

Diffusion Acceleration - Cheat Sheet

Fora Soft Learn · AI for Video Engineering

1 - The step tax (why this matters)

cost per clip $\approx$ cost per step $\times$ number of steps
 50 steps $\rightarrow$ $\sim$ 50 units of GPU work 4 steps $\rightarrow$ $\sim$ 4 units $\Rightarrow$ $\sim$ 12x less work, cost & wait

2 - The one idea: distillation

- ☐ A slow, accurate TEACHER model trains a fast few-step STUDENT to jump to the same answer.
- ☐ LoRA = a small stackable patch (tens of MB) you clip onto a base model without retraining it.
- ☐ CFG (classifier-free guidance) = the prompt-pushing dial; distilled models bake it in - turn it OFF (usually 1.0).

3 - The methods (read 'best for' first)

Method	Steps	Best for
LCM / LCM-LoRA	1-4	Fastest add-on to an existing image model (stackable)
SDXL-Turbo (ADD)	1-4	Single-step proof of speed
SDXL-Lightning	1-8	Tunable speed/quality on SDXL
Hyper-SD	1-8	Best single-step image quality
AnimateDiff-Lightning	1-8	Fast short video across many styles
Wan2.2-Lightning	4-8	Accelerate an open video model you run
CausVid / Self Forcing	few	Real-time, streamed, interactive video

4 - Pick the step budget from the experience, then the method

- ☐ Batch / overnight (nobody waiting): use full steps (25-50), no acceleration - quality is free.
- ☐ Interactive (user taps, watches a spinner): 4-8 steps via a distilled model. Image $\rightarrow$ LCM-LoRA, or Hyper-SD for best 1-step. Video $\rightarrow$ AnimateDiff-Lightning, or a Lightning distill on your model.
- ☐ Live / real-time (reacts as the user moves): causal streaming - CausVid / Self Forcing, or a real-time-native model like LTX-2.

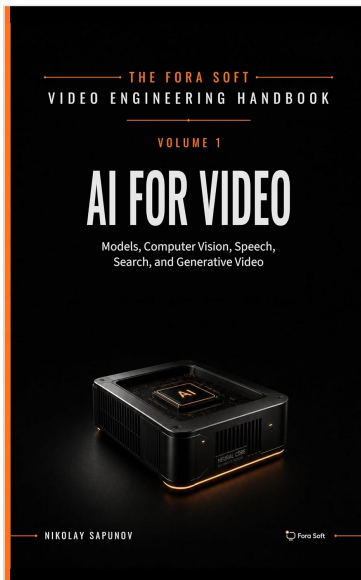
5 - Pitfalls to check before you build

- ☐ Washed-out distilled clips almost always mean CFG was left turned up - set it to the distilled model's value (usually 1.0).
- ☐ Fewer steps trade away fine detail and motion richness; 4 steps is the common production sweet spot, 1 step is the sharpest trade.
- ☐ Licences move: a distilled LoRA or model version can carry a different licence than its base - confirm the EXACT version you ship.

Companion to: 'Acceleration Tricks - LCM, Hyper-SD, AnimateDiff-Lightning'

www.forasoft.com/learn/ai-for-video-engineering/articles-ai/diffusion-acceleration-lcm-hypersd-animatediff-lightning

Fora Soft · www.forasoft.com



GO DEEPER · VOLUME 6 OF 7

AI for Video: A Complete Guide to Intelligence in Video Products

From Models and Inference to Computer Vision, Speech, Search, and Generative Video

By Nikolay Sapunov, CEO at Fora Soft · 1,444 pages · Kindle & paperback on Amazon

[Read it on Amazon →](#)
amazon.com/dp/B0HH6TKPZ1

This sheet maps the step budget to the experience and picks an acceleration method from LCM to Lightning distills. Volume 6 covers generative video in depth, together with model artefacts, deployment topologies and the ML minimum for choosing between models.

WHAT YOU WILL FIND INSIDE

- Generative video models
- Model artefacts and licences
- Deployment topologies
- ML minimum for model choice

THE FULL HANDBOOK · 7 VOLUMES · 9,000+ PAGES · ALL ON AMAZON

Vol. 1 · Video Encoding — pixels and color to H.264, HEVC, AV1, quality metrics and AI

Vol. 2 · Video Streaming I — network transport, contribution and ingest, HLS, MPEG-DASH, CMAF

Vol. 3 · Video Streaming II — ABR, CDNs, players, WebRTC, DRM and the economics of delivery

Vol. 4 · Audio for Video I — PCM, codecs, loudness, accessibility, real-time voice processing

Vol. 5 · Audio for Video II — jitter buffers, lip sync, Atmos, quality metrics, neural codecs

Vol. 6 · AI for Video I — models and inference, computer vision, speech, search, generative video

Vol. 7 · AI for Video II — real-time pipelines, agents, serving, cost, compliance, verticals

Need generated video in seconds, not minutes?

We will pick the distillation for your step budget, check licences and benchmark quality and cost on your GPUs. 250+ video products shipped since 2005.

[Book a 30-min call →](#)
[WhatsApp →](#)
[Email us →](#)