

■ // MINIMAX H3 · LOKALES SETUP

H3 lokal laufen lassen — und die Lizenz-Falle für die EU

MiniMax hat H3 als Open Weights veröffentlicht: Video mit nativem Stereo-Ton, bis 15 Sekunden am Stück. Quantisiert passt das Modell auf eine einzelne Grafikkarte. Die Community-Lizenz nimmt die EU allerdings ausdrücklich aus. Hier stehen beide Seiten — das echte Setup und der Weg, der von Deutschland aus offen bleibt.

OPEN WEIGHTS

GGUF AB 6,3 GB

EIN BEFEHL

EU-LIZENZ GEPRÜFT

QUELLEN INKLUSIVE

Worum es geht: H3 erzeugt Bild und Ton **gemeinsam in einem Durchgang** — der Ton wird nicht nachträglich daruntergelegt. Das Modell ist ein 33-Milliarden-Parameter-Transformer; als Text-Encoder hängt Qwen3-VL-32B daran. In voller Genauigkeit ist das Rechenzentrums-Hardware. Über quantisierte Community-Builds schrumpft der Teil, der auf die Grafikkarte muss, auf gut 6 GB. Bevor du das aufsetzt, lies Abschnitt 2 — er entscheidet, ob der lokale Weg für dich überhaupt offensteht.

1 Was H3 tatsächlich ausgibt

Die Eckdaten aus der offiziellen Model Card: **4 bis 15 Sekunden** Länge, **24 Bilder pro Sekunde**, **32 kHz Stereo**. Die kurze Bildkante liegt standardmäßig bei 768 Pixeln. Seitenverhältnisse von 21:9 bis 9:16 sind möglich. Für Dialog nennt MiniMax elf stabil unterstützte Sprachen, Deutsch ist dabei.

Es gibt zwei getrennte Checkpoints — kein Schalter, sondern zwei Downloads. **FL2VA** nimmt Text und wahlweise ein erstes und/oder letztes Einzelbild. **Ref2VA** nimmt Referenzen: bis zu neun Bilder, drei Videoclips und drei Audioclips, zusammen höchstens zwölf Dateien und in Summe 15 Sekunden Material.

Die 2K-Ausgabe kommt nicht aus dem lokalen Modell. Dafür schickt MiniMax das 768p-Ergebnis mit dem Originalkontext erneut durch H3 — dieses Modul heißt **H3-Regenerate-2K** und ist **nicht** Teil der Veröffentlichung. Lokal endest du bei 768p.

2 Die Lizenz — und warum das für dich in Deutschland zählt

H3 steht nicht unter Apache oder MIT, sondern unter der **MiniMax H3 Community License Agreement**. Deren Reichweite ist räumlich begrenzt. Wörtlich definiert der Vertrag:

```
// Abschnitt I, Definition 5
"Excluded Territories" means the European Union, the United Kingdom,
the Republic of Korea and the United States of America.
```

Die Rechte werden ausdrücklich nur innerhalb des „Applicable Territory“ eingeräumt, und das ist weltweit *ohne* diese vier Gebiete. Abschnitt V.4 zieht das noch weiter: Auch die **erzeugten Videos** sollen außerhalb des Gebiets nicht verwendet werden. Für dich in Deutschland heißt das: Die offenen Gewichte sind zwar herunterladbar, die Lizenz deckt die Nutzung hier aber nicht ab.

MiniMax nennt in einem eigenen Q&A den Grund — die Regulierung rund um Video- und Abbild-Erzeugung, unter anderem der EU AI Act — und schreibt dazu: die Einschränkung bedeute „*not yet*“, *nicht* „*not ever*“.

Was in der EU offensteht

- **Die API:** weltweit nutzbar, weil MiniMax die Infrastruktur selbst betreibt — platform.minimax.io
- **Die Web-App Hailuo:** hailuoai.video/tools/minimax-h3
- **Eine eigene Lizenz beantragen:** MiniMax vergibt Genehmigungen nach Prüfung des Einsatzszenarios — platform.minimax.io/h3-license

Zwei Punkte gelten unabhängig davon: Ab **20 Millionen US-Dollar Jahresumsatz** brauchst du eine gesonderte schriftliche Freigabe, und kommerzielle Produkte müssen „MiniMax H3“ sichtbar in der Oberfläche nennen. Eine KI-Kennzeichnung an erzeugten Dateien empfiehlt der Vertrag ausdrücklich.

3 Hardware — der offizielle Weg und der Consumer-Weg

Das offizielle Deployment-Beispiel von MiniMax startet mit `--num-gpus 4`. Vier Karten, volle BF16-Genauigkeit — das ist der Referenzpfad, nicht dein Wohnzimmer.

Der Consumer-Weg führt über quantisierte Builds. Unsloth veröffentlicht GGUF-Varianten, bei denen der Denoiser deutlich schrumpft. Die Größen für den FL2VA-Denoiser:

```
// Denoiser (fl2va_pruned) – Dateigröße
Q2_K          6,26 GiB    Q5_0          12,97 GiB
UD-Q2_K_XL    7,51 GiB    Q6_K          15,45 GiB
Q3_K          8,16 GiB    Q8_0          19,97 GiB
Q4_K          10,64 GiB
```

Dazu kommt der Text-Encoder Qwen3-VL-32B als eigene Datei: **12,20 GiB** in Q2_K_M oder **16,97 GiB** in Q4_K_M. Der entscheidende Kniff: Er muss nicht auf die Grafikkarte. Mit `--backend te=cpu` bleibt er im Arbeitsspeicher, und auf der Karte liegt nur der Denoiser. Die kleinste Stufe paart man mit dem Q2-Encoder, alles ab Q3 mit dem Q4-Encoder. Die VAEs sind für beide Checkpoints dieselben und liegen bei `Comfy-Org/MiniMax-H3`.

Unsloth dokumentiert als Beispiel: `UD-Q2_K_XL`, 960×544, 124 Bilder, 24 FPS, 8 Schritte — **auf einer einzelnen Karte**.

4 Der Befehl — und die drei Flags, ohne die es abbricht

So sieht ein Durchlauf mit `stable-diffusion.cpp` aus:

```
sd-cli --mode vid_gen \
  --diffusion-model minimax_h3_fl2va_pruned-UD-Q2_K_XL.gguf \
  --llm qwen3vl_32b_minimax_h3-Q2_K_M.gguf \
  --vae minimax_h3_video_vae_fp16.safetensors \
  --audio-vae minimax_h3_audio_vae_fp32.safetensors \
  --prompt "a red fox trotting through falling snow, cinematic" \
  --width 640 --height 384 --video-frames 25 --steps 4 --cfg-scale 1.0 \
  --backend te=cpu --diffusion-fa \
  --output out.webm
```

Drei Angaben sind nicht optional, und an genau denen scheitern die meisten ersten Versuche:

- `--mode vid_gen` — fehlt es, interpretiert H3 den Bildpfad falsch und bricht ab.
- `--cfg-scale 1.0` — die veröffentlichten Gewichte sind CFG-destilliert und arbeiten ohne Guidance. Der Standardwert 7.0 führt zum Abbruch.
- `--backend te=cpu` — hält den 12-GB-Text-Encoder von der Grafikkarte fern.

Reicht der Speicher trotzdem nicht, kommt `--offload-to-cpu` dazu. Wer lieber klickt als tippt: In ComfyUI gibt es offizielle Vorlagen für Text-zu-Video und Referenz-zu-Video, die Gewichte liegen unter `Comfy-Org/MiniMax-H3`.

5 Warum deine Ergebnisse schlechter aussehen als die Demos

Zwischen deinem Prompt und dem Modell sitzt bei MiniMax noch eine Stufe: `H3-Context-IR`. Sie zerlegt die Eingabe, verknüpft Text, Bild und Ton miteinander und schreibt daraus eine strukturierte Fassung, die H3 direkt versteht. MiniMax nennt sie „entscheidend für die Qualität des Ergebnisses“ — und hat sie **nicht** mitveröffentlicht.

Lokal fällt diese Stufe also weg. Wer die Lücke schließen will, arbeitet die Prompting Guidance im Repo durch oder nutzt die offiziellen Skills unter `github.com/MiniMax-AI/MiniMax-H3/tree/main/skills`. Das erklärt den größten Teil des Unterschieds zwischen den Demo-Clips und dem ersten eigenen Versuch.

6 Quellen

- **Model Card und Gewichte:** huggingface.co/MiniMaxAI/MiniMax-H3
- **Lizenztext:** huggingface.co/MiniMaxAI/MiniMax-H3/blob/main/LICENSE
- **Q&A von MiniMax zur Lizenz:** huggingface.co/MiniMaxAI/MiniMax-H3/blob/main/docs/QA-about-License.md
- **GGUF-Quantisierungen:** huggingface.co/unsloth/MiniMax-H3-GGUF
- **ComfyUI-Gewichte und Vorlagen:** huggingface.co/Comfy-Org/MiniMax-H3
- **Repo:** github.com/MiniMax-AI/MiniMax-H3

■ // AUF EINEN BLICK

Die Kurzfassung

Vier Punkte, bevor du etwas herunterlädst.

- ✓ H3 erzeugt Bild und **Stereo-Ton gemeinsam** — bis 15 Sekunden, 24 FPS, lokal bei 768p
- ✓ Die Lizenz nimmt die **EU ausdrücklich aus** — hier führen API, Hailuo oder ein Lizenzantrag zum Ziel
- ✓ Quantisiert reicht **eine Karte:** Denoiser ab 6,26 GiB, Text-Encoder per Flag auf die CPU
- ✓ `--cfg-scale 1.0` ist Pflicht — die Gewichte sind destilliert, der Standardwert bricht ab

■ // BLEIB DRAN

KI-Releases, mit dem Kleingedruckten.

Ich lese bei solchen Veröffentlichungen auch die Lizenz, bevor ich etwas empfehle. Folg mir für die nächste Einordnung.

@ki.borish

Fakten verifiziert am 31.08.2026 aus der Model Card und dem Lizenztext unter huggingface.co/MiniMaxAI/MiniMax-H3 sowie der Model Card von huggingface.co/unsloth/MiniMax-H3-GGUF. Die Angaben zu Dateigrößen und Befehlen stammen von den jeweiligen Anbietern und können sich mit neuen Releases ändern. Dieser Guide fasst öffentlich einsehbare Lizenzbedingungen zusammen und ist keine Rechtsberatung — prüfe den Lizenztext für deinen Fall selbst oder lass ihn prüfen.