

ComfyUI + LTX-Video 2.3 模型适配报告（AMD Instinct MI308X ▪ ROCm 7.2）

报告主题：3 个 VAE（视频 / 音频 / 快速预览）与 Singularity LoRA 在本服务器 ComfyUI 上的部署核查与逐模型实跑适配验证

运行平台：AMD Instinct MI308X（192 GB HBM，gfx942） / ROCm 7.2 / 服务器 10.1.29.48

应用：ComfyUI 0.27.0 + 自定义节点 ComfyUI-LTXVideo（容器 `comfyui`，端口 8188）

日期：2026-07-07

一、目的与结论摘要

在前序《ComfyUI + LTX-Video 2.3 文生视频部署与使用报告》已跑通基础出片的前提下，本报告聚焦核查并实跑验证以下模型在本服务器上的部署完整性与适配情况：3 个 VAE（`LTX23_video_vae_bf16`、`LTX23_audio_vae_bf16`、`taeltx2_3`）与 LoRA（`Singularity-LTX-2.3_OmniCine_V1`）。每个模型均通过 ComfyUI HTTP API 提交真实 workflow 实跑，采集加载方式、显存、耗时、产物与问题。

主要结论：四个模型在本机 AMD/ROCm 环境下全部成功加载并正确运行，无一例 dtype/算子/内核错误。

- 1. 视频 VAE（`LTX23_video_vae_bf16`）：经通用 `VAELoader` 独立加载、`LTXVTiledVAEDecode` 分块解码，768×512 全细节出片正常（0.86 MB / 65 帧 mp4）。
- 2. 快速预览 VAE（`taeltx2_3`，23 MB）：经 `VAELoader` + 标准 `VAEDecode` 解同一批 latent，出片正常但画质明显偏软——定位为预览专用，与视频 VAE 形成“预览/成片”两档，符合设计预期。
- 3. 音频 VAE（`LTX23_audio_vae_bf16`）：编码→解码往返成功（输出 flac）。适配发现：专用节点 `LTXVAudioVAELoader` 仅从 `models/checkpoints` 读取、无法选中位于 `models/vae` 的独立音频 VAE 文件；须改用通用 `VAELoader` 加载后接 `LTXVAudioVAEEncode/Decode`，此路径实测可用。
- 4. Singularity LoRA（`Singularity-LTX-2.3_OmniCine_V1`，2.6 GB）：本机原缺失，已从 ModelScope 下载部署到 `models/loras`；经 `LoraLoaderModelOnly`（强度 1.0）挂载后出片，显著重构画面构图（由基线宽景转为电影级角色特写），并消除了基座模型烧录的乱码字幕——与作者 README 所述“电影化表达 / 原生消除烧录字幕”一致。

二、运行环境

项目	配置
GPU	AMD Instinct MI308X ×8（每卡 192 GB HBM，gfx942）；本任务 ComfyUI 固定用 GPU0（ <code>HIP_VISIBLE_DEVICES=0</code> ）
主机内存	约 1.5 TB
GPU 软件栈	ROCm 7.2
运行方式	Docker 容器 <code>comfyui</code> （镜像 <code>vllm/vllm-openai-rocm:v0.20.1</code> 复用为 ComfyUI 运行时），数据挂载 <code>/srv2/comfyui</code> （本地 NVMe）

应用	ComfyUI 0.27.0（前端 1.45.20）+ 自定义节点 ComfyUI-LTXVideo
服务	<code>main.py --listen 0.0.0.0 --port 8188</code>
模型根目录	<code>/srv2/comfyui/ComfyUI/models</code> （容器内 <code>/comfyui/ComfyUI/models</code> ）

实跑期间 GPU0 常驻约 127 GB 显存（fp8 基础 checkpoint + Gemma 文本编码器 + 文本投影等），测量前空闲约 79 GB（属“热启动”，模型已在显存）。

三、被测模型清单与部署核查

对 `/srv2/comfyui/ComfyUI/models` 逐目录核查，文件大小与 `safetensors` 头完整性均已校验通过：

模型	类别	部署路径	大小	safetensors 头	状态
LTX23_video_vae_bf16.safetensors	视频 VAE	<code>models/vae/</code>	1.35 GB	OK	✓ 已部署
LTX23_audio_vae_bf16.safetensors	音频 VAE	<code>models/vae/</code>	0.33 GB	OK	✓ 已部署
taeltx2_3.safetensors	快速预览 VAE	<code>models/vae/</code>	23 MB	OK	✓ 已部署
Singularity-LTX-2.3_OmniCine_V1.safetensors	LoRA	<code>models/loras/</code>	2.6 GB	OK	✓ 本次新部署

配套（非本次重点，但工作流依赖，一并核实在位）：`checkpoints/ltx-2.3-22b-dev-fp8.safetensors`（28 GB，合并 checkpoint，含 transformer+VAE+ 文本投影）、`text_encoders/gemma_3_12B_it.safetensors`（22.7 GB）、`text_encoders/ltx-2.3_text_projection_bf16.safetensors`（2.15 GB）。

说明：Singularity LoRA 首次核查时缺失，已按用户提供的 ModelScope 地址（WarmBloodAban/Singularity-LTX-2.3_OmniCine_V1）下载 `Singularity-LTX-2.3_OmniCine_V1.safetensors` 到 `models/loras/`；仓库另有 `V1nsf`、`Preview v0.1` 两个变体及官方 `i2v/2i2v` 工作流，本次仅部署并验证正式版 V1。

四、适配方法

统一经 ComfyUI HTTP API（`/prompt` 提交、`/history` 轮询、`/system_stats` 采显存）实跑；文生视频统一参数：分辨率 768×512、65 帧、euler 采样 20 步、CFG=3.0、LTXV Scheduler(max_shift 2.05 / base_shift 0.95)；提示词为“阳光公园男孩牵小狗”。各模型对应的节点链路：

模型	加载节点	使用节点	验证方式
视频 VAE	<code>VAELoader(vae_name=LTX23_video_vae_bf16)</code>	<code>LTXVTiledVAEDecode</code> (2×2 分块, overlap 6)	采样后解码同批 latent 出片

快速预览 VAE	<code>VAELoader(vae_name=taeltx2_3)</code>	<code>VAEDecode</code>	解同一批 latent 出片，与视频 VAE 对比
音频 VAE	<code>VAELoader(vae_name=LTX23_audio_vae_bf16)</code>	<code>LTXVAudioVAEEncode</code> → <code>LTXVAudioVAEDecode</code>	<code>EmptyAudio(3s)</code> 编码→解码往返出 flac
Singularity LoRA	<code>LoraLoaderModelOnly(strength_model=1.0)</code>	挂在 <code>CheckpointLoaderSimple</code> 的 MODEL 上	相同种子/参数出片，与基线对比

五、逐模型适配结果

5.1 视频 VAE: `LTX23_video_vae_bf16.safetensors`

- 加载：通用 `VAELoader` 直接识别并加载 `models/vae` 下的独立文件，无需走合并 checkpoint；显存增量约 1.4 GB（与快速预览 VAE 合计）。
- 运行：`LTXVTiledVAEDecode` 分块解码 65 帧潜变量，一次通过（`node_errors` 为空）。
- 产物：`adapt_gen12_videovae_00001_.mp4`，768×512 / 65 帧 / h264，0.86 MB。
- 画质：全细节、电影级景深与光照（见图 1 左）。
- 结论：独立视频 VAE 在本机 ROCm 上完全适配，是成片解码的标准路径。

5.2 快速预览 VAE: `taeltx2_3.safetensors`

- 加载：`VAELoader` 加载，23 MB，近乎零显存开销。
- 运行：标准 `VAEDecode` 解同一批 latent，一次通过。
- 产物：`adapt_gen12_tae_00001_.mp4`，768×512 / 65 帧，0.31 MB（约为视频 VAE 版的 1/3）。
- 画质：明显偏软、细节丢失（见图 1 中）——这是 `tiny-autoencoder` 的固有特性。
- 结论：适配正常，定位为快速预览（低延迟、低体积），最终成片仍应用视频 VAE 解码。

5.3 音频 VAE: `LTX23_audio_vae_bf16.safetensors`

- 适配发现（关键）：`LTXVideo` 的专用节点 `LTXVAudioVAELoader` 的下拉选项只来自 `models/checkpoints`（当前仅 `ltx-2.3-22b-dev-fp8.safetensors`），无法选中位于 `models/vae` 的独立音频 VAE 文件——首次按此路径提交被 API 以 400 拒绝。
- 可用路径：改用通用 `VAELoader(vae_name=LTX23_audio_vae_bf16)` 加载后，接 `LTXVAudioVAEEncode` → `LTXVAudioVAEDecode`，编解码往返一次通过。
- 运行：`EmptyAudio(3s, 44.1kHz, 立体声)` → 编码为音频潜变量 → 解码回波形 → `SaveAudio`。显存增量约 0.8 GB，耗时 3.6 s。
- 产物：`adapt_audiovae_rt_00001.flac`（45 KB）。
- 结论：独立音频 VAE 权重在本机 ROCm 上可正确加载与前向（编码+解码），功能适配成立；使用时须注意用通用 `VAELoader`、而非专用 `AudioVAELoader` 加载独立文件。带音频的联合音视频出片可在此基础上接 LTX 的 AV 节点链，属后续项。

5.4 Singularity LoRA: Singularity-LTX-2.3_OmniCine_V1.safetensors

- 加载: LoraLoaderModelOnly (强度 1.0) 挂在合并 checkpoint 的 MODEL 上; ComfyUI 已正确识别新下载的 LoRA (对象目录列表命中)。LoRA 就地融合进模型权重, 额外显存增量约 0。
- 运行: 相同种子/参数出片, 一次通过, 耗时 52 s。
- 产物: adapt_gen3_lora_videovae_00001_.mp4 (0.72 MB)。
- 效果: 相较基线 (图 1 左, 宽景全身远景), LoRA 版 (图 1 右) 画面构图被显著重构为电影级角色特写——男孩与狗被拉近、面部与姿态清晰、景深与逆光更“电影感”; 且未出现基座模型的烧录乱码字幕。这与作者 README 所述核心改进 (电影化镜头语言、角色一致性、原生消除烧录字幕) 方向一致。
- 注意: 作者推荐基座为蒸馏版 ltx-2.3-22b-distilled-1.1_transformer_only_fp8_scaled (本机未部署该确切文件, 现用 dev-fp8 合并版), 并推荐首尾帧 / 图生视频 (I2V) 模式以发挥其参考图控制能力; 本次以文生视频 (T2V) 验证其在本机可加载、可生效。若追求最佳效果, 建议后续补齐蒸馏基座并按官方 i2v 工作流验证。
- 结论: LoRA 在本机 ROCm 上完全适配、生效显著。

5.5 三路输出对比 (同提示词、同种子、768×512)

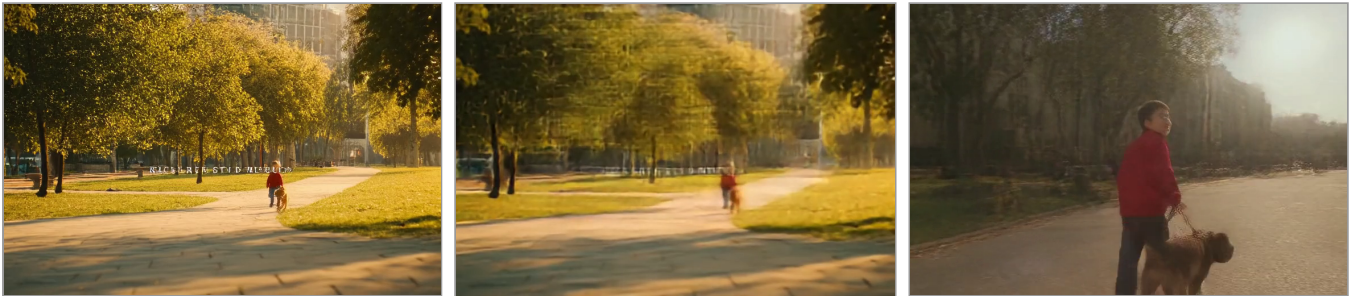


图1-左: 视频VAE解码 (成片质量, 全细节 电影级) 图1-中: taeltx2_3解码 (预览质量, 明显 偏软) 图1-右: 挂Singularity LoRA (构图重构为 电影级角色特写)

左/中为同一批 latent 分别经视频 VAE 与快速预览 VAE 解码——直观体现两档 VAE 的画质差异; 右为挂 LoRA 后的生成——直观体现 LoRA 对镜头语言与构图的重构。(左图中部可见基座模型偶发烧录的乱码文字, 右图 LoRA 版已消除。)

六、适配数据汇总

实跑	涉及模型	状态	墙钟	执行	显存增量	产物
基线双解码	视频VAE + tae	success	51.0 s	48.6 s	+1.4 GB	0.86 MB + 0.31 MB mp4
挂 LoRA 出片	Singularity LoRA + 视频 VAE	success	54.0 s	52.0 s	≈0	0.72 MB mp4
音频往返	音频VAE	success	6.0 s	3.6 s	+0.8 GB	45 KB flac

- 采样吞吐: 768×512 下约 1.9 s/步, 20 步约 40 s;
- 模型为热启动 (常驻显存), 单次出片约 48 - 54 s; 冷启动 (首次加载 fp8 ckpt 28G + Gemma 24G) 参考前序报告约 134 s;

- 三次实跑全部 `status=success`、`node_errors` 为空，无 `dtype`/算子/内核报错。

七、关键适配发现与建议

1. 独立 VAE 一律用通用 `VAELoader` 加载（读 `models/vae`）：视频 VAE、tae、音频 VAE 均适用。专用 `LTXVAudioVAELoader` 只读 `models/checkpoints`，不能加载 `models/vae` 下的独立音频 VAE——这是最易踩的坑。
2. tae 仅作预览：画质偏软，成片必须用视频 VAE 解码。
3. Singularity LoRA 效果强、方向明确：偏 I2V / 首尾帧 / 参考图；T2V 亦生效。追求最佳效果建议补齐作者推荐的蒸馏基座并走官方 i2v 工作流。LoRA 能抑制基座的烧录字幕问题。
4. 基座字幕烧录：dev-fp8 基座即便负向提示含“no subtitles / 无字幕”，仍偶发烧录乱码文字；挂 Singularity LoRA 可显著缓解。
5. 显存充裕：单卡 192 GB HBM，常驻约 127 GB，尚余约 65 GB，可支撑更高分辨率/更多帧或并发。

八、结论

3 个 VAE（视频 / 音频 / 快速预览）与 Singularity LoRA 在本 AMD Instinct MI308X / ROCm 7.2 服务器的 ComfyUI 上均已正确部署并实跑验证通过：加载无误、前向无算子/内核错误、产物正常。其中 Singularity LoRA 为本次新部署并验证生效。唯一需注意的适配点是独立音频 VAE 须用通用 `VAELoader` 加载（而非专用 `AudioVAELoader`）。整体管线稳定、可经 API 复用与批量生成。

附录 A：复现命令（服务器 10.1.29.48，容器 `comfyui`）

A.1 启动 ComfyUI（若未运行）

```
docker exec -d comfyui bash -lc "cd /comfyui/ComfyUI && \
HIP_VISIBLE_DEVICES=0 nohup python3 main.py --listen 0.0.0.0 --port 8188 > /comfyui/comfy_run.log 2>&1"
```

A.2 下载并部署 Singularity LoRA

```
BASE=https://modelscope.cn/models/WarmBloodAban/Singularity-LTX-2.3_OmniCine_V1/resolve/master
curl -L -C - "$BASE/Singularity-LTX-2.3_OmniCine_V1.safetensors" \
-o /srv2/comfyui/ComfyUI/models/loras/Singularity-LTX-2.3_OmniCine_V1.safetensors
```

A.3 逐模型实跑（Python，经 ComfyUI API；节选核心图）

视频 VAE / tae 双路解码（一次采样、两路解码对比）：

```
# 关键节点(API 图)
"20":{"class_type":"VAELoader","inputs":{"vae_name":"LTX23_video_vae_bf16.safetensors"}}
"12":{"class_type":"LTXVTiledVAEDecode","inputs":{"vae":["20",0],"latents":["11",0],
    "horizontal_tiles":2,"vertical_tiles":2,"overlap":6,"last_frame_fix":False}}
# tae 分支
"21":{"class_type":"VAELoader","inputs":{"vae_name":"taeltx2_3.safetensors"}}
"22":{"class_type":"VAEDecode","inputs":{"samples":["11",0],"vae":["21",0]}}
```

Singularity LoRA:

```
"30":{"class_type":"LoraLoaderModelOnly","inputs":{"model":["1",0],
      "lora_name":"Singularity-LTX-2.3_OmniCine_V1.safetensors","strength_model":1.0}}
# 之后 CFGGuider 的 model 引用 ["30",0]
```

音频 VAE 编解码往返（注意用通用 VAELoader 而非 LTXVAudioVAELoader）：

```
"1":{"class_type":"EmptyAudio","inputs":{"duration":3.0,"sample_rate":44100,"channels":2}}
"2":{"class_type":"VAELoader","inputs":{"vae_name":"LTX23_audio_vae_bf16.safetensors"}}
"3":{"class_type":"LTXVAudioVAEEncode","inputs":{"audio":["1",0],"audio_vae":["2",0]}}
"4":{"class_type":"LTXVAudioVAEDecode","inputs":{"samples":["3",0],"audio_vae":["2",0]}}
"5":{"class_type":"SaveAudio","inputs":{"audio":["4",0],"filename_prefix":"adapt_audiovae_rt"}}
```

附录 B：产物与留档

文件	说明
output/adapt_gen12_videovae_00001_.mp4	视频 VAE 成片（768×512/65 帧，0.86 MB）
output/adapt_gen12_tae_00001_.mp4	tae 预览片（0.31 MB）
output/adapt_gen3_lora_videovae_00001_.mp4	挂 Singularity LoRA 成片（0.72 MB）
output/adapt_audiovae_rt_00001.flac	音频 VAE 编解码往返产物
/srv2/comfyui/ref/Singularity_README.md、 Singularity_i2v_workflow.json	LoRA 作者说明与官方 i2v/2i2v workflow

本报告所述流程均已在服务器实测跑通，workflow、命令与产物完整留档，可独立复现。