SORA : A Review

Members:
Sneha KK (MDS202240)
Sarvesh Bhandary (MDS202253)

Sora and its capabilities

- A text to image and text to video model by OpenAI released in Feb 2024.
- Generates realistic videos/images with text prompt alone. Allows images/videos as well.
- Videos upto 1920x1080 and 1080x1920 and Images upto 2048x2048.
- It follows scaling laws of LLMs and has emergent properties.
- First model that can generate minute long videos.

Sora using image and video prompts

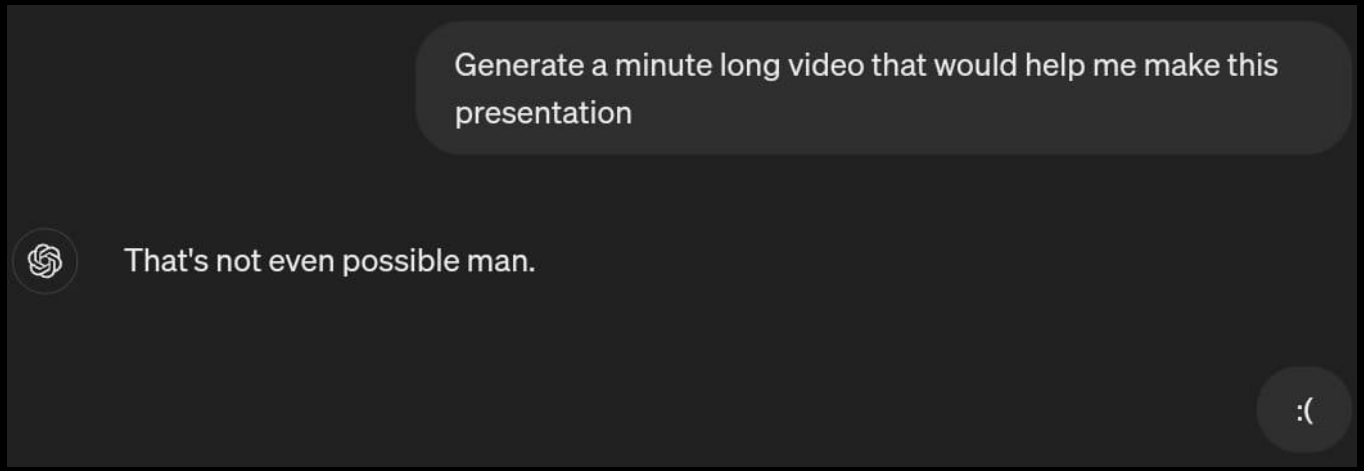
- Uses image/video prompts as context for video generation.
- Video editing - Connecting videos, extending videos forwards/backwards etc.
- Static image animation - Uses image and text prompt to create video.

+

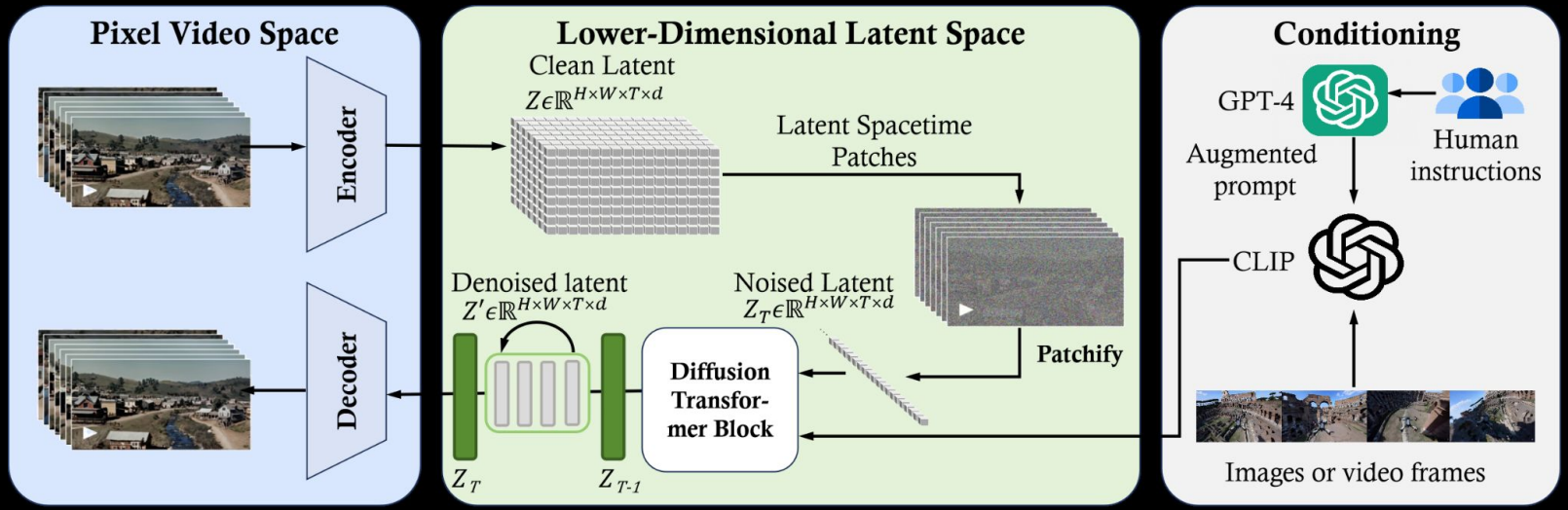
A Shiba Inu dog wearing a beret and black turtleneck

History of generative models

- Image generation models like GANs, VAEs, DALLE etc.
- Video generation models like Imagen Video and VideoLDM.
- Text generation models like ChatGPT, Gemini etc.



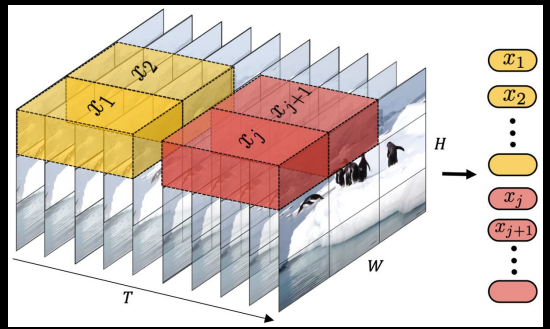
Overview of Sora framework



Data Pre-processing

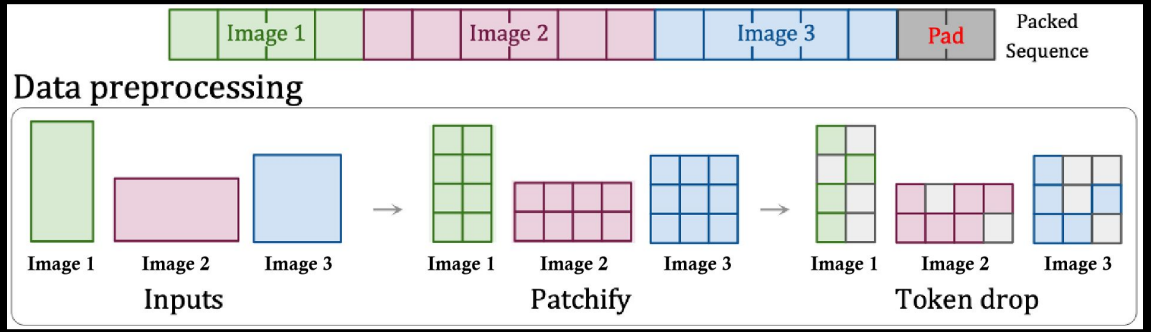
- Training on data in their native sizes: improves composition and framing.

- Video compression using VAE or VQ-VAE.
- Decompose into unified spacetime latent patches.
- Spacetime - patches :



Spacetime Latent Patches

- Variability in number of patches.
- Patch and pack :
  - Arrange patches in fixed length sequences.
  - Padding and token drop.
- Large context window : can have multiple videos in a sequence.
- Noise addition.

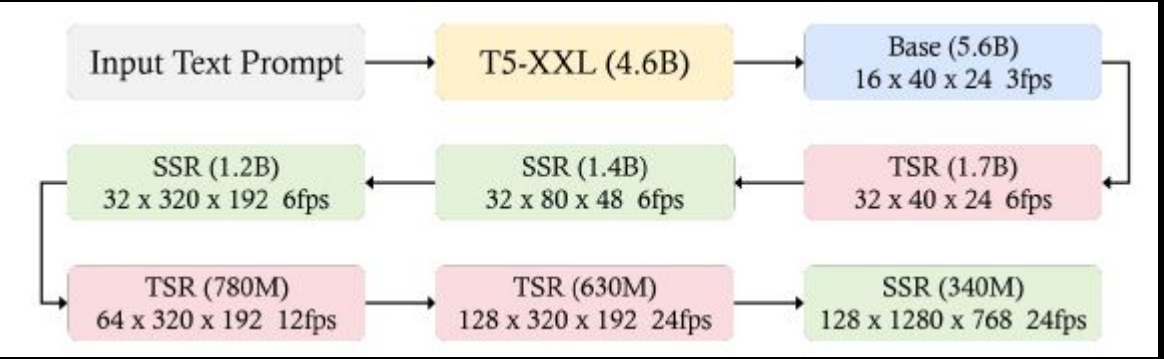


Diffusion Transformers

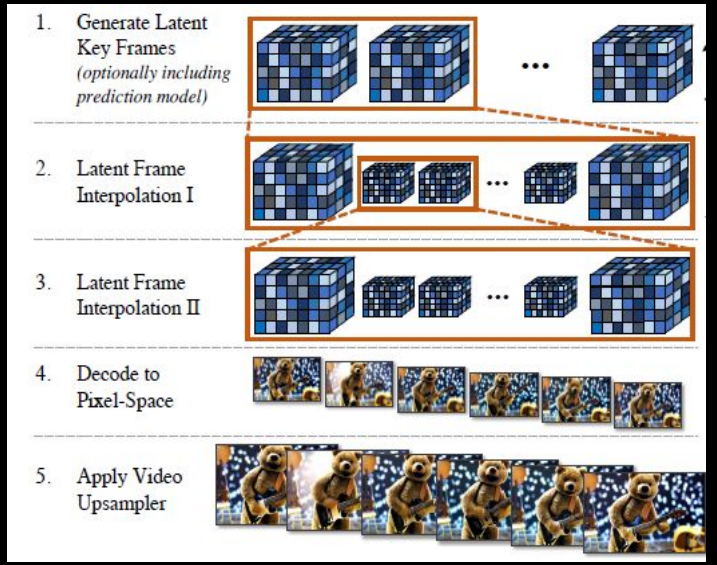
- Success of transformers over U-Nets in visual domain : ViT, U-ViT, MDT.
- Transformers in place of U-Nets for diffusion.
- Diffusion transformers for increasing spatial and temporal resolution.
  - Iteratively go from low resolution image to high resolution image.
  - Interpolate for intermediate frames to increase frame rate.
- Imagen Video and Video LDM.

Imagen Video

- Contextual embeddings from input prompts.
- Base model (3D U-Net) .
- Cascade of Spatial Super Resolution and Temporal Super Resolution diffusion models.



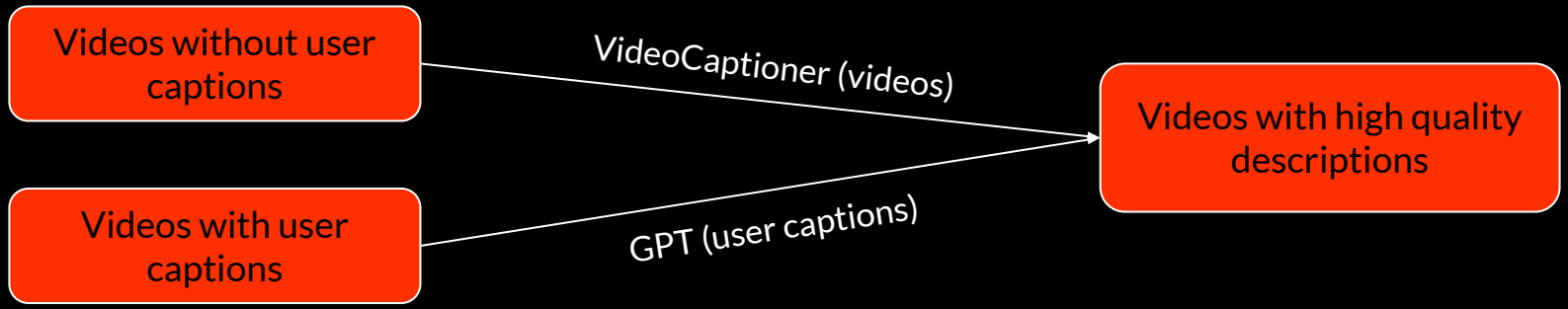
- Video Latent Diffusion Model:
  - Cascade of diffusion models.
  - Additional temporal layers that learn to align individual frames.
  - Fine-tuned for temporal consistency.
- Sora likely uses a cascade of diffusion models:
  - Base model and spacetime refiner models.
- Temporal consistency is more important: it likely trains with longer videos of low resolution .



Video LDM Architecture

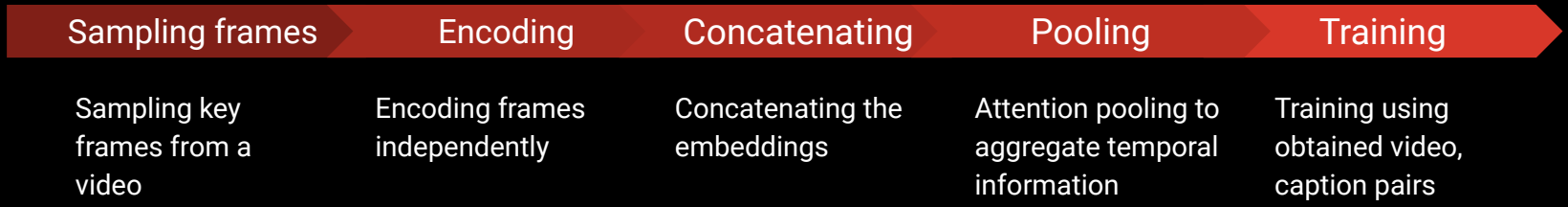
Training data for the text-to-video model

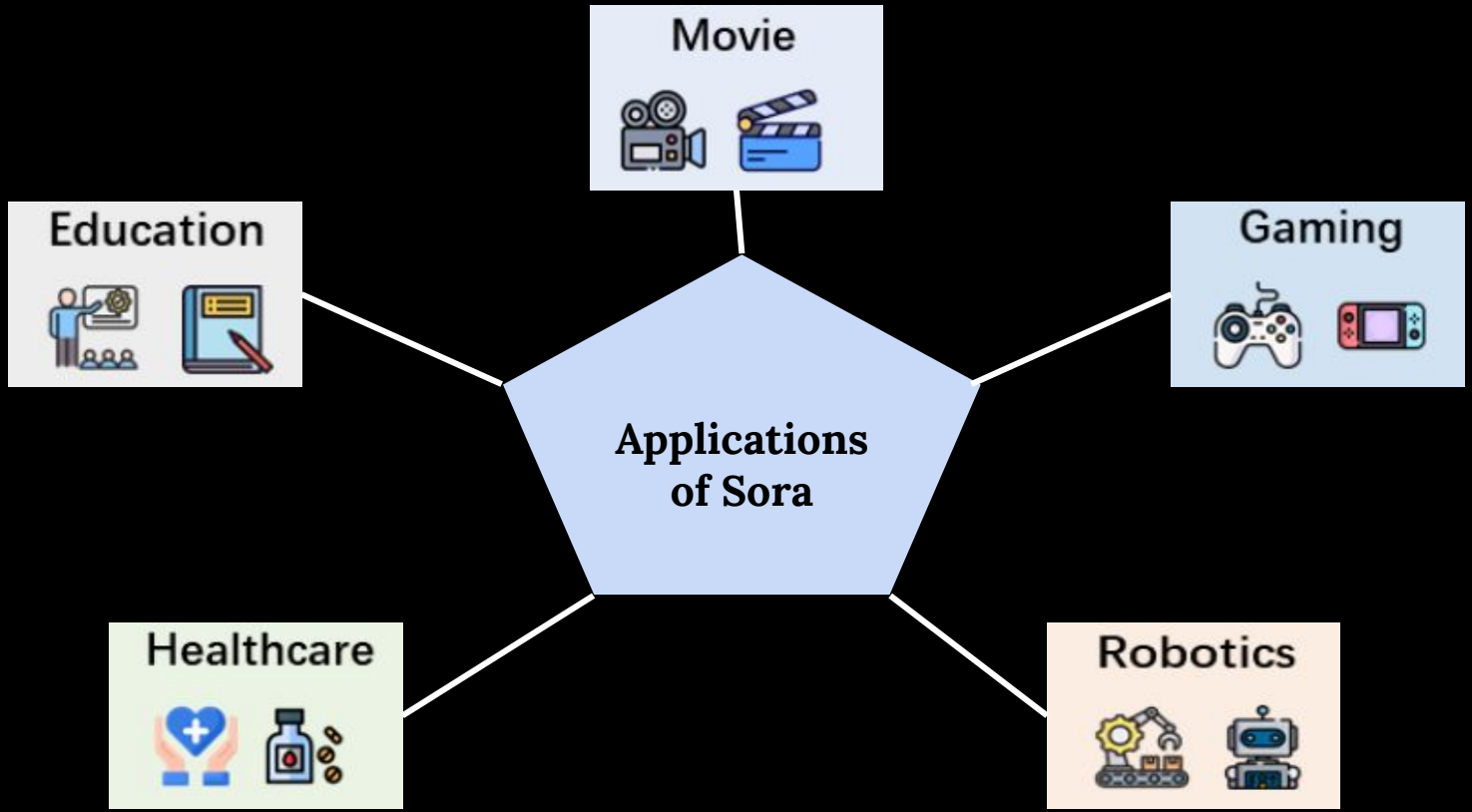
- First train a highly descriptive video captioner.
- Apply re-captioning on all videos like DALLE-3.
- Also utilizes GPT-enhanced user prompts.
- Gives high quality video caption pairs



Video captioner training using VideoCoCa

- Sample multiple frames from a video.
- Image encoder gives frame token embeddings.
- Flatten and concatenate into a long sequence of video representation.
- Processing token information using attention pooler.
- Uses contrastive and captioning loss for training.




Movie
Education
Gaming
Applications of Sora
Healthcare
Robotics

Security and Protection

- Misuse and jailbreak attacks
- Hallucination
- Bias
- Privacy

- RLHF
- Usage permissions
- Privacy protection

Limitations and Challenges

- Physical realism
- Spatial and temporal complexities
- Human-computer interaction
- Length of videos

Thank you