ComfyUI cheat sheet

The files you download and the nodes that load them

plain meaning · the photographer version in yellow · what to remember in grey

A · Files

Model / checkpoint — the big file of numbers that actually draws. A checkpoint bundles model + text encoder + VAE. *The photographer you hire; she brings her helpers.* models/checkpoints

Diffusion model — the same model shipped alone, without helpers. Most new models come this way. *She shows up with just her camera.* models/diffusion_models

Text encoder (CLIP) — turns your words into numbers. 'CLIP' is the slot name; the file may be T5, Qwen, Gemma… *The translator between you and the photographer.* models/text_encoders

VAE — converts between the small working image (latent) and real pixels. Encode = in, Decode = out. *The darkroom.* models/vae · wrong family = odd colours or shape error

Latent — the compressed working image, ~8× smaller per side. Generation happens here. *The negative.* sizes in multiples of 8 (some models 16/32/64)

LoRA — a small add-on that nudges a model toward a style, face or skill. Must match the model family. *She took a weekend class. Same camera, new trick.* models/loras · trigger word · strength 0.6–1.0

ControlNet — steer with an image (pose, edges, depth) instead of only words. *You hand her a reference photo: 'stand like this'.* prompt = WHAT, ControlNet = WHERE

Upscaler model — enlarges an image and invents plausible detail. *The print lab.* models/upscale_models

A · Formats

FP16 / FP8 / GGUF Q8→Q4 — same model, fewer decimals. Smaller file, less VRAM, slightly rougher. Q8 ≈ perfect, Q4 visibly rough. *Same photo saved at 4K / 1080p / 720p.* file size + 2 GB ≈ VRAM needed · GGUF needs the GGUF loader nodes

safetensors — modern model format that cannot hide code inside. Prefer it over .ckpt. it is safer, not faster

Read the filename — family_version · precision (fp8) · variant (scaled) · packager tag (KJ) · extension. Only precision and extension change what you need. z_image_turbo_fp8_scaled_KJ.safetensors

B · Nodes in a basic workflow

Node — a box that does one job. Inputs left, outputs right, settings inside. *One crew member, one task.* double-click the canvas to search

Link — a wire from an output to an input. Colour = data type; wrong colour won't plug. output = power strip (many) · input = wall socket (one)

Workflow — all nodes + links, saved as JSON. Also hidden inside every PNG ComfyUI saves. *The shot list.* workflows never contain models · drag a PNG in to restore it

Load Checkpoint — one node, three outputs: MODEL, CLIP, VAE. *Hire the photographer with her helpers.*

Load Diffusion Model / Load CLIP / Load VAE — the three-node version for models shipped in pieces. Load CLIP's 'type' must match the model family. *Hire everyone separately.* GGUF files: Unet Loader (GGUF) + CLIP Loader (GGUF)

CLIP Text Encode ×2 — prompt → conditioning. One for what you want, one for what you don't. *The brief, and the 'please no' list.* the two nodes are identical; the KSampler socket decides

Conditioning — the encoded prompt (orange cable). turbo models ignore the negative → 'Conditioning Zero Out' as an empty plug

Empty Latent Image — a blank latent of the size you pick. *Choosing the blank photo paper.* swap for Load Image → VAE Encode to start from a picture

KSampler — the node that generates: model + positive + negative + latent → finished latent. *The shoot.* all the settings you'll argue about live here

VAE Decode / Encode — latent → pixels at the end; pixels → latent when starting from an image. *The darkroom, both directions.*

Save Image / Preview Image — Save writes to output/. Preview shows it (temp/, cleared). 'prompt has no output' = you forgot one

Load Image — a picture from input/. Outputs IMAGE and MASK; mask editor on right-click.

The seven-node pipeline

1 Load Checkpoint — hire

2 CLIP Text Encode (+ / −) — brief the translator

3 Empty Latent Image — choose the paper

4 KSampler — the shoot

5 VAE Decode — the darkroom

6 Save Image — deliver

The folders you actually touch

models/ — checkpoints · diffusion_models · text_encoders · vae · loras · controlnet · upscale_models wrong subfolder = 'value not in list' · press R after adding

input/ output/ temp/ — what you load · what you save · previews (cleared).

custom_nodes/ — one folder per node pack (Python from GitHub). Delete folder = uninstall. red node → Manager → Install Missing Custom Nodes → restart

user/default/workflows/ — where your saved JSON workflows live.

First-week errors

Value not in list — file not in the folder, or not refreshed (R).

Red node — missing custom node → Manager.

CUDA out of memory — too big: lower size/batch or use FP8/GGUF.

Shape mismatch — wrong VAE / encoder / LoRA family.

Error at node X — X is where it failed, not why. Walk one node upstream.

ComfyUI cheat sheet

The settings, the canvas, and the hardware

plain meaning · the photographer version in yellow · what to remember in grey

C · Inside the KSampler

Seed — the random starting noise as a number. Same seed + same settings = same image. *Which roll of the dice you started with.* control after generate: fixed / increment / randomize · fix it when comparing

Steps — how many denoising passes. More = cleaner, slower, then no better. *How many wipes of the foggy window.* classic 20–30 · turbo 4–8 · use the model page's number

CFG — how hard the model is pushed toward the prompt and away from the negative. *Tension on the rope: positive pulls, negative pushes.* turbo models: CFG 1 (negative lets go) · too high = burnt colours, six fingers

Sampler — HOW noise is removed each step (Euler, DPM++ 2M…). *Her walking style down the stairs.* Euler is the safe default

Scheduler — WHEN noise is removed: how much per step (simple, karras, beta…). *The height of each stair.*

Denoise — how much the input may change, 0–1. Only matters when starting from an image. *'Redraw this sketch but keep 70% of it.'* 1.0 = ignores the input (the usual mistake) · 0.3–0.7 for img2img

Batch size vs batch count — size = images in one go (more VRAM). count = run N times (more time).

How to test anything

The compare ritual — fix the seed → change ONE thing → run → compare with an Image Compare node. Bypass the node you're testing (Ctrl+B) for a clean before/after. *Lab coat optional.*

Where to download what

Civitai — LoRAs and community fine-tunes. Filter by BASE MODEL or it won't load. The page lists the trigger words. *The style agency.* → models/loras

Hugging Face — base models and their pieces: checkpoints, text encoders, VAEs. Open Files, pick your precision. *The warehouse.* → models/checkpoints · Comfy-Org repos are packed for ComfyUI

GitHub — custom nodes, and ComfyUI itself. Install through the Manager, not by hand. *The tool shop.* → custom_nodes · node packs are code: check who wrote it

D · Canvas words

Group — a coloured rectangle that moves nodes together. Cosmetic.

Bypass (purple, Ctrl+B) — wire runs through, node skipped. Downstream still works. use for A/B tests

Mute (grey, Ctrl+M) — scissors on the wire; downstream errors. *The node called in sick and told nobody.* rarely useful

Subgraph — nodes folded into one box. Double-click to open.

Template — a saved copy-paste of nodes.

Custom node / Manager — Python from GitHub in custom_nodes/. Manager is the app store for them. Install Missing Custom Nodes → restart

Queue / Run — Run adds the workflow to a queue. Only nodes whose inputs changed re-run.

Shortcuts — double-click: search · R: refresh files · Ctrl+B / Ctrl+M: bypass / mute · Alt-drag: clone · Ctrl+Shift+V: paste with links · Space+drag: pan

Keeping it working

Update rules — close ComfyUI first · never before a deadline · if it works, don't touch it · rollback exists. *It updates like a toddler: daily, and something breaks.*

PNG trick — every saved PNG carries its workflow. Drag it onto the canvas. Discord and Twitter strip it.

E · Hardware words

VRAM — memory on the graphics card. Everything fast happens here. *The photographer's desk.* file size + 2 GB ≈ what you need

RAM — system memory. Overflow goes here and slows down. *The filing cabinet in the hallway.* 32 GB comfortable, 64 GB rescues small cards

Offloading / dynamic VRAM — ComfyUI shuffling pieces between desk and hallway. Leave it alone in Part 1.

Attention (sage / flash / pytorch) — different maths engines for the same job. Faster ones can break some models. black image or buzzing audio → try the default

CUDA — Nvidia's language for the GPU. 'CUDA out of memory' = desk is full

Your tier

6–8 GB — everything in Part 1 with FP8 or GGUF.

12–16 GB — everything, FP8 comfortably; most video at 720p later.

24 GB+ — everything, FP16 if you like.

Mac / AMD — images via the Desktop app; fp8 files may not load.

No GPU — cloud (RunPod / RunningHub), pay per hour.

Model classes

Text → image · Edit · Inpaint · Video · Audio · LLM — six kinds of model. Most 'it doesn't work' is asking an edit model to inpaint, or a text-to-image model to edit. *Pick the right photographer for the job.* Part 2 goes deep on this

You don't need to remember any of this. You need to know where to look.