Diffusion Acceleration - Cheat Sheet

Fora Soft Learn · AI for Video Engineering

1 - The step tax (why this matters)

```
cost per clip  ~=  cost per step  x  number of steps
50 steps -> ~50 units of GPU work    4 steps -> ~4 units    =>  ~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 -> LCM-LoRA, or Hyper-SD for best 1-step. Video -> 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.