimieren und dir das Ergebnis abrechnen. Die Bildmodelle der Gemini- und GPT-Familien kennen Auflösung und Qualitätsstufen, aber keine Sampling-Schritte; Midjourney hat einen Qualitätsparameter ohne offene Schnittstelle. Die Ausnahme ist FLUX.2 [flex], wo die Schrittzahl von 4 bis 50 offensteht. Wer die Kontrolle über diesen Kostenhebel braucht, landet in der Regel bei einem lokal betriebenen Modell.

Kann ich eine gefundene Schrittzahl einfach übernehmen?

Innerhalb desselben Aufbaus ja, darüber hinaus nicht. Die Zahl gilt für eine bestimmte Kombination aus Modellversion, Sampler, Scheduler, Auflösung und Guidance. Ein Modellwechsel, ein anderer Scheduler oder ein deutlich größeres Format verschieben den

Sättigungspunkt. Deshalb gehört die Zahl nie allein ins Protokoll, sondern immer mit dem Kontext, unter dem sie gemessen wurde.

Fazit

Sampling-Schritte sind der Regler, bei dem das Bauchgefühl am zuverlässigsten in die Irre führt. Er sieht aus wie ein Qualitätsregler, verhält sich aber wie die Auflösung eines Näherungsverfahrens: Er bestimmt, wie genau ein feststehender Weg abgetastet wird. Ab einer gewissen Feinheit ist genauer nicht mehr besser, sondern nur noch länger.

Die Merksätze passen auf eine Karteikarte. Ein Schritt ist eine Stützstelle, keine Schicht Lack. Jede Schrittzahl erzeugt ihren eigenen Rauschplan, deshalb ist anders nicht gleich besser. Nicht destillierte Modelle sättigen typischerweise zwischen 20 und 30 Schritten, schritt-destillierte haben einen Sollwert von vier bis acht, den man nicht anfasst. Und der Sättigungspunkt ist messbar – in zwanzig Minuten, einmal pro Modell und Auflösung.

Der Rest ist Betriebsdisziplin: die gemessene Zahl zusammen mit Modell, Sampler und Auflösung festhalten, sie in der Vorlage hinterlegen und beim Modellwechsel neu prüfen, statt sie blind mitzuschleppen. Das kostet einmal Aufwand und spart bei jeder Bildstrecke Rechenzeit, die sonst nichts einbringt außer einer längeren Kaffeepause. Wenn dein Team das sauber lernen soll, ist [KI-Bildgenerierung Schulung](#) der kürzeste Weg dorthin.

Weiterlesen

- › Alles Übrige zum Thema im Überblick: [KI-Bildgenerierung](#)
- › Das Verfahren hinter den Schritten: [Sampler & Scheduler erklärt](#)
- › Der Regler, der oft für das Steps-Problem verantwortlich ist: [CFG-Scale: Guidance in KI-Bildgeneratoren](#)
- › Die Zahl, die beim Messen festbleiben muss: [Der Seed bei der KI-Bildgenerierung](#) und [batch_size und Reproduzierbarkeit](#)
- › Welches Modell wie viele Schritte braucht: [Die 15 wichtigsten Modelle zur lokalen Bildgenerierung](#) und [Übersicht der wichtigsten KI-Bildgeneratoren](#)
- › Vom Einzelbild zur Produktion: [Aufbau einer skalierbaren Bild-Maschine mit ComfyUI und Flux](#) und [KI-Bildgenerierung im Unternehmen – Die strategische Alternative zur Cloud](#)
- › Wenn Bilder zur Marke passen sollen: [Markenkonsistente KI-generierte Bilder](#)
- › Warum die Bilder hier selbst erzeugt sind: [Dogfooding: Ich setze ein, was ich berate](#)