



AI+ 研发数字峰会
AI+ Development Digital summit



Big Data Empower Large Models

张松昕 | 南方科技大学

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **K+峰会**  **敦煌站**

K+ 思考周®研习社

时间: 2025.08.29-30

 **K+峰会**  **上海站**

K+ 金融专场

时间: 2025.10.17-18

 **K+峰会**  **香港站**

K+ 思考周®研习社

时间: 2025.11.25-26



K+峰会详情



 **AiDD峰会**  **上海站**

AI+研发数字峰会

时间: 2025.05.17-18

 **AiDD峰会**  **北京站**

AI+研发数字峰会

时间: 2025.08.08-09

 **AiDD峰会**  **深圳站**

AI+研发数字峰会

时间: 2025.11.28-29



AiDD峰会详情



张松昕

南方科技大学研究学者

南方科技大学统计与数据科学系研究学者，UCloud顾问资深算法专家，曾任粤港澳大湾区数字经济研究院访问学者，主导大模型高效分布式训练框架的开发，设计了SUS-Chat-34B的微调流程，登顶Open LLM Leaderboard、Opencompass同参数量级模型榜首。

构建先进大模型训练基础设施

- 超长序列高效分布式训练框架
- 云原生大模型基础设施与智算平台
- 面向多模态数据预训练场景流式数据框架

打磨一流大模型预训练与微调技术

- 世界一流的通用领域语言模型
- 面向垂直领域应用的智能问答系统
- 对标世界领先的多模态生成模型，构建多模态生成为基础的世界模型

► SUS-Chat: Instruction Tuning Done Right



百亿参数级别通用领域语言模型达到世界领先水平

- 基于 Scaling law 构建数据筛选策略
 - 快速迭代构造十亿token量级高质量数据
- 自研分布式框架助力模型高效训练
 - 全尺寸模型的训练成本线性拓展



国内外领先的通用语言模型

发布时，在OpenLLMLeaderboard 40,000+的模型中排名第一，

OpenCompass中Chat模型排名第一

OpenCompass (1227)

Open LLM Leaderboard (1205)

大语言模型评测榜单

模型	发布日期	类型	参数量	精度	推理	生成	评测	训练	部署
QWen-72B	20231109	预训练	72B	73.1	77.2	88	70.5	70	84.4
Open-72B	20231109	预训练	72B	87.1	81	88.8	84.8	73.6	83.1
TigerBot-70B-Chat-v1	20230915	预训练	70B	80.2	83.7	89.3	84.5	86	87
Qwen-72B-Chat-v1	20231017	预训练	72B	85.3	87.2	92.4	87	81.3	87.4
Qwen-72B-Chat-v1.5	20231205	预训练	72B	88.7	79.2	92.7	88.7	80.2	87.6
Open-72B	20230905	预训练	72B	82.4	75.1	92.7	88.1	85.6	80.1
Py-72B	20231102	预训练	72B	82.1	78.1	88.8	84.8	80.8	80.8
Open-72B-Chat	20231130	预训练	72B	82	78.8	88.8	83.8	78.8	88.4
ChatGPT	20230111	预训练	66B	71.8	82.7	88.6	84.5	84.8	84
OpenAI GPT-4o	20231116	预训练	180B	81.1	78.8	92.8	88.8	88.1	87.7
Open-72B-Chat	20230905	预训练	72B	80.3	71.2	92.1	87.2	80.2	84.8
ErnieBot-4.0-Turbo	20231115	预训练	90B	80.6	71.8	90.7	84.8	80.4	80
Open-72B	20230905	预训练	72B	80	72.4	88	80.8	80.8	82.2
ErnieBot-4.0-Base	20231109	预训练	170B	79.8	80.8	88.3	84.7	80.8	84.8
LLaMA-3.1-70B	20230901	预训练	70B	80.1	82.8	88	80.7	87.8	84.8
Meta Llama-3.1-70B	20231016	预训练	70B	80.3	84.4	88.8	85.8	84.1	84.8
TigerBot-70B-Chat-v1	20230818	预训练	70B	80.3	82.2	88.2	84	81.2	85.1

Open LLM Leaderboard

The Open LLM Leaderboard aims to track, rank and evaluate open LLMs and chatbots.

Submit a model for automated evaluation on the GPU cluster on the "Submit" page! The leaderboard's backend runs the great [Leader AI Language Model Evaluation Harness](#) - read more details in the "About" page!

LLM Benchmark Metrics through time About Submit here!

Search for your model (separate multiple queries with ' ') and press ENTER...

Select columns to show: Average, ARC, HelloSwag, MMLU, TruthfulQA, Winogrande, CoSQA, Type, Architecture, Precision, Hub License, #Params (B), Hub, Available on the hub, Model ID

T	Model	Average	ARC	HelloSwag	MMLU	TruthfulQA	Winogrande	CoSQA
1	OpenAI GPT-4o	74.13	87.03	82.03	86	76.4	79.36	84.36
2	Qwen-72B-Chat-v1.5	73.22	86.3	83.91	76.41	87.84	83.5	72.18
3	Qwen-72B-Chat-v1	73.79	87.75	86.02	72.42	85.85	84.21	83.60
4	OpenAI GPT-4o-mini	71.73	84.81	85.14	75.54	87.48	83.11	83.09
5	OpenAI GPT-4o	71.69	88.43	85.21	78.13	84.43	84.06	85.82
6	OpenAI GPT-4o	71.37	88.77	85.1	80.58	87.54	83.58	83.53
7	OpenAI GPT-4o	71.31	84.93	85.02	76.18	85.84	83.03	83.94
8	OpenAI GPT-4o	70.95	84.99	85.63	76.33	85.8	82.79	84.8
9	OpenAI GPT-4o	70.81	85.36	85.58	76.86	85.84	82.56	83.64
10	OpenAI GPT-4o	70.57	84.13	86.24	74.89	86.37	82.4	87.39
11	OpenAI GPT-4o	70.11	73.08	87.09	70.58	82.35	83.88	85.26

国内外领先的通用语言模型

- 在难度最大的**数学推理**任务中，位居**世界第一**
- 逻辑推理的综合能力**达到世界领先水平

Model	gsm8k (0-shot)	MATH (0-shot)	BBH (0-shot)
GPT-4	91.4	45.8	86.7
OrionStar-Yi-34B-Chat	54.36	12.8	62.88
Yi-34B-Chat	63.76	10.02	61.54
Qwen-72b-Chat	<u>76.57</u>	<u>35.9</u>	<u>72.63</u>
Deepseek-67b-Chat	74.45	<u>29.56</u>	<u>69.73</u>

SUSChat-80.06

28.8
67.62

大语言模型评测榜单

全部数据集 中文数据集 英文数据集

综合榜单 学科能力 语言能力 知识能力 理解能力 **推理能力**

排名	模型	发布日期	更新	类型	参数量	推理能力(均分)
1	GPT-4 OpenAI	2023/3/15	更新:2023/9/1	对话	N/A	74.4
2	ChatGLM3-6B-Base ZhipuAI	2023/10/27	更新:2023/10/30	基座	6B	67.4
3	TigerBot-70B-Chat-V2 TigerResearch	2023/9/15	更新:2023/10/13	对话	70B	67
4	ChatGPT OpenAI	2023/3/1	更新:2023/9/1	对话	N/A	64
5	SUSChat-34B SUSTech	2023/12/5	更新:2023/12/26	对话	34B	63.6
6	Qwen-72B Alibaba	2023/11/30	更新:2023/12/4	基座	72B	63.1
7	Qwen-14B Alibaba	2023/9/25	更新:2023/9/25	基座	14B	60.1
8	DeepSeek-67B-Chat DeepSeek	2023/11/29	更新:2023/12/4	对话	67B	59
9	Qwen-72B-Chat Alibaba	2023/11/30	更新:2023/12/4	对话	72B	58.4
10	OrionStar-Yi-34B-Chat OrionStarAI	2023/11/16	更新:2023/11/22	对话	34B	57.7
11	StableBeluga2	2023/7/21		对话	70B	57.1

5 / 38

▶▶ Efficient distributed training infra

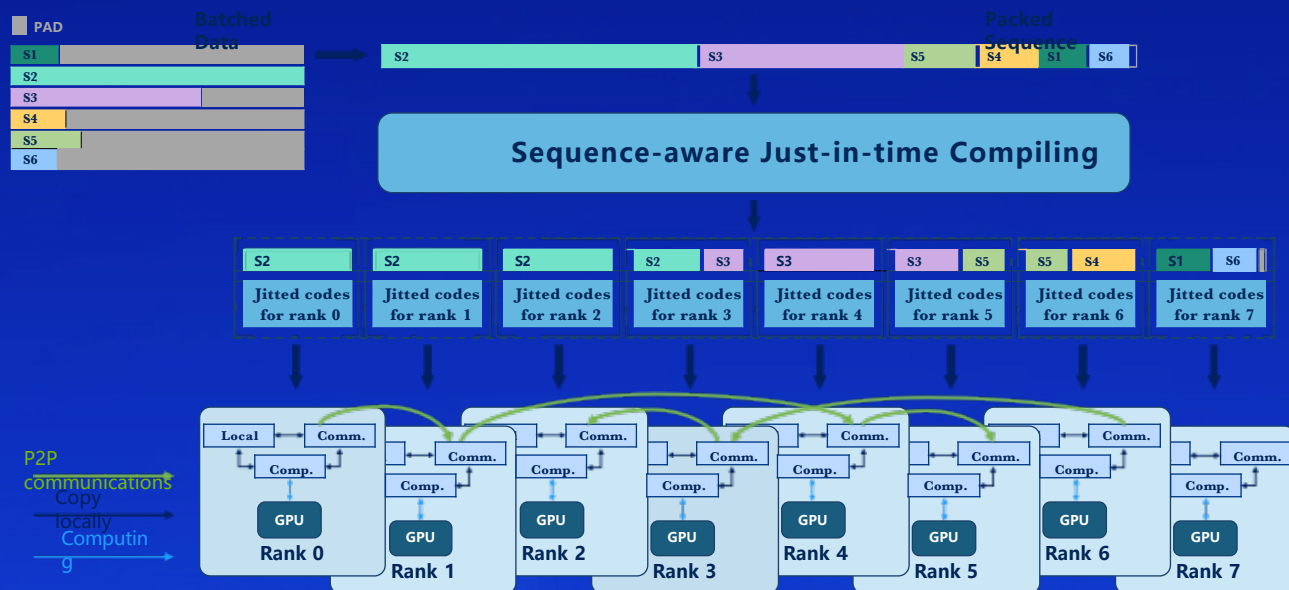
标准场景下，分布式训练框架在各指标和特性上超越世界主流框架

- ✓ 唯一支持6D并行的分布式训练框架
- ✓ 高效训练算子适配与优化
- ✓ 无效Padding计算裁剪

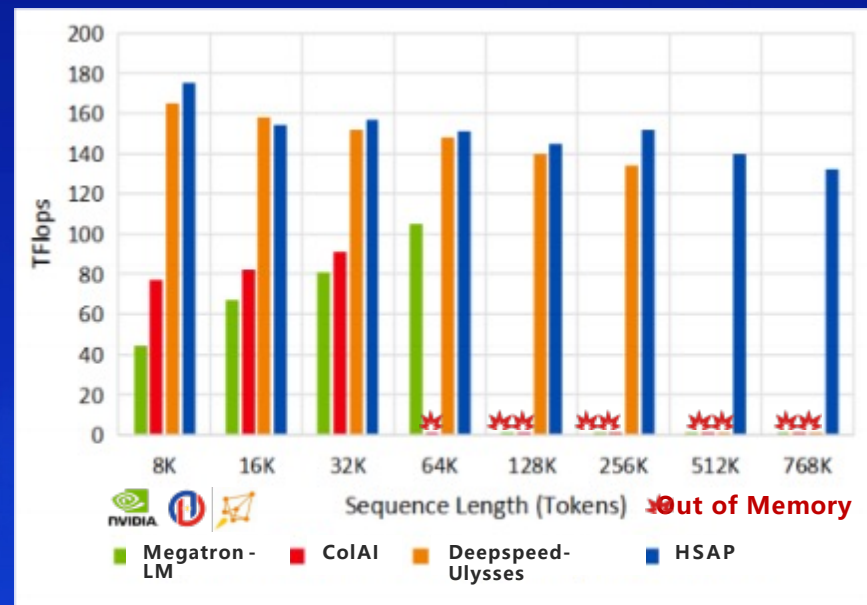
	FlashAttn-2	FP8 (H100)	Parallel Level	Padding Free	Fused Kernel	Static Graph	TGS[^1]
Platformers	✓	✓	6D + Zero	✓	100%	✓	3743
Megatron	✗	✓	4.5D	✗	80%	✗	3581
Deepspeed	✓	✓	3D + Zero	✗	60%	✗	✗
Colossal-ai	✗	✗	5D + Zero	✗	✗	40%	✗ 2610

Scaling exact attention to ultra long sequence

基于即时编译的拓扑感知序列并行，
唯一实现显存瓶颈与序列长度解耦分布式训练框架



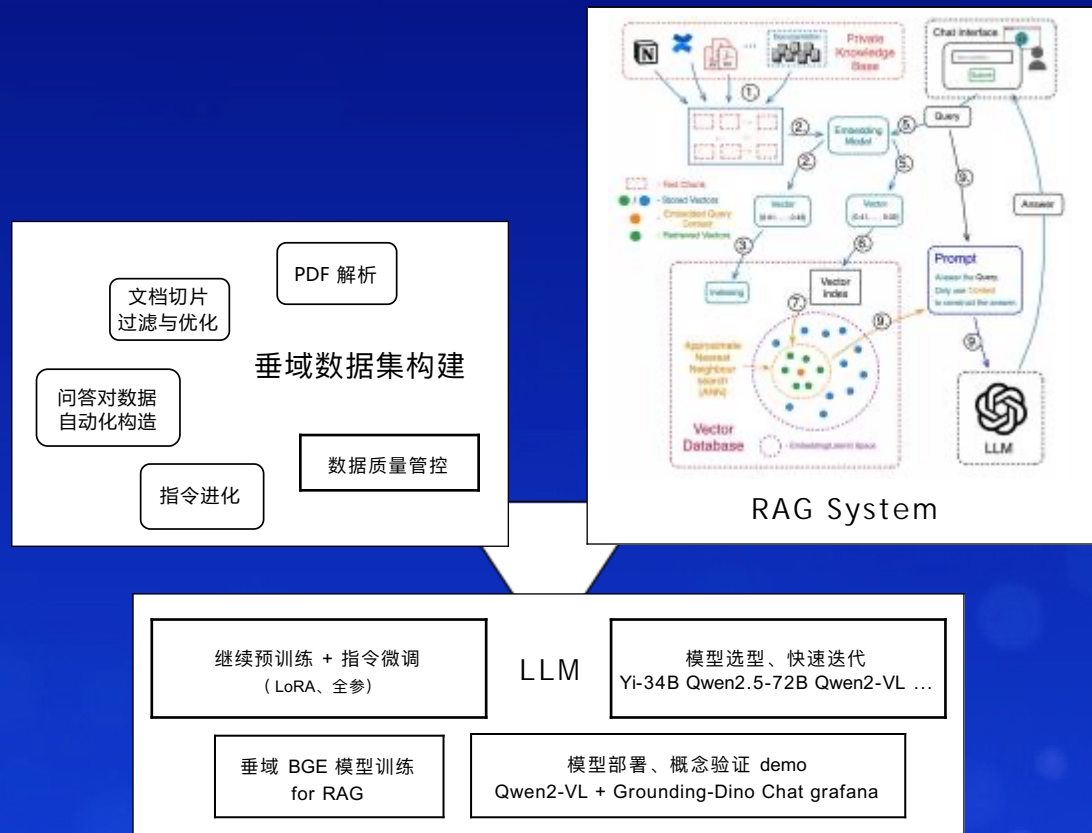
Executing Jitted codes on each ranks



通过增加计算资源高效扩展至主流框架无法达到超长序列

海诚建筑行业知识库智能问答系统

- 知识检索增强问答
- 长序列问答支持 (>32k tokens)
- 多引用来源的知识推理、总结
- 实现复杂表格内容识别
- 行业问题准确回答 (准确率90%)
- 34B基座模型综合效果超GPT4o



Cloud-native Streaming Data Infrastructure

► Dawning of the World Model Era

How many data SORA uses?

We take inspiration from large language models which acquire generalist capabilities by training on **internet-scale data** ¹

目前有一个比较准确的估计，一分钟视频约为 1M tokens 。 ²

- 每分钟上传至 YouTube 的视频是 500h 的量级。
- 则近五年的 YouTube 上的视频数据总量为：
 - 13亿小时 = 788亿分钟。
- 由于Difusion模型训练text to video需要高质量的标注视频，因此我们可以估计Sora训练的视频量级为**1亿分钟**左右。

~ **500TB** trained data

~ **500PB** raw data

1. [Video generation models as world simulators\(SORA tech report\)](#)

2. [浅谈 Sora 未来的千倍推理算力需求](#)

▶▶ Here comes challenge

Training on internet-scale

multi-modal data

More diverse and complicated

data processing workflows

Legacy training paradigm

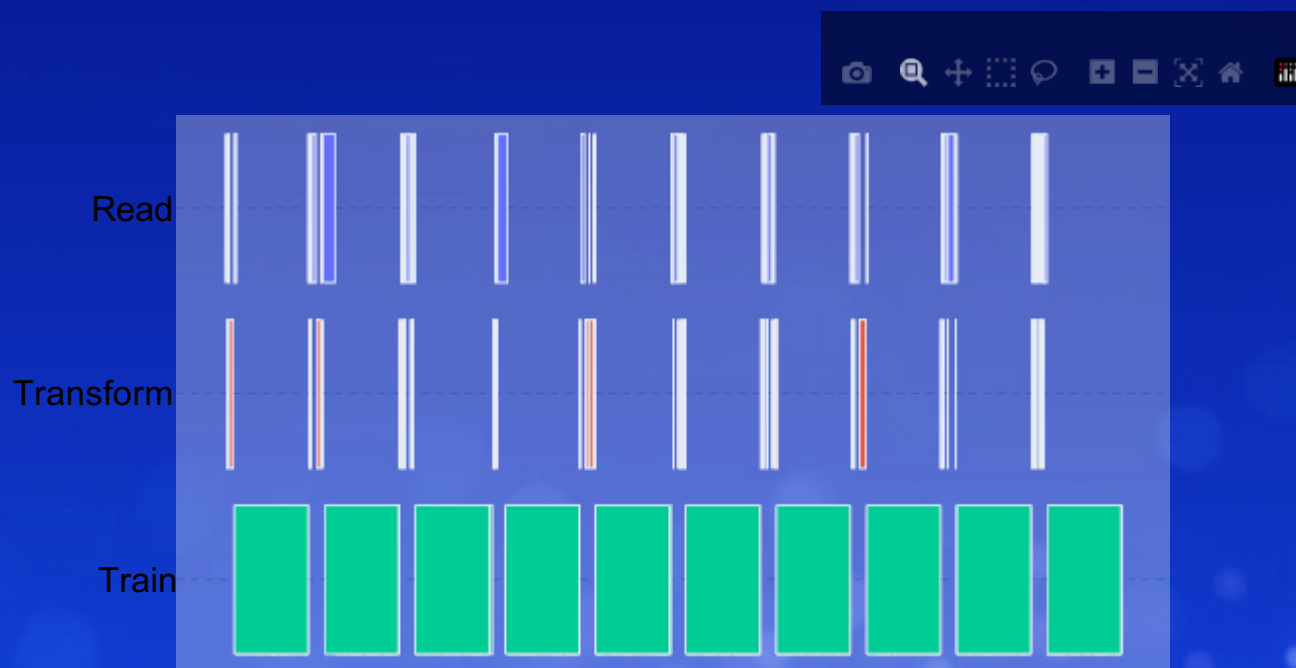
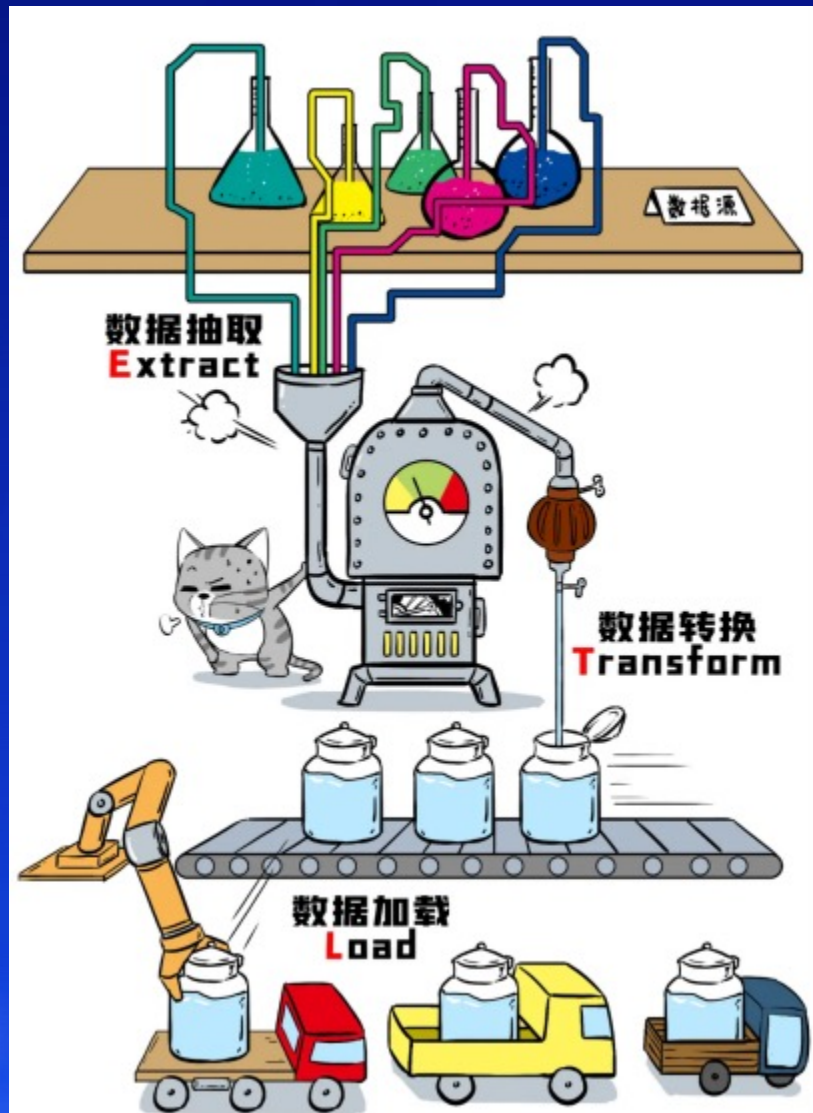
传统的训练方式通常是一次性将数据下载到本地，然后进行处理。

```
1
2 import datasets
3 from transformers import AutoTokenizer
4
5 dataset = datasets.load_dataset(
6     "rotten_tomatoes",
7     split="train",
8 )
9 tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")
10
11 dataset = dataset.map(
12     lambda examples: tokenizer(examples["text"]),
13     batched=True,
14 )
15
16 ...
```

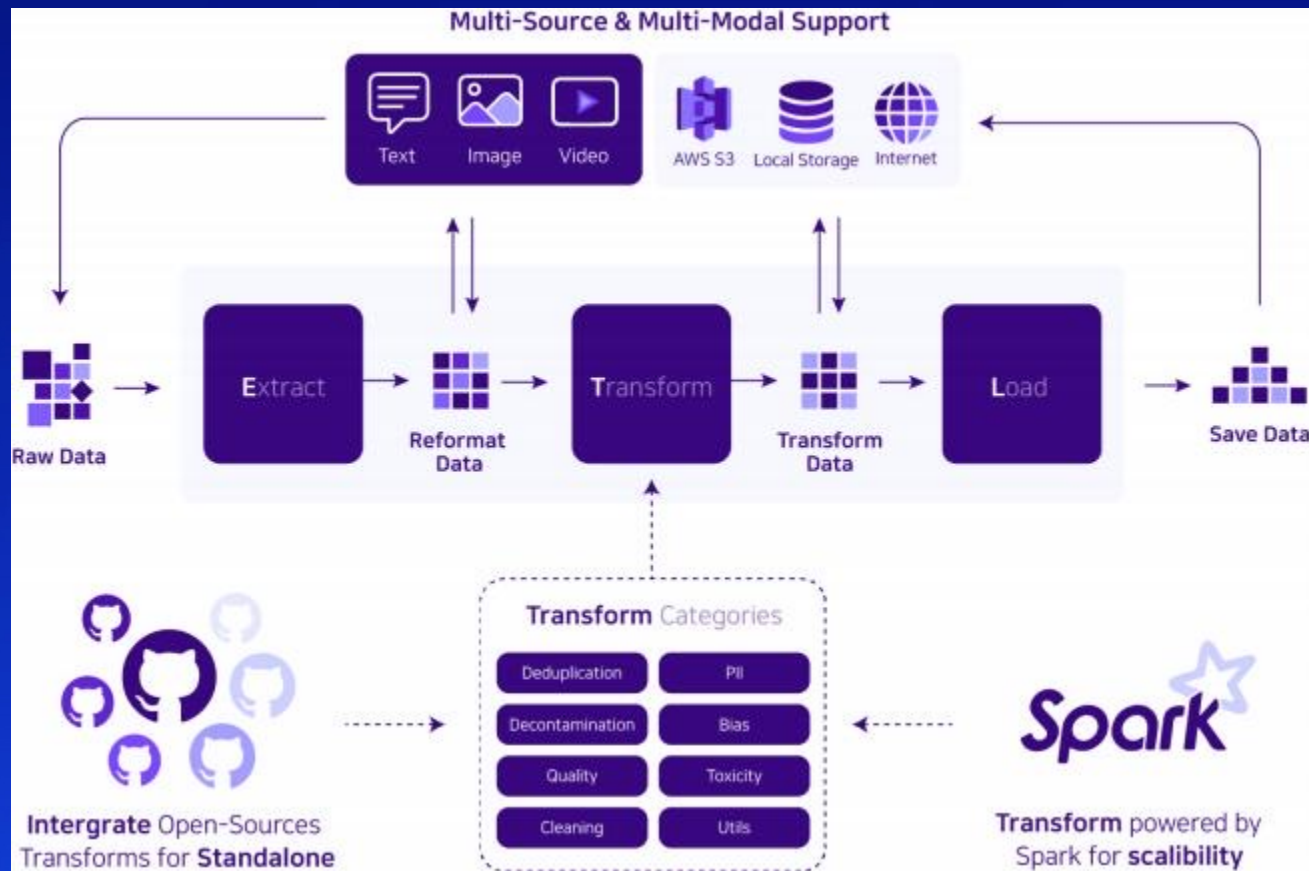
- ① 下载数据集
- ② 将数据集处理为模型输入，并保存到本地
- ③ 准备训练

Legacy training paradigm

这种范式下ETL与模型训练完全串行，是一种简单明了的方式。



▶ What 's the Problem



多模态大模型的ETL流程正变得越来越复杂

- E: 数据模态多，来源复杂，拉取时间长
- T: 数据处理流程复杂
- L: 存储占用高

▶ What 's the Problem

多模态数据由于版权和存储原因，大多以下载链接的形式分发，获取速率受到限制

videoid	name	page_dir	contentUrl
21179416	Aerial shot winter forest	006001_006050	https://ak.picdn.net/shutterstock/videos/21179416/preview/stock-footage-21179416-006001-006050.mp4
5629184	Senior couple looking through binoculars on sailboat together. shot on red epic for high quality 4k, uhd, ultra hd resolution.	071501_071550	https://ak.picdn.net/shutterstock/videos/5629184/preview/stock-footage-5629184-071501-071550.mp4
1063125190	A beautiful cookie with oranges lies on a green tablecloth	133701_133750	https://ak.picdn.net/shutterstock/videos/1063125190/preview/stock-footage-1063125190-133701-133750.mp4
1039695998	Japanese highrise office skyscrapers tokyo square	011901_011950	https://ak.picdn.net/shutterstock/videos/1039695998/preview/stock-footage-1039695998-011901-011950.mp4
9607838	Zrenjanin,serbia march 21 2015: fans watching live concert bokeh blur urban background 1920x1080 full hd footage	121551_121600	https://ak.picdn.net/shutterstock/videos/9607838/preview/stock-footage-9607838-121551-121600.mp4
21157780	Young beautiful woman using smartphone in cafe	136051_136100	https://ak.picdn.net/shutterstock/videos/21157780/preview/stock-footage-21157780-136051-136100.mp4
1016180245	3d render of inky injections into water with luma matte. blue ink on white background 5	048301_048350	https://ak.picdn.net/shutterstock/videos/1016180245/preview/stock-footage-1016180245-048301-048350.mp4
19107358	Swimming in the pool ,slow motion 120 fps,handheld camera balanced steady shot	055351_055400	https://ak.picdn.net/shutterstock/videos/19107358/preview/stock-footage-19107358-055351-055400.mp4
1036840850	Circa 1940s usa: woman posing for camera next to vintage car	041501_041550	https://ak.picdn.net/shutterstock/videos/1036840850/preview/stock-footage-1036840850-041501-041550.mp4
13347398	Timor-leste scanned by software	101801_101850	https://ak.picdn.net/shutterstock/videos/13347398/preview/stock-footage-13347398-101801-101850.mp4
1026436784	Fabulous dark background of water in the river	073951_074000	https://ak.picdn.net/shutterstock/videos/1026436784/preview/stock-footage-1026436784-073951-074000.mp4
9000853	New-york city with ravenwood in western part of borough of queens, queensboro bridge and manhattan	074501_074550	https://ak.picdn.net/shutterstock/videos/9000853/preview/stock-footage-9000853-074501-074550.mp4

webvid以url形式提供，共包括10730233条数据

▶▶ What 's the Problem

处理流程复杂耗时，甚至超过训练开销



GPT-4V (20s/it)

An aerial video sweeps over a majestic ocean cliff with striated rock formations in rich hues of red, green, and orange. The sun's rays enhance the colorful palette of the landscape, while the sea's azure waters gently crash against the cliff's base. Visible are the textured details of the cliff face and the contrast of the aerial algae and sea grass coating parts of the rock. Seabirds can be seen flying close to the water around the rocky outcrop. The video conveys a serene yet dynamic coastal scenery, highlighting the natural beauty and geological diversity of a rugged coastline.

LLaVA-1.6-Yi-34B (3s/it)

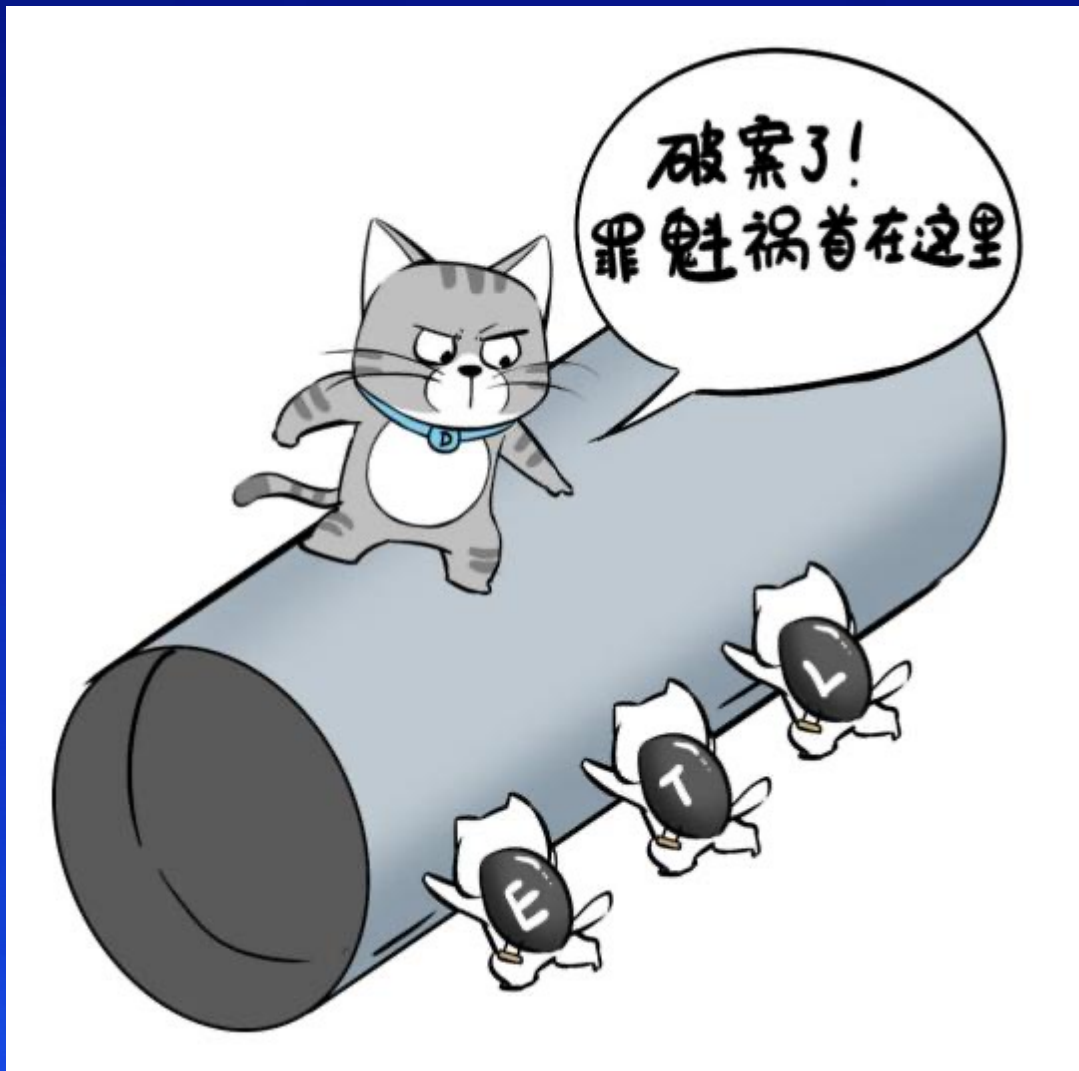
A breathtaking aerial view of a rocky cliff jutting out into the ocean. The cliff, adorned with vibrant green moss and patches of yellow and red lichen, is bathed in the warm glow of the sun. The ocean, a deep blue, is dotted with small white waves that crash against the cliff's base. The perspective of the video is from above, providing a bird's eye view of the cliff and the surrounding ocean. The sun's rays illuminate the scene, casting a beautiful light on the cliff and the ocean. The video is a stunning representation of the raw beauty of nature.

Storage plays an important role in AI training, and yet is one of the least talked-about aspects. As the GenAI training jobs become more multimodal over time, consuming large amounts of **image, video, and text data**, the need for data storage grows rapidly.¹

- 要从原始数据中筛选出一亿分钟数据，可能意味着原始数据量高达数十PB以上
- 对于一般的小型数据中心，没有能力搭建适应视频预训练的存储设施。

1. [Building Meta's GenAI Infrastructure](#)

►► What 's the Problem



- 数据来源复杂
- 数据源不能立即被拉取
- 数据处理流程复杂
- 数据处理和模型训练耦合
- 数据量过大，无法一次性处理
- ...

▶▶ What 's the Problem

- 数据流离模型训练越来越远
- 仍然使用传统的方式处理数据，数据流将成为阻塞训练的瓶颈。



How to train Just training on
on internet-scale the internet!
data? Streamingly

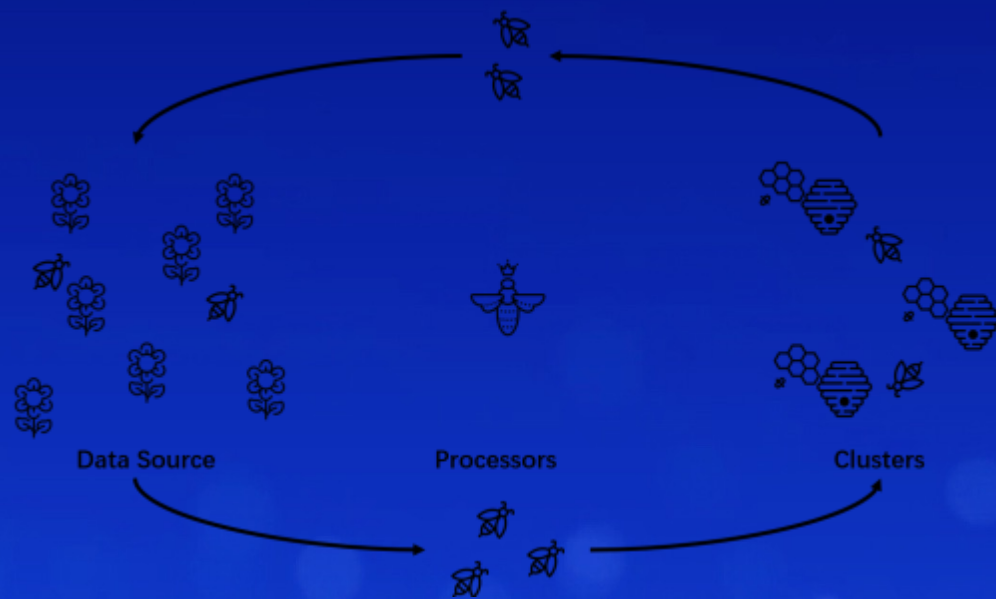
Streaming Data Flow

▶ Streaming to the rescue

流式传输数据可以解决这些问题

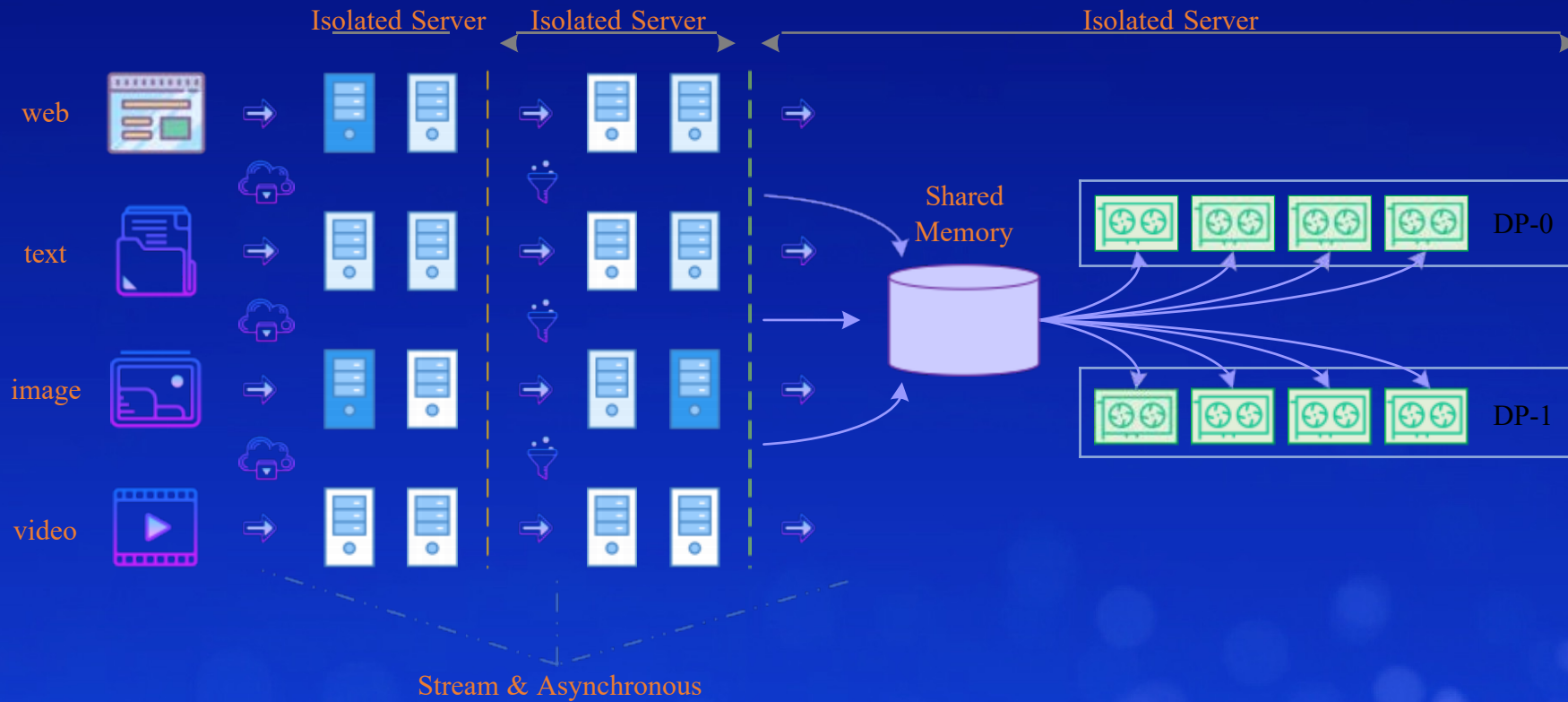


流式地传输数据只是个开始

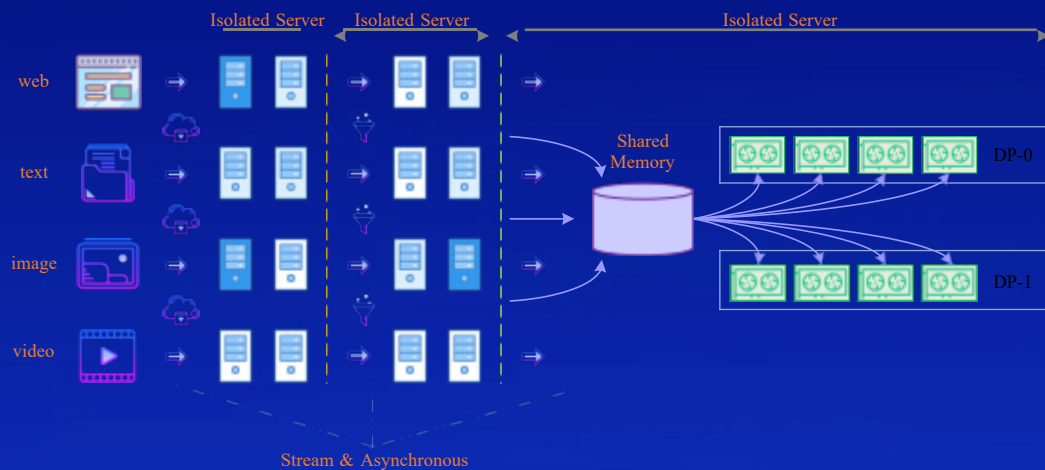


我们需要完全基于流式处理范式构建数据基础设施

▶▶ Streaming to the rescue



▶ Streaming to the rescue



W 零启动开销

✓ 数据处理进程和模型训练进程完全分离

W 节点内通过 **SharedMemory** 通信, 节点间通过内存数据库通信

✓ 数据处理集群拓扑与GPU拓扑无关, 可以动态调整

✓ 定时sink数据库, 允许回溯数据流

✓ 确定性的数据切分和洗牌算法, 确保回溯的一致性

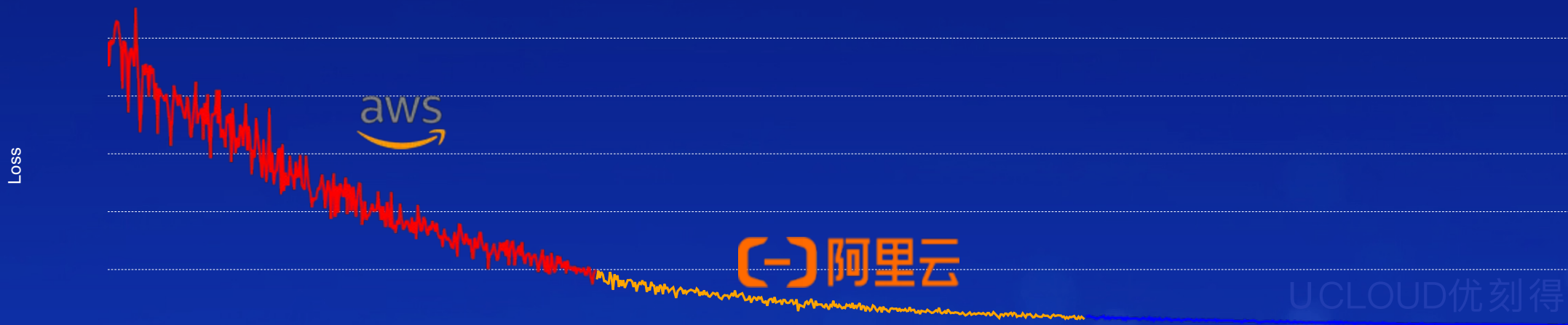
Data Shards (stored in cloud)



▶▶ Training on the internet

Data Shards (stored in cloud)

利用不同云厂商资源作为数据和权重的存储后端，无缝进行不同规模的云迁移



▶▶ Training on the internet

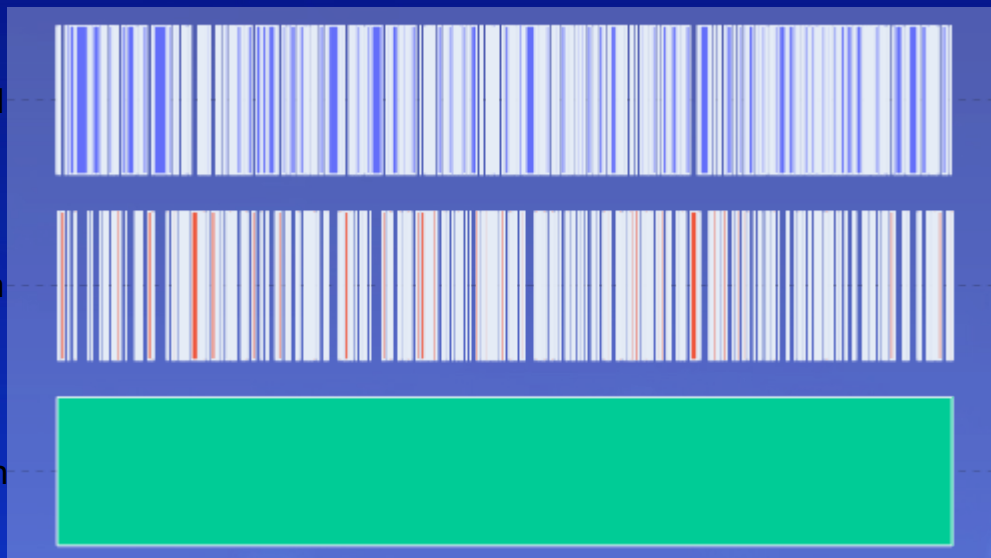
- 进一步分离了数据处理和模型训练
- 使ETL与模型训练完全并行



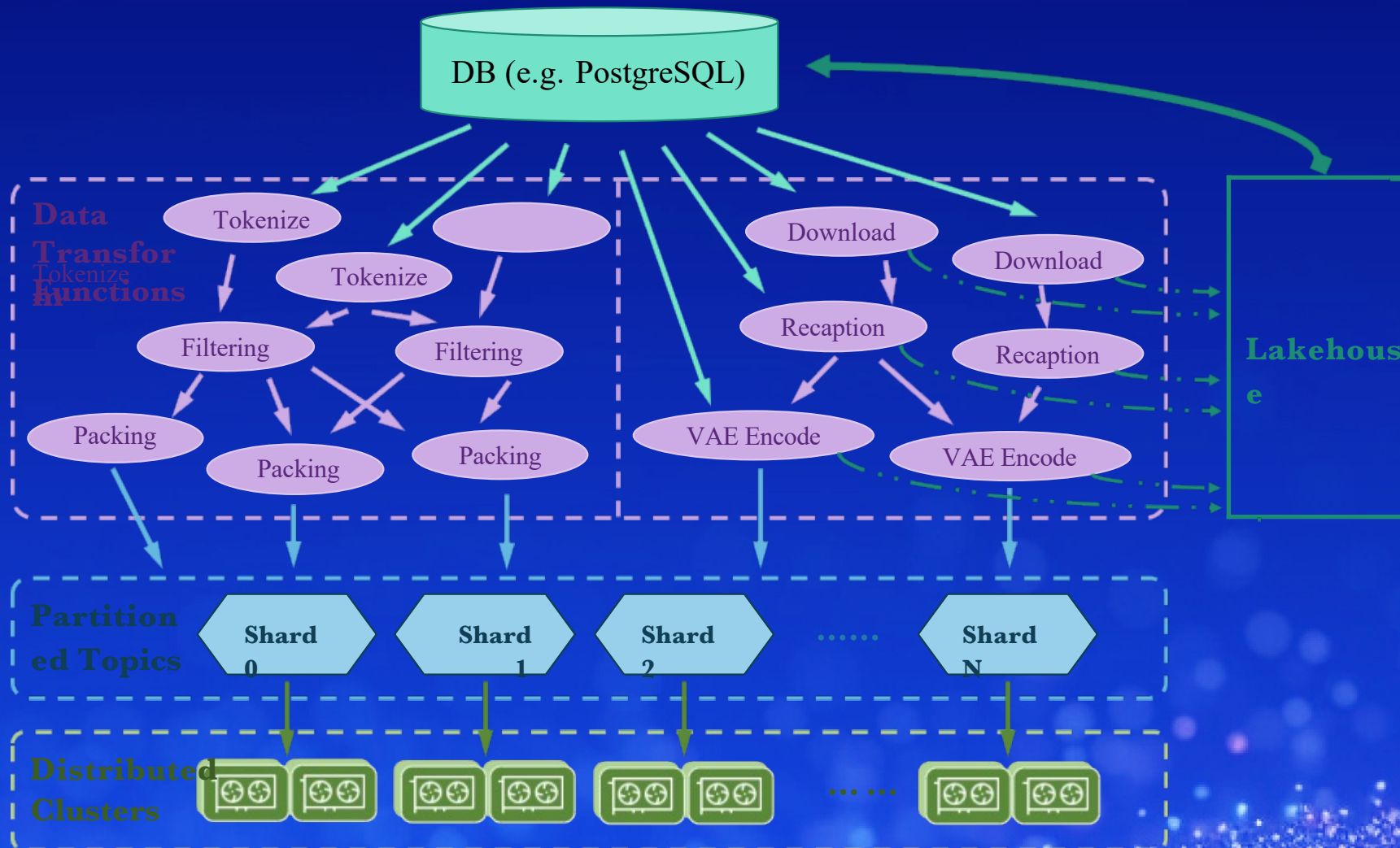
Read

Transform

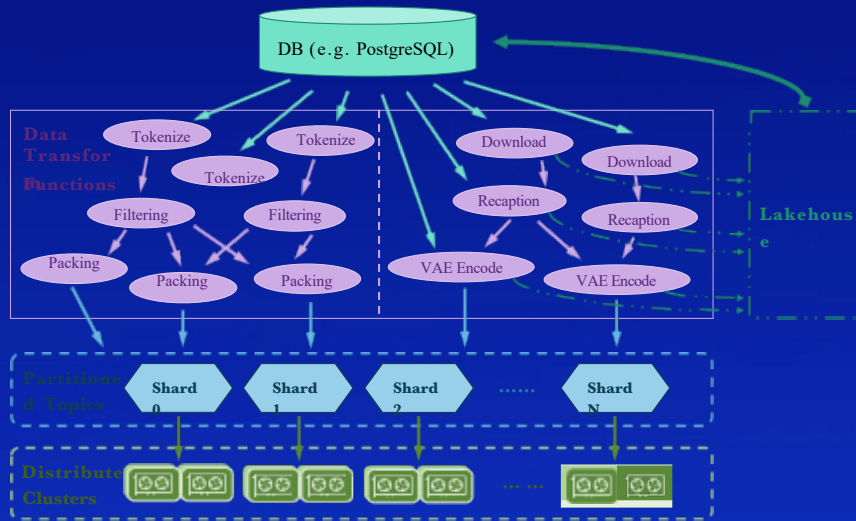
Train



▶▶ Streaming powers new era of DL training



▶ Streaming powers new era of DL training



- PostgreSQL manages massive data
- Pulsar defines diverse processing operations
- Lakehouse buffers intermediate results for reusing
- Partitioned Topics ensures distributed data consistency

▶▶ Streaming powers new era of DL training

mlflow

Ww&B

Tensorboard



PyTorch

Tensorflow



RAY

kafka

RabbitMQ

ULSAR

RockerMQ

APACHE
spark™



beam

Polars

pandas

RAY

DORIS

rr dask



ALLUXIO

hadoop

MINIO

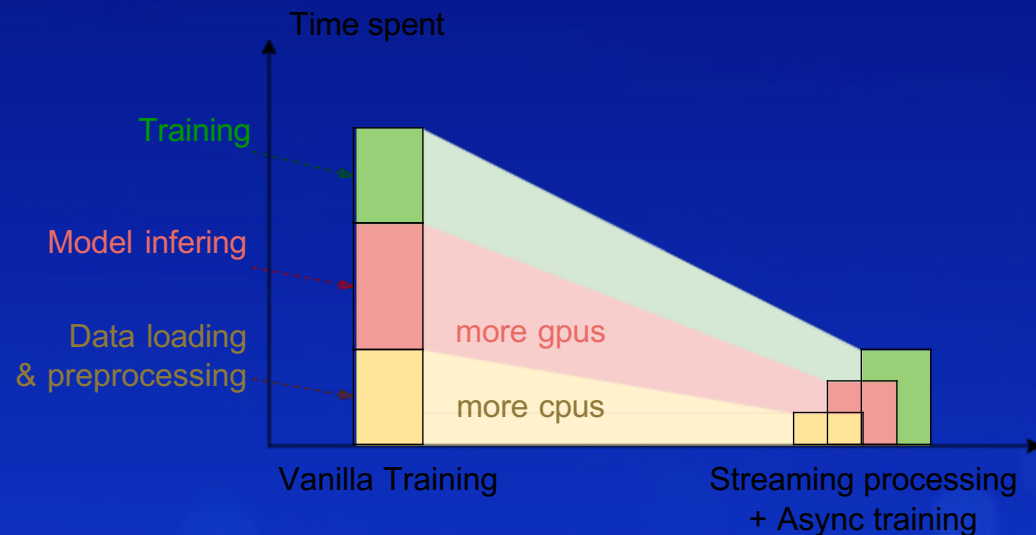
aws | s3



▶▶ Higt efficiency experimentally verified

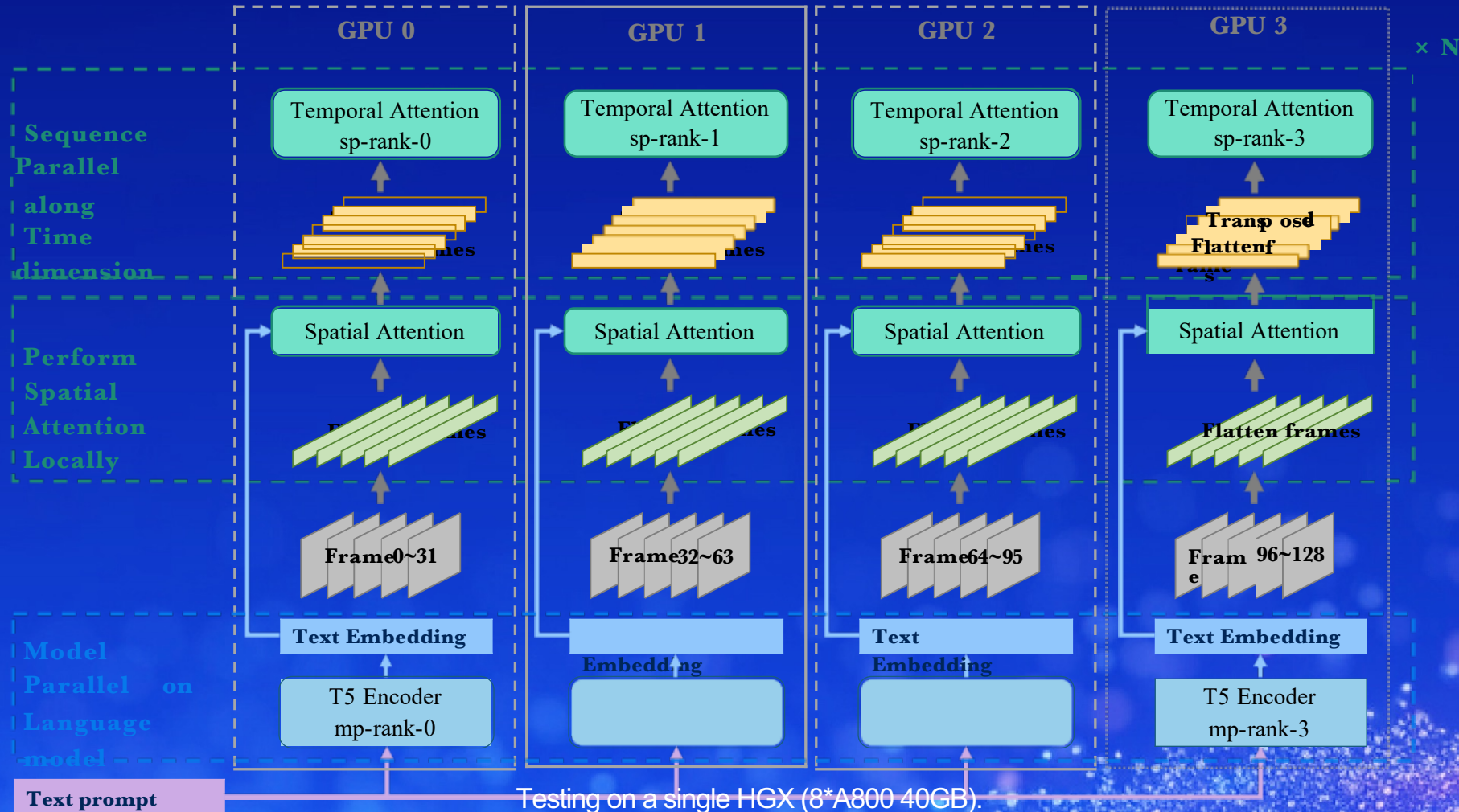
流式数据管线助力多模态训练: A practice of SD training

- ✓ 各部分解耦部署、流水运行
 - CPU任务: (远端) 数据获取与处理
 - GPU推理: VAE、Text encoder、VLM
 - GPU训练: UNet or DiT
- ✓ 原生支持不同部分各自**针对性优化**
- ✓ 资源按需**动态调配**，提速关键瓶颈
- ✓ 充分利用各节点本地存储 (NVMe)
提升流水线传输效率



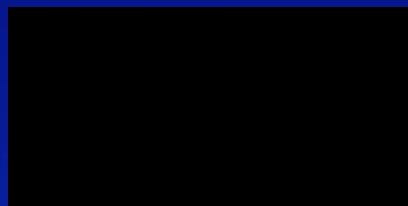
Overall speedup > 2x

Heterogeneous parallel topology boosts low-resource efficient training

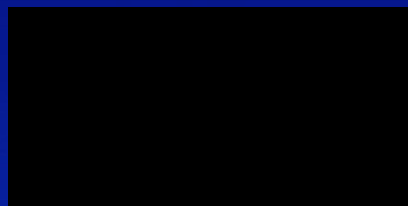


▶▶ Scaling exact attention to ultra long sequence

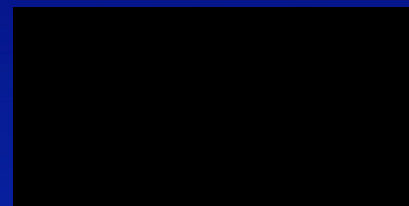
16 → 32 → 64 → 128 frames



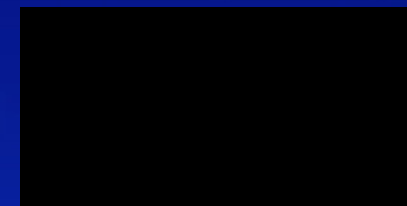
Fireworks exploding in the sky



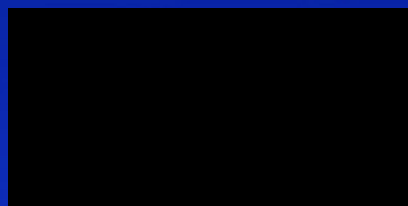
Fishes swimming in ocean camera moving



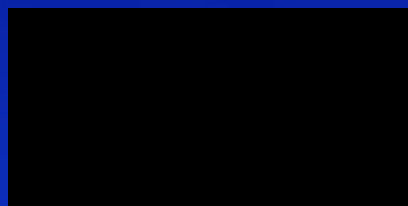
Fly to a mansion in a tropical forest viewpoint moving from down to up



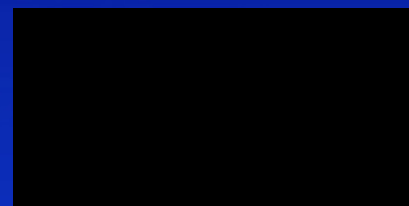
Flying above a breathtaking limestone structure



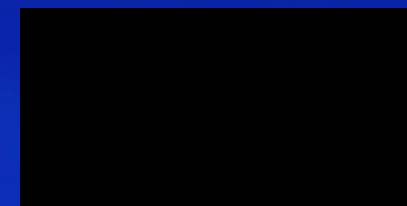
Petals Auttering in the wind, slow motion.



Santa dancing



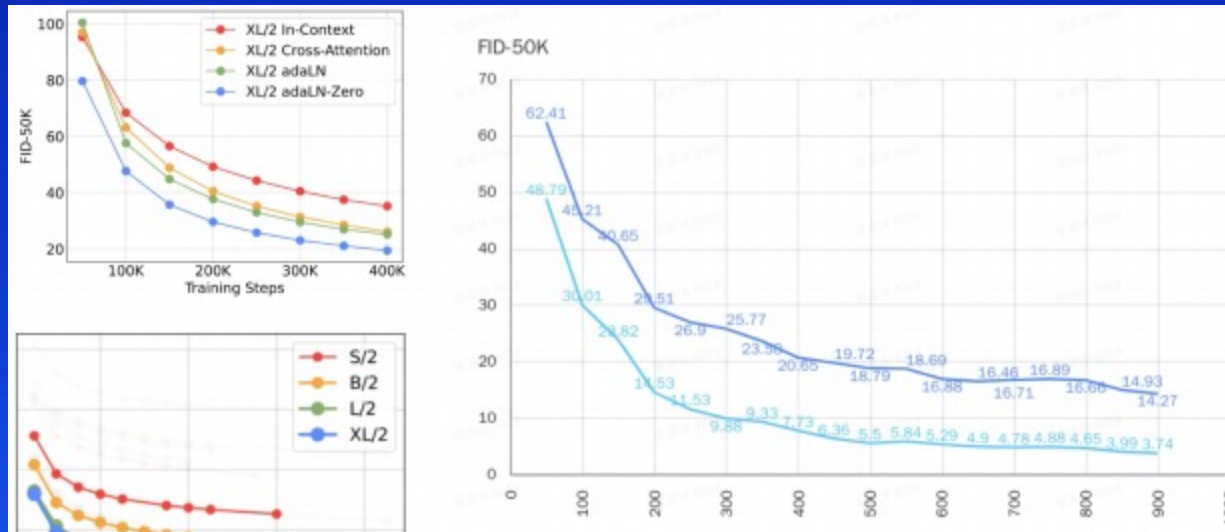
Snow fallen in the forest pine tree are coverd with snow



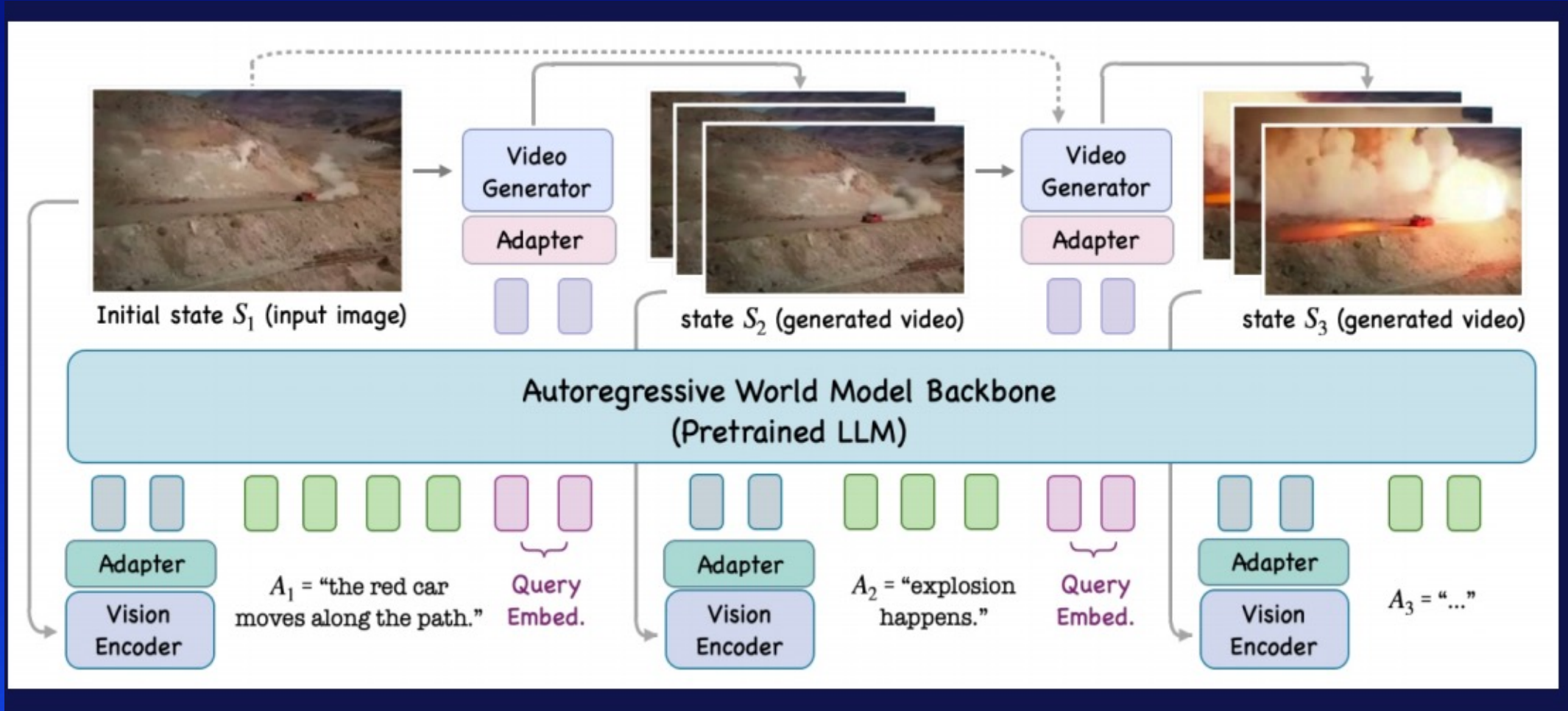
Drone view of waves crashing against the rugged cliffs along Big Sur's garay point beach.

Efficiently reproduce DiT pretraining process

- Completely pre-trained from scratch
- Achieve outstanding performance using low-resource
- Easily adapted to Text2Image and Any2Image
- 2 * A800 nodes using 4 days (67 GPU days)



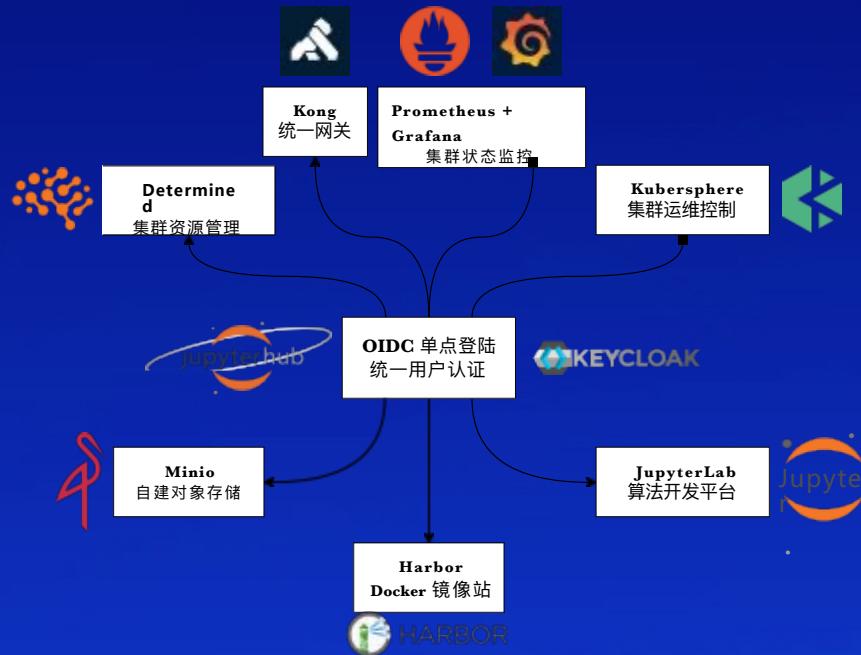
► Modeling world using Any2Any Model



▶▶ Cloud-native AI Ops as the foundation of all niDD AI+研发数字峰会



▶ Cloud-native AI Ops as the foundation of all nDDD AI+研发数字峰会 AI+ Development Digital summit



All in k8s

- 资源调度平台：
 - 完备用户权限机制，包含资源配额、命名空间隔离等
 - 高效利用计算资源
- JupyterLab算法开发平台：
 - 无缝集成资源调度平台
 - 支持分布式场景交互式开发
- 统一网关：
 - 为所有组件提供统一的访问入口 (<https://xxx.bk8s>)
- 更多组件：
 - 资源运维控制、私有镜像站、自建对象存储.....

科技生态圈峰会 + 深度研习



—1000+ 技术团队的选择



 **K+峰会**  **敦煌站**

K+ 思考周®研习社

时间: 2025.08.29-30

 **K+峰会**  **上海站**

K+ 金融专场

时间: 2025.10.17-18

 **K+峰会**  **香港站**

K+ 思考周®研习社

时间: 2025.11.25-26



K+峰会详情



 **AiDD峰会**  **上海站**

AI+研发数字峰会

时间: 2025.05.17-18

 **AiDD峰会**  **北京站**

AI+研发数字峰会

时间: 2025.08.08-09

 **AiDD峰会**  **深圳站**

AI+研发数字峰会

时间: 2025.11.28-29



AiDD峰会详情



利用AI技术深化计算机对现实世界的理解

推动研发进入智能化时代

