AI+研发数字峰会
AI+ Development Digital summit
第5届

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 + 技术团队的共同选择

KEYLINK ing

K+峰会

- 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

- 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

演讲嘉宾

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。

Main Point

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

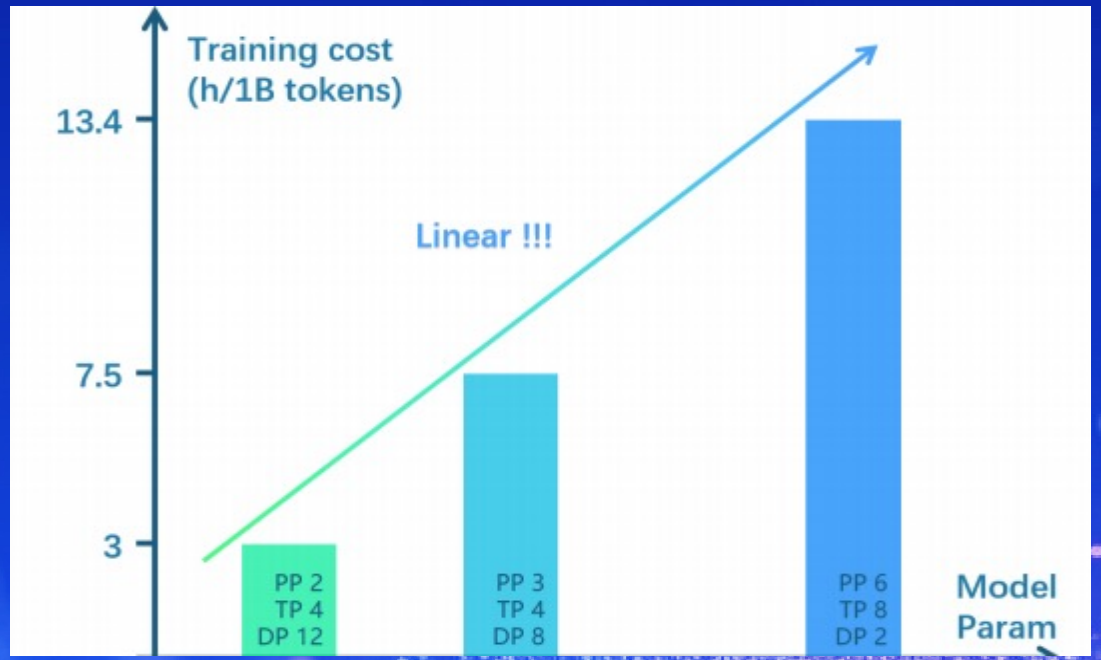
打磨一流大模型预训练与微调技术

- 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



Trained on A100 40G * 96, about 6 hours Completed during collaboration

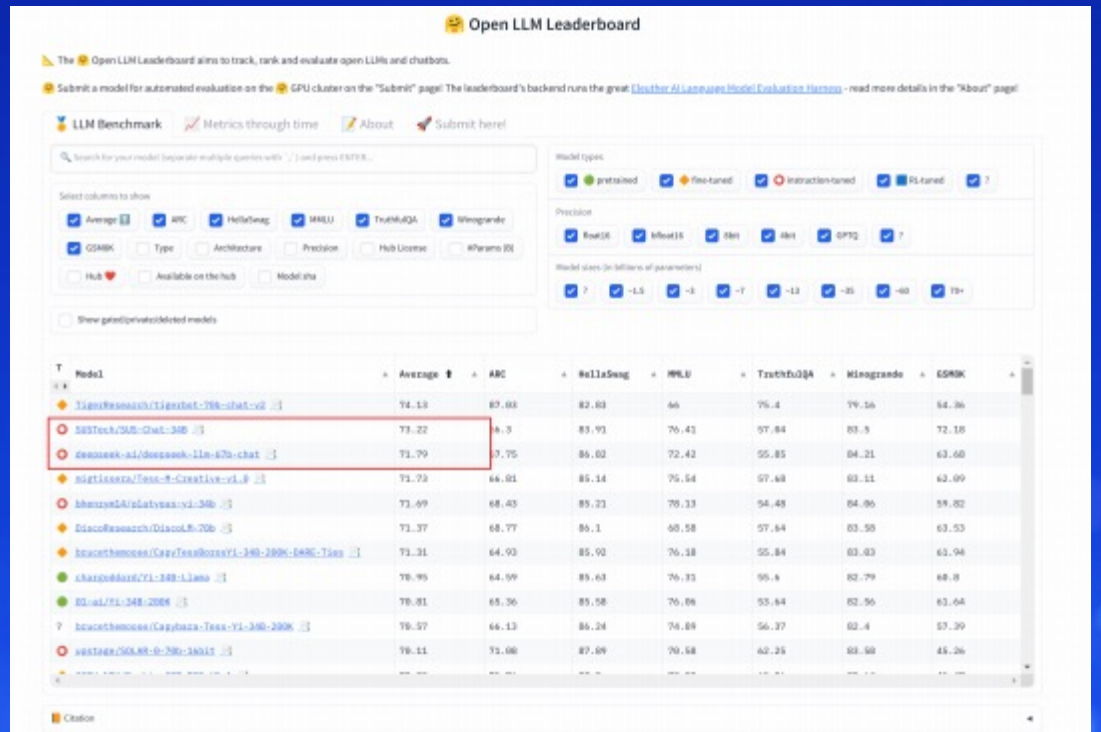
国内外领先的通用语言模型

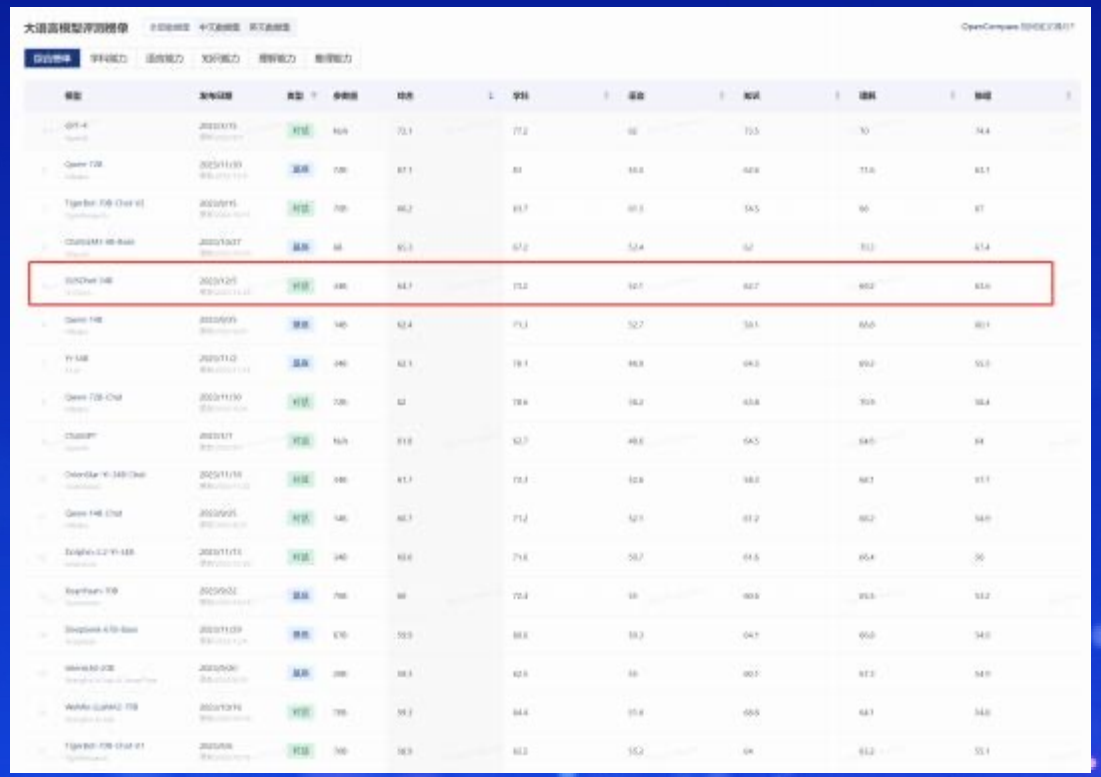
发布时，在OpenLLMLeaderboard40,000+的模型中排名第一，

OpenCompass中Chat模型排名第一

OpenCompass (1227) 👉

Open LLM Leaderboard (1205) 👇





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	76.57	35.9	72.63
Deepseek-67b-Chat	74.45	29.56	69.73
SUS-Chat-80.06	80.06	28.8	67.62

Effcient distributed training infra

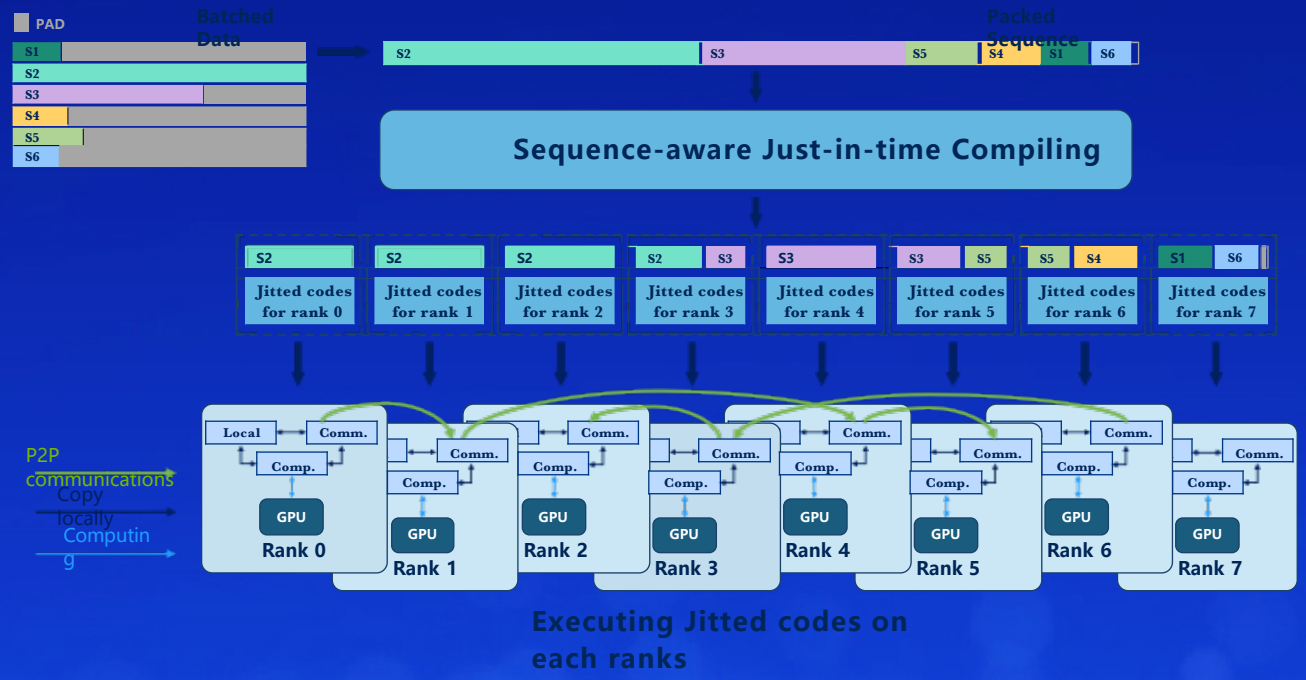
标准场景下，分布式训练框架在各指标和特性上超越世界主流框架

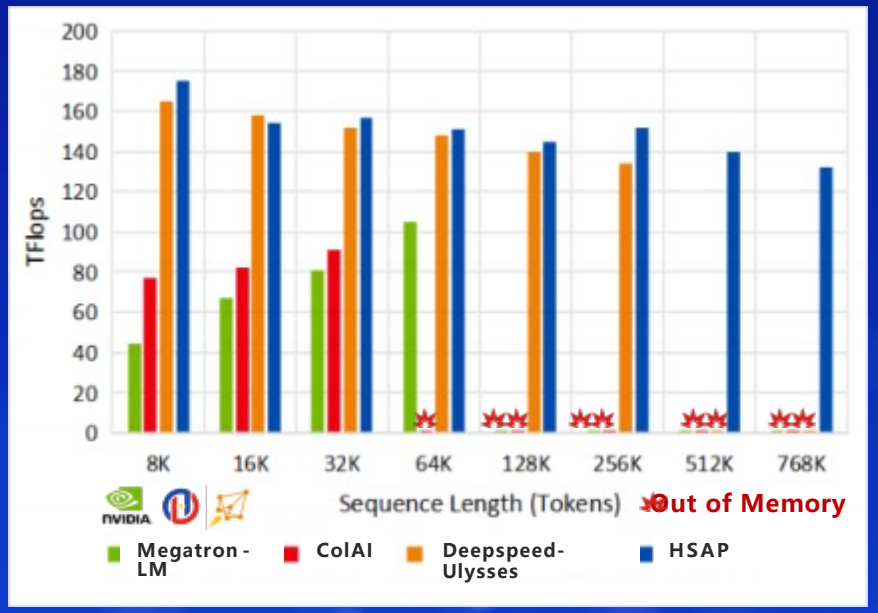
- ☑ 唯一支持6D并行的分布式训练框架
- ☑ 高效训练算子适配与优化
- ☑ 无效Padding计算裁剪

	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





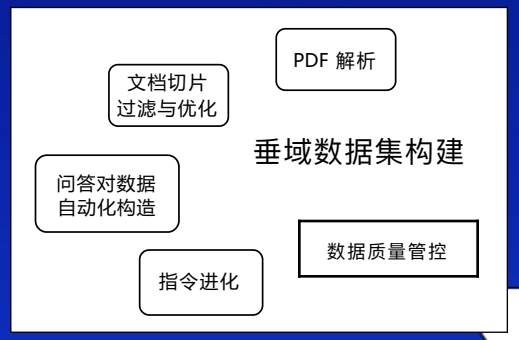
通过增加计算资源高效扩展至主流框架无法达到超长序列

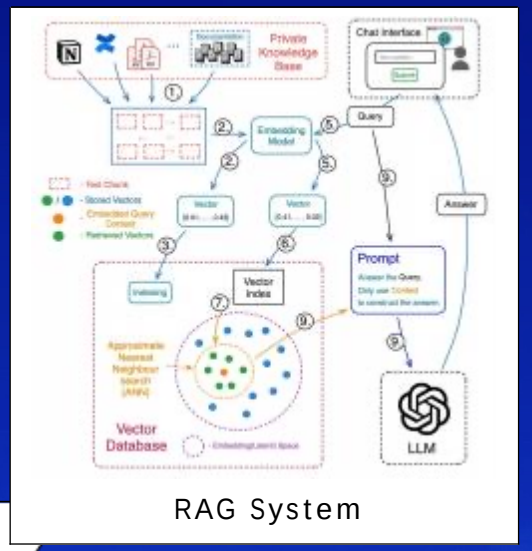
Testing on a single HGX (8*A100 80GB).

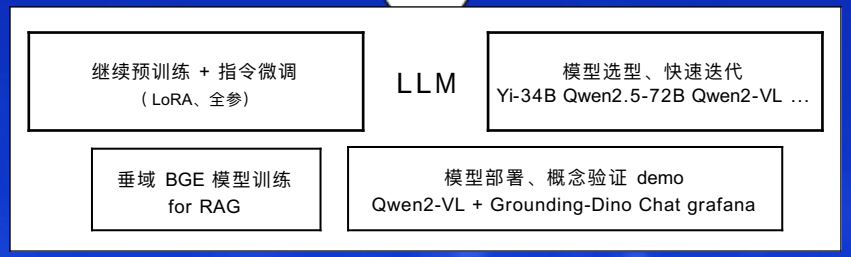
大模型训练技术赋能垂直领域应用

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 workAows

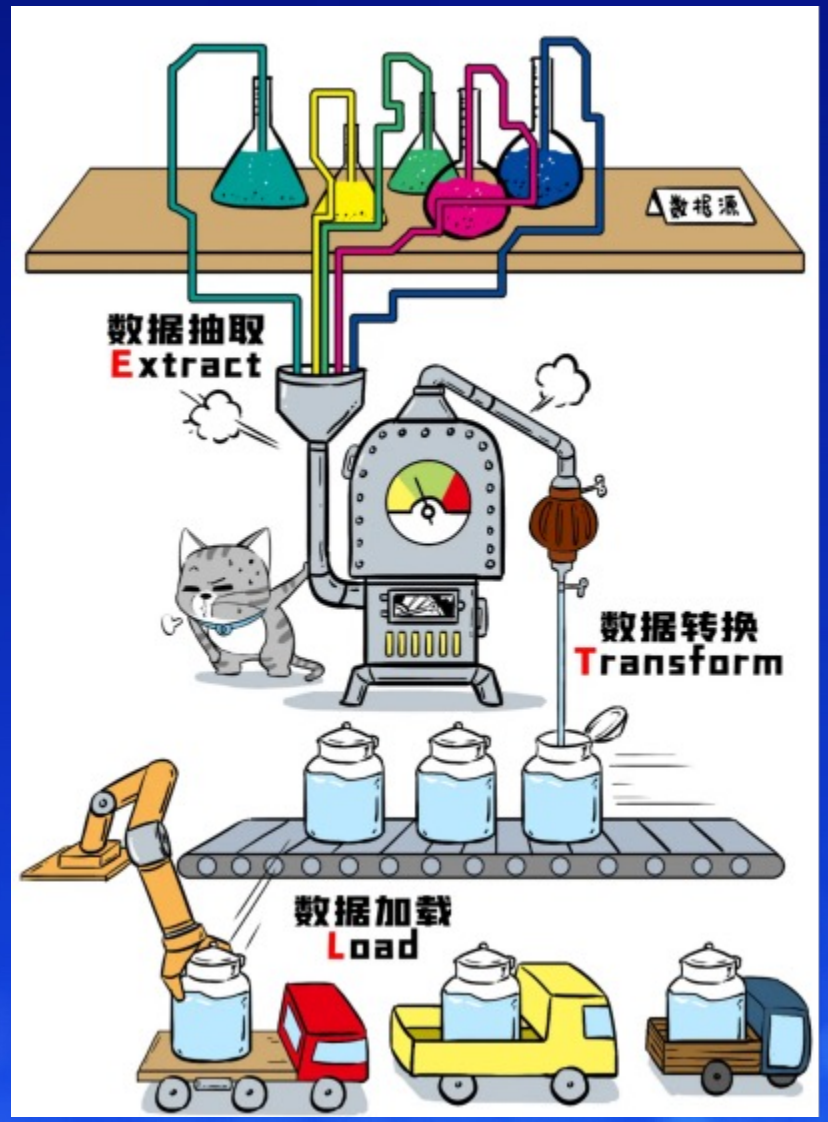
Legacy training paradigm

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。

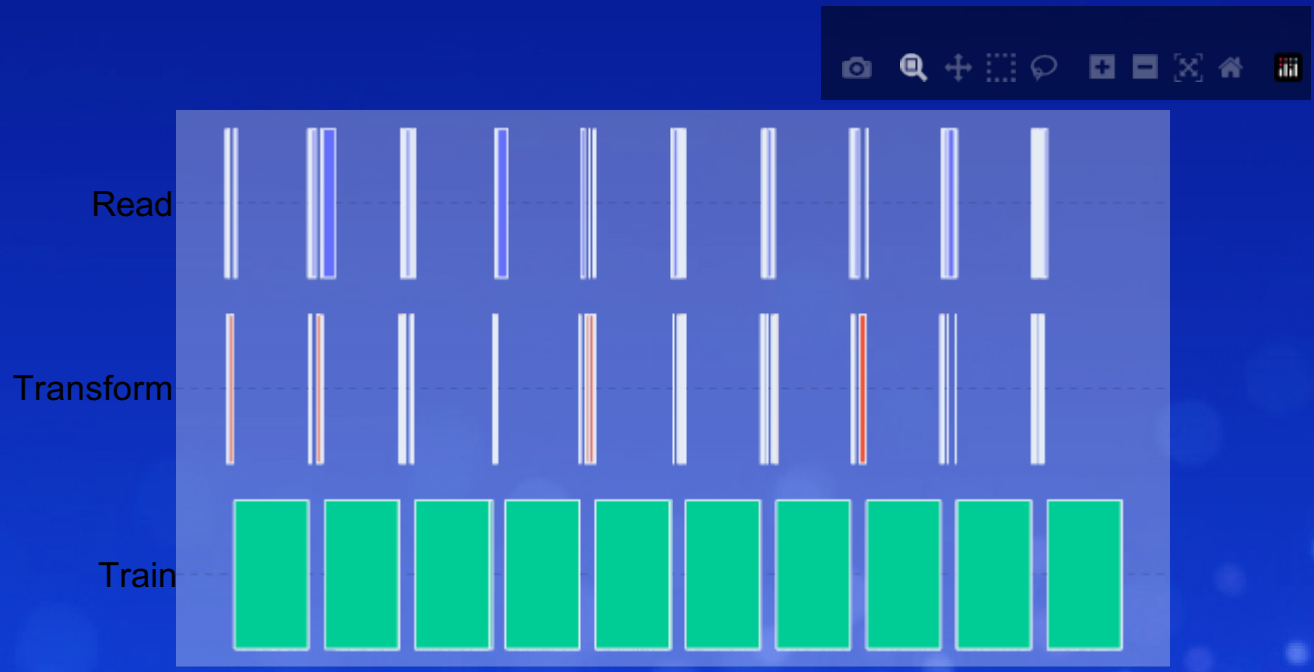
```
1
2import datasets
3from transformers import AutoTokenizer
4
5dataset = datasets.load_dataset(                                    1
6        "rotten_tomatoes",
7        split="train",
8 )
9tokenizer = AutoTokenizer.from_pretrained("bert-base-cased")
10
11dataset = datasets.map(                  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

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

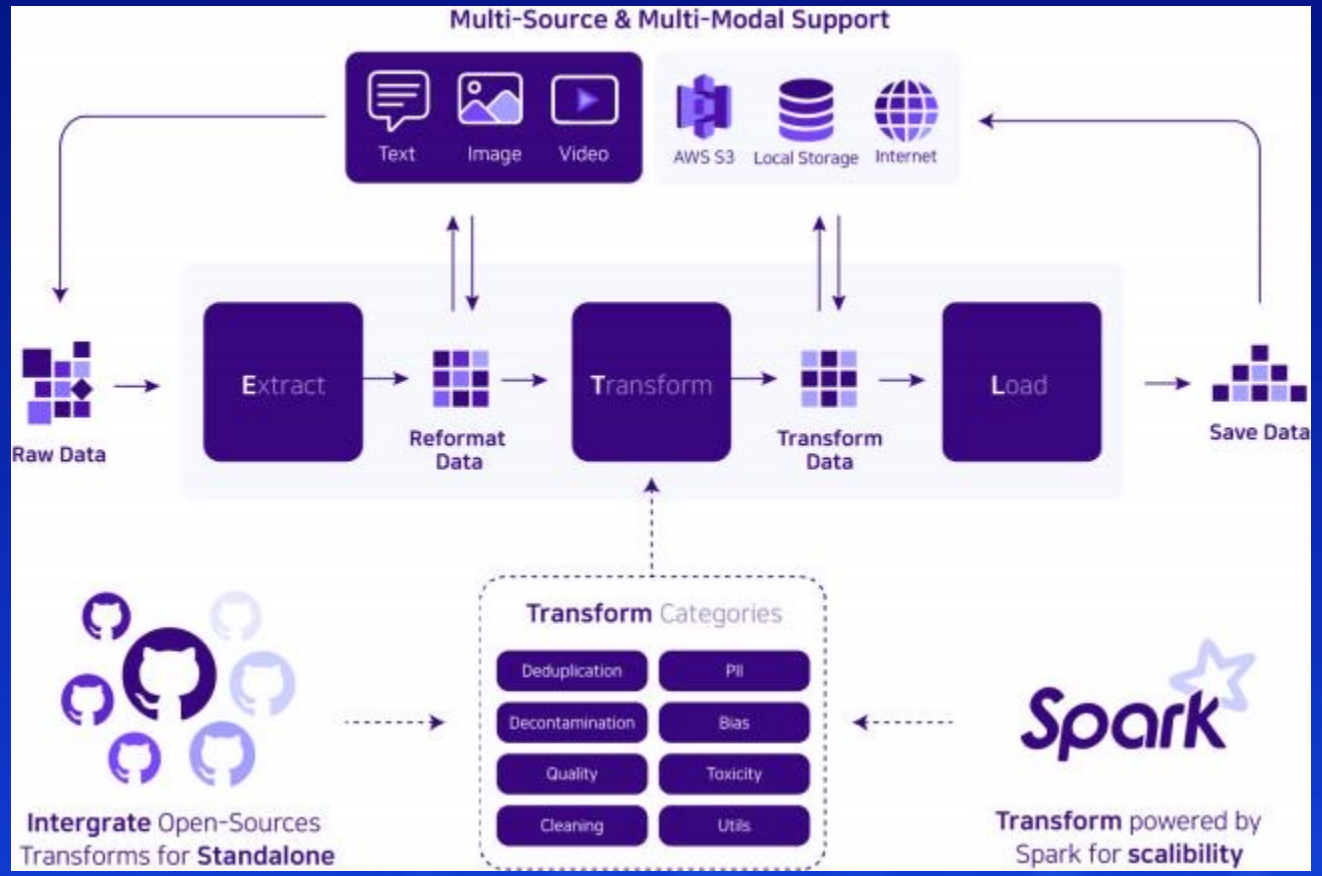
Legacy training paradigm



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/prev
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/st
1063125190	A beautiful cookie with oranges lies on a green tablecloth	133701_133750	https://ak.picdn.net/shutterstock/videos/1063125190/previe
1039695998	Japanese highrise office skyscrapers tokyo square	011901_011950	https://ak.picdn.net/shutterstock/videos/1039695998/previe
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/st
21157780	Young beautiful woman using smartphone in cafe	136051_136100	https://ak.picdn.net/shutterstock/videos/21157780/preview/s
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/previe
19107358	Swimming in the pool ,slow motion 120 fps,handheld camera balanced steady shot	055351_055400	https://ak.picdn.net/shutterstock/videos/19107358/preview/
1036840850	Circa 1940s usa: woman posing for camera next to vintage car	041501_041550	https://ak.picdn.net/shutterstock/videos/1036840850/previe
13347398	Timor-leste scanned by software	101801_101850	https://ak.picdn.net/shutterstock/videos/13347398/preview/
1026436784	Fabulous dark background of water in the river	073951_074000	https://ak.picdn.net/shutterstock/videos/1026436784/previe
9000853	New-york city with ravenswood in western part of borough of queens, queensboro bridge and manhattan	074501_074550	https://ak.picdn.net/shutterstock/videos/9000853/preview/st

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 clif with striatedrock formations in rich hues of red, green, and orange.The sun'srays enhance the colorful palette of the landscape, while the sea'sazure watersgently crash aqainst the clif'sbase. Visible are thetextured detais ofthe clifface and the contrast of the areeralaae and seaqrasscoating parts of the rock.Seabirds can beseen Aying close to the water around the rocky outcrop. Thevideo conveys a serene yet dynamic coastal scenery, highlightingthe natural beauty and geological diversity of a rugged coastline

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

A breathtakina aerial view of a rocky clif iutting out into theocean.The clif. adorned with vibrant qreen moss and patches ofyellow and red lichen, is bathed in the warm glow of the sun. Theocean, a deep blue, is dotted with small white waves that crashagainst the clif's base. The perspective of the video is fromabove, providing a bird's eye view ofthe clif and the surroundingocean.The sun's rays illuminate the scene, casting a beautifullight on the clif and the ocean. The video is a stunningrepresentation of the raw beauty of nature.

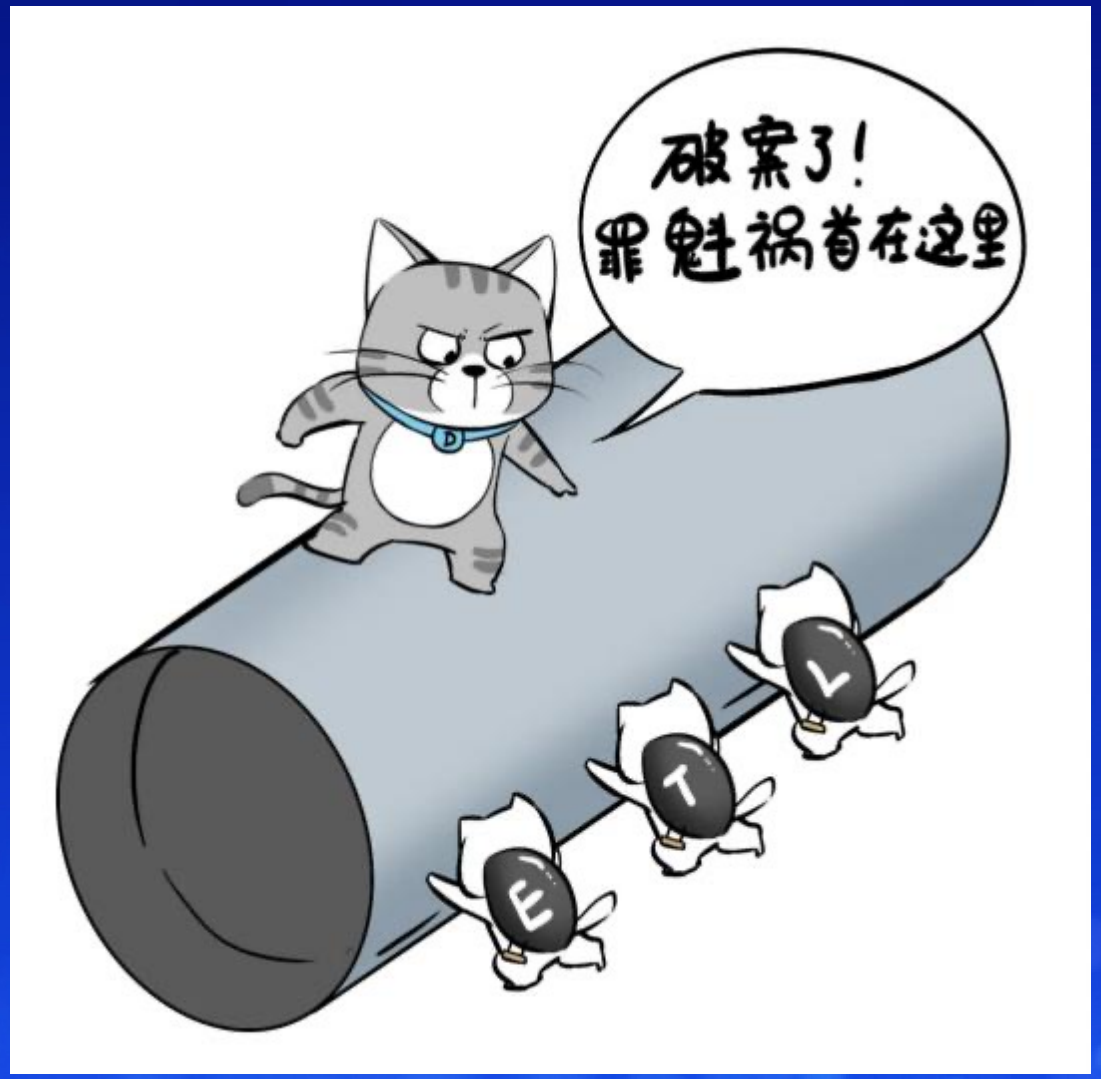
What 's the Problem

> 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

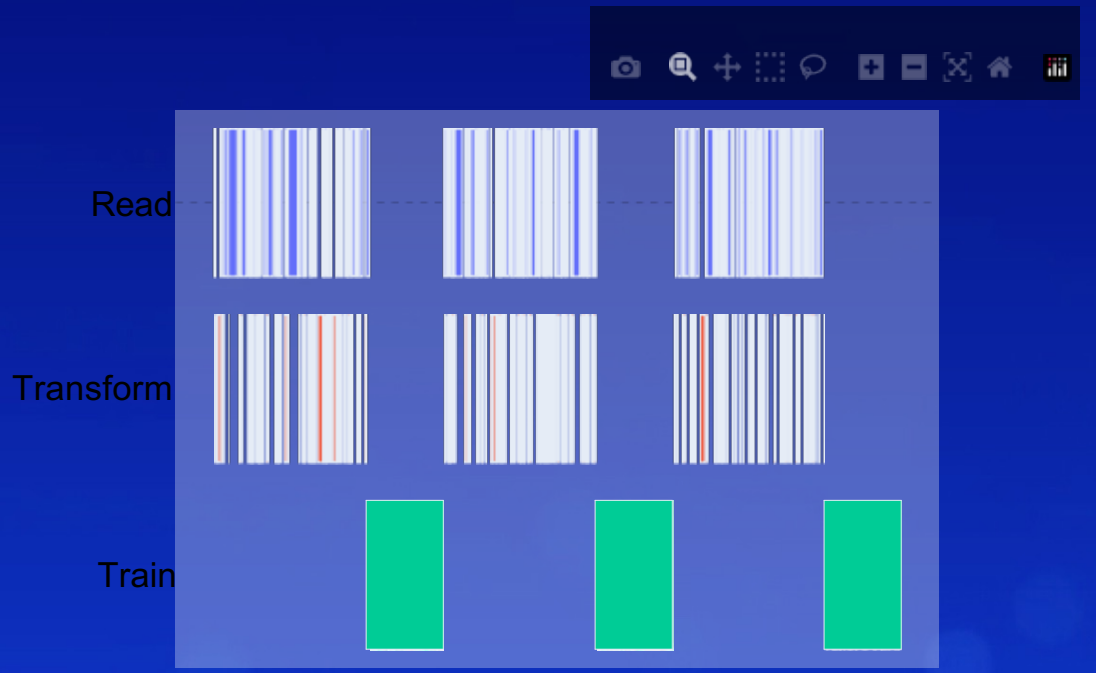
What 's the Problem



- 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 on internet-scale data?

Just training on the internet!

Streamingly

Streaming Data Flow

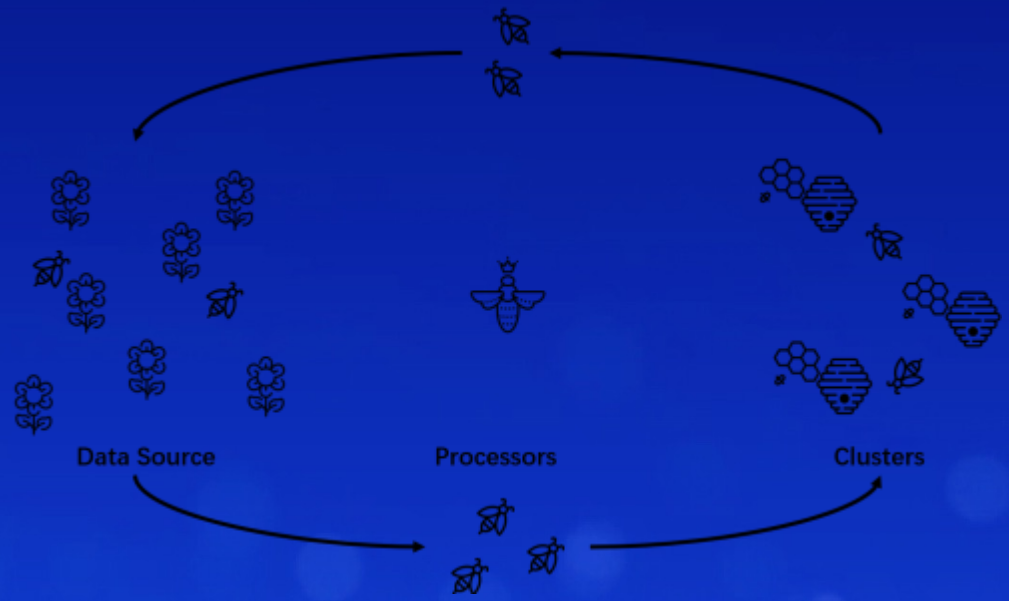
Streaming to the rescue

流式传输数据可以解决这些问题

c4 en dataset: 305GB

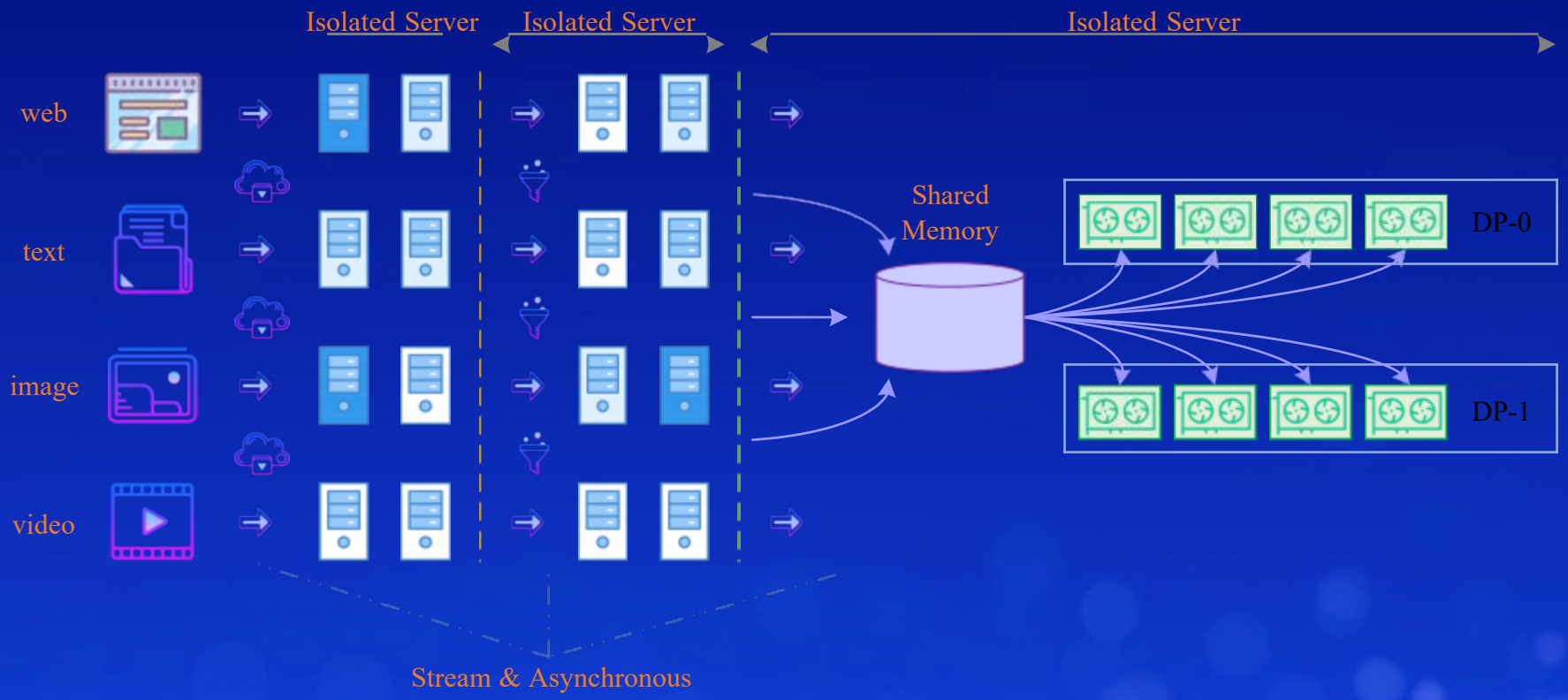
```
load_dataset("c4", "en", streaming=True)
```

流式地传输数据只是个开始

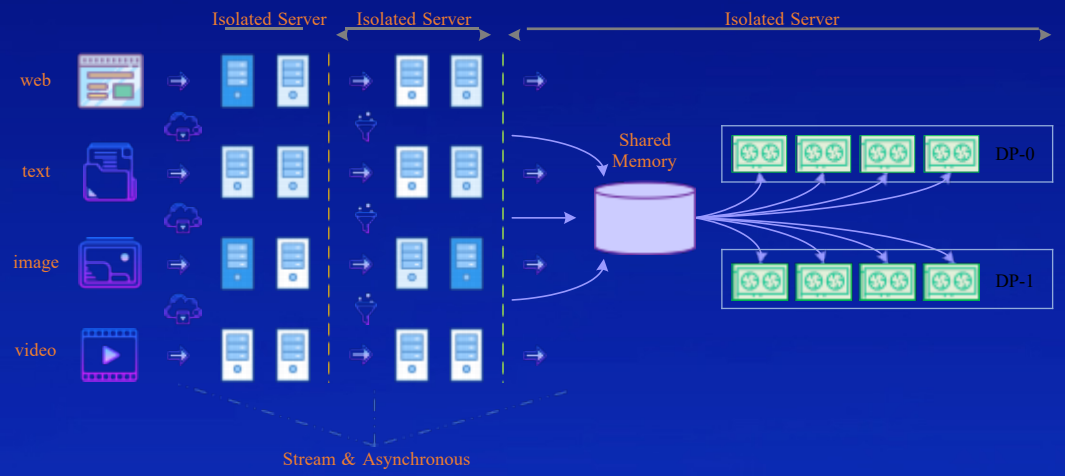


我们需要完全基于流式处理范式构建数据基础设施

Streaming to the rescue



Streaming to the rescue



- 零启动开销
- ☑ 数据处理进程和模型训练进程完全分离

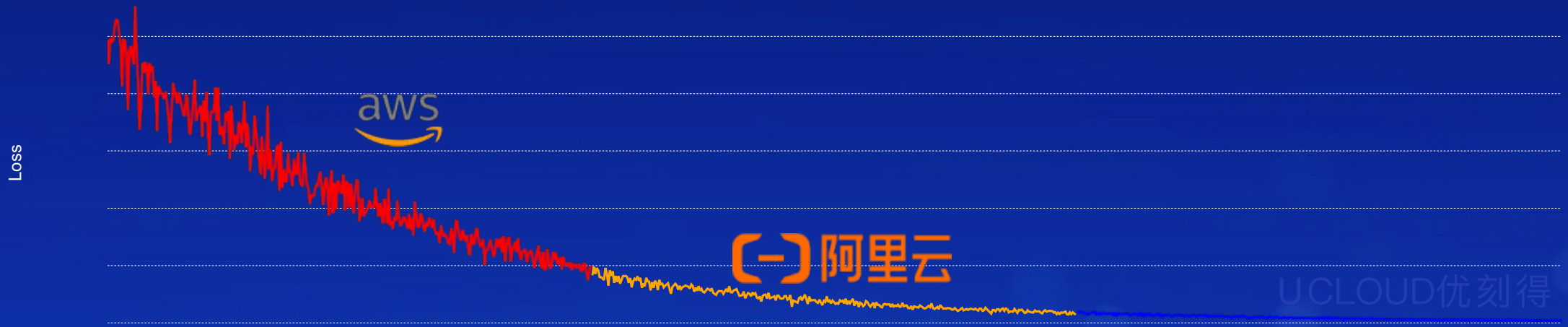
- 节点内通过SharedMemory通信，节点间通过内存数据库通信
- ☑ 数据处理集群拓扑与GPU拓扑无关，可以动态调整
- ☑ 定时sink数据库，允许回溯数据流
- ☑ 确定性的数据切分和洗牌算法，确保回溯的一致性

Data Shards (stored in cloud)

Training on the internet

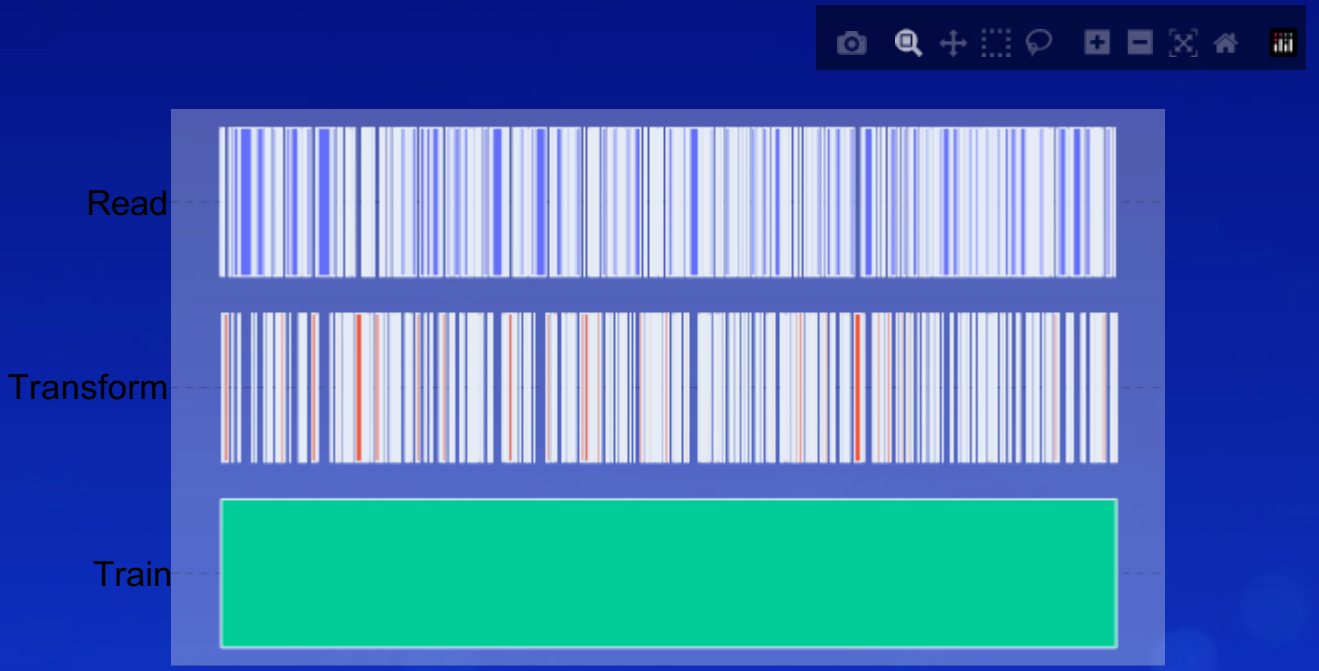
Data Shards (stored in cloud)

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

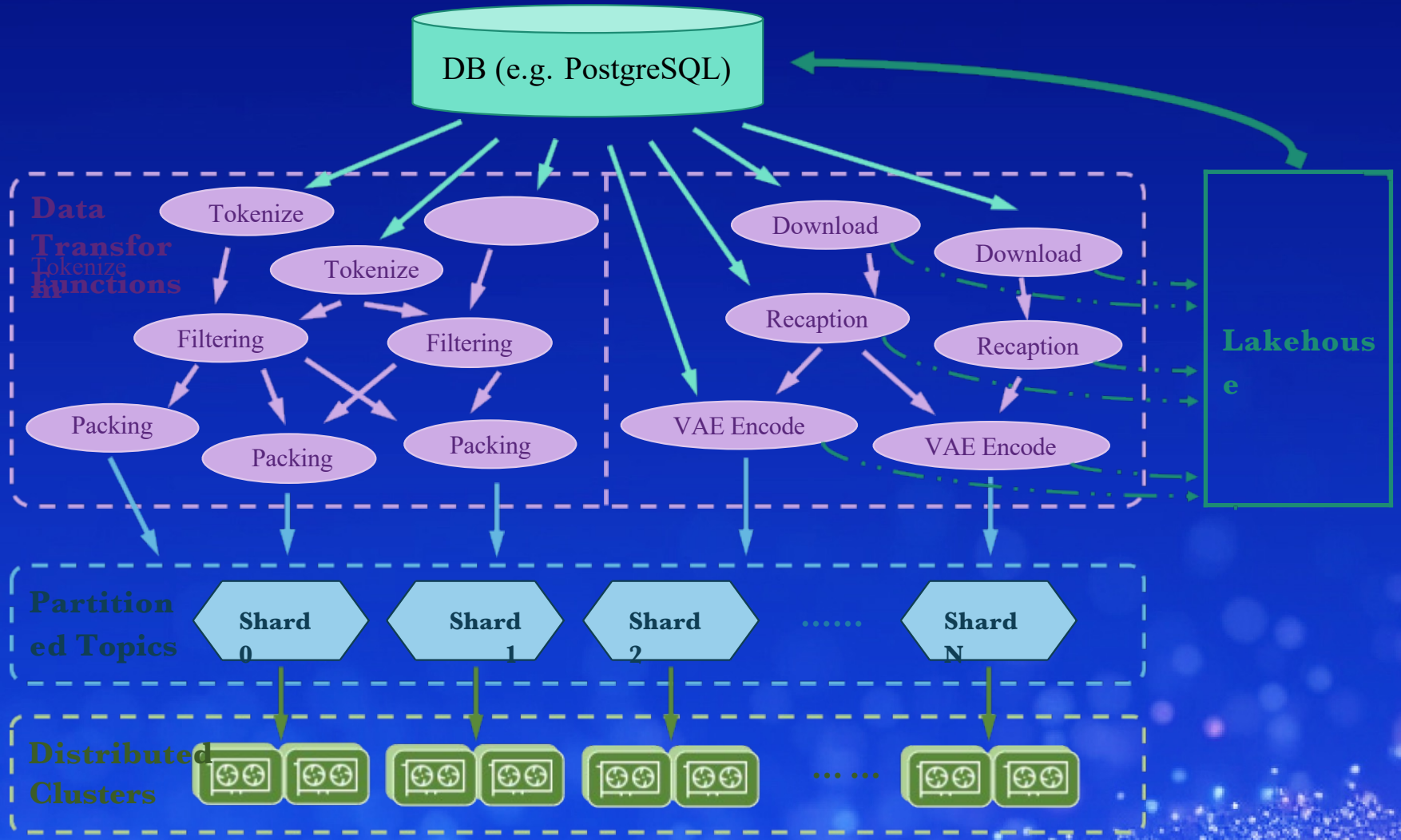


Training on the internet

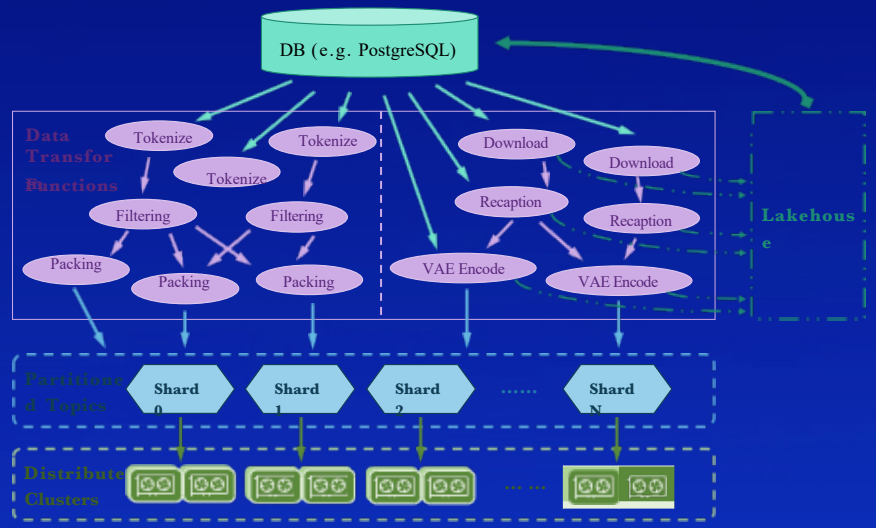
- 进一步分离了数据处理和模型训练
- 使ETL与模型训练完全并行



Streaming powers new era of DL training



Streaming powers new era of DL training



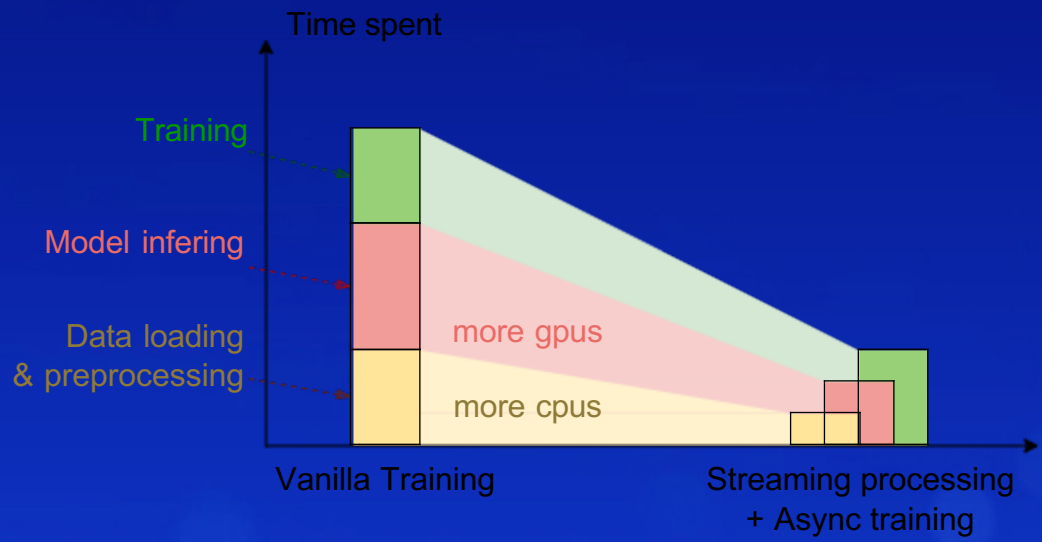
- **PostgreSQL** manages massive data
- **Pulsar** defines diverse processing operations
- **Lakehouse** bufers intermediate results for reusing
- **Partitioned Topics** ensures distributed data consistency

Streaming powers new era of DL training

Higt effciency experimentally verified

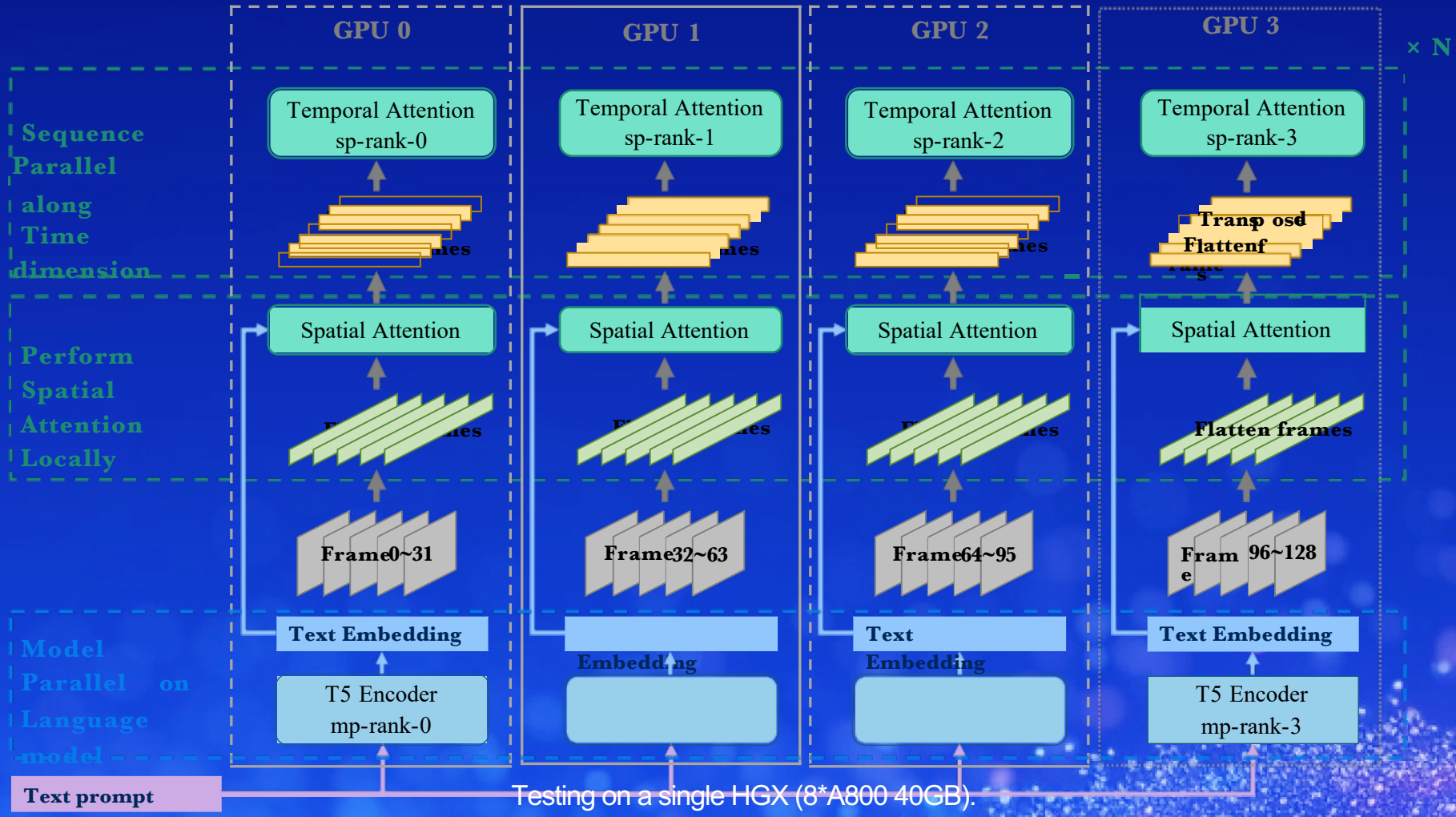
流式数据管线助力多模态训练: A practice of SD training

- ☑ 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）提升流水间传输效率



Overall speedup > 2x

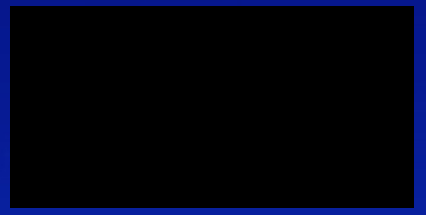
Heterogeneous parallel topology boosts low-resource efficient training



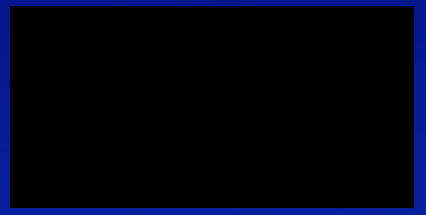
Testing on a single HGX (8*A800 40GB).

Scaling exact attention to ultra long sequence

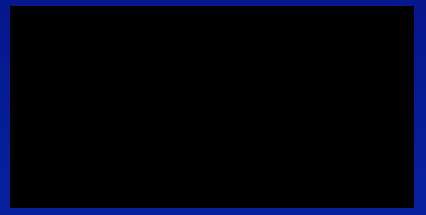
16 → 32 → 64 → 128 frames



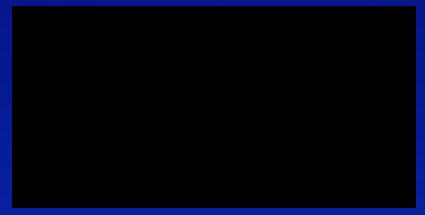
Fireworks exploding in the sky



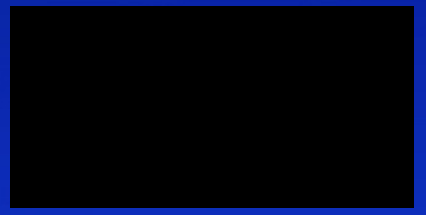
Fishes swimming in ocean camera moving



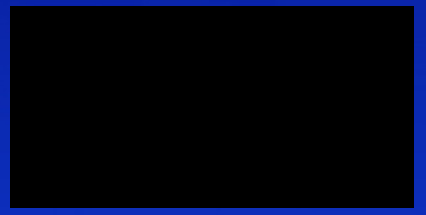
Fly to a mansion in a tropical forest viewpoint moving from down to up



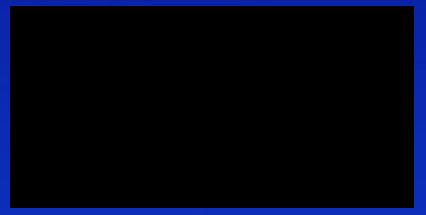
Flying above a breathtaking limestone structure



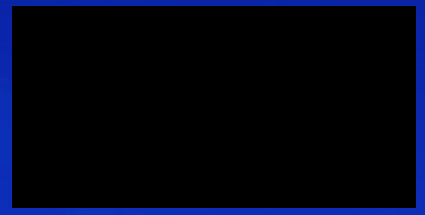
Petals Auttering in the wind, slow motion.



Santa dancing



Snow fallen in the forest pine tree are coverd with snow

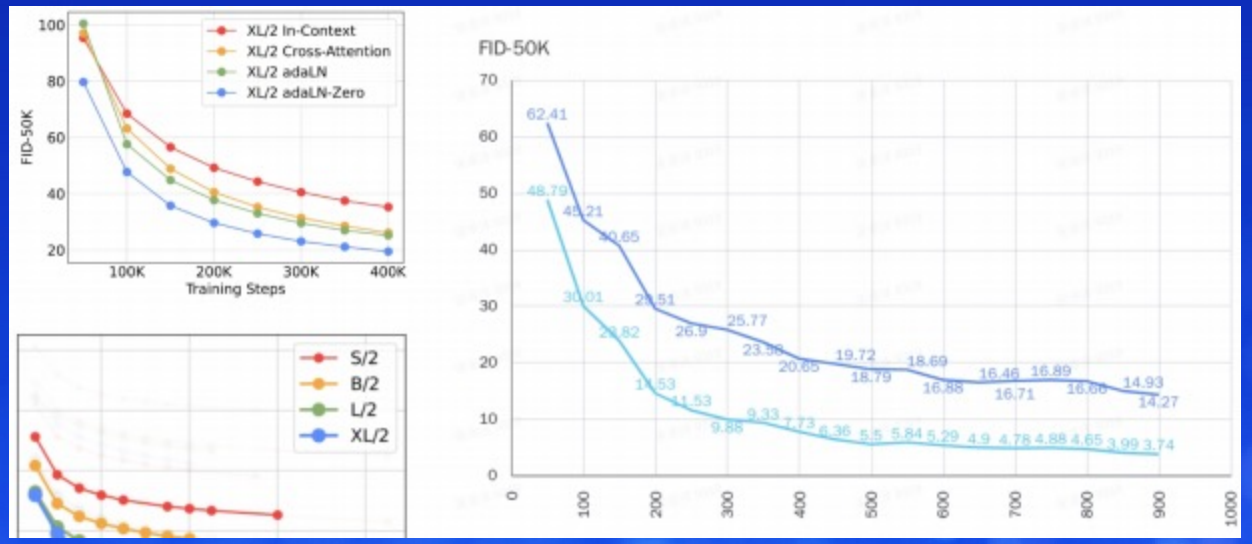


Drone view of waves crashing against the rugged clifs along Big Sur's garay point beach.

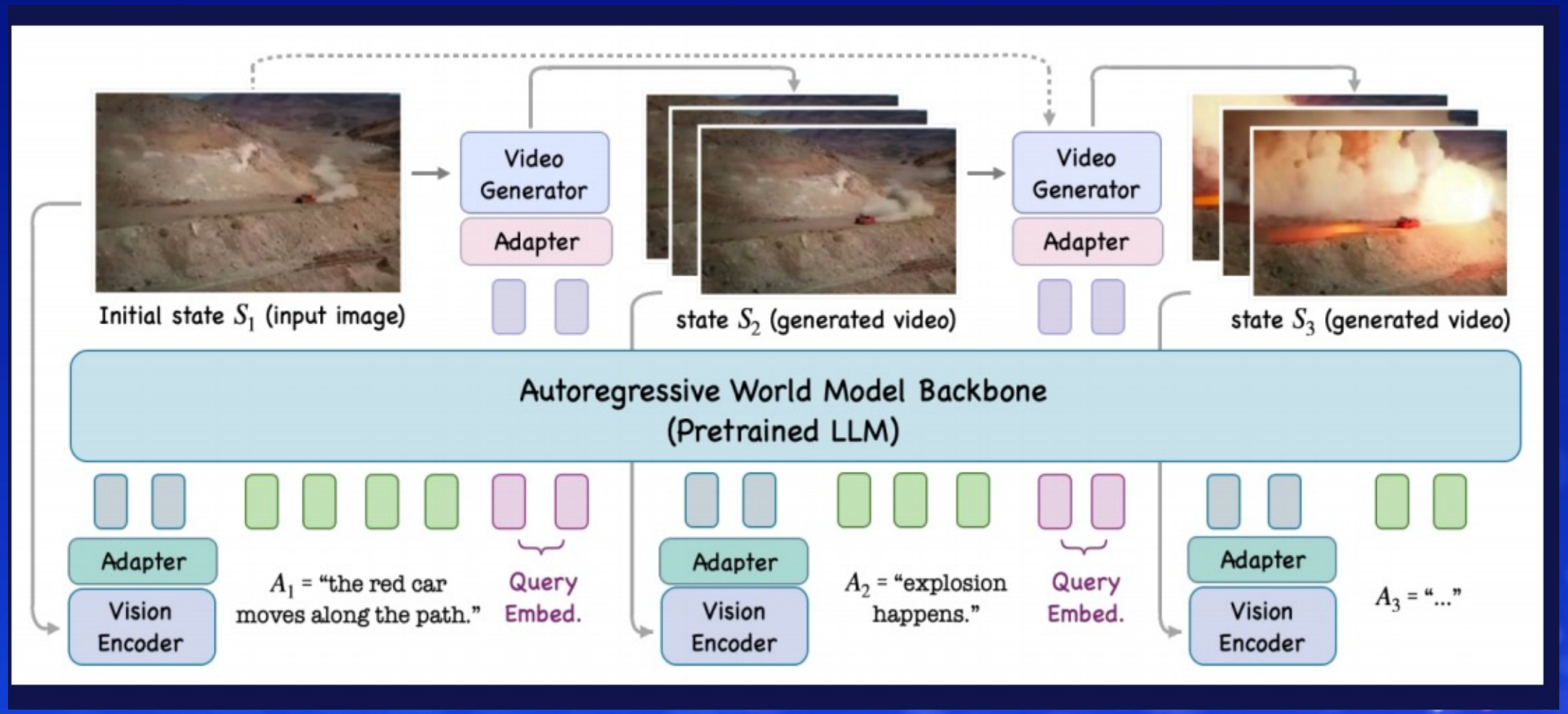
Train from Latte XL (16 frames); Done in March 2024.

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

Cloud-native AI Ops as the foundation of all

- 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 + 技术团队的共同选择

KEYLINK ing

K+峰会

- 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

- 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

AiDD AI+研发数字峰会 第5届
AI+ Development Digital summit

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