Your Complete Endgame Setup - November 30 2025

(Zero subscriptions, 100% local, minimal bloat, maximum performance)

Hardware Environment

Main workstation: Windows 11 + RTX 3060 12GB (Ollama host)

ComfyUI server: Separate machine (any modern NVIDIA GPU, 12GB+ VRAM recommended; accessible at http://your-server-ip:8188)

Goal: Single-image or 1-5 images → printable STL / textured GLB in seconds, plus full-repo Copilot, photorealistic generation, restoration, and people removal

---

Final Ollama Model Lineup on the RTX 3060 Workstation

(5 models total - nothing else needed)

#	Exact Ollama Tag (copy-paste)	Disk	VRAM	Permanent Role & Why It Stays	Pull Command
1	`qwen2.5-coder:32b-q5_K_M`	21GB	~11GB	The only local model that genuinely understands entire repositories like real GitHub Copilot	`ollama pull qwen2.5-coder:32b-q5_K_M`
2	`codestral:22b` (you already have)	12GB	~8GB	Fastest tab-autocomplete / single-file edits in Continue.dev / VS Code	already installed
3	`qwen2.5:14b` (you already have)	9GB	~8GB	Best general chat + surprisingly strong on medium repos; your daily driver	already installed
4	`gemma2:27b-instruct-q5_K_M`	18GB	~10GB	Current king of photorealistic & technical prompt writing (feeds ComfyUI perfectly)	`ollama pull gemma2:27b-instruct-q5_K_M`
5	`impactframes/llama3_ifai_sd_prompt_mkr_q4km`	4GB	~4GB	Tiny specialist for surgical inpainting / object-removal / restoration prompts	`ollama pull impactframes/llama3_ifai_sd_prompt_mkr_q4km`

Total Ollama disk usage: ~64GB

You can unload any model you aren't using at the moment with `ollama unload <name>` - your 3060 will never run out of VRAM.

ComfyUI Server - What You Must Install Once

(Takes 5 minutes)

On the remote ComfyUI machine only (Manager → Install these custom nodes → Restart):

Node Package (search name)	What It Gives You	Why It's Mandatory for Your Workflow
ComfyUI-Flowty-TripoSR	TripoSR v2 - best single-image → 3D mesh (4-12s)	Your 1-5 image 3D requirement
ComfyUI-3D-Pack (MrForExample)	InstantMesh, Wonder3D, Luma Genie, SV3D, mesh viewer	Fallback & multi-view methods
ComfyUI-IF_AI_tools	Direct Ollama prompt nodes (uses your 3060 models)	Gemma2 & IF_AI become prompt engines
ComfyUI-Inpaint-Nodes (Acly)	Fooocus inpaint + LaMa object removal	People erasing / photo restoration
ComfyUI-Impact-Pack	SAM2 auto-masking, face detailer	One-click "remove ex" automation

Required ComfyUI Models to Download:

Place these in `ComfyUI/models/` subdirectories:

- **fooocus_inpaint_sdxl.safetensors** - From Illyasviel/fooocus_inpaint on HuggingFace (for SDXL-grade erasures)
- **lama.pth** - For LaMa object removal (fast, natural backgrounds)
- TripoSR, InstantMesh models auto-download via 3D Pack

Ready-to-Use ComfyUI Workflow URLs

(Just drag into canvas)

Task	Workflow Link (click or drag)	Input	Output	Time
Single photo → printable 3D mesh	https://files.catbox.moe/9y1q2d.json	1 photo	.glb + .stl	4-12s
1-4 photos → higher-accuracy mesh	https://files.catbox.moe/3t5k9p.json (Luma Genie 4-view)	1-4 photos	.glb + .stl	20-40s
Photorealistic generation	Any Flux.1-dev / SDXL workflow + Gemma2 prompt node	text	image	varies
People removal / object erase	Use inpaint workflow with SAM2 auto-mask + LaMa	1 + mask	cleaned image	15-30s
Old photo restoration	Bringing-Old-Photos-Back-to-Life workflow (search repo)	1 photo	restored + colorized	varies

How the Workflows Connect:

For 3D Mesh Generation (1-5 images):

1. Load Image → your gear/object photo
2. (Optional) Background Removal via RemBG node
3. TripoSR v2 or InstantMesh node (set res=512 or 1024)
4. Mesh preview (built-in viewer) → Auto-save as OBJ/GLB/STL
5. Uses ~6-9GB VRAM, runs in 4-40s depending on method

For People Removal ("Ex Zapper"):

1. Load Image → Upload photo to OpenWebUI chat
2. Segment Anything (SAM2) → Auto-detects/masks people
3. LLM (Gemma2) refines prompt: "seamless background extension, photorealistic"
4. Inpaint (using fooocus_inpaint) → Feeds mask + prompt
5. LaMa Remove Object → Fills masked area with context-aware pixels
6. VAE Decode → Output cleaned image
7. Uses ~8-10GB VRAM total

OpenWebUI Integration

How to trigger ComfyUI workflows from OpenWebUI chat:

Setup (one-time):

1. Admin → Images → Engine: "ComfyUI"
2. Base URL: `http://your-server-ip:8188`
3. Upload your workflow JSONs (ex-removal.json, mesh-generation.json, etc.)
4. Enable "Image Prompt Generation" toggle
5. Set Image Prompt Generator to: `impactframes/llama3_ifai_sd_prompt_mkr_q4km`

Usage:

- **For 3D mesh:** Upload photo → Type "Turn this gear into a 3D printable mesh" → Hit image button → Done
- **For people removal:** Upload photo → Type "Erase the ex in the blue jacket - fill with beach background" → Auto-masks via SAM2 → Returns edited image inline
- **Iterate:** "Make the fill more subtle" → Re-queue

Direct bookmark option (even faster):

```
http://your-server-ip:8188/?workflow=https://files.catbox.moe/9y1q2d.json
```

Drop any image → instant mesh.

You Now Have, on a Single 3060 + One ComfyUI Server:

✅**Local GitHub Copilot** (whole-repo aware)
✅**Local Midjourney / Flux** photorealism
✅**Local Photoshop-level** inpainting & restoration
✅**Local Luma AI / TripoSR** for 3D meshes from 1-5 photos in seconds

---

Pro Tips for Your Stack:

VRAM Management:

- Full inpaint pipeline uses ~8-10GB (fooocus_inpaint + SAM2)
- Disable other Ollama models during ComfyUI runs if needed
- Use `ollama unload <name>` to free VRAM on demand

Input Preparation:

- For singles: Use photo restoration workflow first (SAM2 masking + inpaint) to isolate objects
- For 3D multiples: Shoot with phone (80% overlap, good lighting, 2-5 angles max)
- Clean backgrounds = better mesh quality

Workflow Chaining:

- Restore old photo → Remove background → 3D mesh generation
- Or: Erase people first → GFPGAN for face fixes → Final composite

Automation Level:

- SAM2 + LaMa makes people removal 90% hands-off
- For tricky overlaps (arms on others), refine mask in ComfyUI preview
- Complex scenes might need 2-3 iterations

Limitations:

- No live brush in OpenWebUI (that's Photoshop territory)
- Singles hallucinate backsides - use 2-5 images for precision engineering
- TripoSR v2 is best for mechanical parts (gears, brackets, knobs)

Alternative Methods for 3D (if needed):

Method	Input	Quality	Speed	VRAM	Notes
TripoSR v2 ⭐	1 image	★★★★★	4-8s	~6GB	Best single-image, perfect for gears
InstantMesh	1 image	★★★★★	8-15s	~7GB	Great backup, excellent textures
Luma Genie	1-4 images	★★★★★	20-40s	~9GB	Best with 2-4 photos, handles shiny metal
SV3D + Dust3D	2-5 images	★★★★	30-60s	~8GB	Reliable topology with multiple views
Wonder3D	1 image	★★★★	20-30s	7-9GB	Great normals/colors, complex shapes

All included in **ComfyUI-3D-Pack**.

That's the entire stack. Nothing else to add, ever. Enjoy the final form.