ComfyUI WAN 2.5 Serverless Setup Guide

This comprehensive tutorial covers the complete setup process for ComfyUI WAN 2.5 on RunPod, including network storage, model downloads, JupyterLab configuration, serverless deployment, Postman testing, and web app environment setup.

---

1. Create Required Accounts

- **RunPod** – for GPU pods and serverless functions: Referral link (Sign up and deposit $10 to receive potential credit bonus $5–$500)
- **HuggingFace** – required for accessing and downloading AI models: https://huggingface.co/
- **Civitai** – for community model downloads: https://civitai.com/

2. Fund Your RunPod Account

Ensure you have at least $10 in your RunPod account to spin up GPU pods.

3. Create Network Storage

- Navigate to **Network Volumes** in RunPod.
- Create a volume of at least 40 GB for persistent storage.

4. Deploy a Pod with GPU

- Click **Deploy Pod** from the dashboard.
- Select GPU (e.g., RTX A5000).
- Choose template: **ComfyUI Manager – Permanent Disk – torch 2.4**.
- Attach the network volume.

5. Launch the Environment

- Wait for pod installation.
- Open **Jupyter Notebook** via **Connect**.
- Run terminal command: `./run_gpu.sh`

6. Select Template

- Choose **WAN 2.2 Text to Video** template for ComfyUI setup.

7. Install Necessary Models

Download and install all required WAN 2.2 models for video generation.

a. wan2.2_t2v_high_noise_14B_fp8_scaled.safetensors

Location: /workspace/ComfyUI/models/diffusion_models/

```
curl -L -o
/workspace/ComfyUI/models/diffusion_models/wan2.2_t2v_high_noise_14B_
fp8_scaled.safetensors \
  "https://huggingface.co/Comfy-Org/Wan_2.2_ComfyUI_Repackaged/resolv
e/main/split_files/diffusion_models/wan2.2_t2v_high_noise_14B_fp8_sca
led.safetensors"
```

b. wan2.2_t2v_low_noise_14B_fp8_scaled.safetensors

Location: /workspace/ComfyUI/models/diffusion_models/

```
curl -L -o
/workspace/ComfyUI/models/diffusion_models/wan2.2_t2v_low_noise_14B_f
p8_scaled.safetensors \
  "https://huggingface.co/Comfy-Org/Wan_2.2_ComfyUI_Repackaged/resolv
e/main/split_files/diffusion_models/wan2.2_t2v_low_noise_14B_fp8_scal
ed.safetensors"
```

c. wan2.2_t2v_lightx2v_4steps_lora_v1.1_high_noise.safetensors

Location: /workspace/ComfyUI/models/loras/

```
curl -L -o
/workspace/ComfyUI/models/loras/wan2.2_t2v_lightx2v_4steps_lora_v1.1_
high_noise.safetensors \
  "https://huggingface.co/Comfy-Org/Wan_2.2_ComfyUI_Repackaged/resolv
e/main/split_files/loras/wan2.2_t2v_lightx2v_4steps_lora_v1.1_high_no
ise.safetensors"
```

d. wan2.2_t2v_lightx2v_4steps_lora_v1.1_low_noise.safetensors

Location: /workspace/ComfyUI/models/loras/

```
curl -L -o
/workspace/ComfyUI/models/loras/wan2.2_t2v_lightx2v_4steps_lora_v1.1_
low_noise.safetensors \
  "https://huggingface.co/Comfy-Org/Wan_2.2_ComfyUI_Repackaged/resolv
e/main/split_files/loras/wan2.2_t2v_lightx2v_4steps_lora_v1.1_low_noi
se.safetensors"
```

e. wan_2.1_vae.safetensors (VAE model)

Location: /workspace/ComfyUI/models/vae/

```
curl -L -o /workspace/ComfyUI/models/vae/wan_2.1_vae.safetensors \
  "https://huggingface.co/Comfy-Org/Wan_2.2_ComfyUI_Repackaged/resolv
e/main/split_files/vae/wan_2.1_vae.safetensors"
```

f. umt5_xxl_fp8_e4m3fn_scaled.safetensors

Location: /workspace/ComfyUI/models/text_encoders/

```
curl -L -o
/workspace/ComfyUI/models/text_encoders/umt5_xxl_fp8_e4m3fn_scaled.sa
fetensors \
  "https://huggingface.co/Comfy-Org/Wan_2.1_ComfyUI_repackaged/resolv
e/main/split_files/text_encoders/umt5_xxl_fp8_e4m3fn_scaled.safetenso
rs"
```

"

8. Test the Workflow

- Ensure all model names match downloaded files.
- Open ComfyUI and verify workflow.
- Test workflow to confirm no errors.

9. Move the Models Folder

```
mv /workspace/ComfyUI/models /workspace/
```

10. Clean Up Workspace

```
rm -rf /workspace/ComfyUI
```

Keep only the **models** folder.

11. Terminate the Pod

- Terminate pod to save costs.
- All files remain in Network Storage for serverless endpoint.

12. Upload to Private GitHub Repository

- Create private GitHub repo.
- Upload WAN serverless folder, Dockerfile, and snapshot.
- Avoid tracking large model files unless necessary.

13. Deploy as Serverless Endpoint

- Connect GitHub repo to RunPod.
- Add HuggingFace token for model access.

14. Configure Endpoint Settings

- Attach network storage.
- Add environment variables:

```
COMFY_POLLING_MAX_RETRIES=2000
COMFY_POLLING_INTERVAL_MS=500
```

15. Save and Wait

- Save configuration.
- Wait for deployment to complete.

16. Testing the Endpoint

Method 1: Postman

- Set Postman variables to match your serverless API endpoint.
- Update input keys and output paths according to deployed workflow.

Method 2: Custom Web App

- Update `.env` with proper API URL and tokens.
- Match environment variables with endpoint settings.

Notes & Best Practices

- If ComfyUI is modified, update Postman requests and web app code.
- Ensure model paths, API keys, and workflow names match deployed setup.
- Regularly check endpoint status and update models if needed.
- Use network storage for persistent data between serverless invocations.
- For web app, update `.env` variables such as `COMFY_API_URL`, `HF_TOKEN`, `STORAGE_PATH`.

End of ComfyUI WAN 2.5 Serverless Setup Guide