⚛ ComfyUV

This guide explains how to deploy the Docker image:

[sandichhuu/comfyuv on Docker Hub]

An optimized, production-ready ComfyUI Docker image built with `uv` for ultra-fast Python package management. It comes pre-configured with modern performance optimizations like Triton, SageAttention.

> ⚠ **Important Note**
> □□ I only have RTX 20, 30 series (SM75 & SM86). So newer series untested.
> □□ If you facing any issue, please send the error log to:
>
> □□ LinkedIn
>
> □□ Facebook
>
> □□ Whatsapp

□□ **Compatible GPU**	
□ **Newer**	Runnable, but slower.
□ **SM120**	RTX 50 Series.
□ **SM89**	RTX 40 Series.
□ **SM86**	RTX 30 Series.
□ **SM75**	RTX 20 Series.
□ **Unsupported**	GTX 10 Series and older, AMD or Intel GPU.

> ⚠ Unsupport **xformers**.
> Because cu132 have no pre-build for this package & have no reason to use the slowest method.
> □ You can use sage-attention or triton instead.

>

> ⚠ No longer support **flash-attn** (because the compile time is super slow).
> You can do it your self with this command:
>
> ```
> MAX_JOBS=10 NVCC_THREADS=1 uv pip install flash-attn --no-build-isolation
> ```

Why use this Docker image for ComfyUI?

Nodes Compatibility	Python 3.13.13, widely compatible with extensions.
Blazing Fast Speed	Uses `uv` and **CUDA 13.2** for ultra-fast compilation and execution.
Precompiled Nodes	Comes with precompiled ecosystem, saving your setup and compile time.
Strict Privacy	Completely stripped of telemetry. Your data and workflows remain private.
Cloud Optimized	Ready for public deployment. Use caddy or nginx for authentication.

PRE-INSTALLED Custom Nodes

calcuis/gguf	(calcuis/gguf)
ComfyUI-Crystools-MonitorOnly	(BobRandomNumber)
ComfyUI-VideoHelperSuite	(Kosinkadink)
rgthree-comfy	(rgthree)
ComfyUI-Lora-Manager	(willmiao)
Nvidia_RTX_Nodes_ComfyUI	(Comfy-Org)
ComfyUI-KJNodes	(kijai)

◻◻ Quick Start (Docker Compose)

Install **Docker**, **Nvidia Driver** (refer Studio version), **CUDA Toolkit** (version 13.2).
Next, copy **docker-compose.yml** content below and put to empty folder.
Then, open **Terminal/CMD/Windows PowerShell** and navigate to directory contain `docker-compose.yml` file.
Finally, run this command `docker compose up -d` to start the container. Go to `http://localhost:8188` and enjoy.

> ## Requirements
>
> ⬇ Docker
>
> ⬇ nvidia-driver
>
> ⬇ cuda-toolkit (CUDA 13.2)

docker-compose.yml

```
services:
  comfyuv:
    image: sandichhuu/comfyuv:sm86cu132
    container_name: comfyuv
    ipc: host
    port:
      - 8188:8188
    deploy:
      resources:
        reservations:
          devices:
            - driver: nvidia
              count: 1
              capabilities: [gpu]
    volumes:
      - ./input:/comfy/input
      - ./output:/comfy/output
      - ./workflows:/comfy/user/default/workflows
      - ./models:/comfy/models
      - ./workflows:/comfy/user/default/workflows

      - comfyuv:/comfy
      - uv_cache:/root/.cache/uv
    command: >
      --enable-triton-backend
#     --use-flash-attention unsupport becase it take long times to compile.
#     --lowvram
      --use-sage-attention
      --disable-pinned-memory
      --fast fp8_matrix_mult autotune
```

```
      --front-end-version Comfy-Org/ComfyUI_frontend@latest

volumes:
  comfyuv:
  uv_cache:
```

□□ Setup Structure

```
comfyuv/
??? docker-compose.yml
??? input/
??? models/
??? workflows/
??? output/
```

□□ Why have no all-in-one image ?

All GPU support int4, int8, fp16, fp32.
But each GPU series has different physics structure due to different kernels and different attention optimization.

For example RTX4090 support fp8 native, RTX5090 support fp8 & nvfp4 native.
RTX40xx fastest with FlashAttention3, RTX50xx with FlashAttention4.
If I compile everything on one image, the file size come super heavy.
And the compile time is slow as hell.

References

- ComfyUI Official Docs
- Docker Hub Image
- Docker Hub Documentation

Revision #14
Created 2026-05-23 23:22:15 UTC by sandichhuu
Updated 2026-05-28 06:03:25 UTC by sandichhuu