

# Solutions for MeloTTS and Open-ai Whisper models, on web discussion for errors.

---

Run both models: Solution Using Method-3.

Melotts: <https://huggingface.co/myshell-ai/MeloTTS-Japanese>

Whisper Turbo: <https://huggingface.co/openai/whisper-large-v3-turbo>

using Rocm as a framework on the device provided, you are able to run on Radeon GPU.

References : Rocm docs

<https://rocm.docs.amd.com/projects/install-on-linux/en/latest/install/3rd-party/pytorch-install.html>

## Environment Setup

-  **Author:** `iilly@PC-AMD` ⇒ To check my environment “**wsl -d Ubuntu-22.04**” in Windows powershell
-  **Host OS:** Windows 11
-  **WSL2 Distro:** Ubuntu 22.04
-  **GPU:** AMD Radeon RX 9070 (Confirmed functional via ROCm & PyTorch)
-  **ROCM Version:** 6.4.1
-  **PyTorch Version:** 2.6.0 (Nightly ROCm Build with HIP 6.4.x)
-  **Python Version:** 3.10.18
-  **Frameworks & Tools:**
  - Hugging Face Transformers (`transformers`)
  - SoundFile (`soundfile`)

- TorchAudio (`torchaudio`)
- ROCm Toolkit (`rocminfo`, `hip-runtime`)
-  **Model Inference Tools:**
  -  `myshell-ai/MeloTTS-Japanese` (.Japanese TTS model — tested and exported `.wav` file)
  -  `openai/whisper-large-v3-turbo` (ASR model — successful speech-to-text inference)
-  **Conda Environments:**
  - `rocmtest` – ROCm/PyTorch validation and setup
  - `melotts` – MeloTTS text-to-speech pipeline
  - `whisperturbo` – Whisper Turbo transcription

## Task Progress Summary

Step	Task	Status
1	Set up WSL2 (Ubuntu 22.04)	 Done
2	Verified AMD Radeon RX 9070 GPU	 Done
3	Installed ROCm 6.4.1 via <code>.deb</code> package	 Done
4	Installed PyTorch ROCm wheels in Conda ( <code>rocmtest</code> )	 Done
5	Verified <code>rocminfo</code> detects CPU and GPU agents	 Done
6	Installed Python 3.10 + Conda env + ROCm PyTorch (2.6.0)	 Done
7	Verified <code>torch.cuda.is_available()</code> → <code>True</code>	 Done
8	Created <code>melotts</code> Conda environment, cloned MeloTTS from GitHub, and successfully ran inference on ROCm-enabled Radeon GPU (via <code>CUDA</code> )	 Done
9	Saved generated audio output ( <code>melotts_output.wav</code> ) successfully (some non-blocking warnings observed)	 Done
10	Created <code>whisperturbo</code> Conda environment, installed dependencies, and successfully ran <b>Whisper Large v3</b>	 Done

Step	Task	Status
	<b>Turbo</b> model on ROCm GPU with valid audio transcription	

# Outputs

## 1. Set up WSL2 (Ubuntu 22.04)

```
*** Done ***
(lmelotts) lilly@PC-AMD:~/miniconda3/envs/melotts/lib/python3.10/site-packages/torch/lib/MeloTTS$ cat /etc/os-release
PRETTY_NAME="Ubuntu 22.04.5 LTS"
NAME="Ubuntu"
VERSION_ID="22.04"
VERSION="22.04.5 LTS (Jammy Jellyfish)"
VERSION_CODENAME=jammy
ID=ubuntu
ID_LIKE=debian
HOME_URL="https://www.ubuntu.com/"
SUPPORT_URL="https://help.ubuntu.com/"
BUG_REPORT_URL="https://bugs.launchpad.net/ubuntu/"
PRIVACY_POLICY_URL="https://www.ubuntu.com/legal/terms-and-policies/privacy-policy"
UBUNTU_CODENAME=jammy
(melotts) lilly@PC-AMD:~/miniconda3/envs/melotts/lib/python3.10/site-packages/torch/lib/MeloTTS$ ..
```

conda environments :

```
(melotts) lilly@PC-AMD:~/miniconda3/envs/whisperferturbo/lib/python3.10/site-packages/torch/lib$ conda deactivate
(rocmtest) lilly@PC-AMD:~/miniconda3/envs/whisperferturbo/lib/python3.10/site-packages/torch/lib$ conda deactivate
(base) lilly@PC-AMD:~/miniconda3/envs/whisperferturbo/lib/python3.10/site-packages/torch/lib$ conda deactivate
# conda environments:
# 
base          * /home/lilly/miniconda3
melotts      /home/lilly/miniconda3/envs/melotts
rocmtest      /home/lilly/miniconda3/envs/rocmtest
whisperferturbo /home/lilly/miniconda3/envs/whisperferturbo
(base) lilly@PC-AMD:~/miniconda3/envs/whisperferturbo/lib/python3.10/site-packages/torch/lib$ ..
```

## 2. Verified rocminfo detects CPU and GPU agents

```
TRACT POLICY URL="https://www.ubuntu.com/legal/terms-and-policies/privacy-policy"
UBUNTU_CODENAME=jammy
(melotts) lilly@PC-4M0:~/miniconda3/envs/melotts/lib/python3.10/site-packages/torch/lib/MeloTTS$ rocminfo
WSL environment detected.

HSA System Attributes
Runtime Version: 1.1
Runtime Ext Version: 1.7
System Timestamp Freq.: 1000.000000MHz
Sig. Max Wait Duration: 18446744073709551615 (0xFFFFFFFFFFFFFFFFF) (timestamp count)
Machine Model: LARGE
System Endianness: LITTLE
Hwaitx: DISABLED
ENACK enabled: NO
DMAbuf Support: YES
VMM Support: YES

HSA Agents
*****
Agent 1
*****
Name: AMD Ryzen 9 3900X 12-Core Processor
Uid: CPU -A
Marketing Name: AMD Ryzen 9 3900X 12-Core Processor
Vendor Name: CPU
Feature: None specified
Profile: FULL_PROFILE
Float Round Mode: NEAR
Max Queue Number: 0(0x0)
Queue Min Size: 0(0x0)
Queue Max Size: 0(0x0)
Queue Type: MULTI
Node: 0
Device Type: CPU
Cache Info:
```

```
*****  
Agent 2  
*****  
  
Name: gfx1201  
Marketing Name: AMD Radeon RX 9070  
Vendor Name: AMD  
Feature: KERNEL_DISPATCH  
Profile: BASE_PROFILE  
Float Round Mode: NEAR  
Max Queue Number: 128(0x80)  
Queue Min Size: 64(0x40)  
Queue Max Size: 131072(0x20000)  
Queue Type: MULTI  
Node: 1  
Device Type: GPU  
Cache Info:  
    L1: 32(0x20) KB  
    L3: 65536(0x10000) KB  
Chip ID: 30032(0x7550)  
CacheLine Size: 64(0x40)
```

3. Verified `torch.cuda.is_available()` → True

#### 4. Successfully ran MeloTTS model on CUDA (Radeon GPU).

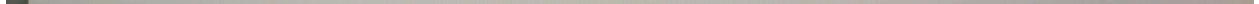
```
Environment Summary
PyTorch version : 3.10.18
PyTorch version 2.6.0+rocm6.4.1.git1ded221d
CUDA Available (ROCM) : True
CUDA Device count : 1
Device0 Name: AMD Radeon RX 9070
device0 Properties : _CudaDeviceProperties(name='AMD Radeon RX 9070', major=12, minor=0, gcnArchName='gfx1201', total_memory=16304MB
ROCM HIP version: 6.4.43483-a187df25c
(lilly@PC-AMD:~/miniconda3/envs/melotts/lib/python3.10/site-packages/torch/lib/MeloTTS$ python melotts_test.py
/home/lilly/miniconda3/envs/melotts/lib/python3.10/site-packages/huggingface_hub/file_download.py:943: FutureWarning: resume_download
  warnings.warn(
Using device:cuda
GPU:AMD Radeon RX 9070
> Text split to sentences.
Hello, this is Melo speaking in English!
> =====
0%
Some weights of the model checkpoint at bert-base-uncased were not used when initializing BertForMaskedLM: ['cls.seq_relationship.weight']
- This IS expected if you are initializing BertForMaskedLM from the checkpoint of a model trained on another task or with another architecture.
MIOpen(HIP): Warning [IsEnoughWorkspace] [GetSolutionsFallback WTI] Solver <GemmFwdRest>, workspace required: 5283840, provided ptr: 0
MIOpen(HIP): Warning [IsEnoughWorkspace] [EvaluateInvokers] Solver <GemmFwdRest>, workspace required: 5283840, provided ptr: 0
MIOpen(HIP): Warning [IsEnoughWorkspace] [GetSolutionsFallback WTI] Solver <GemmFwdRest>, workspace required: 49315840, provided ptr: 0
MIOpen(HIP): Warning [IsEnoughWorkspace] [EvaluateInvokers] Solver <GemmFwdRest>, workspace required: 77496320, provided ptr: 0
MIOpen(HIP): Warning [IsEnoughWorkspace] [GetSolutionsFallback WTI] Solver <GemmFwdRest>, workspace required: 77496320, provided ptr: 0
MIOpen(HIP): Warning [IsEnoughWorkspace] [EvaluateInvokers] Solver <GemmFwdRest>, workspace required: 77496320, provided ptr: 0
100%
TTS audio saved in the location:english-melotts.wav
pid:5712 tid:0x785a6542db80 [-VaMgr] frag_map_size is not 1.
(melotts) lilly@PC-AMD:~/miniconda3/envs/melotts/lib/python3.10/site-packages/torch/lib/MeloTTS$
```

#### 5. successfully ran Whisper Large v3 Turbo model on ROCm GPU with valid audio transcription

```
NameError: name 'Python' is not defined
(whispturbo) [1119@PC-AMD:~/miniconda3/envs/whispturbo/lib/python3.10/site-packages/torch/] $ nano whisptest.py
(whispturbo) [1119@PC-AMD:~/miniconda3/envs/whispturbo/lib/python3.10/site-packages/torch/] $ python whisptest.py
Running Whisper turbo

using device: cuda
GPU:AMD Radeon RX 9070
HIP version: 6.4.43483-a187df25c
Using custom 'forced_decoder_ids' from the (generation) config. This is deprecated in favor of the 'task' and 'language' flags/config options.
Transcription using a multilingual Whisper will default to language detection followed by transcription instead of translation to English. This might be a breaking change for your use case. I
for more details.
/home/lilly/miniconda3/envs/whispturbo/lib/python3.10/site-packages/transformers/integrations/sdpa_attention.py:66: UserWarning: Flash Efficient attention on Current AMD GPU is still experi
mers/hip/sdp_utils.cpp:256.)
    attn_output = torch.nn.functional.scaled_dot_product_attention(
/home/lilly/miniconda3/envs/whispturbo/lib/python3.10/site-packages/transformers/integrations/sdpa_attention.py:66: UserWarning: Mem Efficient attention on Current AMD GPU is still experi
mers/hip/sdp_utils.cpp:302.)
    attn_output = torch.nn.functional.scaled_dot_product_attention(
The attention mask is not set and cannot be inferred from input because pad token is same as eos token. As a consequence, you may observe unexpected behavior. Please pass your input's 'attenti
Transcription: Mr. Quilter is the apostle of the middle classes, and we are glad to welcome his gospel.
Whisper model is working well

pid:7174 uid:0x767611286b80 [-WaMqr] frag_map_size is not 1.
(whispturbo) [1119@PC-AMD:~/miniconda3/envs/whispturbo/lib/python3.10/site-packages/torch/] $
```



## Wonders.ai company test