OXEN.AI – Arxiv Dives
“Road to Sora” Paper Reading List

Greg Schoeninger
Mar 5. 2024

Suk-Hwan Lee

Artificial Intelligence

Creating the Future

Dong-A University

Division of Computer Engineering &
Artificial Intelligence

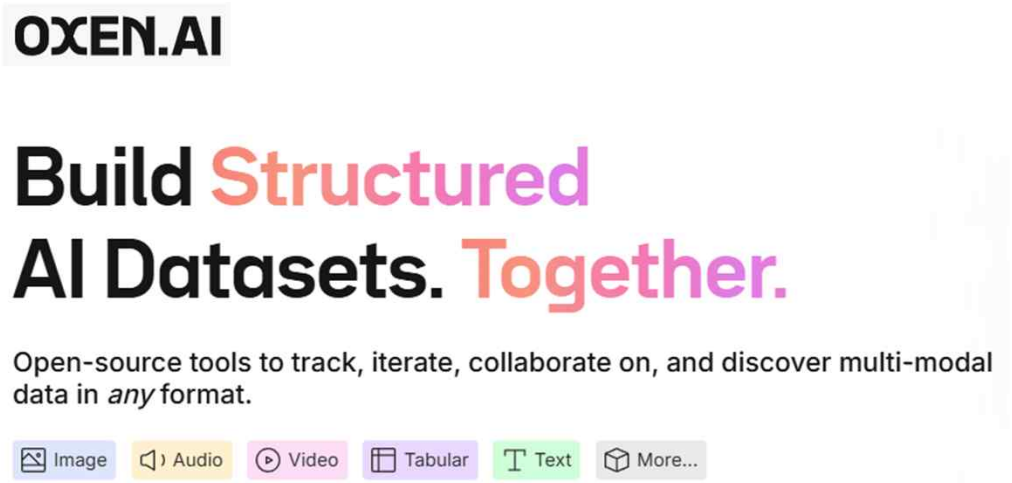
References

OXEN.AI - Arxiv Dives
"Road to Sora" Paper Reading List

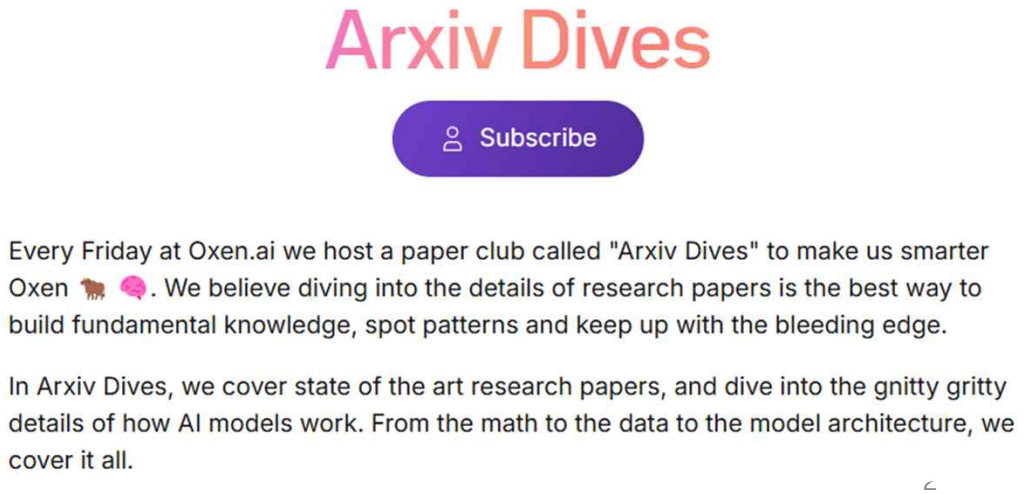
https://www.oxen.ai/blog/road-to-sora-reading-list

Road to Sora: OpenAI의 Sora를 이해하기 위한 선행 연구 소개 (feat. Oxen.AI) - 읽을거리&정보공유 - 파이토치 한국 사용자 모임 (pytorch.kr)

Since there has not been an official paper released yet for Sora, the goal is follow the bread crumbs from OpenAI's technical report on Sora.



https://www.oxen.ai/community/arxiv-dives



Road to Sora

What is Sora?

- Sora has taken the Generative AI space by storm with it's ability to **generate high fidelity videos from natural language prompts**.

☞ a turtle swimming in a coral reef

- OpenAI has not released an official research paper on the technical details

- Technical Report : OpenAI – Sora

https://openai.com/index/video-generation-models-as-world-simulators/

Sora Architecture Overview

- "Sora: A Review on Background, Technology, Limitations, and Opportunities of Large Vision Models" that gives a **high level diagram of a reverse engineered architecture.**

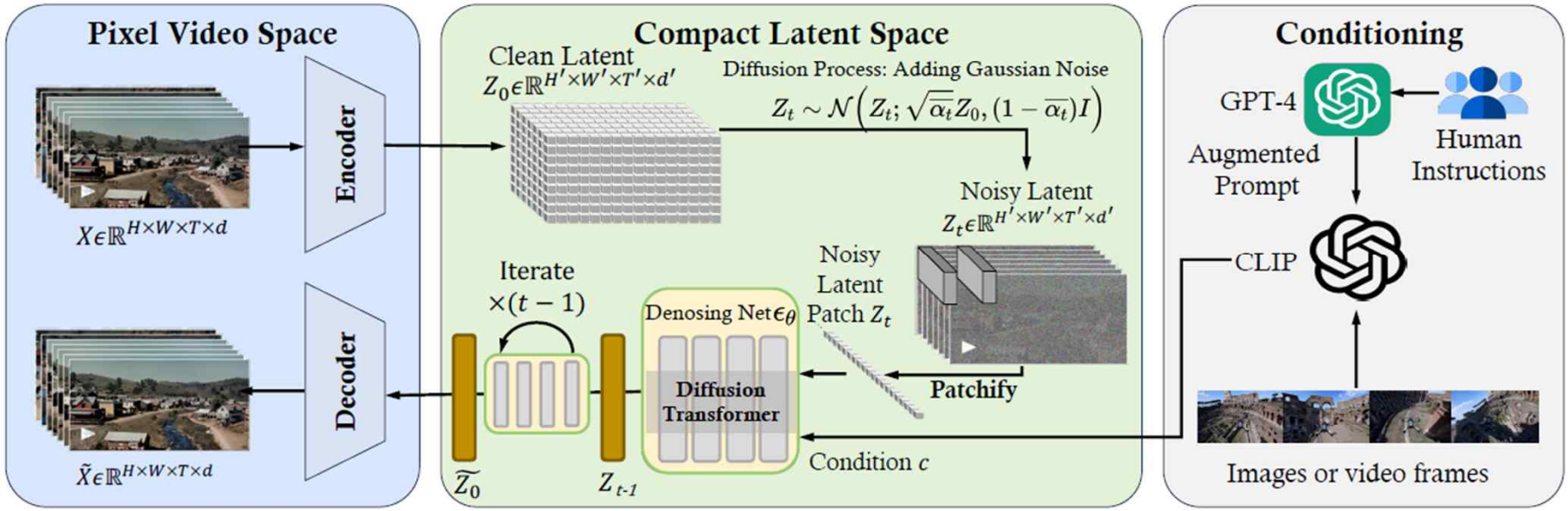


Figure 4: **Reverse Engineering**: Overview of Sora framework

- OpenAI states that **Sora is a "Diffusion Transformer"** which combines many of the concepts listed in the papers above, but **applied to latent spacetime patches generated from video.**

Sora Architecture Overview

- This is a combination of **the style of patches used in the Vision Transformer (ViT)** paper, with **latent spaces similar to the Latent Diffusion** Paper, but combined in the style of the Diffusion Transformer.
- They not only have **patches in width and height of the image** but extend it to the **time dimension** of video.

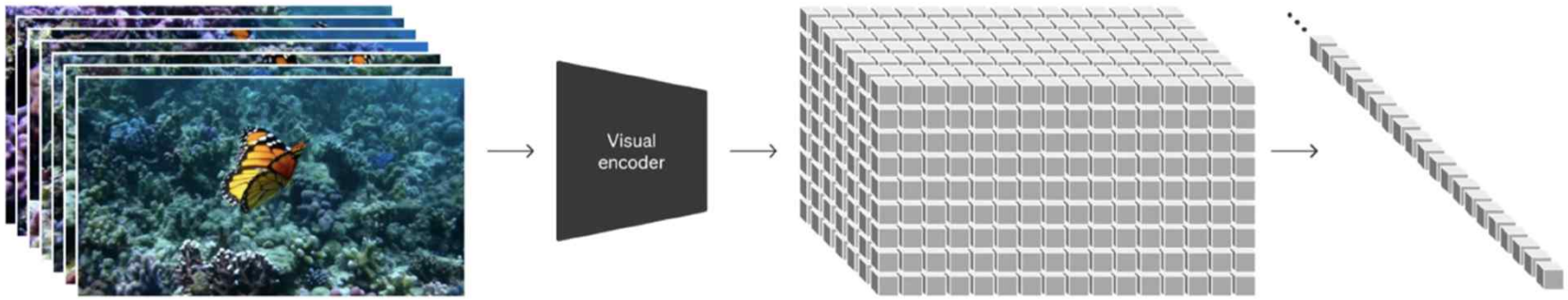


Figure 7: At a high level, Sora turns videos into patches by first compressing videos into a lower dimensional latent space, and subsequently decomposing the representation into spacetime patches. Source: Sora's technical report [3].

Sora Architecture Overview

- It's hard to say how exactly they collected the training data for all of this, but it seems like a combination of the techniques in the Dalle-3 as well as using GPT-4 to elaborate on textual descriptions of images, that they then turn into videos.
- **Training data** is likely the **main secret sauce** here, hence has the least level of detail in the technical report.

Use Cases

- There are many interesting use cases and applications for **video generation technologies** like **Sora**.

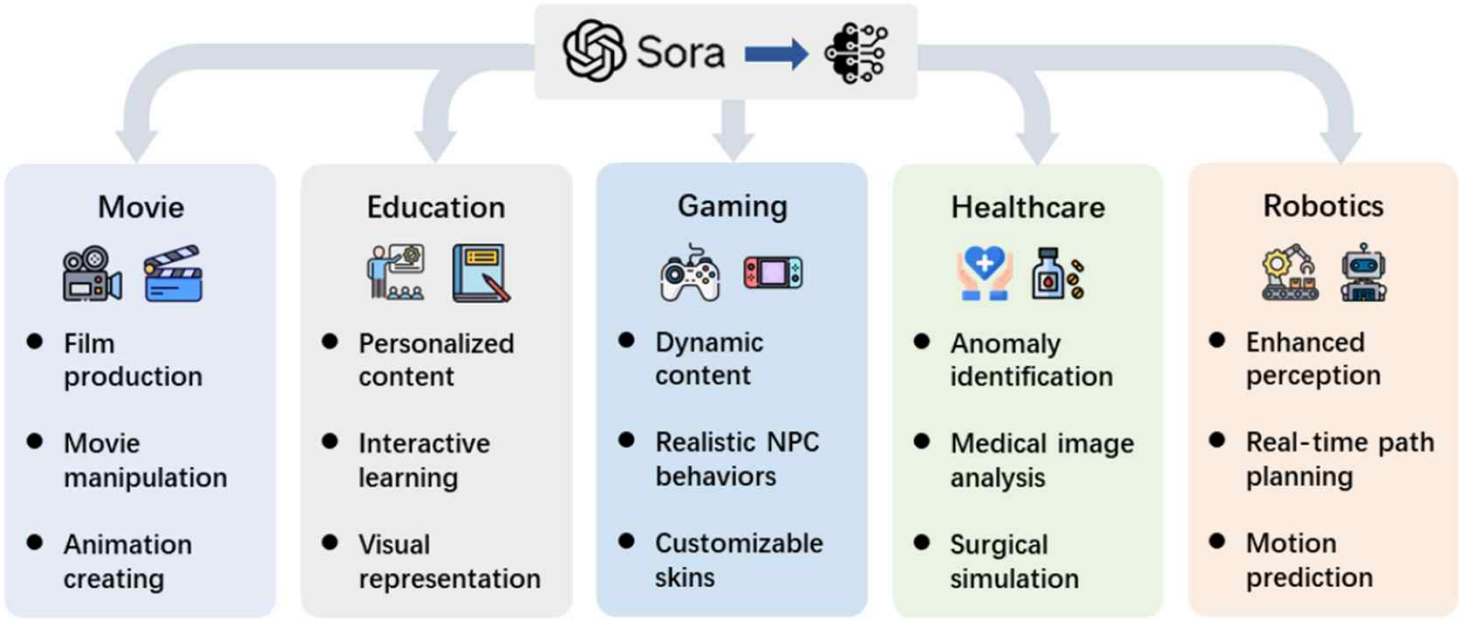


Figure 18: Applications of Sora.

- (3) The Need for Interdisciplinary Collaboration: Ensuring the safety of models is not just a technical issue but also requires cross-disciplinary cooperation. To address these challenges, experts from various fields such as law [142] and psychology [143] need to work together to develop appropriate norms (e.g., what's the safety and what's unsafe?), policies, and technological solutions. The need for interdisciplinary collaboration significantly increases the complexity of solving these issues.

➢ **Oxen.ai** : We are building open source tools to help you collaborate on and evaluate data the comes in and out of machine learning models.

Paper Reading List

- Selected what are the most impactful and interesting ones to read

1) Background Papers

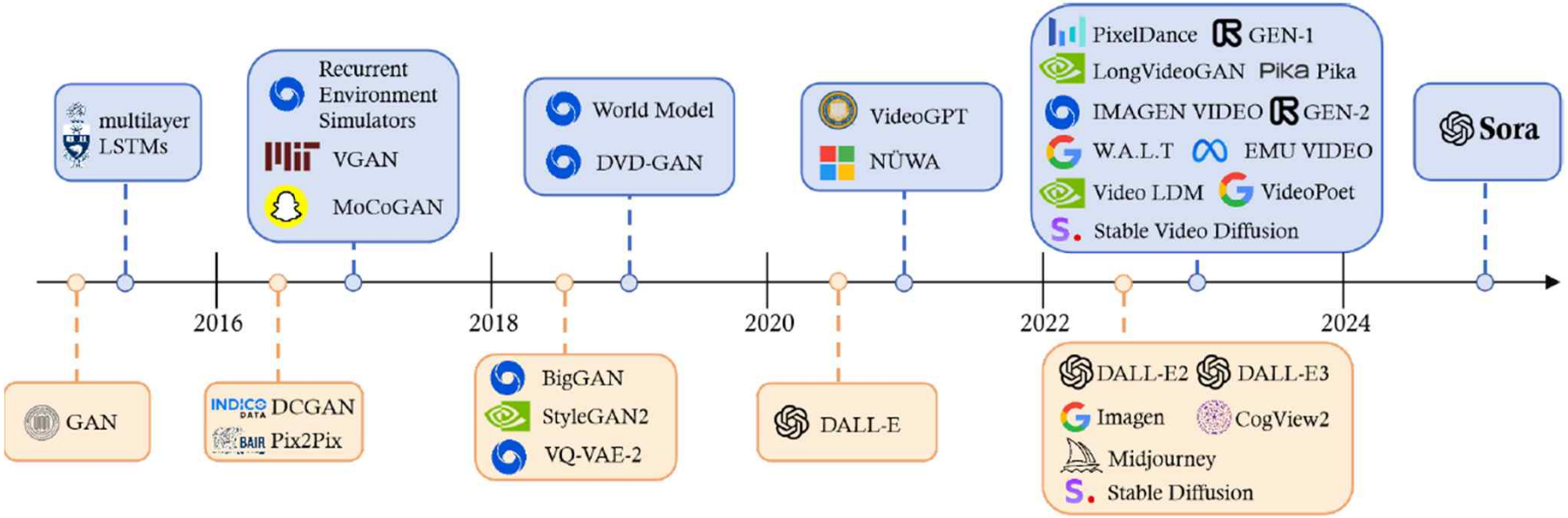


Figure 3: History of Generative AI in Vision Domain.

Paper Reading List

➢ U-Net

- “U-Net: Convolutional Networks for Biomedical Image Segmentation”
- Most notably is the backbone many diffusion models such as Stable Diffusion to **facilitate learning to predict and mitigate noise at each step**

➢ Language Transformers

- Transformer : "Attention Is All You Need"
- Transformers are now the backbone of many LLM applications such as ChatGPT.
- Transformers end up being extensible to many modalities and are used as a component of the Sora architecture.

➢ Vision Transformer (ViT)

- "An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale" Proving that they can outperform ResNets and other Convolutional Neural Networks if you train them **on large enough datasets**.
- This takes the architecture from the "Attention Is All You Need" paper and makes it work for computer vision tasks. Instead of the inputs being text tokens, ViT uses 16x16 image patches as input.

Paper Reading List

➢ Latent Diffusion Models (LDM)

- "High-Resolution Image Synthesis with Latent Diffusion Models" is the technique behind many image generation models such as **Stable Diffusion**.
- They show how you can reformulate the image generation as **a sequence of denoising auto-encoders** from **a latent representation**.
- They use the **U-Net** architecture as **the backbone of the generative process**.
- These models can generate photo-realistic images given any text input.

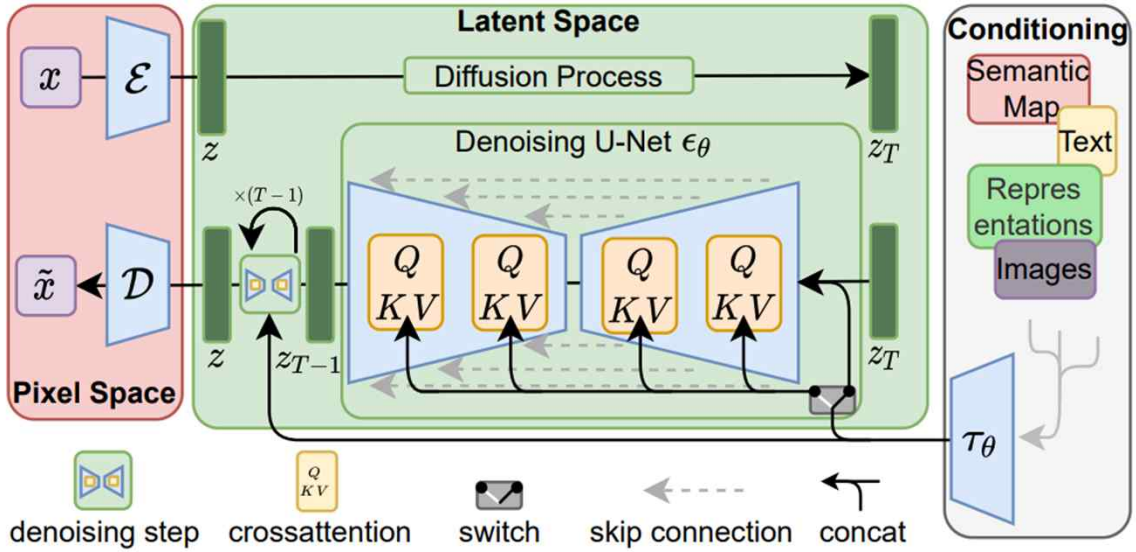


Figure 3. We condition LDMs either via concatenation or by a more general cross-attention mechanism

Paper Reading List

➢ CLIP

- "[Learning Transferable Visual Models From Natural Language Supervision]" often referred to as **Contrastive Language-Image Pre-training (CLIP)** is a technique for **embedding text data and image data into the same latent space as each other**.
- This technique helps connect the language understanding half of generative models to the visual understanding half by making sure that the **cosine similarity between the text and image representations** are high between text and image pairs.

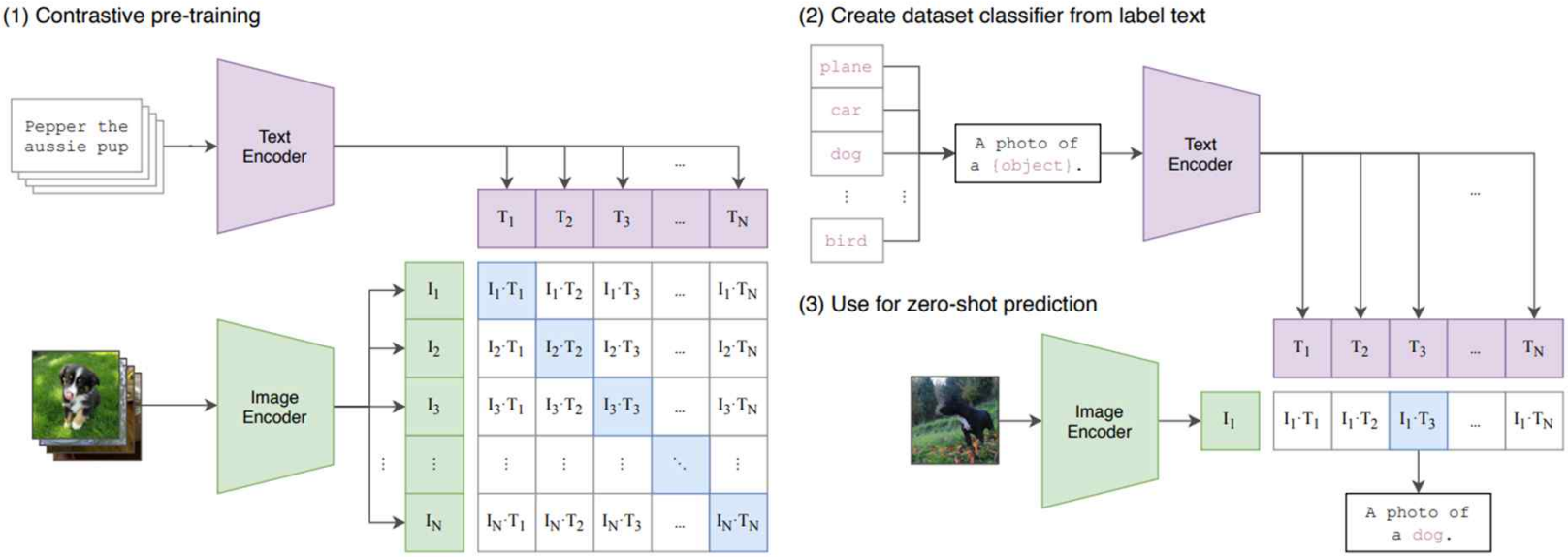


Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Paper Reading List

➢ VQ-VAE

- "Neural Discrete Representation Learning"
- According to the SORA technical report, they **reduce the dimensionality of the raw video** with a **Vector Quantized Variational Auto Encoder (VQ-VAE)**.
- **VAEs** have been shown to be a **powerful unsupervised pre-training** method to **learn latent representations**.

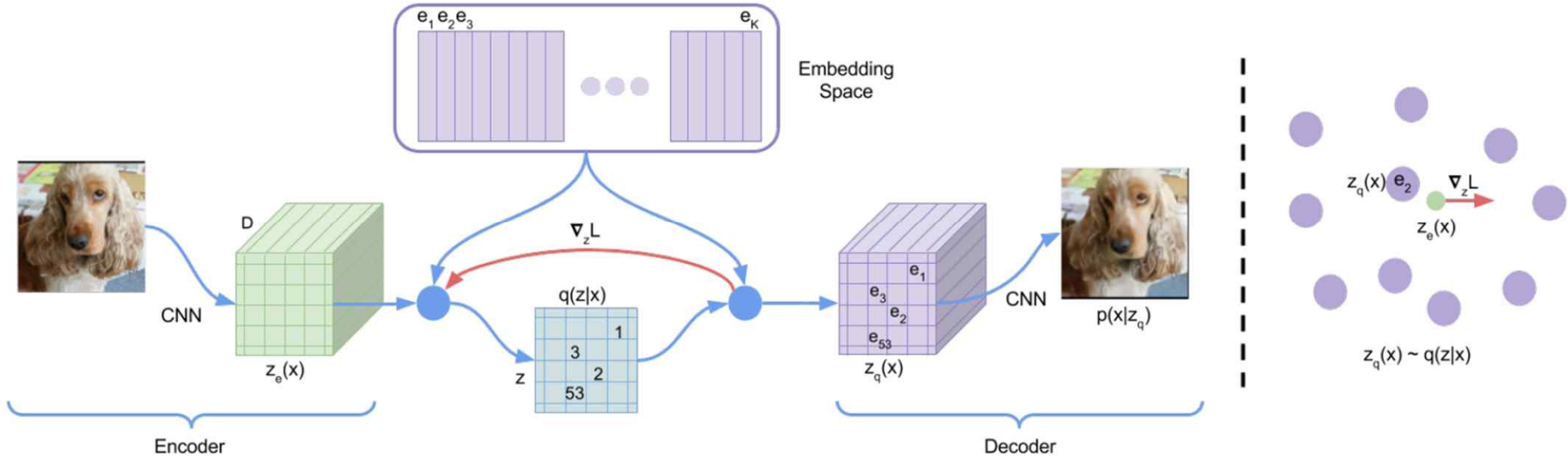


Figure 1: Left: A figure describing the VQ-VAE. Right: Visualisation of the embedding space. The output of the encoder $z(x)$ is mapped to the nearest point e_2. The gradient $\nabla_z L$ (in red) will push the encoder to change its output, which could alter the configuration in the next forward pass.

Paper Reading List

➢ **Patch n' Pack: NaViT**

- "Patch n' Pack: NaViT, a Vision Transformer for any Aspect Ratio and Resolution"
- The Sora technical report talks about **how they take in videos of any aspect ratio**, and **how this allows them to train on a much larger set of data**.
- **The more data they can feed the model without having to crop it, the better results they get**. This paper uses the same technique but for images, and Sora extends it for video.

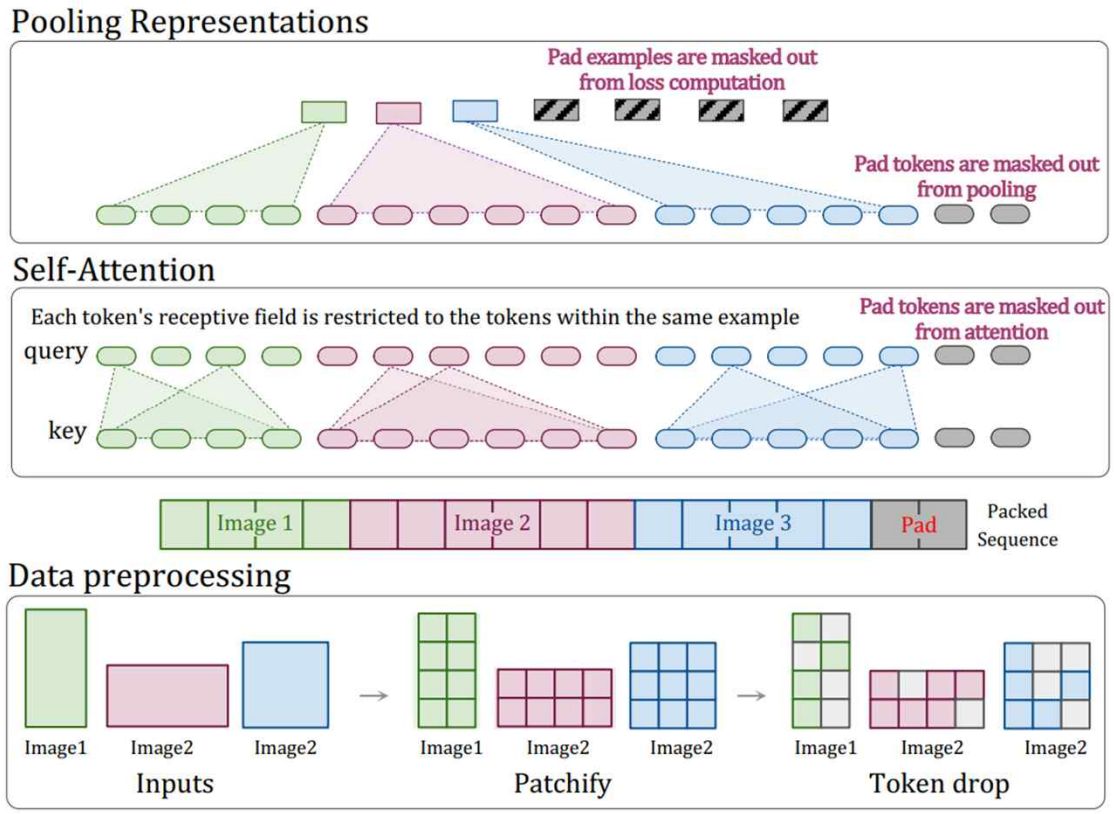


Figure 2: Example **packing** enables variable resolution images with preserved aspect ratio, reducing training time, improving performance and increasing flexibility. We show here the aspects of the data preprocessing and modelling that need to be modified to support **Patch n' Pack**. The position-wise operations in the network, such as MLPs, residual connections, and layer normalisations, do not need to be altered.

Paper Reading List

➢ Diffusion Transformer (DiT)

- "Scalable Diffusion Models with Transformers"
- OpenAI states that Sora is a "Diffusion Transformer" which combines many of the concepts listed in the papers above.
- The DiT **replaces** the **U-Net backbone in latent diffusion models**, with **a transformer that operates on latent patches**.
- This paper generates state of the art high quality images by using a more efficient model than a U-Net, therefore **increasing the amount of data and compute that could be used to train these models**.

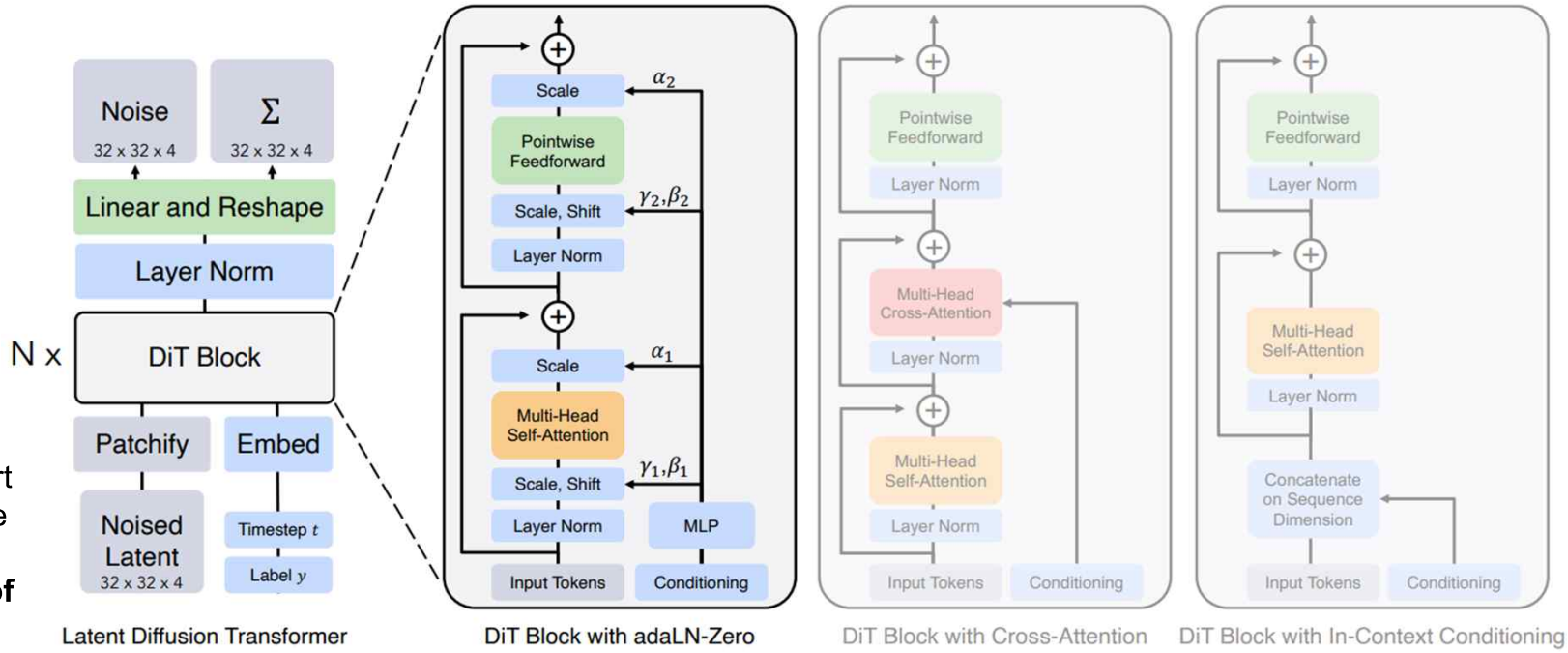


Figure 3. The Diffusion Transformer (DiT) architecture. Left: We train conditional latent DiT models. The input latent is decomposed into patches and processed by several DiT blocks. Right: Details of our DiT blocks. We experiment with variants of standard transformer blocks that incorporate conditioning via adaptive layer norm, cross-attention and extra input tokens. Adaptive layer norm works best.

Paper Reading List

2) Video Generation Papers

➢ **ViViT: A Video Vision Transformer**

- "ViViT: A Video Vision Transformer"
- This paper goes into details about **how you can chop (분할) the video into "spatio-temporal tokens"** needed for video tasks.
- The paper **focuses on video classification**, but **the same tokenization can be applied to generating video**.

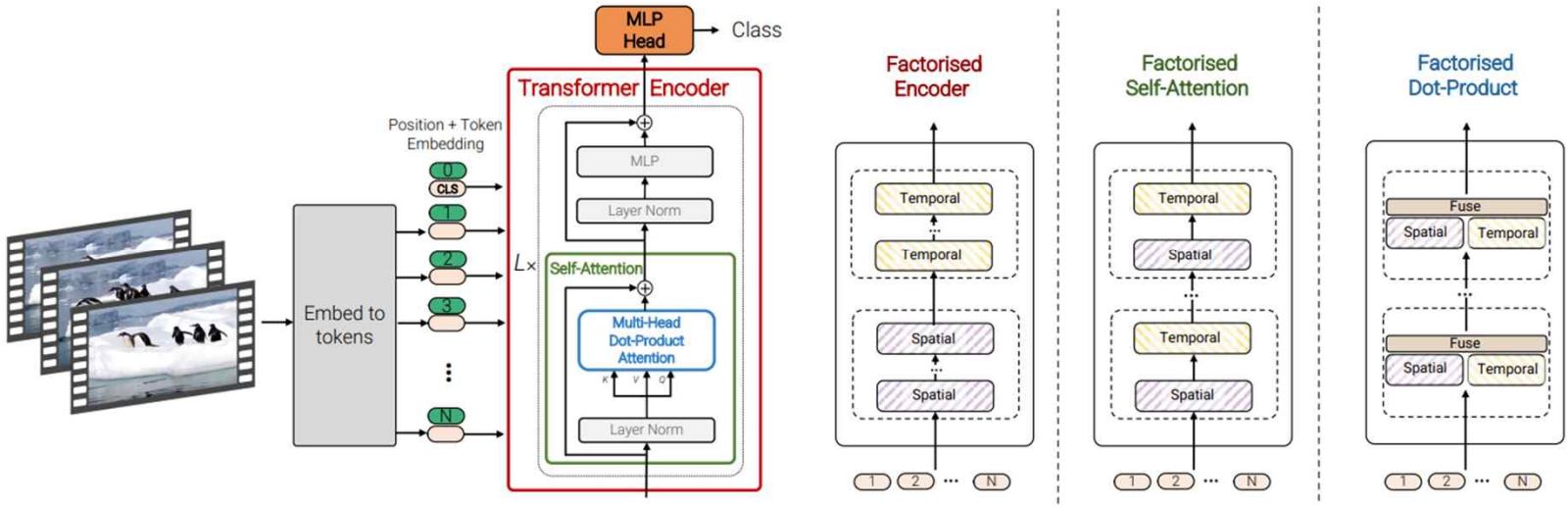


Figure 1: We propose a **pure-transformer architecture for video classification**, inspired by the recent success of such models for images [18]. To effectively process **a large number of spatio-temporal tokens**, we develop several **model variants which factorise different components of the transformer encoder over the spatial- and temporal-dimensions**. As shown on the right, these factorisations correspond to different attention patterns over space and time.

Paper Reading List

➢ Imagen Video

- "Imagen Video: High Definition Video Generation with Diffusion Models"
- Imagen is a **text-conditional video generation system** based on **a cascade of video diffusion models**.
- They use **convolutions in the temporal direction and super resolution** to generate high quality videos from text.

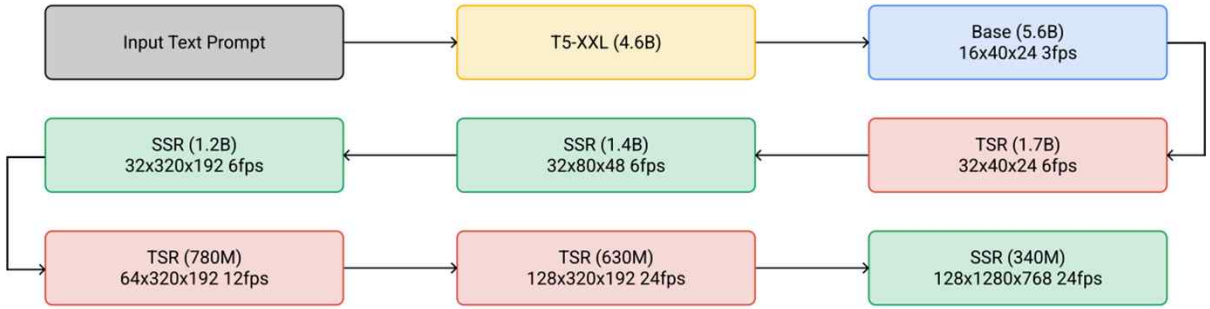


Figure 6: The cascaded sampling pipeline starting from a text prompt input to generating a 5.3-second, 1280×768 video at 24fps. "SSR" and "TSR" denote spatial and temporal super-resolution respectively, and videos are labeled as frames×width×height. In practice, the text embeddings are injected into all models, not just the base model.

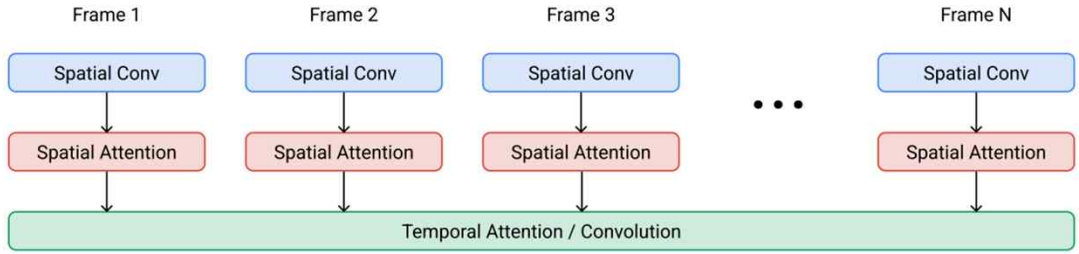


Figure 7: Video U-Net space-time separable block. Spatial operations are performed independently over frames with shared parameters, whereas the temporal operation mixes activations over frames. Our base model uses spatial convolutions, spatial self-attention and temporal self-attention. For memory efficiency, our spatial and temporal super-resolution models use temporal convolutions instead of attention, and our models at the highest spatial resolution do not have spatial attention

Paper Reading List

➢ Align your Latents (Video LDM)

- "Align your Latents: High-Resolution Video Synthesis with Latent Diffusion Models"
- This paper takes the latent diffusion models from the image generation papers above and introduces **a temporal dimension to the latent space**.
- They apply **some interesting techniques** in **the temporal dimension by aligning the latent spaces**, but does **not quite have the temporal consistency of Sora yet.**

Paper Reading List

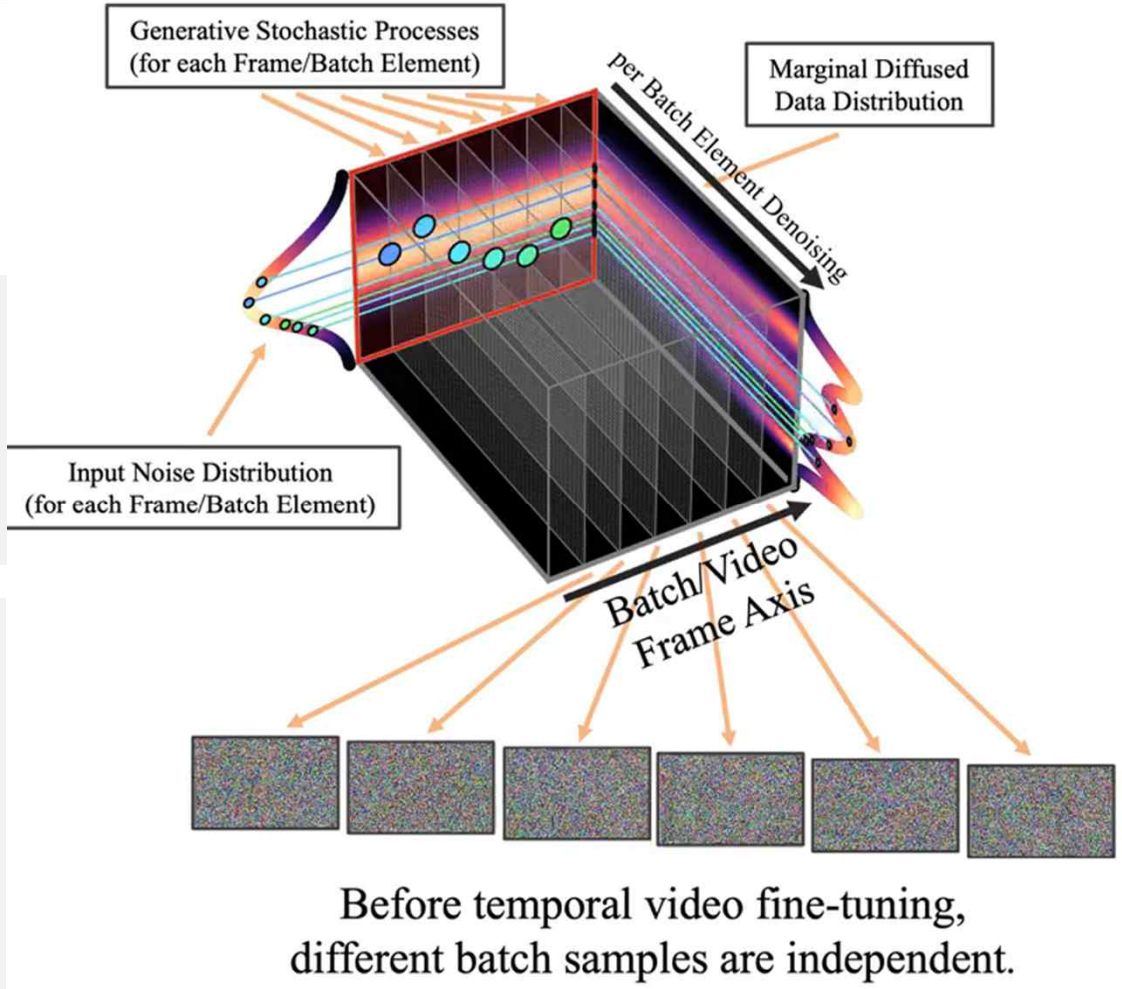
➢ **Align your Latents (Video LDM)**

https://research.nvidia.com/labs/toronto-ai/VideoLDM/

[Abstract]

- **Video Latent Diffusion Models (Video LDMs)** for computationally efficient high-resolution video generation.
- To alleviate the **intensive compute** and **memory demands** of high-resolution video synthesis, we leverage the **LDM** paradigm and extend it to video generation.

- Our Video LDMs **map videos into a compressed latent space** and **model sequences of latent variables corresponding to the video frames** (see animation above).
- We **initialize the models from image LDMs** and **insert temporal layers into the LDMs' denoising neural networks** to temporally model encoded video frame sequences. The temporal layers are based on **temporal attention** as well as **3D convolutions**.
- We also fine-tune the model's decoder for video generation (see figure below).



Before temporal video fine-tuning,
different batch samples are independent.

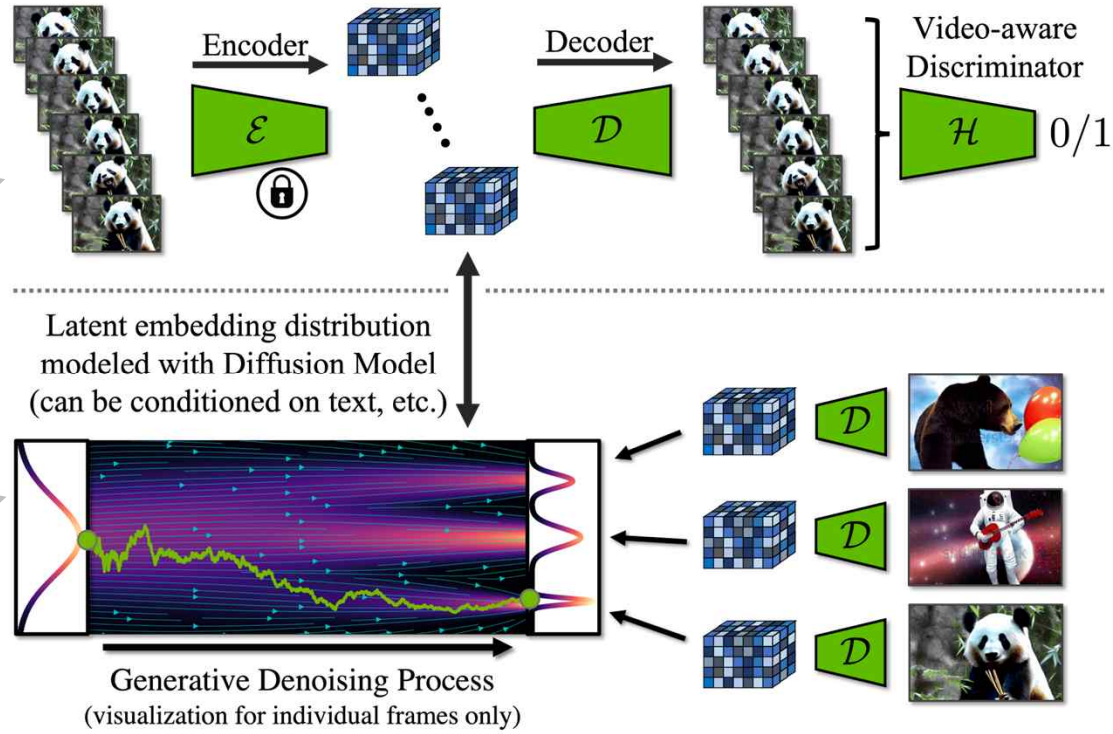
Paper Reading List

➢ **Align your Latents (Video LDM)**

- During temporal decoder fine-tuning, we process video sequences with a frozen per-frame encoder and enforce temporally coherent reconstructions across frames.
- We additionally employ a video-aware discriminator.

- In LDMs, a diffusion model is trained in latent space. It synthesizes latent features, which are then transformed through the decoder into images.
- Note that in practice we model entire videos and video fine-tune the latent diffusion model to generate temporally consistent frame sequences.

Latent diffusion model framework and video fine-tuning of decoder.

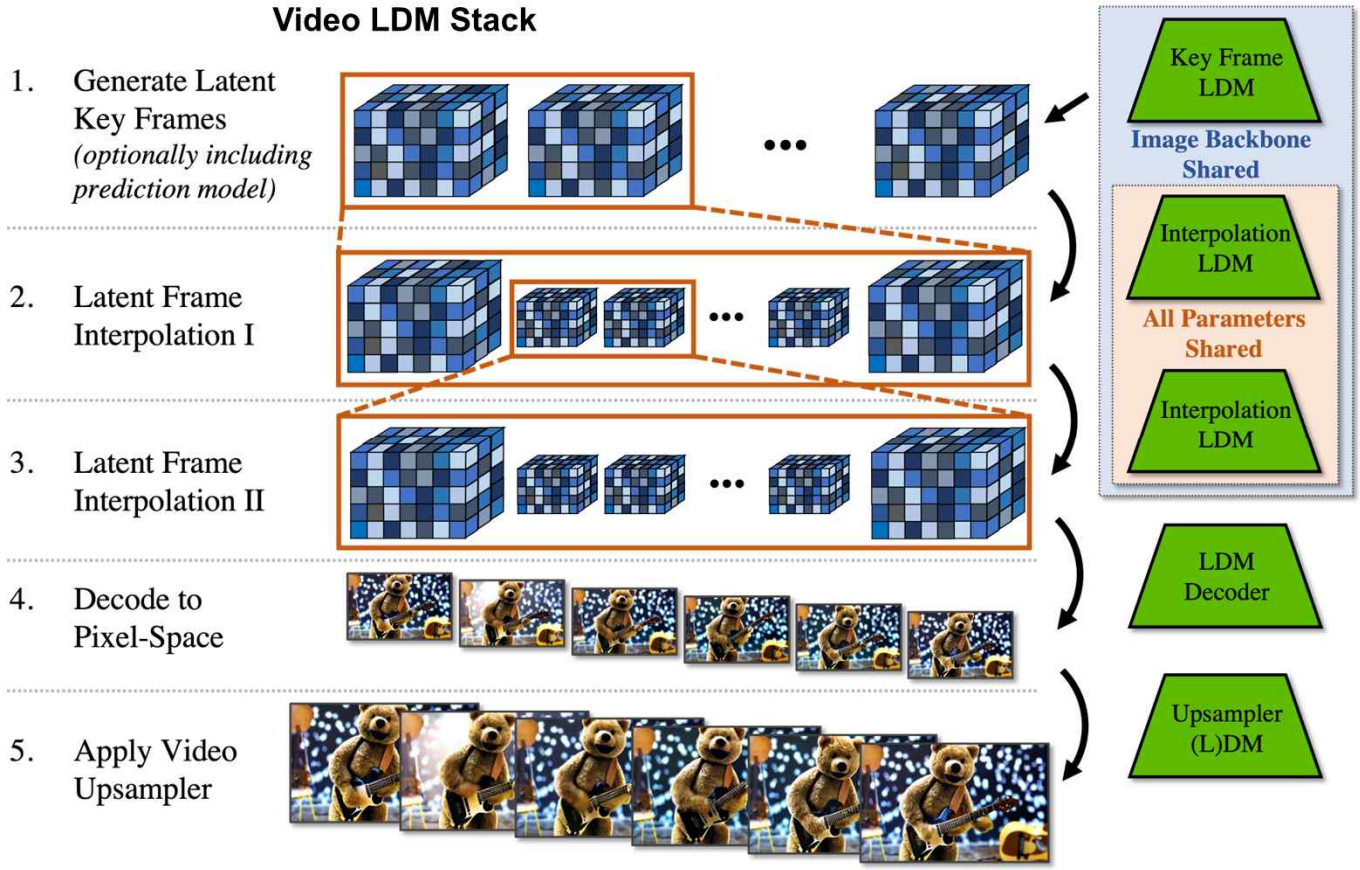


Paper Reading List

➢ **Align your Latents (Video LDM)**

- 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 **use 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.

- To achieve high-resolution generation, we further leverage spatial diffusion model upsamplers and temporally align them for video upsampling.

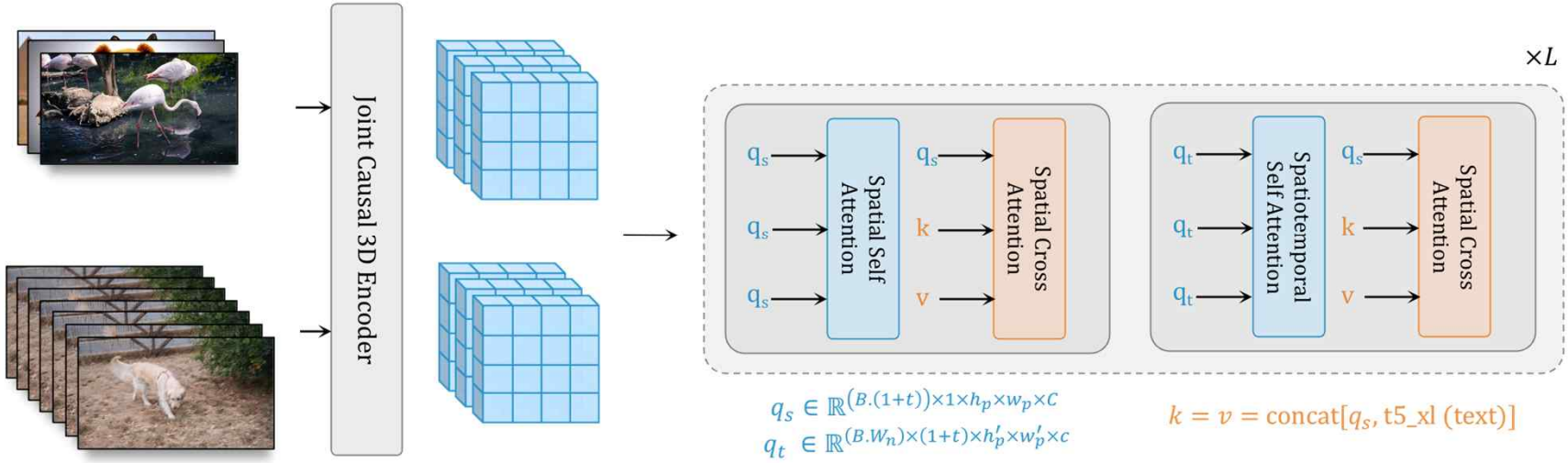


Paper Reading List

➢ Photorealistic video generation with diffusion models

- "Photorealistic video generation with diffusion models"
- They introduce **W.A.L.T**, **a transformer-based approach** for **photorealistic video generation via diffusion modeling**.
- This feels like the closest technique to Sora and was released in December of 2023 by the teams at Google, Stanford and Georgia Tech.

- **W.A.L.T** (Window Attention Latent Transformer) : A transformer-based method for latent video diffusion models (LVDMs).



W.A.L.T: We **encode** images and videos into a **shared latent space**. The transformer backbone processes these latents with blocks having **two layers of window-restricted attention**: **spatial layers** capture spatial relations in both images and video, while **spatiotemporal layers** model temporal dynamics in videos and passthrough images via identity attention mask. **Text conditioning** is done via **spatial cross-attention**.

Paper Reading List

3) Vision-Language Understanding

- In order to **Generate Videos from text prompts**, they **need to collect a large dataset**.
- It is not feasible to have humans label that many videos, so it seems they use some **synthetic data techniques** similar to those described in the **DALL·E 3** paper.

➢ **DALL·E 3**

- https://openai.com/index/dall-e-3/
- **Training text-to-video generation systems** requires **a large amount of videos with corresponding text captions**.
- They apply the **re-captioning technique** introduced in **DALL·E 3** to **videos**.
- Similar to DALL·E 3, they also **leverage GPT** to **turn short user prompts into longer detailed captions** that are sent to the video model.

Paper Reading List

➢ LLava

- "Visual Instruction Tuning"
- In order for the model to be able to follow **user instructions**, they likely did some **instruction fine-tuning** similar to the **LLava** paper.
- This paper also shows **some synthetic data techniques to create a large instruction dataset** that could be interesting in combination with the Dalle methods above.

The primary goal is to effectively leverage the capabilities of both the pre-trained LLM and visual model. The network archtecture is illustrated in Figure 1. We choose Vicuna [9] as our LLM $f_\phi(\cdot)$ parameterized by ϕ, as it has the best instruction following capabilities in language tasks among publicly available checkpoints [48, 9, 38].

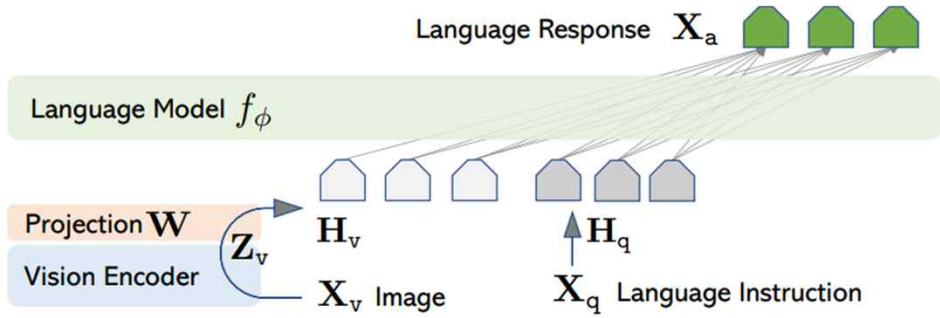


Figure 1: LLaVA network architecture.

For an input image $\mathbf{X}_v$, we consider the pre-trained CLIP visual encoder ViT-L/14 [40], which provides the visual feature $\mathbf{Z}_v = g(\mathbf{X}_v)$. The grid features before and after the last Transformer layer are considered in our experiments. We consider a simple linear layer to connect image features into the word embedding space. Specifically, we apply a trainable projection matrix $\mathbf{W}$ to convert $\mathbf{Z}_v$ into language embedding tokens $\mathbf{H}_v$, which have the same dimensionality as the word embedding space in the language model:

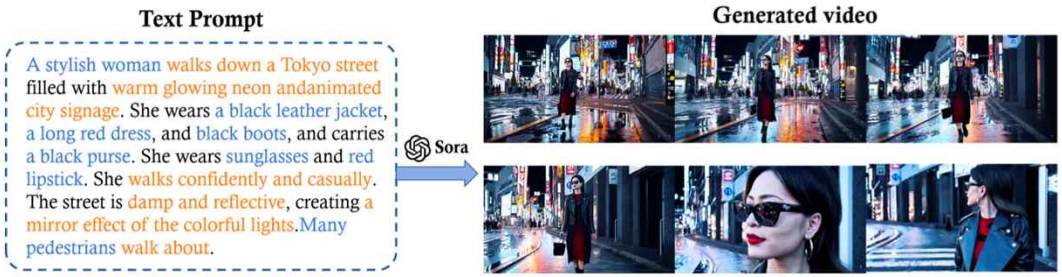
$$\mathbf{H}_v = \mathbf{W} \cdot \mathbf{Z}_v, \text{ with } \mathbf{Z}_v = g(\mathbf{X}_v) \quad (1)$$

Thus, we have a sequence of visual tokens $\mathbf{H}_v$. Note that our simple projection scheme is lightweight, which allows us to iterate data centric experiments quickly. More sophisticated schemes to connect the image and language representations can also be considered, such as gated cross-attention in Flamingo [2] and Q-former in BLIP-2 [28]. We leave exploring possibly more effective and sophisticated architecture designs for LLaVA as future work.

Paper Reading List

➢ Make-A-Video & Tune-A-Video

- "Make-A-Video: Text-to-Video Generation without Text-Video Data"
- "Tune-A-Video: One-Shot Tuning of Image Diffusion Models for Text-to-Video Generation"
- Papers like **Make-A-Video** and **Tune-A-Video** have shown how **prompt engineering leverages model's natural language understanding ability to decode complex instructions** and **render them into cohesive, lively, and high-quality video narratives**.
- For example: taking a simple user prompt and extending it with adjectives and verbs to more fully flush out the scene.



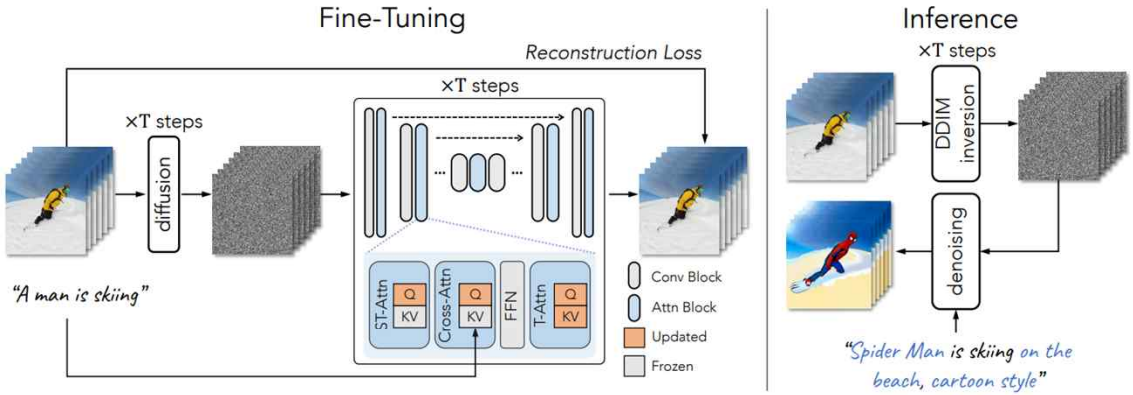


Figure 4: Pipeline of **Tune-A-Video**: Given a text-video pair (e.g., "a man is skiing") as input, our method leverages the pretrained T2I diffusion models for T2V generation. During fine-tuning, we update the projection matrices in attention blocks using the standard diffusion training loss. During inference, we sample a novel video from the latent noise inverted from the input video, guided by an edited prompt (e.g., "Spider Man is surfing on the beach, cartoon style").