
SORA

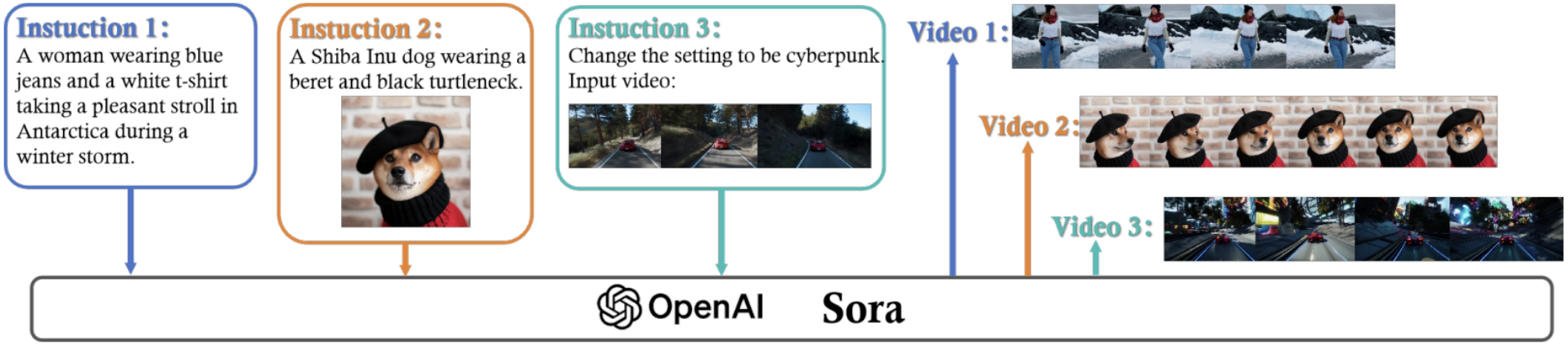

SORA

Contents

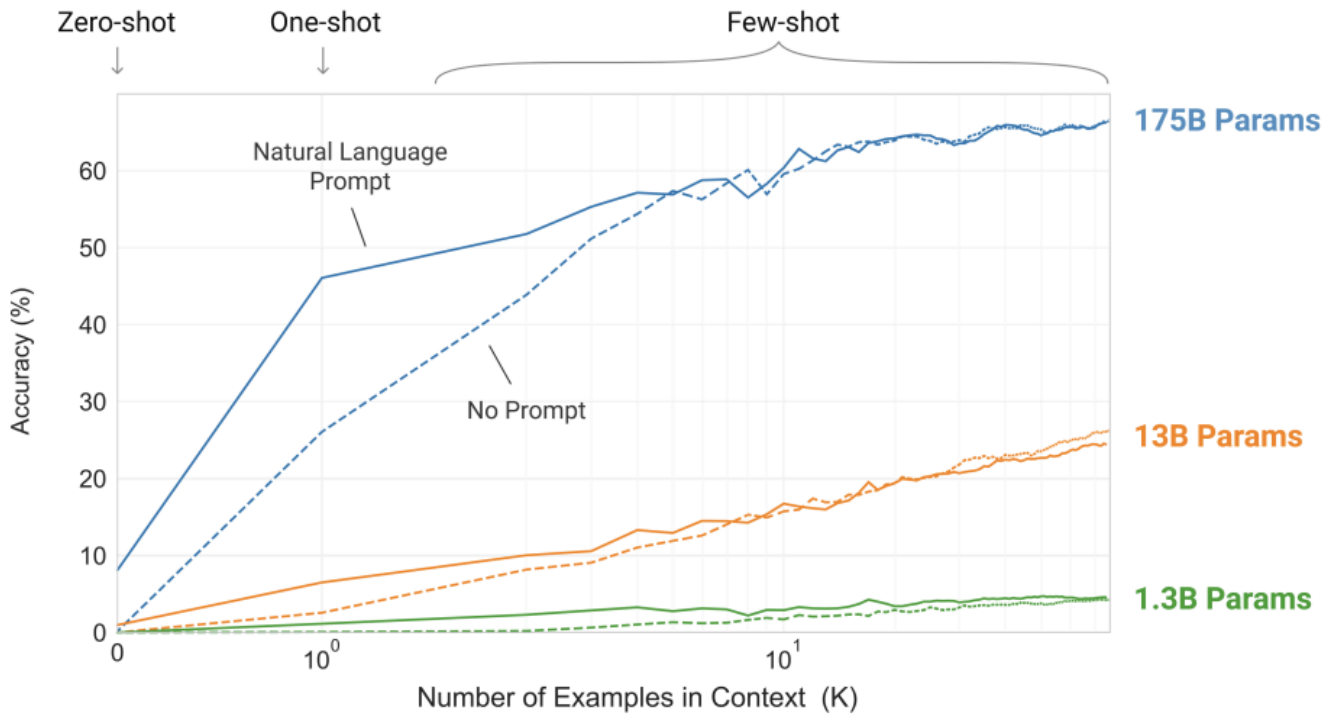
- Introduction
- Applications
- Data
- Architecture
- Challenges/limitations

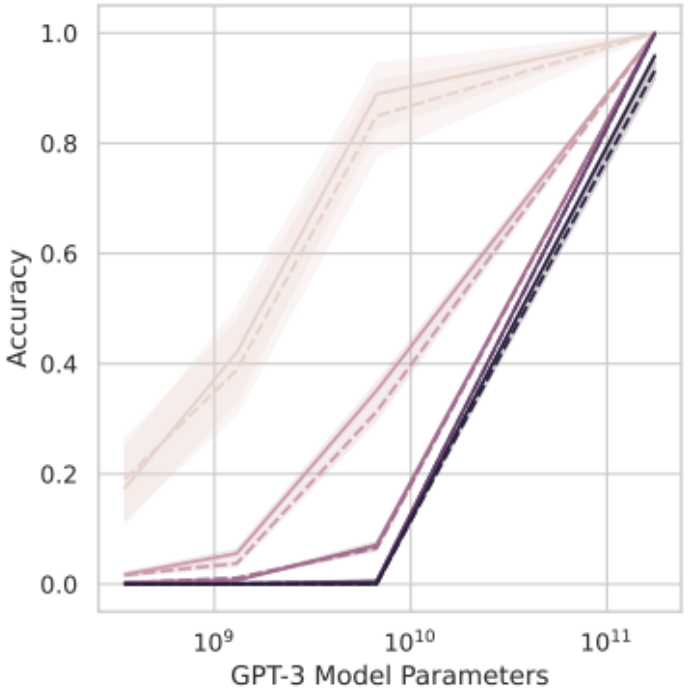
What is SORA?

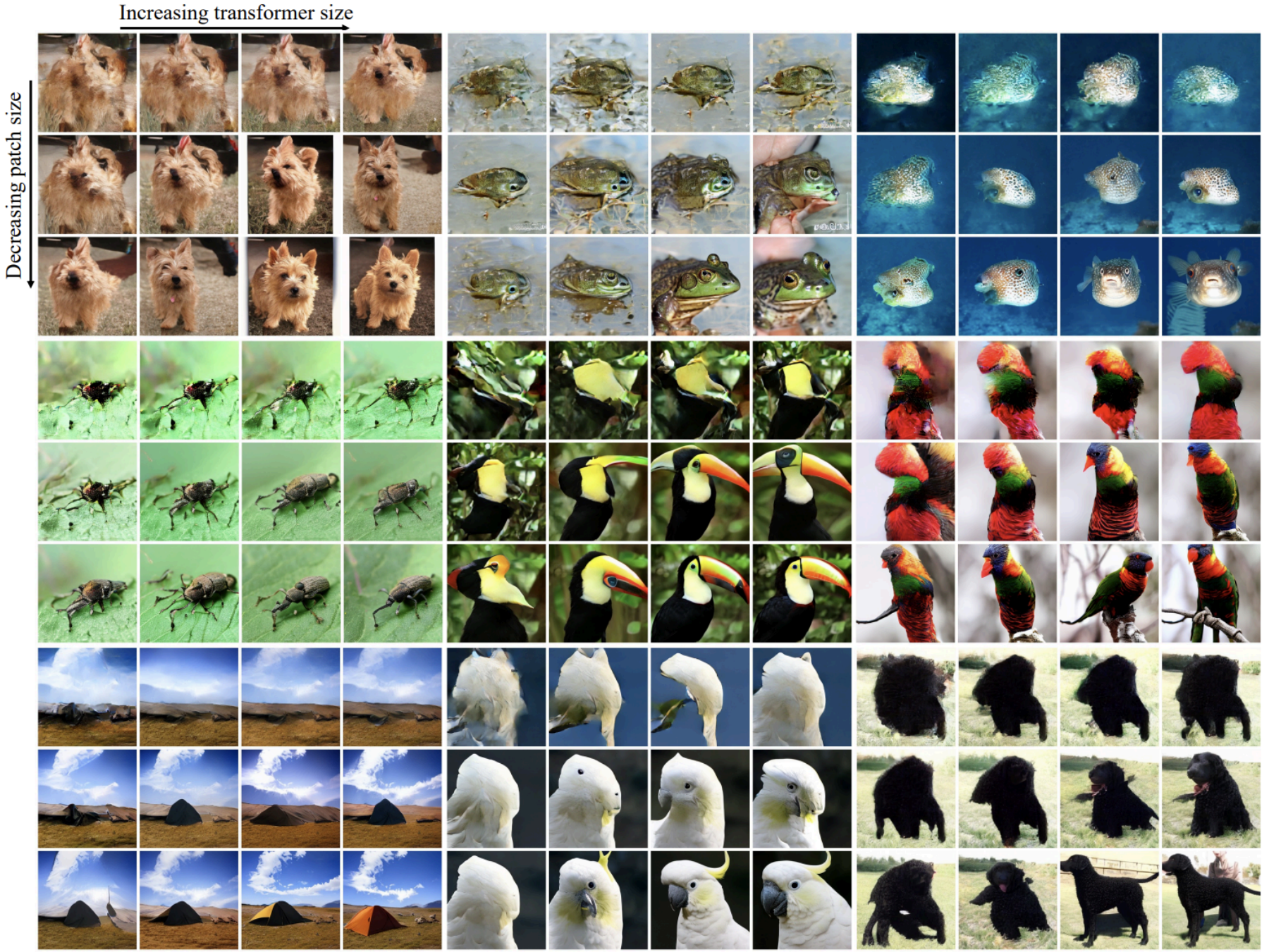
- SORA means sky in Japanese.
- A text-to-video **multimodal** generative diffusion transformer model by OpenAI (February, 2024)
- **Not** an official technical report



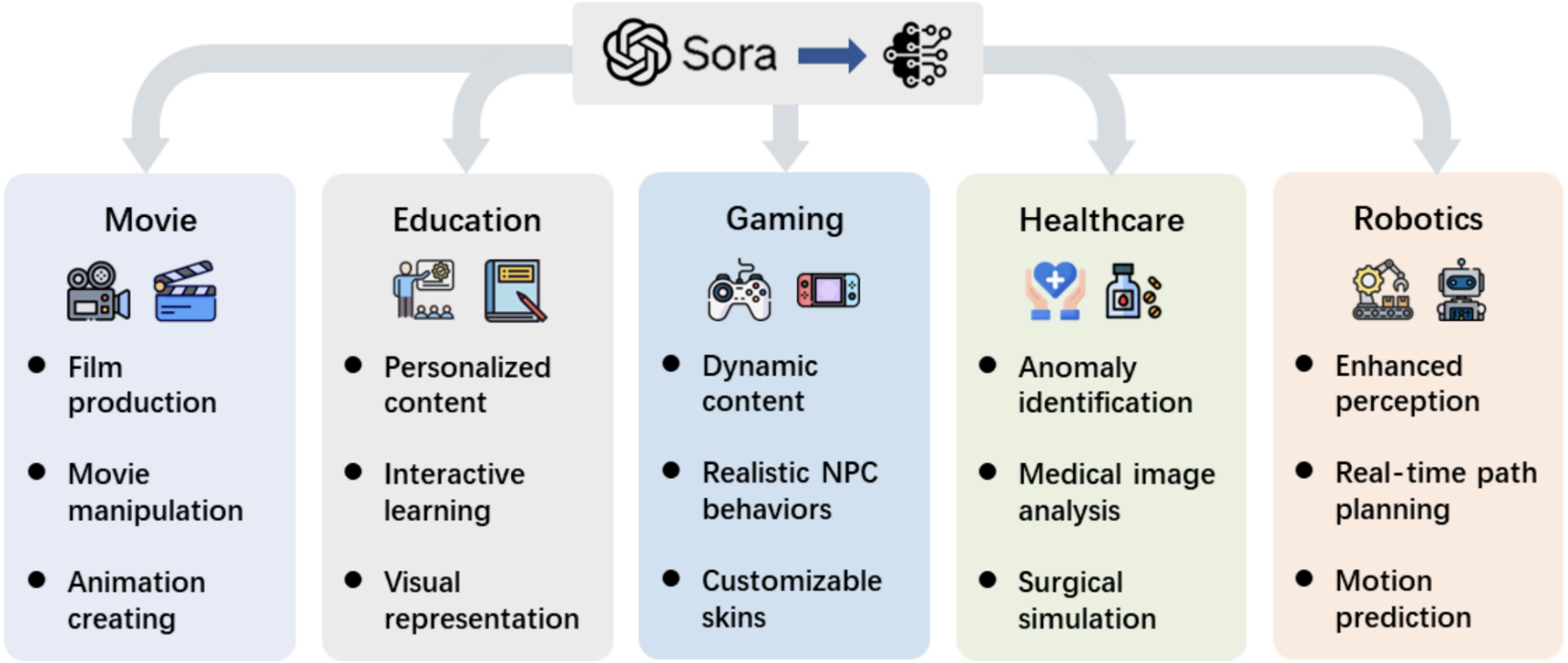
Emergent properties






Increasing transformer size
Decreasing patch size

Applications



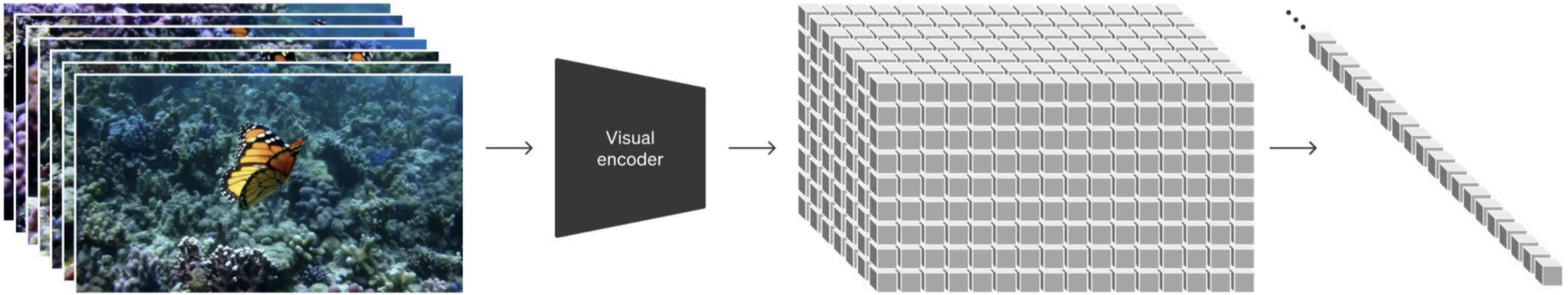
Data

- The main challanges of variable durations, resolutions and aspect ratios
- Traditional methods often resize, crop or adjust the aspect ratios of videos to fit a uniform standard

(a) Training on videos that are cropped to squares leads to unnatural compositions and framing.

(b) Training in native sizes improves framing.

Unified Visual Representation



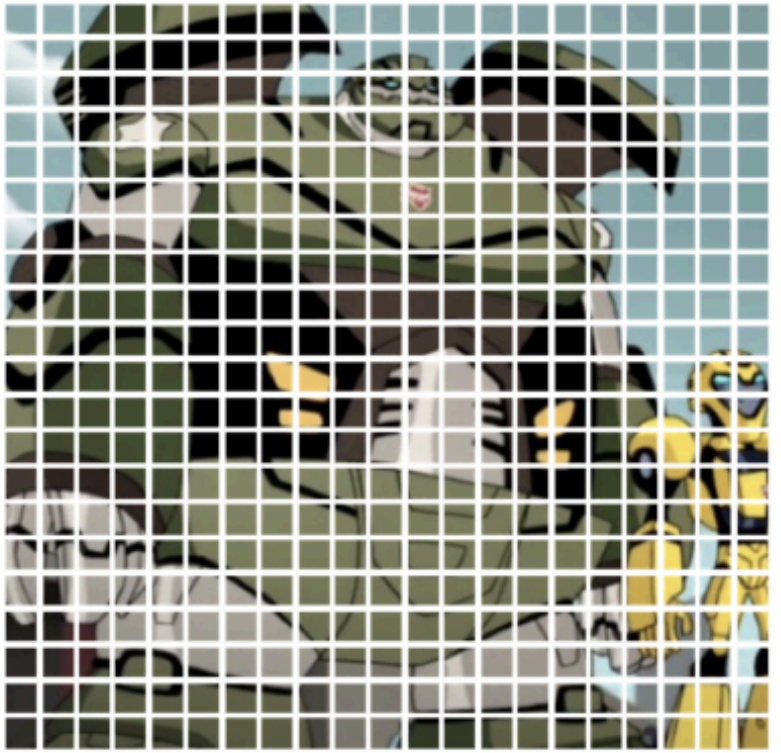
Video Compression Network

Spatial Patch Compression

- Convert video frames into fixed size patches
- Temporal dimension variability, Pre-trained visual encoders, Temporal information aggregation

Cropped Image

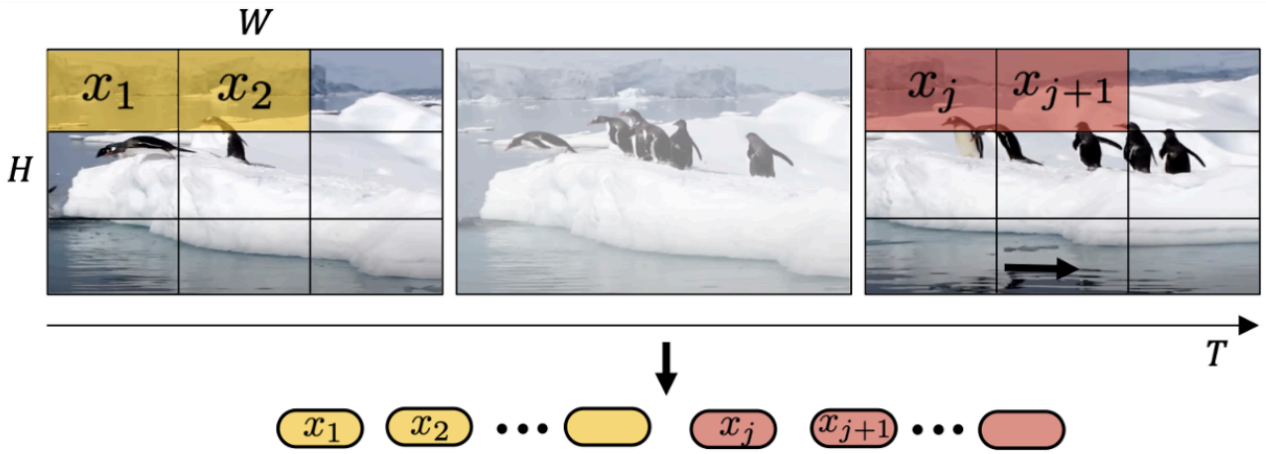
Image Patches

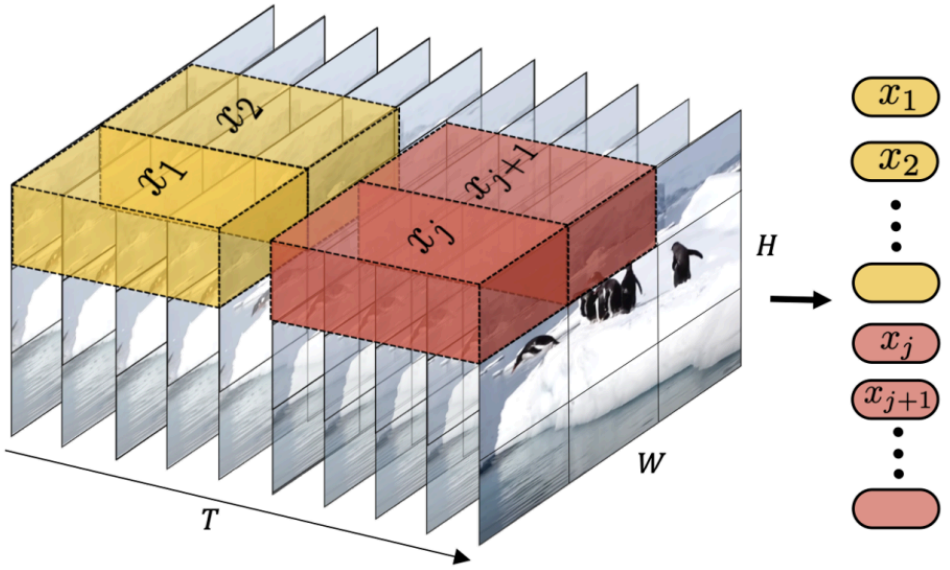


Flattened Image Patches

Spatial-Temporal Patch Compression

- Captures dynamic changes across the frames.
- 3D convolution - fixed kernel sizes, strides and output channels
- Since, Sora aims to generate high fidelity videos, a large patch size/ kernel size is used for efficient compression.
- Varying-size patches could also be used but it would compromise the positional embeddings.
- Still a concern how to handle varying number of patches.


W
H
T
x_1
x_2
x_j
x_{j+1}


H
W
T
x_1
x_2
x_j
x_{j+1}

Spacetime Latent Patches

- Solution - Patch n' pack
- Enables variable resolution, preserves aspect ratio.
- Addresses two challenges:-
  1. Compact Packing
  2. Token Dropping

Why It Matters:

- The PNP technique optimizes transformer-based models for variable-length inputs by:
  1. Efficiently handling sequence lengths through compact packing.
  2. Reducing unnecessary computational load by discarding redundant tokens.
  3. Using the fixed sequence lengths necessary for batched operations.

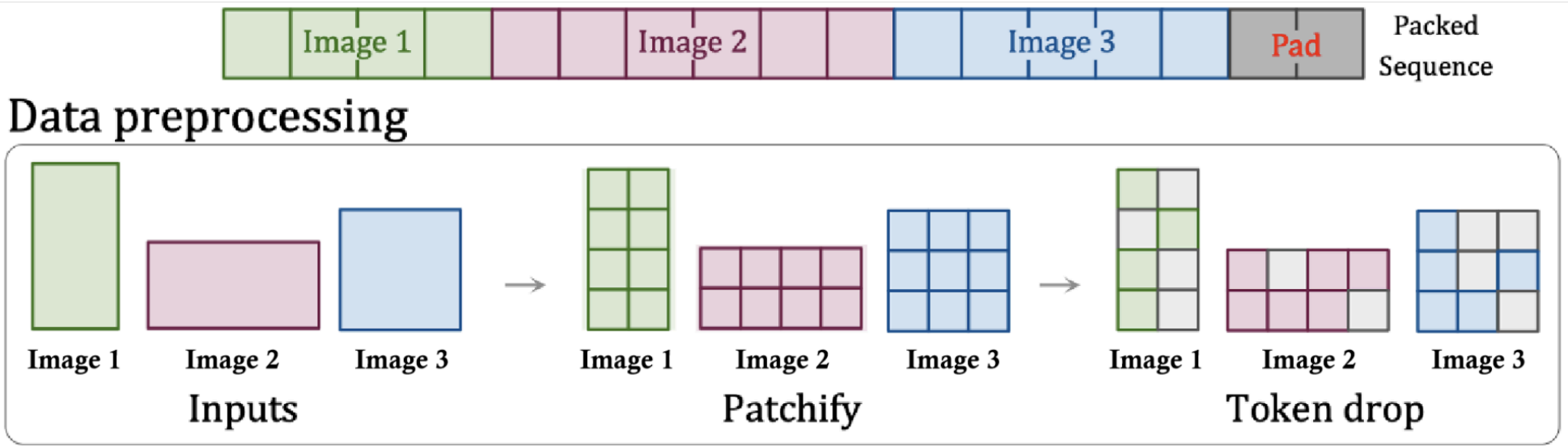
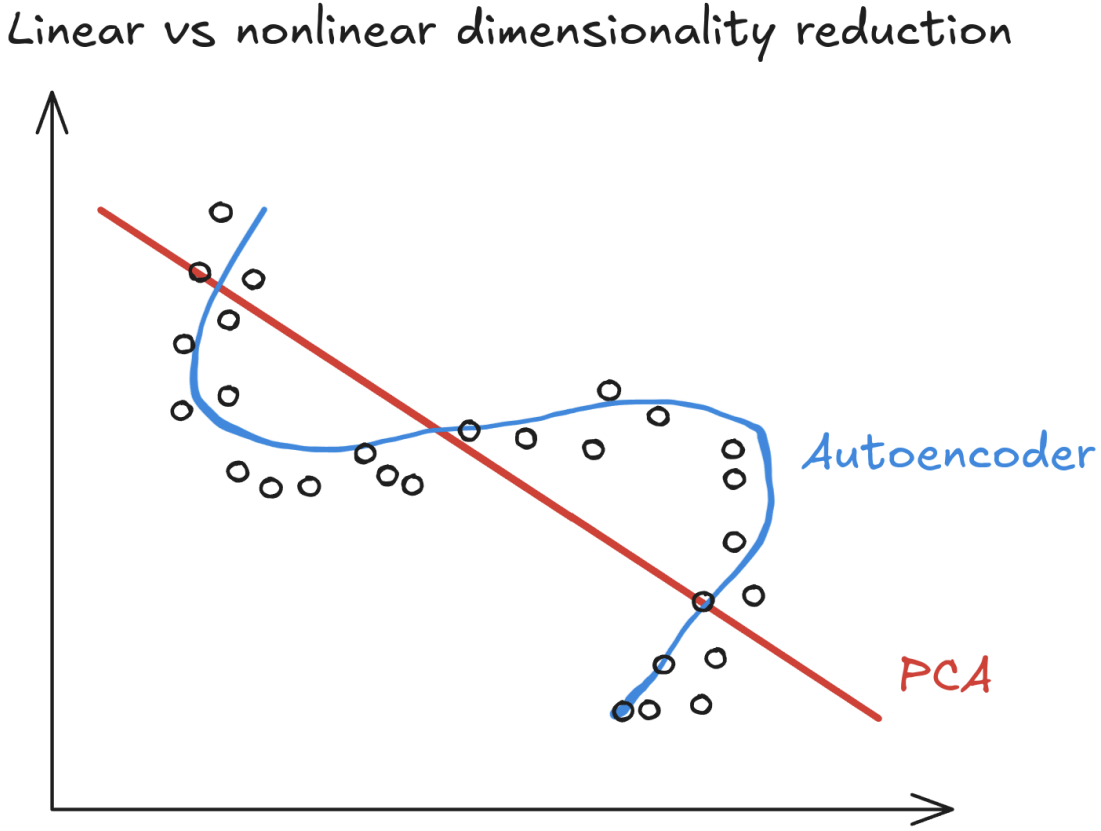


Image 1
Image 2
Image 3
Pad
Packed Sequence
Data preprocessing
Image 1
Image 2
Image 3
Inputs
Image 1
Image 2
Image 3
Patchify
Image 1
Image 2
Image 3
Token drop

Presumed Architecture

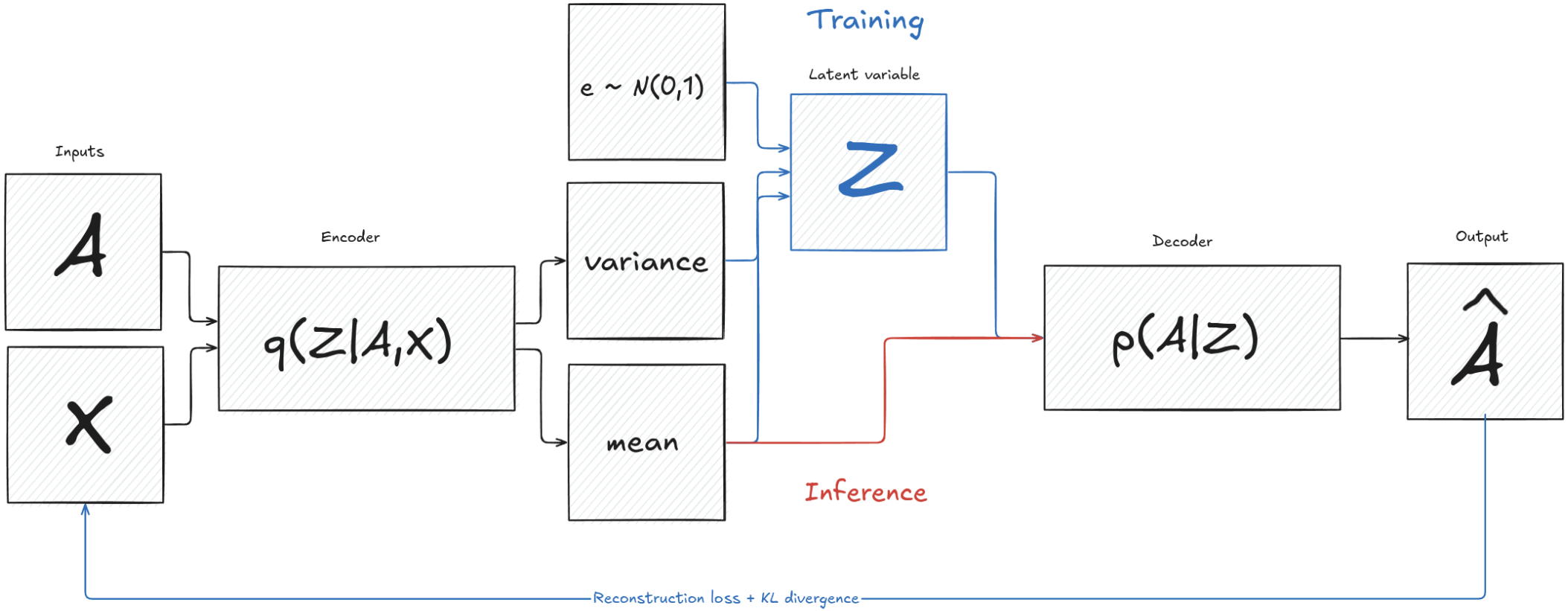
How do the experts think the SORA works?



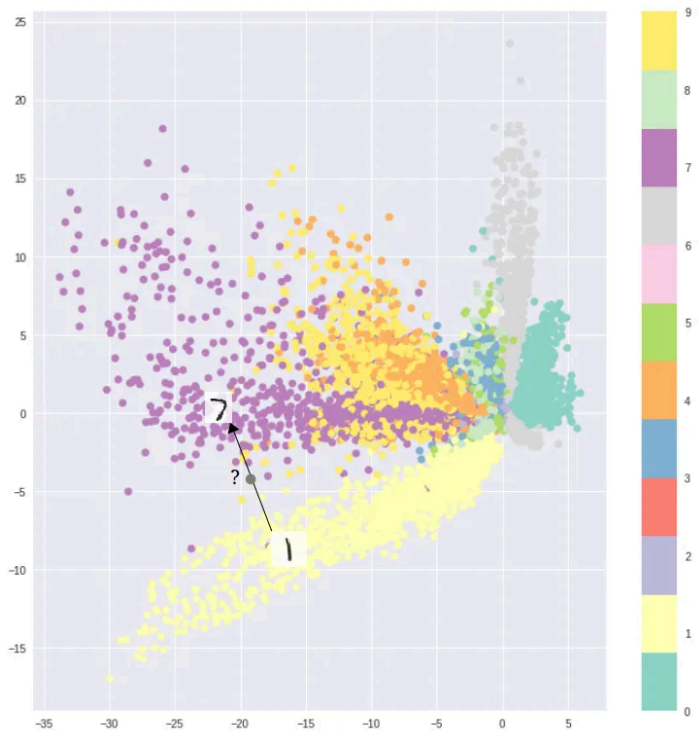
Video Compression

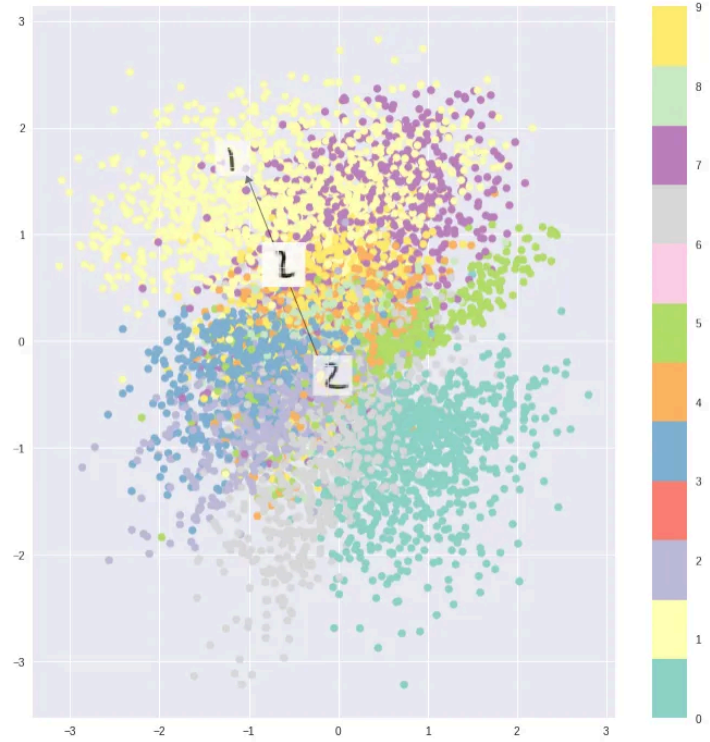
> An encoder aims aiming to reduce the dimensionality of input data and output a latent representation that is compressed both temporally and spatially.

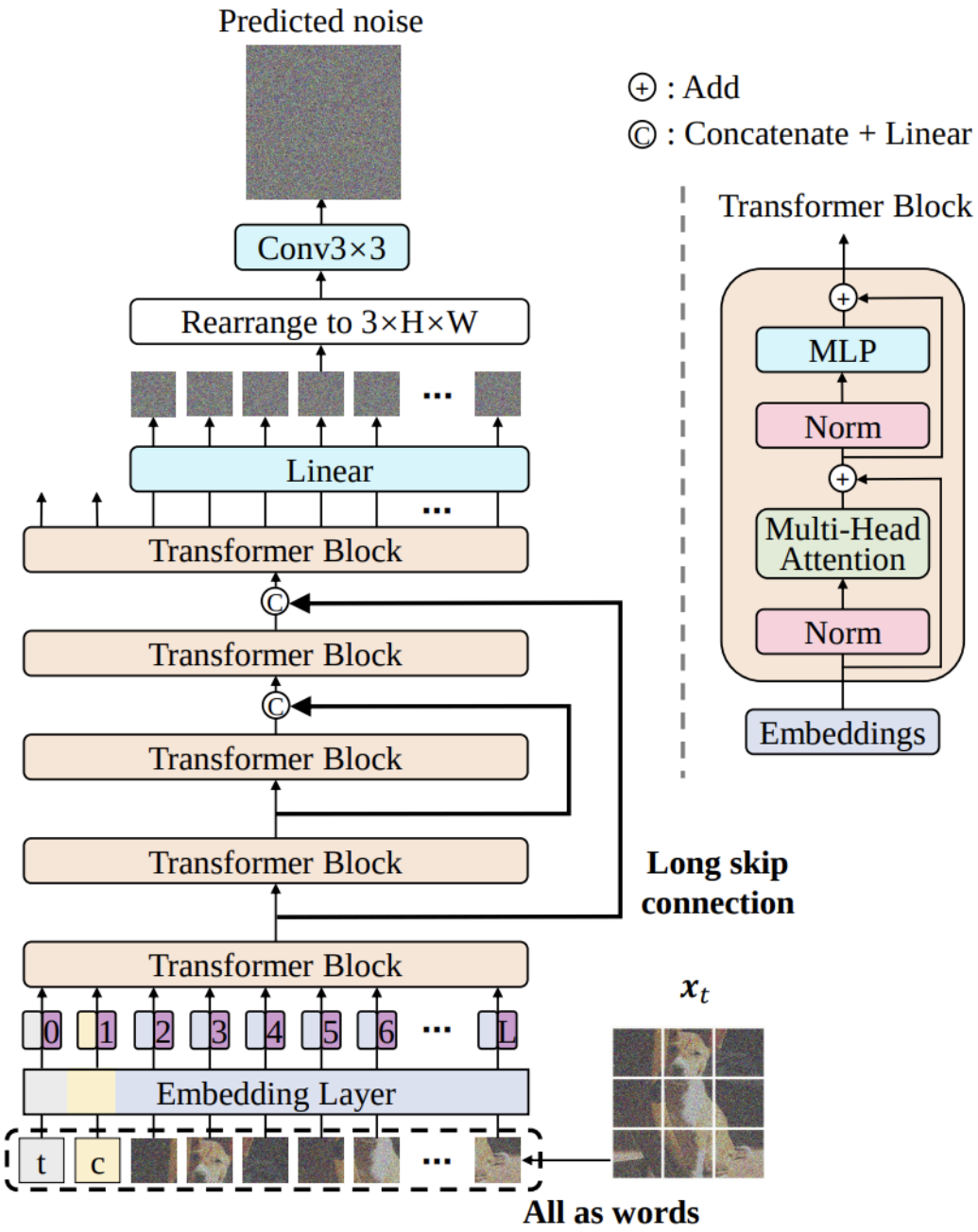
Variational Autoencoder



Regular Autoencoder Variational Autoencoder





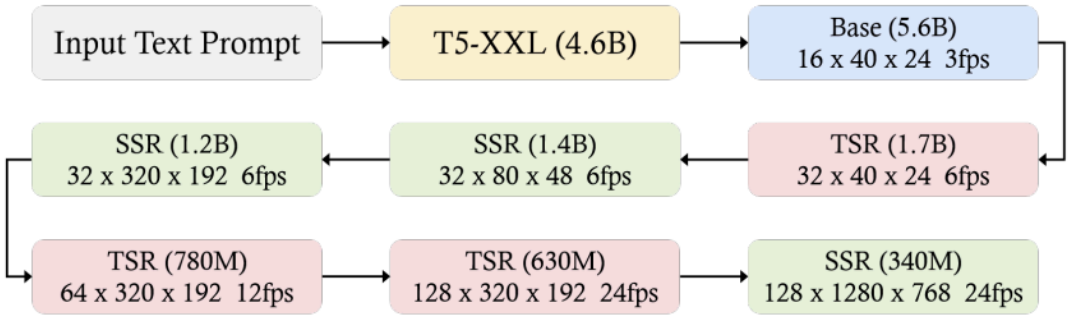


Vision Transformers

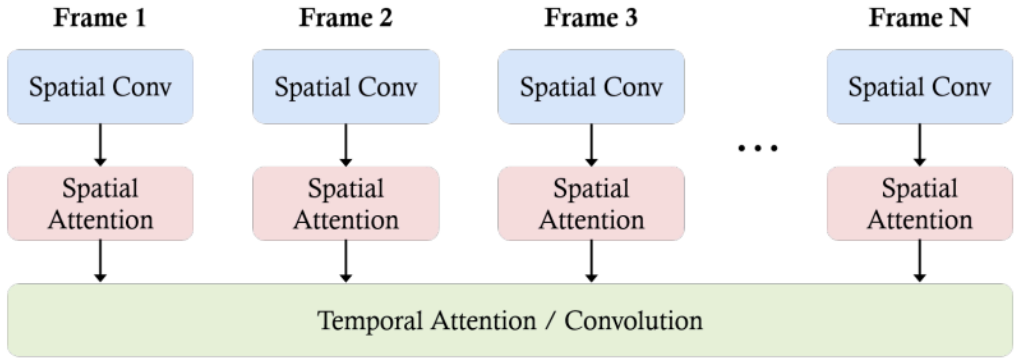
> The U-ViT architecture for diffusion models, which is characterized by treating all inputs including the time, condition and noisy image patches as tokens and employing skip connections known from U-Net.

Imagen Video Framework

Allows the system to focus computational resources on producing fine details only where needed, rather than generating high-resolution outputs in a single, computationally expensive pass.



(a) **Cascaded diffusion models.** The cascaded sampling pipeline with a base diffusion model and six up-sampling models that operate spatially and temporally. The text embeddings are injected into all the diffusion models.



(b) **Video U-Net space-time separable block.** Spatial operations are performed independently over frames with shared parameters, whereas the temporal operation mixes activations over frames. Temporal attention is only used in the base model for memory efficiency.

Figure 13: The overall framework of Imagen Video. Source: Imagen Video [29].

A colorful professional animated logo for 'Diffusion' written using paint brush in cursive. Smooth animation.

Sprouts in the shape of text 'Imagen' coming out of a fairytale book.

Thousands of fast brush strokes slowly forming the text 'Imagen Video' on a light beige canvas. Smooth animation.

Figure 10: Snapshots of frames from videos generated by Imagen Video demonstrating the ability of the model to render a variety of text with different style and dynamics.

Latent Diffusion Models

1. Convert input into latent representation
2. Perform diffusion
3. Convert back to the regular representation

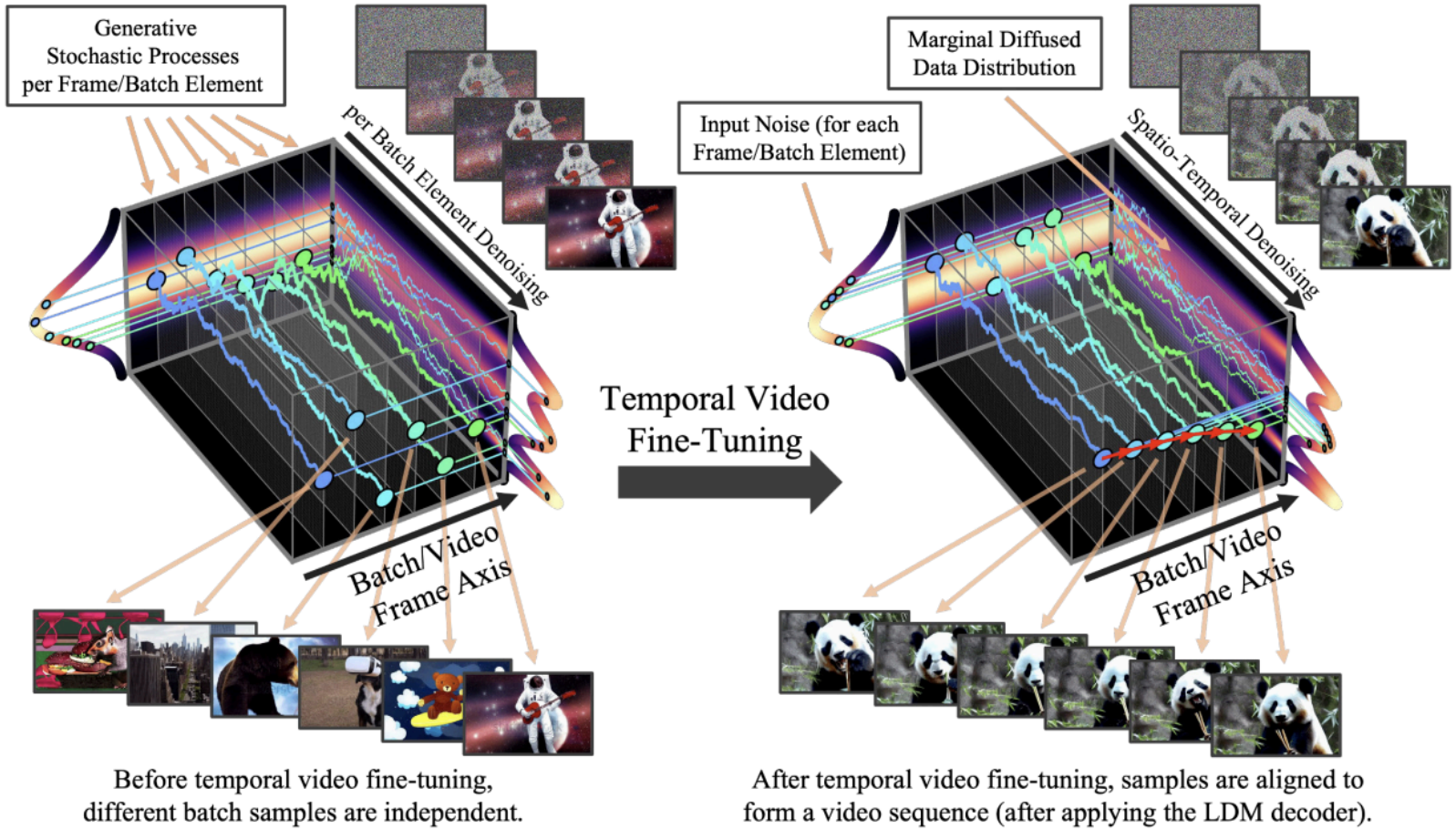


Figure 2. **Temporal Video Fine-Tuning.** We turn pre-trained image diffusion models into temporally consistent video generators. Initially, different samples of a batch synthesized by the model are independent. After temporal video fine-tuning, the samples are temporally aligned and form coherent videos. The stochastic generation process before and after fine-tuning is visualised for a diffusion model of a one-dim. toy distribution. For clarity, the figure corresponds to alignment in pixel space. In practice, we perform alignment in LDM's latent space and obtain videos after applying LDM's decoder (see Fig. 3). We also video fine-tune diffusion model upsamplers in pixel or latent space (Sec. 3.4).

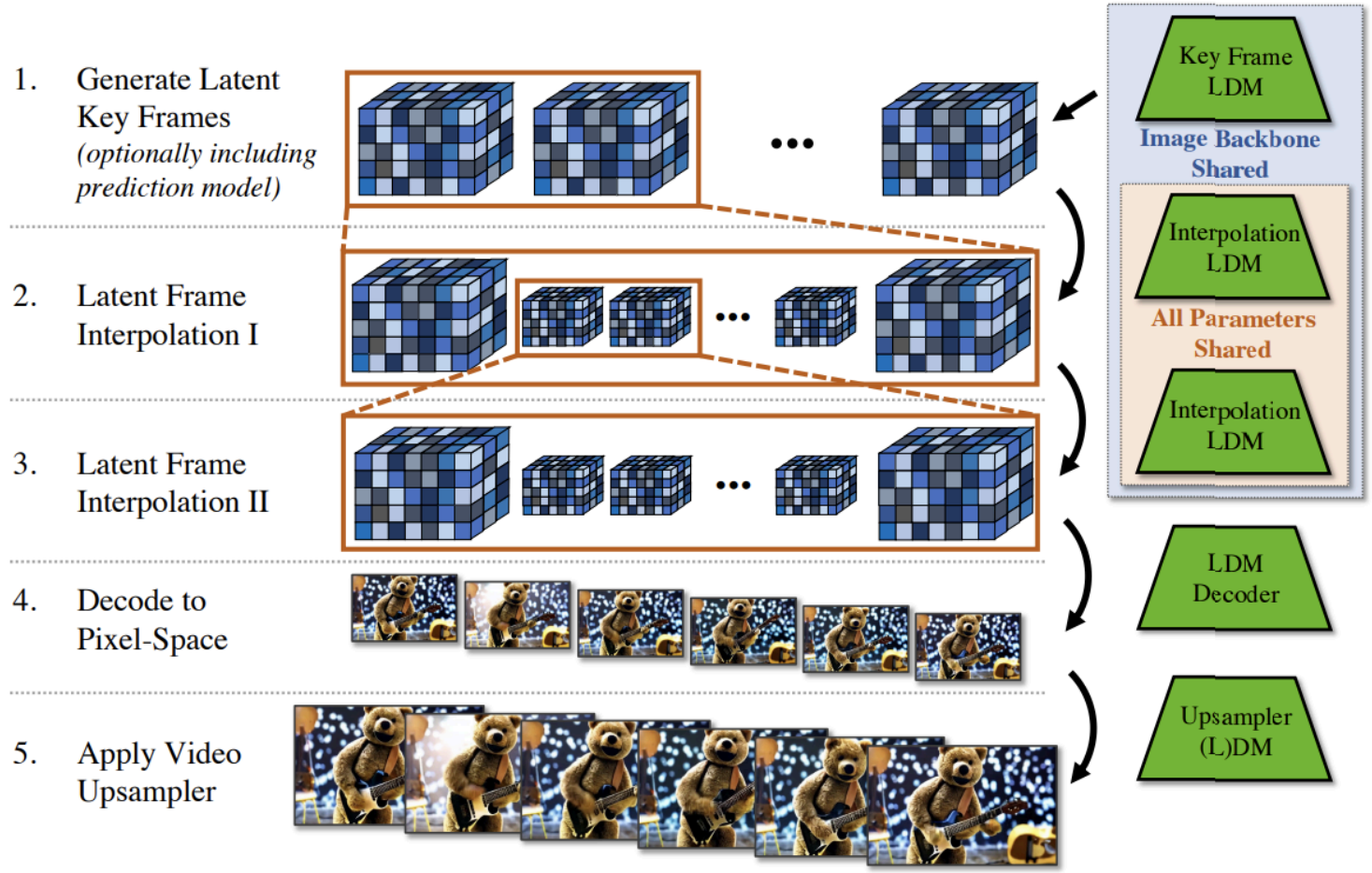


Figure 5. **Video LDM Stack.** We first generate sparse key frames. Then we temporally interpolate in two steps with the same interpolation model to achieve high frame rates. These operations are all based on latent diffusion models (LDMs) that share the same image backbone. Finally, the latent video is decoded to pixel space and optionally a video upsampler diffusion model is applied.

"A horse galloping through van Gogh's 'Starry Night'"

"A dog wearing virtual reality goggles in sunset, 4k, high resolution"

"The Orient Express driving through a fantasy landscape, animated oil on canvas"

Prompt Engineering

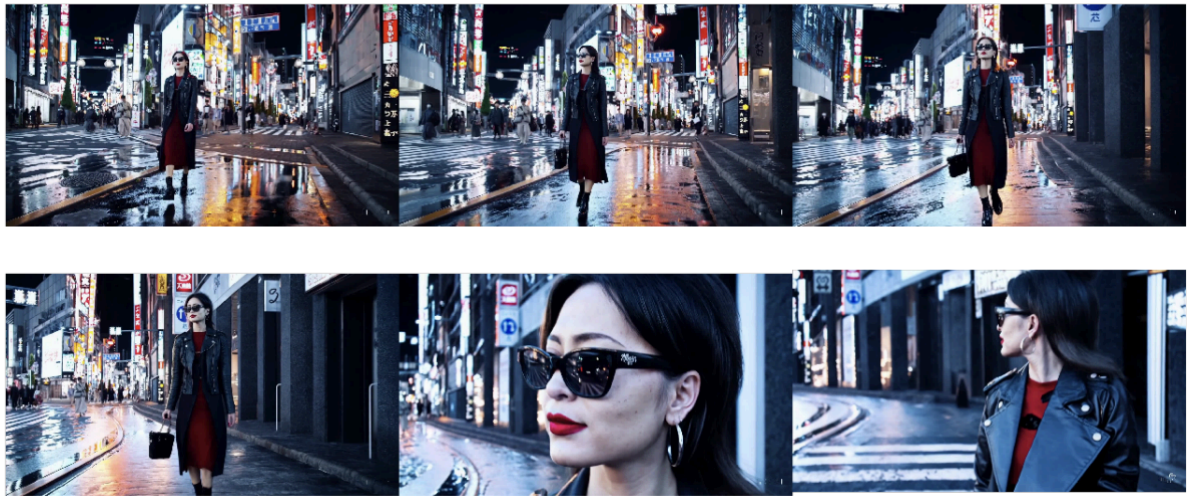
How do we talk with SORA?

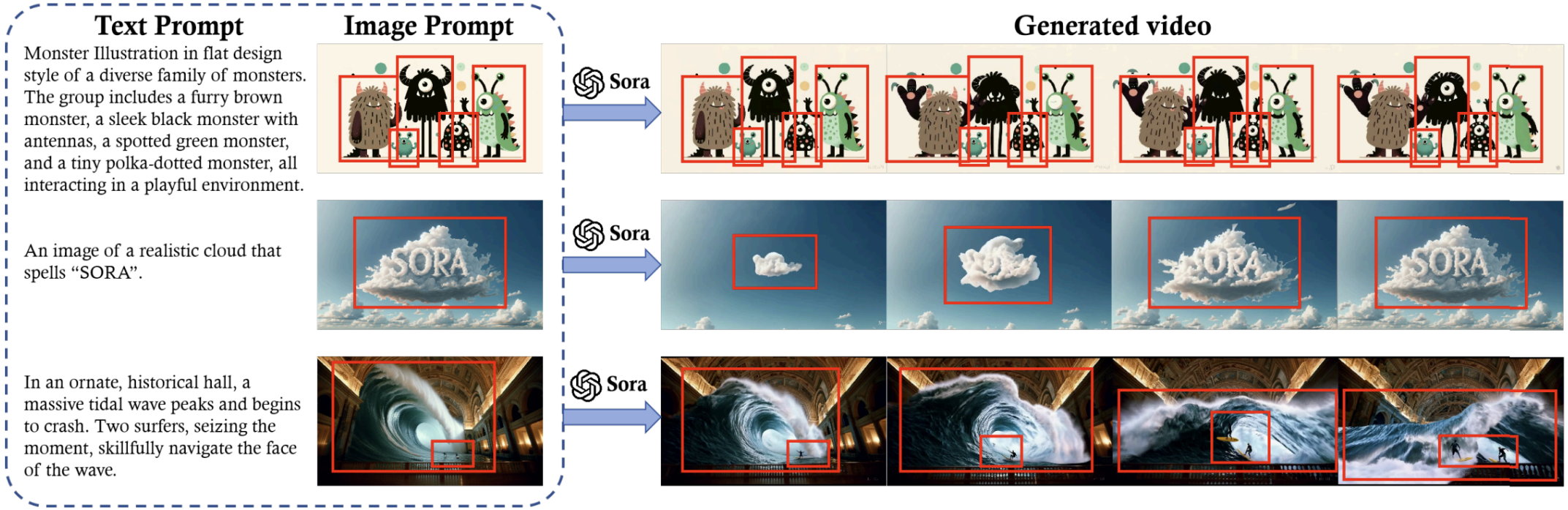
Text Prompt

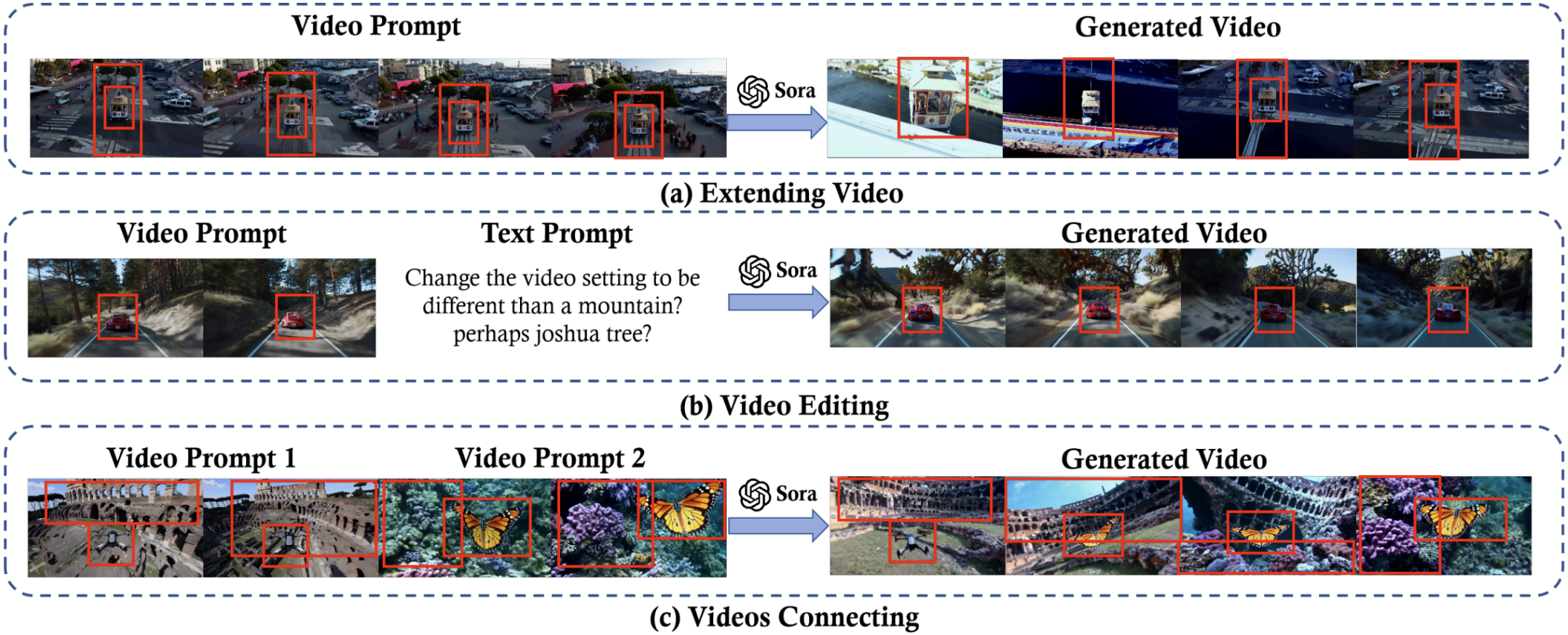
A stylish woman walks down a Tokyo street filled with warm glowing neon andanimated city signage. She wears a black leather jacket, a long red dress, and black boots, and carries a black purse. She wears sunglasses and red lipstick. She walks confidently and casually. The street is damp and reflective,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

Sora

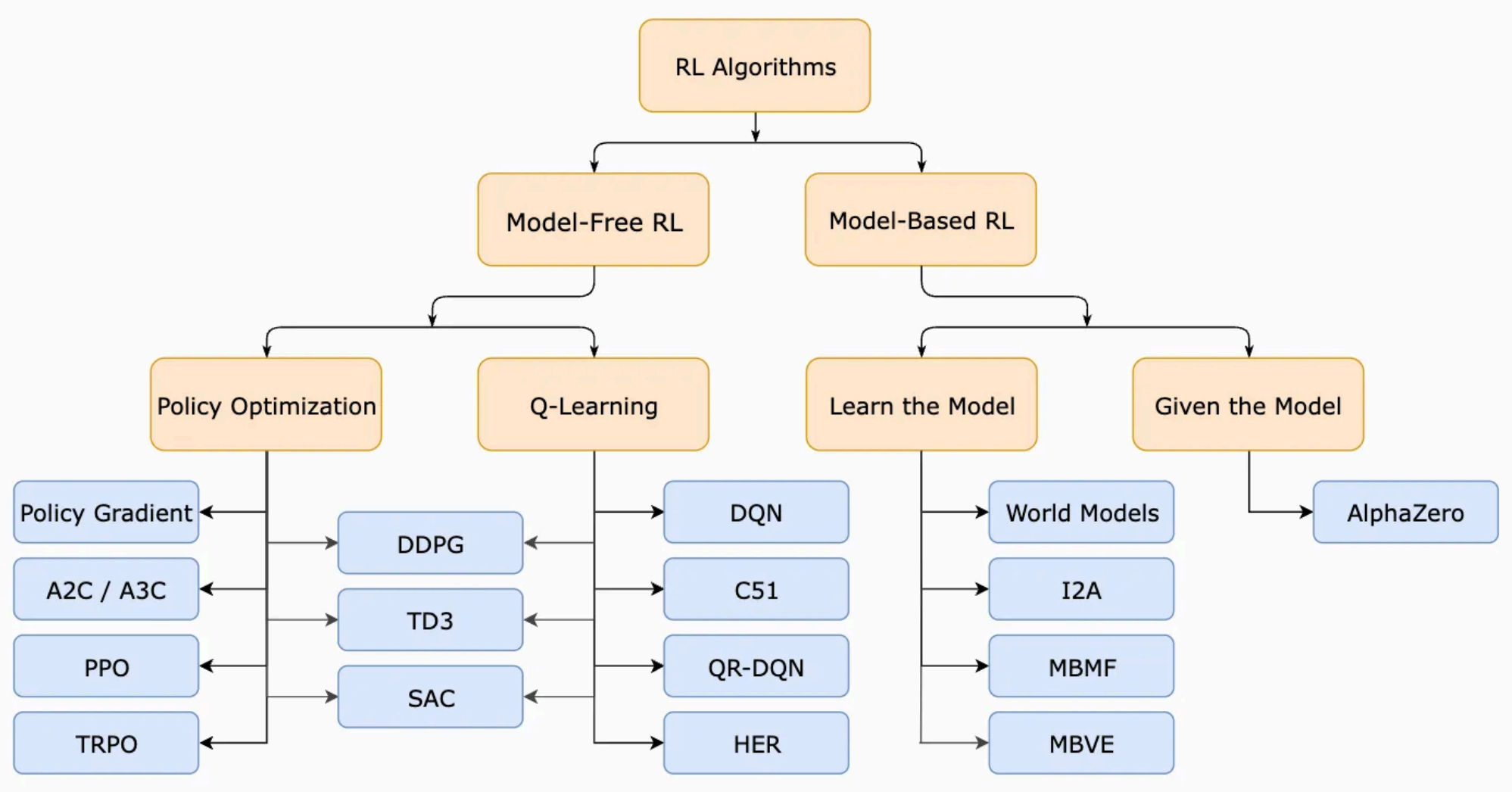
Generated video




Text Prompt
Image Prompt
Generated video
Monster Illustration in flat design style of a diverse family of monsters. The group includes a furry brown monster, a sleek black monster with antennas, a spotted green monster, and a tiny polka-dotted monster, all interacting in a playful environment.
Sora
An image of a realistic cloud that spells "SORA".
SORA
Sora
In an ornate, historical hall, a massive tidal wave peaks and begins to crash. Two surfers, seizing the moment, skillfully navigate the face of the wave.
Sora


Video Prompt
Sora
Generated Video
(a) Extending Video
Video Prompt
Text Prompt
Change the video setting to be different than a mountain? perhaps joshua tree?
Sora
Generated Video
(b) Video Editing
Video Prompt 1
Video Prompt 2
Sora
Generated Video
(c) Videos Connecting

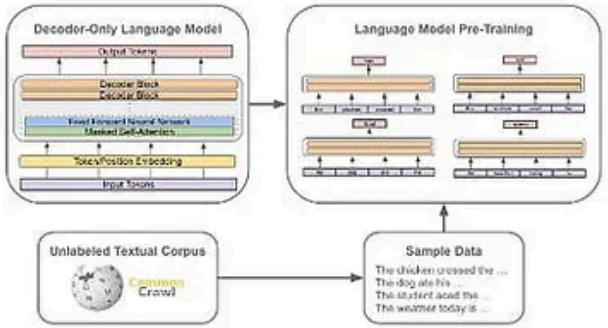
Alignment and Learning with Human Feedback


RL Algorithms
Model-Free RL
Model-Based RL
Policy Optimization
Q-Learning
Learn the Model
Given the Model
Policy Gradient
A2C / A3C
PPO
TRPO
DDPG
TD3
SAC
DQN
C51
QR-DQN
HER
World Models
I2A
MBMF
MBVE
AlphaZero

“Proximal Policy Optimization (PPO) is an algorithm in the field of reinforcement learning that trains a computer agent’s decision function to accomplish difficult tasks. PPO was developed by John Schulman in 2017,had become the default reinforcement learning algorithm at American artificial intelligence company OpenAI.”

Alignment

Pre-Training



SFT

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

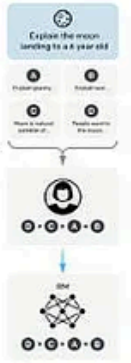
This data is used to fine-tune GPT-3 with supervised learning.

RLHF

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.

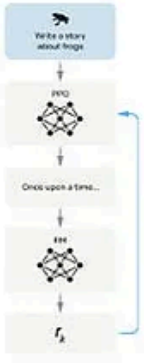


A new prompt is sampled from the dataset.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.



Final Remarks

Sources

- SORA
- U-Vit
- Imagen Video
- Latent Diffusion Models

End