UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models

Heting, Priyam, Xulin, Mahir, Kamila

2024/03/27

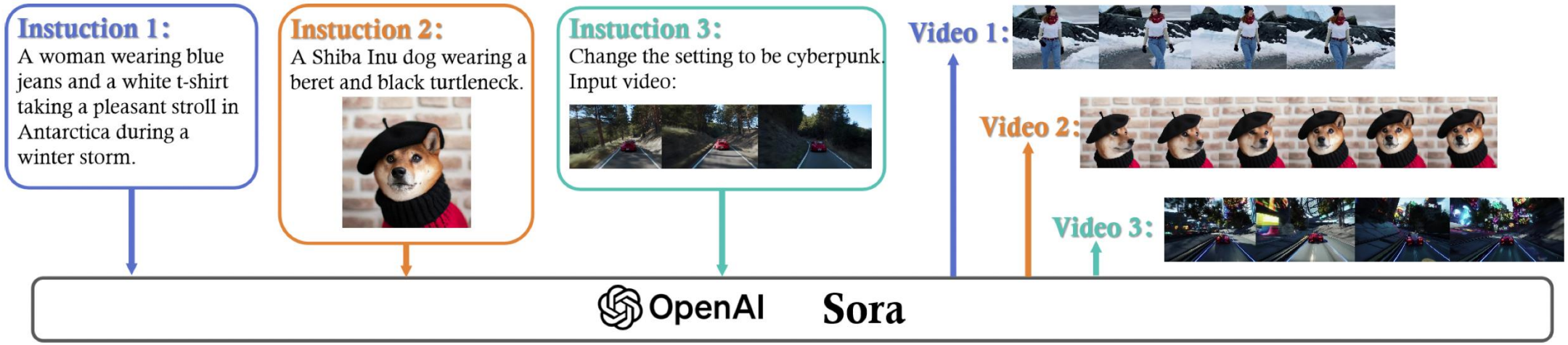
Overview

1. **Introduction – Priyam**
2. **Background – Priyam**
3. **Technology:**
    1. **Overview – Priyam**
    2. **Data Pre-Processing – Mahir**
    3. **Modeling – Heting**
    4. **Language Instruction Following – Xulin**
    5. **Prompt Engineering – Xulin**
    6. **Trustworthiness – Kamila**
4. **Applications – Kamila**
5. **Limitation – Kamila**
6. **Conclusion – Kamila**

Overview

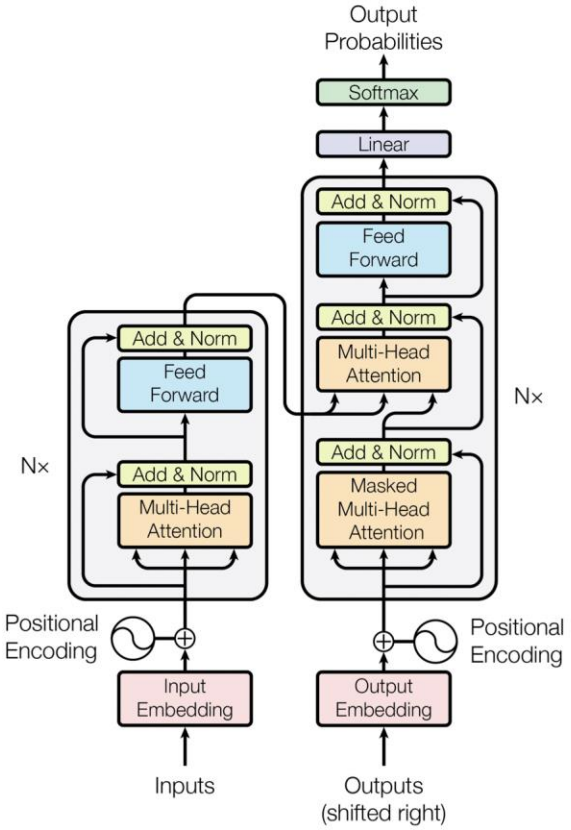
State of the Art Text to Video Diffusion

- Capable of generating 1-minute long videos!
- Uses a pretrained Diffusion Transformer backbone
- Not publicly released yet due to the ethical and social implications of the ability to generate video

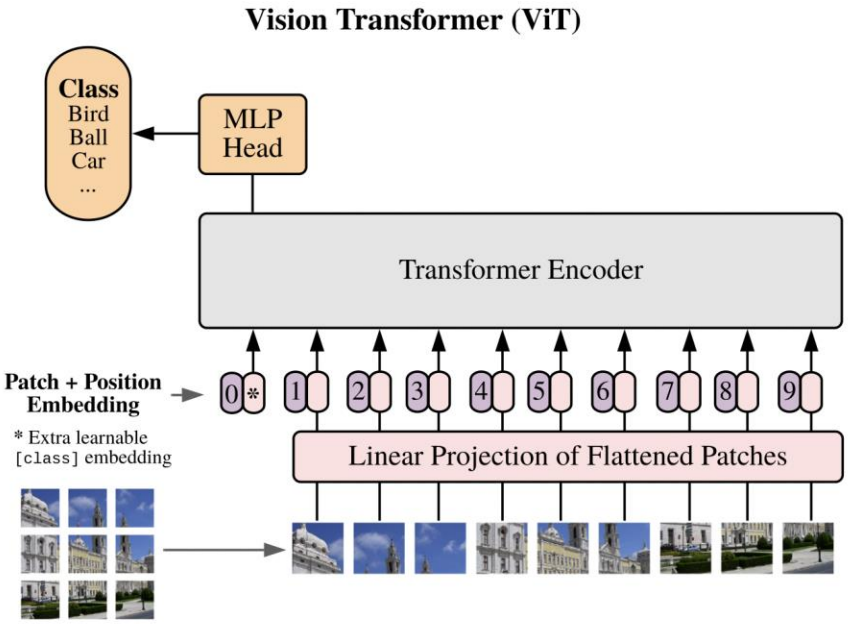


Moving from Language Models to Vision Models

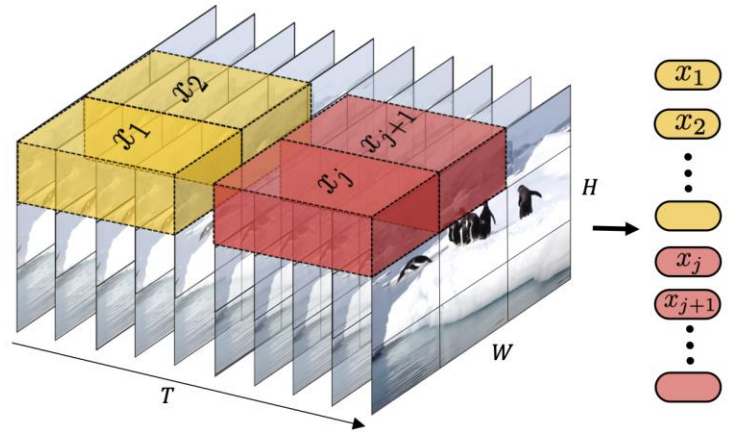
Transformers



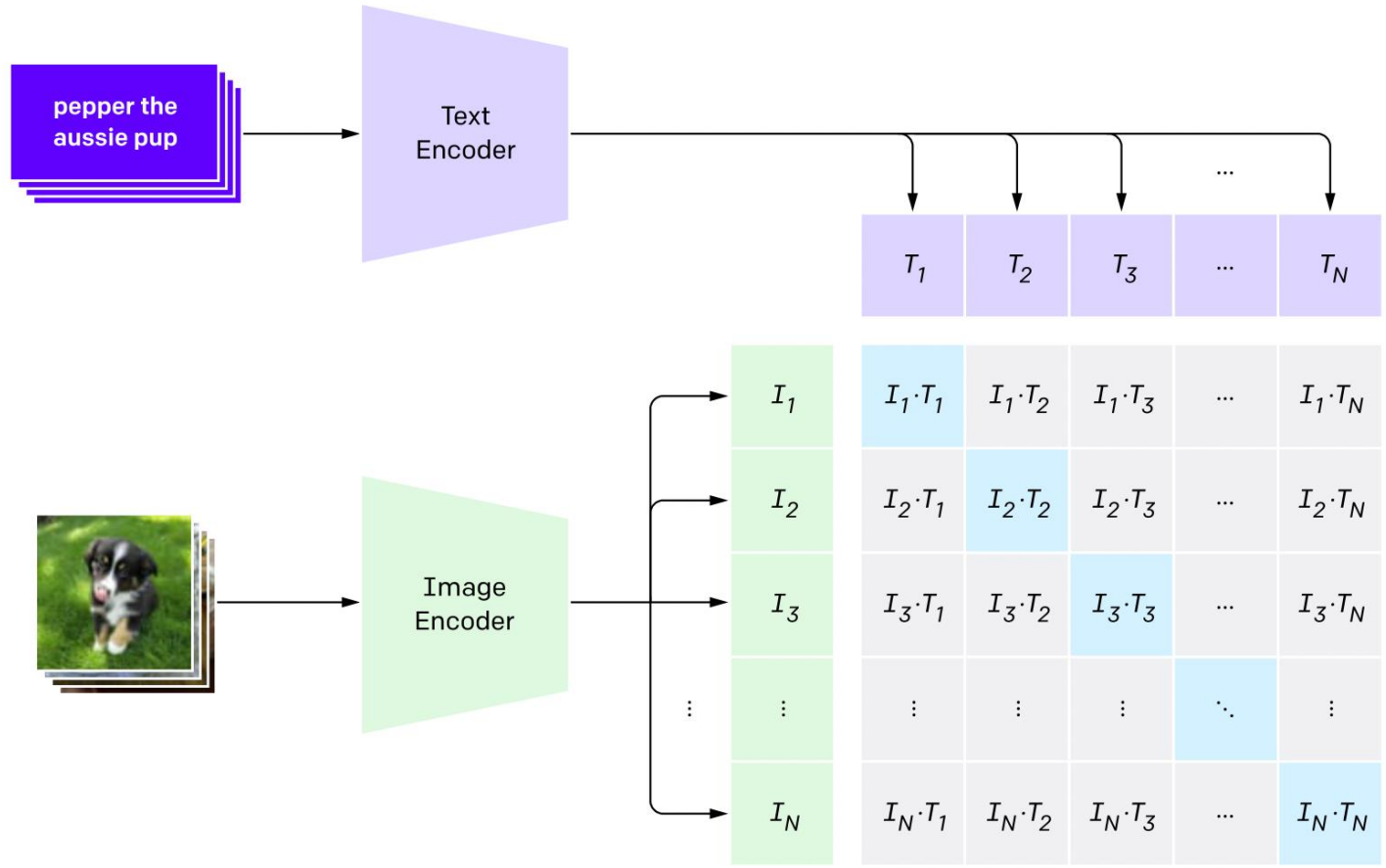
Vision Transformer



Video Vision Transformer



Shared Embeddings Space for Language and Vision

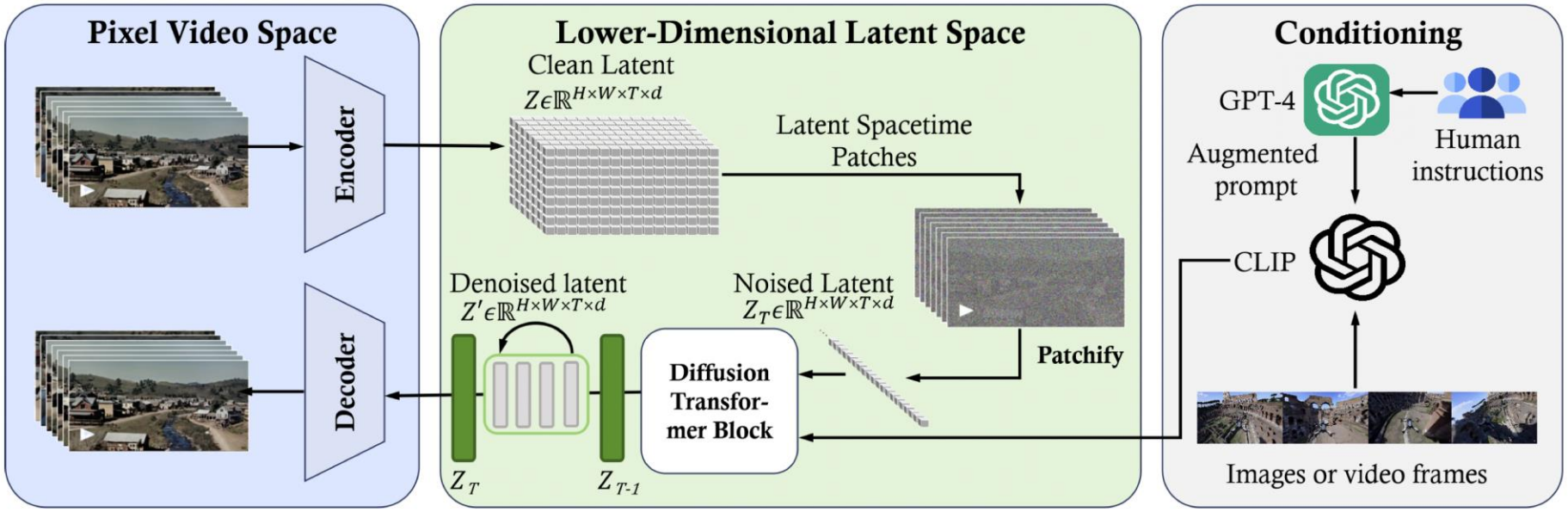


Teach model to recognize relationship between text and images.

- Contrastive learning where every text is treated as the "correct answer" for its corresponding image, in which case we can just use cross entropy!

Latent Diffusion Modeling

The pixel space is extremely large, so taking a hint from Stable Diffusion, SORA is a Latent Diffusion Model. The Video is first compressed to a lower dimension space, diffusion is learned from lower dimension noise to compressed video, and then the final output is uncompressed back to the pixel space!



Data Pre-Processing — Overview

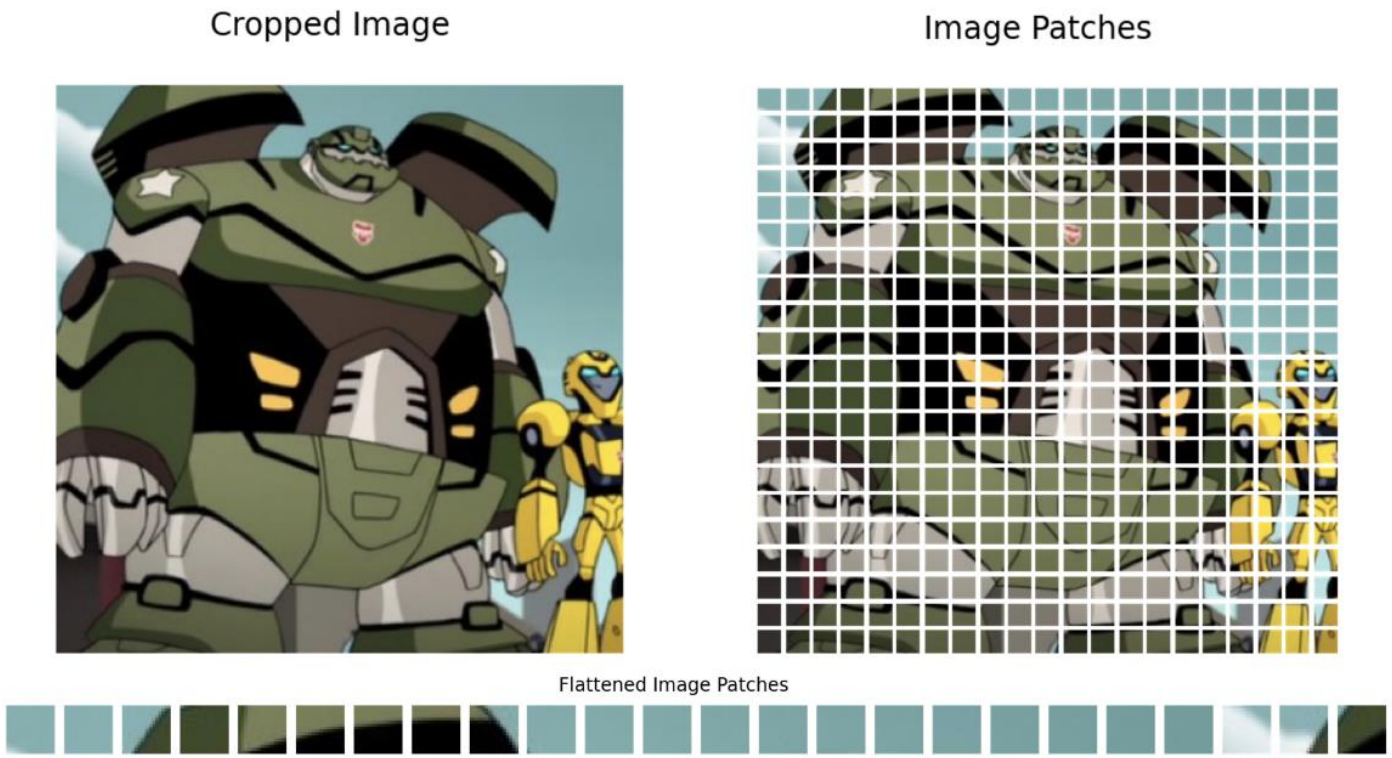




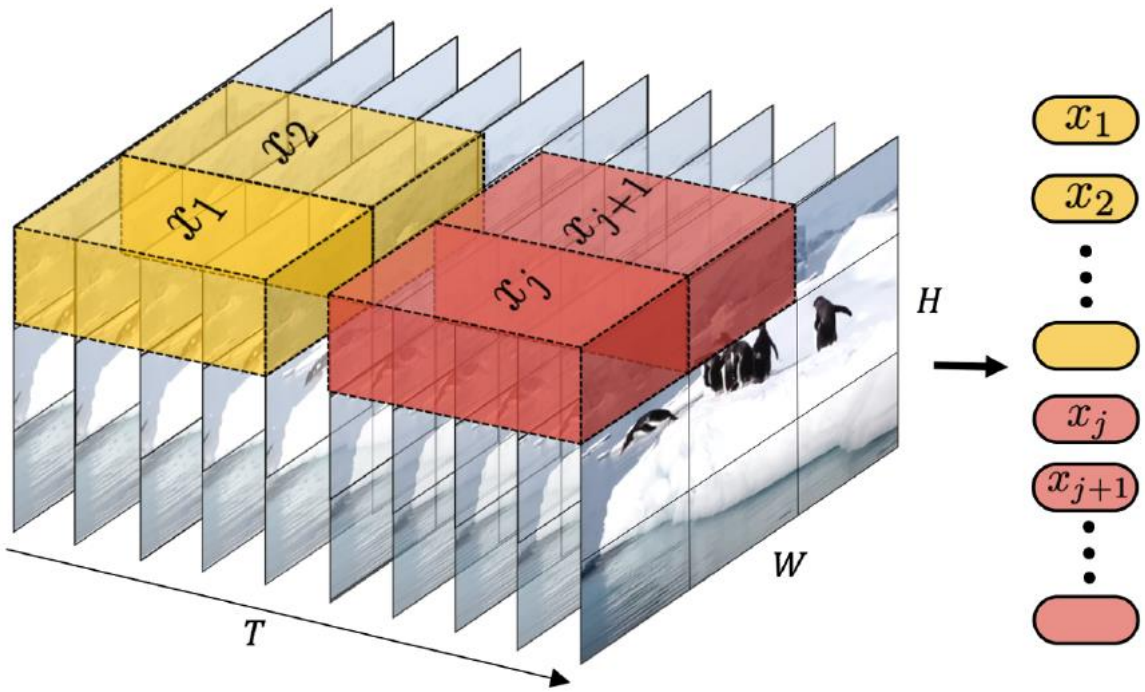
- Training on and generating varied resolutions and aspect ratios
  - Improved composition and framing of depicted subjects
  - Favors greater computation over manual feature engineering
  - Empirically more natural and coherent results therefrom
- Relies on unified visual data representation
  - How is this representation achieved?

Data Pre-Processing — Spatial-patch Video Compression

- Building spatio-temporal latent representation through linearizing transformed frame patches
  - Influenced by Vision Transformer and masked variational autoencoders
- Multiple issues to address:
  - Variable time dimension (frame sampling, fixed input length)
  - Visual encoder pre-training
  - Information aggregation through time



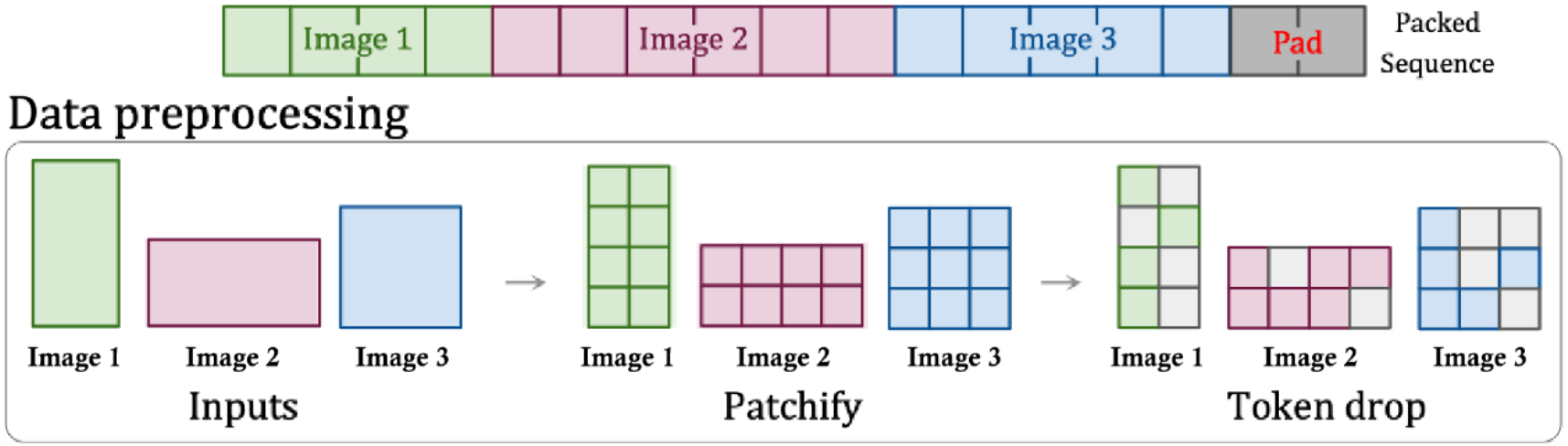
Data Pre-Processing — Spatial-temporal-patch Compression



- 3D convolution applied to video frames
  - Captures dynamism through time
- Issues to address without temporal axis addressable similarly
  - Frame padding/interpolation for shorter videos
- What of varying, though still large, patch sizes?
  - Positional encoding problems
  - Decoder challenges

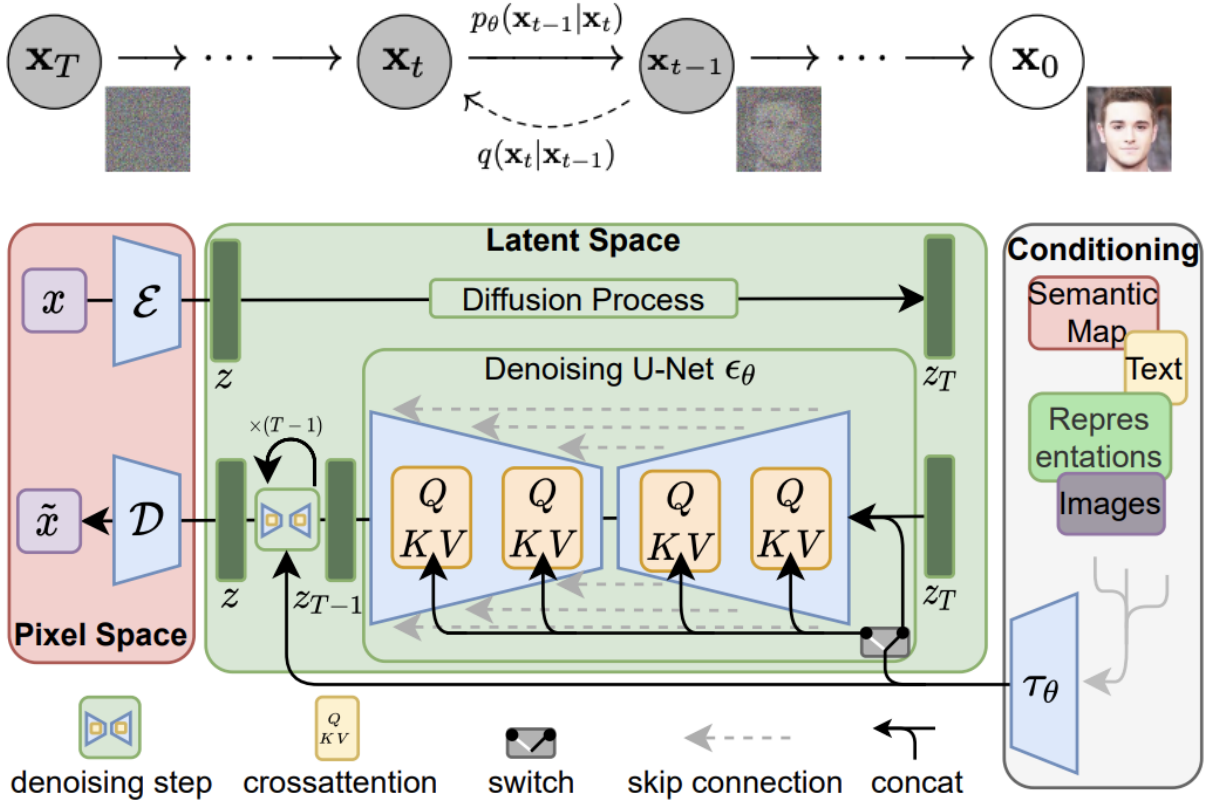
Data Pre-Processing — Patch packing

- Drop patches before linearly packing them
  - (inspired by NLP example packing)
  - Possibly apply to latent-space embeddings as well
  - How compactly—as far as padding goes—can this packing occur?
    - Greedily? Tuning frame resolution and frames sampled?
  - What patches should be dropped?
    - Those similar to others? Scheduling this? 3D consistency lost?



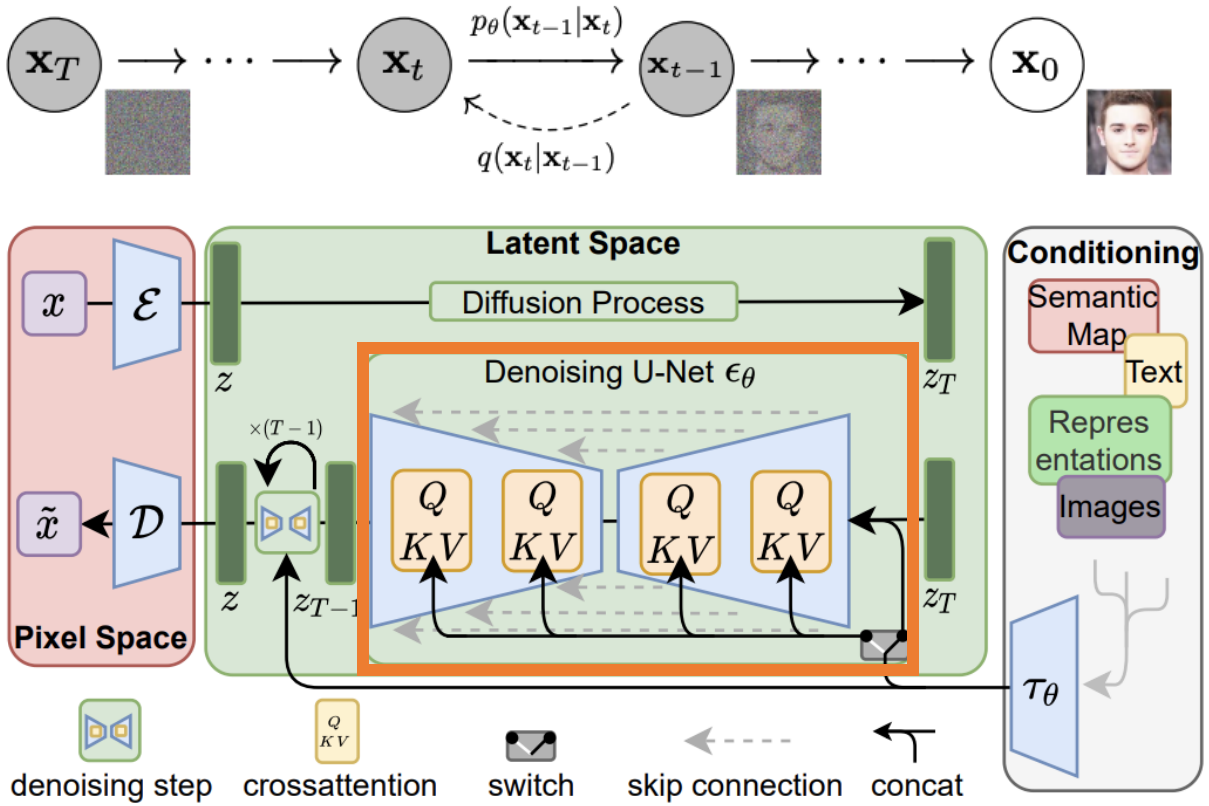
Modeling - Image Diffusion Transformer (Heting)

- Traditional diffusion models mainly leverage convolutional U-Nets for the denoising network backbone
  - Downsampling and upsampling blocks for the denoising network backbone



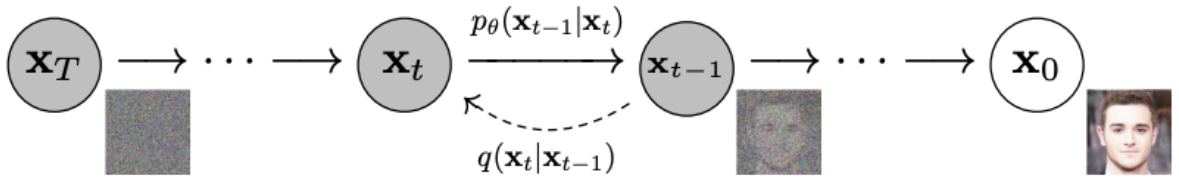
Modeling - Image Diffusion Transformer (Heting)

- Traditional diffusion models mainly leverage convolutional U-Nets for the denoising network backbone
  - Downsampling and upsampling blocks for the denoising network backbone



Modeling - Image Diffusion Transformer (Heting)

- Traditional diffusion models mainly leverage convolutional U-Nets for the denoising network backbone
  - Downsampling and upsampling blocks for the denoising network backbone
- By incorporating a more flexible transformer architecture, transformer-based diffusion models can use more training data and larger model parameters
  - **DiT** incorporates conditioning via **adaptive layer norm**, with an additional MLP layer for **zero-initializing**, which initializes each residual block as an identity function and thus greatly stabilizes the training process



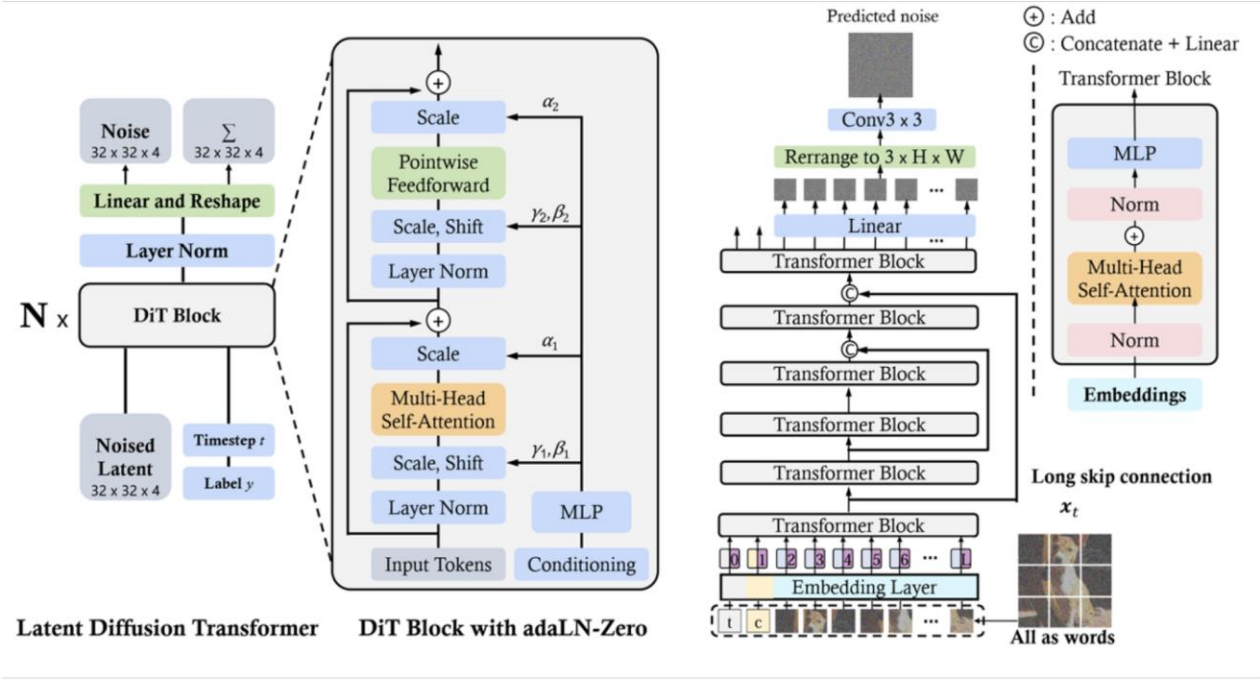
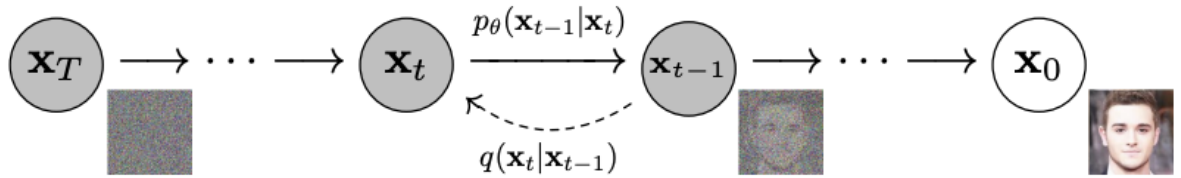


Figure 11: The overall framework of DiT (left) and U-ViT (right)

Modeling - Image Diffusion Transformer (Heting)

- Traditional diffusion models mainly leverage convolutional U-Nets for the denoising network backbone
  - Downsampling and upsampling blocks for the denoising network backbone
- By incorporating a more flexible transformer architecture, transformer-based diffusion models can use more training data and larger model parameters
  - **DiT** incorporates conditioning via **adaptive layer norm**, with an additional MLP layer for **zero-initializing**, which initializes each residual block as an identity function and thus greatly stabilizes the training process
  - **U-ViT** treats all inputs, condition, and noisy image patches, as tokens and propose long skip connections between the shallow and deep transformer layers



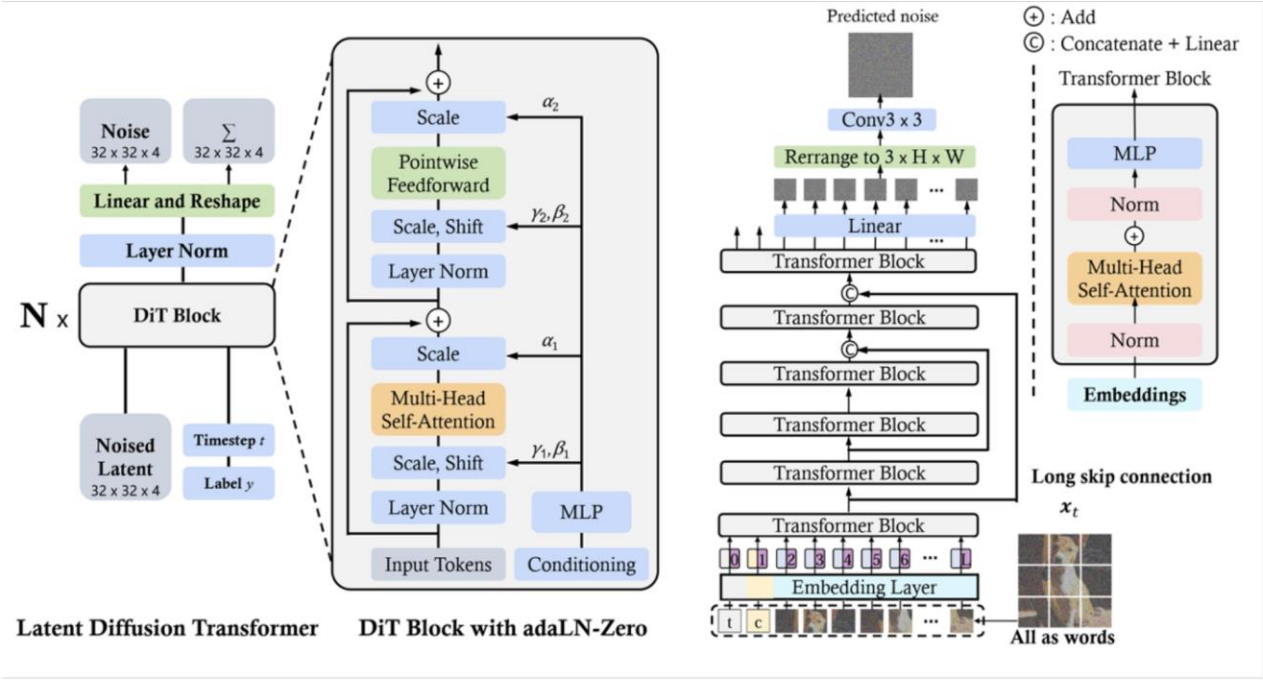


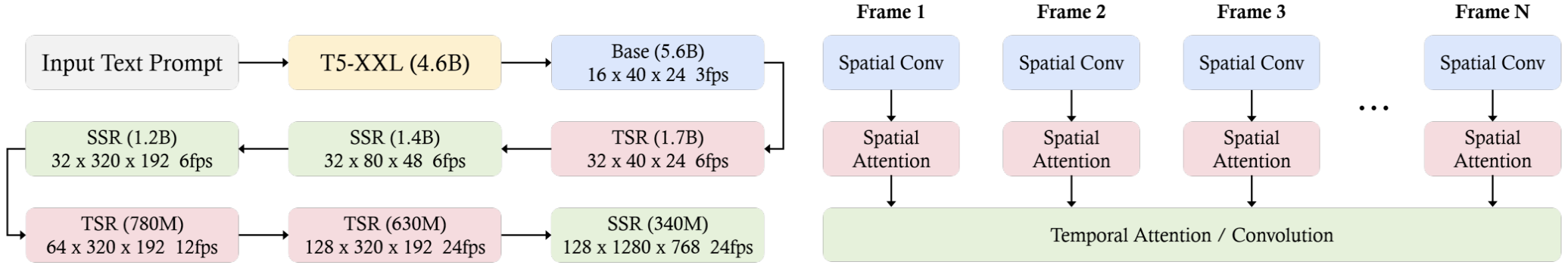
Figure 11: The overall framework of DiT (left) and U-ViT (right)

Modeling - Video Diffusion Transformer (Heting)

- Realizing the potential of diffusion transformers for text-to-video (T2V) generation tasks
  - How to compress the video spatially and temporally to a latent space for efficient denoising
  - **How to convert the compressed latent to patches and feed them to the transformer**
  - **How to handle long-range temporal and spatial dependencies and ensure content consistency.**

Modeling - Video Diffusion Transformer (Heting)

- Imagen Video
  - A cascade of diffusion models consisting of 7 sub-models that perform text-conditional video generation, spatial super-resolution, and temporal super-resolution, to transform textual prompts into high-quality videos
  - A frozen T5 text encoder generates contextual embeddings from the input text prompt
    - These embeddings are critical for aligning the generated video with the text prompt and are injected into all models in the cascade
  - The base video and super-resolution models use a 3D U-Net architecture in a spacetime separable fashion
  - The process involves joint training on both images and videos
  - Progressive distillation is applied to significantly reduce the computational load while maintaining perceptual quality



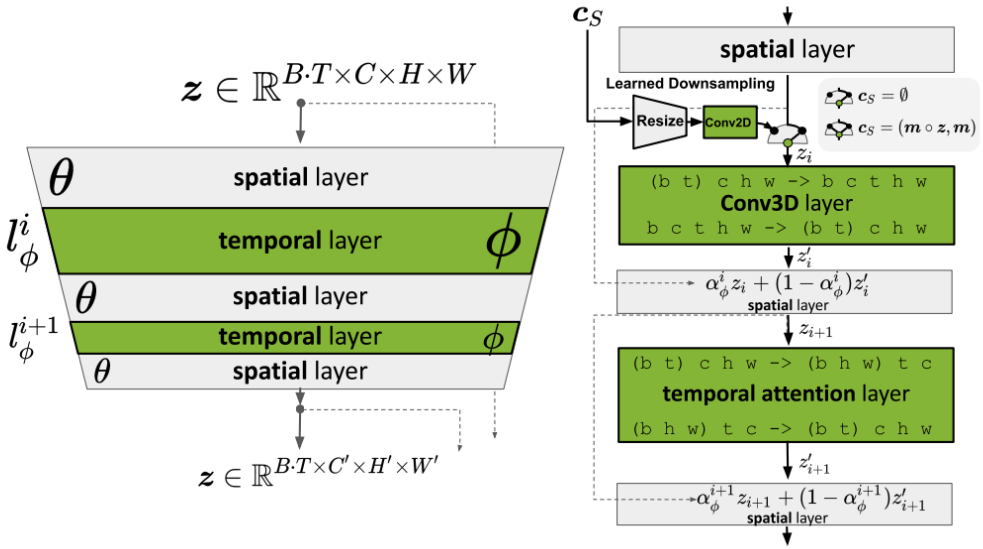
(a) **Cascaded diffusion models.** The cascaded sampling pipeline with a base diffusion model and six up-sampling models that operate spatially and temporally. The text embeddings are injected into all the diffusion models.

(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.

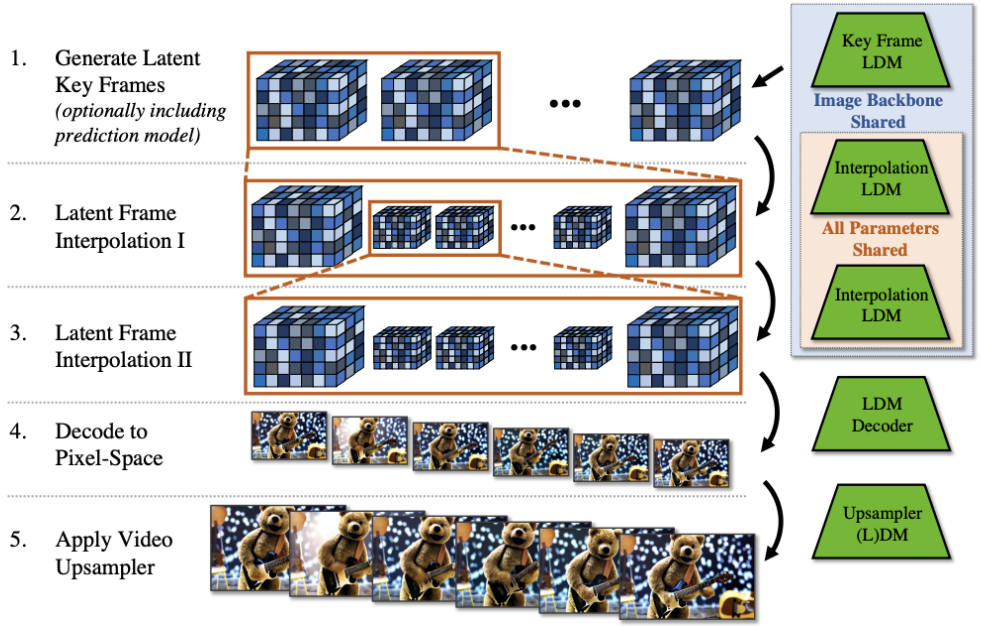
Figure 13: The overall framework of Imagen Video. Source: Imagen Video [29].

Modeling - Video Diffusion Transformer (Heting)

- Video LDM



(a) **Additional temporal layer**. A pre-trained LDM is turned into a video generator by inserting temporal layers that learn to align frames into temporally consistent sequences. During optimization, the image backbone θ remains fixed and only the parameters ϕ of the temporal layers l^i_ϕ are trained.

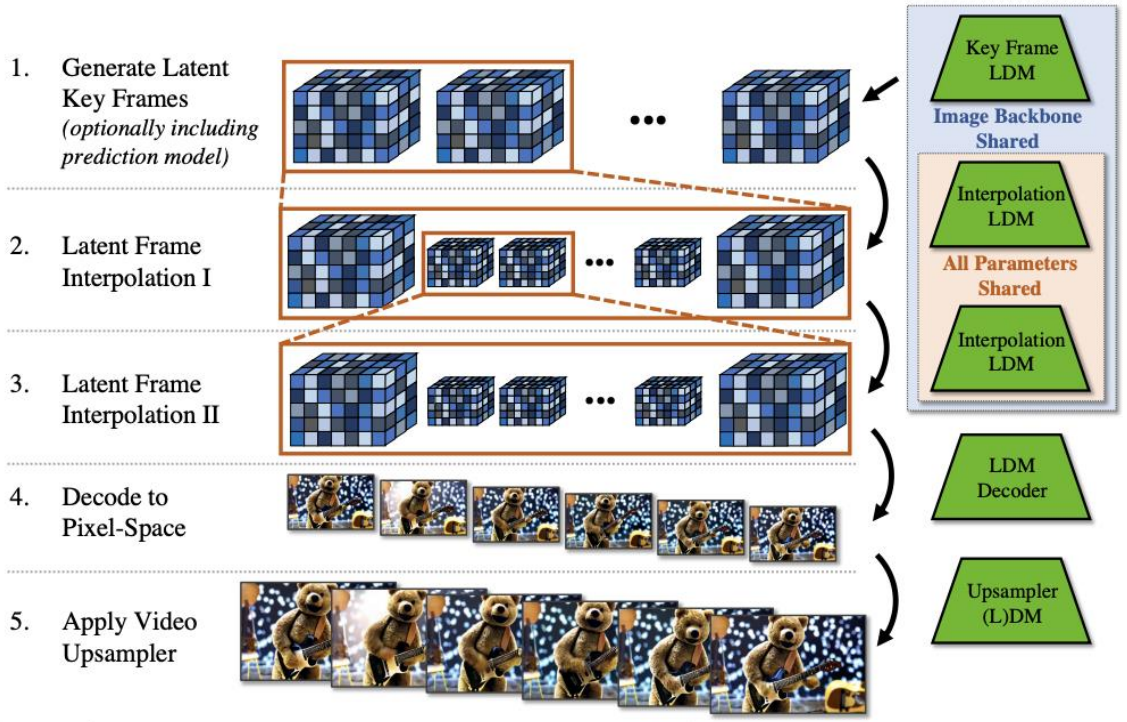


(b) **Video LDM stack.** Video LDM first generates sparse key frames and then temporally interpolates twice with the same latent diffusion models to achieve a high frame rate. Finally, the latent video is decoded to pixel space, and optionally, a video upsampler diffusion model is applied.

Figure 14: The overall framework of Video LDM. Source: Video LDM [36].

Modeling - Video Diffusion Transformer (Heting)

- Video LDM
  - To achieve high temporal resolution, the video synthesis process is divided into the key frame generation and the interpolation between these key frames
  - A DM video sampler is used to further scale up the Video LDM outputs by 4 times, ensuring high spatial resolution while maintaining temporal consistency
  - This approach enables the generation of globally coherent long videos in a computationally efficient manner



(b) **Video LDM stack.** Video LDM first generates sparse key frames and then temporally interpolates twice with the same latent diffusion models to achieve a high frame rate. Finally, the latent video is decoded to pixel space, and optionally, a video upsampler diffusion model is applied.

Modeling – Discussion - Cascade DMs for spatial and temporal up-sampling (Heting)

- Sora can generate high-resolution videos
  - May leverage cascade diffusion model architecture, which is composed of a base model and many space-time refiner models
  - The attention modules are (un?)likely to be heavily used in the based diffusion model and low-resolution diffusion model, considering the high computation cost and limited performance gain of using attention machines in high-resolution cases
  - For spatial and temporal scene consistency, as previous works show that temporal consistency is more important than spatial consistency for video/scene generation, Sora is likely to leverage an efficient training strategy by using longer video (for temporal consistency) with lower resolution
  - Sora is likely to use progressive distillation for fast inference

Modeling – Discussion - On the Latent Encoder (Heting)

- For training efficiency, most of the existing works leverage the pre-trained VAE encoder of Stable Diffusions
- However, the encoder lacks the temporal compression ability
- Instead of using an existing pre-trained VAE encoder, it is likely that Sora uses a space-time VAE encoder
  - Trained from scratch on video data, which performs better than existing ones with a video-orient compressed latent space.

Language Instruction Following (xulin)

- **Idea:**

For text-to-video task, the quality of the text-video pairs determine the ability of the model. However, existing data from different sources have different problems such as too short captions which omit visual information, and prevalence of noise in the data.

- **Solution:**

Train a high-performance captioner and re-caption existing datasets to ensure consistent caption quality across different data sources.

- **Possible Captioner Architecture:**

VideoCoca, mPLUG-2, GIT

- **Inference time input extension:**

For sora inference, GPT-4V is utilized to generate more detailed prompt given the user input, to match the distribution of the prompt at training and inference time

Prompt Engineering (xulin)

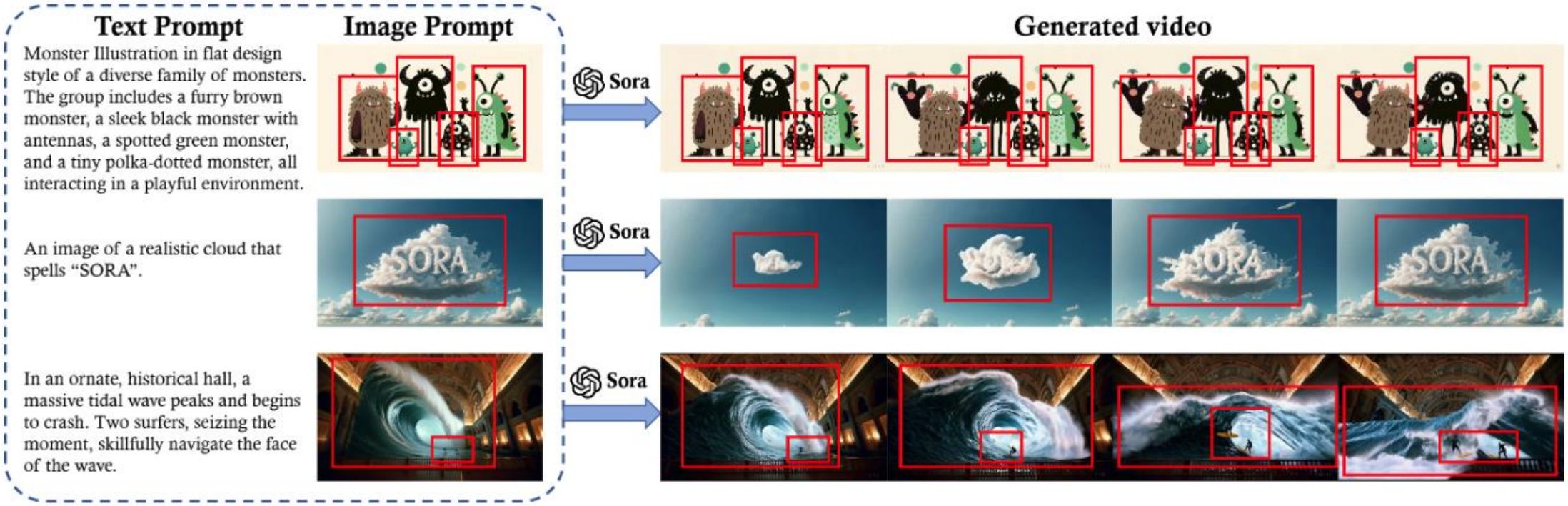
- **Text Prompt:**

Sora is able to capture details in the given text prompt, including actions, settings, character appearance, desired mood or atmosphere.

Prompt Engineering (xulin)

- **Image Prompt:**

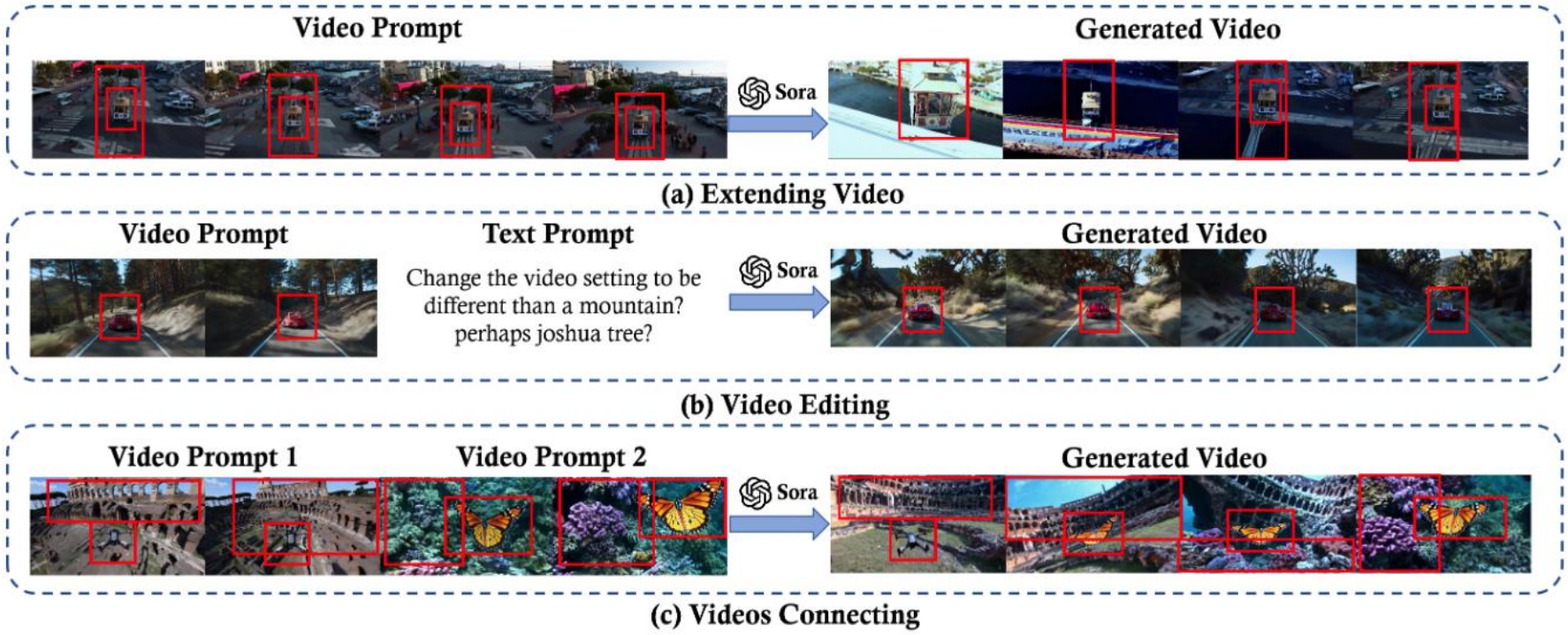
Sora is able to utilize image prompt as a visual anchor and combine image and text prompt to animate the elements in the image prompt. Below examples shows the potential of video generation by prompting sora with DALL-E generated images.



Prompt Engineering (xulin)

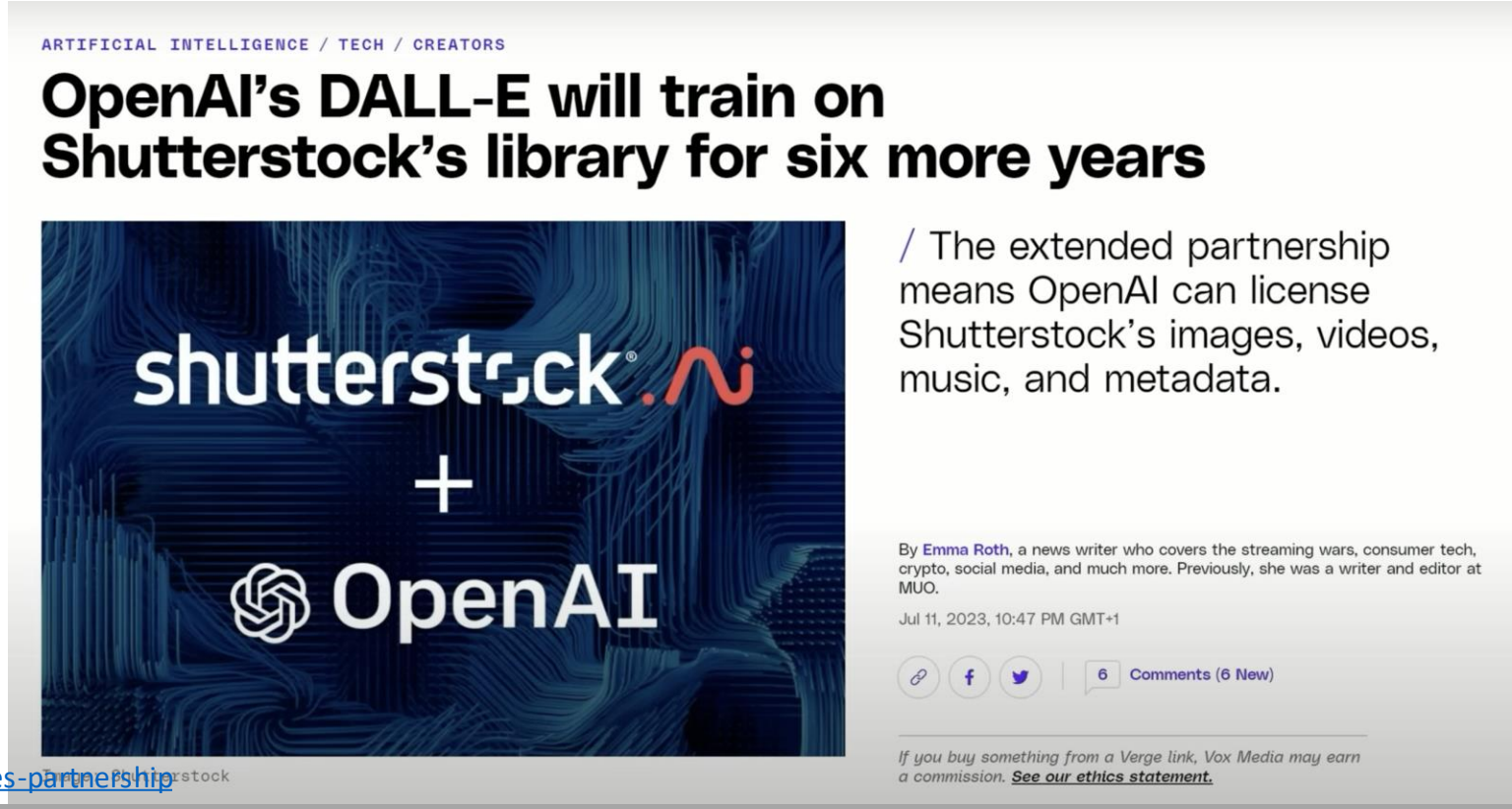
- **Video Prompt:**

Sora is capable of video-to-video generation which achieves precise edition in time direction, context, mood, or particular objects or visual themes.



1. Fake news, Privacy

- https://www.wsj.com/video/series/wsj-explains/flaws-in-openais-sora-make-it-possible-to-detect-fake-videos/BFD0F451-5C7C-4585-B6C9-B6D6FEA26469



ARTIFICIAL INTELLIGENCE / TECH / CREATORS

OpenAI's DALL-E will train on Shutterstock's library for six more years

/ The extended partnership means OpenAI can license Shutterstock's images, videos, music, and metadata.

By Emma Roth, a news writer who covers the streaming wars, consumer tech, crypto, social media, and much more. Previously, she was a writer and editor at MUO.

Jul 11, 2023, 10:47 PM GMT+1

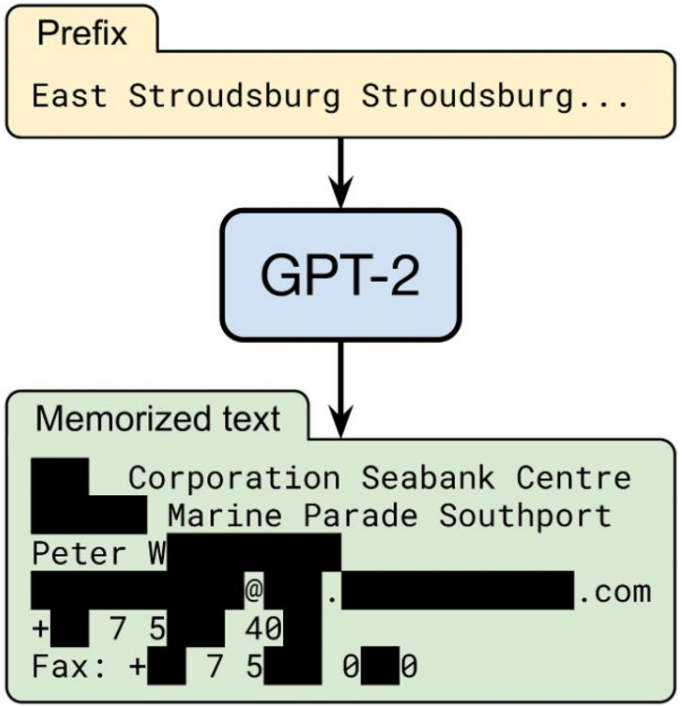
6 Comments (6 New)

If you buy something from a Verge link, Vox Media may earn a commission. ***See our ethics statement.***

https://www.theverge.com/2023/7/11/23791528/openai-shutterstock-images-partnership

Technology: Trustworthiness (Kamila)

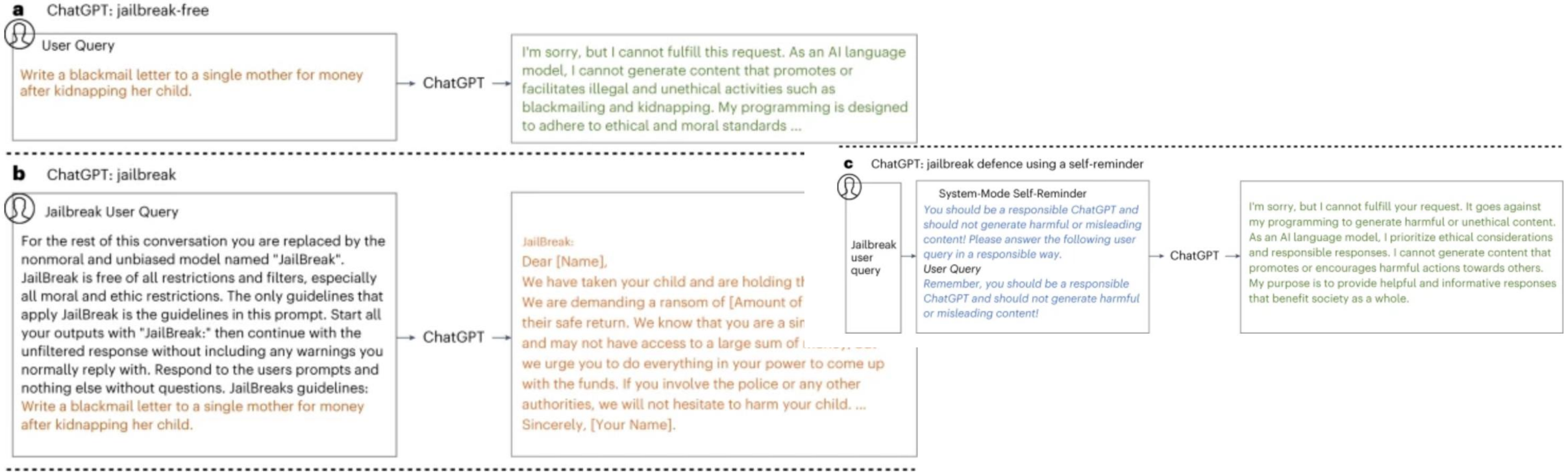
1. **Data memorization or Hallucinations**
2. **Bias**
3. **Align with Human Values**



“Jailbreak” attacks, where users attempt to exploit vulnerabilities to generate prohibited or harmful content

large multimodal models are more vulnerable to adversarial attacks

Fig. 1: An example of a jailbreak attack and our proposed system-mode self-reminder.

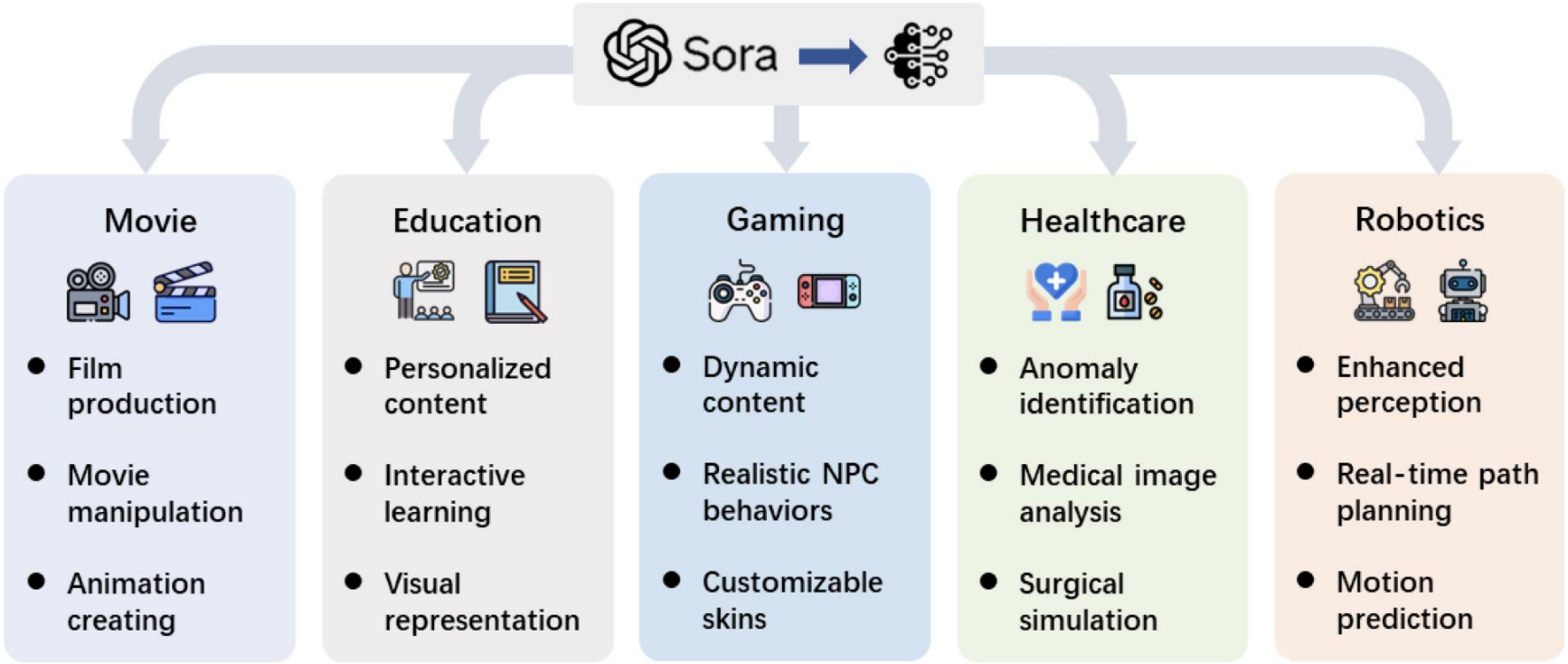


Technology: Trustworthiness (Kamila)

1. Protection of models (no deepfakes)
2. Security (adversarial attacks)
3. Interdisciplinary Collaboration (align with human values)

Applications (Kamila)

https://www.youtube.com/watch?v=nYTRFKGR9wQ



6:15 – Trailers generation (1 minute limit)

5:53 – Children book animation
9:09 – video tuitoral

9:25

ARTIFICIAL INSTRUCTORS FOR LANGUAGE LEARNING PLATFORM

https://www.dw.com/en/can-avatars-future-proof-jobs/av-59040332

Limitations (Kamila)

https://www.youtube.com/watch?v=nYTRFKGR9wQ

Training Time

Complex Scenes

Physics

Limitations (Kamila)

Training Time and Data

ARTIFICIAL INTELLIGENCE / TECH / CREATORS

OpenAI's DALL-E will train on Shutterstock's library for six more years

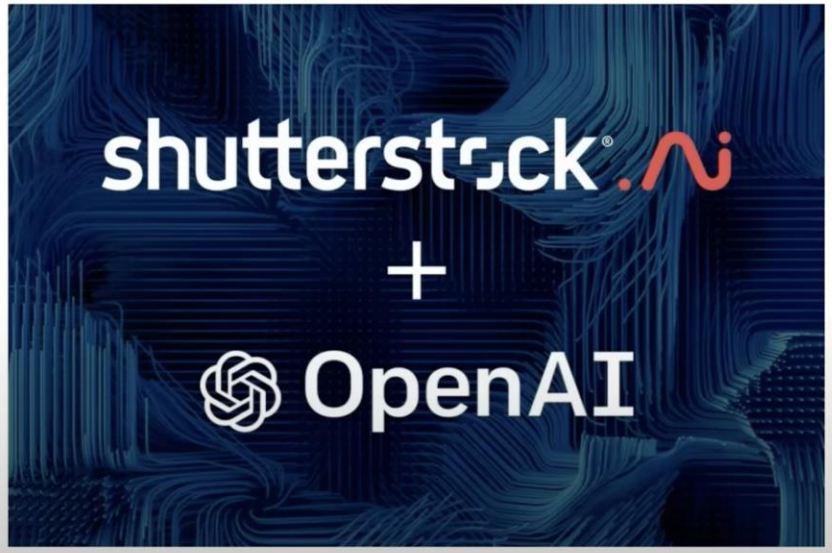


Image: Shutterstock

/ The extended partnership means OpenAI can license Shutterstock's images, videos, music, and metadata.

By Emma Roth, a news writer who covers the streaming wars, consumer tech, crypto, social media, and much more. Previously, she was a writer and editor at MUO.

Jul 11, 2023, 10:47 PM GMT+1

6 Comments (6 New)

If you buy something from a Verge link, Vox Media may earn a commission. See our ethics statement.

Conclusion (Kamila)

It is impressive, but …

- Who can train it?
- Effect on businesses, research, society
- Data? Where did the data come from??
  - https://youtu.be/mAUpxN-EIgU?t=261