摩尔线程
MOORE THREADS

国芯国模：世界模型在 MTT S5000 上的原生全流程训练实践

基于 MUSA 软件栈的多模态生成与分布式训练

联合编制：EvoPhys AI 团队、摩尔线程

版本：v2026.09.15

第一章

参与方介绍

1.1 摩尔线程

摩尔线程围绕全功能 GPU（Universal GPU）构建算力产品和 MUSA 软件栈，支持图形计算、人工智能训练与推理等任务。在 EvoPhys-World 项目中，摩尔线程提供基于 MTT S5000 的夸娥（KUAE）智算平台、MUSA 软件栈及应用工程支持，支撑模型在国产基础设施上完成训练。

灯塔计划是摩尔线程的学术合作项目，面向国内领先高校的合作学者，优先提供 GPU 算力、MUSA 软件栈和应用工程资源，支持前沿人工智能研究。

1.2 EvoPhys 研究团队

EvoPhys 团队的核心研发成员主要来自北京大学，长期从事世界模型与具身智能基础模型研究，也是灯塔计划首批合作学者团队之一。

团队研发的 EvoPhys-World 是以人为中心的自演化 5D 世界模型，面向人与物体、场景之间的交互，支持场景生成与动作策略学习。团队负责模型研发与训练，摩尔线程提供算力平台和工程支持。

第二章

背景与目标

视频生成与世界模型将文本、图像、视频和音频融入同一计算流程，对算力平台的要求也随之延伸：既要支持模型训练和推理，也要衔接媒体处理，并适应集群规模的增长。世界模型常以视频生成模型为基座：基座模型提供通用的画面与时序生成能力，世界模型在此基础上结合动作与交互数据继续训练，学习动作如何改变场景。

在本次实践中，Wan2.2 与 MiniMax-H3 均可作为世界模型的可共用基座模型：Wan2.2 是常用的开源视频生成基座，MiniMax-H3 则是支持音视频联合生成的新一代开源基座。

围绕这些需求，本次基于自主 MUSA 软件栈，在训推一体智算卡 MTT S5000 上完成了 MiniMax-H3 的音视频推理与 SFT 训练适配，以及 Wan2.2 的视频模型训练与多机扩展实践。两项适配均沿用模型原有的计算流程，通过 MUSA 软件栈组件完成设备与算子对接，为同类世界模型在国产算力平台上的落地提供可复用的工程实例。

世界模型在 MTT S5000 上的全流程训练，可以拆分为三个环节：平台适配、基座模型的训练与扩展、世界模型的训练与评测。本白皮书依次展开：第 3 章介绍 MTT S5000 与 MUSA 软件栈的适配方案；第 4、5 章分别记录 MiniMax-H3 与 Wan2.2 两个基座模型的适配与训练实践，第 6 章总结可复用的适配经验；第 7 章介绍世界模型 EvoPhys-World 的训练支撑、评测结果与应用。

在世界模型层，EvoPhys-World 基于约 4 万小时纯人手第一视角（EGO）数据，在基于 MTT S5000 的夸娥（KUAE）智算平台与 MUSA 软件栈上历时约一个月完成端到端训练。模型一经推出便在斯坦福大学 WorldScore（ICCV 2025）“世界生成”维度登顶，并连续 37 天蝉联榜首；它可以根据用户动作生成并持续更新场景，已展示倒水、开关灯、移动杯子和场景漫游等应用。

2.1 MiniMax-H3：联合音视频任务

MiniMax-H3 支持文本、图像、视频和音频输入，并生成原生立体声音视频。用户可以通过首尾帧或参考视频控制生成内容，完成音视频创作。

本次实践完成了 MiniMax-H3 的训练与推理适配，涵盖音视频生成和 SFT 训练，并优化了通用分布式训练框架中的通信缓冲区管理 [5–7]，降低了多机微调的显存峰值。

开展这项实践的原因在于：音视频联合生成是基座模型演进的新方向，SFT 则是在基座上结合特定数据继续训练的常用方式。在 MTT S5000 上完成 H3 的推理、SFT 训练与多机扩展，说明新一代世界模型基座可以在国产算力平台上完成评估与定制训练；世界模型升级基座时，训练可以继续在同一平台上进行。实践中对通信缓冲区管理的优化（单卡通信缓冲区峰值由 7.712 GiB 降至约 0.451 GiB，见 4.4 节），也适用于其他采用 HSDP 的多机训练。

2.2 Wan2.2：视频 DiT 训练任务

Wan2.2 支持文本和图像条件下的视频生成。以 TI2V-5B 为代表的适配工作覆盖视频模型计算、训练数据处理和分布式训练。

在完成注意力等核心算子适配后，训练任务扩展至多机集群，训练吞吐随集群规模增长而提高。

对以视频 DiT 为基座的世界模型而言，若要在同一平台上开展训练，需具备两项基础条件：核心算子可迁移至 MUSA，训练规模可随集群扩展。Wan2.2 的算子适配与多机扩展实践，分别对应这两项条件。实测中，训练从 2 节点 /16 卡扩展到 14 节点 /112 卡，扩展效率为 92.5%，单卡显存保持稳定（见 5.4 节）。

第三章

MTT S5000 与 MUSA 软件栈与适配方案

MTT S5000 提供模型计算资源，MUSA 软件栈连接 PyTorch、计算算子、媒体处理和多机通信，为 H3 的推理与 SFT 训练，以及 Wan2.2 的分布式训练提供支持。

3.1 软件组件

组件	作用	对应任务
Torch-MUSA	为 PyTorch 提供 MUSA 张量计算与自动求导	模型前向计算、反向传播和参数更新
MATE	提供优化后的注意力算子	H3 与 Wan2.2 的 Attention 计算
MCCL	提供多设备集合通信	分布式训练中的参数与梯度通信
TorchCodec、Moore FFmpeg/MUVID	解码视频并输出设备侧张量	视频输入与媒体处理

表 1：MTT S5000 与 MUSA 软件栈主要软件组件与对应任务。

计算、通信与媒体组件共同衔接模型的输入、训练和输出。各项任务使用的软件版本见**附录 A**。

3.2 设备与精度适配

适配层负责模型权重和输入数据的设备放置，并根据计算环节处理精度转换。上层模型保留原有计算流程，由底层接口对接 MUSA 设备。

H3 在推理时按计算阶段调度模型权重，在训练时通过分片存储管理模型状态。Wan2.2 统一处理各计算模块的输入，并在归一化等环节保留所需精度。

3.3 分布式训练

分布式训练通过分片存储分担模型状态的显存开销，再以数据并行扩展集群处理能力。MCCL 负责设备间通信，支持参数与梯度同步。

H3 与 Wan2.2 均基于这一方式完成了多机训练。具体配置和测试方法见**附录 B**。

第四章

MiniMax-H3 基于 MTT S5000 与 MUSA 软件栈的实践

MiniMax-H3 已基于 MTT S5000 与 MUSA 软件栈完成训练与推理适配，支持首尾帧控制、参考视频生成、SFT 训练及多机扩展。这些能力为音视频内容生产、模型定制和世界模型基座的训练提供了国产算力上的实践参考。

4.1 模型与任务

MiniMax-H3 在统一上下文中处理文本、图像、视频和音频，并生成带原生立体声的视频。在推理方面，本次基于 DiffSynth-Studio 的公开实现，完成了两类生成任务的适配。模型介绍

DiffSynth-Studio 为 FL2VA 和 Ref2VA 提供独立的权重和任务入口。本次分别采用首尾关键帧和单个参考视频作为条件。

任务	支持的条件输入	本次输入
FL2VA	文本、首帧和尾帧	文本、首帧和尾帧
Ref2VA	图像、视频、音频或带音频视频	单个参考视频

表 2：MiniMax-H3 生成任务与条件输入。

两种任务分别加载对应权重，经过条件编码和联合去噪，由两个 VAE 分别解码视频和音频。

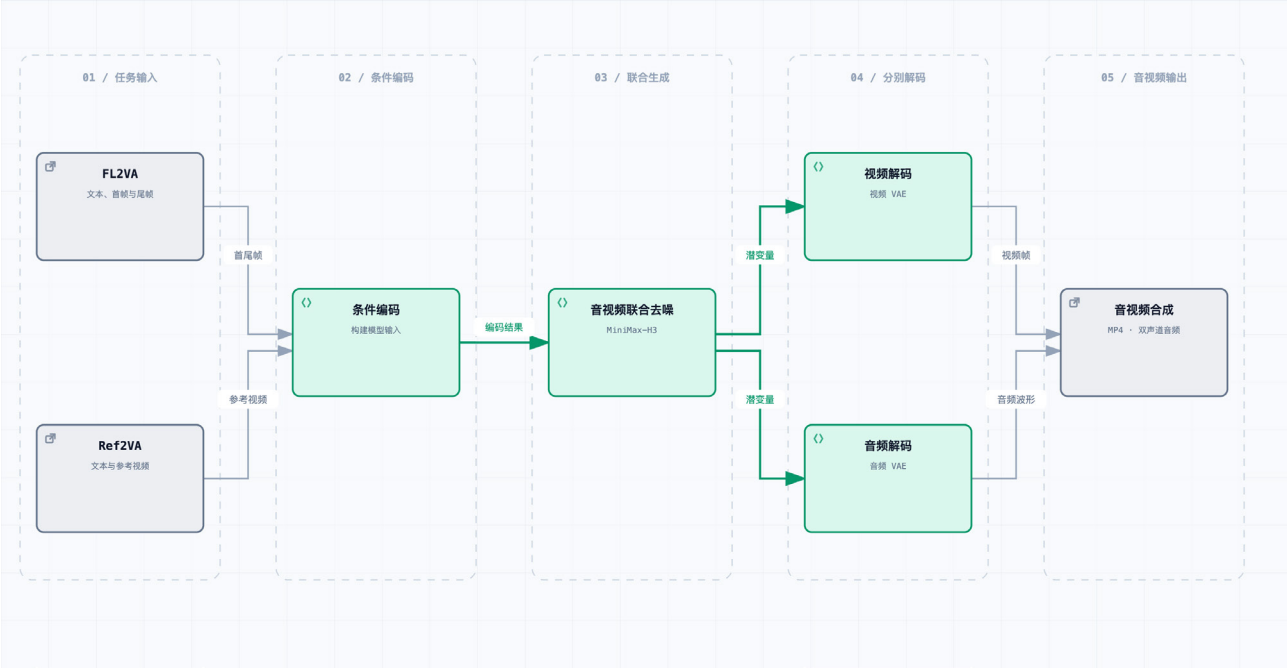

图 1：FL2VA/Ref2VA 音视频生成流程。

4.2 模型与媒体适配

适配工作沿用 DiffSynth 的模型管线，集中处理设备识别、权重迁移和张量精度对齐。模型组件按计算阶段在主机与加速卡之间调度，满足音视频联合生成的显存需求。

注意力计算接入 MATE FA3，视觉、音频和视频组件分别完成图像处理、音频重采样及视频解码，共同支撑音视频生成。

4.3 音视频生成

首尾帧控制和参考视频生成均已完成端到端运行，输出包含视频与双声道音频。两项任务采用相同的输出规格，测试结果如下。

任务	条件	采样设置	总时长 (s)	输出
FL2VA	首帧 + 尾帧	640×352，192 帧，50 步，seed 42	331.001	8 秒视频与双声道音频
Ref2VA	192 帧参考视频	640×352，192 帧，50 步，seed 42	1972.178	8 秒视频与双声道音频

表 3：固定输入测试结果，各运行一次。主流程总时长含模型加载、条件处理、生成与写出。

完整数据见推理运行数据（附件 1）。连续请求的补充测试见**附录 B**。

4.4 SFT 训练实践

H3 训练采用 FSDP2/HSDP，将模型状态分布到多张加速卡，并通过跨机通信同步梯度。

针对多机反向传播中的显存占用 [5]，适配方案优化了临时通信缓冲区的释放机制 [6,7]：在通信完成后按顺序回收，限制同时保留的缓冲区数量。优化后，原有训练配置恢复运行，优化器和训练精度保持不变。

对照	并行配置	单卡通信缓冲区峰值
修复前	2 节点/16 卡，dp_replicate=2、dp_shard=8	7.712 GiB
修复后	2 节点/16 卡，dp_replicate=2、dp_shard=8	约 0.451 GiB

表 4：HSDP all-reduce 缓冲区生命周期修复前后对照。

数据来源：缓冲区对照表（附件 2）、训练验证数据（附件 3）。

4.5 多机训练性能

多机扩展测试采用统一的分片规模，随集群规模增加训练副本。测试期间保持训练精度和优化器一致，各规模均进行复测。结果显示，训练吞吐随集群扩展持续增长。

节点/卡数	dp_replicate× dp_shard	吞吐几何均值	相对 2 节点/16 卡加速比	扩展效率	复测漂移
2/16	1×16	1.013179 samples/s	1.000×	100.000%	0.2826%
4/32	2×16	1.954544 samples/s	1.929×	96.456%	0.0523%
8/64	4×16	3.464289 samples/s	3.419×	85.481%	0.0058%
10/80	5×16	4.286872 samples/s	4.231×	84.622%	0.0098%
14/112	7×16	5.951876 samples/s	5.874×	83.921%	0.1171%

表 5：H3 多机训练扩展结果。完整测试数据（附件 4）。

第五章

Wan2.2 基于 MTT S5000 与 MUSA 软件栈的适配实践

Wan2.2 面向文本生成视频、图像生成视频等应用。本次适配以 TI2V-5B 为基础，基于 MTT S5000 与 MUSA 软件栈完成模型计算适配与多机训练，为视频生成模型的定制训练和以其为基座的世界模型训练提供实践参考。

5.1 模型与应用

TI2V-5B 共用一套模型主干，既能根据文字描述生成视频，也能根据参考图像生成视频。适配沿用公开模型结构和输入输出方式，增加面向目标场景的训练入口。Wan2.2 的主要模型见下表及 Wan2.2 官方仓库、官方模型集合。

模型	任务	分辨率	结构
Wan2.2-T2V-A14B	文本生成视频	480P、720P	两专家扩散 MoE
Wan2.2-I2V-A14B	图像生成视频	480P、720P	两专家扩散 MoE
Wan2.2-TI2V-5B	文本/图像生成视频	720P (1280×704 或 704×1280)	5B 稠密 DiT，高压缩 VAE
Wan2.2-S2V-14B	语音驱动视频	480P、720P	语音条件视频生成
Wan2.2-Animate-14B	角色动画与替换	按任务管线设定	角色参考与动作条件

表 6：Wan2.2 模型变体。本次训练采用 TI2V-5B。

5.2 模型计算与算子适配

推理时，模型先编码输入条件，再通过迭代去噪生成视频的潜在表示，最后解码并保存视频。文本和图像生成任务共用模型主干，图像生成任务增加参考图像条件。

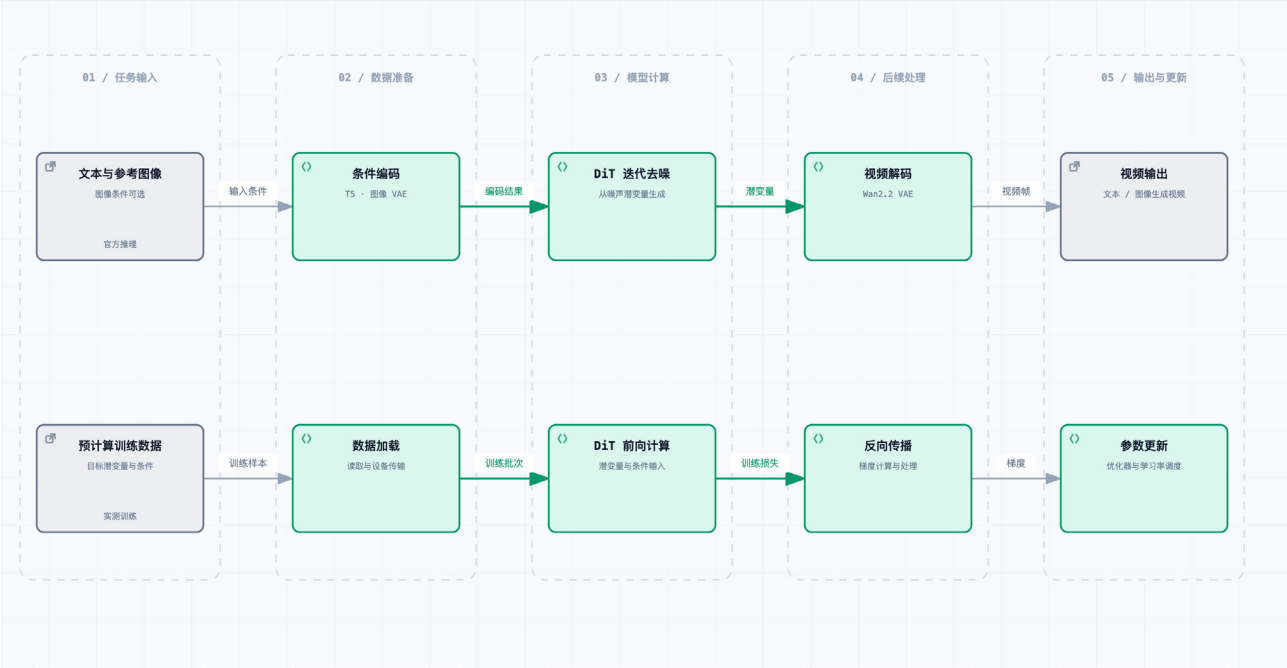

图 2：Wan2.2 TI2V 数据流程。上：官方推理；下：基于预计算数据的实测训练。

MUSA 适配集中在注意力、位置编码和归一化算子。通过统一设备与精度管理、调整算子实现，并对模型计算块进行编译，完成视频生成主干的迁移。

5.3 分布式训练

训练采用 FSDP2/HSDP，将模型状态分片存储，并通过增加数据并行组扩大训练规模。通信库自动选择网络设备，完成跨节点通信初始化。

训练数据按索引组织，模型读取视频潜变量和文本等条件后，完成前向计算、反向传播及参数更新。不同规模使用相同的模型和单卡训练配置。

5.4 多机训练性能

随着集群规模扩大，训练吞吐持续增长，单卡显存占用保持稳定。下表记录了共享存储环境中的实测表现。

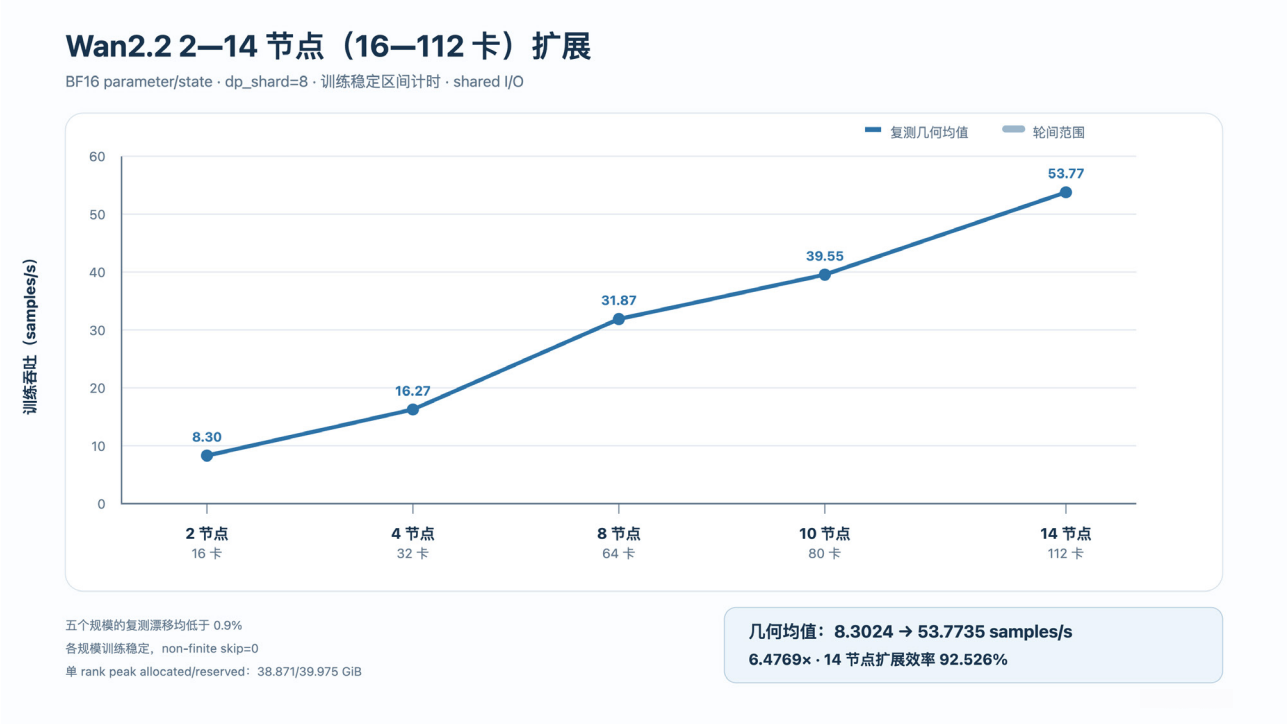

图 3：Wan2.2 TI2V-5B 训练扩展结果。

节点/卡数	吞吐 (samples/s)	相对 2 节点/16 卡加速	扩展效率	复测漂移	单卡峰值显存 allocated/reserved (GiB)
2/16	8.302412	1.000000×	100.000%	0.8984%	38.871/39.975
4/32	16.266617	1.959264×	97.963%	0.0399%	38.871/39.975
8/64	31.873995	3.839125×	95.978%	0.2208%	38.871/39.975
10/80	39.546185	4.763216×	95.264%	0.7035%	38.871/39.975
14/112	53.773488	6.476851×	92.526%	0.2691%	38.871/39.975

表 7：多机训练性能。完整数据见多机训练数据表（附件 5）和发布数据（附件 6）。
吞吐取稳定训练区间的几何均值，各规模均完成多轮复测。

更大规模的补充测试结果如下。

节点/卡	吞吐 (samples/s)
7/56	27.325
9/72	35.212
11/88	42.522
12/96	约 46.1
16/128	60.236
20/160	73.880

表 8：Wan2.2 高卡数训练吞吐。
高卡数阶段结果。完整数据见规模测试数据（附件 7）。
补充测试计时包含预计算潜变量与条件张量的读取、设备传输及训练计算。

第六章

可复用的 MTT S5000 与 MUSA 软件栈适配经验

H3 与 Wan2.2 采用不同的模型结构，适配工作都以保留上层计算流程、对接底层平台组件为主。这种分工让模型实现与设备相关处理各有归属，也便于复用已有适配。以这二者为基座的世界模型，同样可沿用相关适配。

6.1 设备放置与计算精度

将设备调度与精度转换集中到适配层，上层模型便可沿用原有计算流程。H3 在推理时按阶段调度权重，在训练时通过分片管理模型状态；Wan2.2 在计算模块内统一输入处理。设备与精度管理根据推理、训练的实际需要分别配置。

6.2 算子适配

两项适配均接入 MATE 注意力算子；Wan2.2 还完成了位置编码与归一化适配。算子实现与模型结构分开维护，为同类计算模块的接入提供了参考。

6.3 分布式通信

多机适配既涉及计算任务的分配，也涉及运行时资源管理。H3 通过及时释放通信缓冲区降低显存占用，Wan2.2 通过通信库的设备选择机制完成集群初始化。这些经验可用于同类模型的分布式部署。

第七章

EvoPhys-World 具身世界模型的训练与应用

第 4、5 章记录的是基座模型在 MTT S5000 与 MUSA 软件栈上的适配与训练实践。本章介绍 EvoPhys-World 的应用定位、交互方式、训练支撑与评测结果。

7.1 应用定位

EvoPhys-World 是面向人与场景交互的可控 5D 世界模型。用户可以通过动作控制场景变化，比较不同动作可能产生的结果，并继续与生成的场景交互。

系统由 World Engine 和 World Policy 组成：前者根据动作预测场景的下一状态，后者结合场景信息和候选结果预测下一动作。

7.2 动作驱动的场景交互

模型以首帧或场景记忆为基础，结合头部姿态、手部骨架、手势和接触动态生成视频。生成结果可用于下一轮交互，使场景随用户动作持续变化。

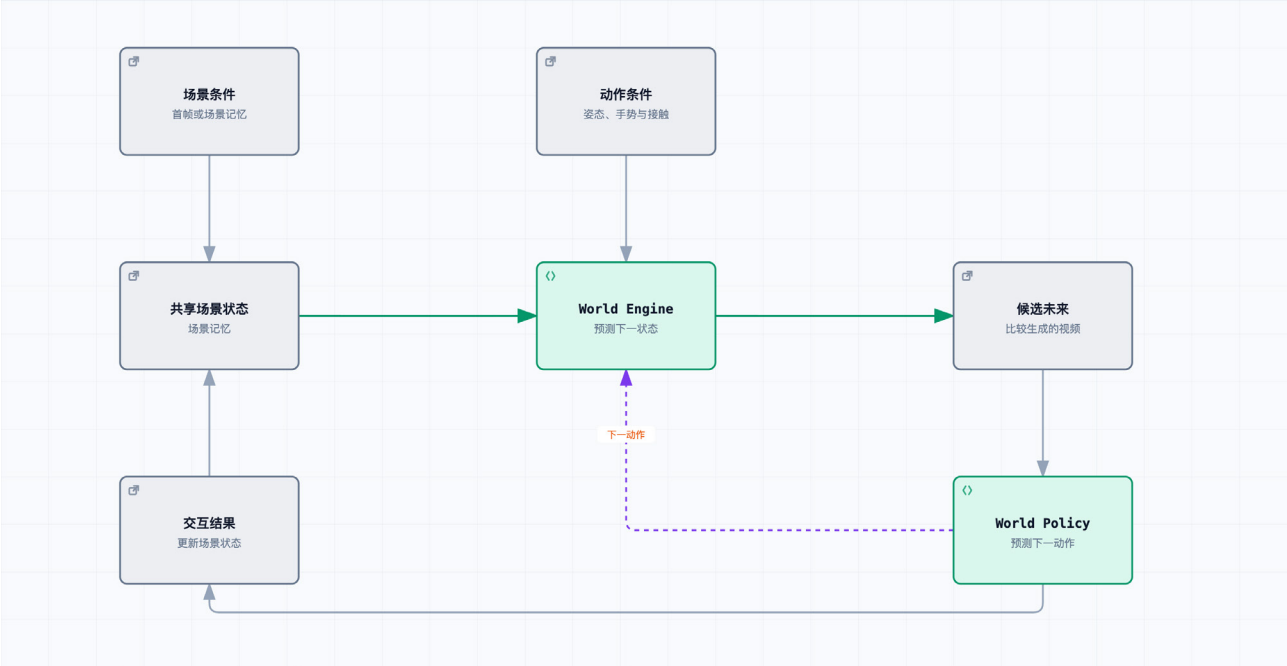

图 4：EvoPhys-World 的动作控制与场景交互流程。

7.3 MTT S5000 与 MUSA 软件栈训练支撑

EvoPhys-World 基于约 4 万小时纯人手第一视角（EGO）数据训练，学习人与物体、场景之间的交互。大规模、多模态数据训练要求平台兼顾数值精度与吞吐，并具备高效的跨节点通信和持续训练的稳定性。其中数值精度、吞吐与跨节点通信，在前文均有对应的实践：设备与精度适配见 3.2 节，注意力等核心算子适配见 4.2、5.2 节，分片存储与多机扩展见 3.3 节及 4.4、4.5、5.3、5.4 节。

EvoPhys-World 的训练围绕第一视角下的人与环境交互展开。在数据准备环节，对视频及相关交互信息进行时序整理和统一预处理，将连续记录组织为包含场景、动作及上下文信息的训练样本。可复用的预处理结果提前生成，减少训练中的重复计算，为大规模数据的持续读取提供支持。

在模型训练环节，结合视频生成能力，引入动作条件和历史场景信息，学习不同交互条件下的场景变化。训练关注的不仅是单段视频的生成效果，也包括动作与画面变化之间的对应关系，以及连续生成时的场景一致性。通过样例生成和连续交互验证，观察模型在不同场景下的表现，为后续迭代提供依据。

在工程实现上，团队依托 MTT S5000、夸娥智算平台与 MUSA 软件栈，围绕数据读取、显存管理和跨节点通信组织训练任务，将数据准备、模型训练与生成验证衔接起来。模型研发与平台适配协同推进，使交互建模的需求能够落实到国产算力平台上的实际训练流程中。

摩尔线程提供基于 MTT S5000 的夸娥（KUAE）智算平台与 MUSA 软件栈。在少量现场工程支持下，EvoPhys 团队约用一个月完成了端到端训练，将前沿模型研究落实到国产 GPU 平台。EvoPhys-World 以扩散式生成为核心、融合多模态时空建模，MUSA 软件栈支撑了其从适配到稳定训练的全链路；训练从单机扩展到数十节点规模时，吞吐保持近线性增长。

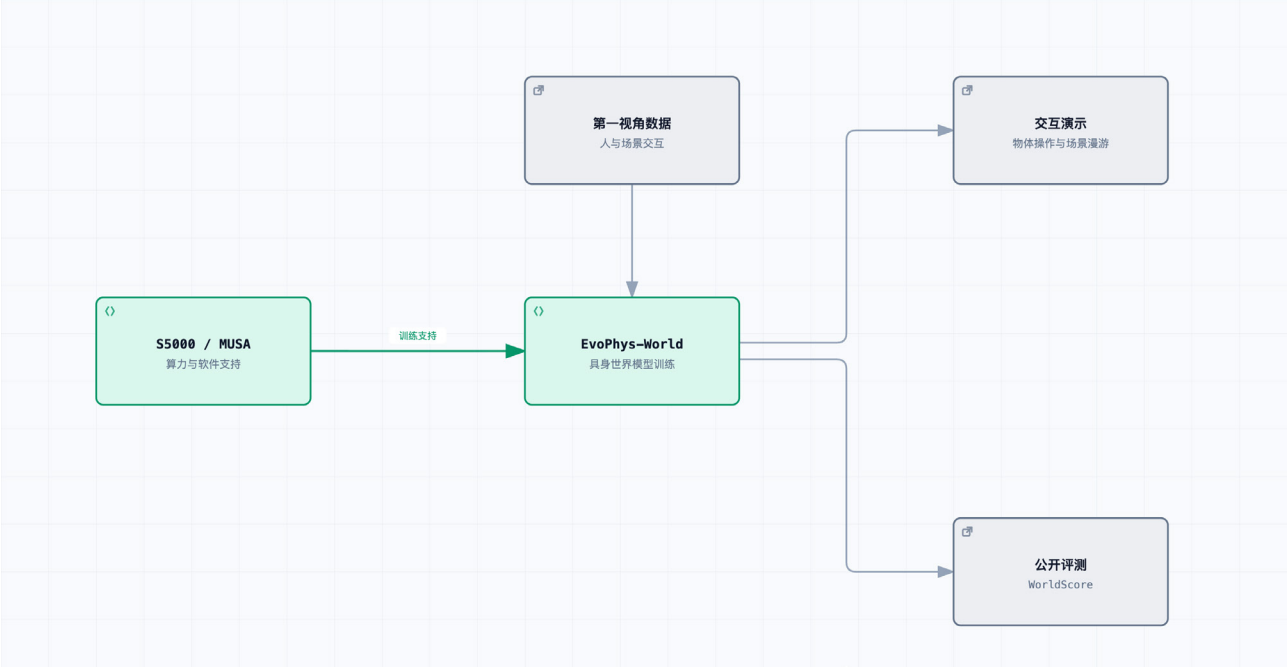

图 5：MTT S5000 与 MUSA 软件栈对 EvoPhys-World 训练的软硬件支持。

7.4 演示与评测结果

项目页展示了倒水、开关灯、移动杯子和场景漫游等应用。摩尔线程公开案例还通过相同起始画面和提示词下的生成视频，呈现抓握、笔迹、手与物体接触及物理反馈效果。

EvoPhys-World 于 2026 年 6 月 3 日发布，同日在斯坦福大学 WorldScore 基准的世界生成（World Generation）类别中排名第一，此后连续 37 天保持榜首。摩尔线程公开案例

WorldScore 是发表于 ICCV 2025 的世界生成统一评测框架。下表列出 EvoPhys-World 在公开榜单中的 IT2V 评测结果。

指标	数值	指标	数值
WorldScore-Static	83.45	WorldScore-Dynamic	73.78
Camera Control	94.56	Object Control	85.52
Content Alignment	83.37	3D Consistency	86.78
Photometric Consistency	85.56	Style Consistency	83.24
Subjective Quality	65.12	Motion Accuracy	62.47
Motion Magnitude	25.18	Motion Smoothness	66.00

表 9：EvoPhys-World 的 WorldScore IT2V 结果。数据表（附件 8）。

第八章
结语

本次实践围绕世界模型在 MTT S5000 上的原生全流程训练展开：在基座模型层，完成了 H3 的训练与推理适配、Wan2.2 的视频模型训练与多机扩展；在世界模型层，EvoPhys-World 基于 MTT S5000 与 MUSA 软件栈完成了端到端训练，并通过公开评测与应用演示检验了训练成果。模型团队与平台团队共同积累的适配实现和测试依据，为同类模型的训练与部署提供了参考。

附录 A

软件环境与配置

A.1 软件环境

H3 推理、H3 训练与 Wan2.2 训练分别使用以下软件组合

任务	框架与组件	主要配置
H3 推理	MUSA Toolkit 5.2.0；PyTorch / Torch-MUSA 2.9.1.post1+musa5.2.0；MATE 0.2.5	FL2VA、Ref2VA 与 Turbo；BF16
H3 SFT	PyTorch / Torch-MUSA 2.11.0.post1+musa5.2.0；DiffSynth 2.1.0	lazy_musa_fused_adamw；BF16 计算与归约；FP32 常驻参数与最终梯度
Wan2.2 TI2V-5B 训练（主要复测）	PyTorch / Torch-MUSA 2.7.1；Accelerate 1.12.0	musa_fused_param_adamw；BF16 参数与优化器状态；8 卡分片

表 10：各任务的软件环境与主要配置。

媒体组件使用 TorchCodec MUSA 0.5、Moore FFmpeg 7.0.2/MUVID、MCCL package 2.4.0， 容器 driver capabilities 为 compute,utility,video。H.264/MUVID 样本解码后输出 musa:0 张量。

详细组件版本见软件版本表（附件 9）和平台配置（附件 10）。

A.2 模型与测试配置

任务	配置与结果
H3 推理	FL2VA 配置（附件 11）、Ref2VA 配置（附件 12）、Turbo 配置（附件 13）、推理结果（附件 1）
H3 HSDP 修复	dp_replicate=2, dp_shard=8、window=2 配置（附件 14）、缓冲区对照（附件 15）、修复结果（附件 3）
H3 多机扩展	固定 dp_shard=16 的扩展配置（附件 16）、扩展结果（附件 4）
Wan2.2 多机扩展	主要复测配置（附件 17）、主要复测结果（附件 6）、高卡数测试配置（附件 18）、高卡数测试结果（附件 7）

表 11：模型测试配置与结果索引。

H3 的 HSDP 修复采用 dp_replicate=2, dp_shard=8；多机扩展实验固定 dp_shard=16。Wan2.2 训练使用 TI2V-5B，主要复测覆盖 2/4/8/10/14 节点，高卡数测试覆盖 56—160 卡。

A.3 H3 多媒体组件

组件	作用	适配结果
musa_patches	设备与张量适配	完成设备识别、权重迁移和 LoRA 张量对齐
torchvision 0.24.1.post1+musa5.2.0	图像处理	PIL resize 正常运行，NMS 在 MUSA 上返回有效结果
torchaudio 2.9.1+musa5.2.0	音频读取与重采样	CPU 读取音频，MUSA 执行重采样
TorchCodec MUSA 0.5 + Moore FFmpeg 7.0.2/MUVID	视频解码	H.264 样本在 MUSA 上完成 16 帧解码
safetensors 0.7.0	权重读取	在 CPU 读取后迁移至 MUSA

表 12：MiniMax-H3 组件适配结果。组件验证表（附件 19）。

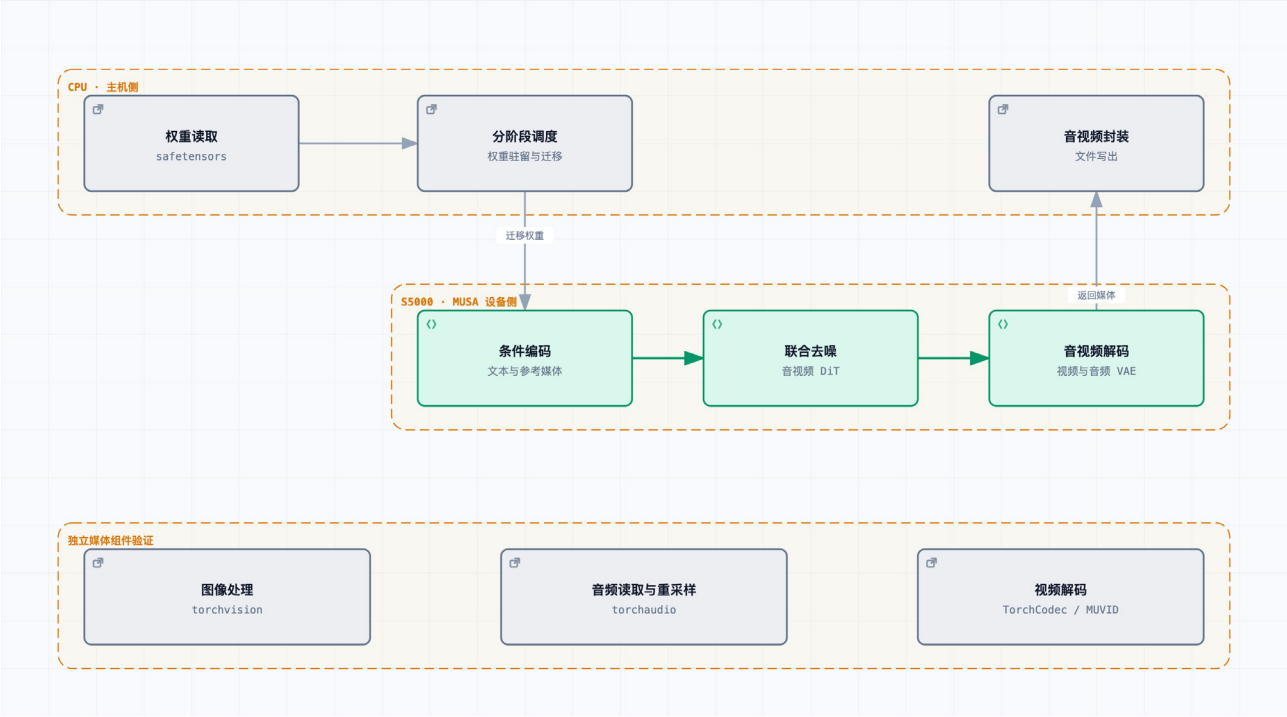

图 6：H3 的 CPU/MUSA 放置关系；下方为独立媒体组件验证。

附录 B
性能指标与补充数据

请联系摩尔线程 jinyue@mthreads.com 获取。

附录 C

术语表

术语	全称	含义
MUSA	Moore Threads GPU software stack	摩尔线程 GPU 软件栈，包含运行时、编译工具和计算库，PyTorch 中的设备名称为 musa。
Torch-MUSA	PyTorch extension for MUSA	为 PyTorch 提供 MUSA tensor、autograd 与分布式接入的扩展。
MCCL	MUSA Collective Communication Library	多设备进程组使用的集合通信 backend。
MATE	MUSA AI Tensor Engine	提供与上层接口兼容的优化算子。
FA3	FlashAttention-3	注意力计算的优化实现。
MUVID	MUSA video decode path	Moore FFmpeg 使用的视频解码实现。
TorchCodec	PyTorch media decoding library	将媒体解码结果转换为张量的 PyTorch 库。
FSDP2	Fully Sharded Data Parallel 2	基于 DTensor 的逐参数分片接口。
HSDP	Hybrid Sharded Data Parallel	组合 replicate 与 shard 维度的二维分布式布局。
FL2VA	First/Last-frame-to-Video-and-Audio	H3 接收文本和首尾帧条件，生成带音频的视频。
Ref2VA	Reference-to-Video-and-Audio	H3 接收图像、视频或音频作为参考，生成带音频的视频。
DiT	Diffusion Transformer	在扩散生成过程中处理 latent 与条件信息的 Transformer 主干。
TI2V	Text-Image-to-Video	接收文本或图像条件的视频生成接口。
Cold start	冷启动	包括进程创建、权重读取、首次编译和缓存建立等初始化过程。
Warmup	预热	正式计时前用于稳定程序执行和缓存状态的迭代。
Stable window	稳定统计窗口	程序运行稳定后用于统计性能的时间或迭代区间。
Strong scaling	强扩展	保持全局工作量一致，比较设备规模变化后的执行时间与效率。
Weak scaling	弱扩展	保持每台设备的工作量一致，根据吞吐量相对于设备数量的增长计算扩展效率。
dtype	Data type	张量的数据类型，如 BF16、FP32。
backend	执行后端	算子、媒体处理或集合通信所调用的底层实现。
latent	潜变量	编码器或生成模型内部使用的紧凑张量表示。
offload	卸载	将权重或参数分片放到主机内存，并按配置迁入目标设备的资源管理方式。

表 18：常用术语及含义。

参考文献

官方模型与公开实现

1. **MiniMax.** MiniMax H3 release. 2026-07-31.

2. **ModelScope DiffSynth-Studio.** MiniMaxH3Pipeline implementation.

3. **ModelScope DiffSynth-Studio.** MiniMax-H3 FL2VA example.

4. **ModelScope DiffSynth-Studio**. MiniMax-H3 Ref2VA example.

5. **PyTorch.** FSDP2 HSDP all-reduce tensors are kept alive until the end of backward. Issue #179128.

6. **PyTorch.** FSDP2: early release of HSDP all-reduce buffers. PR #179443.

7. **PyTorch.** FSDP2: limit HSDP all-reduce tensors kept alive. PR #183617.

8. **Wan Team.** Wan2.2 official repository.

9. **Wan-AI.** Wan2.2 model collection.

平台、框架与工具链

10. **Moore Threads.** torch_musa README.

11. **Moore Threads.** MUSA AI Tensor Engine README.

12. **Moore Threads.** MT Container Toolkit installation guide. version 1.9.0.

13. **PyTorch.** FSDP2 fully_shard documentation. tag v2.9.1.

14. **Moore Threads.** MUSA SDK.

应用案例与公开评测

15. **EvoPhys Team.** EvoPhys-World project page.

16. **摩尔线程 .** 摩尔线程 × 北京大学 | EvoPhys-World 国产 GPU 全栈原生训练 . 2026-06-08.

17. **摩尔线程 .** 具身智能社区 EvoPhys-World 应用案例 .

18. **WorldScore.** WorldScore Leaderboard.

19. **Duan H, Yu H X, Chen S, et al.** WorldScore: A Unified Evaluation Benchmark for World Generation. ICCV 2025.