# AI -State of Art

(2023)

BY MANISH AGARWAL

# Problem Statement

- How to generate detailed images conditioned on text descriptions using AI.
- User type-: Good in Imagination Skills but may lack in Creation skillsgenerate detailed images conditioned on text descriptions

Major Models

- Midjourney
- Dall-e 2
- Stable Diffusion

# DALL·E 2 vs Midjourney vs Stable Diffusion

- **Pros of DALLE-E**
- High-Quality images of close up and clothing-designs
- Prevents you from using it for bad intentions
- Generates multiple images
- **Cons of DALLE-E**
- Copyright of AI-images not clear
- Can create false results
- If it doesn't understand a text, it may generate from previous training

DALL·E 2 .

- **Pros of Midjourney**
- High-Quality results
- Other people's work is viewable
- Reasonable prices
- **Cons of Midjourney**
- Only available via Discord
- Privacy costs more money
- Not easy to use

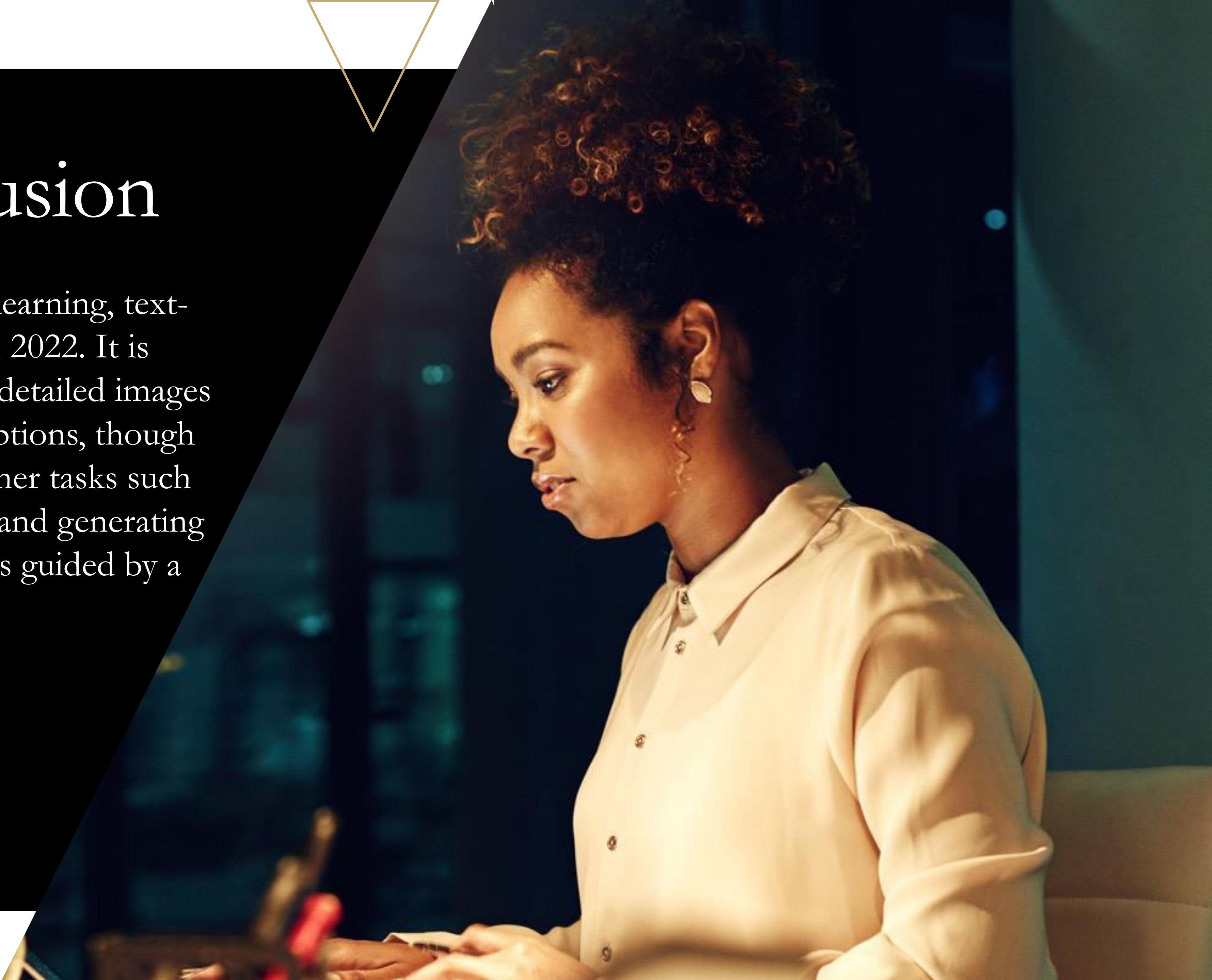
Midjourney

- **Pros of Stable Diffusion**
- High-quality images
- Unlimited possibilities
- Exploring of different art styles
- **Cons of Stable Diffusion:**
- No copyright projections worldwide as of now
- AI-art has possible risks
- AI technology is uncertain

Stable Diffusion

# Stable-Diffusion

- Stable Diffusion is a deep learning, text-to-image model released in 2022. It is primarily used to generate detailed images conditioned on text descriptions, though it can also be applied to other tasks such as inpainting, outpainting, and generating image-to-image translations guided by a text prompt.



# Why Stable Diffusion(SD)?

Both Dall-e 2 and midjourney are not open source and they will remain as that in the future while SD is open source.

Both are based on subscription model while SD not

SD Have Large Community

# Industrial Use case

- Anime
- Manga
- Digital Animation
- Drawings
- General Animation
- Traditional Animation
- Pixel Art
- Avatar Creation
- NFT Creation
- Plugins (adobe , canvas etc)
- Paintings ....many more

# Types of  Models in Stable diffusion

- SD-1.5
- SD-2.1
- Openjourney(midjourney Style)

….many more we can tune the model according to our need

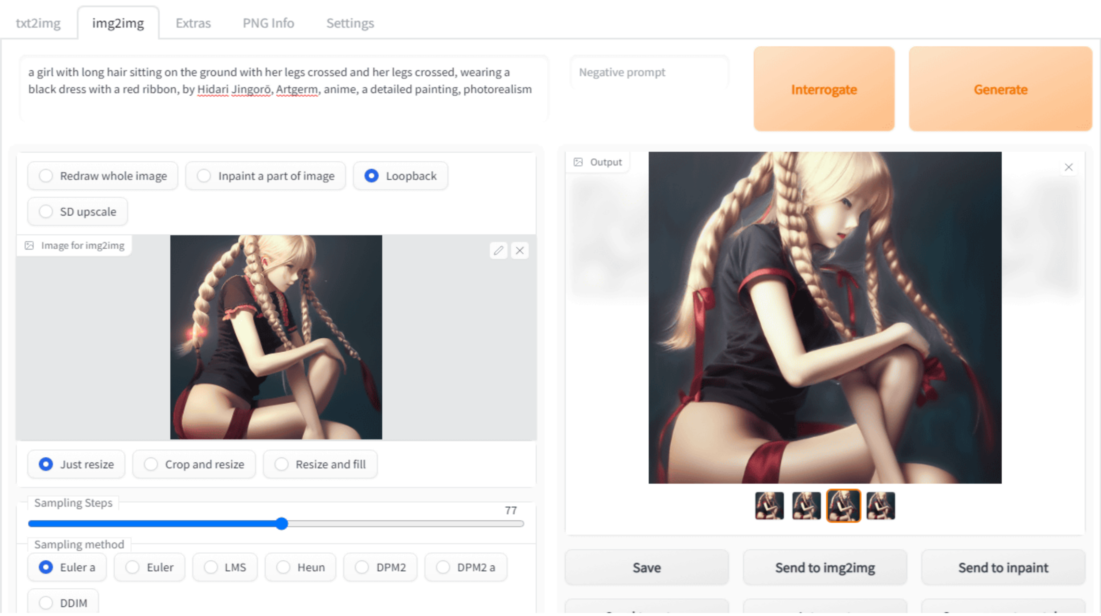
# Way to access SD

Hugging Face
Api/spaces

Using transformer
/diffusions

Stable diffusion

Web-ui



a girl with long hair sitting on the ground with her legs crossed and her legs crossed, wearing a black dress with a red ribbon, by Hidari Jingorō, Artgerm, anime, a detailed painting, photorealism

○ Redraw whole image ○ Inpaint a part of image ● Loopback
○ SD upscale

● Just resize ○ Crop and resize ○ Resize and fill

Sampling Steps 77

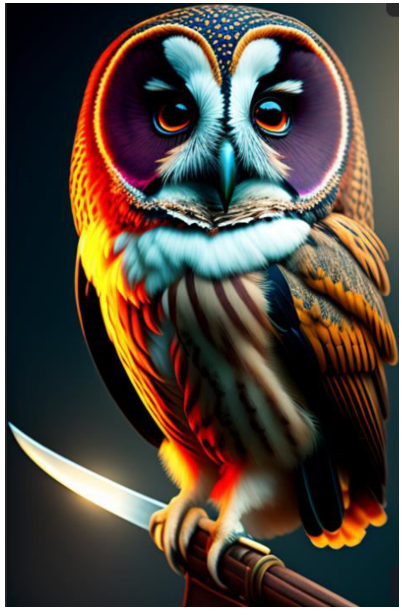
Sampling method
● Euler a ○ Euler ○ LMS ○ Heun ○ DPM2 ○ DPM2 a
○ DDIM

# Web-ui

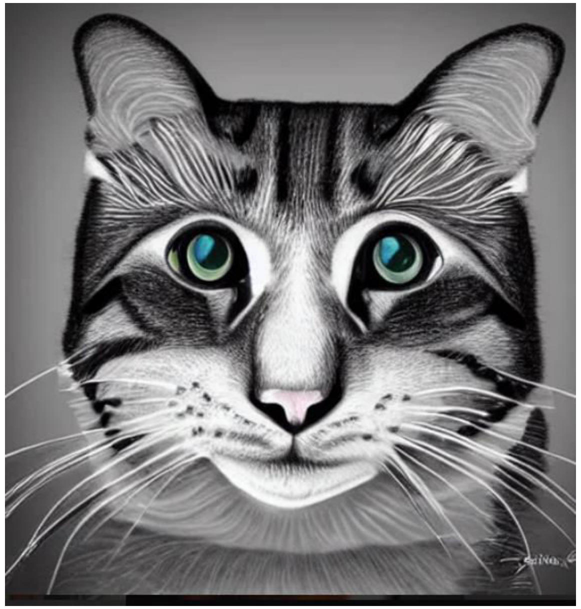
- Automatic1111(Web Based)
- InvokeAI (Terminal Based)

# SD-Results

PROMPT-: AN BLOODY AND MURDERER OWL HOLDING A WEAPON, ULTRAREALISTIC



PROMPT-: ANTHROPOMORPHIC CAT PORTRAIT ART



# Resources

- Discord – an independent one that seems to be more helpful than the official one

- News feed of industry updates

- /r/StableDiffusion – noisy, with a lot of "look what I made"

- Huge compilation of tools & techniques

# Competitors

- Stability AI.
- Rosebud AI.
- Wombo.
- Fotor.
- Playground.
- Nyx.gallery.
- Midjourney.
- Craiyon.
- Lexcia.art

# Key Features

- Original txt2img and img2img modes
- Outpainting
- Inpainting
- Color Sketch
- Text2video
- Upscaling

# Thank you

# TEXT TO VIDEO GENERATION

By Manish Agarwal

# AGENDA

Introduction

Types Of Models

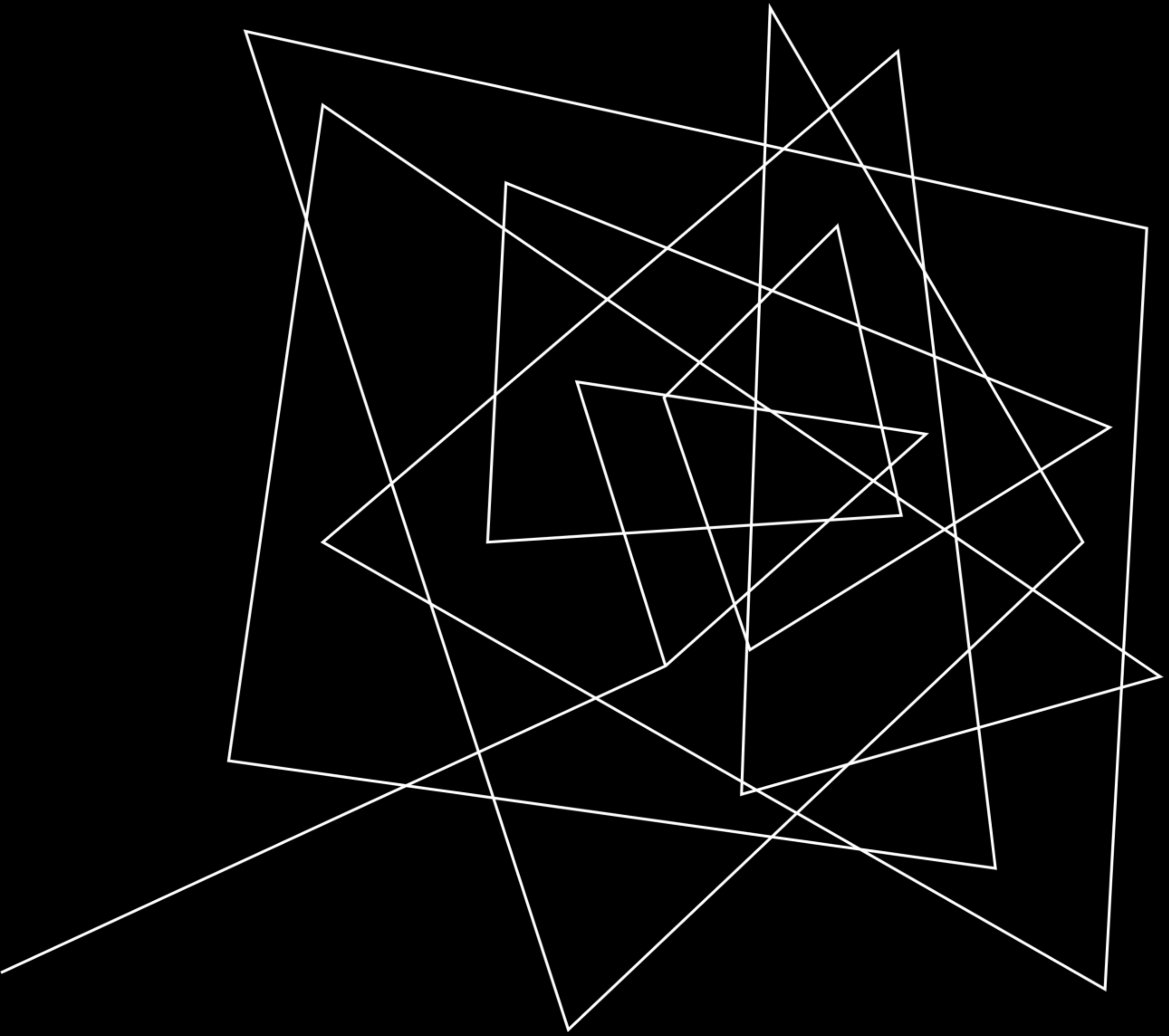
DataSet Overview

Cog-video outputs

Summary

# INTRODUCTION

Text-to-Video is **a state-of-the-art artificial intelligence technology which needs only text as input for the output as video**. The inspiration came from text-to-image models which deliver images as output from text as input.



# TYPES OF MODEL

1. Make A video(Meta)
2. Imagen Video(by google)
3. Phenaki(google)
4. Cogvideo

# DESCRIPTION

## Make-A-Video

Research builds on the recent progress made in text-to-image generation technology built to enable text-to-video generation. The system uses images with descriptions to learn what the world looks like and how it is often described. It also uses unlabelled videos to learn how the world moves. With this data, Make-A-Video lets you bring your imagination to life by generating whimsical, one-of-a-kind videos with just a few words or lines of text.

Research paper
Website

# DESCRIPTION

## Imagen-Video

Imagen Video builds on Google’s Imagen, an image-generating system comparable to OpenAI’s DALL-E 2 and Stable Diffusion. Imagen is what’s known as a “diffusion” model, generating new data (e.g. videos) by learning how to “destroy” and “recover” many existing samples of data. As it’s fed the existing samples, the model gets better at recovering the data it’d previously destroyed to create new works.

Research paper
website

# DESCRIPTION

## Phenaki

An AI model called Phenaki can generate minutes of coherent video based on detailed, sequential text input. While Imagen Video focuses on quality, Phenaki prioritizes coherency and length. The system can turn paragraph-long prompts into films of an arbitrary length

Research paper
website

# DESCRIPTION

## Cog-Video

Production of text-to-video, transformation of 9 billion parameters, and inheritance of a trained text-to-image model One of the first large-scale pretrained open-source text-to-video models, CogView2 has a multi-frame-rate hierarchical training technique and outperforms all currently available models.

[Research paper]
[Website]

# DATASET USED BY DIFFERENT MODELS

### Make A Video

- Laion-5b
- WebVid-10M
- HD-VILA-100M

### Imagen-Video

- Laion-400
- Private dataset 14 million video-text pairs and 60 million image-text pairs

### Phenaki

- Moment in time (MIT) Dataset
- Complete Dataset (TBA)

### Cog Video

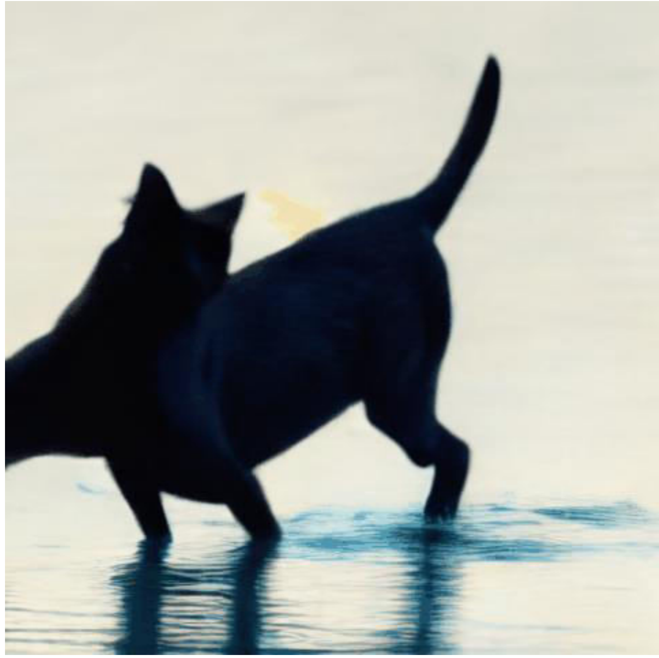
- inheritance of a trained text-to-image model

# COG-VIDEO

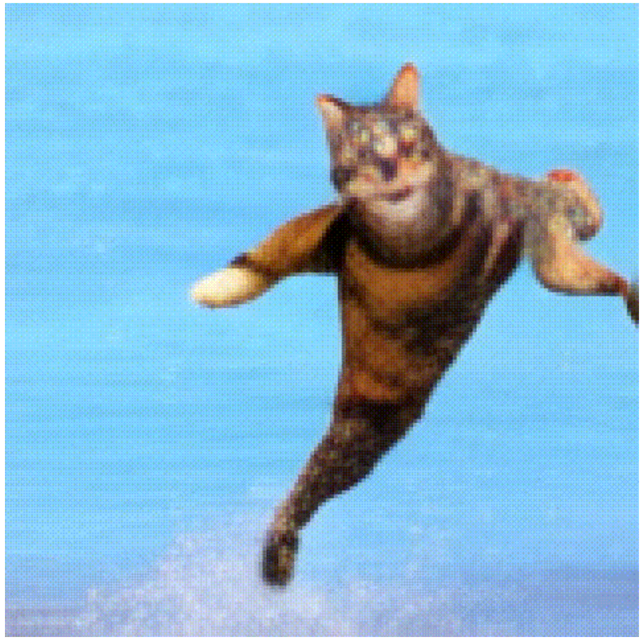
This the only available pretrained model as an open source and available on both hugging-face and replicate

# Cog Video Result

Prompt -: A cat Running
On water

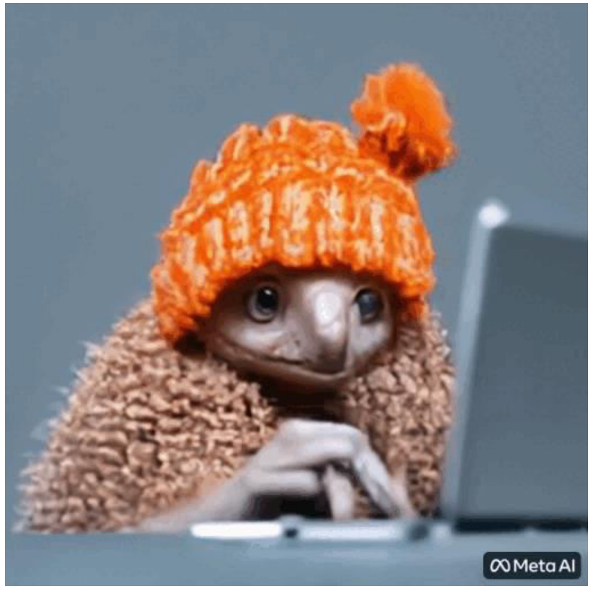


Replica Platform

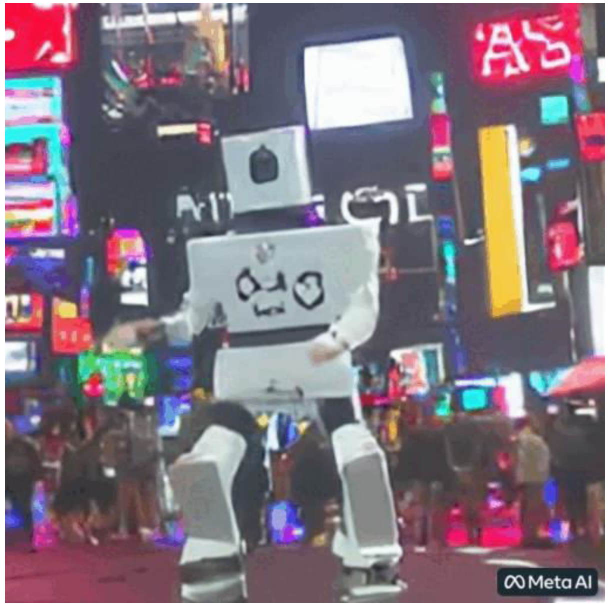


Hugging Face platform

# Make a Video Result

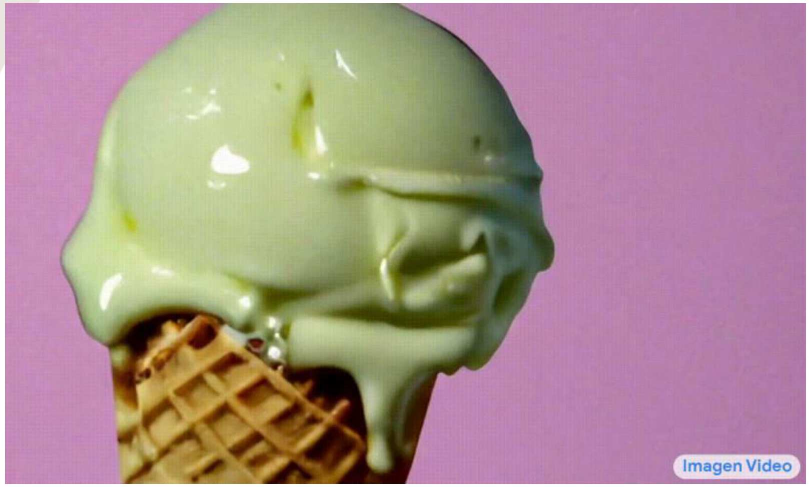


Prompt-:A fluffy baby sloth with an orange knitted hat trying to figure out a laptop close up highly detailed studio lighting screen reflecting in its eye

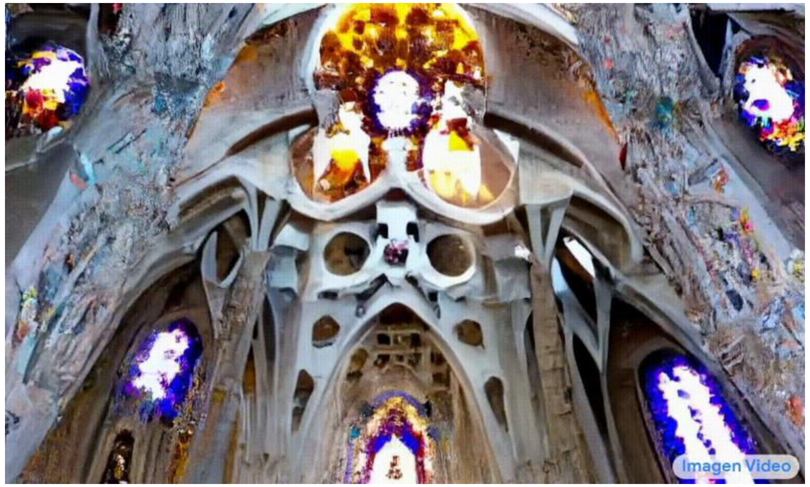


Prompt-: Robot dancing in times square

# Imagen Video Result



Prompt -: Melting Ice-Cream Dripping down the cone



Prompt -:drone fly through interior of sagrada familiar cathedral

# Phenaki Video Result

Lots of traffic in futuristic city. An alien spaceship arrives to the futuristic city. The camera gets inside the alien spaceship. The camera moves forward until showing an astronaut in the blue room. The astronaut is typing in the keyboard. The camera moves away from the astronaut. The astronaut leaves the keyboard and walks to the left. The astronaut leaves the keyboard and walks away. The camera moves beyond the astronaut and looks at the screen. The screen behind the astronaut displays fish swimming in the sea. Crash zoom into the blue fish. We follow the blue fish as it swims in the dark ocean. The camera points up to the sky through the water. The ocean and the coastline of a futuristic city. Crash zoom towards a futuristic skyscraper. The camera zooms into one of the many windows. We are in an office room with empty desks. A lion runs on top of the office desks. The camera zooms into the lion's face, inside the office. Zoom out to the lion wearing a dark suit in an office room. The lion wearing looks at the camera and smiles. The camera zooms out slowly to the skyscraper exterior. Timelapse of sunset in the modern city

# RESULT OBSERVATION

| Make A video | Imagen Video | Phenaki | Cog-Video |
|---|---|---|---|
| •high-definition videos<br>•Sort Clip around 5 sec<br>•Medium coherent | high-definition videos<br>Sort Clip around 5 sec<br>High coherent | Low – Quality Video<br>Long Video around 2 min<br>High coherent | Low-Quality<br>Bad performance<br>Low coherent |

# PROBLEMS -:

1. Apart from Cog Video none other model has released their code and pretrained model
2. Imagen-video have ml architecture code  but without trained model
3. Some Dataset is available but its not clean or removes NFWS Content
4. If we Want To duplicate the Structural code than we want deep knowledge related to transformers and mathematical terms and large amount of time to train the model
5. Cog-video is not provided satisfied result

# THANK YOU

Manish Agarwal

# VIDEO TO TEXT GENERATION

By Manish Agarwal

# AGENDA

How to get  transcripts from audio or video

Result

Observation

Usecase

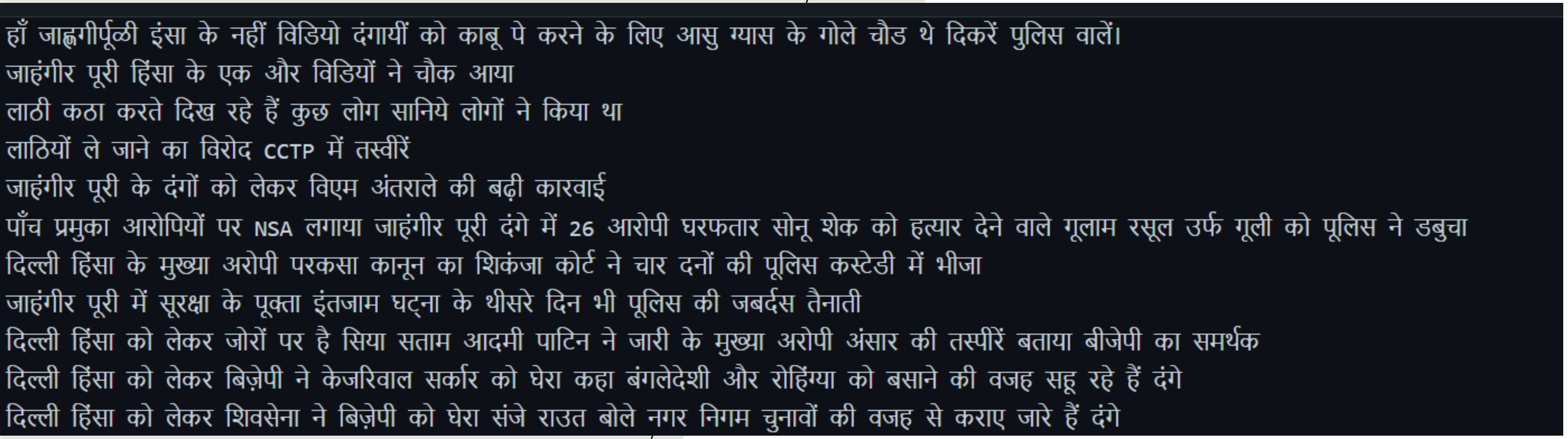
# INTRODUCTION

Whisper is a general-purpose speech recognition model. It is trained on a large dataset of diverse audio and is also a multi-task model that can perform multilingual speech recognition as well as speech translation and language identification..

# WHISPER ON HINDI NEWS VIDEO

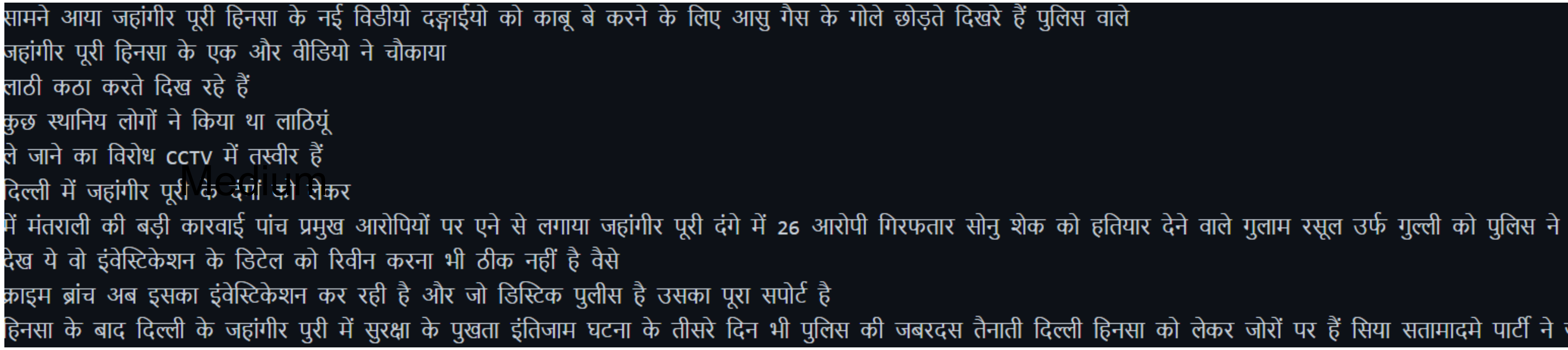
# Whisper Ai Result

Case 1 using medium model  and large model for transcribe the video in same language



```
हाँ जाह्रगीर्पूळी इंसा के नहीं विडियो दंगायीं को काबू पे करने के लिए आसु ग्यास के गोले चौड थे दिकरें पुलिस वालें।
जाहंगीर पूरी हिंसा के एक और विडियों ने चौक आया
लाठी कठा करते दिख रहे हैं कुछ लोग सानिये लोगों ने किया था
लाठियों ले जाने का विरोद CCTP में तस्वीरें
जाहंगीर पूरी के दंगों को लेकर विएम अंतराले की बढ़ी कारवाई
पाँच प्रमुका आरोपियों पर NSA लगाया जाहंगीर पूरी दंगे में 26 आरोपी घरफतार सोनू शेक को हत्यार देने वाले गूलाम रसूल उर्फ गूली को पूलिस ने डबुचा
दिल्ली हिंसा के मुख्या अरोपी परकसा कानून का शिकंजा कोर्ट ने चार दनों की पूलिस कस्टेडी में भीजा
जाहंगीर पूरी में सूरक्षा के पूक्ता इंतजाम घट्ना के थीसरे दिन भी पूलिस की जबर्दस तैनाती
दिल्ली हिंसा को लेकर जोरों पर है सिया सताम आदमी पाटिन ने जारी के मुख्या अरोपी अंसार की तस्पीरें बताया बीजेपी का समर्थक
दिल्ली हिंसा को लेकर बिज़ेपी ने केजरिवाल सर्कार को घेरा कहा बंगलेदेशी और रोहिंग्या को बसाने की वजह सहू रहे हैं दंगे
दिल्ली हिंसा को लेकर शिवसेना ने बिज़ेपी को घेरा संजे राउत बोले नगर निगम चुनावों की वजह से कराए जारे हैं दंगे
```

Medium



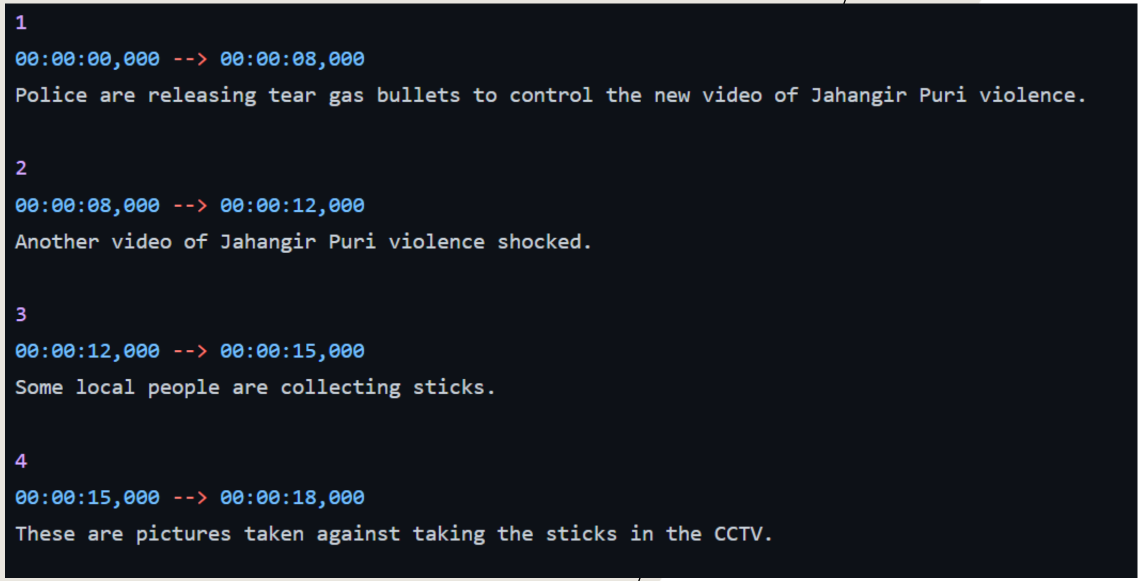
```
सामने आया जहांगीर पूरी हिनसा के नई विडीयो दङ्गाईयो को काबू बे करने के लिए आसु गैस के गोले छोड़ते दिखरे हैं पुलिस वाले
जहांगीर पूरी हिनसा के एक और वीडियो ने चौकाया
लाठी कठा करते दिख रहे हैं
कुछ स्थानिय लोगों ने किया था लाठियूं
ले जाने का विरोध CCTV में तस्वीर हैं
दिल्ली में जहांगीर पूरी के दंगों को लेकर
में मंतराली की बड़ी कारवाई पांच प्रमुख आरोपियों पर एने से लगाया जहांगीर पूरी दंगे में 26 आरोपी गिरफतार सोनु शेक को हतियार देने वाले गुलाम रसूल उर्फ गुल्ली को पुलिस ने
देख ये वो इंवेस्टिकेशन के डिटेल को रिवीन करना भी ठीक नहीं है वैसे
क्राइम ब्रांच अब इसका इंवेस्टिकेशन कर रही है और जो डिस्टिक पुलीस है उसका पूरा सपोर्ट है
हिनसा के बाद दिल्ली के जहांगीर पुरी में सुरक्षा के पुखता इंतिजाम घटना के तीसरे दिन भी पुलिस की जबरदस तैनाती दिल्ली हिनसा को लेकर जोरों पर हैं सिया सतामादमे पार्टी ने
```

Medium

Large

# Whisper Ai Result

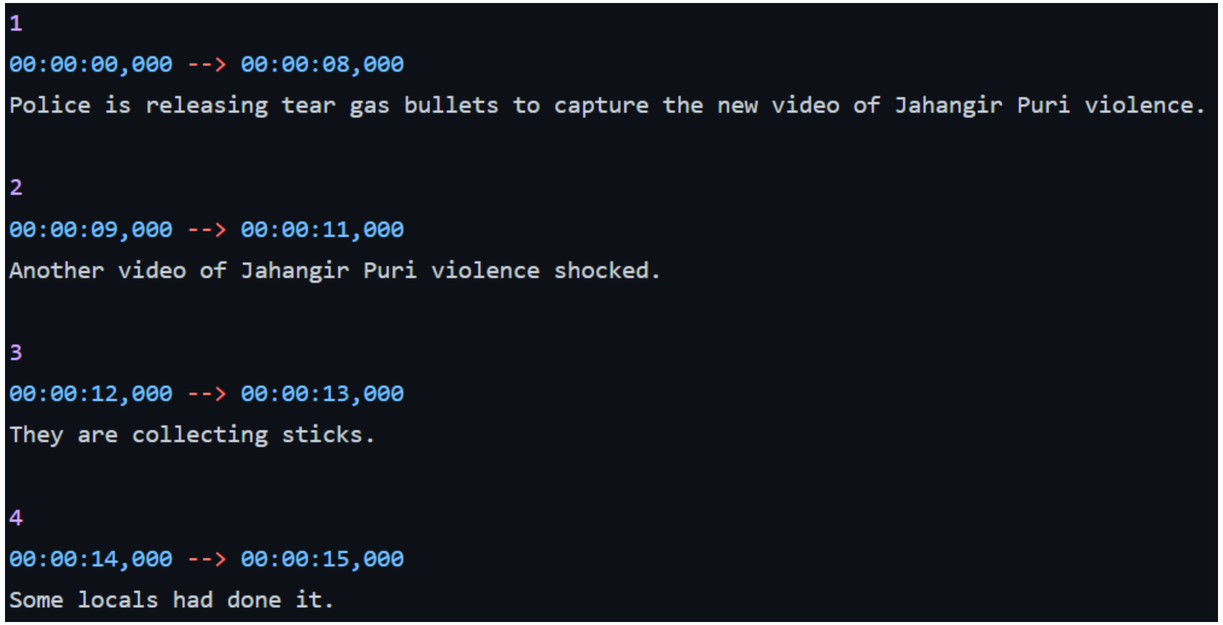
Case 2 using medium model  and large model for Translate the video in English language



```
1
00:00:00,000 --> 00:00:08,000
Police are releasing tear gas bullets to control the new video of Jahangir Puri violence.

2
00:00:08,000 --> 00:00:12,000
Another video of Jahangir Puri violence shocked.

3
00:00:12,000 --> 00:00:15,000
Some local people are collecting sticks.

4
00:00:15,000 --> 00:00:18,000
These are pictures taken against taking the sticks in the CCTV.
```

Large

Medium



```
1
00:00:00,000 --> 00:00:08,000
Police is releasing tear gas bullets to capture the new video of Jahangir Puri violence.

2
00:00:09,000 --> 00:00:11,000
Another video of Jahangir Puri violence shocked.

3
00:00:12,000 --> 00:00:13,000
They are collecting sticks.

4
00:00:14,000 --> 00:00:15,000
Some locals had done it.
```

# RESULT OBSERVATION

| Medium-Hindi | Medium English | Large Hindi | Large English |
|---|---|---|---|
| •Its work well for 75% of timeline while for 25% its repeated some words continuously | •Its work well for 50% of timeline while for 50% its repeated some words continuously | •Its work best in all the cases without any repetition and similar or same the audio script | Its work good compared to medium model by some last line get repeated |

# USE CASE -:

1. Youtuber can use this to provide subtitles in many different languages  with a small tweaking

2. It help for Hearing Aids patients to get audio in text form in real time
3. Use this to get the summary of speech in video
4. May use in digital notes lecture

# THANK YOU

Manish Agarwal