# **Task : Replicate Sora (Create Videos of humans)**

## **Approach : Text to Video Generation**

This project aims to create realistic videos of humans based on descriptive textual prompts. We have explored various models and tools for this purpose, including open-source models from Hugging Face and paid tools like Runway Gen-2, ImagineArt, and Haiper. We evaluated their performance using two specific prompts.

### **Descriptive Prompts:**

**Prompt 1:** A stylish woman walks down a Tokyo street filled with warm glowing neon and animated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective, creating a mirror effect of the colorful lights. Many pedestrians walk about.

**Prompt 2:** A young man in his 20s is sitting on a piece of cloud in the sky, reading a book.

### 

### **Text to Video Models:**

#### **Model-1:** [AnimateDiff-Lightning](https://huggingface.co/ByteDance/AnimateDiff-Lightning)

**Videos:**

* [Video for Prompt 1](https://drive.google.com/file/d/1X9X7AgW-Np2k2vn-g3mfhmX_PQZu0_Xr/view?usp=drive_link)
* [Video for Prompt 2](https://drive.google.com/file/d/1isnuGoQg2K9yuehQu2wUuTMb59O-Yeyx/view?usp=sharing)

**Insights:**

* Rapid generation of videos.
* Clear images in generated GIFs.

**Limitations:**

* Accuracy in Prompt Interpretation:
  + First Prompt: The overall ambiance is captured, but specific details (e.g., attire) do not match perfectly. The reflective street surface is well-executed, but finer details lack precision.
  + Second Prompt: Minor details in posture, expression, or surrounding elements are not accurately rendered. The main concept is clear, but interactions between the character and the cloud or background elements are not refined.
* Struggles with generating longer videos.

#### **Model-2:** [AnimateLCM](https://huggingface.co/wangfuyun/AnimateLCM)

**Videos:**

* [Video for Prompt-1](https://drive.google.com/file/d/1FnzAncvvKueAJvuKN4gbv03XYa3o8Tys/view?usp=sharing)
* [Video for Prompt-2](https://drive.google.com/file/d/1s4ill3FNcXv1Qb1zi1ntFaT0ByBSKMZL/view?usp=sharing)
* [10 second Trial Video for Prompt-1](https://drive.google.com/file/d/1WKPJ6MeUFce9FJSNWBGnFC1IA7whGHON/view?usp=sharing)

**Insights:**

* High level of detail in the final output, closely matching the prompts.
* Clear and visually appealing images.

**Limitations:**

* GIF Length and Consistency:
  + Struggles to generate all frames in a single batch, requiring smaller batches.
  + Maintaining consistency across batches is challenging, leading to noticeable differences in appearance or quality, affecting overall animation smoothness.
* Slight deviations from the expected visual representation of specific prompt elements.

#### **Model-3:** [Text-to-video-Zero](https://huggingface.co/docs/diffusers/en/api/pipelines/text_to_video_zero)

**Videos:**

* [Video for Prompt-1](https://drive.google.com/file/d/1dtTwpGSPySp91Bbi3FTpf5LmKsI52zOl/view?usp=sharing)
* [Video for Prompt-2](https://drive.google.com/file/d/1F8xE-ZXL42UyQ90FrYws0KJAD7k6S-Zr/view?usp=sharing)

**Insights:**

* Able to create short-length videos (2 seconds).
* The generated videos capture the general themes of the prompts.

**Limitations:**

* The model misses finer details from the prompts. For example, the exact appearance of the woman's black boots, black purse, and red lipstick is not precisely depicted.
* The model struggles with complex scenes involving multiple dynamic elements.
* The smoothness and natural flow of movement in the videos is limited. The walking motion in the first video or the gentle floating of the cloud in the second video appear less fluid compared to real-world footage.

#### 

#### **Model-4:** [Text-to-video-ms-1.7b](https://huggingface.co/ali-vilab/text-to-video-ms-1.7b)

**Videos:**

* [Video for Prompt-1](https://drive.google.com/file/d/1-EdWvqObg-0_tnXXfWRykgAPDDqfC_Wp/view?usp=sharing)
* [Video for Prompt-2](https://drive.google.com/file/d/1ptz5jyoEUJVM4PsMG7G39dvIviJV_hWM/view?usp=sharing)

#### **Insights:**

* The model performs well in creating appropriate settings and atmospheres that match the prompts.
* Basic movements are captured effectively.

#### **Limitations:**

* The model cannot achieve perfect film and television quality generation. The animations lack smoothness, making some actions appear stiff or unnatural.
* There is an inability to generate finer details of characters and objects.

#### 

#### 

#### **Model-5:** [Animatediff-motion-adapter](https://huggingface.co/Warvito/animatediff-motion-adapter-sdxl-v1-0-beta)

**Videos:**

* [Video for Prompt-1](https://drive.google.com/file/d/17PhOSfF4sOMlQj8Yyxy2KkjYH03YlKXB/view?usp=sharing)
* [Video for Prompt-2](https://drive.google.com/file/d/1hx5RRSmL42a6bG5GxCaLHuA5fHzYCU4K/view?usp=sharing)

**Insights** :

* It maintains a relatively consistent output in terms of style and animation across different prompts.
* The model handles lighting effects and reflective surfaces well as seen in the GIF that a damp street with colorful light reflections is produced.

**Limitation** :

* The model struggles with understanding and accurately rendering long or complex prompts.
* There is inaccuracy in depicting specific character features.
  + The specifics of the woman's attire, such as the red dress is not accurate or detailed.
  + The woman's facial details, like the red lipstick, lacks precision.
  + The specifics of the man's appearance such as his age is not rendered with accuracy.
* It is not able to identify all the objects mentioned in the prompt clearly. It also lacks to build the interaction between the objects like the man is sitting on a chair instead of a piece of cloud.

### 

### **Text to Video Tools:**

#### **Tool-1:** [Runway-Gen2](https://runwayml.com/research/gen-2)

**Videos:**

* [Video for Prompt-1](https://drive.google.com/file/d/1JHEyX4HoWCWZQF1z_OuVP0l-gKWvrfzk/view?usp=sharing)
* [Video for Prompt-2](https://drive.google.com/file/d/12ayuEG-7MCzTKlJlqMHzJtUlRrG0dGRi/view?usp=sharing)

**Insights** :

* The tool generates short videos of 4 seconds.
* The tool makes an attempt to capture detailed backgrounds and multiple elements.

**Limitation** :

* There are noticeable deformities in human figures, particularly in their limbs and facial features.
* The tool struggles to keep the main subject in focus in videos with moving background elements, such as in the "woman walking" video where pedestrians and other distractions cause the primary subject to lose focus.

#### **Tool-2:** [ImagineArt](https://www.imagine.art/dashboard/videos)

**Videos:**

* [Video for Prompt-1](https://drive.google.com/file/d/1Ew4sX32EAJVvftUG9yfG3w92us-0xVPF/view?usp=sharing)
* [Video for Prompt-2](https://drive.google.com/file/d/1iXcoc8XTRqu8cbETH8d4Qp3RoSNnHfvg/view?usp=sharing)

**Insights** :

* The tool generates 2-second videos.
* The clothing and accessories of the characters match the descriptions provided in the prompts.

**Limitation** :

* The videos feature faces that are blurry and lack clear details, resulting in an abstract and defaced appearance.
* The overall visual quality is poor, making it hard to recognize facial features.

#### 

#### **Tool-3:** [Haiper](https://haiper.ai/)

**Videos:**

* [Video for Prompt-1](https://drive.google.com/file/d/1CDVzwSYM-jtgYbrjM0UBrc3KJrcnE6qd/view?usp=sharing)
* [Video for Prompt-2](https://drive.google.com/file/d/1OAcQ8ZBFA2qIyTOIVLtdrS5QxkQZfw_-/view?usp=sharing)

**Insights** :

* 4 second videos are created.
* The videos effectively capture the overall atmosphere and themes described in the prompts.

**Limitation** :

* Some details, particularly in facial features and background elements, lack sharpness and detail.
* The video has an almost animated quality to it.