NUS
National University
of Singapore

OPEN-SORA

Democratizing Efficient Video Production for All

GitHub: *https://github.com/hpcaitech/Open-Sora*

CONTENTS

Introduction of OpenAI’s Sora

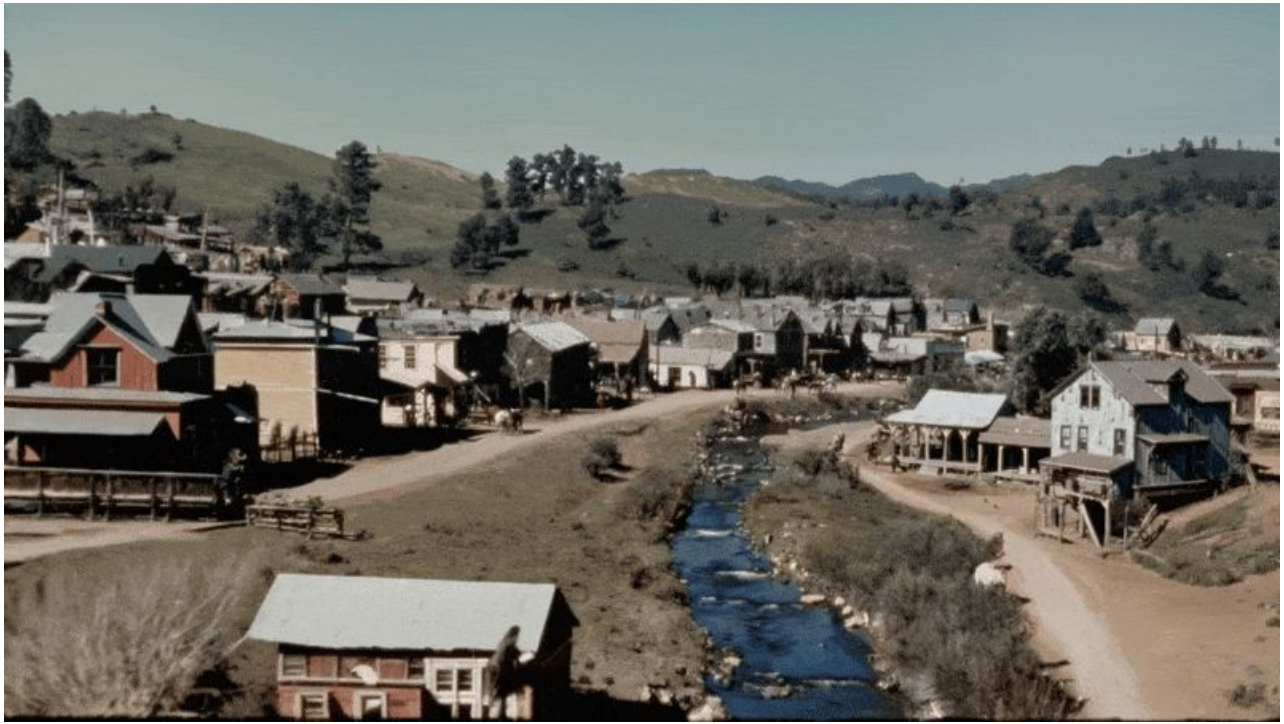
Introduction of OpenAI's Sora

Impact and Performance

Outperformed Pika, Runway, Stable Video considerably

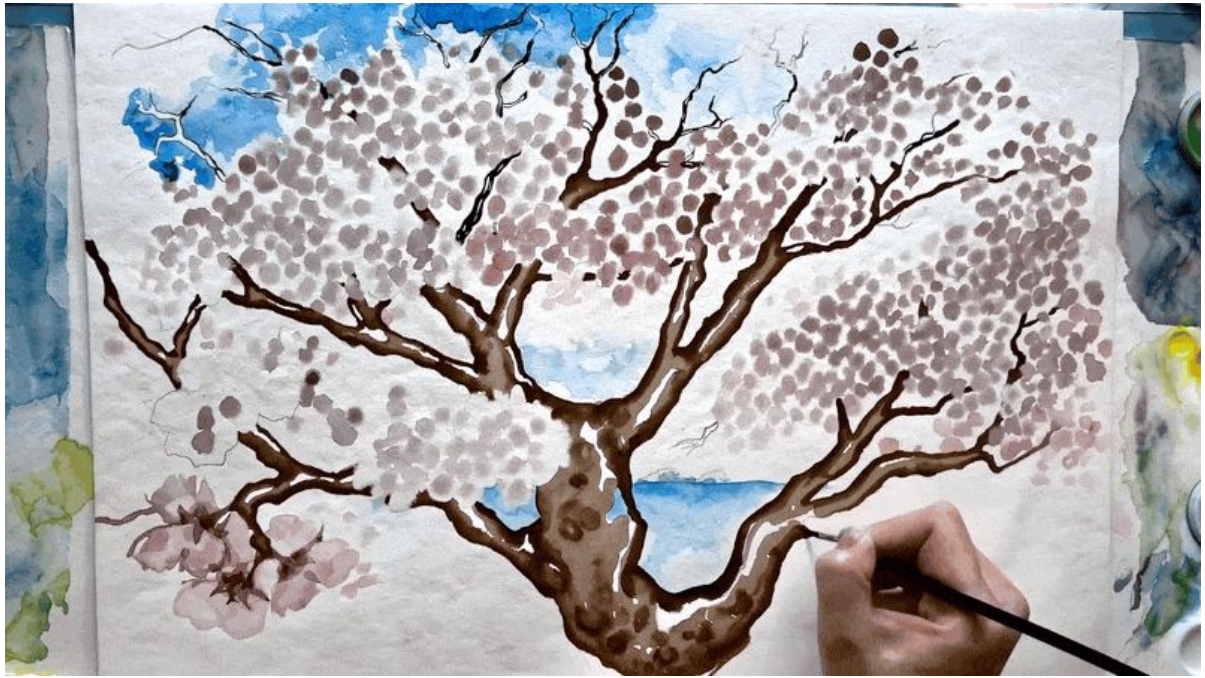
source: https://twitter.com/DailyUpdatesNet/status/1758646902751670355
https://twitter.com/gabor/status/1758282791547232482

Impact and Performance



Generated by OpenAI's Sora

Impact and Performance



Generated by OpenAI's Sora

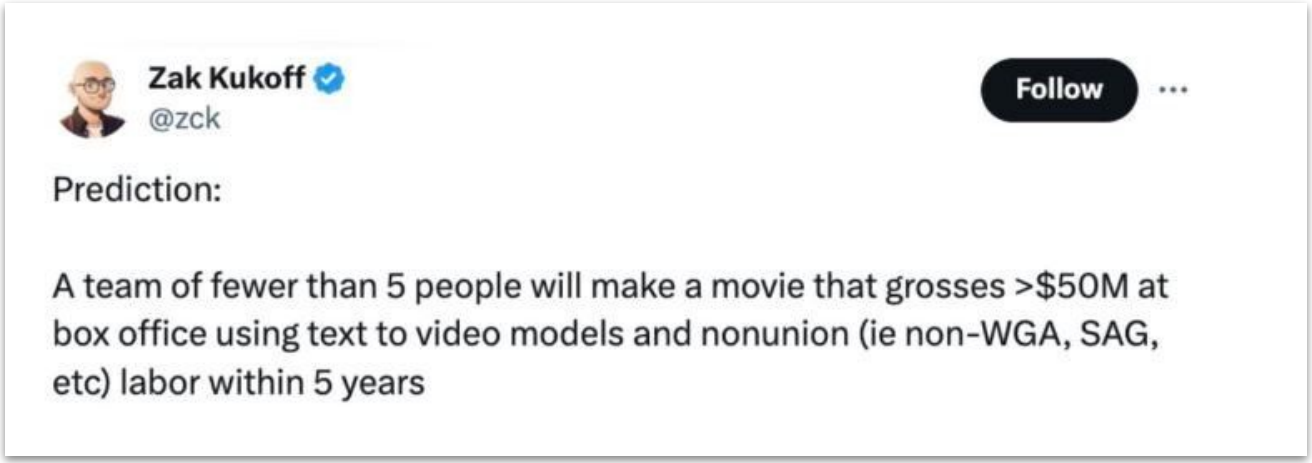
Impact and Performance

Generated by OpenAI's Sora

Applications and Use Cases

- Gaming and Virtual Reality
- Art and Creative Exploration
- Media Production
- Advertising and Marketing
- Education and Training

Applications and Use Cases



Overview of Open-Sora

Open-Sora: the First Open Source Sora-like Video Generation Model

Bringing OpenAI's Sora model to the community with low-cost, fully open-source replication:

- **Model Architecture**
- **Trained Model Checkpoints**
- **Training Process Details**
- **Data Preprocessing**
- **Video Demonstration and Tutorial**

Practical Principles for Large Model Training

What you need for training AI models?

- Core (by priority)
  - Data
  - Computation & System
  - Algorithm (Architecture)
- Other
  - Evaluation
  - Application
  - Deployment
  - ……

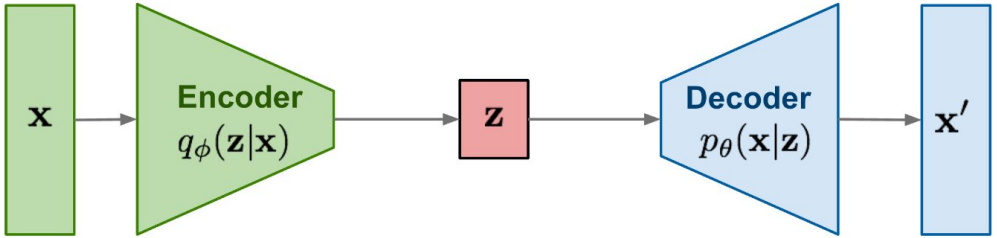
What you need for training AI models?

- Data
- Computation & System Architecture
- Model Architecture

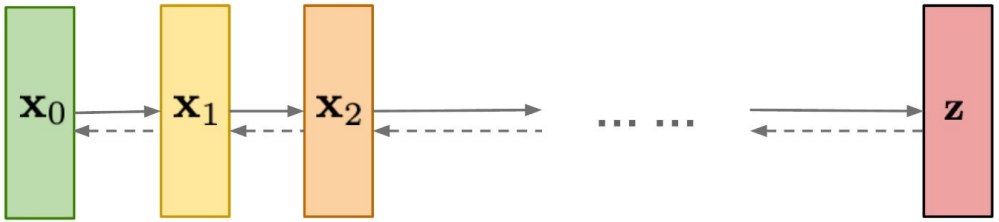
Backbone models

VAE and Diffusion Models

VAE: maximize variational lower bound



Diffusion models: Gradually add Gaussian noise and then reverse



source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/#forward-diffusion-process

Open-Sora Model Architecture

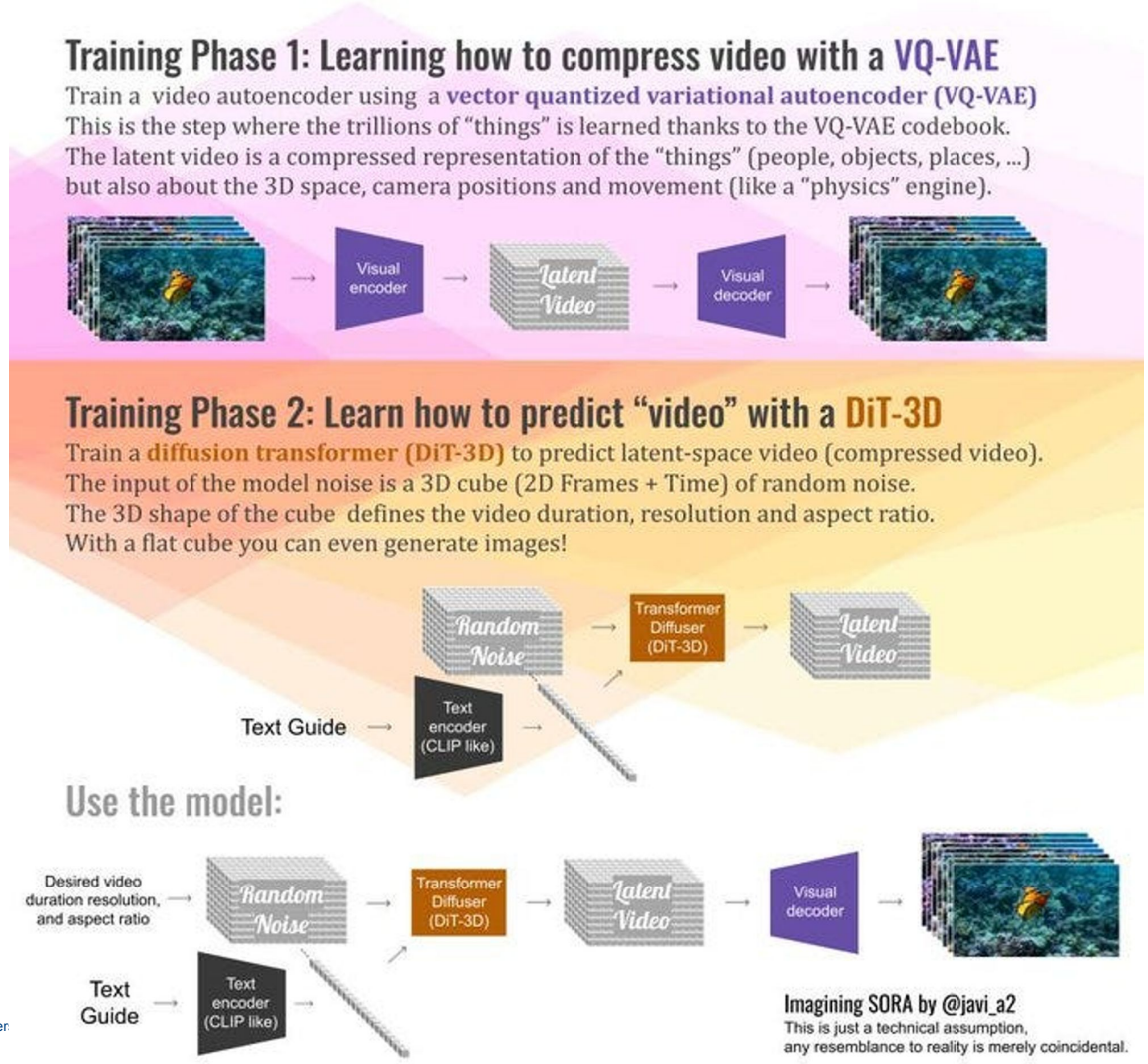
Open-Sora: Training and Inference

Training Stage:

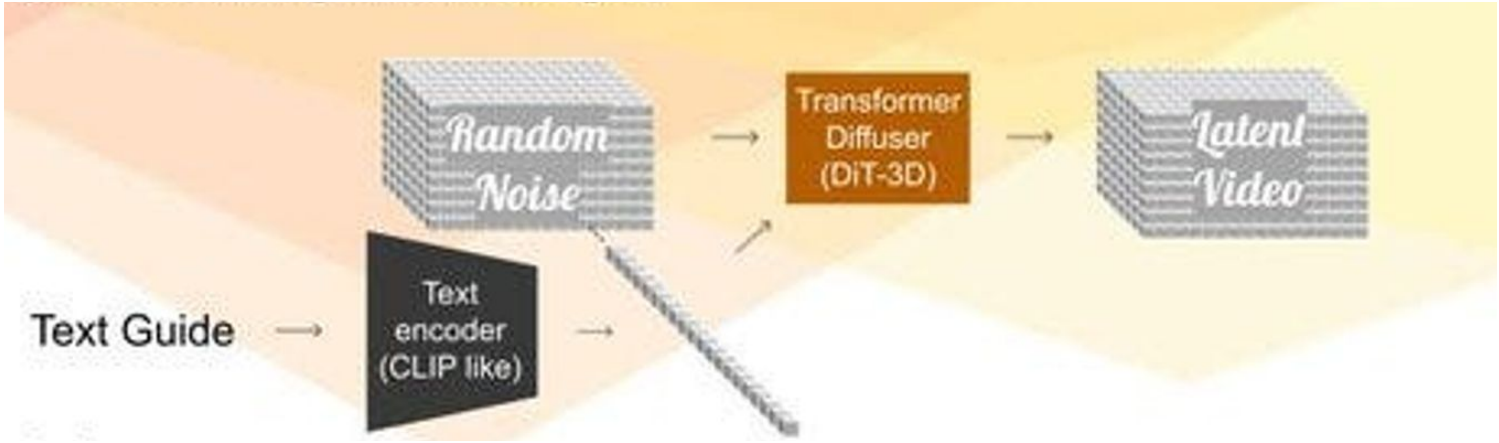
- Compression of video to the latent space.
- Train generation model in the latent space.

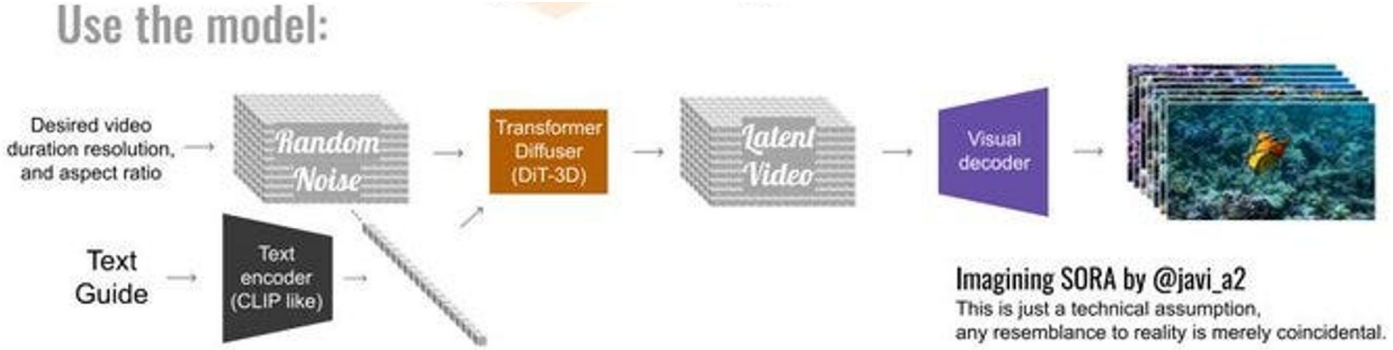
Inference Stage:

- Generation model generates latent-space video based on text prompt.
- Latent-space video decompressed into normal video.



Open-Sora roughly follows Sora's general design



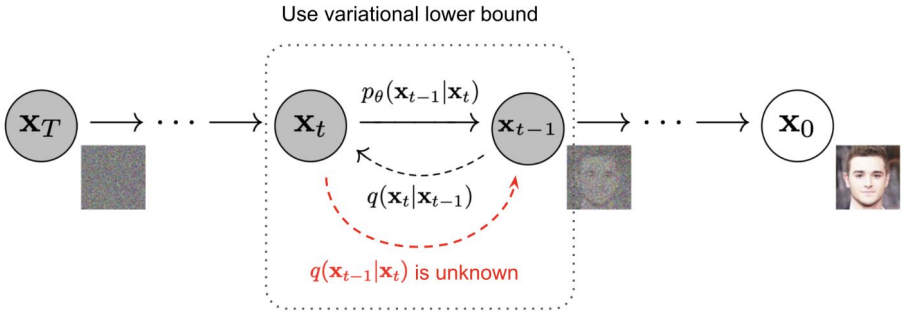


Diffusion Models: Forward Diffusion Process

Given a data point sampled from a real data distribution $\mathbf{x}_0 \sim q(\mathbf{x})$, let us define a forward diffusion process in which we add small amount of Gaussian noise to the sample in ***T*** steps, producing a sequence of noisy samples $\mathbf{x}_1, \ldots, \mathbf{x}_T$. The step sizes are controlled by a variance schedule $\{\beta_t \in (0,1)\}_{t=1}^T$

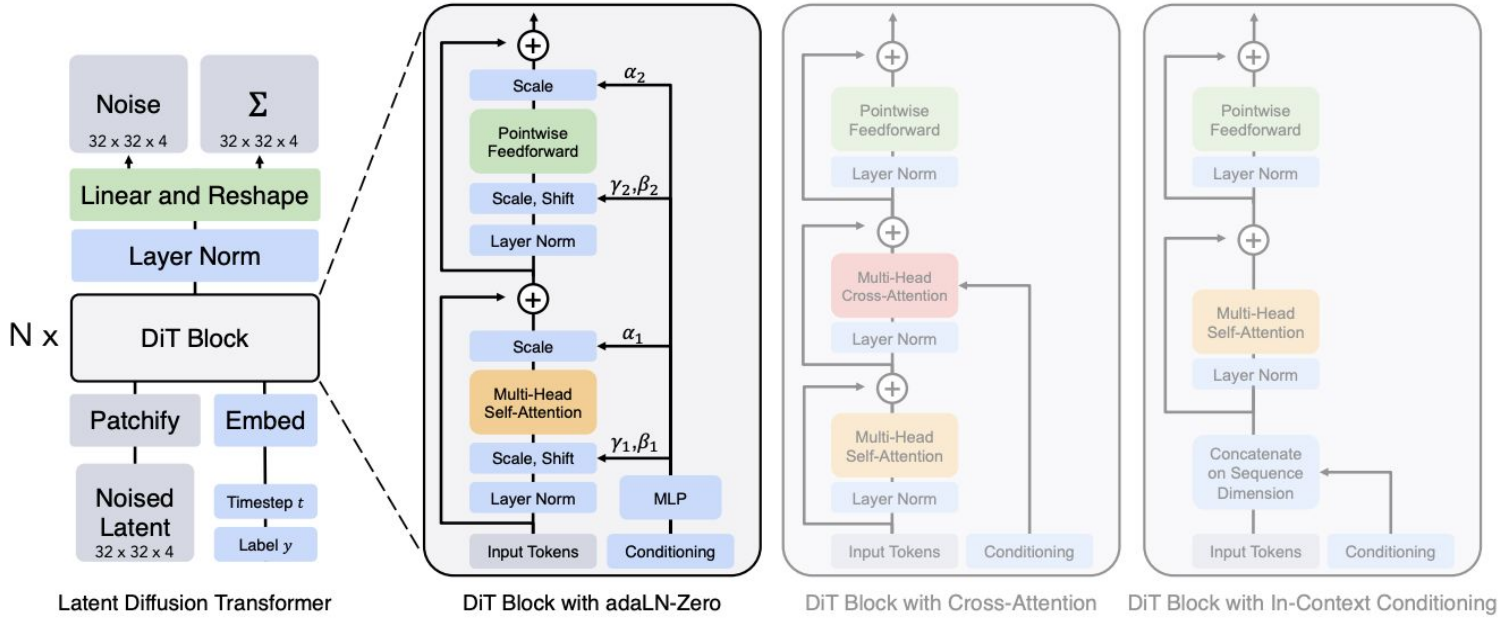
$$q(\mathbf{x}_t|\mathbf{x}_{t-1}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \quad q(\mathbf{x}_{1:T}|\mathbf{x}_0) = \prod_{t=1}^{T} q(\mathbf{x}_t|\mathbf{x}_{t-1})$$

The data sample $\mathbf{x}_0$ gradually loses its distinguishable features as the step ***t*** becomes larger. Eventually when $T \to \infty$, $\mathbf{x}_T$ is equivalent to an isotropic Gaussian distribution.



source: https://lilianweng.github.io/posts/2021-07-11-diffusion-models/#forward-diffusion-process

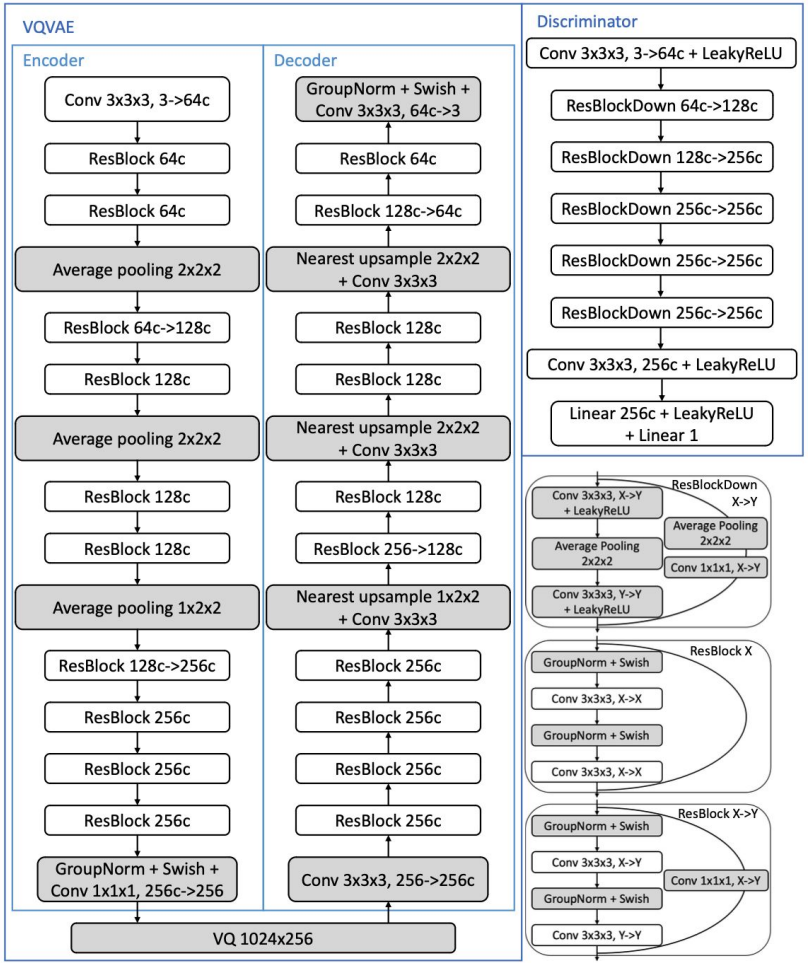
Diffusion Transformers



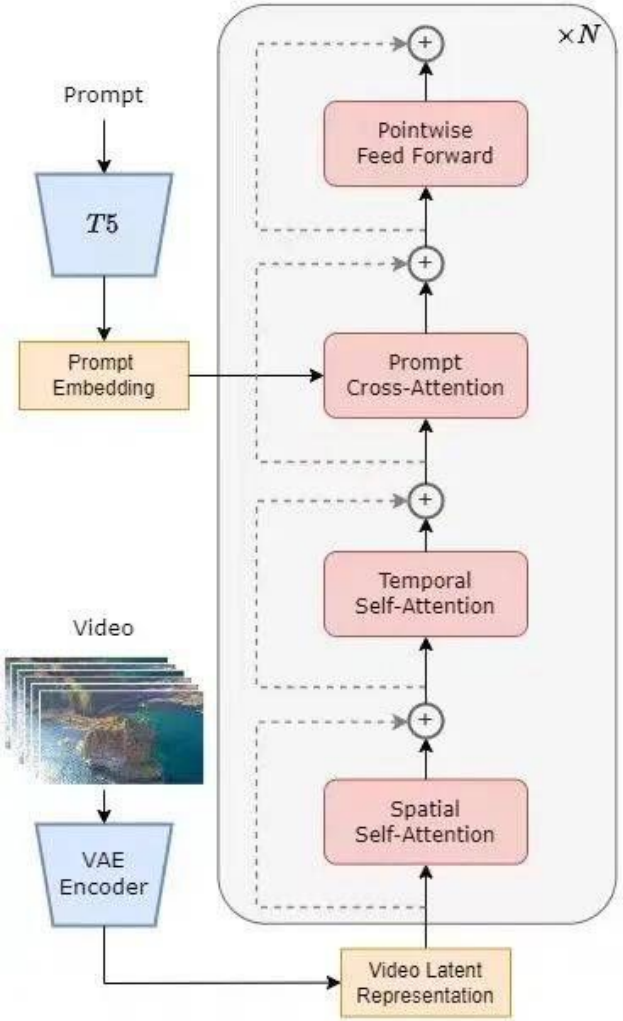
Source: Peebles, William, and Saining Xie. "Scalable diffusion models with transformers." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 4195-4205. 2023.

VAE

- **VAE:** relies on a surrogate loss.
  - **Encoder:** Uses repeated 3D convolutions to downsample the video into a much smaller latent space
  - **Decoder**: Uses repeated upsampling to transform latent space vectors back into the input video-space (i.e., reconstruction).
  - **Training**:
    - *Reconstruction loss* with KL regularization (input to decoder is sampled from a Gaussian Distribution of the latent space)
    - *Adversarial loss* to distinguish the reconstructed and original input video

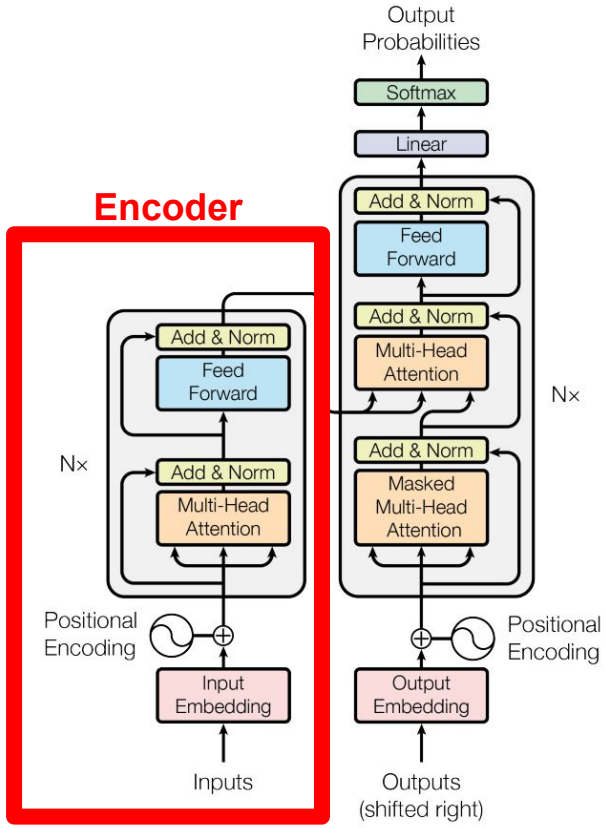


Source: Yu et al., "MAGVIT: Masked Generative Video Transformer." In Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 10459-10469. 2023.



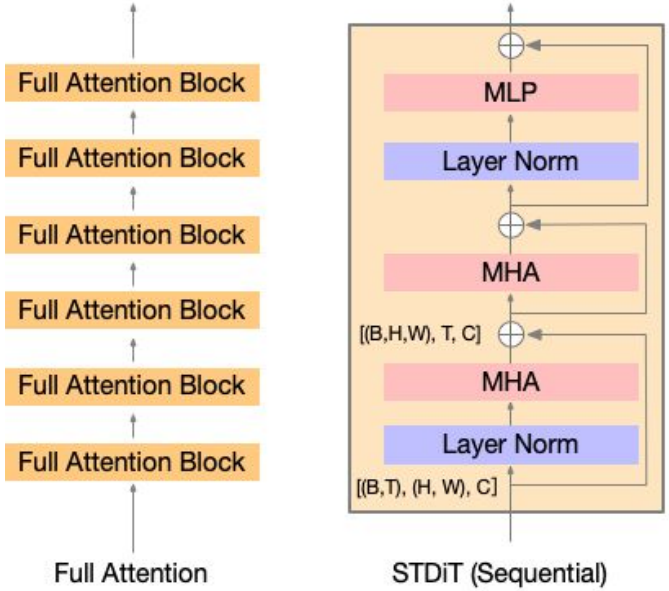
Text Encoder (T5)

- Input text tokens into an pre-trained T5 model and obtain the last hidden states from the encoder



Source: Vaswani et al., "Attention is all you need." Advances in neural information processing systems 30 (2017).

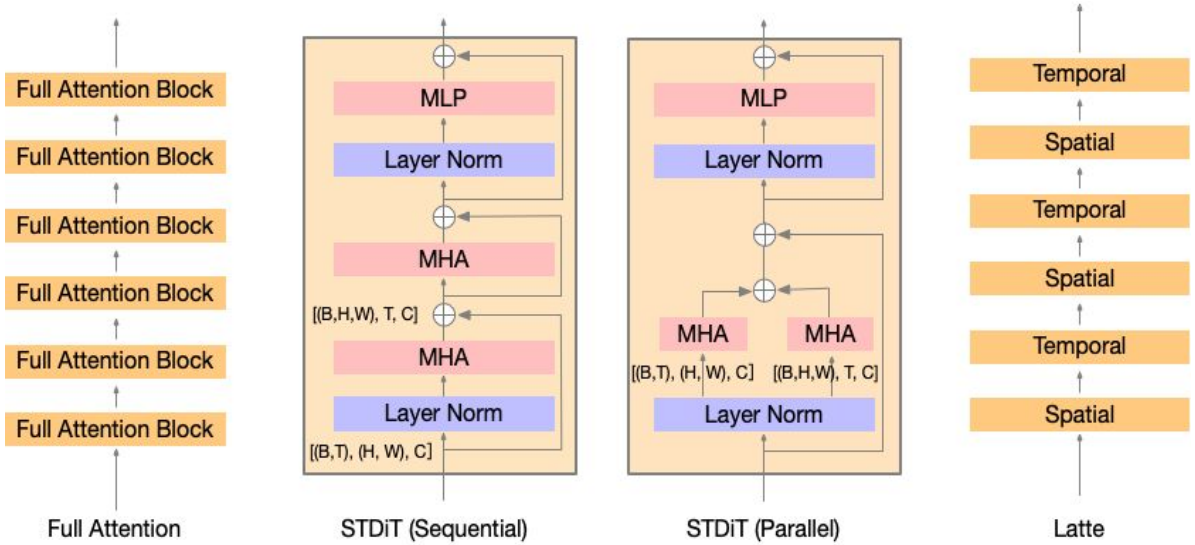
STDiT (Sequential) chosen for efficiency



- Video training involves ~1.5M tokens for each
- Attention has quadratic complexity!
- Spatial-temporal attention reduces computational costs.

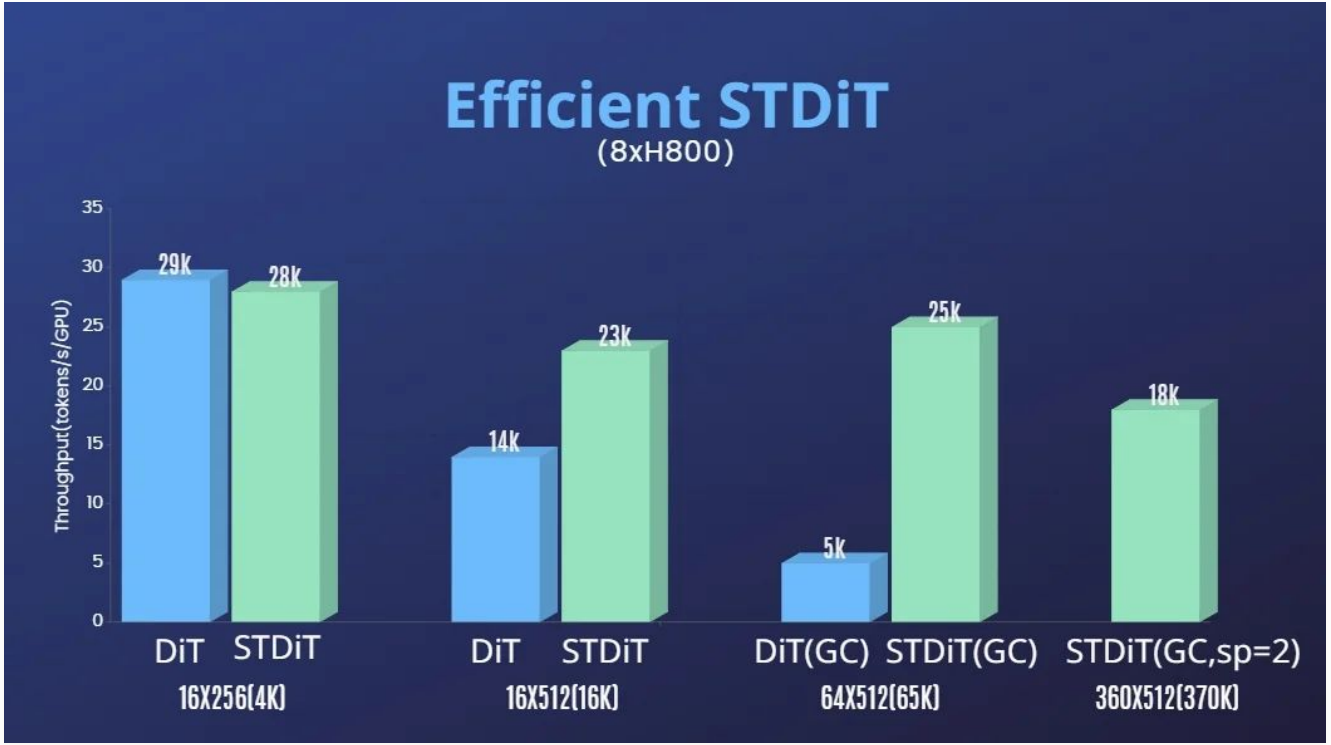
***B**: batch_size*
***H**: height*
***W**: width*
***T**: num of frames*
***C**: hidden dimension size*

STDiT (Sequential) chosen for efficiency



- Video training involves ~1.5M tokens for each
- (Attention has quadratic complexity)
- Spatial-temporal attention reduces computational costs.

Efficiency in Choosing the Architecture



Tokens Count (video clip -> token count)

fps 30, frame_interval 3, effective_fps 10

16x256x256 / 1x8x8 / 1x2x2 = 4096

Model hyperparameters

STDiT: 724M

VAE: 300M -> stability AI

T5: 4B -> huggingface

Open-Sora: Training and Inference

Training Stage:

- Utilize pre-trained VAE encoder for video data compression.
- Train STDiT model with text embedding in the latent space.

Inference Stage:

- Randomly sample Gaussian noise from VAE's latent space.
- Input noise and prompt embedding into STDiT for denoising.
- Pass denoised features into VAE decoder to generate video.

Open-Sora: Training Phases

Open-Sora's training process consists of three phases

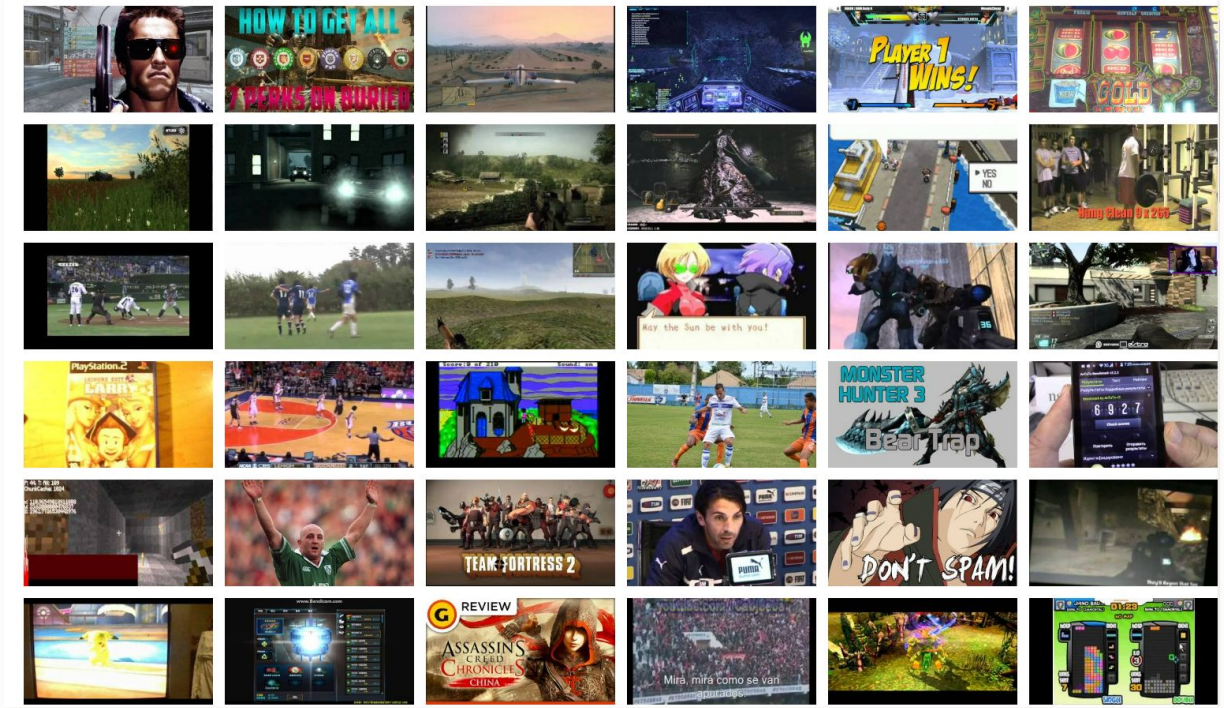
- Phase 1: Large-scale image pre-training; (PixArt-α)
- Phase 2: Large-scale video pre-training;
- Phase 3: High-quality video data fine-tuning.

Phase 1: Large-scale Image Pre-training

- High-quality model initialization reduces costs.
- **PixArt-alpha** (initializes our DiT)

16x256x256 on 366K pretraining datasets

Phase 2: Large-scale Video Pre-training



Diverse video data training

- Enhances model generalization.

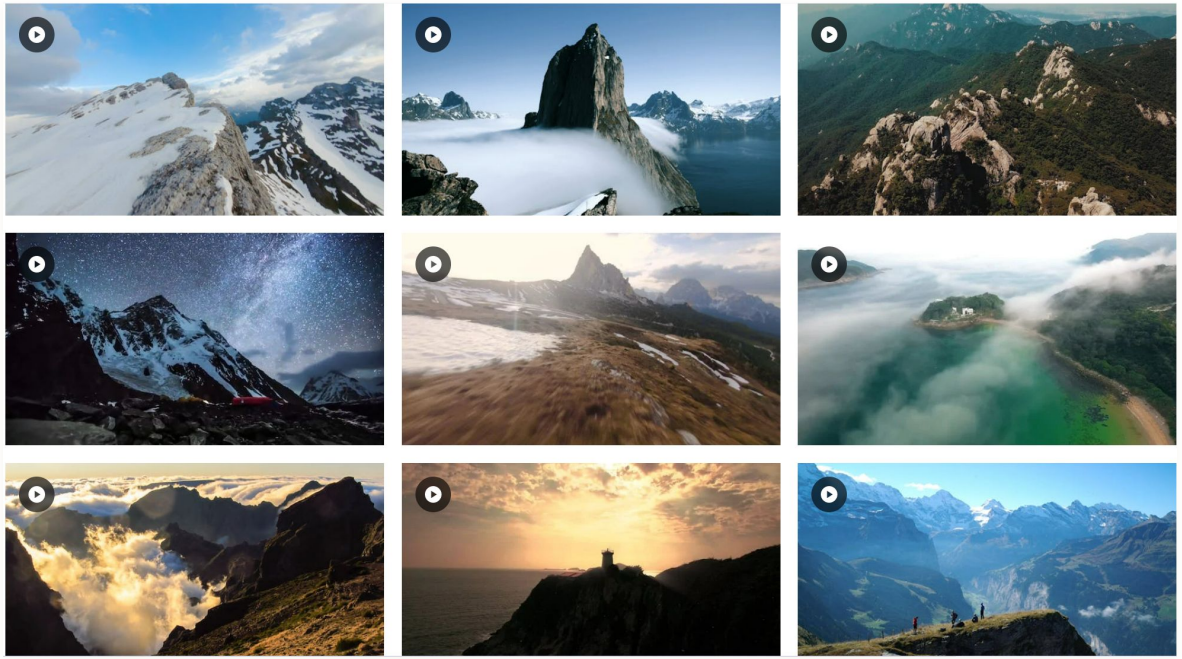
Temporal attention module

- Improves temporal relationship learning.

Optimized resolution

- Speeds up convergence and reduces costs.

Phase 3: High-quality Video Data Fine-tuning



Quality improvement

- Enhances video generation quality significantly via fine-tuning

Reduced data size

- Third-phase data smaller but higher duration, resolution, and quality.

Efficient scaling achieved

- Enables generation from short to long videos, low to high resolution, and low to high fidelity.

Data Preprocessing for Open-Sora

Data Is the Key to High Quality

- **Data quantity and quality** significantly impact video generation quality.
- Initial dataset (366K video clips) from HD-VG-130M shows varying quality and inaccurate captions.
- Additional collection of 20k high-quality videos from Pexels, with free licenses.
- Videos labeled using LLaVA image captioning model, utilizing three frames and a designed prompt.
- Designed prompt enhances LLaVA's caption quality for improved accuracy.

Datasets Used for Now

- **HD-VG-130M**

  HD-VG-130M comprises 130M text-video pairs. The caption is generated by BLIP-2. We find the cut and the text quality are relatively poor. It contains 20 splits. For OpenSora 1.0, we use the first split. We plan to use the whole dataset and re-process it.

- **Inter4k**

  Inter4k is a dataset containing 1k video clips with 4K resolution. The dataset is proposed for super-resolution tasks. We use the dataset for HQ training. The videos are processed as mentioned here.

- **Pexels.com**

  Pexels.com is a website that provides free stock photos and videos. We collect 19K video clips from this website for HQ training. The videos are processed as mentioned here.

Datasets Watching List

We are also watching the following datasets and considering using them in the future, which depends on our disk space and the quality of the dataset.

Name	Size
Panda-70M	70M videos
WebVid-10M	10M videos
InternVid-10M-FLT	10M videos
EGO4D	3670 hours
OpenDV-YouTube	1700 hours
VidProM	6.69M videos

Video Captioning

Our Prompts: A video is given by providing three frames in chronological order. Describe this video and its style ..... Here are some examples of good descriptions:

1. A stylish woman walks down a Tokyo .....
2. Several giant wooly mammoths .....
3. Drone view of waves crashing against .....

Sora's examples

GPT-4V

A serene cascade of waterfalls flows into an emerald green pond surrounded by lush greenery. The falls create soft white curtains of water, stirred by a gentle breeze, blending with the tranquil sounds of nature. The water is crystal clear, allowing visibility of the rocky bottom, hinting at the purity of this natural setting. Sunlight filters through the canopy above, creating dappled patterns of light that dance across the surfaces of both the rocks and water. This setting evokes a sense of peacefulness and the timeless beauty of an untouched forest.

LLaVA-1.6-Yi-34B

The serene beauty of a waterfall cascading into a secluded, lush green pool. The waterfall, with its multiple tiers, is surrounded by verdant foliage, creating a sense of tranquility and natural harmony. The camera angle is from a low perspective, looking up towards the waterfall, emphasizing its grandeur and the power of nature. The sunlight filters through the leaves, casting a warm glow on the water and the surrounding vegetation. The overall style of the video is a peaceful, naturalistic depiction of a waterfall in its natural habitat, with a focus on the interplay of light, water, and greenery.

- Human video labeling is costly and time-intensive.
- We employ image captioning models for video caption generation.
- Despite GPT-4V's superior performance, its 20s/sample speed is impractical.
- Batch inference with LLaVA achieves 3s/sample speed with comparable quality.
- LLaVA ranks second in MMMU and supports videos of any resolution.

Video Captioning

- **GPT-4V Captioning**

  Run the following command to generate captions for videos with GPT-4V:

```
python -m tools.caption.caption_gpt4 FOLDER_WITH_VIDEOS output.csv --key $OPENAI_API_KEY
```

  The cost is approximately $0.01 per video (3 frames per video). The output is a CSV file with path and caption.

- **LLaVA Captioning**

  First, install LLaVA according to their official instructions. We use the `liuhaotian/llava-v1.6-34b` model for captioning. Then, run the following command to generate captions for videos with LLaVA:

```
CUDA_VISIBLE_DEVICES=0,1 python -m tools.caption.caption_llava samples output.csv
```

  The Yi-34B requires 2 80GB GPUs and 3s/sample. The output is a CSV file with path and caption.

Video Filtering

Clip matching: https://openai.com/research/clip

Aesthetic Score: https://laion.ai/blog/laion-aesthetics/

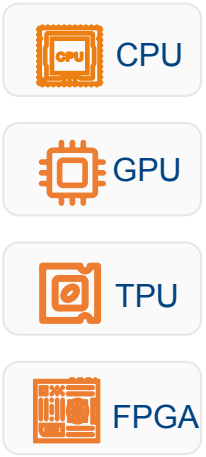
Optical Flow

OCR

Rule filtering

Efficient Training Strategies from Colossal-AI

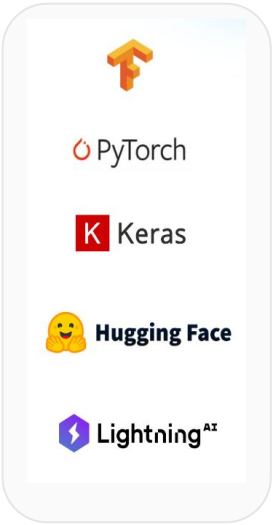
Efficient Training Strategies from Colossal-AI



Layer 3: Low Latency Inference System

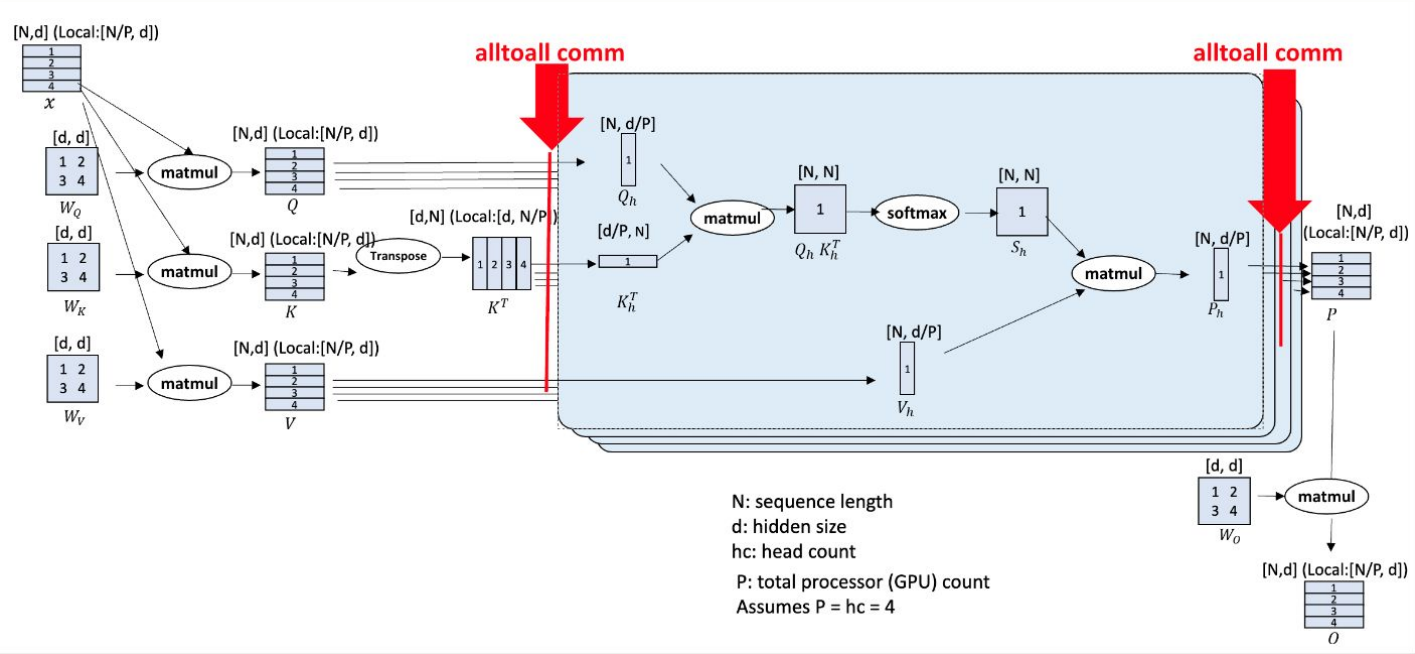
Layer 2: N-Dim Parallelism System

Layer 1: Efficient Memory System



- Maximize computational efficiency
- Minimize deployment cost

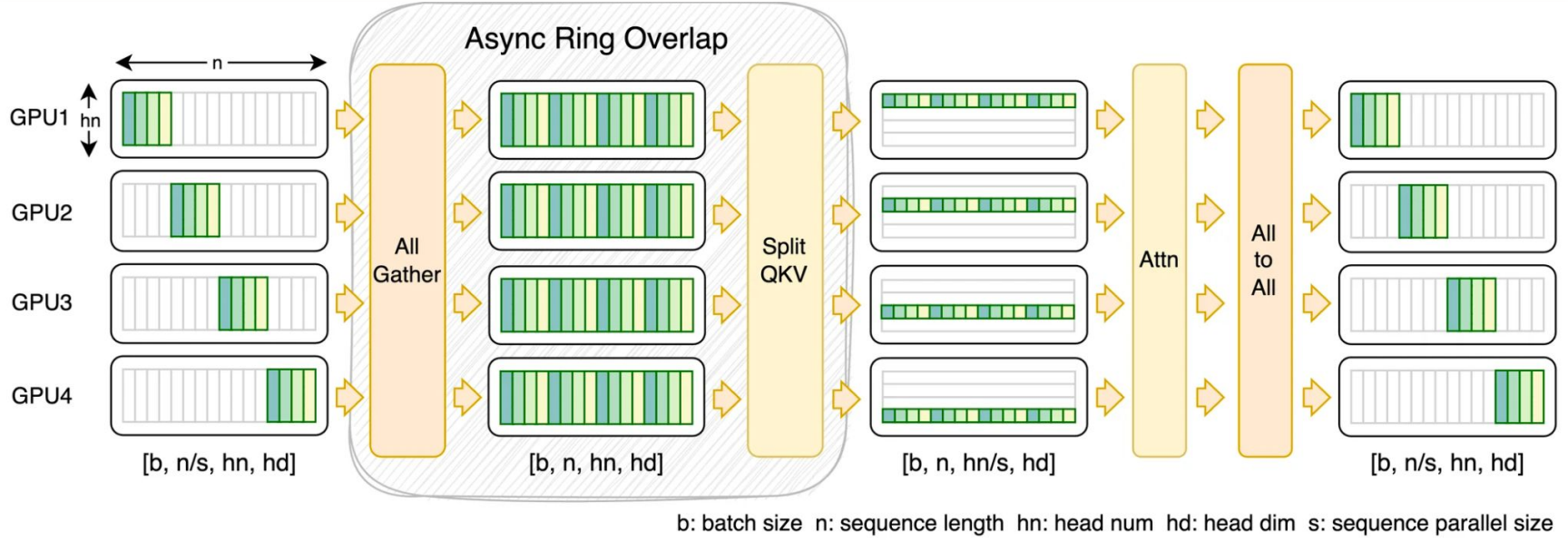
Efficient Training Strategies



- Enhanced performance with Ulysses: Ideal for small or long sequences

source: Jacobs, Sam Ade, et al. "Deepspeed ulysses: System optimizations for enabling training of e long sequence transformer models." arXiv preprint arXiv:2309.14509 (2023).

Efficient Training Strategies



FastSeq enhances training efficiency:

- Overlaps qkv projection and all-gather communication.
- Minimal additional memory usage.

source: https://github.com/NUS-HPC-AI-Lab/OpenDiT/tree/master

Efficient Training Strategies

Open-Sora
Longer sequences for better performance with Colossal-AI

Model	GPUs	Num Patches	SP Size	Parallel Mode	Throughput
DiT-XL/2	8	614400	1	Colossal-AI ZeRO2	0.79
	8		2	Colossal-AI FastSeq	**1.15**
	4		1	Colossal-AI ZeRO2	0.44
	4		2	Colossal-AI Ulysses	**0.62**

- More than 40% performance improvement and cost reduction over the baseline solution at a sequence length of 600K

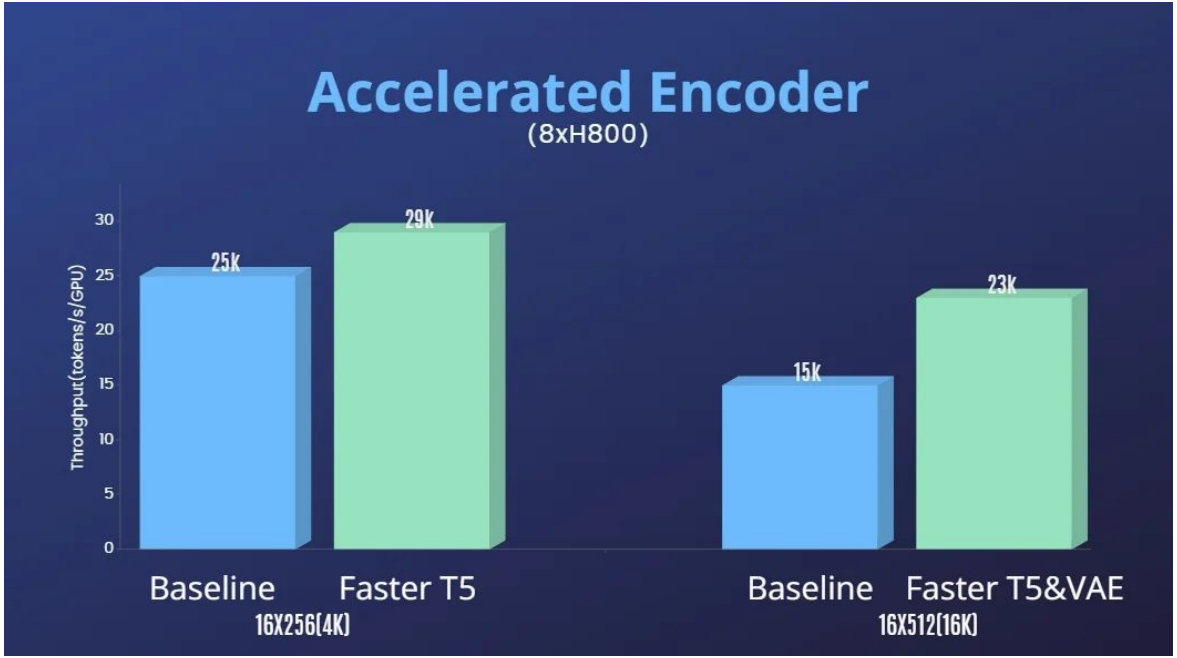
Efficient Training Strategies

Open-Sora
Longer sequences for better performance with Colossal-AI

Model	GPUs	Num Patches	SP Size	Parallel Mode	Throughput
DiT-XL/2	8	614400	1	Colossal-AI ZeRO2	0.79
		819200	2	Colossal-AI Ulysses	**0.88**

- Supports training with 30% longer sequences up to 819K+ while guaranteeing faster training speeds

Efficient Training Strategies



- Achieves 1.55x training acceleration for 64-frame, 512x512 resolution videos.
- Seamless training of 1-minute 1080p HD video task on a single server (8*H800).

Profiling

CPU computation

Dataloading

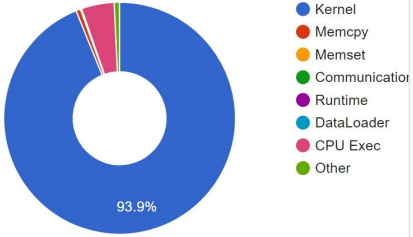
Memory copy

GPU computation

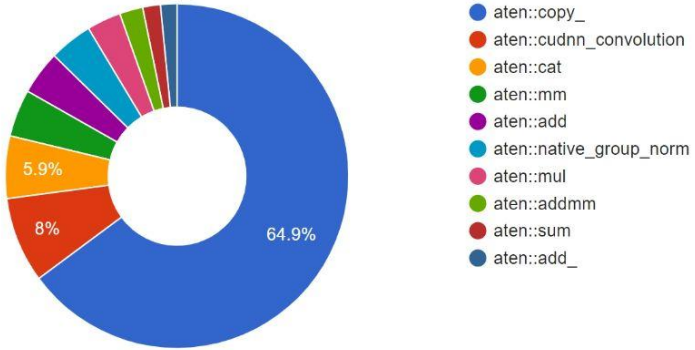
Communication

Execution Summary

Category	Time Duration (us)	Percentage (%)
Average Step Time	8,411,007	100
Kernel	7,894,108	93.85
Memcpy	53,476	0.64
Memset	1,412	0.02
Communication	12,878	0.15
Runtime	0	0
DataLoader	0	0
CPU Exec	384,606	4.57
Other	64,528	0.77

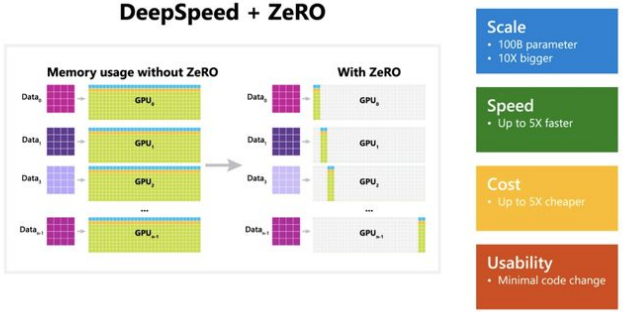


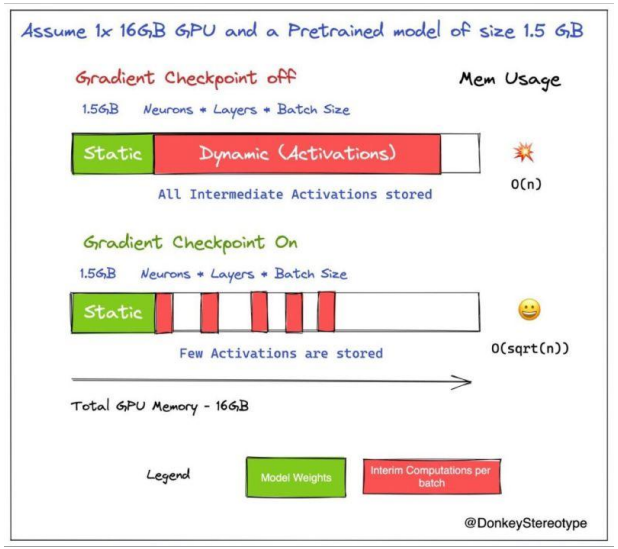
Host Self Time (us)



Methods

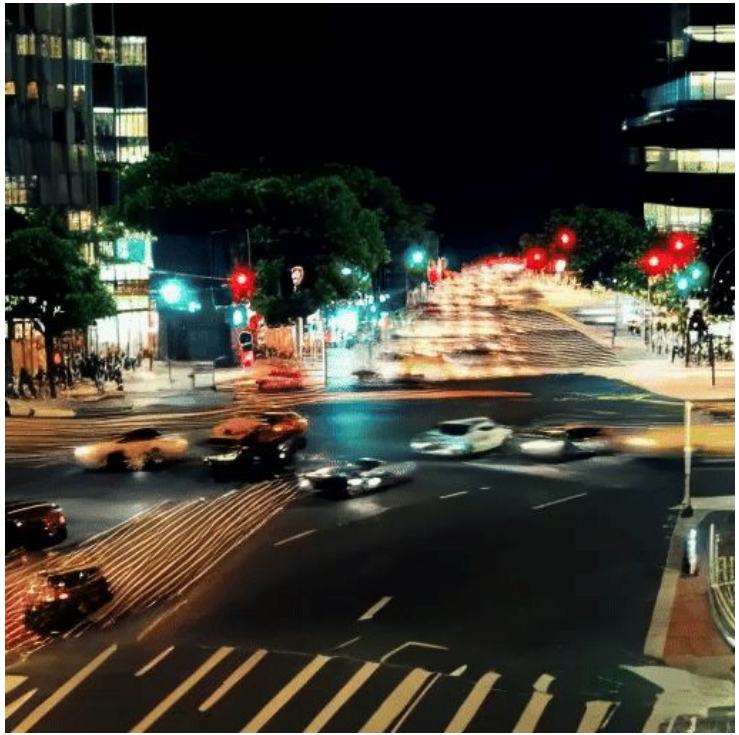
https://github.com/hpcaitech/Open-Sora-dev/blob/dev/v1.0.1/docs/acceleration.md





Performance

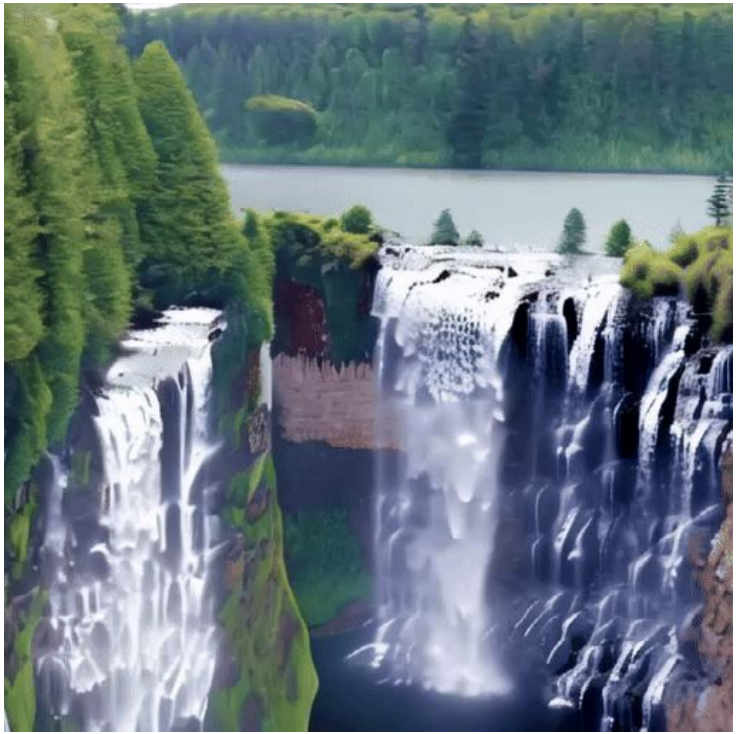
Demo



A View of A Bustling City at Night

Prompts: A bustling city street at night, filled with the glow of car headlights and the ambient light of streetlights. [...]

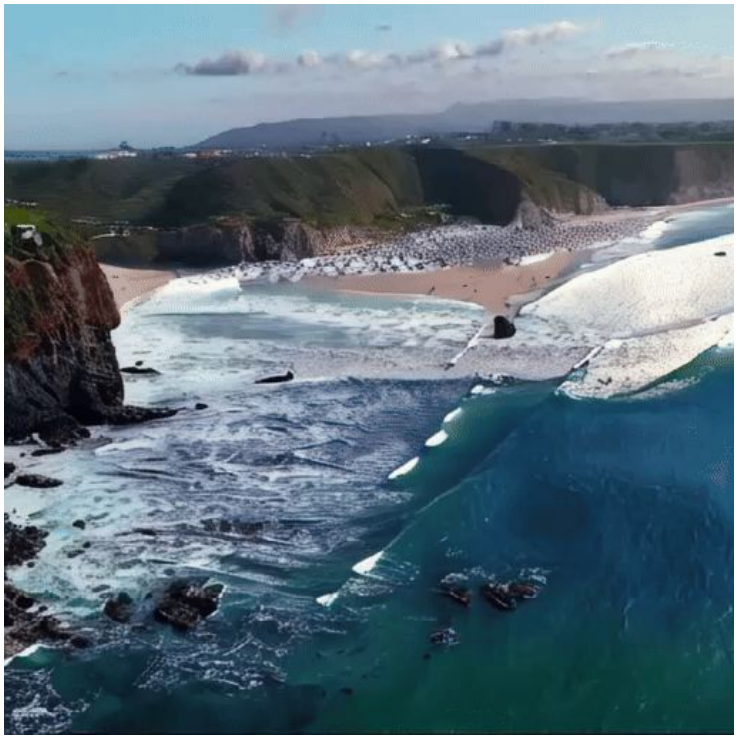
Demo



A Bird’s Eye View of A Waterfall Cascading down A Cliff

Prompts: The majestic beauty of a waterfall cascading down a cliff into a serene lake. [...] The camera angle provides a bird's eye view of the waterfall.

Demo



An Aerial Views of the Majestic Beauty of a Coastal

Prompts: A soaring drone footage captures the majestic beauty of a coastal cliff, [...] The water gently laps at the rock base and the greenery that clings to the top of the cliff.

Demo

A Picturesque Snowy Mountain with Hot Air Balloons

Prompts: A vibrant scene of a snowy mountain landscape. The sky is filled with a multitude of colorful hot air balloons, [...]

Demo

A Sea Turtle's Serene Swim Among Coral Reefs

Prompts: A serene underwater scene featuring a sea turtle swimming through a coral reef. The turtle, with its greenish-brown shell, is the main focus of the video, swimming gracefully [...]

Demo

Capturing the Milky Way's Timeless Beauty in Time-Lapse

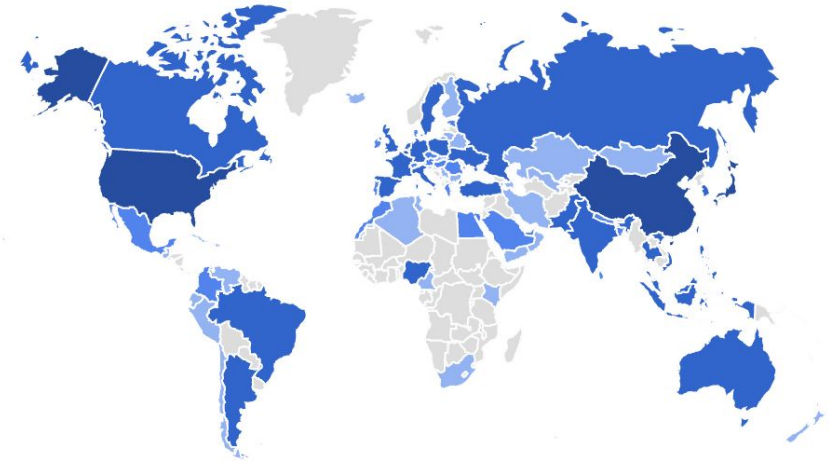
Prompts: A serene night scene in a forested area. [..] The video is a time-lapse, capturing the transition from day to night, with the lake and forest serving as a constant backdrop. [...]

Demo

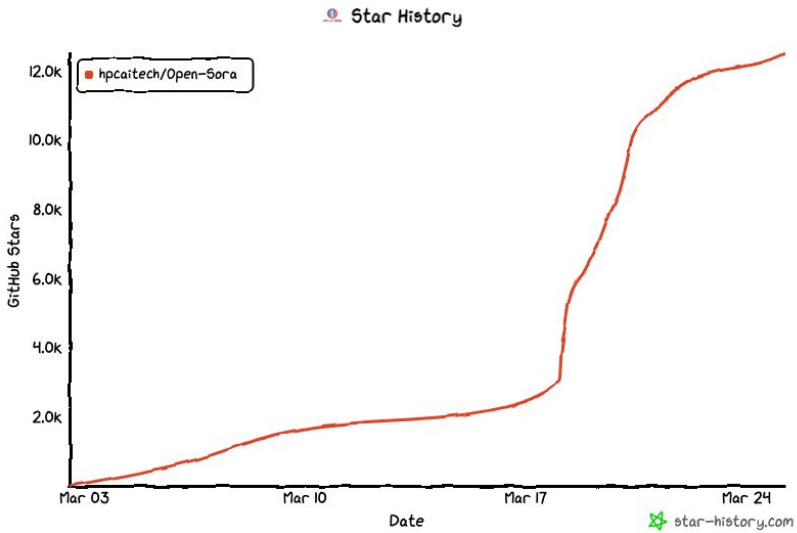
SunFlower Field Captured from A Low Angle

Prompts: The vibrant beauty of a sunflower field. The sunflowers, with their bright yellow petals and dark brown centers, are in full bloom, creating a stunning contrast [...]

Rapid Growth in Community



Global Followers Everywhere



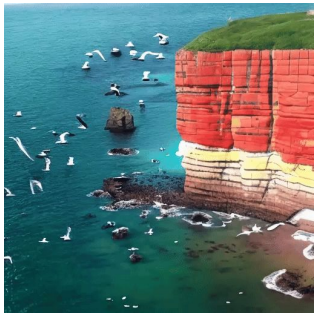
GitHub Stars Surging Rapidly

Open-Sora is public and available at https://github.com/hpcaitech/Open-Sora

Reproduction Demo from Community





Tutorial

Inference

Inference with DiT pretrained on ImageNet

The following command automatically downloads the pretrained weights on ImageNet and runs inference.

```
python scripts/inference.py configs/dit/inference/1x256x256-class.py --ckpt-path DiT-XL-2-256x256.pt
```

Training

To resume training, run the following command. `--load` different from `--ckpt-path` as it loads the optimizer and dataloader states.

```
torchrun --nnodes=1 --nproc_per_node=8 scripts/train.py configs/opensora/train/64x512x512.py --data-path YOUR_CSV_PATH --load YOUR_PRETRAINED_CKPT
```

To enable wandb logging, add `--wandb` to the command.

```
WANDB_API_KEY=YOUR_WANDB_API_KEY torchrun --nnodes=1 --nproc_per_node=8 scripts/train.py configs/opensora/train/64x512x512.py --data-path YOUR_CSV_PATH --wandb True
```

Training Hyperparameters

`dtype` is the data type for training. Only `fp16` and `bf16` are supported. ColossalAI automatically enables the mixed precision training for `fp16` and `bf16`. During training, we find `bf16` more stable.

Future Plans

Completing the Data Processing Pipeline

- Integrating dense optical flow analysis
- Incorporating aesthetics scores assessment
- Implementing text-image similarity metrics
- Addressing deduplication concerns

Training Video Compression Network

- Work in progress to train and open-source a high-quality video compression network that encodes videos into latent representations both temporally and spatially