**AUDICO – AI**

**TEXt2VIDEO PIPELINE**

1. **Goals:**

The objective of this project is to develop an innovative Generative AI strategy that focuses on both text and video-based content, with a primary goal of generating audio files that exhibit distinct personalities and reading styles. Through experimentation with various models, we aim to create diverse and engaging audio content that resonates with audiences. Additionally, our endeavor extends to generating accompanying video files that complement the audio content, providing a seamless and immersive user experience. The ultimate aim is to propose a robust end-to-end pipeline encompassing text-to-audio and video generation models, showcasing a working prototype that exemplifies the potential of Generative AI in shaping the future of multimedia content creation.

1. **Problem statement:**

Design an AI solution to generate personalized video stories for care home residents based on collected text stories. Analyse story context, user preferences, and sentiment to recommend relevant stories. Utilize text-to-speech and video synthesis models to create engaging audio narrations and corresponding visuals. Continuously improve the system through user feedback and provide an interactive interface for residents to access and enjoy the personalized content.

1. **Requirement analysis:**

* Determine the types of input data supported, such as text stories, scripts, or other content sources? *Input data for now will be plain text, I will provide a JSON file containing metadata about the story (e.g. name), as well as the main text*
* Identify the desired formats and characteristics of the input data, including text length, language, or specific content elements? *Language will be English to start. Text length may vary, but in general we try to use short stories. To give you an idea, the generated audio files are between 10-20 minutes long currently and the underlying stories are between 1500 - 3500 words. During testing, we can provide shorter samples of text to help speed up the development times.*
* Define the desired style, visual aesthetics, and overall feel of the generated videos? *As mentioned in the work plan, let's try to aim for 2 styles. A simple narration style, where a generated person is reading the story. Then also an illustration based style.*
* Specify the required elements to be included in the videos, such as background images, animations, or transitions?*No specific requirements here, although we could add a "watermark" with our logo at the start & end. If so, I will provide the necessary files*
* Determine the target video duration, aspect ratio, resolution, and other technical specifications? *Duration will be based on the length of the text. Aspect ratio should be 16:9. Let's experiment with this, I think that the illustration-based videos could potentially get away with a lower resolution, perhaps let's start with 360p for this and see how it looks. For narration-based videos, this may require a bit more resolution, maybe start with 480p*
* Consider the potential need for personalization based on user profiles, location, or other relevant factors? *Yes this is a good idea, this is certainly something we can think about*

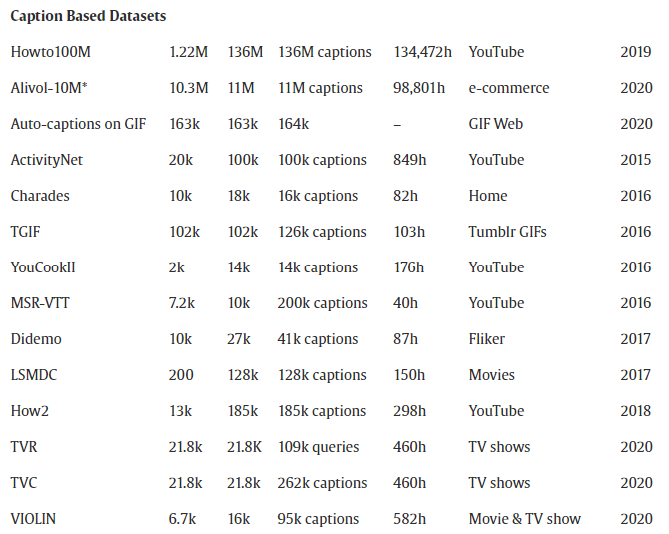
1. **Possible Solution:**

Employ advanced TTS models like **WaveNet**or Tacotron to generate high-quality and natural-sounding audio narration for the stories and to Implement techniques like **prosody transfer**to adjust the speaking style, intonation, or emotional expression of the TTS output based on the context or user preferences.

     Utilize deep learning models like **Generative Adversarial Networks (GANs) or Variational Autoencoders (VAEs)** to generate relevant visual content, such as scene images or background visuals, based on the story context along with exploring **video synthesis** techniques like VideoGPT or Video Transformer to generate dynamic visuals that align with the audio narration, incorporating the context and personalized elements.

1. **Meeting Notes on July 3rd:**
2. Which Pipeline will be better?
   1. Text – Speech – Video
   2. Text – Video:
3. Pipeline definition:
4. How to match the video’s frames per sec and audio’s sample/sec?
5. Preferred - Pre-trained models
   1. PoC
   2. Available on hugging face
   3. Feasibility study
6. Deployment
   1. to create the microservices or docker containers
   2. API which should be available to outside world
7. Perform the research, analyze the different models and implementation idea
8. Plan for next 4 to 5 weeks
9. **Datasets:**

* Need Image and Text pairs along with Video and Text pairs (to find Spatial-temporal dependencies)
* **Video and Text pairs:**
  + WebVid
  + [QuerYD](https://www.robots.ox.ac.uk/~vgg/data/queryd/) - focuses on the event localization task
  + [CelebV-Text](https://celebv-text.github.io/) is a large-scale facial text-video dataset of over 70K videos to generate videos with realistic faces, emotions, and gestures
  + Max-Planck Institute for Informatics (MPII) Movie Description dataset.



1. **Which Pipeline will be better?**
   1. **Text – Speech – Video**

* [microsoft / speecht5\_tts](https://huggingface.co/microsoft/speecht5_tts)
* TEXT-TO-SPEECH WITH TACOTRON2
* Wav LM
* **VALL-E – Problem – not available as pretrained model**
* Tacotron2, Deep Voice 3, and Transformer TTS
* Concatenative speech synthesis.
* Vocoders - convert spectrogram into a waveform
* Researchers have also used **GANs** to produce synthesized speech from text input. Advanced deep learning technologies like [Amazon Polly](https://aws.amazon.com/polly/) and [DeepMind](https://www.deepmind.com/) synthesize natural-sounding human speech. Such models operate directly on character or phoneme input sequences and produce raw speech audio outputs
* **Models: (NVdia)**
  1. **FastPitch and HiFiGAN:**

### The combination of FastPitch and HiFiGAN delivers end-to-end speech synthesis, where the FastPitch model produces a mel spectrogram from raw text, and HiFiGAN can **generate audio** from a **mel spectrogram**. Collectively, these pretrained models are ideal for a wide range of text-to-speech (TTS) applications such as audiobooks, voice cloning, and music generation.

<https://catalog.ngc.nvidia.com/models?filters=application|uscs_text_to_speech|Text%20to%20Speech&orderBy=scoreDESC&query=fastpitch%20hifigan>

* 1. **Text – Video:**

[Phenaki](https://huggingface.co/papers/2210.02399), [Make-A-Video](https://huggingface.co/papers/2209.14792), [NUWA](https://huggingface.co/papers/2111.12417), [VideoGPT](https://huggingface.co/papers/2104.10157) and [CogVideo](https://huggingface.co/papers/2205.15868)

* GAN/ VAE -- Transformer architectures – Diffusion based architectures
* Visualizations of thought tend to bring out a lot of the implicit context present in the explicit text.
  + Descriptions of kicking a ball can lead us to image soccer, green grass, and shorts

1. **Comparison:**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **S.No** | **Paper Name** | **Input** | **Differentiators** | **Based on** | **Git hub** | **Pretrained model** | **Benefits** |  |
| 1 | CogVideo | Text and Video | Sequential generation model and a frame interpolation model.  Multi frame rate hierarchical training. | Transformers |  |  | Long-range correlation.  Better align text and video frames | May 2022 (Pretrained on Text – Image model) |
| 2 | MakeAVideo | Text and Image pairs (Without Video data) | High resolution video generation without video data | Diffusion models |  |  | HR image with interpolation for coherent video. | Sep 2022 |
| 3 | Text to Video – CLIP | Text and Image |  | Cycle GANs |  |  |  | Oct 2022 |
| 4 | Make it move | Text and Image | Based on I2V |  |  |  |  | Not suitable for us. We want **T2V** |
| 5 | Tune A Video | Image and Text  One video | Sparse spatial-temporal attention  One shot video tuning | Stable diffusion | Yes |  |  | 17 Mar 2023 |
| 6 | **Text2Video-Zero** |  | without any training or optimization |  | Yes |  |  | 23 Mar 2023 |
| 7 | Align your latents | Finetune on Video data |  |  | Yes |  |  | Apr 2023 |
| 8 | **TaleCrafter** | S2P, T2L, C-T2I, I2V |  |  | **No** |  |  | May 2023 |
| 9 | Story Dall-e |  |  | transformers |  |  |  |  |
| 10 | Imagen Video |  |  |  |  |  |  |  |
| 11 | LDM |  |  |  |  |  |  |  |
| 12 | **Make-a-story** |  | LDM, coreference resolution |  | **Yes** |  |  |  |

1. **DLSS – nVidia**

Pioneering generative AI advances, NVIDIA presented [DLSS](https://developer.nvidia.com/rtx/dlss) (Deep Learning Super Sampling). It is a neural graphics technology to reconstruct images. The 3rd generation of DLSS increases performance for all GeForce RTX GPUs using AI to create entirely new frames and display higher resolution through image reconstruction.

1. **GAN/VAE:**

**Inputs needed –** Text and images

**Pipelines:**

* Vectorize and embed the text into latent space (fastText)
  + VAEs – to concatenate the vectors and **synthesize the new meaningful vectors**.
  + Siamese network – to find the relation between **text and image.**
* Use GANs to expand the text embeddings into a **series of images**.
  + [Text to Photo-realistic Image Synthesis with Stacked Generative Adversarial Networks](https://arxiv.org/pdf/1612.03242v1.pdf)
  + [Generating Videos with Scene Dynamics](http://carlvondrick.com/tinyvideo/paper.pdf)
  + Two GANS – first one to create images and the second one to create frames.
* Convert the **series into a GIF**

1. **Transformers (CogVideo): Published in Dec 2020**

**Inputs needed –** Needs video data

**Repo:** https://github.com/THUDM/CogVideo

**Pipelines:**

1. **Meta: (Make-a-Video) – Sep 2022**

**Inputs needed –** Text and Image pairs

**Pipelines:**

4 layer of diffusion models

1. **TuneAVideo – Mar 2023**
2. **Audio GEN:**

After introducing ‘Make-A-Video,’ a team of scientists from Meta have released ‘AudioGen’, jointly with the University of Jerusalem. An auto-regressive generative model that generates audio samples based on text inputs. It operates on learnt discrete audio representations;

1. **Development of text to Video field:**

**Hugging face:**

* Using Hugging Face Diffusers, you can easily download, run and fine-tune various pretrained text-to-video models, including Text2Video-Zero and ModelScope by [Alibaba / DAMO Vision Intelligence Lab](https://huggingface.co/damo-vilab). We are currently working on integrating other exciting works into Diffusers and Transformers.

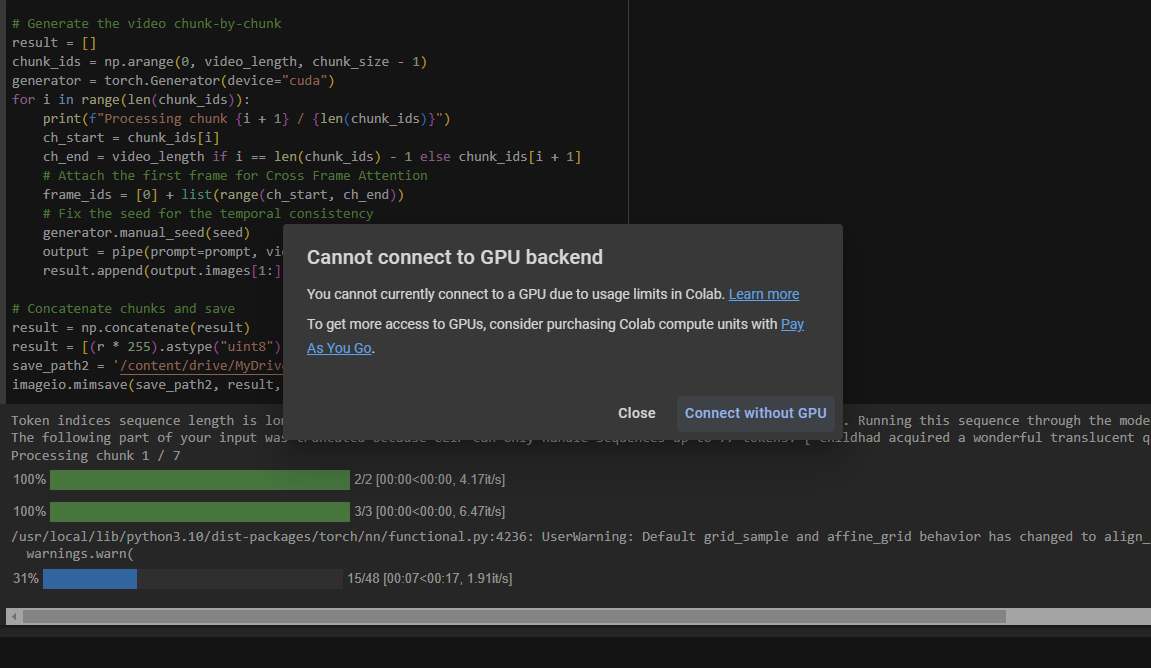
[VideoGPT](https://huggingface.co/spaces/akhaliq/VideoGPT), [CogVideo](https://huggingface.co/spaces/THUDM/CogVideo), [ModelScope Text-to-Video](https://huggingface.co/spaces/damo-vilab/modelscope-text-to-video-synthesis), and [Text2Video-Zero](https://huggingface.co/spaces/PAIR/Text2Video-Zero)

**Development plan:**

* Mostly comparative study of multiple models – find out which one will be the best models.
* Finding out the best pipelines
* Work on different embeddings

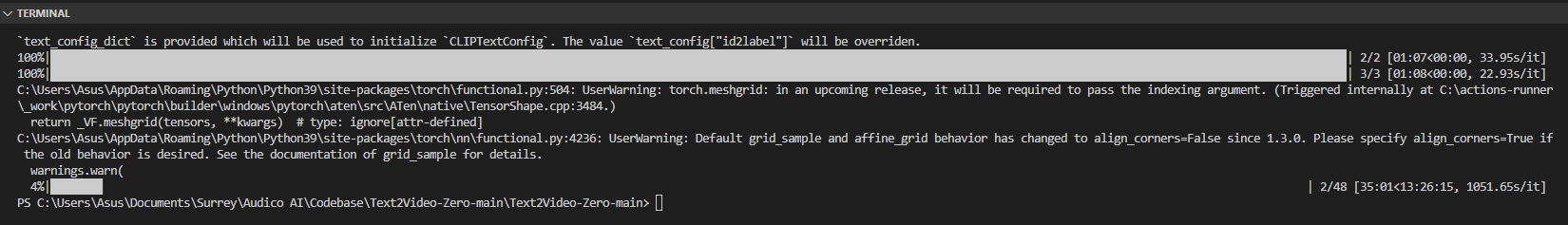
1. First Text2Video pipeline:
2. Second Text2Video pipeline:
3. Third Text2Video pipeline:
4. Debug Issues Faced:

**Text2Video:**

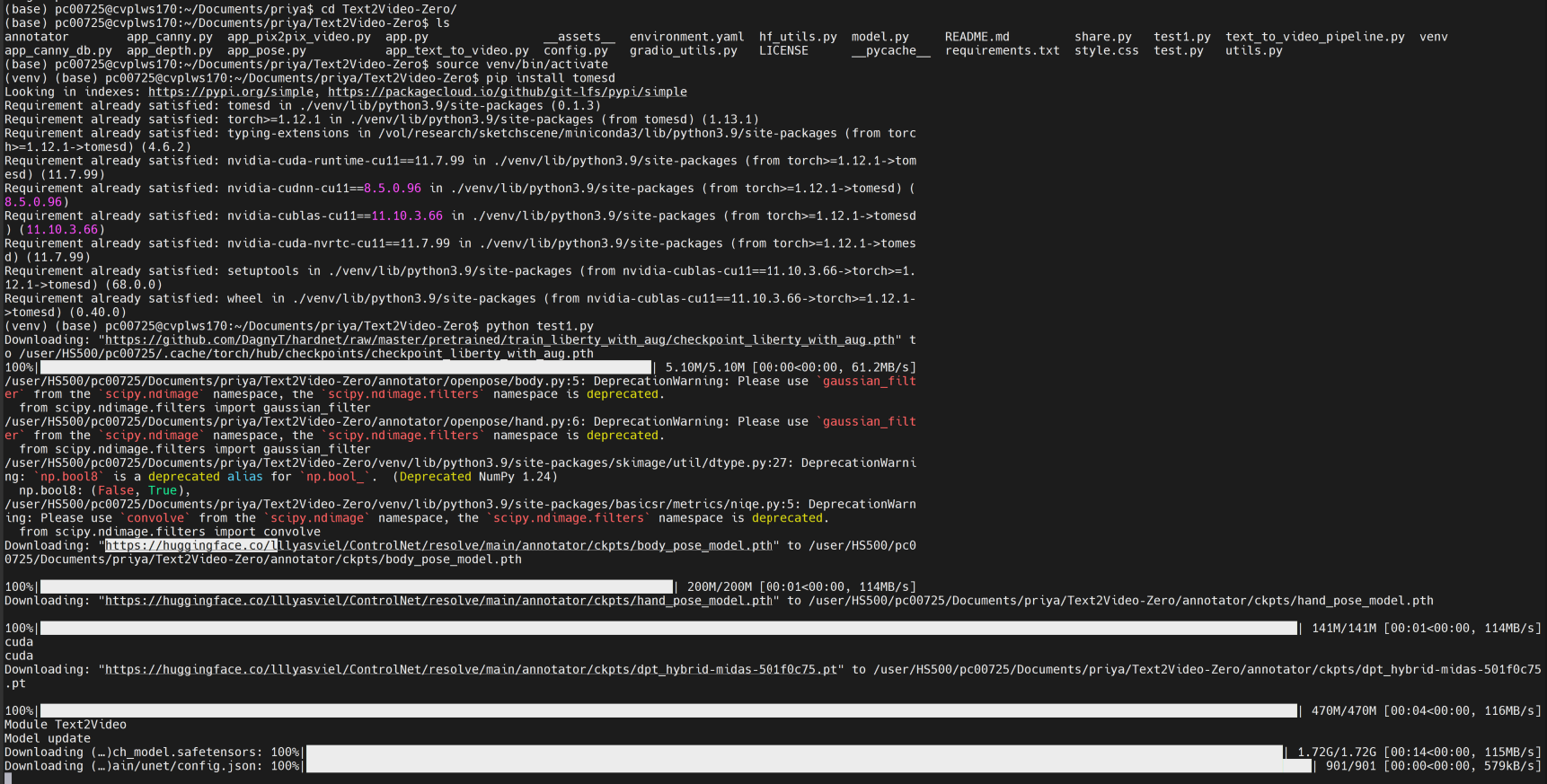
****

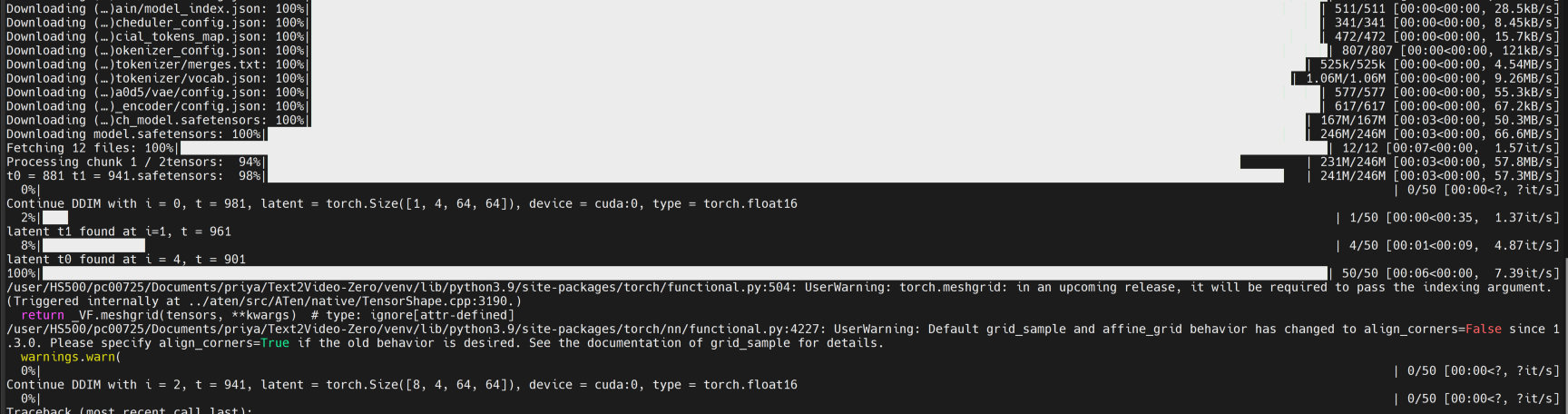
**Text2Video:**

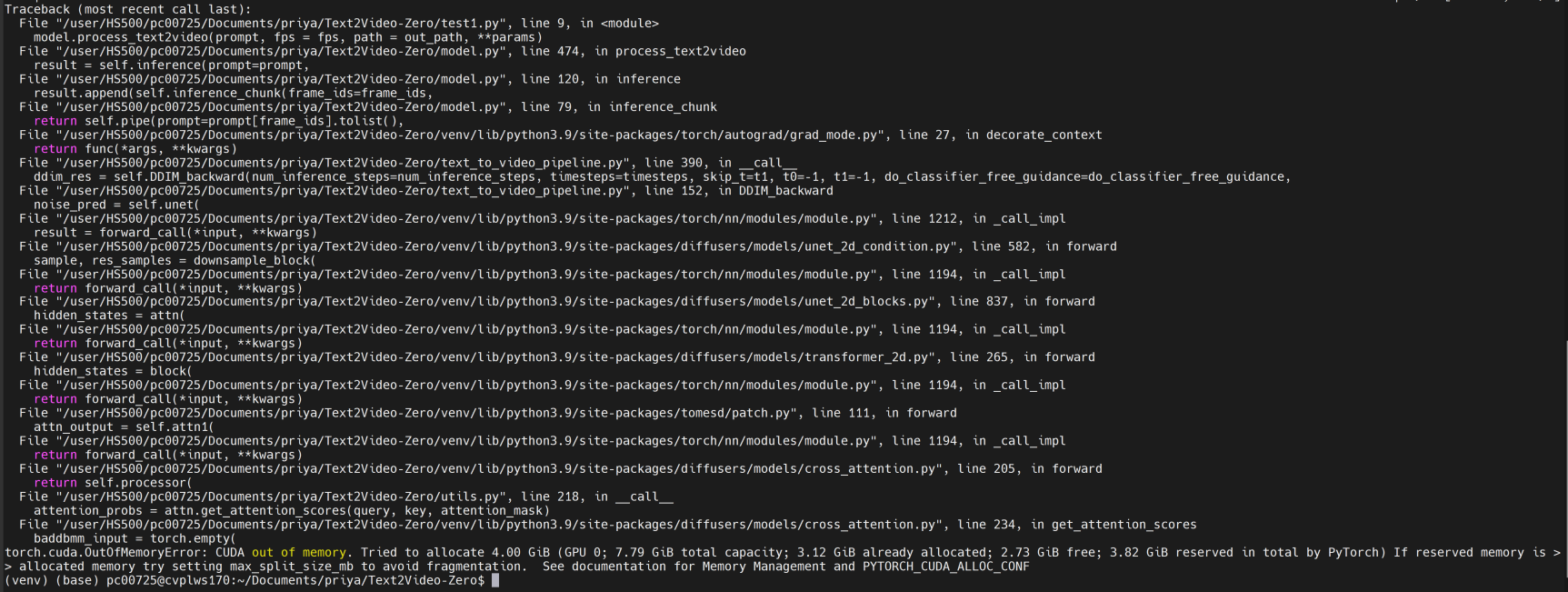
In my local:

****

In remote:

****

****

****

**Issues Encountered:**

1. **Exception has occurred: AssertionError**

**Cuda not available.**

Torch not compiled with CUDA enabled

File "C:\Users\Asus\Documents\Surrey\Audico AI\Codebase\Text2Video-Zero-main\Text2Video-Zero-main\VideoTest.py", line 6, in <module> pipe = TextToVideoZeroPipeline.from\_pretrained(model\_id, torch\_dtype=torch.float16).to("cuda") Assertion Error: Torch not compiled with CUDA enabled