Assessment 9 : CCCS660 Computational Intelligence

Quentin Lao — Matricule : 261233739

HANDS-ON OPTION

The goal of this assessment is to make a catchy video created by Gen AI that highlights the importance of making small everyday choices to reduce one's carbon footprint. This project can be broken down into small tasks. The overall pipeline for our video generation is illustrated in figure 1.

The completed video is accessible via the provided link : https://youtu.be/-vla5R3UHnk

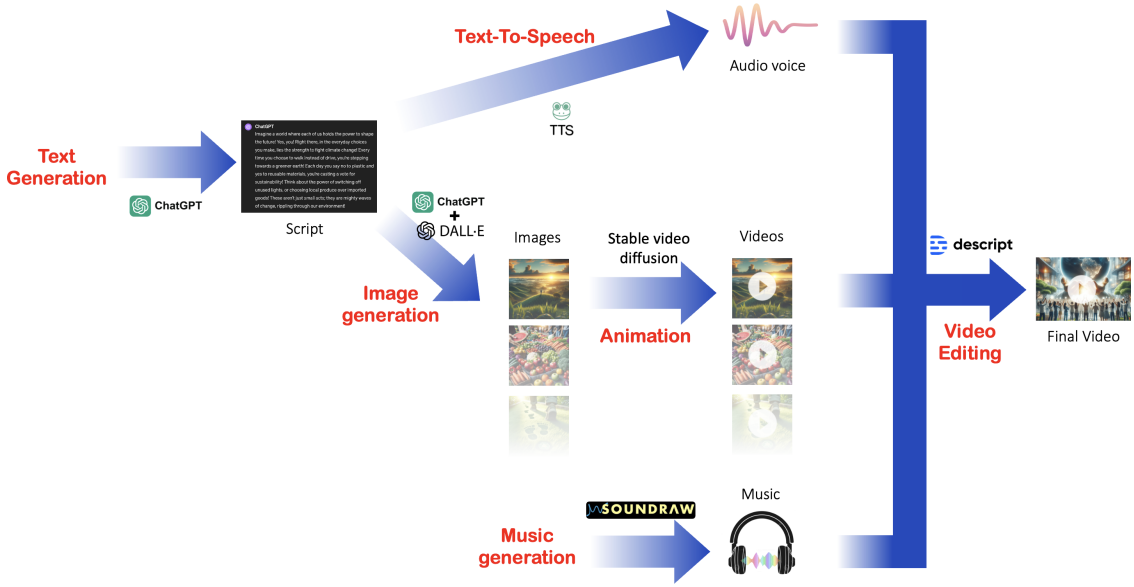

FIGURE 1 – Pipeline for our video generation

1 Text generation

In this phase, we simply used ChatGPT [1] for generating the script. We made some prompt engineering :

— First, we define its identity : "You are a video maker". This is very important to provide the maximum context to get the closer results as possible to our expectations.

— Then, we defined its task : "Please write a 1 mins long monologue for an inspiring video emphasizing the importance of making small everyday choices to reduce one's carbon footprint.". We specified the duration of the monologue here to ensure that ChatGPT produce a concise script. Our earlier attempts used terms like "video", "speech" or "script" instead of "monologue" but ChatGPT usually introduced addition and unnecessary directions such as "[Background music starts]" or speaker introductions like in a play. We find out that using "monologue" gives just the script as we wanted.

1. https://chat.openai.com/

— Lastly, we told it to use exclamations : "Use exclamations to wake people !". Without this additional prompt, our previous text attempts were too flat and lacked the drive to inspire action.

The figure 2 shows our entire prompt and ChatGPT's response that we deemed correct.

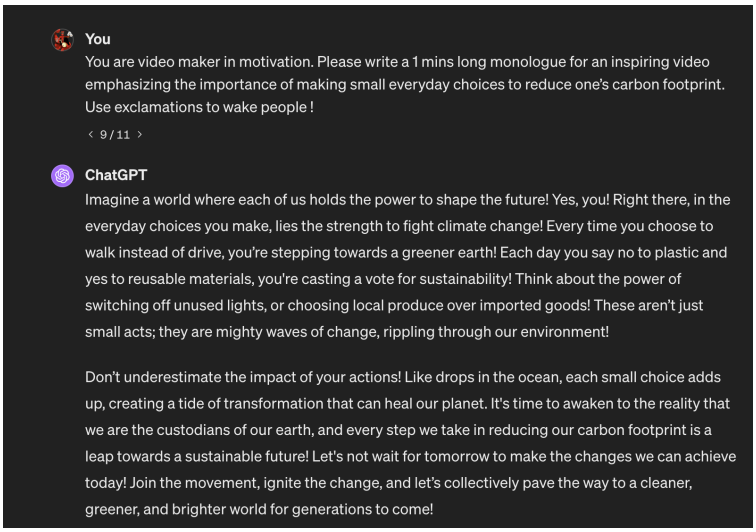

FIGURE 2 – Our prompt and ChatGPT's response for the script generation

2 Image generation

For image creation, we used both Dall-E and ChatGPT from OpenAI. In the basic stage for the hands-on option, the task was to create a single still background image. This could have been done by using ChatGPT that generates a Dall-E prompt that accurately represents the essence of the text it generated. However, as we are aiming for the in-depth stage, we produced instead multiple images corresponding to different sections of the text.

To do so, we asked ChatGPT to place in the text some instructions in brackets, detailing image descriptions that aligned well with the text segments. The figure 3 shows our prompt and ChatGPT's response.

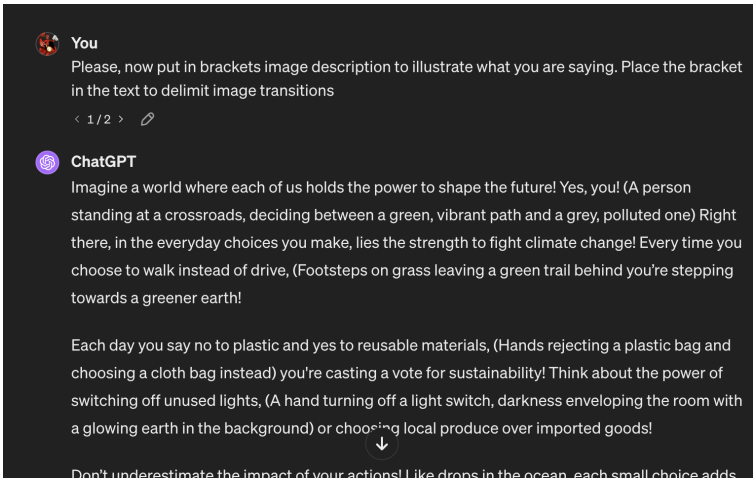

FIGURE 3 – Our prompt and ChatGPT's response for describing potential images

We then used these descriptions as prompt for Dall-E to generate images. The figure 7 illustrates the 10 figures generated by Dall-E for our video.

3 Image animation (in-depth stage)

Displaying a series of static images can look dull and unengaging. To add interest, for each images generated in the previous section, we slightly animated them. To accomplish this, we used a model trained for Stable Video Diffusion (SVD). My code is very similar to the code available on the Hugging Face website [2] [3]. I ran my code on Google Colab for its GPU capabilities.

The short videos produced are approximately 4 seconds long, at 7 FPS. I experimented Topaz [4], an AI-based video enhancement software, particularly for its frame interpolation service to increase the FPS. The software works very well and is actually quite impressive but in its free version, there is a large watermark in the middle of the resulting videos. Consequently, I decided to keep the videos at 7 FPS.

4 Text-To-Speech (in-depth stage)

To make the video more catchy, we wanted the text to be read. There are many text-to-speech (TTS) tools available on the Internet, some of which allow users to train their own model to clone any voice unless having some voice samples. For this assessment, we wanted to make things simple and use pre-trained models. We used Turtoise TTS [5] which proposes in addition an web user interface to generate voice samples from text [6]. We chose the voice of Morgan Freeman as we thought it fit well with our video's theme : calm, inspiring, not too fast with a motivational tone. The figure 4 shows the web user interface of Turtoise TTS.

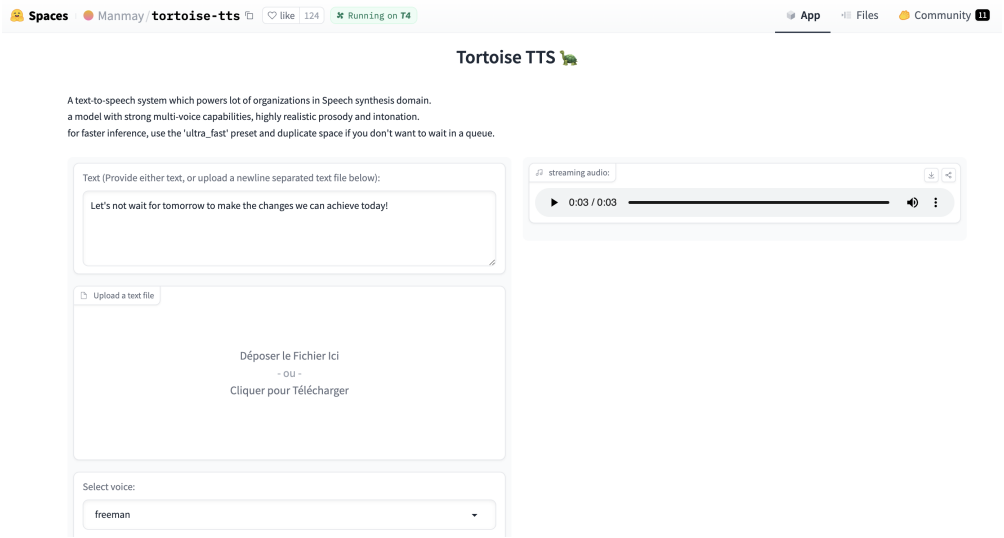

FIGURE 4 – The web user interface of Turtoise TTS

2. My code : https://colab.research.google.com/drive/1COmZJtob3qWuRb74apTZfLThOw_iPM8o?usp=sharing
3. The hugging face website : https://huggingface.co/docs/diffusers/main/en/using-diffusers/svd
4. https://www.topazlabs.com/
5. https://github.com/neonbjb/tortoise-tts
6. https://huggingface.co/spaces/Manmay/tortoise-tts

5 Music generation

For music generation, we used Soundraw, an online tool to generate musics with AI. To produce a music on Soundraw, one must specify the genre, the mood, the theme, the duration and the tempo. Our aim was to create a very motivating video with hope and emotions, we then set the parameters as follows :

— Genre : Orchestra
— Mood : Epic, Sentimental, Hopeful
— Theme : Nature
— Length : 1 minutes 30 seconds
— Tempo : Fast

Once the music is generated, the website allows for small tweeks like slight modification of a specific part of the music. The figure 5 displays the Soundraw interface with the music we chose.

FIGURE 5 – Soundraw interface with the music we chose

6 Video editings

Once we have every piece of the project, we gathered all of them to create the final video. We used the free online video editor Descript [7]. A significant advantage of Descript is its powerful audio transcription tool which allowed us to make automatically subtitles from the TTS audio.

We then manually assembled the pieces together : music, short videos from SVD and voice audio from TTS. Actually, we didn't find any free automation tools to this. However, we are convinced that this final process can be easily automated using Python with more time, as it involves merely assembling various elements without the need for complex decision-making. Figure 6 displays the interface of Descript.

7. https://www.descript.com/

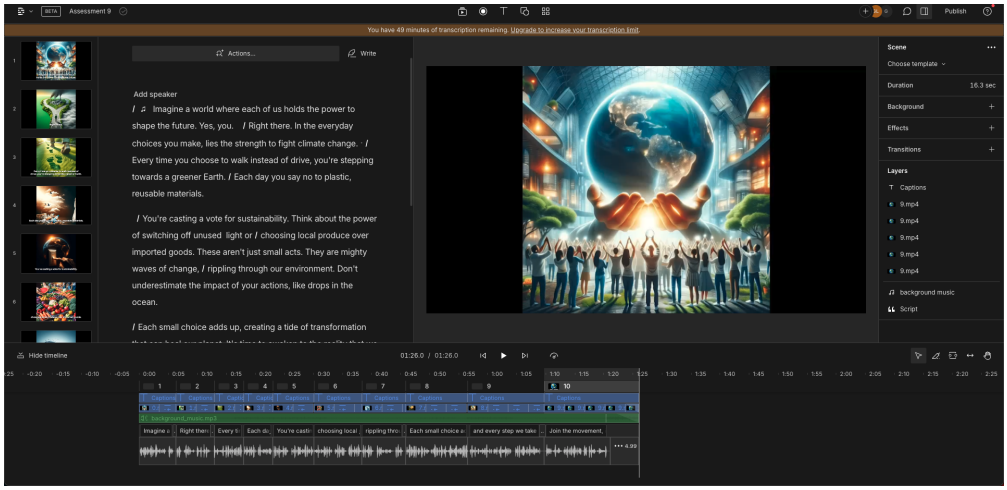

FIGURE 6 – Interface of Descript

(a) *A diverse group of people holding a glowing globe together, symbolizing unity and empowerment*

(b) *A person standing at a crossroads, deciding between a green, vibrant path and a grey, polluted one*

(c) *Footsteps on grass leaving a green trail behind*

(d) *Hands rejecting a plastic bag and choosing a cloth bag instead*

(e) *A hand turning off a light switch, darkness enveloping the room with a glowing earth in the background*

(f) *A person selecting fruits and vegetables at a local farmer's market*

(g) *Close-up of water droplets joining to form a big wave*

(h) *A person standing on a hilltop, overlooking a flourishing landscape at sunrise*

(i) *Image of diverse people planting trees and cleaning up nature*

(j) *Image of a bright, healthy earth as seen from space, reflecting a sustainable future*

FIGURE 7 – All images generated by Dall-E with their respective prompt