Cosmos 3:面向物理 AI 的全模态世界模型

Cosmos 3: Omnimodal World Models for Physical AI

NVIDIA

技术报告 · 2026-06-22 · 原文 139 页 · research.nvidia.com/labs/cosmos-lab/cosmos3
代码与权重:github.com/nvidia/cosmos · huggingface.co/collections/nvidia/cosmos3

译稿说明。本稿为全文中文译稿,由 Claude 翻译,供内部研究阅读。原报告以 Linux Foundation `OpenMDW-1.1` 许可开源发布,版权归 NVIDIA(© 2026 NVIDIA)。译文完整保留原文结构与全部图表;附录 G(贡献者名单)与参考文献列表按惯例不译,请查原文。文末「**译者解读**」一**章不属于原文**,是面向本团队(车载 IMS/DMS 合成数据,现用 Cosmos-Transfer2.5 管线)的工程评估。

核心术语对照:omnimodal=全模态 · world model=世界模型 · Reasoner=推理器(理解面)· Generator=生成器(生成面)· Mixture-of-Transformers(MoT)=Transformer 混合架构 · mid-training=中期训练 · post-training=后训练 · flow matching=流匹配 · token=词元 · embodiment=具身形态 · video transfer=视频迁移(结构控制)· SDG=合成数据生成 · CFG=无分类器引导

摘要 Abstract

我们提出 Cosmos 3——一族全模态(omnimodal)世界模型,在统一的 Transformer 混合架构(mixture-of-transformers)中联合处理并生成语言、图像、视频、音频与动作序列。凭借高度灵活的输入-输出配置,Cosmos 3 无缝统一了物理 AI(Physical AI)的关键模态——实际上把视觉-语言模型、视频生成器、世界模拟器与世界-动作模型收纳进同一个框架。我们的评测表明,Cosmos 3 在一系列多样的理解与生成任务上确立了新的最先进水平,证明全模态世界模型可以作为具身智能体可扩展的通用骨干。截至本技术报告撰写时,经后训练的 Cosmos 3 模型被 Artificial Analysis 评为最佳开源文生图(Text-to-Image)与图生视频(Image-to-Video)模型,并被 RoboArena 评为最佳策略模型。为加速物理 AI 的开放研究与部署,我们以 Linux Foundation 的 OpenMDW-1.1 许可开放代码、模型检查点、精选合成数据集与评测基准,见 `github.com/nvidia/cosmos` 与 `huggingface.co/collections/nvidia/cosmos3`;项目主页 `research.nvidia.com/labs/cosmos-lab/cosmos3`。

开放资源一览(扉页资源表全译)

类别	名称	地址
开源代码	Cosmos	`github.com/nvidia/cosmos`
	Cosmos-Framework	`github.com/nvidia/cosmos-framework`
开放权重检查点	Cosmos3-Super	`huggingface.co/nvidia/Cosmos3-Super`
	Cosmos3-Nano	`huggingface.co/nvidia/Cosmos3-Nano`
	Cosmos3-Super-Text2Image	`huggingface.co/nvidia/Cosmos3-Super-Text2Image`
	Cosmos3-Super-Image2Video	`huggingface.co/nvidia/Cosmos3-Super-Image2Video`
	Cosmos3-Nano-Policy-DROID	`huggingface.co/nvidia/Cosmos3-Nano-Policy-DROID`
开放合成数据集	SDG-PhyxSim	`…/PhysicalAI-WorldModel-Synthetic-Physical-Interaction-Scenes`
	SDG-RobotSim	`…/PhysicalAI-WorldModel-Synthetic-Embodied-Robot-Scenes`
	SDG-DriveSim	`…/PhysicalAI-WorldModel-Synthetic-Autonomous-Driving-Scenarios`
	SDG-SynHuman	`…/PhysicalAI-WorldModel-Synthetic-Digital-Human-Scenes`
	SDG-Warehouse	`…/PhysicalAI-WorldModel-Synthetic-Warehouse-Operations-Scenes`
开放评测基准	Cosmos-HUE	`huggingface.co/datasets/nvidia/Cosmos-HumanEval-v1`

1. 引言 Introduction

物理AI(Physical AI)智能体通过感知、推理并采取行动来与真实世界交互。然而,直接在真实世界中训练这类智能体既缓慢、昂贵,还可能带来危险。为克服这些瓶颈,我们必须构建一座"训练设施",在模拟世界中实现安全且可扩展的学习;在那里,物理AI智能体需要习得两种在根本上相互耦合的能力:*理解*(understanding)与*生成*(generation)。理解使智能体能够从部分观测中推断潜在表征、语义与动态;生成则赋予智能体预测并模拟合理未来的能力,预判世界将如何演化、以及智能体应如何相应地采取行动。以往工作大多将这两大支柱割裂对待,由此形成了彼此独立的模型:用于感知与推理的判别式模型,如视觉-语言模型(Vision-Language Model, VLM);用于世界模拟的生成式模型,如视频生成模型(Video Generation Model)与前向动力学模型(Forward Dynamics Model);以及动作预测模型,如视觉-语言-动作模型(VLA)与世界-动作模型(World-Action Model, WAM)。

我们认为这种范式割裂从根本上构成了限制:理解需要对世界的未来演化及动作的后果进行推理,而生成依赖对世界与智能体行为的紧凑、结构化表征。因此,把二者统一进一个可扩展的单一框架,对物理AI至关重要。设想一台被指示在晚餐后清理餐桌的通用家用机器人。在当前范式下,机器人必须拼接一整套彼此脱节的模型:用 VLM 定位餐具并生成可执行计划,用 VLA 或 WAM 生成动作序列,再用前向动力学模型或"世界模型(World Model)"来模拟并评估未来状态。这种碎片化架构既非最优,又浪费算力。我们能否转而设计一个单一的统一模型,原生地覆盖物理AI智能体所需的全部核心能力?

我们提出 Cosmos 3——一族全模态(omnimodal)世界模型,对语言、图像、视频、音频与动作进行联合建模,同时服务于理解与生成。作为物理AI的通用主干,Cosmos 3 把大量彼此不同的模型类别统一到单一框架之中(图 1)。依据输入-输出配置的不同,Cosmos 3 可在多种运行模式之间无缝切换:它可以作为进行多模态理解与推理的视觉-语言模型;作为文生图(Text-to-Image)生成器;作为视频生成器,支持文生视频合成、图像动画化(图生视频)、未来预测(视频生视频)或音视频同步生成;还可以作为世界-动作模型,联合完成动作预测与环境模拟。通过在不做架构修改的前提下统一感知、模拟与执行,Cosmos 3 消除了对碎片化、任务专用管线的需求,借助共享表征与联合多任务监督实现可扩展的学习。

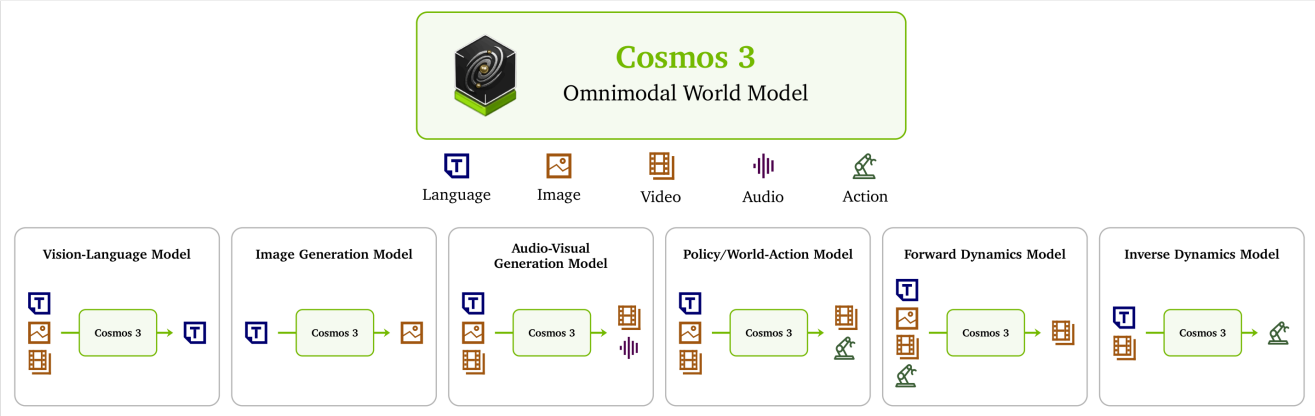


图 1:Cosmos 3 充当物理AI的通用主干。通过对语言、图像、视频、音频与动作进行理解与生成的联合建模,Cosmos 3 在单一网络架构内统一了众多模型类别,包括视觉-语言模型、图像生成模型、音视频生成模型、策略(policy)/世界-动作模型、前向动力学模型与逆动力学模型。

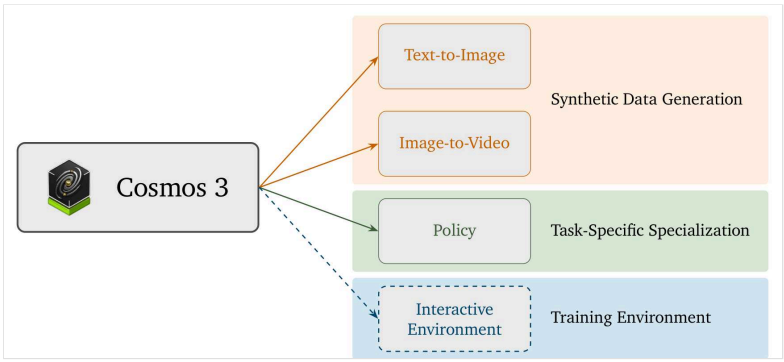


图 2:Cosmos 3 为训练物理AI智能体提供了强有力的起点。Cosmos 3 可以在不修改架构的情况下,针对不同应用在目标数据上进行后训练(post-training)。本文展示了我们如何把 Cosmos 3 后训练成更好的合成数据生成器(§4.2.3 与 §4.2.4)以及更好的机器人策略(§4.2.5)。我们预期 Cosmos 3 未来将在为物理AI智能体生成高质量、复杂环境方面发挥关键作用。

表 1:Cosmos 3 结果总览。Cosmos 3 在所有能力上持续优于各专用开源基线。详细结果见 §6。表中 * 表示后训练的 Cosmos 3 变体;† 表示闭源模型;每行的灰色单元格表示该模型不具备对应能力。

Model	Reasoning(推理)				Generation(生成)					
	General	Robotics	Smart infra.	Driving	Text2Image	Text2Video	Image2Video	Audio	FD: Robot	Policy: Robot
Cosmos3-Super	73.7	57.8	**62.6**	**79.3**	**91.36***	**80.0**	**82.8**	7.31	**26.0***	-
Cosmos3-Nano	69.6	55.1	61.0	76.0	84.61	79.4	82.7	7.34	25.5*	**39.7***
Gemini 3.1 Pro †	**77.5**	**58.2**	58.6	47.2						
Qwen3-VL-32B	72.8	52.6	56.1	40.7						
Qwen3-VL-8B	68.9	48.5	52.7	46.4						
Gemma-4-31B	69.8	51.0	51.3	36.6						
Gemma-4-E4B	53.1	39.3	29.4	26.0						
Gemini 3 Pro Image †					90.85					
Qwen-Image-2512					84.25					
Veo-3.1 †						79.1	82.6	**7.45**		
Wan2.2-A14B						78.0	81.3			
Ctrl-World									23.0	
$\pi_{0.5}$										28.1

译注 表中加粗为该列最优,下划线为该列次优(与原文的粗体/下划线标注一致);各分数出自本报告自建的评测体系(详见 §6),而非各对比模型论文的原始报告值。另注意 Cosmos3-Super 的 Policy: Robot 一栏原文为空("-"),即机器人策略结果目前仅有 Nano 的后训练变体;Text2Image 与两项 Robot 指标带 *,即这些最优值来自后训练变体而非基座模型。

为物理AI智能体扩展训练数据与环境的规模,始终是一个顽固的瓶颈。Cosmos 3 从三个方面为应对这一挑战提供了强有力的起点:(i)合成数据生成(SDG);(ii)任务专用特化;(iii)训练环境(图 2)。短期内,Cosmos 3 可合成高保真、多样化的视觉数据,以增强物理AI智能体的训练。我们在 §4.2.3 与 §4.2.4 展示了如何把 Cosmos 3 后训练成更好的合成数据生成器。由于智能体通过多样的具身形态(embodiment)与任务来感知环境并与之交互,Cosmos 3 支持在共享模型之上进行任务与具身形态专用的特化。作为面向物理AI的强大中期训练(mid-training)模型,Cosmos 3 通过建模通用的世界动态与动作先验建立了更好的起点,同时保持对下游适配的高度友好。实践中,该模型可在不修改架构的情况下在目标数据上做后训练;得益于其全模态设计,这种数据驱动的特化仍能保留共同的世界表征。§4.2.5 介绍了我们如何把 Cosmos 3 在 DROID 上后训练为能力很强的世界-动作模型。长远来看,Cosmos 3 有望为物理AI智能体生成高质量、复杂的训练环境。为加速物理AI的开放研究与部署,我们以 OpenMDW-1.1 许可证发布代码、模型检查点、整备后的合成数据集与一套评测基准,地址为 `github.com/nvidia/cosmos` 与 `huggingface.co/collections/nvidia/cosmos3`。

我们在覆盖物理AI核心理解与生成能力的大量基准上评测了 Cosmos 3 及其后训练变体。表 1 给出了基准结果摘要,详见 §6。如表所示,Cosmos 3 在多数能力上确立了新的最先进水平,与各专用模型相比极具竞争力或更胜一筹。

Cosmos 3 的技术细节按如下方式组织:§2 介绍模型架构,包括各模态的编码器、用于实现不同生成模式的多模态词元(token)排布、Transformer 混合架构(Mixture-of-Transformers, MoT)主干、多模态位置嵌入以及模型变体;§3 概述 Reasoner(推理器)与 Generator(生成器)训练所用数据;§4 详述 Reasoner 与 Generator 的训练配方;§5 描述基础设施,包括数据、训练、推理服务与评测;§6 呈现实验结果;§7 讨论相关工作;§8 总结全文。

2. 模型架构 Model Architecture

Cosmos 3 能够处理多模态输入并生成多模态输出。除语言、视觉(图像与视频)与音频之外,Cosmos 3 把动作也当作一种核心模态,引入了专门的一类动作词元。这些动作词元把物理世界与基于语言的推理、基于视频的世界建模连接起来,直接对应物理落地的控制信号,用于真实世界交互。Cosmos 3 集成各模态专用的编码器,把不同模态投影到统一表征空间,再交由 MoT 主干处理。推理时,语言词元通过下一词元预测生成,其他模态则通过迭代去噪生成。

2.1 编码器

给定由语言、视觉、音频与动作组成的输入序列,第一步是用各模态专用编码器把它们嵌入到统一表征空间。为了让共享的 transformer 参数与位置嵌入能够区分不同模态,我们在每个非语言模态送入 MoT 主干之前,为其加上一个可学习的、模态专用的嵌入向量。

2.1.1 图像与视频

我们为视觉输入采用两个相互独立的编码器。视觉理解使用经视觉-语言对齐预训练的 ViT 编码器;视觉生成使用来自 `Wan2.2-TI2V-5B`(Wan et al., 2025a)的视频 VAE 编码器。ViT 编码器的 patch 大小为 16×16,其后接一个两层 MLP,把 2×2 个词元合并并投影到 transformer 的潜空间(latent space)。参照 Qwen3-VL(Bai et al., 2025b),我们还通过 DeepStack(Meng et al., 2024)聚合 ViT 的视觉特征,并在视频帧之间交错插入文本形式的视频时间戳(Chen et al., 2024b)。VAE 对输入视频做 4× 时间压缩与 32×32 空间压缩,实现方式为 16×16 空间压缩后接 2×2 patch 合并。每个 VAE 词元先经一个线性层投影到 transformer 隐藏维度,再把潜变量送入 MoT 主干。用于理解的 ViT 编码器与主干联合训练,而用于生成的 VAE 编码器在训练中保持冻结。

> **译注** 生成侧 VAE 原样取自 Wan2.2-TI2V-5B 且全程冻结,意味着 Cosmos 3 的视频生成潜空间与 Wan2.2 完全一致。32×32 的总空间压缩(16×16 VAE 压缩 + 2×2 patch 合并)远高于常见视频 VAE 的 8×8,单帧词元数因此大幅减少,这是长视频序列能进同一 transformer 的重要前提。

2.1.2 音频

音频生成采用 Lee et al. (2025b) 的音频 VAE 架构。48 kHz 采样的原始立体声音频以 1920 个采样点的跳距(hop size)编码,得到每秒 25 个音频词元。音频 VAE 在训练中冻结。与其他非文本模态一样,音频词元先用一个线性层投影到 transformer 隐藏维度,再进入 MoT 主干。

2.1.3 动作

我们支持跨多种具身形态的动作建模,包括自动驾驶车辆、相机运动、机器人,以及第一人称人体运动(头部与双手)。由于每个领域都暴露自己原生的控制空间——如关节轨迹、转向指令、身体姿态或相机变换——我们把它们映射到一个统一的动作接口,从而在各领域间实现一致的多模态推理、生成与策略学习。

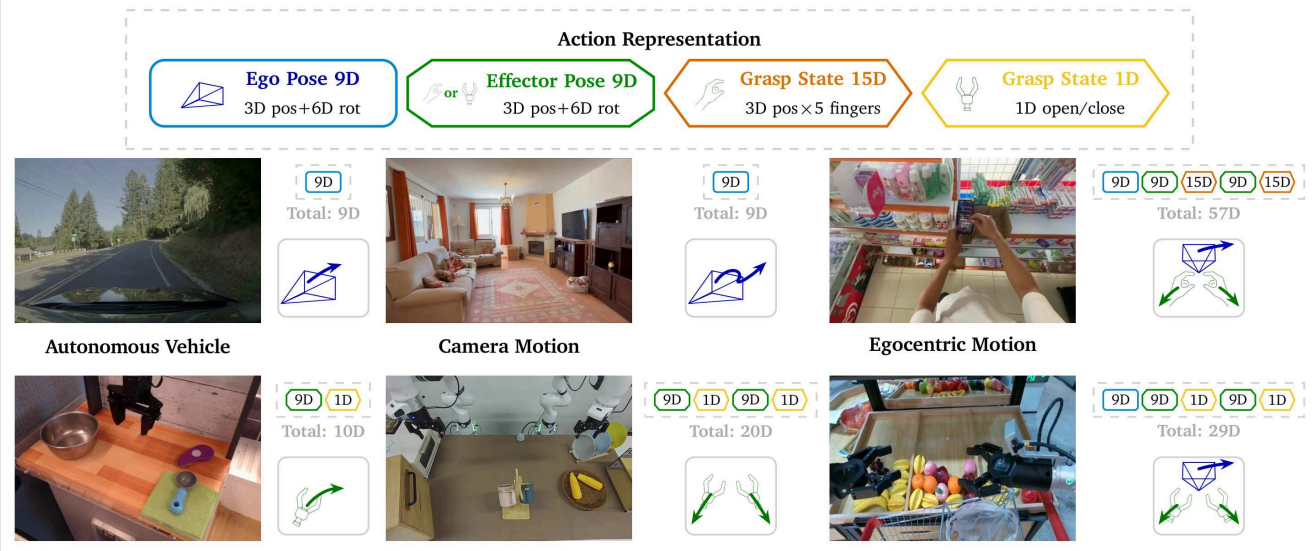


图 3:统一动作表示。我们把异构的具身控制映射为由共享几何组件构成的紧凑动作向量。自身(ego)运动与执行器(effector)运动用 3D 平移 + 6D 旋转编码为相对位姿伪动作(6D 旋转是 Zhou et al. (2019) 提出的过参数化旋转表示,因为旋转的自由度实为 3);而抓取状态(grasp state)直接编码当前操作状态,例如手的指尖位置或机器人夹爪的开合值。领域感知的输入/输出投影用于处理异构的动作向量长度,同时保持共享的语义空间。

动作表示。我们用"动作"指代引起世界状态变化的因果变量。给定连续的视频词元,一个动作词元 a_t 表示从上一状态 v_{t-1} 到当前状态 v_t 的转移。每个具身数据源都被变换为一种紧凑表示,以捕捉不同动作领域之间共享的底层几何结构,如图 3 所示。宏观上,动作最多包含三个组成部分:表示智能体主观测坐标系的自身位姿(ego pose)、表示智能体各执行器的执行器位姿,以及表示操作状态的抓取状态。为避免绑定具身特有的控制器细节(如比例-积分-微分 PID 参数或底层驱动接口),自身位姿与执行器位姿用由状态差分导出的伪动作表示:对连续的 SE(3) 位姿 $\mathbf{T}_{t-1}$ 与 $\mathbf{T}_t$,运动表示为相对变换 $\Delta\mathbf{T}_t = \mathbf{T}_{t-1}^{-1}\mathbf{T}_t$。旋转采用 Zhou et al. (2019) 的 6D 表示,并遵循 OpenCV 约定:z 轴沿手指/夹爪方向,x 轴指向右方。抓取状态的处理方式则不同:它不表示时间差分,而是直接编码 t 时刻的当前操作状态。

对相机与自动驾驶车辆,动作仅由自身位姿表示,没有任何执行器位姿或抓取状态。对第一人称数据,我们用头部相机位姿增量作为自身位姿、腕部位姿增量作为执行器位姿、各腕坐标系下的指尖位置作为抓取状态(Yang et al., 2025d)。对机器人数据,我们用头部相机位姿增量作为自身位姿、末端执行器(end-effector)法兰位姿增量作为执行器位姿(Lyu et al., 2026),以及连续的夹爪开合值作为抓取状态。

动作词元化。我们的动作表示把多样的具身形态映射进一个共享的潜动作空间,同时保留各具身特有的结构与语义。因此,我们为每个具身领域使用带独立权重矩阵的领域感知输入/输出投影层(Zheng et al., 2026),而 MoT 主干保持共享。对输入 $\mathbf{x} \in \mathbb{R}^{d_{in}^{(k)}}$(例如把头部位姿增量、左右腕位姿增量与指尖坐标拼接而成的第一人称动作向量)以及领域标识 $k \in \{1, \cdots, K\}$,输入投影为:

$$\boldsymbol{z} = \boldsymbol{W}_{in}^{(k)} \boldsymbol{x} + \boldsymbol{b}_{in}^{(k)} \tag{1}$$

其中 $\mathbf{z} \in \mathbb{R}^{d_{model}}$ 是潜动作词元,$\mathbf{x}$ 表示归一化后的动作向量,$\mathbf{W}_{in}^{(k)} \in \mathbb{R}^{d_{model} \times d_{in}^{(k)}}$ 与 $\mathbf{b}_{in}^{(k)} \in \mathbb{R}^{d_{model}}$ 是该领域专用的输入投影矩阵与偏置。

为把词元解码回原始动作空间,我们使用领域专用的输出投影:

$$\boldsymbol{x} = \boldsymbol{W}_{out}^{(k)} \boldsymbol{z} + \boldsymbol{b}_{out}^{(k)} \tag{2}$$

其中 $\mathbf{W}_{out}^{(k)} \in \mathbb{R}^{d_{in}^{(k)} \times d_{model}}$ 与 $\mathbf{b}_{out}^{(k)} \in \mathbb{R}^{d_{in}^{(k)}}$ 是该领域专用的输出投影矩阵与偏置。所有投影参数均从头随机初始化,并与 MoT 主干联合优化。预测出的 6D 旋转经奇异值分解(SVD)转换回 3×3 的 SO(3) 旋转矩阵。

2.2 词元排布与生成模式

Cosmos 3 是一个支持多种模态与任务的统一模型。不同任务都可以表述为交错的多模态序列,每个序列由来自不同模态的一系列片段组成。给定一个任务,所有片段先用上文所述的模态专用编码器编码为嵌入。嵌入完成后,来自不同模态的词元按一种适用于所有任务的统一格式打包,下面对此加以说明。

2.2.1 词元排布

输入词元序列由两个子序列组成:自回归(autoregressive, AR)子序列在前,扩散(diffusion, DM)子序列在后。

AR 子序列负责推理与理解。它包含语言词元,以及由 ViT 编码器嵌入的视频与图像词元。所有 AR 词元被路由到 transformer 解码器各层中一组专属的参数。

扩散子序列紧随 AR 子序列之后,包含来自 VAE 编码器的视频与图像词元,以及音频与动作词元。生成时,模型对含噪的扩散词元迭代去噪,得到对应的干净词元。扩散词元被路由到与 AR 词元不同的另一组参数,但仍通过每一层 transformer 解码器中的联合注意力与 AR 词元交互。

对任意给定任务,我们都用同一格式排布这些词元:(1)自回归词元置于扩散词元之前;(2)在扩散子序列内,对每个模态,干净的条件词元置于含噪的扩散词元之前;(3)在条件与扩散两部分子序列内部,词元均按视觉、音频、动作的模态顺序排列。借助这一统一格式,Cosmos 3 可以支持多种生成任务,详述如下。

2.2.2 生成模式

Cosmos 3 支持语言、视觉、音频与动作等不同模态。我们把干净的视觉、音频、动作词元分别记作 v、s、a,把它们的含噪对应记作 $\tilde{v}$、$\tilde{s}$、$\tilde{a}$。基于这些模态,支持的生成模式列举如下:

- **语言**。语言生成时,输入只包含自回归子序列,生成专用的扩散参数不被激活。若存在图像与视频输入,则由 ViT 编码器嵌入并放入自回归子序列。此时,Cosmos 3 的行为如同一个标准 VLM。
- **文生图**。此模式下,自回归子序列包含语言词元,扩散子序列包含由 VAE 编码器嵌入的含噪目标图像词元。整个词元序列为:

$$\boldsymbol{S}_{T2I} = [\boldsymbol{S}_{AR}, \tilde{v}_1] \tag{3}$$

  其中 $\mathbf{S}_{\mathrm{AR}} \triangleq [l_1, \cdots, l_n, \langle \mathrm{EOS} \rangle, \langle \mathrm{BOG} \rangle]$ 是下列所有模式共享的 AR 前缀($l_1, \cdots, l_n$ 为语言词元;$\langle \mathrm{EOS} \rangle$ 与 $\langle \mathrm{BOG} \rangle$ 分别是句末与生成起始特殊词元),$\tilde{v}_1$ 是含噪图像词元。
- **文生视频(+音频)**。此模式与文生图类似,但扩散子序列改为包含含噪的目标视频词元。当(可选地)联合生成音频时,含噪音频词元追加在含噪视觉词元之后。打包后的序列为:

$$\boldsymbol{S}_{T2V+Audio} = [\boldsymbol{S}_{AR}, \tilde{v}_{1:N}, \tilde{s}] \tag{4}$$

  其中 N 为潜视频帧数。
- **图生视频/视频生视频(+音频)**。此模式引入一张初始条件图像或若干初始视频帧,模型在其与文本提示的条件下生成完整的后续内容。在扩散子序列中,干净的条件图像或视频词元后面跟着含噪的目标视频词元:

$$\boldsymbol{S}_{V2V} = [\boldsymbol{S}_{AR}, v_{1:P}, \tilde{v}_{P+1:N}] \tag{5}$$

  其中 P 是条件潜帧数。$P = 1$ 时任务为图生视频,$P > 1$ 对应视频生视频。当同时生成音频时,音频词元的追加方式与文生视频情形相同。
- **视频迁移(transfer)**。此任务的输入为一段控制视频(例如 edge 边缘或深度)加一段文本描述,模型生成对应的 RGB 视频。词元布局与视频生视频类似:控制视频词元用作条件词元,RGB 视频词元用作含噪目标词元:

$$\boldsymbol{S}_{Transfer} = [\boldsymbol{S}_{AR}, {v^{ctrl}}_{1:N}, \tilde{v}_{1:N}] \tag{6}$$

  其中 ${v^{\mathrm{ctrl}}}_{1:N}$ 为控制视频经 VAE 编码的干净词元。

> **译注** 此处的结构控制(edge/深度)是把控制视频当作上下文条件词元原生打包进同一序列(in-context 条件化),不像 ControlNet/VACE 那样外挂适配器分支、也不引入额外控制参数——这正是 Cosmos 3 中 transfer 能力"架构原生"的含义。

- **动作**。Cosmos 3 支持三种与动作相关的生成模式——前向动力学、逆动力学,以及视频-动作联合预测(策略)。对包含连续视频词元的轨迹,每个动作词元 a_t 表示从 v_{t-1} 到 v_t 的转移。前向动力学在已观测上下文与干净动作词元的条件下预测未来视觉状态;逆动力学则推断能解释所观测视觉转移的动作词元。在策略模式下,模型联合预测动作与视频词元,使其能够在同一个序列模型中同时生成干预动作及其预期的视觉后果。各模式的条件方向汇总见图 4。

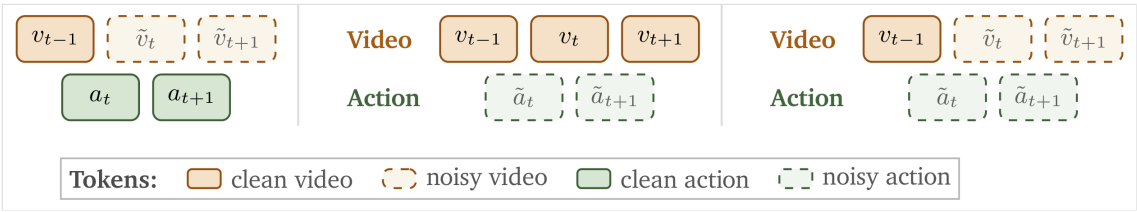


图 4:动作序列配置。对一个视频-动作数据样本,Cosmos 3 通过改变哪些词元为干净、哪些为含噪来构造不同的训练模式。图示为一个局部时间窗口,其中动作词元位于相邻视频词元之间:a_t 连接 v_{t-1} 与 v_t,a_{t+1} 连接 v_t 与 v_{t+1}。前向动力学模式以干净动作词元为条件对视觉词元去噪;逆动力学模式以干净视觉词元为条件对动作词元去噪;视频-动作(策略)模式对视觉与动作词元同时去噪。为简洁起见,图中省略了语言与特殊词元。

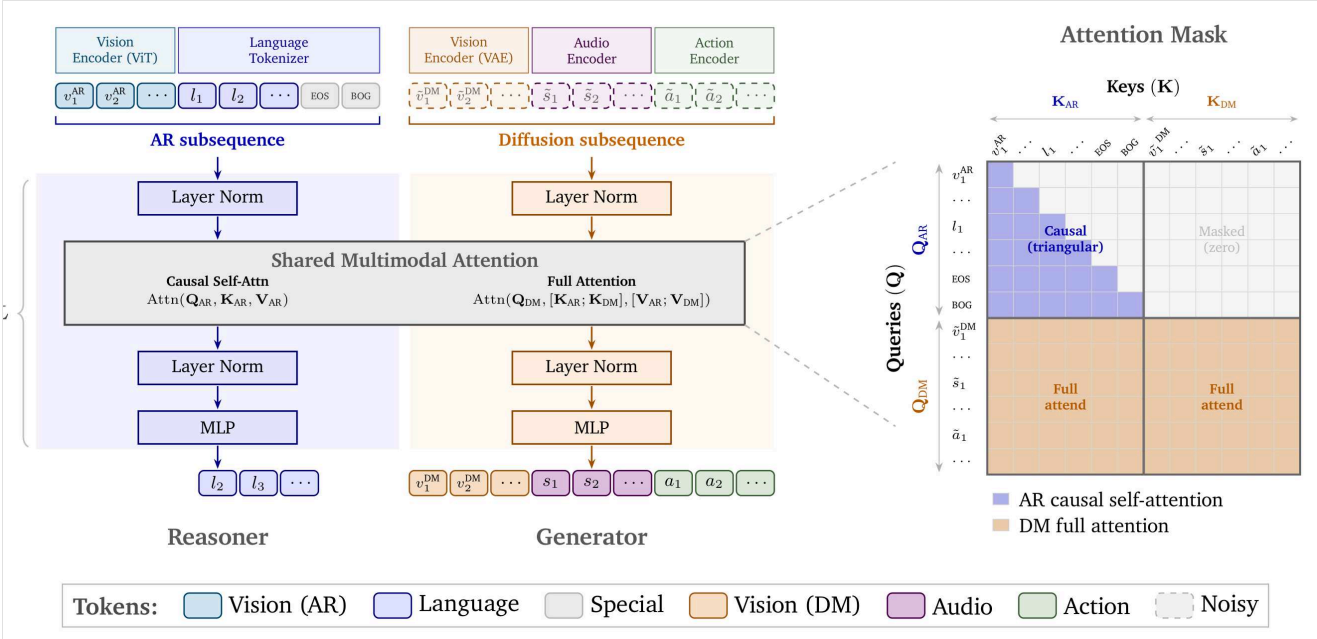


图 5:Cosmos 3 的 Transformer 混合架构(MoT)。左:单个 transformer 作用于一条由自回归(AR)与扩散(DM)两个子序列组成的词元序列:AR 携带离散文本词元及(可选的)ViT 编码视觉词元,以 `<EOS>` 与生成起始词元 `<BOG>` 结尾;DM 携带来自各自编码器的连续词元,训练时被噪声扰动。此处为简明起见把所有输入词元画成含噪;对图生视频或视频迁移等生成模式,DM 内的干净条件词元位于含噪目标之前,见 §2.2.2。在每个 transformer 块内,AR 词元与 DM 词元由相互独立的 LayerNorm 与 MLP 处理(全部由同一预训练 VLM 共同初始化),仅在共享的自注意力算子处交汇。设 Q、K、V 为注意力中的查询、键、值向量,下标标明其所属的塔。Q_{AR} 仅对 K_{AR}、V_{AR} 做因果注意力,而 Q_{DM} 对拼接后的 $[K_{AR}; K_{DM}]$ 与 $[V_{AR}; V_{DM}]$ 做双向注意力。如此,扩散以 AR 上下文为条件,而 AR 保持自回归自洽。输出方面,Reasoner 给出下一词元预测,Generator 给出去噪后的词元(实际训练采用预测速度场的流匹配(flow matching)目标;此处为清晰起见画出干净目标)。**右:**注意力掩码,AR 为因果掩码,扩散为全连接。

2.3 Transformer 混合架构(MoT)

Cosmos 3 采用 **Transformer 混合架构(MoT)**,处理来自不同模态词元组成的统一序列。在层的级别上,每个 transformer 解码器层包含两套参数:一套面向推理任务,处理 AR 子序列的词元(reasoner);一套面向生成任务,处理扩散子序列的词元(generator)。尽管 Cosmos 3 在解码器层结构上与 Deng et al. (2025) 等统一生成模型有相似之处,但它在训练策略、位置嵌入与整体能力上均有不同。

2.3.1 双塔层结构

标准的 transformer 解码器层由一个自注意力操作、一个前馈网络和若干归一化层组成。MoT 设计不用同一套参数处理所有词元类型,而是使用两条通路,如图 5 所示。每条通路都是一个拥有自身参数的标准 transformer 层,包括层归一化模块、注意力投影矩阵与前馈网络。两条通路均由同一个预训练视觉-语言模型(VLM)的权重初始化,使 Cosmos 3 在学习生成高保真视频的同时,继承强大的语言与视觉推理能力。在训练与推理期间,序列前部的 AR 子序列被路由到 reasoner 塔,后部的扩散子序列被路由到 generator 塔。

2.3.2 双流联合注意力

尽管两座塔使用相互独立的参数,扩散子序列的词元仍通过一次双流联合注意力(dual-stream joint attention)操作与 AR 子序列交互。这里我们把 AR 与扩散子序列的查询、键、值向量分别记作 Q_{AR}、K_{AR}、V_{AR}、Q_{DM}、K_{DM} 与 V_{DM}。

自回归子序列注意力。AR 子序列中的词元仅以*因果自注意力*关注 AR 子序列内部的词元,即每个词元只能关注同一序列中位于它前面的词元。这与从 VLM 主干继承的自回归性质完全一致,使模型得以保留预训练 VLM 的文本生成能力:

$$\boldsymbol{O}_{AR} = Attn_{causal}(\boldsymbol{Q}_{AR}, \boldsymbol{K}_{AR}, \boldsymbol{V}_{AR}). \tag{7}$$

扩散子序列注意力。DM 子序列的词元使用*全双向注意力*,以 AR 与 DM 词元的并集作为键与值。这使每个扩散词元都能自由关注自回归子序列中的文本提示,以及序列中所有其他条件词元与扩散词元,从而保持时间与空间一致性:

$$\boldsymbol{O}_{DM} = Attn_{full}(\boldsymbol{Q}_{DM}, [\boldsymbol{K}_{AR}; \boldsymbol{K}_{DM}], [\boldsymbol{V}_{AR}; \boldsymbol{V}_{DM}]), \tag{8}$$

其中 $[\cdot;\cdot]$ 表示沿序列维度拼接。需要指出,AR 词元从不基于 DM 词元更新,这保持了条件通路的因果完整性。

2.4 多模态位置嵌入

位置嵌入向注意力机制注入时间与空间结构,促使词元更强地关注语义与几何上相关的词元——它们往往在空间或时间上相邻。由于 Cosmos 3 在统一注意力框架内联合建模语言、视觉、音频与动作词元,设计一种能在各模态之间一致泛化的位置嵌入方案本身就具有挑战性。受 3D 多模态 RoPE(MRoPE)(Bai et al., 2025a) 启发,我们设计了带绝对时间索引的 3D MRoPE,把视频、音频与动作词元对齐到同一条物理时间轴上。原始 3D MRoPE 把每个注意力头的隐藏维度划分为时间、高度、宽度三个分量,其中时间分量只记录离散的词元索引。这一设计对图像与视频理解任务已经足够,但对我们的设定并不适用:视频、音频与动作词元可能以不同的帧率或采样率同时生成,此时不同模态的词元必须对齐到一条绝对物理时间轴。我们先介绍沿用原始 3D MRoPE 设计的基础形式,再描述我们的扩展与修改,尤其是用于对齐绝对时间轴的绝对时间调制。

2.4.1 位置索引分配

自回归词元。为了向后兼容语言生成与图像/视频理解模型,AR 子序列中所有语言词元与 ViT 编码媒体词元的位置索引沿用原始 3D MRoPE 设计。对语言词元,t = h = w 被设为同一个单调递增值,使 3D MRoPE 退化为标准的 1D RoPE 行为。对来自 ViT 编码器的词元,同一帧的所有词元共享 t,而 h 与 w 索引则按各词元的空间位置独立变化。自回归子序列中的位置索引分配与 Qwen3-VL(Bai et al., 2025a)的 3D MRoPE 设计完全相同。

扩散词元。如图 6 所示,视频词元在全部三个轴上变化:t 随时间潜帧索引推进,h 与 w 在每帧内独立铺满空间网格(0 ⋯ H−1, 0 ⋯ W−1)。图像词元被视为单帧视频,只在 (h, w) 上变化。空间与时间索引在每个视觉片段的起始处都重置为零,因此模型把 t、h、w 当作视频内部的绝对坐标,而非全局序列中的位置。例如,在视频迁移任务中,用户提供文本提示与受控视频帧(如深度图),干净的控制视频词元与含噪的生成视频词元都从自回归子序列末词元的时间偏移处开始。所有**音频词元**与**动作词元**只携带时间坐标,空间索引置零(h = w = 0)。音频词元的时间索引随每个音频跳距推进;动作词元的时间索引随每个采样步推进。

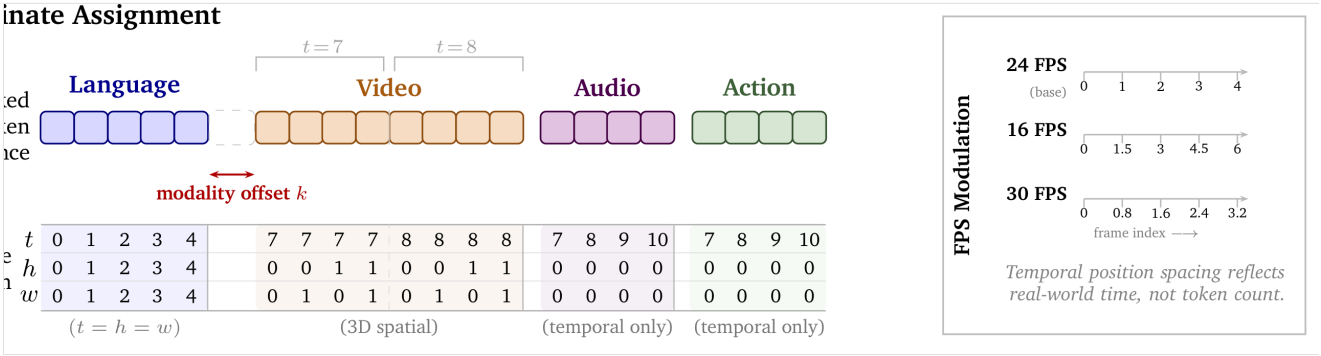


图 6:3D MRoPE 下坐标分配示意。**左**:一条打包词元序列,包含语言、视频(两帧,各为 2×2 空间网格)、音频与动作词元。每个词元获得一个 (t, h, w) 三元组。语言词元取 t = h = w;视频词元在全部三个轴上变化;动作与音频词元只使用时间坐标(h = w = 0)。模态偏移 k 把文本与视觉的时间范围分隔开。**右**:FPS 调制把帧索引映射到缩放后的时间位置,使相同的真实世界时长在 16、24、30 FPS 下占据相同的位置范围;其中 24 FPS 是我们的基准帧率。

自回归与扩散词元间隔。实践中我们发现,若直接让扩散词元从最后一个自回归词元的时间偏移处开始,会导致初始视频帧出现过饱和与棋盘伪影。该现象在 Cosmos 3 的较大变体(如 Super 模型)上尤为明显。我们推测其原因是:最后一个语言词元与首帧视觉词元占据相邻的时间位置,导致二者的时间嵌入几乎相同。为解决此问题,受 Cao et al. (2025) 启发,我们在自回归与扩散子序列之间插入一个固定的时间间隔,把后续所有视觉、音频与动作词元的时间索引整体平移。这在位置空间中制造出一个缓冲带,提供了更清晰的文本到视觉过渡信号,且无需修改架构或增加可学习嵌入。在我们所有模型中,该间隔均设为 15000。

2.4.2 绝对时间调制

沿时间维度的一个单位步长,在不同模态或数据源之间可能对应不同的物理时间间隔。例如,分别以 60 FPS 与 24 FPS 编码视频时,24 FPS 视频词元一次时间索引增量所对应的物理时长,是 60 FPS 视频词元的 2.5 倍。动作与音频词元也会出现类似差异,因为不同数据源可能采用不同的采样率。FPS 调制旨在通过调节每次时间增量的有效大小,把具有不同时间分辨率的词元对齐到共享的物理时间轴上。

我们先定义每秒时间步数(temporal steps per second, TPS)来刻画物理时间分辨率。对视频词元,TPS 为视频帧率除以时间压缩因子——由于视频 VAE 编码器,本文中该因子为 4。对音频词元,TPS 计算为 $\mathrm{TPS}_{\mathrm{audio}} = 48000/1920 \approx 25$(48 kHz,跳距 1920)。对动作词元,TPS 恰为动作数据的采样频率。

随后,我们把时间维度上的单位长度关联到一个基准 TPS,记作 $\mathrm{TPS}_{\mathrm{base}}$。对给定扩散子序列中的词元,我们先计算其对应的 TPS;当时间索引需要增加一个单位步长时,经调制后的时间增量 δt 计算为

$$\delta t = TPS_{base} / TPS. \tag{9}$$

由于视频占我们训练数据的大多数,而 24 FPS 是我们设定中最常见的帧率,我们取 $\mathrm{TPS}_{\mathrm{base}} = 24/4 = 6$,其中 4 是视频 tokenizer 的时间压缩比。

2.5 模型变体

Cosmos 3 按三个模型规模训练:**Edge**、**Nano** 与 **Super**,覆盖从端侧部署到大型数据中心推理的广泛算力预算。Edge 是建立在 2B 参数稠密 transformer 之上的 4B 参数模型;Nano 是建立在 8B 参数稠密 transformer 之上的 16B 参数模型;Super 是建立在 32B 参数稠密 transformer 之上的 64B 参数模型。所有变体都由预训练视觉-语言模型(VLM)初始化,并采用上文所述的 MoT 架构。表 2 汇总了各变体的关键架构超参数。本文发布 Cosmos3-Nano 与 Cosmos3-Super 模型;Cosmos3-Edge 模型将在后续版本中发布。

> **译注** 总参数约为底座 LLM 的两倍(2B/8B/32B → 4B/16B/64B),因为 MoT 双塔在每个解码器层各持一套完整的独立参数;另注意 Edge 与 Nano/Super 出身不同——Edge 的底座是从头训练的,Nano/Super 由 Qwen3-VL 初始化。

Cosmos3-Edge 采用 2B 稠密 transformer 设计:28 层、隐藏维度 2048、16 个注意力头、8 个键值头、头维度 128、FFN 维度 9216。我们使用 Megatron 代码库从头训练该 LLM。其设计大体沿用 `Qwen3-1.7B` 架构,但有两点显著不同:去掉了 QK 归一化,并把 FFN 激活改为 ReLU 平方(ReLU-squared),与表 2 中 Edge 的 FFN 维度配套。

Cosmos3-Nano 改造自 Qwen3-VL 8B(Bai et al., 2025b)架构:LLM 为 36 层、隐藏维度 4096、32 个注意力头、8 个键值头、头维度 128、FFN 维度 12,288。

Cosmos3-Super 改造自 Qwen3-VL 32B(Bai et al., 2025b)架构:LLM 为 64 层、隐藏维度 5120、64 个注意力头、8 个键值头、头维度 128、FFN 维度 25,600。

表 2:Cosmos 3 MoT 模型变体。所有模型共享双塔(dual-tower)MoT 架构。"LLM Layers" 指 transformer 解码器层数;每层为 reasoner 与 generator 两塔各携带一套独立参数。Edge 使用从头训练的 2B 参数稠密 transformer,而 Nano 与 Super 由预训练 Qwen3-VL 权重初始化。

Variant	LLM Layers	Hidden Dim	Attn Heads	KV Heads	Head Dim	FFN Dim
Cosmos3-Edge	28	2,048	16	8	128	9,216
Cosmos3-Nano	36	4,096	32	8	128	12,288
Cosmos3-Super	64	5,120	64	8	128	25,600

3. 数据 Data

训练 Cosmos 3 需要服务于两个互补目标的数据:Reasoner(推理器)通路学习理解并推理世界,而 Generator(生成器)通路学习合成与模拟世界,或在其中行动。尽管两条通路共享同一个 transformer 与相同的词元(token)表征,它们依赖的训练数据类型并不相同。Reasoner 在成对的视觉-语言数据(如图像-文本对与视频-文本对)上训练,以支持问答、空间定位(spatial grounding)、时序推理与动作理解等任务。相反,Generator 使用基于重建的目标函数(而非显式标注)在大规模多模态语料上训练,语料涵盖图像、视频、音频与动作。

因此,两条通路遵循不同但互补的训练课程(curriculum)。二者均采用多阶段训练策略,数据构成随训练进程演化。Reasoner 先进行广泛的视觉-语言预训练(pre-training),随后通过在物理AI(Physical AI)任务上的监督微调(SFT)实现专门化,任务覆盖机器人、自动驾驶与空间智能。这一分阶段课程先建立强大的通用能力,再逐步引入更专门的领域知识。Generator 先进行大规模图像、视频与音频预训练,然后逐步纳入更多模态,例如动作、控制条件迁移(control-conditioned transfer)以及面向特定能力提升的定向合成数据。

3.1 Reasoner 数据

我们的 Reasoner 数据课程共包含约 24.2M 个样本:22.0M 用于预训练,2.2M 用于监督微调,后者来自领域专用的物理AI数据集及合成生成的数据。表 3 汇总了两个阶段所用的数据模态。预训练阶段以图像-文本与纯文本数据为主,提供广泛的通用视觉理解能力。相比之下,监督微调阶段转向物理AI专门化,视频-文本样本占混合数据的 50%,以强化时空理解能力以及机器人、智能基础设施(smart infrastructure)与自动驾驶车辆领域的能力。图 7 按能力类别汇总了两个阶段的数据混合构成。

> **译注** 表 3 中 SFT 阶段视频-文本样本为 1,079,200 / 总计 2,171,673 ≈ 49.7%,与正文"约 50%"的说法一致。

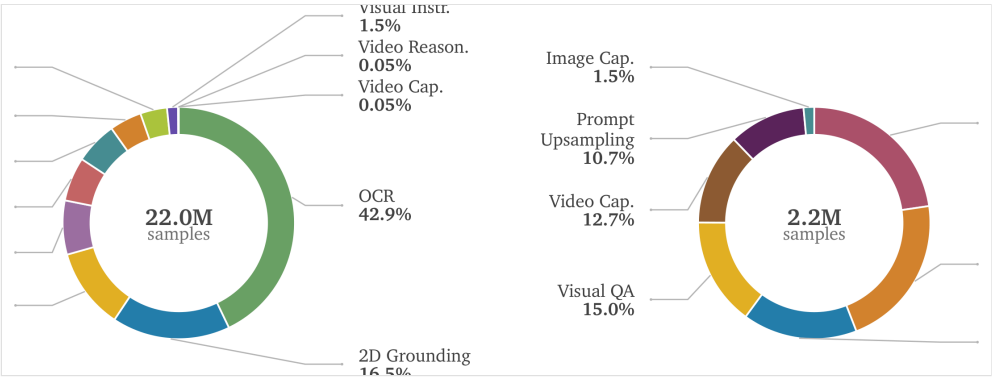


图 7:Cosmos 3 Reasoner 按能力类别划分的数据构成。我们汇总了用于训练 Cosmos 3 Reasoner 的整备数据混合,覆盖预训练与监督微调两个阶段。该混合包含 22.0M 预训练样本与 2.2M 监督微调样本,横跨图像-文本、视频-文本与纯文本类别;每个圆环展示了各主要能力流的相对占比,如 OCR、视觉问答、推理、描述文本(caption)、定位(grounding)与指令微调。

表 3:Cosmos 3 Reasoner 按模态与训练阶段划分的数据课程。Image-text 与 Video-text 行的表值为媒体样本(图像或视频)数量,纯文本行为对话(conversation)数量。

模态	预训练	监督微调
Image-text	18,814,952	1,051,513
Video-text	1,016,299	1,079,200
Text only	2,170,762	40,960
总计	22,002,013	2,171,673

3.1.1 预训练

我们的预训练数据混合由两部分构成:从 Nemotron Nano 2 数据集合 (NVIDIA et al., 2025a) 中子选出的 19.7M 样本,以及为增强数学、视频、空间定位与指令跟随能力而额外整备的 2.3M 样本。详见表 3。所采集的数据集在进入最终训练混合之前,会经过一条两阶段数据整备(curation)流水线:先做语义去重(semantic deduplication),再做 AI 评审(AI-judge)质量过滤。

语义去重。第一阶段在对话(conversation)层面移除多模态近重复样本;这里的"对话"指完整的训练样例:一张图像或一段视频与其指令-回复文本配对,或在无媒体时为纯文本的指令-回复样本。对每条对话,我们计算一个联合嵌入,将媒体表征(若存在)与相应的指令-回复文本表征结合起来。图像-文本与纯文本对话使用 `Qwen3-VL-Embedding-8B` (Li et al., 2026b) 进行嵌入,而视频-文本对话使用 Perception Encoder `PE-Core-G14-448` (Bolya et al., 2025) 进行嵌入。所得的拼接嵌入表征同时刻画视觉与语言语义,使流水线能够区分"视觉相似但任务意图不同的样本"与"真正冗余的监督信号"。

为了将重复检测扩展到生产规模的数据集,我们使用聚类。图像-文本、视频-文本与纯文本样本先经 K-means 聚类划分;然后在每个聚类簇内部,基于对话嵌入空间中的余弦相似度识别近重复组。相似度高于 0.95 这一较高阈值的样本被移除。这种层次化设计使大规模重复检测在计算上可行,同时保持了对视觉内容与任务语

义两方面冗余的敏感性。

AI 评审质量过滤。第二阶段在去重后的语料上应用一个 AI 评审来评估标注质量。我们使用 Gemma-4 作为视觉-语言评审 (Google DeepMind, 2026b),具体为 `Gemma-4-31B-it` 模型 (Google DeepMind, 2026a)。评审被提示扮演训练数据审计员的角色,按评分细则(rubric)在三个主要质量维度上给出 1 到 5 的整数分:

- *忠实性(Faithfulness)*:回复中的所有论断是否都有所给图像、视频或文本上下文的依据。
- *完整性(Completeness)*:回复是否完整地回应了指令,没有重要遗漏。
- *正确性(Correctness)*:回复在事实、逻辑与任务层面是否准确。

忠实性对 Cosmos 3 尤其重要,因为缺乏依据的视觉论断可能教会模型幻觉出物理状态、物体属性或时序事件。同时,完整性维度过滤掉描述不足或不完整的回复,正确性维度则移除最终答案或推理与输入不一致的监督信号。除标量分数外,评审还会给出简短的、基于证据的理由说明,便于对过滤行为进行有针对性的抽查与审计。

基于阈值的数据集构建。我们从同一份去重后的基础语料出发,使用对三个质量维度的最低阈值规则,构建多个经评审过滤的数据集变体。只有当样本的完整性、正确性与忠实性得分全部达到或超过指定阈值时才会被保留。换言之,每个被保留的样例都必须在全部三个维度上同时满足最低质量要求。

这一过滤策略有意比"对分数取平均"更严格。在任何单一维度上存在严重失败模式的样本都会被移除,即使它在其余标准上得分很高。例如,一个高度详细、逻辑正确、但包含无依据视觉论断的回复,仍会因忠实性得分低而被过滤掉。实践中,我们采用保守的阈值设定,以剔除明显低质量的监督信号,同时尽量减小各能力域上过度的分布偏移。

多模态去重移除了 4.23% 的数据(视为近重复监督)。AI 评审的分数分布显示,质量失败在各维度上并不均匀。我们还按能力类别分析了保留率,以确保质量过滤在改善语料的同时不会无意间压垮技能分布。在更严格的阈值下,剪除会变得强烈地"类别选择性":例如,当阈值从 2 提高到 5 时,指代表达定位(referring-expression grounding)被移除得最为激进,图像描述与视觉问答也大幅下降。这里的阈值指每个质量维度上可接受的最低 AI 评审分:只有当样本的完整性、正确性与忠实性得分全部高于该阈值时才被保留。OCR 数据在阈值为 2 时相对稳健,但在阈值为 5 时显著下降。这些趋势促使我们对主混合数据采用最低的非平凡评审阈值,即 2:它能过滤明显的标注失败,同时保留推理、定位、OCR、描述与 VQA 能力的原始覆盖面,从而在预训练数据集上取得最优的质量-数量权衡。然而,在 SFT 阶段,我们使用阈值 5,只保留置信度最高的监督样例——在该阶段,标注精度与回复可靠性比广泛覆盖更为关键。AI 评审过滤在阈值 2 和 5 下分别保留 78% 和 46% 的数据。

最终的预训练混合包含约 22M 样本,涵盖 OCR、定位、问答、推理、描述与指令跟随数据,如表 3 所示。就构成而言,OCR 是最大的组成部分,贡献 9.44M 样本(42.9%);其后是 2D 定位 3.62M 样本(16.5%)、视觉问答 2.48M 样本(11.3%)与图像推理 1.66M 样本(7.5%)。其余部分通过文本问答(1.35M,6.1%)、图像描述(1.30M,5.9%)、视频问答(0.99M,4.5%)、文本指令数据(0.82M,3.7%)、视觉指令数据(0.34M,1.5%),以及少量视频描述与视频推理数据(各 0.01M)提供广泛的多模态覆盖。这一构成强调强图像-文本对齐、文字阅读与空间定位能力,同时保留一个轻量的视频成分,为后续在时序与视频推理任务上的监督微调做好准备。

3.1.2 监督微调

在监督微调阶段,我们增强通用的空间与时间理解能力,并聚焦于为以下三个领域整备数据:自动驾驶车辆、机器人与智能基础设施。该阶段总计使用 2.2M 样本训练。

通用空间理解。我们通过 2D 与 3D 定位来增强通用空间理解,并同时用真实数据与仿真数据加以扩充。

2D 与 3D 定位。2D 定位支撑需要定位(物体、区域、部件与关键点)、指点(pointing)、计数与多图对应的物理AI任务。我们整备的样本覆盖检测、指代表达、OCR/版面、视觉提示描述与 VQA、指点、轨迹、计数与稠密定位。框与点经归一化后转换为统一的 JSON 格式。2D 定位数据整备自合成数据与既有数据,其中大部分来自 LocateAnything (Wang et al., 2026a) 数据,我们从中采样出一个高质量子集。对于 3D 定位数据,我们将 3D 扫描场景转换为带相机相对 3D 框的指令跟随训练样本。每个物体在规范化(canonicalization)与内参归一化 (Brazil et al., 2023) 之后,标注有类别标签、中心、尺寸与朝向。我们没有把 2D 定位与 3D 推断串联起来 (Cheng et al., 2026; Man et al., 2025),而是直接对最终的结构化 3D 框进行监督。

真实世界空间理解与定位。我们组合了多种互补的问答类型。图像指代问答任务要求在给定自然语言表达时,以归一化图像坐标指向目标物体或空闲自由空间 (Zhou et al., 2025b)。机器人空间问答要求预测自由空间中的放置点,并回答横跨自身中心(ego-)、世界中心与物体中心参考系的二元相对位置与可达性问题 (Song et al., 2025)。我们进一步纳入带位姿场景的多图问答、视频帧空间问答,以及从固定集合中预测两物体间关系(如左侧、后方、上方/之上)的多选题,包括视角替换后的观察点 (Liu et al., 2025b)。这些数据合在一起,覆盖物体指代、自由空间、跨视角对应、相机运动、尺寸、距离、方向、路线、计数与房间尺度推理。

仿真器锚定的具身空间推理。我们通过从可执行的仿真器状态中推导标签,来衔接视觉空间问答与具身动作。答案由相机、物体位姿/包围框、深度、掩码、可见性与可行区域计算得出,再经程序化校验器与 VLM 评审校验。该课程涵盖度量几何、空间参考系、物理语义、可操作定位、视角动态与具身组合,输出形式包括多选题、数值、点、框、二元与文本。

通用时间理解。我们沿三个轴增强时间能力:时序事件理解、物理合理性判断,以及结构化时空场景升采样(upsampling)。

时序事件理解。我们用三个互补的监督数据源强化时序与运动理解。第一,由人工标注者创建稠密时序描述:对日常室内任务的第一人称(egocentric)视频标注原子级人类动作描述及起止时间戳,动作平均时长 1.8 秒、平均 14.2 个词。此外,一个更广泛的视频语料(55K 视频,2.6K 小时)提供了 743K 个事件三元组 (t_{start}, t_{end}, caption),用于事件枚举与查询条件下的定位。第二,为增加问题多样性,我们用 FoundationMotion 流水线 (Gan et al., 2025) 整备训练数据,为每个片段生成十道四选一选择题,考查动作身份、时序演化与细粒度运动差异。最后,我们标注平移(panning)、变焦(zooming)等相机运动模式,使模型能学习自身相机(ego-camera)的运动。

物理合理性判断。我们用两个互补的监督数据源强化 Cosmos 3 Reasoner 判断生成视频物理合理性的能力。第一,我们纳入附录 F 所述 Cosmos 人工评估的标注,其中包含来自 1K 条生成视频的 13.5K 条人工评分的 (视频, 问题, 答案) 三元组,覆盖视觉完整性、时间稳定性、几何、解剖结构、运动合理性与物理常识,答案为"是/否/不确定"的类别标签。第二,我们采用 VideoPhy-2 (Bansal et al., 2025b),它提供 3.4K 条以动作为中心的视频,横跨 200 种动作,每条按物理定律符合程度打 1–5 分;其标注覆盖守恒定律、重力、碰撞动力学、时间因果与空间约束,被转换为监督问答对,由模型预测物理符合度分数。

结构化时空场景升采样。我们整备"欠指定的用户输入"与"稠密结构化描述"的成对数据。输入可以是文本提示词、图像或二者兼有,目标则是 §3.2.1 中描述的成对结构化描述文本标注。该升采样器(Upsampler)能力可以恢复输入中隐含的空间、时间与视觉细节,例如主体属性、场景布局、相机取景、物体交互与合理的未来运动。为使模型对不同请求格式保持稳健,我们合成了多种指令变体——其中一个变体为规范形式、采样频率更高——并将不同输入提示词长度(如描述应多详细:简短、详细等)与多种提示风格(如描述应如何刻画场景:请求式、陈述式等)交叉组合。所得监督信号鼓励模型遵循简洁、直接与风格化的用户请求,生成详细的场景描述,同时保持相同的结构化输出契约。§6.3.2 描述了将该能力用于 Cosmos 3 图像/视频生成。

自动驾驶车辆(AV)。我们纳入的 AV 数据集横跨人工标注与自动标注的思维链(chain-of-thought, CoT)推理、时序事件理解与 3D 车辆定位。

动作 CoT。来自内部驾驶日志的人工标注 CoT 数据包含超过 10K 条带显式驾驶决策的视频,覆盖天气、光照、路况、交通规则、自车(ego-vehicle)行为、关键物体,以及场景元素与自车行为之间的因果关联。为扩展该信号,我们从内部日志中自动标注了约 1.1M 条额外的富决策视频。对每条视频,我们将元动作(meta-action)转换的关键帧识别为决策时刻,并利用最先进的 VLM、原始视频、自车轨迹、动态状态与元动作,产出结构化决策、关键要素与简洁的推理轨迹。

时序事件定位。我们从 Nexar 行车记录仪素材 (Moura et al., 2025) 中推导时序事件定位数据,包含超过 24K 条视频,覆盖碰撞、险些碰撞、急刹车、急加速与急转弯。片段按 6 FPS 采样,时长上限 300 秒。我们在人工/自动标注数据之外,补充了关于场景、交通参与者、交互、自车行为与时空上下文的稠密描述。

3D 车辆定位。我们用 MADS 数据集 (Ren et al., 2025a) 训练度量级 3D 车辆定位;该数据集提供同步的多相机序列、世界场景图(world-scenario-map)3D 标注、相机内外参与自身位姿(ego pose)。我们以约 1 FPS 采样帧,构建开放词汇检测与指代定位问答,要求模型枚举车辆,或按相对位置、车道、运动状态或距离定位实例。答案是以位置、尺寸、roll、pitch、yaw 与类别标签参数化的相机坐标系 3D 框,过滤后仅保留 100 m 以内可见或轻度遮挡的物体,覆盖多样的地区、光照与天气。

机器人与具身 AI。我们为机器人操作中的动作 CoT、具身推理与医疗机器人手术理解整备数据。

机器人动作思维链。动作 CoT 教会 Cosmos 3 Reasoner 将高层具身指令与当前帧转化为用于机器人末端执行器(end-effector)控制的 2D 图像平面运动规划。它不采用自由形式的推理说明,而是通过任务相关位置、定位点、移动推理与 2D 路径点(waypoint)来结构化推理过程,以紧凑、可检视的轨迹从感知走向动作。该轨迹识别被操作物体、上下文物体、可供性(affordance)点与无碰撞区域,将它们定位为坐标,并把规划解析为像素空间的末端执行器路径点。由于定位、坐标化与操作规划需要不同技能,数据用一条模块化流水线构建:`Qwen3-VL-72B-Instruct` (Bai et al., 2025a) 生成定位理由与移动推理,`Molmo-7B` (Deitke et al., 2025) 对指代表达进行坐标定位,运动规划目标则来自 MolmoAct (Lee et al., 2025a) 或经跟踪的 DROID (Khazatsky et al., 2024) 回合(episode)。

具身推理。我们用时序定位、任务规划与机器人具身问答数据强化 Cosmos 的具身推理。在机器人操作方面,我们针对 MimicGen (Mandlekar et al., 2023) 的瓶颈——将演示切分为以物体为中心、带精确时间戳边界的子任务:使用 60 条留出(held-out)视频做零样本评估,并构建一个含 3.6K 条 Omniverse 重渲染视频、横跨六个任务的监督微调集,时间戳由物体轨迹与关节运动学推导并经人工校验。在长时程规划方面,我们整备了 83K 条 BEHAVIOR-1K (Li et al., 2023a) 样本,将场景帧与候选动作列表映射到真值动作。我们还加入了来自 EO-Data-1.5M (Qu et al., 2025) 的 ERQA 机器人问答,覆盖任务规划、可供性、故障检测、物理常识、坐标定位、指代、关系与轨迹预测。

医疗机器人手术理解。我们整备了一个机器人辅助手术 VQA 数据集,含 398K 条多轮对话、覆盖 2.2M 张图像。数据采集自手术室的外部(exocentric)相机、一台第一人称机器人细节相机与操控台显示器,借鉴了 ORQA (Özsoy et al., 2026) 的思路。多数样本组合多个视角,并包含跟踪器元数据(器械状态、3D 平移、欧拉旋转)与机器人元数据(手术阶段与步骤)作为上下文提示。任务覆盖器械识别、定位、是/否分类、场景图、监视器文字转录、人员计数、距离/时间估计、动作标注与手术步骤识别。

智能基础设施。我们整备了三个互补的智能基础设施数据源,覆盖仓库空间智能、稠密行人定位以及交通与异常推理。

仓库空间智能。我们使用 PhysicalAI-Spatial-Intelligence-Warehouse (Tang et al., 2025),这是一个合成 Omniverse 语料,横跨 44 个仓库场景集合与 40 个相机视角。从 93K 张 RGB–D 图像与 873K 个问答对中,我们子采样出 80K 个类别均衡的样例,覆盖对托盘、纸箱、叉车、货架与操作员区域的物体计数、度量距离、定位与二元空间关系。

稠密行人定位。我们整备了含 208K 张图像的标注,来自 44 个场景,共 5.6M 个人工标注的人体框。所有人体包围框均由人工标注者手工标注,且在标注与发布之前通过模糊化处理去除个人可识别信息(PII),以确保被摄者匿名。

交通与异常推理。我们组合了合成 ITS 碰撞监督、真实交通事件推理与监控异常校验三类数据。CARLA (Dosovitskiy et al., 2017) 片段用于训练无保护左转与侧面碰撞(T-bone)场景中被标记车辆对之间的二元碰撞预测,经 Cosmos-Transfer2.5 (NVIDIA, 2025b) 增广后得到 3.4K 条带标签的车辆对查询。TAR (NVIDIA, 2026b) 贡献了 3.6K 条交通摄像头视频(26 小时)与 44K 条标注,横跨问答、时序推理、因果关联、场景描述与摘要,其层次化 CoT 自动标签与人工标注进行了交叉核验。为把异常覆盖面拓宽到交通之外,我们还整备了 1K 条内部监控片段,用于尾随进入(tailgating)的二元校验。

> **译注** 原文此处写作"3,6K traffic-camera videos",按上下文(26 小时)应为 3.6K(3,600 条)之误,系欧式小数逗号排版笔误。

3.2 Generator 数据

我们的 Generator 训练遵循一个渐进式多阶段课程,在训练过程中逐步引入新模态:预训练阶段从图像、视频与音频开始,随后在中期训练(mid-training)阶段纳入动作与交错多模态内容。Cosmos 3 被定位为各类物理AI应用的良好起点。为展示其能力,我们取中期训练检查点 `Cosmos3-Nano` 与 `Cosmos3-Super`,对其进行后训练(post-training),用专门的后训练数据集产出领域专家模型,包括 `Cosmos3-Super-Text2Image`、`Cosmos3-Super-Image2Video` 与 `Cosmos3-Nano-Policy-DROID`。这些模型与其对应的中期训练模型共享相同的架构。图 8 汇总了 Generator 跨模态、跨阶段的训练课程。

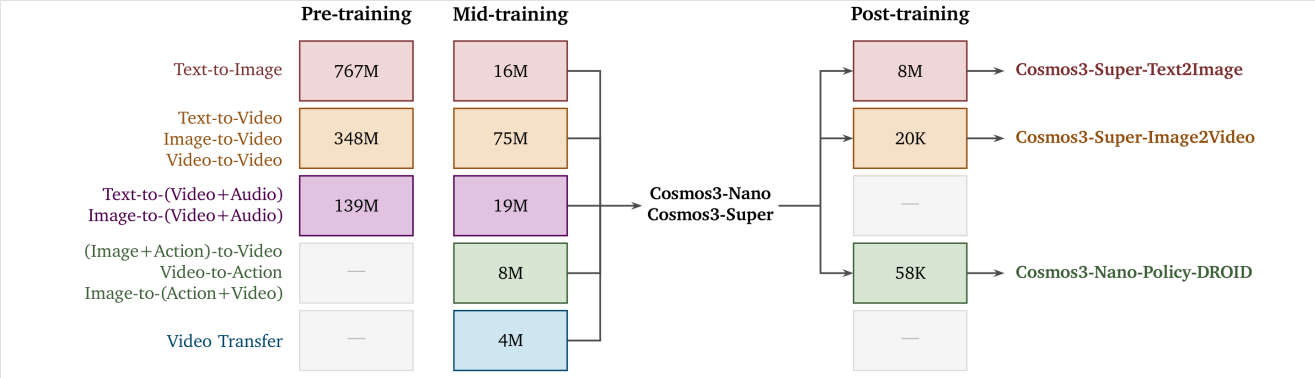


图 8:Generator 数据课程。每一行是一种训练模式;每一列是一个训练阶段。有色单元格给出该阶段所用训练样本数;灰色单元格(—)表示该模式未启用。视频行涵盖文生视频、图生视频与视频到视频(video-to-video)续写;V2V 使用干净的条件视频前缀与加噪的未来视频目标。动作与视频迁移(transfer)数据在中期训练阶段首次引入。中期训练产出基础模型 `Cosmos3-Nano` 与 `Cosmos3-Super`(展示于中期训练与后训练两列之间),它们随后进入后训练。后训练对每种模态独立进行,产出右侧列出的专门化模型:`Cosmos3-Super-Text2Image`、`Cosmos3-Super-Image2Video` 与 `Cosmos3-Nano-Policy-DROID`。我们注意到,这些专门化模型与其对应的中期训练模型架构完全相同。

> **译注** 图 8 各行样本量——文生图 767M / 16M / 8M;视频(文生视频+图生视频+V2V)348M / 75M / 20K;视频+音频 139M / 19M / —;动作 — / 8M / 58K;视频迁移 — / 4M / —。图中视频预训练 348M 为约数,与 §3.2.1 正文的 347.7M 一致(四舍五入)。

3.2.1 图像与视频

图像与视频数据整备遵循一组精心设计的处理、标注与过滤流水线:(1) 采集原始数据并做预处理;(2) 计算嵌入并去重;(3) 对样本分类并做基础过滤;(4) 标注数据;(5) 按分辨率与时长将样本分组为训练就绪的分片(shard)。为提升物理AI场景及其他困难案例的生成质量,我们在视觉生成数据混合中引入了合成数据。最终数据被组织为预训练数据,以及质量更高的中期训练与后训练数据。

预训练。在预训练阶段,我们使用 767M 张图像与 347.7M 个视频片段,它们处理自 7.8B 张原始图像与 3B 条原始源视频。在所得语料中,720p 与 480p 是图像与视频的主导分辨率。具体而言,720p 占图像的 26.8%、视频的 36.4%;480p 占图像的 26.0%、视频的 30.8%。此外,25.2% 的图像与 12.2% 的视频为 1080p 或更高分辨率。最常见的宽高比是 16:9,占图像的 52.0%、视频的 97.3%。对图像而言,第二常见的宽高比是 1:1,占保留图像语料的 25.2%。原始数据经由下述流水线处理与过滤:

- *原始数据采集与处理*。我们从多样的数据源采集数十亿张原始图像与视频,记录原始媒体内容及相关元数据(如原始描述文本与说明)。对视频,我们额外用 TransNetV2 (Souček and Lokoč, 2024) 做镜头切换检测,把长视频切分为时间上连贯的片段。随后用 `ffmpeg cropdetect` 检测并去除黑边,并把所有视频片段重编码为规范格式,以标准化存储并确保可正常回放。
- *嵌入与去重*。原始数据含有大量重复的图像与视频内容,且其概念分布往往高度不均衡。为移除重复内容并为概念均衡与评估打下基础,我们把媒体内容嵌入为向量。图像使用 `Qwen3-VL-Embedding-8B` (Li et al., 2026b);视频使用 `nvidia/Cosmos-Embed1-448p` (NVIDIA et al., 2025b)。我们从完整数据语料中采样 147M 张图像与 400M 个视频片段,对每种数据类型独立运行 cuML KMeans,聚为 20,000 个簇 (Raschka et al., 2020; McQueen, 1967)。然后把每张图像或每个视频片段分配到最近的簇,并在每个簇内基于余弦相似度分数做近重复移除。
- *分类与基础过滤*。我们使用一小套自研 VLM 模型做语义打标与质量过滤。图像与视频数据都被划分为 47 个层次化类别,包括通用域与物理AI域。对图像过滤,一个专用模型标注拼贴图(collage)等属性,并产出美学分与写实度(photorealism)分。只有美学分超过预设阈值的图像才被保留;被标注为拼贴图、水印、白底或 NSFW 的图像被丢弃。对于不以文字渲染为目的的合成图像,我们额外按写实度分过滤,只保留高于指定阈值者。对视频过滤,我们使用三个连续质量分——DOVER 美学质量 (Wu et al., 2023a)、DOVER 技术质量与 VTSS 训练适用性 (Wang et al., 2025d),均为 0–9 分制——并配合约 100 个二元瑕疵标签。重大瑕疵(分屏布局、旋转画面、静止视频)导致直接拒绝;轻微瑕疵(文字叠加、运动模糊、压缩噪声)则被标记但保留在预训练数据中。

中期训练。中期训练阶段的目标是:用精心挑选的高质量数据提升生成质量,并为模型配备额外能力——既包括物理AI等领域专门能力,也包括视频迁移(video transfer)等新任务。数据来自三个来源:(1) 高质量图像与视频;(2) 合成图像与视频;(3) 视频迁移数据。

- *高质量图像与视频*。我们用更严格的过滤规则选择数据,并对困难案例概念加大采样力度,以缓解长尾分布。对真实图像,样本必须满足按宽高比设定的分辨率阈值与严格的 DOVER 美学分下限。我们还纳入合成子集与文字渲染子集:经仔细拒绝采样(rejection sampling)整备的合成图像拓宽了对罕见视觉概念与物体组合的覆盖,而文字渲染图像弥补了图内可读文字的代表性不足。所得中期训练图像混合的有效占比为:真实图像 60%、合成图像 36%、文字渲染图像 4%。对视频,我们类似地施加更严格的分辨率与美学过滤,从预训练视频池中选出片段,占中期训练视频混合的 46.0%。随后我们并入来自机器人、自动驾驶、人类活动与第一人称人-物交互序列的高质量领域专用片段,面向具身AI与操作场景,占混合的另 43.9%。为进一步提升在困难与边角案例概念(如人体运动、高速复杂运动、细粒度操作)上的稳健性,我们围绕这些困难案例采集了额外的能力导向数据,占中期训练视频剩余的 10.1%。
- *合成数据*。尽管我们的预训练语料高度多样,但其概念分布仍呈长尾。因此,模型对一些罕见却重要的物理AI领域与场景(如机器人、自动驾驶与仓库环境)接触相对有限。在这些环境中,模型往往难以理解场景动态、物理交互与长时程行为。为解决这些局限,我们构建了一个大规模合成数据(SDG)语料,包含以下子集:1) 物理交互场景 (`SDG-PhyxSim`),聚焦刚体碰撞、铰接物体动力学、可形变材料、流体动力学与光学效应;2) 具身机器人场景 (`SDG-RobotSim`),覆盖 6–8 种机器人具身形态(embodiment)与多样任务类别的操作与移动序列;3) 自动驾驶场景 (`SDG-DriveSim`),涵盖常规与边角交通场景;4) 数字人场景 (`SDG-SynHuman`),旨在改善人体动态、相机运动先验与多角色交互的建模;5) 仓库作业场景 (`SDG-Warehouse`),面向仓库安全,包含人-叉车交互场景。合成数据的构建细节与分析请参见附录 C。我们发布全部 SDG 数据集以支持社区。
- *视频迁移数据*。迁移数据为 Generator 配备控制条件生成能力。给定空间控制信号(如边缘图、模糊帧、深度图、分割图或世界场景图)与文本描述,模型被训练生成 RGB 视频。我们从预训练视频池中选出 3M 条视频,侧重高质量视频与机器人、自动驾驶等物理AI领域。对边缘与模糊控制,控制信号在训练时用随机采样参数的 Canny 边缘检测、高斯模糊与双边滤波在线(on the fly)计算。对深度与分割控制,我们分别用 Video Depth Anything (Chen et al., 2025c) 与

SAMv2 (Ravi et al., 2025) 预计算控制信号。对世界场景图控制,我们使用由 Cosmos-Drive-Dreams (Ren et al., 2025a) 采集的 MADS 数据集。MADS 含 1.1M 个样本,每个样本带七路同步相机视角:前广角、前长焦、交叉左、交叉右、后左、后右与后长焦。视频以 30 FPS 录制,并附带逐相机的世界场景图控制输入,编码车道线、道路边界、交通灯以及车辆与行人的动态 3D 包围框。该数据集覆盖 14 个地理区域,包括美国、德国、日本与英国,共 25 个"国家-时长"分区。

后训练。我们为训练领域专门化的 Cosmos 3 模型(如 `Cosmos3-Super-Text2Image` 与 `Cosmos3-Super-Image2Video`)构建后训练数据集。

后训练图像语料是一个紧凑、精心整备的集合,取自三个来源:合成图像、文字渲染图像与高质量真实图像。通用的网络规模预训练数据被完全排除,转而采用直接针对生成质量与能力短板的高保真内容。

后训练视频语料由两部分组装而成。第一部分是预训练视频语料的一个子采样子集,起正则化作用:防止模型对更窄的后训练分布过拟合,同时保留预训练获得的更广泛视觉知识。第二部分也是主要部分,是一个紧凑的监督微调(SFT)视频集,专门为弥合系统性评估中识别出的生成质量差距而整备。SFT 视频集由三个子来源构成:第一是合成视频,覆盖难以从真实世界素材中获取的多样视觉概念、运动类型与场景构成;第二是人工整备的真实 SFT 视频,由人工挑选与标注,为各关键视觉域的生成保真度提供直接的质量信号;第三是从预训练语料中检索到的真实视频,经嵌入相似度选出,以确保覆盖常见失败案例。

结构化描述文本标注。描述文本质量是生成质量的关键因素。高质量的描述文本应忠实刻画图像或视频中的实体、属性、关系与整体场景内容。对视频,描述还应进一步捕捉时序动态,包括物体运动、人类动作、物理变化、交互与相机运动。为提升描述质量,我们进行了多轮设计迭代,并在所有训练阶段的全部数据上采用结构化 JSON 标注格式,而非稠密的自由形式自然语言描述。我们的实验表明,自由形式描述往往精确但不完整:它们倾向于准确描述可见内容,却在复杂场景中遗漏重要细节。相反,丰富的预定义结构鼓励对物体、属性、关系与场景级信息的系统性覆盖,在保持高精确率的同时提升召回率。

我们的结构化格式捕捉一组广泛的视觉属性,包括主体、背景、光照、美学与摄影语言(cinematography)。对视频,我们额外引入时序动态字段,范围从物理形变与物体交互到复杂人体运动。我们在结构化标注数据上微调了两个 `Qwen3-VL-8B` 模型,分别作为图像与视频的自研打标模型(captioner)。打标模型与完整结构化描述模式(schema)的更多细节请参见附录 A。

为定量且严格地评估标注质量,我们为图像与视频设计了一个专门的描述质量基准。该基准聚焦难以描述的样例,强调物理AI等对物体、空间关系、动作与时序动态的准确描述至关重要的领域。对每条模型生成的描述,我们在断言(assertion)层面用两种不同方法计算精确率与召回率。精确率直接对照源媒体评估,以惩罚幻觉:由一个 VLM 将生成的描述分解为原子论断,并逐条校验其是否得到图像或视频本身的视觉支持。召回率则衡量全面性,依赖人工整备的真值:为实现可靠、可追溯的召回评估,我们把视频或图像的视觉内容分解为一份原子断言清单,覆盖实体、属性、关系、事件及其他相关细节;然后由一个 LLM 将生成的描述与这份真值断言清单交叉比对,判定哪些关键细节被成功捕捉。该协议使我们不仅能评估描述的事实准确性,还能评估其是否包含训练高质量生成模型所需的关键视觉信息。在该基准上,我们的结构化标注方法在保持高精确率的同时显著提升了召回率。

3.2.2 音频

音视频成对数据教给 Generator 的不仅是"应该出现什么声音",还有"该声音相对可见事件应在何时发生"。原始网络视频的音频对此颇具挑战:旁白与解说往往只是描述视频而非由视频引起,而人为添加的背景音乐(BGM)可能掩盖屏上事件产生的物理声音。因此我们在不同训练阶段以不同方式使用音频:预训练保留广泛的声学覆盖,而中期训练通过一套针对语音与非语音音频的显式筛选策略,构建更高精度的音视频对。

预训练。音频预训练语料完全派生自预训练视频池。总计 138.9M 个预训练片段含有可用音轨,覆盖画内(diegetic)与画外(non-diegetic)语音、解说、BGM、环境声、音乐与物理事件声的广泛混合。其中 62.5M 个片段短于 30 秒;对这一子集,我们用 `Qwen3-Omni-Captioner` (Xu et al., 2025) 生成合成音频描述。该阶段偏重规模与多样性,让模型先接触真实视频音频的长尾分布,再在中期训练中施加更严格的整备。

中期训练。中期训练音频池从预训练音视频语料中过滤而来,以改善因果性的音画对齐。最终池含 18.8M 个片段:12.8M 个非语音片段用于环境与物理声音生成,6M 个语音同步片段用于视觉锚定的语音生成。整备流水线围绕一条简单原则组织:仅当语音与可见人脸同步时才保留语音;从非语音样本中移除画外语音;当非器乐 BGM 会主导目标音频时将其移除。

- *音源分离*。对每个候选片段,SAM-Audio (Shi et al., 2025a) 把原始音频分离为语音声部(speech stem)与剩余声部(remaining stem)。语音声部用于识别视觉锚定的语音,剩余声部则为非语音音频生成提供一个人声被抑制的候选。
- *唇同步打分*。在语音声部与原始视频上运行 SyncNet (Chung and Zisserman, 2016),产出 `has_face` 与 `lip_sync_confidence`。我们把 `speech_synced` 定义为 `has_face` = True 且 `lip_sync_confidence` ⩾ 3.0(如 LatentSync (Li et al., 2024b) 所示)。
- *音频事件检测*。在原始音频上运行 FireRedASR2S (Xu et al., 2026),估计 `speech_ratio`(被标为语音或歌唱的时间占比)与 `music_ratio`(被标为音乐的时间占比)。我们把 `high_music` 定义为 `music_ratio` ⩾ 0.1。
- *乐器检测*。对所有 `high_music` 片段(包括语音同步的片段),用 Qwen3-VL (Bai et al., 2025b) 预测 `is_music_instrument`。这保护了乐器演奏视频——其音乐本身就是可见事件的一部分,而非可移除的 BGM。
- *语音分支*。满足 `speech_synced` 的片段构成语音中期训练池。我们保留原始音频,除非该片段为 `high_music` 且 `is_music_instrument` 为 False;此时用 SAM-Audio 从原始波形中移除音乐。该编辑路径仅当以下条件满足时才被保留:对候选音频的第二次 FireRedASR2S 检测报告 `speech_ratio` ⩾ 0.05 且 `music_ratio` = 0,且候选按 `max_abs` ⩾ 0.007、`p50_db` ⩾ −80、`active_ratio` ⩾ 0.2 判定为非近似静音。这样既保留了唇同步的语音,又抑制了非器乐 BGM。
- *非语音分支*。未被分入语音同步分支的片段,被整备用于非语音的音视频生成。我们首先选择基础波形:若 `speech_ratio` ⩾ 0.05,则使用 SAM-Audio 的剩余声部以移除人声;否则保留原始音频以保存物理声音。若片段为 `high_music` 且 `is_music_instrument` 为 False,则对所选基础波形施加 SAM-Audio 音乐移除。派生自剩余声部的候选必须通过第二次 FireRedASR2S 检测,要求 `speech_ratio` < 0.05;其 `music_ratio` 在施加了音乐移除时必须 = 0,否则须 < 0.1。派生自"原始波形+音乐移除"的候选必须满足 `music_ratio` = 0。处理后的候选还必须通过与语音分支相同的非静音阈值。
- *描述文本标注*。描述与最终用于训练的波形绑定。若所选波形为原始音频,则保留原始音频描述;若音源分离或音乐移除改变了波形,则对最终候选音频重新打标,使文本不再描述已被移除的语音或伴奏。对语音同步池,我们用 `Qwen3-ASR` (Shi et al., 2026) 转写每条选中的语音轨,再用 `GPT-OSS-120B` (Agarwal et al., 2025) 将转写文本与音频描述合并。所得描述同时指明说话内容与非语言的声学上下文,并保持与视频配对音频的忠实性。

3.2.3 动作

动作提供了连接跨时间世界观测状态的因果变量。仅用视频训练虽然能教会生成器外推可能的运动,却不能让模型接触可控的干预:同一初始观测在不同机器人指令、相机轨迹、车辆路线或人手运动下可能演化出不同结果。因此我们在中期训练阶段引入文本-视频-动作成对数据,使 Cosmos 3 能学习"世界-动作"关系的两个方向:在动作条件下预测未来观测,推断能解释一段观测轨迹的动作,以及联合生成动作与未来视频。

数据统计。我们把动作中期训练聚焦于四个物理AI支柱:第一人称运动(egocentric motion)、机器人、自动驾驶车辆与相机运动。最终整备的数据包含 8.4M 个回合(episode)、共 61.3K 小时,如图 9 所汇总。

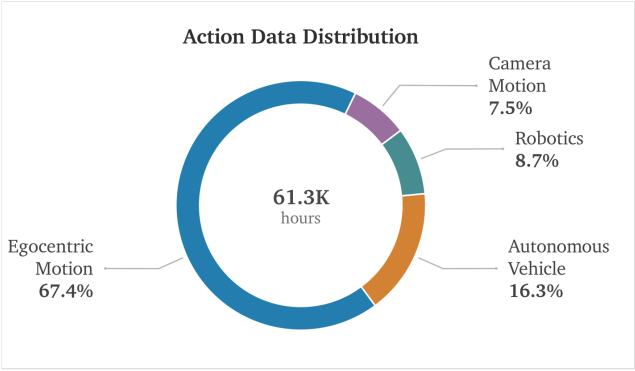


图 9:动作数据分布。小时数按最终整备的动作中期训练集中四个主要动作数据支柱聚合,该数据集含 8.4M 个回合与 61.3K 小时。

表 4:机器人数据明细。按机器人具身形态分组。

Embodiment	数据来源	任务数	回合数	小时数
AgiBot	Bu et al. (2025)	338	239.4K	4.37K
Franka Panda	Wu et al. (2025a); Khazatsky et al. (2024)	67.5K	76.3K	442
Google Robot	Brohan et al. (2023b)	599	87.2K	351
WidowX-250	Walke et al. (2023)	21.8K	50.4K	100.1
UMI	Lin et al. (2025a); Ha et al. (2024); Liu et al. (2024e); Chi et al. (2024); Liu et al. (2025a); Wu et al. (2024)	43	38.3K	67
UR	Wu et al. (2025a)	114	25.0K	35
总计	–	90.4K	516.7K	5.36K

- *第一人称运动*。第一人称运动数据贡献 41.3K 小时(67.4%),是最大的组成部分。它包含来自一个专有数据集的 1.7M 个回合,内容为头戴式 RGB 相机采集的双手操作。每帧标注有同步的头部相机位姿,以及每只手的 21 关键点 3D 姿态 (Zimmermann and Brox, 2017; Simon et al., 2017)——提供相机坐标系下的逐关节位置与朝向——使模型能够联合学习第一人称自身运动与细粒度灵巧手部运动。
- *自动驾驶车辆*。自动驾驶车辆数据贡献 10.0K 小时(16.3%),派生自使用 NVIDIA Hyperion 平台采集的高质量内部驾驶日志。该数据集通过在大规模语料中挖掘、以匹配一个横跨多样驾驶场景的目标分布来构建。所选场景覆盖广泛条件,包括多样的天气、光照与路况,以及丰富的纵向与横向操控动作,而不局限于以近直线巡航为主。为与其他领域对齐,我们把驾驶轨迹从车辆坐标系变换到前广角相机坐标系。
- *机器人*。机器人数据贡献 5.4K 小时(8.7%),汇聚自开源数据集。该子集含 90.4K 个任务与 516.7K 个回合,按具身形态与来源的明细见表 4。为避免具身形态相关的控制器细节(如 PID 参数或底层驱动接口),我们使用由状态差分导出的伪动作(pseudo-action)。我们同时整备成功与失败的回合,使模型不仅观察到按预期完成的动作效果,也观察到偏离正常(off-nominal)的动作效果。
- *相机运动*。相机运动数据贡献 4.6K 小时(7.5%),挖掘自我们的预训练视频数据集。我们用 ViPE (Huang et al., 2025) 与 DepthAnything3 (Lin et al., 2025b) 估计相机位姿,把这些视频转换为动作轨迹。为确保数据质量,我们对数据集做了严格过滤,移除位姿估计不可靠的片段,例如抖动过大或相机内参异常者。所有相机位姿保持度量尺度,并转换到统一的动作坐标约定。该整备流程产出一个含 1.9M 个片段的数据集。

数据处理流水线。我们用 §2.1.3 描述的统一动作词元化(tokenization)转换每个数据源。为在转换后跨具身形态平衡动作幅值,我们从训练数据计算逐维归一化系数,把动作通道缩放到大致 [−1, 1] 的可比范围。对含多路同步视角的数据,我们把各视角拼接成一张画布(canvas),并将相机布局存入元数据,如图 30 所示。我们没有把空闲(idle)操作过滤掉,而是保留它们并在元数据中记录空闲步数,使下游采样能够显式地平衡活跃段与非活跃段。

4. 训练 Training

我们分两个主要阶段训练 Cosmos 3。首先,Reasoner(推理器)在大规模图文与视频—文本语料上进行预训练(pre-training),随后在精心筛选的物理AI(Physical AI)混合数据上微调,得到一个具备视觉理解与推理能力的强多模态骨干。由于 Reasoner 与 Generator(生成器)共享相同的 transformer 块架构,训练好的 Reasoner 权重随后被用于初始化 Generator,将语义与世界知识迁移到一个能够合成像素、音频与动作的模型中。Generator 采用渐进式多阶段课程(curriculum)训练:先进行大规模图像、视频与音频预训练,再经中期训练(mid-training)逐步引入动作与迁移(transfer)数据;最后在规模更小、精心筛选的物理AI数据集上进行后训练,以改进下游行为、物理一致性与动作保真度。

4.1 Reasoner 训练

Cosmos 3 Reasoner 的训练分两个阶段:大规模多模态预训练,以及在精选物理AI任务上的监督微调(SFT)。在预训练阶段,模型从大规模图文与视频—文本语料中学习通用的多模态表征;监督微调则使模型专门化到物理AI领域——包括机器人、自动驾驶与智慧基础设施应用——同时保留预训练阶段获得的广泛能力。

4.1.1 预训练

Reasoner 预训练从一个语言模型和一个通过多模态投影器(projector)相连的 ViT 编码器出发。在初始化上,我们同时试验了附录 D 中描述的内部预训练模型与开源的 `Qwen3-VL` 系列模型:Edge 模型使用我们的内部模型,Nano 模型使用 `Qwen3-VL-8B`,Super 模型使用 `Qwen3-VL-32B`。与以往做法不同——先冻结 VLM 其余参数、单独训练投影器完成对齐阶段 (Bai et al., 2025a)——我们发现这种分阶段对齐并无必要,因而从预训练一开始就联合训练所有组件。

模型在 §3.1.1 描述的大规模多模态语料上以下一词元(token)预测为目标进行训练。我们使用不放回采样器均匀拼接所有数据集,在完整预训练混合数据上训练两个 epoch。由于许多物理AI应用要求高效推理与低延迟,我们将训练序列长度限制在最多 16k 词元,单样本上限为 2048 个图像词元、8192 个视频词元。

参照先前工作 (Bai et al., 2025a; Wang et al., 2025e),我们采用平方根归一化的逐词元损失加权,以平衡短序列与长序列的贡献。我们发现这一归一化策略显著改善了下游基准得分和整体训练稳定性。

优化使用 AdamW:语言模型与投影器的峰值学习率为 5×10^{-5},ViT 为 5×10^{-6}。所有学习率在 10% 的线性预热(warm-up)之后按余弦衰减到峰值的 0.1 倍。Adam 系数取 $(\beta_1, \beta_2) = (0.9, 0.999)$。训练还使用 0.05 的权重衰减,并按全局范数阈值 1.0 做梯度裁剪。

4.1.2 监督微调

为使模型适配下游物理AI任务,我们在一套精选的高质量多模态混合数据上进行监督微调。与预训练中各数据集跨 epoch 均匀采样不同,监督微调采用重要性感知的采样策略:根据重要性、质量与规模为每个数据集分配固定的采样预算。这使优化能聚焦于高价值的下游任务,同时仍维持跨领域、跨能力的多样性。

为防止下游专门化损害模型的通用推理与视觉理解能力,我们额外掺入经过筛选的高质量预训练数据子集,预训练与 SFT 的采样预算比固定为 1:4。保留一小股预训练数据流可提高鲁棒性、保持指令遵循行为,并在多个基准上维持较强的通用领域能力。我们还在监督微调混合数据中加入一个轻量级指令遵循数据集(80万样本),以在任务适配期间进一步稳定对话与指令遵循能力。

训练共进行 8200 次迭代,全局 batch size 为 512。优化器为 AdamW:语言模型与投影器峰值学习率 1×10^{-5},ViT 为 1×10^{-6}。所有学习率在 1000 步线性预热后按余弦衰减到峰值的 0.1 倍。Adam 系数取 $(\beta_1, \beta_2) = (0.9, 0.95)$。权重衰减为 0.1,梯度裁剪全局范数阈值为 1.0。

> **译注** SFT 阶段与预训练阶段的 Adam β_2 不同(0.95 对 0.999),权重衰减也从 0.05 提高到 0.1——这是微调阶段常见的"更快适应+更强正则"配置,原文未加评论,但两组超参并非笔误。

4.2 Generator 训练

Cosmos 3 Generator 采用渐进式多模态课程训练,旨在跨多种分辨率、时长与条件模态,联合建模视觉、听觉以及动作条件下的世界动态。训练配方强调可扩展性、高保真生成与高效的长上下文学习。预训练阶段,模型从覆盖图像、视频与音频的大规模数据中学习通用生成先验;后续训练阶段逐步引入更丰富的多模态监督——包括动作与迁移序列——使模型学会时间上连贯的世界演化与有物理根据的交互。

训练目标。Cosmos 3 Generator 在所有模态上均以修正流匹配(rectified flow matching)为优化目标。对任一模态的目标潜变量(latent),我们通过直线插值构造带噪潜变量:

$$x_\sigma = \sigma \cdot \varepsilon + (1 - \sigma) \cdot x_0\,,\ \ \varepsilon \sim \mathcal{N}(0, I),\ \ \sigma \in [0, 1]$$

其中 x_0 是干净目标,$\varepsilon \sim \mathcal{N}(0, I)$,$\sigma \in [0, 1]$ 为噪声水平。单个去噪器 $v_\theta(x_\sigma, \sigma, c)$ 被训练为通过带掩码的均方误差预测恒定速度 $v^* = \varepsilon - x_0$;条件词元(例如图生视频任务中的干净条件帧)被从损失中屏蔽。我们对各模态独立采样时间——对图像、视频、音频、动作分别独立抽取噪声水平 σ。参照 Waver (Zhang et al., 2025c),图像、音频与动作批次使用 logit-normal 噪声分布,视频批次使用众数采样(mode sampling);我们发现众数采样能带来更好的生成质量。我们进一步通过修正流平移(shift)重参数化对 t 做映射:

$$\sigma = s \cdot \bar{t} \,/\, (1 + (s - 1) \cdot \bar{t}),\ \ \bar{t} = 1 - t$$

其中 $s \geqslant 1$ 使噪声的边缘分布偏向更高噪声水平。

4.2.1 预训练

在预训练阶段,我们联合训练模型在多种分辨率与生成任务下生成图像、视频与音频。为此,我们采用多分辨率训练策略,并在多个生成任务上联合优化模型,包括文生图(Text-to-Image)、文本生成"视频+音频"、图像生成"视频+音频"、以及视频生成"视频+音频"。

多分辨率训练。我们不固定单一输出分辨率,而是同时在三个分辨率档(256p、480p、720p)、五种宽高比和可变帧数上训练,如表 5 所示。这既让模型接触高保真内容,又鼓励其学到分辨率无关的表征。训练数据据此划分:256p 流从完整数据集抽取(所有原生分辨率均可);480p 流仅限原生分辨率不低于 480p 的素材;720p 流只用不低于 720p 的内容,以在最高档保留锐度和细节。各分辨率档的最大帧预算不同:256p 与 480p 最多 400 帧,720p 最多 300 帧;720p 限制到 300 帧是出于序列长度约束。训练批次按 1:1:2:1 的比例混合四个流——纯图像、视频-256p、视频-480p、视频-720p。我们发现这一分布在高保真学习与样本多样性之间取得了良好平衡,使模型既能看到更多训练样例,又仍侧重较高分辨率内容。我们使用随分辨率自适应的平移值:256p 时 $s = 1$,480p 时 $s = 3$,720p 时 $s = 5$。

为在支持可变序列长度的同时避免无谓的重编译开销,我们采用词元打包(token packing),每条序列固定预算 74,000 词元。不同分辨率的序列被一起打包填满每个批次,在不引入 padding 的情况下最大化 GPU 利用率(如图 10 所示)。

表 5:图像/视频模型规格。图像与视频模态支持的配置。每行给出该分辨率下的 FPS 范围、帧数范围(仅视频),以及五种受支持宽高比下的图像/视频尺寸 (w, h)。

分辨率	视频		各宽高比下的图像/视频尺寸 (w, h)				
	FPS	帧数	16:9	4:3	1:1	3:4	9:16
256p	10–30	5–400	(320, 192)	(320, 256)	(256, 256)	(256, 320)	(192, 320)
480p	10–30	5–400	(832, 480)	(736, 544)	(640, 640)	(544, 736)	(480, 832)
720p	10–30	5–300	(1280, 720)	(1104, 832)	(960, 960)	(832, 1104)	(720, 1280)

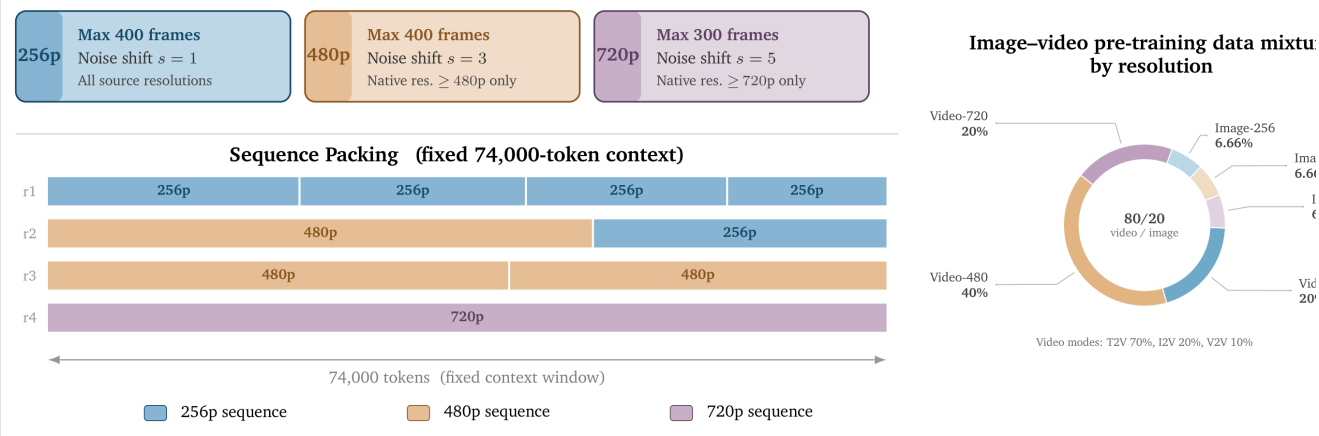


图 10:左:多分辨率训练与序列打包。三个分辨率档(256p、480p、720p)在最大帧预算、可用源素材与修正流噪声平移值上各不相同;来自不同档的变长序列被一起打包,填满固定的 74,000 词元上下文窗口,在不引入 padding 的情况下最大化 GPU 利用率。右:Generator 预训练所用的数据混合。我们采用图像—视频联合训练:80% 的时间采样视频,其余 20% 采样图像。在每一部分内部,我们在多个分辨率(256p、480p、720p)上训练。对视频批次,我们进一步在三种条件模式——文生视频、图生视频、视频生视频——之间均匀采样。具体数据混合见右图。

> **译注** 图 10 右图的精确配比为 Video-480 占 40%、Video-256 与 Video-720 各占 20%,三档图像各占 6.66%(合计 20%);视频条件模式为 T2V 70%、I2V 20%、V2V 10%。这与下文正文的 20%/56%/16%/8% 一致:80% × 70% = 56%,80% × 20% = 16%,80% × 10% = 8%。

训练模式。对形状为 C × T × H × W 的潜空间视频张量,设 T_{cond} 为条件潜帧数,T_{noised} 为加噪潜帧数($T = T_{cond} + T_{noised}$)。训练时,前 T_{cond} 帧不加噪声,作为条件输入;只有其余 T_{noised} 帧被加噪,模型学习对其去噪。T_{cond} 与 T_{noised} 的不同取值产生不同的训练模式。我们使用四种生成模式——文生图、文生视频、图生视频、视频生视频——彼此仅在前置于每个样本的条件视觉帧数上不同,采样比例分别为 20%、56%、16%、8%。所有模式均使用 §3.2 描述的结构化 JSON 描述(caption)格式。

- **文生图(T2I)**。图像被视为时间维度收缩为 T = 1 的视频特例。此模式下,图像从全部三个分辨率档与各宽高比中随机抽取,再经序列打包送入模型。由于单张图像产生的词元数远少于视频,一条典型序列中包含的样本数远多于视频序列。
- **文生视频(T2V)**。文生视频训练中 $T_{cond} = 0$,模型学习仅以文本为条件对整段视频去噪。除描述文本外,模型还在 JSON caption 中接收时长、FPS 与时间戳元数据作为附加字段,从而能够生成指定长度与时间跨度的视频。
- **图生视频(I2V)**。单帧条件($T_{cond} = 1$)时,第一个潜帧保持干净,后续帧被加噪。模型学习生成与初始帧和描述文本都一致的未来帧。
- **视频生视频(V2V)**。多帧条件($T_{cond} = 2$)时,模型以视频的前五帧(等价于前两个潜帧)为条件,学习预测与条件帧和输入提示词都一致的未来帧。

> **译注** "前五帧等价于前两个潜帧"透露了 tokenizer 的时间压缩方式:首帧单独编码为 1 个潜帧,其后每 4 帧压成 1 个潜帧(1 + 4 = 5),与 Cosmos 系列因果视频 tokenizer 的惯例一致。

FPS 调制。我们以变化的 FPS 值训练模型,因此词元之间的物理时间间隔因样本而异:以 30 FPS 采样的片段在真实时间上比同样词元数、16 FPS 采样的片段排布更密。为反映这一点,我们调制 3D MRoPE 位置编码的时间轴,按真实世界时间而非词元下标来分配时间坐标(见 §2),基准帧率为 24 FPS。时长与 FPS 同时被追加到文本提示词中,使模型在推理时可以按指定的时间特性进行条件生成。

优化。Generator 预训练期间只更新生成专属参数;Reasoner 塔保持冻结,以保留语言与视觉理解能力。我们使用 FusedAdamW,学习率 10^{-4},$(\beta_1, \beta_2) = (0.9, 0.99)$,权重衰减 0.05,按范数 1.0 做梯度裁剪。学习率调度为带预热的线性衰减,在 n 次迭代内从峰值降到峰值的 0.30 倍下限。为支持无分类器引导(CFG),我们在所有模态上使用 10% 的文本丢弃(text-dropout)率。

训练词元量。在预训练阶段,`Cosmos3-Nano` 使用 1024 块 NVIDIA GB200 GPU 训练了 31.05T 词元;`Cosmos3-Super` 使用 2048 块 NVIDIA GB200 GPU 训练了 17.86T 词元。

> **译注** 小模型 Nano 见过的词元(31.05T)反而比大模型 Super(17.86T)多近一倍——这符合"小模型用更多数据补容量"的常见算力分配,读对比数字时不要误以为 Super 训练更充分。

4.2.2 中期训练

中期训练弥合广谱预训练与下游部署之间的差距。此时 Generator 已从大规模数据中学会了通用的图像、视频与音频生成,但目标物理AI应用要求对稀有动态、具身场景、控制接口和高质量视觉领域有更强的覆盖。因此我们从预训练检查点继续训练,采用一套既保留原有视觉生成模式、又引入新监督来源的精选混合数据。该阶段有两个互补目标:领域专门化——加大高价值物理AI领域的曝光;多模态整合——把模型从视觉与音频生成扩展到动作与控制条件下的世界建模。

领域专门化。在保留通用知识的同时,让模型接触高度精选的专门数据集,以提升其在应用关键的物理AI场景中的质量与可靠性。图像方面,我们使用一个 1560 万样本的中期训练数据池,以高质量真实影像为主,并加入合成与文字渲染数据,以扩大概念覆盖并保持可辨认的文字生成能力。视频方面,我们纳入 7470 万条精选片段,覆盖机器人、自动驾驶、人类活动、物理现象与合成仿真数据。这些数据源针对的是通用网络规模预训练中覆盖不足的失效模式,例如长时程交互、细粒度的人与机器人运动、物理物体动力学,以及安全攸关的驾驶或仓储场景。通过把这些领域聚焦数据集与既有的图像、视频训练模式混合,中期训练在不丢弃预训练所学广谱视觉先验的前提下提升了物理AI相关性,如 §3.2.1 所述。

多模态整合。中期训练把 Generator 从图像、视频与音频生成扩展为一个统一的物理AI模型,使其还能读入并合成动作与控制信号。我们沿用预训练中 T2I、T2V、I2V、V2V 的"干净前缀/加噪目标"形式,既有视觉能力保持激活,同时在扩散子序列中引入新的模态专属词元。这使动作、音频、控制与视频词元共享 §2 所述的同一时间坐标系与双向注意力模式。在预训练模式之外,我们新增两族多模态监督:动作与视频迁移(transfer)。

表 6:Generator 中期训练数据混合。预训练完成后,我们在中期训练阶段引入新模态(动作与迁移),数据比例如下。

训练流	模式 / 条件	占比
图像	T2I	10%
视频	T2V、I2V、V2V	32%
视频 + 音频	T2(V+Audio)、I2(V+Audio)、V2(V+Audio)	8%
动作	正向动力学、逆向动力学、策略(policy)	25%
通用迁移	边缘、模糊、深度与分割控制	20%
驾驶迁移	世界场景图(world-scenario-map)控制	5%

- **动作**。我们引入使用 §2.1.3 统一动作表示的文本—视频—动作配对训练数据。模型不仅被训练为在动作条件下预测未来视频,还要从观测轨迹中推断动作,以及联合生成动作与视觉未来。这教会 Generator 一个连接"可控干预"与"世界演化"的因果接口。
- **视频迁移**。我们加入控制条件的迁移数据:干净的控制信号作为输入,模型对相应的目标图像或视频去噪。控制信号包括来自高质量视频语料的边缘、模糊、深度与分割图,以及驾驶场景的世界场景图。这让模型在保持文本条件与视觉生成质量的同时,学会遵循空间上有依据的约束。

各模态的混合比例见表 6。

多分辨率训练。与预训练类似,中期训练在固定的 74K 上下文窗口内跨 256p、480p、720p 多分辨率进行。为更好地处理动态并减少时间维与高分辨率伪影,我们将修正流平移值提高:256p、480p、720p 分别为 3、5、10。

训练目标。与预训练类似,所有模态都使用修正流目标。对动作,我们沿用视觉的噪声调度。中期训练的总损失是各模态速度 MSE 按模态专属损失系数加权后的和;其中动作损失乘以 10 倍,以补偿归一化动作向量逐元素 MSE 偏小的问题。

优化。与预训练类似,我们使用 FusedAdamW,学习率 10^{-4},权重衰减 0.05,按范数 1.0 做梯度裁剪,损失系数(loss scale)为 10。学习率采用 LambdaLinear 调度,起始因子 0.4,周期长度 100,000。

训练词元量。中期训练阶段,`Cosmos3-Nano` 使用 1024 块 NVIDIA GB200 GPU 训练了 2.4T 词元;`Cosmos3-Super` 使用 2048 块 NVIDIA GB200 GPU 训练了 1.9T 词元。

4.2.3 文生图后训练

为展示 Cosmos3-Super 的全模态(omnimodal)能力,我们把模型进一步专门化为一个文生图检查点 `Cosmos3-Super-Text2Image`。我们的目标是把模型有物理根据的世界理解迁移到高质量图像生成中,在追求开源第一梯队 T2I 效果的同时提升物理合理性与场景级对齐。

文生图专门化采用两阶段 SFT,遵循文生图基础模型的常见训练范式:先做语义增强,再做面向偏好的精调。

- **第一阶段:广谱 T2I 专门化**。我们在精选的高质量 SFT 数据集上微调 20k 步。训练混合数据按受控比例采样:45% 通用真实图像数据、40% 合成图像数据、15% 纯文字渲染数据,以平衡视觉保真度、图文对齐与语言能力保持。基础学习率为 1×10^{-4},预热 2k 步,学习率线性衰减,其余超参数与 Cosmos 3 中期训练阶段保持一致。
- **第二阶段:高质量精调**。我们用 47 万对精挑细选的超高质量图像—描述对做最后一轮 2k 步的 SFT。该阶段进一步改善视觉美感、提示词遵循、文字渲染质量以及与人类偏好的对齐。
- **分辨率与上下文长度**。两个阶段都使用 70k 词元的固定上下文窗口,并且只在分辨率高于 720p 的图像上训练。

总体上,`Cosmos3-Super-Text2Image` 在语义对齐与英文文字渲染两类基准上都取得了强劲的文生图成绩。在 UniGenBench 上,它在受评模型中总分最高,完整基准得分 91.36(见表 11)。配合一个智能体(agentic)工作流,该模型在 Artificial Analysis 文生图榜单上位列开放权重模型第一(§6.2.1)。这些结果表明,从 Cosmos 3 出发做下游 T2I 模态适配可能是高度有效的:它在提升场景级提示词对齐的同时,保留了模型有物理根据的生成能力。

> **译注** 两处榜单第一均限定"开放权重模型"范围,且 T2I 的 top-1 成绩是在"配合 agentic 工作流"条件下取得的,并非裸模型单次出图的名次;此外后训练上下文窗口为 70k 词元,与预训练/中期训练的 74k 不同。

4.2.4 图生视频后训练

图生视频能力对全面的视觉理解至关重要:它检验模型对物理定律、物体恒存性与复杂场景几何的理解,同时也是具身AI与机器人规划的关键预测机制——模拟合理的未来帧即可得到一个有效的世界模型 (Wiedemer et al., 2025; Chen et al., 2025a)。虽然 Cosmos 3 天生就能原生处理多样任务,我们仍通过 SFT 显式展示并专门化其在 I2V 领域的潜力。为此,我们采用如下流程:

- **数据与训练混合**。我们用经过筛选的预训练数据微调模型:这些数据先为更均衡的主题多样性做了精炼,又通过一个智能体工作流增强——该工作流识别模型薄弱点并从预训练集中检索针对性样例。在此基础上,再结合 1,000 条高质量人工精选视频,以及约 2 万条覆盖多样主题的合成视频片段数据集(约占总词元的

6%)。所有视频序列都只用 I2V 形式训练,但训练混合中还纳入 20% 的 T2I 图像词元,以保持模型的语义对齐。

- **分辨率与时长**。我们把模型专门化到 480p 的目标分辨率、189 帧的目标时长——按 24fps 约合 8 秒。这一配置在推理速度与时间上下文之间取得平衡,使模型能在有意义的时间跨度上快速生成物理上合理的视频。
- **训练调度**。I2V 后训练阶段共 10k 次迭代,学习率 1×10^{-5}。整个 SFT 过程模型约处理 500 亿(50B)词元。

经过后训练,`Cosmos3-Super-Image2Video` 在图生视频生成上达到领先质量;特别地,该模型在 Artificial Analysis 图生视频榜单上位列开放权重模型第一(§6.2.2)。该模型的使用细节请参见 §6.3.1。

4.2.5 机器人策略后训练

我们开展机器人策略(policy)后训练,以考察 Cosmos 3 全模态世界模型能否扩展为强大的机器人策略模型。中期训练已使 Cosmos 3 能建模包含语言、视觉观测与动作的多模态序列,并能联合生成动作与视频。我们进一步为机器人策略学习定制模型:引入本体感知(proprioceptive)信号、降低推理延迟,并使模型能为闭环控制产出可执行的动作。

作为先导研究,我们选用 DROID 机器人平台与数据集 (Khazatsky et al., 2024),因其流行度高、社区采用广泛。DROID 平台使用 Franka Panda 7 自由度机械臂搭配 Robotiq 2F-85 平行夹爪,在多样的真实环境中执行桌面操作任务。DROID 数据集包含 7.6 万条轨迹、350 小时交互数据、86 个任务、564 个场景,为真实世界机器人策略学习提供了可观的规模与广泛的任务多样性。我们以 360×640 的较高分辨率接入 DROID,应用社区提供的静止帧过滤与失败演示剔除,并在训练中使用随机图像增强。

我们从中期训练完成的 `Cosmos3-Nano` 出发继续训练得到 `Cosmos3-Nano-Policy-DROID`,其中动作编码器、动作解码 MLP 与动作嵌入词元均为全新初始化。我们对动作相关参数施加 5 倍学习率乘数,以加快适应。策略输入由当前的本体感知机器人状态和三视角视觉观测组成:腕部视角图像(原始分辨率 360×640)置于上方,两幅外部视角图像(各为原始分辨率 180×320)左右并排拼接在左下与右下,合成画布为 540×640。策略被训练为预测未来 32 个绝对关节位置动作,并附带输出辅助的 RGB 视频帧,运行频率 15Hz。本次后训练研究使用 DROID 官方的简短任务指令作为提示词。学习率为 2×10^{-4},其余超参数沿用中期训练设置。

推理时,我们用 4 步扩散采样,噪声调度平移值为 5;同时施加无分类器引导(CFG),采用 CFG 并行,引导系数为 3,并跳过视频潜变量解码以进一步降低推理开销。这些优化合在一起带来显著的推理加速,使策略服务器可部署在 2 块 NVIDIA RTX Pro 6000 GPU 上。下游关节位置控制器基于 `Franky` (Schneider, 2023) 实现,以 15Hz 执行预测出的 32 个动作。

总体上,`Cosmos3-Nano-Policy-DROID` 在机器人策略任务上取得了强劲结果。如 §6.2.5 详述,在我们提交榜单时,它在 RoboLab (Yang et al., 2026a)、RoboArena (Atreya et al., 2025) 与 MolmoSpaces (Kim et al., 2026) 上均排名第一,表明 Cosmos3 作为机器人策略学习的基础模型骨干可能是有效的。

> **译注** "跳过视频潜变量解码"意味着部署时策略只取动作输出、不真正渲染辅助视频帧——联合生成视频在这里是训练期的辅助监督,推理期可以整段裁掉,这是全模态模型做实时控制(15Hz)的关键省算力手段。榜单第一同样以"提交当时"为限定。

5. 基础设施 Infrastructure

本节介绍为支撑 Cosmos 3 端到端生命周期而设计的一体化基础设施栈。如图 11 所示,该平台统一了 4 大核心支柱:

- **数据工程**。摄取原始多模态数据,并将其转换为 WebDataset 格式的整备数据集(curated datasets),针对可扩展的分布式训练进行了优化。
- **大规模训练**。通过高效的并行化策略、优化的数据加载(data loader)、快速 checkpoint 保存以及集合通信原语,最大化 NVIDIA GPU 集群的利用率。
- **模型推理服务(serving)**。为生成类与推理类工作负载提供高效、低延迟的部署与推理(inference)执行。
- **基准评测(benchmark)与验证**。提供统一的评测框架,评估模型在多样任务上的能力,支持自动化回归追踪与系统化的模型对比。

5.1 数据基础设施

Cosmos 3 的训练语料取自数百亿量级的图像与视频候选样本,横跨多种模态、领域与任务。在这一规模上运作,要求数据基础设施能够同时做到:(1) 通过大规模分布式处理,把原始多模态数据转换为可直接用于训练的样本;(2) 支持基于嵌入(embedding)的检索、聚类与去重(deduplication);(3) 支持交互式的数据集可视化、检查与调试。为满足这些需求,我们开发了 SILA(Scalable Infrastructure for Large-scale data processing and Annotation,大规模数据处理与标注的可扩展基础设施)——一个可扩展的多模态数据基础设施平台,将存储、元数据管理、分布式处理、语义检索(semantic retrieval)与数据集可视化整合进同一个可扩展框架,用于大规模数据整备(curation)与管理。

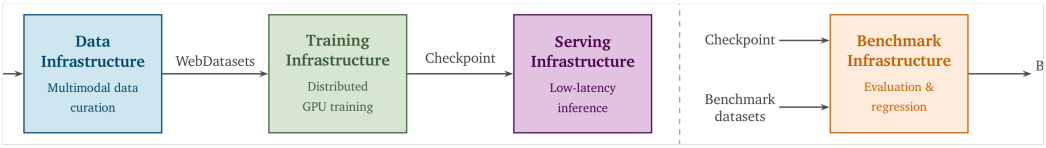


图 11:Cosmos 3 基础设施栈概览。平台横跨四大支柱。*数据基础设施*摄取原始多模态数据流,并将其整备为 WebDataset 格式的训练分片(shard)。*训练基础设施*在 NVIDIA GPU 集群上消费这些分片,配以高效的并行化、数据加载与 checkpoint 机制。产出的 checkpoint 流向两条并行路径(以虚线分隔):*推理服务基础设施*将其部署用于低延迟的生成与推理类推断;*基准评测基础设施*则用标准化的基准数据集评估同一批 checkpoint,以追踪回归并支持系统化验证。

SILA 的核心设计,是把流水线逻辑与基础设施机制干净地分离开来。研究人员只需声明带类型的处理阶段——指明每个阶段消费哪些列、产出哪些输出——平台则透明地处理数据集分片(sharding)、分布式执行、容错、checkpoint、元数据更新与资产注册。这一抽象使得引入新的数据源、基础模型和处理阶段(包括过滤、打标注(captioning)、嵌入生成、打分与打标签)变得非常直接,研究人员无需具备分布式系统方面的专门知识。其结果是一个能让数据整备跟上研究节奏演进的平台:随着模型、质量标准与训练配方的变化,新的信号可以被增量地添加、重算或替换。

5.1.1 大规模数据处理

多模态数据整备是一个迭代式的富化过程,而非一次性的离线预处理。随着模型、质量标准、标签与训练配方的演进,原始的文本、图像与视频样本会被反复地变换、过滤、标注与再处理。将这一工作流规模化颇具挑战,因为该流水线既是低产出率的、又是高度迭代的:原始候选样本中最终只有很小一部分会进入训练,这意味着低效的扫描、拷贝或模型推理,有相当大的比例被花在了后来被丢弃的样本上。与此同时,摄取、切分与转码、嵌入生成、去重、过滤、分类学打标、打标注以及分片等阶段,会反复作用于同一批样本。支撑这一工作流,因此需要在数十亿量级多模态样本上高效地进行重复变换与增量重算的机制。

在分布式执行下,这些挑战变得更为突出。整备负载必须在共享集群上持续运行,而集群的 GPU 可用性往往是碎片化且动态变化的,不能假设存在一次性的整块资源分配。基础设施必须协调众多分布式工作进程、避免重复计算、从故障中恢复、管理 CPU 密集与 GPU 密集的异构阶段,并在资源逐步可用时继续处理未完成的工作。为支持这一执行模型,SILA 组合了统一数据层、片段级(fragment)协调与故障恢复、分阶段分布式执行、节点本地模型服务、机会式集群利用,以及对 AI 智能体友好的运维接口。

统一数据层。SILA 将数据整备组织为一个统一的列式 Lance 数据集(Pace et al., 2025):每一行代表一个数据样本,每一个带类型的列代表一种整备信号,如标注文本、标签、质量分或注释。这取代了早期基础设施(Cosmos-Predict 1.0(NVIDIA, 2025d)与 2.5(NVIDIA, 2025b))采用的"每条流水线一张表"的旧架构——在旧架构中,每条流水线写入自己的 Postgres 表,输出随后经变更数据捕获(Change Data Capture, CDC)同步进 Databricks。随着流水线与元数据字段数量的增长,这种"表随流水线"的设计要求在大表之间做日益复杂的连接(join)才能重建单个样本的状态。这些连接在大规模下代价高昂,连简单的运维查询也变得难以表达,除非详细掌握连接键、表间关系与各流水线特有的模式(schema)。与之相反,SILA 通过向共享的 Lance 表追加新的带类型列,来增量富化同一个逻辑样本。这种统一表示天然契合多模态整备负载:大多数阶段都是在为既有样本补充额外元数据,而不是创建新实体。通过将数据集内容、元数据与处理状态放在同一处,系统能直接依据 Lance 片段元数据高效地支持大规模扫描、点查、增量重算与未完成工作的发现,而无需昂贵的启动期连接。

片段级协调与故障恢复。这一规模的数据整备以高并发运行:单个作业内有众多分布式工作进程,同时还有许多相互独立的作业并行地读写同一个持续演化的 Lance 数据集。若无显式协调,工作进程只能依靠昂贵的启动期查询、随机化采样或对已处理样本的事后过滤来避免重叠。这些做法拖慢作业启动、放任重复劳动,并使长时作业在被抢占或中断后的恢复变得复杂。SILA 转而直接在 Lance 片段(fragment)级别协调分布式整备:工作进程从 Lance 元数据中发现未完成的片段,在处理前先获取限时租约(lease),而 Lance 数据集本身始终是完成状态的权威来源(source of truth)。租约的持有通过周期性心跳维持;心跳一旦停止,租约即过期,其他工作进程即可接管该片段,从而无需人工清理即可从故障、抢占或端点崩溃中自动恢复。由于单个片段可能包含大量样本、需要数小时的模型推理,SILA 进一步把已认领的片段划分为更小的处理段(segment),把完成的段写成持久 checkpoint,并在全部段完成后通过一次元数据更新把整个片段原子地提交回 Lance。这将恢复单元与可见性单元解耦:被中断的作业从已完成的段续跑,而下游读者只会看到完整提交的片段输出。

分阶段 Ray 执行。整备流水线组合了资源画像迥异的异构操作,包括数据加载、解码、模型推理、后处理、写入与提交。若这些操作在单个不加区分的控制循环中执行,快速的上游阶段会不断积压中间输出,而较慢的推理或提交阶段则成为瓶颈。SILA 转而通过分阶段的 Ray 流水线引擎(Moritz et al., 2018)执行每个已认领的片段。框架托管的阶段负责加载、写入与提交,而用户自定义的预处理、计算与后处理阶段在各自独立的 Ray actor 池中执行,worker 数量与资源需求可独立配置。跨阶段边界的背压(backpressure)限制了在途工作量,防止快速的 I/O 密集阶段压垮较慢的下游阶段并迫使 Ray 对象存储溢写到磁盘。

节点本地模型端点。基础模型驱动的整备负载,常常需要在众多流水线 worker 并发处理数据的同时,提供大型标注、嵌入、打标或打分模型的服务。集中式推理服务在大规模下会成为瓶颈,而要求一整块连续的 GPU 分配又会削弱利用碎片化集群可用性的能力。SILA 转而借助 vLLM(Kwon et al., 2023)等系统在节点本地拉起模型服务端点,并把本地端点信息直接传给阶段 worker。worker 随后调用节点本地的服务进行推理,使模型密集型的整备阶段能够跨可用节点扩展,同时将数据并行的流水线执行与模型服务的放置解耦。

机会式集群利用。大规模整备必须与训练负载一起在共享集群上持续运行,而集群的 GPU 可用性往往碎片化且动态变化。SILA 不要求一次性的整块分配,而是把整备拆解为细粒度的分布式作业,可随资源可用而增量执行。系统支持 DGX Cloud Lepton(NVIDIA, 2026a)与 Slurm(Jette and Wickberg, 2023)等执行后端,使流水线得以机会式地利用空闲或部分可用的 GPU 容量。这提升了集群利用率,并使数据处理无需大块连续 GPU 预留即可持续进行。

智能体化作业编排。随着 AI 智能体在工具使用与长时程执行上日益强大,SILA 通过对智能体友好的运维接口——包括可复用技能、命令行接口(CLI)与结构化作业元数据——暴露大规模整备工作流。长期运行的编排智能体周期性地监控整备作业、检查日志与执行元数据、重启失败的阶段,并自动协调运维层面的恢复。通过这一持续监控循环,智能体能够追踪流水线进度、识别停滞或不健康的 worker、核验数据集覆盖率、在引入新模型或新过滤标准时触发增量重算,并在分布式作业随时间演化的过程中,将故障、恢复与执行状态通知工程师。

这些设计选择合在一起,大幅提升了大规模整备的效率。通过消除昂贵的启动期表连接、用片段级发现与协调取代随机化的任务选择,SILA 将作业启动延迟从 30–60 分钟降至约 5 分钟(视阶段与模型配置而定)。再结合分阶段执行、checkpoint 与更高的集群利用率,新基础设施相对旧架构实现了 10 倍的吞吐(throughput)提升。在生产高峰时段,单个 SILA 阶段每天可处理数十亿条行级标注,把大型标注与整备战役从月级作业压缩到周级迭代周期。

在可扩展性与吞吐之外,SILA 还通过以数据集为中心的接口隐藏大部分运维复杂性,简化了流水线开发。流水线作者只需指定所消费的输入列与所产出的输出字段,框架负责模式注册、列创建、片段发现、任务协调、checkpoint 以及把结果提交回共享数据集。因此,新增标注、打分、打标或过滤阶段,不再需要创建新的存储表、编写同步逻辑或手工协调分布式 worker。研究人员可以转而以增量方式迭代:添加新的带类型列、复用先前算好的输出,并只对缺失必需字段的样本进行重算。

5.1.2 嵌入存储与语义检索

语义检索负载既需要向量相似度搜索,也需要在数十亿多模态数据上进行感知元数据的过滤。在旧架构中,把高维嵌入直接存进 SQL 表会显著膨胀表体积,加重连接、扫描与运维查询的 I/O 开销。因此,嵌入被导出到独立的向量数据库,而用于前置过滤、后置过滤与结果解读的元数据仍留在关系型存储中。然而在整备过程中,嵌入、元数据与过滤标准持续演化,每当引入新的嵌入模型、元数据字段或搜索过滤条件时,都需要在两套系统之间频繁地同步与迁移。

SILA 转而把嵌入与样本元数据一起直接存进 Lance,让 LanceDB 直接在主数据集上构建向量索引,而无需独立的向量数据库(LanceDB, 2026)。由于嵌入存放在 Lance 数据文件中而非内联的关系行里,大体积的嵌入负载不会膨胀元数据表,也不会拖慢运维查询。通过把嵌入、元数据与向量索引放在同一个存储层,SILA 直接在整备数据集上支持语义检索、聚类与去重,同时保证搜索结果与最新的整备状态一致。

在生产环境中,SILA 使用 LanceDB 的 IVF_PQ 索引与余弦相似度,在覆盖数百亿行的 4096 维嵌入列上执行语义检索、聚类与去重。部署的近似最近邻(Approximate Nearest Neighbor, ANN)配置使用 64K 个 IVF 分区,配合 PQ 压缩的嵌入,以高效支撑十亿级检索负载。由于向量索引直接构建在主 Lance 数据集之上,语义检索运行于维护着最新整备元数据与过滤状态的同一存储层。这使得感知元数据的过滤、最近邻检索与下游数据集分析,能够与持续演化的嵌入、标注与整备产出保持同步,而无需在独立的向量系统与元数据系统之间做同步。

5.1.3 数据集可视化、检查与调试

在生产规模上,数据整备不仅要可扩展,还必须可观测:研究人员需要了解流水线覆盖情况、追踪整备产出随时间的演化,并诊断单个样本为何通过或未通过质量标准。SILA 因此将可视化、检查与调试视为数据基础设施的一等公民。其工具直接运行于 Lance 表之上,把总体流水线进度、有代表性的开发子集、样本级检查与下游分析视图,连接在同一个共享的整备底座之内。

- **开发与流水线验证**。SILA 提供从生产数据集构建小型开发 Lance 表的工具,同时保留完整语料的模式与有代表性的运行特征。由于这些开发表与生产数据高度相似,同一套流水线可以不加修改地在两种环境中运行,使研究人员能够在启动大规模生产作业之前验证正确性与执行行为。
- **数据集检查与进度分析**。为支持大规模整备监控,SILA 直接在 Lance 表上同时暴露聚合式进度分析器与交互式检查工具。片段级元数据被用来估算逐列覆盖率与流水线完成度而无需全表扫描;交互式查看器则允许研究人员抽样查看行、渲染媒体内容,并检查关联的标注、评分、注释与模式元数据。由于 Lance 支持高效的行级随机访问,这些检查工作流可直接运行在主整备表上,避免了传统基于 Parquet 的数据湖中常见的昂贵扫描与受限的行级检索模式。这让研究人员能在同一数据集内,快速地从流水线级进度监控切换到样本级调试。
- **分析查询集成**。面向大型分析负载,SILA 支持用于大规模聚合、报表与看板的联机分析处理(Online Analytical Processing, OLAP)查询。这些查询之所以更易表达,是因为 SILA 把每个样本及其整备信号存放在一张宽 Lance 表中:标注、标签、评分、嵌入、过滤决策与处理状态,都可以从一个逻辑数据集中直接选取与过滤,而无需通过各流水线专属表的连接来重建。选定相关列与样本群组后,分析后端即可计算覆盖率报告、质量看板、数据集审计与训练集分析所需的聚合。SILA 让 Lance 表始终作为整备状态、媒体、元数据、嵌入、向量索引与行级检查的权威来源,而把下游的分析执行视为一种部署选择。实践中,SILA 可以扫描权威 Lance 表,物化出可直接查询的快照或投影(包括 Parquet 文件(Le Dem, 2013)),并用可用的计算后端——如 Databricks、Spark 集群(Zaharia et al., 2016)或基于 Slurm 的批处理作业——执行 OLAP 负载。

贯穿处理、检索与检查三个方面,SILA 闭环达成了定义 Cosmos 3 数据基础设施的三个目标:把原始多模态语料转换为可训练样本;将其组织起来以支持语义检索与去重;并在整个整备生命周期中保持其可检查性。通过把资产、整备信号、嵌入、向量索引与执行状态统一在同一个 Lance 底座之内,SILA 把数据整备从一连串一次性的预处理作业,变成一个持续演化的生产工作流。新模型与新质量标准可以增量应用,分布式富化阶段可以在共享的异构集群上恢复与扩展,研究人员可以在语料级进度监控与样本级调试之间无缝切换。这一体化工作流使 Cosmos 3 训练语料得以扩展到数百亿多模态候选样本,同时保持可检索、可审计、可持续改进。

5.2 训练基础设施

Cosmos 3 依托一套为多模态基础模型训练的规模化而定制的基础设施平台。这一统一技术栈协调 Reasoner(推理器)与 Generator(生成器)训练的端到端生命周期。该生命周期涵盖原始多模态样本摄取、训练计算与持久化 checkpoint 保存,由下述几个环节构成。

- **数据加载器**。数据加载器以任意原生分辨率与宽高比摄取多模态样本——图像、视频、动作、音频、文本。它执行即时(on-the-fly)增广(如缩放、空间裁剪、颜色抖动与视频时间下采样),对文本条件做 tokenize,并把变长样本打包成批。为隐藏 I/O 与预处理延迟,加载器在并行 worker 进程中异步运行,并通过锁页内存(pinned memory)暂存缓冲区把批数据预取到设备上。
- **分布式训练**。训练通过混合分片数据并行(Hybrid Sharded Data Parallelism, HSDP)与上下文并行(Context Parallelism, CP)的组合来并行化。这一方案支撑了向大模型规模与超长输入序列的扩展。HSDP 在每个副本组内对优化器状态、梯度与模型参数做分片,而在组间做复制。CP 沿序列维度把数据切分到多个设备,以应对单块 GPU 显存容纳不下的巨型上下文窗口。两种策略正交组合,并按实验动态配置,以适配目标模型规模、序列长度与集群拓扑。
- **训练循环**。训练循环以 TorchTitan(Liang et al., 2024b)的风格编排,执行标准的前向、反向、优化与学习率调度周期。它原生支持多种优化器(如 AdamW 及其融合变体)、调度器(如带预热的余弦调度、带预热的常数调度)与损失函数(Reasoner 文本用交叉熵损失,Generator 用 EDM 损失)。同时引入即时运行的变分编码器(如 Wan2.2 VAE),流水线可端到端地直接作用于原始多模态输入。这一设计免去了离线的潜变量抽取阶段,并确保增广、编码与训练在各次运行之间保持步调一致。
- **Checkpoint 保存**。Checkpoint 按可配置节奏记录,并采用异步、脱离关键路径的持久化机制,防止磁盘与网络 I/O 阻塞训练循环。模型参数、优化器状态与随机数/数据加载器状态先在设备上做快照,交给后台写入进程,再序列化到远端存储,训练全程不中断。

Reasoner 与 Generator 都用这一统一框架训练,共享同一套 trainer、并行化架构、优化器、学习率调度器、tokenizer、数据加载器与监控工具。

5.2.1 数据加载器

数据加载器是持久化存储与训练循环之间的桥梁:它摄取原始多模态训练数据、执行即时增广,并组装出每个训练步所消费的批。在 Cosmos 3 中,数据加载器需要同时满足三项要求:

- **流水线饱和**。必须异步地流式产出批数据,防止训练循环因数据 I/O 而停顿或阻塞。
- **分布式负载均衡**。必须在分布式各 rank 间产出均衡的批,最小化跨 rank 同步停顿,最大化总体 GPU 利用率。
- **分布保真**。必须保证长期的模态与分辨率混合比例严格遵循配置的目标分布。

在常规 LLM 训练中,这三项要求都有成熟的解法。

- **流水线饱和**。通过调节 worker 数量、预取深度,并使用锁页内存暂存缓冲区把主机到设备的传输与计算重叠来达成。
- **负载均衡**。由于每个样本贡献固定数量的词元(token),这一问题变得平凡:维持各 rank 相同的样本数,即可保证各 rank 的计算量与激活显存画像一致。

然而,Cosmos 3 在全部三个维度上都令这套配方失效。由于其联合训练语料横跨高度异构的模态,单样本词元数的差异超过两个数量级,词元量成为计算与显存开销的首要驱动因素。例如,一段 720p 的两秒视频片段产生的词元,比几十条短文本标注加起来还多。在这种不对称负载下,给各 rank 分配相同样本数会带来严重的系统性低效:(a) 固定批形状导致大量填充(padding)浪费;(b) 各 rank 模态分配不同时造成严重的负载不均;(c) 在大规模下,极端的步时方差引发 NCCL 集合通信超时。

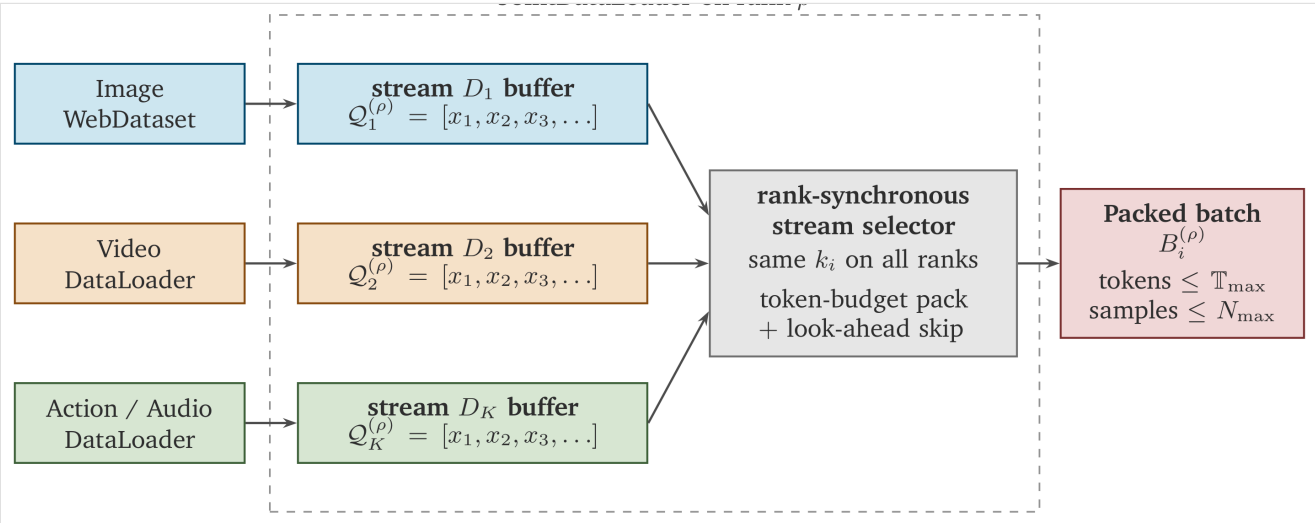


图 12:联合数据加载器(Joint Data-Loader)概览。各数据流专属的加载器在每个 rank 上填充本地的按流缓冲区。在每次全局迭代中,rank 同步的选择器在所有分布式 rank 上选出同一条流 k_i。随后每个 rank 在词元与样本数预算约束下,从其选中的本地缓冲区贪心地把样本打包进 $B_i^{(\rho)}$,并借助有界前瞻降低未用满的词元容量。

为应对这些挑战,Cosmos 3 的数据加载器围绕四个相互协同的机制构建:(i) *词元预算打包序列*,用词元预算而非固定样本数来约束每个 rank 每步的工作量;(ii) *联合数据加载器*,把按流的加载器多路复用为单个统一的训练批;(iii) *rank 同步的流选择*,用全局播种的选择器让所有 rank 每一步都对齐到同一条数据流;(iv) *前瞻打包(look-ahead packing)*,提升词元预算 $T_{\max}$ 的平均利用率。

词元预算打包序列。每个 rank 的工作量不再固定每步样本数,而是受严格的词元预算 $T_{\max}$ 约束。加载器贪心地把样本拼接成单个打包序列——每个样本恰好贡献其序列化形式所需的词元数,样本之间不插入任何填充——直到再追加下一个候选样本就会超出 $T_{\max}$ 为止。通过消除跨样本填充,这一设计使每步计算开销成为 $T_{\max}$ 的直接且可预测的函数,把硬件与模态混合波动引发的执行方差隔离开来。作为防御性的次级约束,每步样本数也被封顶在 $N_{\max}$。

联合数据加载器。每种模态、数据集或更细粒度的数据流,都封装在各自的加载器中,并配有私有的预取缓冲区以隐藏存储延迟。联合数据加载器(如图 12 所示)在这些按流加载器之间做多路复用,每步组装出单个统一的训练批,同时完整保留各流自身的缓冲、预取与可观测性。

rank 同步的流选择。基础模型训练通常取样于同时包含图像与多种空间分辨率视频的数据集。这些流之间的单样本词元数相差数个数量级。例如,视频样本的词元常常是图像的 100 倍以上,720p 视频又是 256p 视频的 10 倍以上。若允许各 rank 独立选择数据流,会在 FSDP 下引发严重的负载不均:单步之内,各 rank 的注意力 FLOPs(因其平方复杂度)大幅发散。我们的缓解办法是,用一个以迭代序号为键、全局播种的选择器来选定当前活跃的流,确保所有 rank 在每一步都处理来自同一模态与分辨率桶的样本。这消除了跨 rank 的计算时间与激活显存方差,而这种确定性的、由种子派生的选择序列在 checkpoint 与重启之间是比特级精确可复现的。rank 同步的流选择相比不同步的基线,把端到端训练吞吐提升了 54%。

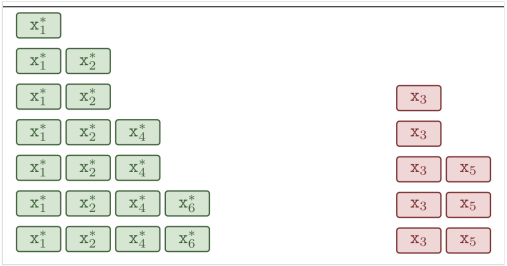


图 13:JointDataLoader 中的前瞻打包。加载器贪心地扫描选中流的样本,把能装入剩余词元预算的样本打包进当前小批(薄荷绿)。超出预算的样本被暂时移入旁置缓冲区(玫红),让后续更小的样本得以填满剩余容量。迭代结束时,被跳过的样本按其原始到达顺序放回流缓冲区头部,从而在减少填充的同时,跨迭代保持数据流的顺序。

前瞻打包。给定第 i 次迭代选定的流 k_i,联合数据加载器贪心地构建本地批:从该流缓冲区头部依次追加样本,直到下一个候选会超出词元预算 $T_{\max}$。然而,纯贪心打包可能留下不小比例的预算未被使用——每当下一个候选大到会溢出预算、而缓冲区更深处还有更小的候选可用时,就会出现这种残余容量;它在功能上等价于填充,并与吞吐损失成正比。我们用有界前瞻策略解决这一问题。当某个候选样本超出*总*词元预算 $T_{\max}$ 时,该样本在当前配置下无法打包,直接丢弃(并记录警告日志)。当候选只是超出*剩余*预算、且当前批已非空时,该候选被暂时移入*旁置缓冲区(look-aside buffer)*,加载器继续向流缓冲区更深处扫描,寻找能装进残余容量的更小样本。

图 13 展示了这一机制。在示例中,样本 a_1 与 a_2 装入当前批;a_3 超出剩余预算,被移入旁置缓冲区;加载器随后继续扫描并打包后续更小的样本(如 a_4 与 a_6)。迭代结束时,旁置缓冲区中剩余的所有样本按原始到达顺序放回流缓冲区头部,因此前瞻在减少填充的同时不会永久性地打乱流的顺序。为限制病态情形的开销——例如某条流一时被超大样本占据——每次迭代的连续前瞻尝试次数由一个按流可配置的上限封顶。在生产中我们把上限设为 10;超过该值后,我们观察到未用容量的进一步下降可忽略不计,而旁置缓冲区的规模(及其内存占用)却会持续增长。总体上,前瞻打包把有效序列长度较基线提升 8%,并带来相应幅度的训练吞吐提升。

> **译注** 原文正文用 $a_1 \cdots a_6$ 描述示例,而图 13 图中标记为 x1⋯x6,二者指同一组样本,系原文符号不一致。

冷启动处理。我们的若干数据流存在可观的首批延迟,主要来自 worker 进程创建、新打开分片的文件系统元数据缓存,以及各流特有的反序列化预热。若这一延迟在第一个训练步才被支付,它将与该步的 NCCL 集合通信竞速,极有可能在最慢的 rank 上触发看门狗超时——在大规模下尤甚,因为此时起作用的统计量是所有 rank 上的最大值。为消除这一故障模式,联合数据加载器在构造期间执行显式的预热阶段:从每条流预取一个批,使 worker 池、文件句柄与反序列化缓存全部就绪,然后在把控制权交还训练循环之前执行一次分布式屏障(barrier)。这保证每个 rank 都在首次前向之前付清了冷启动成本,且第一次迭代面对的是完全预热的数据流。

可观测性与集成。每次迭代,联合数据加载器都会向训练侧的监控回调发出一条结构化的打包统计记录;回调在分布式组内聚合各 rank 的记录,并把结果记录到 Weights & Biases。上报指标包括各流的经验采样比例(与配置的目标混合比对照)、每次迭代打包的样本数、各流的缓冲区占用与等待时间、前瞻饱和率,以及每次迭代的词元预算利用率。这些指标暴露了最可能"悄悄劣化"大规模训练而不使其崩溃的故障模式,使此类问题在出现后的几分钟内就浮现在训练看板上,而不是事后才从模型行为中被发现。

5.2.2 注意力实现

如 §2.3.1 所述,Cosmos 3 的 Mixture-of-Transformers 架构在单次前向中要求两种注意力形态并存:Reasoner 通路仅在 Reasoner 词元上使用因果注意力,而 Generator 通路在 Reasoner 与 Generator 词元的拼接序列上使用双向注意力,使每个 Generator 词元都能以完整上下文为条件。用 FlexAttention 这类通用算子朴素地表达这些异构掩码模式,结果虽然正确,却无法充分利用硬件:掩码结构对内核(kernel)不可见,本应跳过的注意力块内部执行了等价于填充的无效计算。这会降低张量核心(tensor core)利用率并加重显存带宽压力。

为此,我们协同设计了一种定制的双通路扁平注意力(two-way flat attention)机制,把跨通路的掩码结构直接暴露给高性能的变长(variable-length)注意力内核。图 14 展示了该设计。计算被分解为两次独立的内核调用。第一次处理 Reasoner 通路,是一次标准的变长缩放点积注意力(scaled dot-product attention, SDPA)(PyTorch Contributors, 2026b)调用,带因果掩码,只作用于 Reasoner 的查询、键与值。第二次处理 Generator 通路,要求同一样本内的 Reasoner 与 Generator 键/值流对每个 Generator 查询可见,但在打包批中的不同样本之间严格隔离。我们通过在样本粒度上把两条词元流扁平化并交错排布来实现,顺序为

$$[R_0,\, G_0,\, R_1,\, G_1,\, \cdots,\, R_n,\, G_n]$$

其中 R_i 与 G_i 分别表示样本 i 的 Reasoner 与 Generator 键/值词元。每个 Generator 查询在其所属样本的 $[R_i,\, G_i]$ 块上做双向注意力。这一表述以每层两次变长内核调用表达了完整的跨通路注意力,在单个打包表示中同时支持因果与双向掩码,消除了定长实现固有的填充开销,并使 Cosmos3-Nano 模型的端到端训练吞吐较基于 FlexAttention 的基线提升 22%。

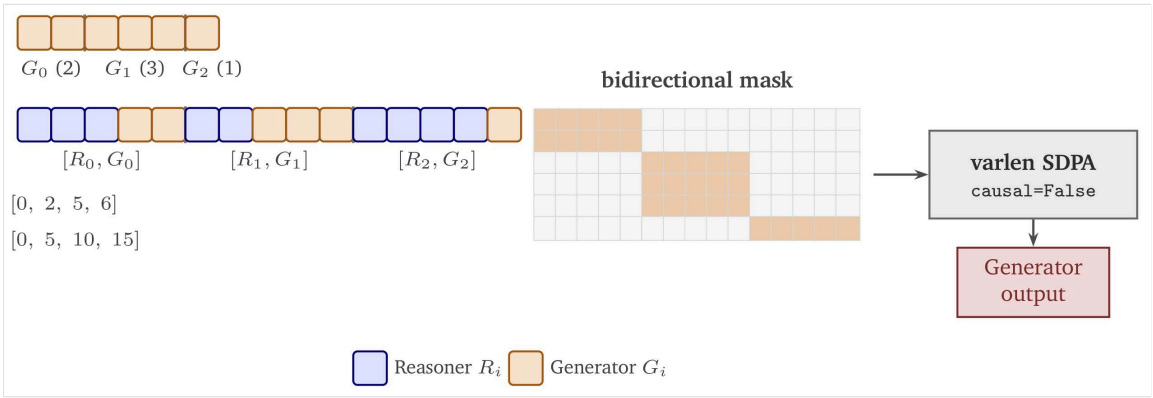


图 14:双通路扁平注意力。每条通路都实现为单次变长 SDPA 调用。(a) Reasoner 通路在打包的 Reasoner 词元上使用标准的因果 `varlen` 调用,产生块对角的因果掩码。(b) Generator 通路把 Generator 查询与交错排布的键/值流 $[R_0, G_0, R_1, G_1, \cdots, R_n, G_n]$ 分开打包。所得掩码呈块对角、且每块内部为矩形,使每个 Generator 查询在其所属样本的 $[R_i, G_i]$ 上下文上做双向注意力而不跨越样本边界。示例使用三个打包样本,$(|R_i|, |G_i|) = (3, 2), (2, 3), (4, 1)$。

变长注意力后端按平台选取,以匹配目标硬件上性能最优且数值上经过验证的实现。在 Hopper 级 GPU(H100、H200)上,我们使用 FlashAttention-3(Shah et al., 2024),它利用 Hopper 架构的 WGMMA 指令与基于 TMA 的异步数据搬运,提供接近峰值的注意力吞吐。在 Blackwell 级 GPU(GB200)上,我们使用 NATTEN(Hassani et al., 2023),其变长内核基于 CUTLASS 模板库构建,专为 Blackwell(SM100/SM103)的第五代张量核心与更新后的存储层级调优。两个后端都通过统一的调度接口接入,因此内核的选择对训练栈的其余部分完全透明,并可随新后端的成熟而重新评估。

5.2.3 分布式训练

Cosmos 3 的 Reasoner 与 Generator 分开训练,共用一套结合了混合分片数据并行(HSDP)与上下文并行(CP)的分布式训练栈。HSDP 在每个副本组内对模型参数、梯度与优化器状态分片、在组间复制,以适度的组内通信换取在常规单卡显存预算上训练数十亿参数模型所需的显存余量。CP 则针对另一类瓶颈:随上下文长度线性增长的单序列激活显存——若不处理,它会对可训练的序列规模强加一个硬上限。两种策略正交组合,(HSDP 度, CP 度) 配置按实验选定,以适配目标模型规模、序列长度与集群拓扑。

基于 Ulysses 方案的上下文并行。CP 采用 Ulysses 方案(Jacobs et al., 2023):在注意力之外,输入序列沿词元维度切分到各 CP rank 设备上;每个注意力层内使用两次 all-to-all 集合通信在两个分片轴之间转换。第一次集合通信把 Q/K/V 激活从序列维度重分布到注意力头维度,使每个 rank 持有一部分不相交注意力头的完整序列,从而无需进一步跨 rank 通信即可本地执行注意力;第二次集合通信把注意力输出恢复为原先按序列分片的布局。

该方案与序列打包(见 §5.2.1)及双通路注意力机制的集成十分干净。用于双向注意力的 Reasoner 与 Generator 键/值流的扁平化与交错操作,被推迟到头维度重分布之后进行——此时每个 rank 已持有其负责的注意力头的完整序列,可在本地完成拼接。因此同一套变长注意力内核在 CP 内可原样复用,无需 CP 专用的内核

变体。此实现所支持的最大 CP 度受模型查询头数的限制——Cosmos3-Nano 为 32,Cosmos3-Super 为 64——就 Cosmos 3 所面向的上下文长度与单卡显存预算而言,我们发现这一上限在实践中并不构成约束。

为何不用环形注意力(ring attention)?我们曾考虑用环形注意力实现 CP 作为替代方案,但发现它在我们的场景中明显缺乏吸引力。环形注意力要求先物化一个包含交错 Reasoner 与 Generator 词元的单一打包序列,再切分到各 CP rank,以便向环形调度暴露一条连续的 K/V 流。这排除了 Ulysses 天然允许的两条通路各自独立分片的可能,并使得在轮转的环形调度下构造每样本的双向/因果掩码变得复杂。再加上 all-to-all 在 NVLink 互连节点上有利的带宽特征,这些因素促使我们选择 Ulysses 作为 Cosmos 3 的 CP 策略。

5.2.4 选择性激活重计算

计算 transformer 的反向传播需要用到前向传播产生的中间激活,但在 Cosmos 3 所面向的模型规模与上下文长度下,把所有激活同时驻留在 GPU 显存中是不可承受的。标准的缓解手段是激活重计算(activation checkpointing)(Chen et al., 2016):前向只保存一组稀疏的"锚点"激活并丢弃其余;反向时从最近的锚点重新运行对应的前向子图,把丢弃的张量重算出来。默认策略只保存每个 transformer 块的输入,块内的一切按需重算;这使激活显存最小化,但在反向阶段引入了一次额外前向,使每步 FLOPs 膨胀约 33%,端到端训练吞吐相应下降。

为了在不超出激活显存预算的前提下降低这一重算开销,我们采用选择性激活重计算(Selective Activation Checkpointing, SAC):额外挑选一部分中间张量常驻显存而不再重算。选择由一个简单的成本收益启发式指导:按 FLOPs/显存比——即每提交一字节激活显存所省下的重算开销——对候选操作排序,优先物化比值最高者,直至耗尽激活显存预算。对 Cosmos 3 而言,注意力输出无疑是这一策略的最大受益者:注意力的重算开销随序列长度平方增长,而注意力输出张量本身相对较小(随序列长度与隐藏维度线性增长),使其成为 FLOPs/显存比最高的操作。用户还可以通过对操作名的正则表达式模式来配置自定义保存集;当剩余激活显存余量允许时,我们用它额外保留少量次要张量。

在我们的实测中,在每批 74,000 词元预算下,对 Cosmos3-Nano 应用"物化注意力输出"的 SAC,使端到端训练吞吐提升 13%,数值结果无任何变化。

5.2.5 面向 Transformer 块的 Torch Compile

我们在训练计算图上使用 `torch.compile`,并开启 `fullgraph=True` 与 `dynamic=True`。`fullgraph` 模式消除 CPU 侧开销并启用算子融合;`dynamic=True` 则处理混合模态批带来的变长序列——例如,不同迭代之间,Generator 通路的词元序列会显著长于图像批。Torch compile 使 Cosmos3-Nano 的 Generator 训练吞吐提升 41%。

5.2.6 视频 Tokenizer

Cosmos 3 训练要求把解码后的视频帧由视频 VAE 即时 tokenize 为潜表示:我们的配置采用 Wan2.2(Wan et al., 2025b)。在初始实现中,我们观察到 tokenizer 在每个训练步中占据了不成比例的份额——对较小的 Cosmos3-Edge 与 Cosmos3-Nano 模型,它甚至主导了前向阶段。在这类模型中,transformer 的计算量小到不足以摊薄 VAE 的开销。由于 tokenizer 位于数据加载器与训练循环之间的关键路径上,它引入的任何延迟都会直接拉低训练吞吐。为此我们实施了一系列针对性优化,合计将 tokenizer 的实际耗时占比显著降低。

分块编码。Wan2.2 因果 tokenizer 先编码一个 1 帧的"prime"块,之后每个潜块对应 4 个像素帧;默认的单次调用粒度是一个潜块(即 prime 之后的 4 帧)。在训练所用的空间分辨率下,这一默认设置使 GPU 严重欠载:每次内核发射处理的工作量太小,无法喂饱张量核心。我们改为每次调用编码可配置数量的像素帧,以额外的激活显存换取每次发射更高的算术强度。最优块大小与分辨率相关:分辨率越高,单帧已占用可观显存,触发 OOM 之前可容纳的帧数就越少;分辨率较低时,则可行且有益于用大得多的块。我们在训练硬件上经验性地确定了如下工作点:256p 为 68 帧,480p 为 24 帧,720p 为 12 帧。这些配置把编码器推到 roofline 曲线的计算受限一侧,同时距单卡显存预算仍有充分余量,贡献了每步加速的主要部分。

提前编译(ahead-of-time compilation)。在分块编码之上,我们对 tokenizer 使用 `torch.compile`,通过融合逐点操作并为编码器的卷积与注意力块挑选优化的内核调度,再取得 52% 的编码延迟下降。要最大化吞吐,输入形状必须是静态的,因此编码器要针对每一种可能被调用的形状分别编译。Cosmos 3 训练涵盖三种空间分辨率(256p、480p、720p),每种分辨率下五种宽高比。在每个(分辨率, 宽高比)组合内,因果 tokenizer 以两类模式被调用:prime 块编码,以及缓存大小为 1 和 2 的分块编码。这产生 $3 \times 5 \times 3 = 45$ 个必须在训练开始前编译完的不同计算图。若在每个 rank 上串行编译全部 45 个图,trainer 启动会被拖慢约 15 分钟。为消除这一开销,我们借助 AOTInductor(PyTorch Contributors, 2026a)把编译工作切分到各数据并行 rank 上(如图 15 所示):AOTInductor 执行提前编译,并把生成的内核与宿主代码序列化到磁盘。只要 rank 数不少于 45(我们所有训练配置都满足),每个 rank 恰好编译一个图;编译产物写入共享文件系统,随后每个 rank 从磁盘加载全部 45 个图。这把预热开销压缩到 1 分钟以内。为在静态形状编译下容纳任意帧数的视频,每段输入片段在编码前先右侧填充到配置的编码块大小的下一个整数倍,所得潜张量再沿时间轴裁剪到模型所需的精确序列长度后才交给模型。

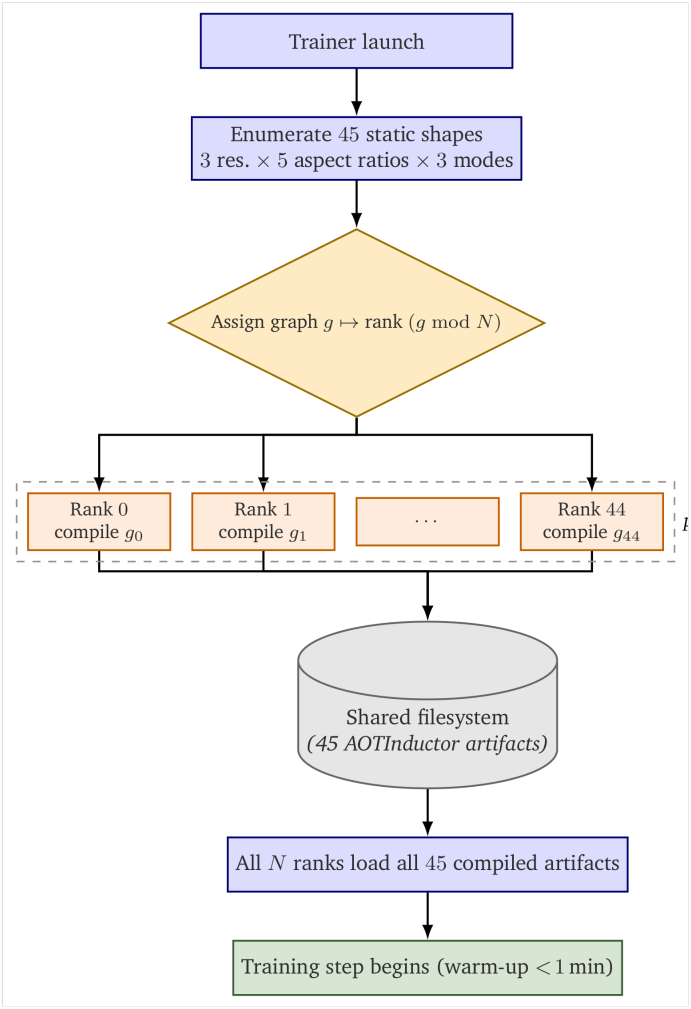


图 15:Wan2.2 tokenizer 的分片 AOT 编译。由 {3 种分辨率} × {5 种宽高比} × {3 种 tokenizer 调用模式} 产生的 45 个静态形状计算图被切分到各 rank;每个 rank 编译其分到的图,把编译产物写入共享文件系统,并在训练开始前加载全套产物。预热时间从约 15 分钟(串行)降到 1 分钟以内(分片)。

面向已知帧数的特化。对于每段视频帧数固定且事先已知的数据集——例如机器人动作数据集,每个 episode 贡献的片段长度完全相同——我们把编译特化到运行时实际出现的精确张量形状,绕过通用情形所用的"填充再裁剪"回退路径。这消除了填充尾部的计算、免去了相应的潜张量裁剪步骤,并在此类数据集上带来虽小但稳定的额外吞吐提升。

5.2.7 Checkpoint 机制

为消除保存引发的停顿,checkpoint 与训练完全重叠执行。沿用 Torchtitan(Liang et al., 2024b)的做法,checkpoint 写入经由专用的 Gloo 进程组,而非承载训练集合通信的 NCCL 通信器,把 I/O 流量与 GPU 侧通信隔离开。这一异步设计以适度增加的主机内存为代价,掩盖了对象存储高度不稳定的写延迟,并消除了几乎全部保存耗时。表 7 量化了收益:相对于 30 分钟间隔的同步 checkpoint,异步 checkpoint 使 Cosmos3-Nano 的端到端训练时间缩短 4%,Cosmos3-Super 缩短 9%。

表 7:异步 checkpoint 的收益。与 30 分钟间隔的同步 checkpoint 相比,异步 checkpoint 分别使 **Cosmos3-Nano** 与 **Cosmos3-Super** 的端到端训练时间缩短 4% 与 9%。**Cosmos3-Super** 上更大的收益反映了其更长的 checkpoint 保存时间。

模型	Checkpoint 保存时间 (s)			相对同步方式的加速
	均值	最小	最大	
Cosmos3-Nano	72	43	250	4%
Cosmos3-Super	167	40	736	9%

异步保存机制。构造时,checkpointer 用 `spawn` 启动方式拉起一个长驻子进程,并通过多进程队列与之通信。子进程加入一个 Gloo 进程组,在 CPU 侧完成构建保存计划所需的归约,让 GPU 始终留给训练。随后它阻塞在入站队列上,直至收到 checkpoint 保存请求或终止哨兵。

保存计划记忆化。为进一步降低开销,checkpoint 保存计划在首次保存操作时计算,后续保存直接复用。这之所以可行,是因为保存计划是状态字典(state dict)拓扑的确定性函数。复用计划避免了跨 rank 反复的元数据通信,使 checkpoint 开销降低约 60%,进一步降低了异步 checkpoint 成为训练瓶颈的可能性。

面向对象存储的优化。保存 checkpoint 时,每个 rank 把自己持有的那份状态字典分片写入对象存储。可能存在于多个 rank 上的复制张量,在写入前会先去重。在默认的轮询(round-robin)分配下,复制张量可能由不同的 rank 写出,导致加载时每个 rank 都要读取全部 checkpoint 文件才能恢复复制状态。这带来显著开销,因为即便只需要文件中的一部分内容,也必须加载整个文件。为降低这一开销,我们设置 `dedup_to_lowest_rank = True`,把复制张量只存放在相应子网格(submesh)中编号最小的 rank 上(通常是 rank 0)。这样加载时,每个 rank 只需读取自己的分片与 rank 0 的分片。这大幅缩短了 checkpoint 加载时间,对包含大量小型复制张量的优化器状态字典尤其明显。

随机状态恢复。trainer 以感知 rank 的方式恢复随机数生成器(RNG)状态。由于 RNG 状态按 rank 建键,恢复的作业会先在 checkpoint 元数据中查找与自身 rank 对应的键,仅当该键存在时才请求该状态。这保持了与早于"按 rank 建键的 RNG 格式"的旧 checkpoint 的兼容性。若 rank 专属的键不存在,该 rank 保留其当前的 RNG 状态。

5.2.8 吞吐总结

表 8 报告了在 NVIDIA GB200 系统上实测的 Cosmos 3 稠密配置的稳态单 GPU 训练吞吐。尽管 Cosmos 3 支持跨异构模态与任务的多种训练模式,这些测量采用文生图与文生视频联合训练配置,以便进行标准化的吞吐对比。图 10 展示了所用的预训练数据混合比。Cosmos3-Nano 使用 1024 块 NVIDIA GB200 GPU 进行基准测试,Cosmos3-Super 使用 2048 块 NVIDIA GB200 GPU。

表 8:Cosmos 3 稠密模型配置的稳态训练吞吐。TFLOPS 与 MFU 按单 GPU 报告。图像与视频词元吞吐分别以每 GPU 小时百万词元计。实验在 NVIDIA GB200 GPU 上进行,Nano 与 Super 的运行分别使用了 2048 与 4096 块 GPU。

模型	单步耗时 (s)	TFLOPS	MFU	迭代/小时	图像词元/hr/GPU (M)	视频词元/hr/GPU (M)
Cosmos3-Nano	7.1	520	0.23	507	4.56	16.23
Cosmos3-Super	19.5	673	0.30	185	1.66	5.91

> **译注** 原文此处正文与表注相互矛盾——正文称 Nano/Super 分别用 1024/2048 块 GB200 GPU,而表 8 表注称 2048/4096 块;论文未作解释。GB200 每个超级芯片(superchip)含 2 个 Blackwell GPU die,两组数字恰为 2 倍关系,或系"超级芯片数"与"GPU 数"两种口径混用,请以官方勘误为准。另按数值反推,MFU 约以 2.25 PFLOPS 的峰值算力为基准(520/0.23 ≈ 2260),与 GB200 单 GPU 的 BF16 稠密峰值一致,即 MFU 按 BF16 口径计算。

Nano 模型取得最高的原始词元吞吐:每小时处理 507 次迭代,达到每 GPU 小时 4.56M 图像词元与 16.23M 视频词元。相比之下,更大的 Super 模型每次迭代执行的计算量大得多,迭代速率降至每小时 185 次,吞吐降至每 GPU 小时 1.66M 图像词元与 5.91M 视频词元。

尽管词元吞吐更低,Cosmos3-Super 却取得了更高的算力利用率:单 GPU 吞吐从 520 提升到 673 TFLOPS,MFU 从 0.23 提升到 0.30。这反映了模型规模与训练吞吐之间意料之中的权衡:Cosmos3-Nano 为最大化词元处理吞吐而优化,而 Cosmos3-Super 凭借更大的模型容量与更高的单词元计算量,更充分地榨取了 GPU 的计算资源。

5.3 推理服务基础设施

Cosmos 3 与多套生产级推理服务框架集成,以支撑广泛的部署场景。Reasoner 由 TensorRT-LLM(NVIDIA Corporation, 2026)与 vLLM(Kwon et al., 2023)支持,两者都通过分页 KV cache 管理、连续批处理(continuous batching)与融合注意力内核提供高度优化的自回归解码。Generator 推理由 vLLM-Omni(vLLM 面向扩散式生成的多模态扩展)(Yin et al., 2026)支持,它在峰值吞吐与多租户调度效率之间提供了互补的权衡选项。在这些生产后端之外,我们还提供一个原生 PyTorch 参考实现,以可读性与可修改性为先,既是推理算法的忠实规范,也是下游适配、研究扩展与集成进自定义应用流水线的起点。

5.3.1 纯 PyTorch 路径

纯 PyTorch 服务路径直接在 eager 模式的 PyTorch 中执行模型及其外围推理流程——不依赖任何专用服务运行时——其设计目标是尽可能忠实地复现训练时的计算。该路径负责完整的端到端推理工作流,由下述阶段组成:

- **输入准备**。解析文本提示词,加载图像、视频或动作等条件输入,构造模型所需的按模态条件字典,并组装相应的输入张量与元数据。
- **自回归循环**。在 Cosmos 3 Reasoner 中,输出词元以自回归方式产生,每一步都以先前生成的词元以及条件上下文缓存的键/值状态为条件。
- **扩散循环**。在 Cosmos 3 Generator 中,PyTorch 路径构造时间步调度表,在每一步调用去噪器,应用无分类器引导(classifier-free guidance, CFG),按采样器更新潜状态,并管理去噪过程外围的请求级控制流。
- **解码与后处理**。去噪完成后,所得潜表示被解码为目标模态——文本、图像、视频、音频或动作——按需后处理,并通过服务接口返回给调用方。

由于原生 PyTorch 后端保留了模型原本的 PyTorch 结构,它是落地新模型特性、采样器修改、KV cache 与激活缓存策略以及调试插桩的首要目标。新能力先在此后端验证,之后才移植到生产运行时(TensorRT-LLM 与 vLLM),从而保证参考实现始终是推理算法的权威规范。PyTorch 后端提供以下特性与优化。

Torch compile 与 CUDA graph。由于 PyTorch 原生推理路径把请求编排与采样逻辑留在 Python 中,推理服务的一大瓶颈是重复去噪步骤中主机侧的内核发射开销。为此我们用 `torch.compile` 与 CUDA graph 重放来优化 PyTorch 路径。CUDA graph 优化在 transformer 层的粒度上实现,捕获重复执行的 transformer 块。每个块以 `reduce-overhead` 模式用 `torch.compile` 编译,当块以兼容的张量形状与内存布局被调用时,PyTorch Inductor 会对其做低层化并启用 CUDA graph 重放。外层推理循环保持为普通 PyTorch:提示词处理、时间步调度、采样器更新、CFG 编排与解码均不在捕获的图内。图中只包含数值密集且高频重复的逐层计算,而动态的服务逻辑留在图外。CUDA graph 的收益在 T2I 类生成中最为明显——生成任务更短、内核更小,CPU 发射开销在端到端延迟中占比更高。在不同硬件后端上,CUDA graph 为 T2I 生成带来 30% 到 60% 的加速。

分布式推理。我们在推理时使用上下文并行(CP),以支持超出单 GPU 显存容量约束的生成任务。我们沿用训练所采用的 Ulysses 方案(Jacobs et al., 2023),保证两种场景的一致性。除了支持长上下文推理,CP 还能通过把前向传播分摊到多块 GPU 来充当降延迟手段;即便单卡显存足以容纳完整上下文,这一收益依然成立,使 CP 成为加速推理的通用工具。

在上下文并行之外,我们还利用无分类器引导(CFG)并行来进一步降低端到端推理延迟。带 CFG 的扩散采样在每个去噪步都需要两次模型前向——一次以输入提示词为条件、一次无条件——两者的噪声预测经线性组合得到引导后的更新。由于条件与无条件两次前向作用于相互独立的输入,且每步只需同步一次以合成引导预测,它们是跨两块 GPU 并行化的理想对象。实践中,我们并发派发条件批与无条件批,并在采样器推进之前只做一次轻量的点对点交换来合并两份预测。这使每步延迟几乎减半,并可与上下文并行干净地组合,在多 GPU 节点上实现相乘的延迟削减。

Reasoner 塔缓存。对于文生图(T2I)、文生视频(T2V)、图生视频(I2V)与视频生视频(V2V)等任务,Reasoner 塔的条件输入——文本提示词,以及(如适用)条件图像或视频——在整条采样轨迹中保持不变。因此,Reasoner 的输出在各扩散步之间是不变的,只取决于条件输入,而与当前噪声水平或部分去噪的样本无关。我们

利用这一性质,在推理开始时只计算一次 Reasoner 前向,并缓存其输出供后续所有去噪步复用。由于缓存的激活与每步重算的结果在数学上完全相同,这一优化在不影响任何生成质量的前提下,大幅降低了每步延迟。

批处理。把多个样本合并进单次前向,可以把每步的开销(内核发射、权重读取与集合通信)摊到更大体量的有效工作上,从而进一步提升推理吞吐。我们的推理批处理器复用了为训练开发的变长序列打包机制:不是把较短序列填充到统一长度——那会在填充词元上同时浪费计算与显存——而是把形状各异的样本拼接为单个打包张量,并把相应的累计序列长度元数据提供给注意力等对形状敏感的算子。用户可指定两种预算模式之一:总词元预算(批内允许的最大词元数)或固定样本数。给定所选预算,批处理器在遵守单设备显存约束的前提下,贪心地打包到来的样本直至预算耗尽。这一设计在离线语料生成与大规模评测等以吞吐为先的部署中提升了 GPU 利用率;但在延迟受限的场景(如机器人负载)中没有收益——那里必须单独处理单个样本,此时批处理器被配置为样本数为 1,实际上等于停用。

表 9 报告了在 189 帧输出的文生视频(T2V)任务上,请求批处理带来的推理吞吐提升。在 256p 下,批处理带来 8% 到 55% 的吞吐增益。到 480p 收益递减:每个样本已提供足够工作量喂饱 GPU,留给额外并行的余地有限。在 720p 下,74,000 词元的上下文窗口只容得下 B=1,批处理加速无从谈起。

表 9:批处理带来的推理加速。在 189 帧输出的文生视频(T2V)任务上评测 **Cosmos3-Nano** 与 **Cosmos3-Super**。在 74k 词元上下文上限下,分别以最大可容纳批大小 $B = 6$ 与 $B = 3$ 报告 256p 与 480p 的结果。720p 因在同一预算下只容得下 $B = 1$ 而省略。

硬件后端	Cosmos3-Nano		Cosmos3-Super	
	T2V-256	T2V-480	T2V-256	T2V-480
H100 80GB	8%	2%	**55%**	5%
GB200	**40%**	2%	9%	1%

带提示词升采样的生成。推理还支持一种提示词升采样(prompt upsampling)模式(详见 §6.3.2):由 Reasoner 把简短、自由形式的用户提示词扩写为更丰富、结构化的 JSON 目标输出描述(涵盖诸如场景构图、主体属性、相机与运动设定、光照与风格等)。升采样后的描述由 Reasoner 自回归生成,再作为额外条件输入与用户提供的图像、视频或动作条件一道交给 Generator,合成最终输出。这条流水线有两重意义:一是把精细提示词工程的负担从终端用户转移给模型自身,并稳定地提升下游生成质量;二是在单次推理调用中同时贯通 Reasoner 与 Generator 两条通路的完整全模态端到端流程——证明两条通路可以无缝组合,从极简的用户输入产出高保真的多模态输出。

吞吐与延迟的权衡。推理负载通常在两个相互竞争的目标中择一优化:吞吐或延迟。批量推理——对固定输入语料生成一大批输出——通常面向吞吐优化,因为端到端墙钟时间与总成本才是关注指标。相反,机器人等延迟敏感应用——动作必须从特定起始上下文出发、在严格的每步截止时间内产出——优先考虑响应性,相关指标是首词元延迟(time-to-first-token,或"首动作延迟")与每步延迟。前文所述优化对这两种场景各有侧重:分布式推理(上下文并行与 CFG 并行)主要通过把单个请求并行到多块设备来降低延迟;批处理则主要通过把每步开销摊到多个并发请求上来提升吞吐。其余优化——torch compile 与 Reasoner 输出缓存——则同时惠及两种场景。

5.3.2 Reasoner 的推理框架:vLLM 与 TensorRT-LLM

由于 Nano 与 Super 的 Reasoner 构建于 Qwen3-VL(Bai et al., 2025b)骨干之上,它们与 vLLM 和 TensorRT-LLM 的集成直接复用了两个框架中已有的上游 Qwen3-VL 支持。这使我们开箱即用地继承了两个后端已为 Qwen3-VL 家族提供的优化注意力内核、分页 KV cache 管理、连续批处理与多模态输入处理,只需极少的配置改动即可把 Cosmos 3 的模型权重与 tokenizer 绑定到既有执行路径。

相比之下,Edge Reasoner 构建于定制的 Nemotron 骨干(NVIDIA, 2025)之上,vLLM 并无原生支持。为此我们按照 vLLM 的模型贡献者(model-contributor)规范实现了专门的集成。该集成在结构上可回馈上游(upstreamable),便于后续维护并支持社区贡献。

5.3.3 Generator 的推理框架:vLLM-Omni

Cosmos 3 Generator 集成进 vLLM-Omni,以借助其为扩散式多模态生成优化的服务栈。该集成把 Cosmos 3 Generator 实现为 vLLM-Omni 中的一等公民模型,支持 Generator 的全部模态,包括图像、视频、音频与动作条件生成。它遵循 vLLM-Omni 框架的模型贡献者规范与代码风格,使实现与该框架既有的调度、分布式执行、显存削减与量化(quantization)特性兼容。

这一集成对 Generator 服务尤其重要:扩散式生成需要在大规模图像、视频或多模态词元序列上进行多次重复的 transformer 求值。因此 vLLM-Omni 后端着力于在保持生成质量的同时,降低每步延迟、降低峰值显存占用并提升吞吐。Cosmos 3 Generator 的集成支持以下 vLLM-Omni 特性:

- **Cache-DiT**。一种免训练的加速方法,在相邻去噪步之间复用缓存的 transformer 块输出,以对生成质量影响可忽略的代价跳过冗余计算。
- **Ulysses 上下文并行**。一种上下文并行执行方案,把长图像与视频词元序列切分到多块 GPU,并用 all-to-all 注意力通信降低单设备显存占用、改善延迟。
- **CFG-Parallel**。一种面向无分类器引导的并行化策略,把条件与无条件前向派发到不同 GPU rank。两份预测每个去噪步同步一次以合成引导更新,降低基于 CFG 的采样延迟。
- **HSDP**。一种显存高效的分布式推理模式,用 FSDP2 把 transformer 权重分片到多块 GPU,并在前向传播中按需聚合参数,降低大型 Generator 模型的峰值显存需求。
- **CPU offload**。一种按层的卸载机制,推理期间在 CPU 与 GPU 内存之间搬运模型参数,以额外的数据传输开销换取大幅更低的峰值显存占用。
- **VAE-Patch-Parallel**。一种并行 VAE 执行模式,把潜张量或像素张量切分为空间瓦片(tile),在多个 rank 上分别编解码,同时降低单设备显存消耗与 VAE 延迟。
- **量化**。一条动态 FP8 量化路径,降低主要计算操作的精度,在保持可接受生成质量的同时降低推理延迟与峰值显存占用。

这些特性合在一起,使 Cosmos 3 Generator 能够从显存受限的单 GPU 部署扩展到高吞吐的多 GPU 服务。Cache-DiT 与量化降低重复去噪计算的成本;上下文并行与 CFG-Parallel 通过把单个请求分摊到多块 GPU 来改善延迟;HSDP、CPU offload 与 VAE-Patch-Parallel 则为高分辨率或长时长的生成任务缓解显存压力。图 16 总结了 Cosmos 3 的服务性能:既对比了不同硬件后端上的单 GPU 运行,也对比了 B200 上的多 GPU 运行。评测覆盖 PyTorch-OSS 与 vLLM-Omni 两套框架。

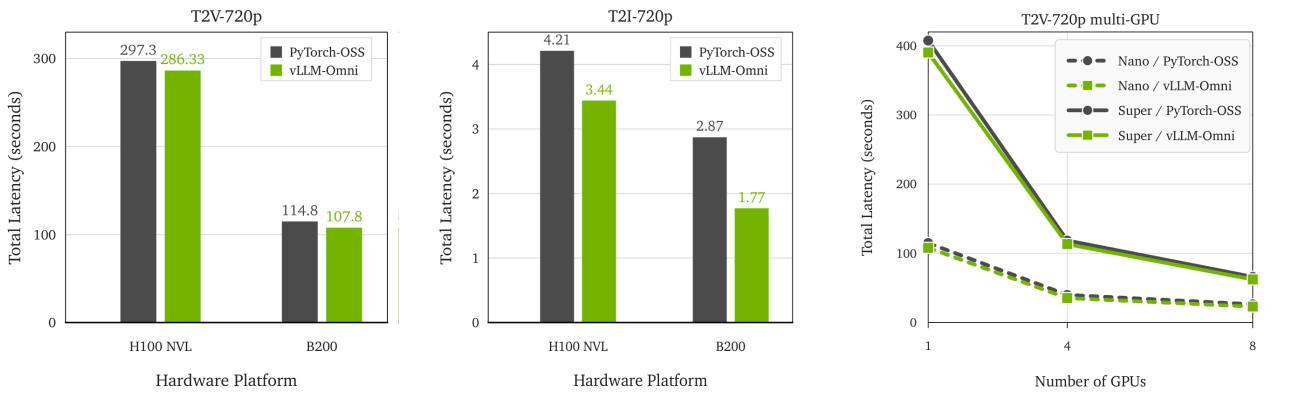


图 16:Cosmos 3 服务性能。(a) Cosmos3-Nano 720p T2V 在 H100 NVL 与 B200 上的单 GPU 延迟,用于观察不同硬件后端上的性能。(b) Cosmos3-Nano 720p T2I 在 H100 NVL 与 B200 上的单 GPU 延迟,用于观察不同硬件后端上的性能。(c) 在 B200 上,Cosmos3-Nano 与 Cosmos3-Super 的 720p T2V 延迟随 1 到 8 块 GPU 的扩展情况。全图数值越低越好。

> **译注** 图 16 在原文中排版于第 50 页 §5.4 正文之间,内容属 §5.3.3 的服务性能总结,此处按其被引用的位置移至 §5.3.3 末尾。

5.4 基准评测基础设施

Cosmos 基准评测系统管理 Cosmos 模型的评测作业,并同时存储生成产物与评测结果。编排层在 Lepton 或 Slurm 集群上调度生成、打分与端点评测作业,并追踪每个阶段的执行状态。对每一次运行,系统都记录元数据,包括模型 checkpoint、代码版本、所选基准、生成设置、基准特有参数与关联数据集。这些记录共同建立起完整的可追溯性:每一个上报的分数,都可以追溯到产生它的确切模型权重、输入、参数配置与评测代码。

该系统支持 §6 所述的异构基准套件,而不要求所有基准共享同一套实现。这些基准从多样的标准出发,评测生成的视频、音频、动作轨迹以及 Reasoner 模型的文本回答,涵盖视觉保真度、音频质量、音画同步、提示词与控制遵循度、动作或轨迹精度、任务完成度、物理合理性与推理正确性。评测器既包括与开源库和公开基准套件的集成,也包括专为 Cosmos 开发的定制评测器。打分方法横跨基于参考的误差指标、感知与时间一致性度量、音画对齐指标、基于视觉语言模型(VLM)的评判、人工标注,以及精确匹配或数值答案评测。

对 Generator 的评测,基准测试被拆分为生成与打分两个阶段。生成作业用 PyTorch 推理流水线或 §5.3 所述的某个服务框架执行模型,并把生成输出写入对象存储。打分作业随后消费这些已存储的产物,计算逐样本与聚合指标,并把结果与运行元数据一并记录。这一解耦设计使得可以用新指标或新评测器对既有输出重新打分,而无需重跑生成。

对 Reasoner 的评测,我们使用 VLMEvalKit(Duan et al., 2024)框架配合 vLLM。这些作业向已部署的模型端点发送提示词与多模态输入,处理模型响应,并记录基准分数与关联元数据。

评测分数与运行元数据存放在关系型数据库中,生成产物存放在对象存储中。人工评测标注与被评测的产物及题目集一起存储,人工评测的聚合结果与自动化指标并列追踪。一个基准评测门户通过看板、排行榜、样例级检查工具,以及模型间或 checkpoint 间的对比功能,提供对这些记录的访问。

6. 实验结果 Results

我们在一系列对物理 AI(Physical AI)至关重要的理解与生成任务上评测 Cosmos 3。与以往聚焦单一模态或单一能力的系统不同,Cosmos 3 被设计为一个统一的全模态世界模型,同时支持推理、感知、模拟与动作生成。因此,我们的评测同时覆盖 Reasoner(推理器)与 Generator(生成器)两个组件,涵盖多模态理解、空间推理(spatial reasoning)与时序推理(temporal reasoning)、图像与视频生成、音视频生成、迁移生成(transfer generation)、正向与逆向动力学,以及机器人策略学习。在这些多样化的基准评测(benchmark)中,Cosmos 3 相对于专用开源模型与领先的闭源系统均持续展现出强劲的结果,凸显了统一世界模型架构对物理 AI 的价值。以下各小节分别给出 Reasoner 与 Generator 的详细结果,并分析二者在机器人、自动驾驶、智慧基础设施与通用多模态领域的能力。

6.1 Reasoner 评测

Cosmos 3 Reasoner 共在 48 个基准评测上进行评测。结果被汇总为四大类:通用(general)、机器人(robotics)、智慧基础设施(smart infrastructure)与驾驶(driving)。表 10 报告了 Edge、Nano、Super 三个型号的结果,及其与现有开源和闭源模型的对比。所有数值均使用 VLMEvalKit(Duan et al., 2024)评测得到,该框架已集成到我们的基准评测基础设施中。

表 10: Cosmos 3 各型号与对比模型在通用多模态理解、机器人、智慧基础设施与自动驾驶基准评测上的 **Reasoner 基准评测结果**。各行报告单项基准评测或分组平均,各列报告模型得分。以**加粗**与下划线标出。† 表示闭源模型。

类别	Benchmark	大型模型块(Super 级)					中型模型块(Nano 级)						
		Cosmos 3 Super	Qwen3-VL 32B	Cosmos-Reason2 32B	Gemma-4 31B	Gemini 3.1 Pro †	Cosmos 3 Nano	Qwen3-VL 8B	Cosmos-Reason2 8B	Gemma-4 E4B	RynnBrain 8B	MiMo-Embodied 7B	Cos Ed
通用19项基准	MMBench-Dev	87.4	87.8	86.3	86.7	**93.2**	85.1	85.3	82.2	66.2	**85.5**	81.4	
	RealWorldQA	79.2	80.3	76.1	71.8	**81.8**	72.2	**73.2**	68.2	61.0	72.5	72.0	
	CVBench	88.0	86.8	88.1	84.1	**88.6**	86.5	85.2	85.6	68.1	87.5	**87.8**	
	VideoPhy2	**47.4**	36.8	43.3	33.1	28.7	**45.6**	28.2	37.1	13.8	10.5	20.8	
	CausalVQA	77.0	81.0	74.5	76.5	**92.0**	70.0	**72.0**	71.5	38.0	68.5	63.0	
	MVPBench	**70.3**	58.6	62.2	27.0	59.4	**66.9**	51.6	54.2	32.8	50.1	43.3	
	CountBenchQA	89.1	93.6	87.5	79.1	**95.3**	84.8	89.5	79.9	55.4	**90.5**	84.6	
	AI2D	87.8	88.2	87.5	88.9	**93.8**	85.0	84.8	83.6	78.2	**85.7**	83.2	
	DocVQA	90.4	**96.0**	95.1	89.6	95.8	94.2	**95.6**	94.3	78.1	95.4	93.9	
	InfoVQA	82.4	**87.6**	85.1	66.5	85.0	81.8	83.4	79.8	46.1	81.8	**84.2**	
	OCRBench-v2	**66.7**	65.9	57.4	61.8	64.5	60.1	**64.1**	56.6	42.4	58.6	41.2	
	LogicVista	55.9	47.9	46.1	57.0	**81.9**	**43.2**	41.8	37.4	31.5	40.3	39.6	
	MMMU-Pro	48.1	49.0	45.6	70.6	**76.7**	41.1	41.1	38.6	**46.9**	42.0	37.3	
	MVBench	**74.5**	72.6	72.5	64.4	72.8	**73.2**	69.1	70.1	48.9	69.5	56.1	
	BlinkSpatial	88.8	88.1	87.4	90.9	**92.3**	81.8	**87.4**	83.9	76.9	76.9	83.2	
	BlinkDepth	**91.9**	82.3	85.5	87.1	79.8	**92.7**	87.1	87.9	77.4	91.1	81.5	
	RefCOCO	89.5	**90.6**	70.7	81.8	84.3	84.3	**87.3**	81.1	71.5	75.9	74.3	
	HallusionBench	49.0	52.8	50.8	57.8	**64.2**	45.3	**50.5**	42.0	42.0	45.7	40.2	
	IFBench	37.0	37.0	28.2	**52.3**	42.5	28.5	32.0	26.0	**34.2**	22.8	25.8	
	通用平均	73.7	72.8	70.0	69.8	**77.5**	**69.6**	68.9	66.3	53.1	65.8	62.8	
机器人17项基准	Cosmos-ER	74.1	61.3	**74.9**	54.3	61.1	69.7	56.9	**71.2**	44.3	54.9	55.6	
	Cosmos-CS	66.4	63.4	65.2	61.1	**69.5**	**63.9**	58.4	63.1	43.2	54.1	53.8	
	RefSpatial	57.0	52.7	48.0	46.6	**70.0**	**53.1**	47.6	41.1	24.6	46.6	41.3	
	VSI-Bench	**60.9**	59.5	58.0	47.6	47.5	54.9	55.1	52.0	28.4	**63.0**	46.7	
	SparBench	**54.9**	48.0	42.9	46.6	51.5	**54.8**	40.1	38.0	28.5	49.5	41.2	
	RynnBrain-Area	53.0	53.1	50.4	58.7	**65.4**	52.0	33.2	43.4	35.7	**56.6**	47.1	
	RynnBrain-Spatial	34.5	16.8	**42.2**	32.4	36.8	26.1	37.5	35.5	33.9	**59.0**	37.2	
	RynnBrain-Trajectory	69.3	61.6	64.6	64.4	**71.3**	**67.9**	54.8	64.4	63.5	61.1	61.1	
	RynnBrain-Affordance	84.6	84.2	86.8	**87.8**	86.8	85.1	82.6	84.6	84.4	85.3	**85.7**	
	RynnBrain-Object	48.3	**57.0**	44.9	47.2	51.5	39.8	49.2	42.0	35.2	**71.6**	33.5	
	RynnBrain-Grounding	74.4	77.0	76.7	72.4	**82.8**	72.9	68.8	72.5	59.3	**74.2**	57.8	
	MMSIBench	**41.8**	33.0	31.4	32.3	40.8	36.2	26.3	29.5	37.6	**38.4**	30.1	
	MMSIVideoBench	26.1	33.8	29.6	33.6	**38.6**	27.2	28.8	29.4	**39.4**	28.2	30.0	
	HealthSurgiBench	44.5	24.4	**53.9**	19.1	24.6	**56.1**	23.0	46.9	20.7	23.3	24.0	
	ERQA	51.2	46.5	42.8	47.8	**65.2**	**46.0**	44.0	44.2	30.2	43.0	42.0	
	RoboSpatialHome	**70.0**	65.1	64.3	63.0	65.1	66.3	64.8	64.5	42.0	**71.4**	66.1	
	Where2Place	**71.0**	56.0	59.0	52.0	61.0	**64.0**	53.0	50.0	17.0	11.0	58.0	
	机器人平均	57.8	52.6	55.0	51.0	**58.2**	**55.1**	48.5	51.3	39.3	52.4	47.7	
智慧基础设施9项基准	VANTAGE-2DGrounding	**76.2**	72.4	45.7	45.1	46.9	**75.6**	73.3	66.9	10.1	67.6	59.4	
	VANTAGE-Astro2D	**81.5**	76.8	22.6	68.7	77.7	78.3	69.2	**81.1**	58.8	10.5	57.8	
	VANTAGE-2DPointing	72.9	75.6	74.0	76.8	**85.6**	**74.8**	68.6	68.7	43.4	64.9	55.0	
	VANTAGE-DVC	29.5	29.4	30.1	29.6	**30.2**	31.4	29.6	**32.5**	14.3	28.4	2.2	
	VANTAGE-EventVerif	71.3	60.0	**73.6**	55.2	67.5	**68.9**	59.4	64.1	40.6	58.9	58.0	
	VANTAGE-SOT	62.2	44.2	33.1	54.7	**72.7**	**59.2**	33.1	37.7	16.0	4.8	9.2	
	VANTAGE-Temporal	**51.9**	46.8	50.5	33.0	41.7	**48.0**	43.3	47.3	16.9	20.1	5.9	
	VANTAGE-VQA	69.5	**71.3**	70.3	67.0	71.2	**69.0**	66.4	68.0	51.5	65.4	67.5	

类别	Benchmark	大型模型块(Super 级)					中型模型块(Nano 级)						
		Cosmos 3 Super	Qwen3-VL 32B	Cosmos-Reason2 32B	Gemma-4 31B	Gemini 3.1 Pro†	Cosmos 3 Nano	Qwen3-VL 8B	Cosmos-Reason2 8B	Gemma-4 E4B	RynnBrain 8B	MiMo-Embodied 7B	Cos… Ed…
	TARBench	**48.4**	28.1	36.6	31.9	33.6	**43.6**	31.5	34.1	12.8	34.8	32.6	…
	智慧基础设施平均	**62.6**	56.1	48.5	51.3	58.6	**61.0**	52.7	55.6	29.4	39.5	38.6	…
驾驶3项基准	LingoQA	**76.8**	66.4	70.0	57.2	71.2	71.4	68.4	**71.6**	24.2	60.4	70.4	…
	AVSpecialCollision	**79.3**	37.3	77.3	32.3	54.0	**79.0**	34.0	74.0	33.3	33.7	37.3	…
	AVSpecialStopBehavior	**81.6**	18.4	69.4	20.4	16.3	**77.5**	36.7	59.2	20.4	36.7	42.9	…
	驾驶平均	**79.3**	40.7	72.2	36.6	47.2	**76.0**	46.4	68.3	26.0	43.6	50.2	…

> **译注** 表 10 中两处原文未着重强调的对比:其一,Edge 档的通用平均分(59.7)略低于同档的 Qwen3-VL 2B(60.3)与 RynnBrain 2B(59.9),Edge 的优势集中在机器人、智慧基础设施与驾驶三个领域类别;其二,RynnBrain-* 系列子基准与对比模型 RynnBrain 同源,RynnBrain 模型在其中多项(如 RynnBrain-Spatial、RynnBrain-Object)明显占优,解读这些子项时需留意同源偏置。

通用。我们选取下列 19 个基准评测来考察模型的通用能力。

- *宽泛的视觉问答(VQA)与多模态理解*由 MMBench_DEV (Liu et al., 2024d)、RealWorldQA (xAI, 2024) 与 AI2D (Kembhavi et al., 2016) 度量,它们检验模型能否在解读自然图像、示意图与真实世界场景的同时,回答多样的语义与组合式问题。
- *空间、定位与数量推理*由 CVBench (Tong et al., 2024)、BlinkSpatial、BlinkDepth (Fu et al., 2024)、RefCOCO (Yu et al., 2016) 与 CountBenchQA (Paiss et al., 2023) 评测,覆盖 2D/3D 空间关系、深度感知、指代表达定位与物体计数。
- *文本密集的视觉理解*由 DocVQA (Mathew et al., 2021)、InfoVQA (Mathew et al., 2022) 与 OCRBench-v2 (Fu et al., 2026) 覆盖,要求对文档、信息图、场景文字与结构化视觉版面进行阅读与推理。
- *视频、物理与因果推理*由 MVBench (Li et al., 2024c)、VideoPhy2 (Bansal et al., 2025b)、MVPBench (Krojer et al., 2025) 与 CausalVQA (Foss et al., 2025) 考察,检验跨帧的时序事件理解、物理合理性与因果推理。
- *高阶推理、鲁棒性与指令遵循*由 LogicVista (Xiao et al., 2024b)、MMMU_Pro (Yue et al., 2025)、HallusionBench (Guan et al., 2024) 与 IFBench (Pyatkin et al., 2025b) 度量,分别探测视觉逻辑、专家级多模态问题求解、幻觉抵抗能力与对用户指令的遵循程度。

这些基准评测合在一起,从感知、定位、文字识别、时序理解到可靠的指令条件化回答生成,为模型的通用推理能力提供了一幅全景视图。

机器人。我们将 17 个机器人与具身推理(embodied reasoning)基准评测归入若干能力族。

- *具身常识与任务推理*由 Cosmos Reason (NVIDIA, 2025) 中的 Embodied Reasoning 与 CommonSense、ERQA (Gemini Robotics Team et al., 2025) 以及 Where2Place (Yuan et al., 2024) 度量,检验模型能否在具身环境中就物体可供性(affordance)、可行动作、放置决策与物理常识进行推理。
- *空间定位与场景几何*由 RefSpatial (Zhou et al., 2025b)、VSI-Bench (Yang et al., 2025b)、SparBench (Zhang et al., 2025b) 与 RoboSpatialHome (Song et al., 2025) 评测,覆盖指代表达定位、度量式与关系式空间理解、室内布局推理、自由空间感知,以及与机器人相关的定位。
- *面向机器人的感知与动作理解*由 RynnBrain (Dang et al., 2026) 与 ERQA (Gemini Robotics Team et al., 2025) 捕捉,它们在真实机器人任务片段上探测与操纵相关的感知、动作可行性、轨迹推理与以物体为中心的决策。MMSIVideoBench (Lin et al., 2025c) 则要求跨多视角或多帧推理,以推断物体对应关系、空间关系、时序变化与场景级结构。我们还自建了 HealthSurgiBench,一个面向手术室理解的内部基准评测,含 23 种题型:对计数、工具、角色、检测框、坐标、时间/距离估计、有序列表、场景图、监视器文字等结构化输出采用基于规则的评分,并用本地部署的 `Qwen3-4B` 判官评判六类自由形式答案。最终得分对每个样本使用相应的评测器,然后报告全部样本的无加权平均。

这些基准评测合在一起,评估模型能否超越静态视觉识别、走向具身推理:理解物体在哪里、它们如何相互关联、哪些动作是可行的,以及空间证据如何随视角与时间演化。

智慧基础设施。我们在 VANTAGE-Bench (NVIDIA, 2026c) 与 Traffic Anomaly Reasoning (TAR) (NVIDIA, 2026b) 上评测,覆盖仓储物流、交通运输与智慧基础设施等固定摄像头场景。VANTAGE-Bench 通过事件验证、视觉问答、指代、指点、定位、时序定位、稠密描述与 VLM 原生单目标跟踪来度量语义、空间、时序与时空理解,共含 3,346 个素材与 35,027 条专家标注,其中包括合成的异常事件录像。TAR 是 AI City Challenge 2026 Track 3 评测套件,以异构问答、时序 IoU 与文本生成指标评测异常验证、时序定位、场景描述、因果分析、摘要与域外泛化。

驾驶。我们用 LingoQA (Marcu et al., 2024) 与两个内部的安全关键驾驶事件分类基准评测来评估驾驶能力。AVSpecialCollisionBench 度量模型能否将每段视频正确归入三类事件之一:碰撞、险些碰撞(near collision)或无碰撞;该基准每类含 100 段视频,先计算逐类准确率,再以三类准确率的均值作为最终得分。AVSpecialStopBehaviorBench 评测五类停车标志行为的分类:完全停车、滑行停车(rolling stop)、未停车、不相关与假标志;该基准每类含 10 段视频,最终得分为五个逐类准确率的均值。

在通用基准评测上,Cosmos 3 与开源模型相比具有竞争力,但仍落后于 Gemini 3.1 Pro (Google DeepMind, 2025a)。与 Cosmos-Reason2 相比,Cosmos 3 展现出更强的通用能力,这得益于额外增加的 20% 预训练数据所带来的数据多样性提升。在机器人、智慧基础设施与驾驶领域,Cosmos 3 优于包括 RynnBrain (Dang et al., 2026)、Mimo-Embodied (Hao et al., 2026) 与 Gemma-4 (Google DeepMind, 2026b) 在内的开源与闭源模型,唯一的例外是在机器人领域与 Gemini 3.1 Pro 尚有微小差距。总体而言,Cosmos 3 在机器人、智慧基础设施与自动驾驶上展现出强劲的领域特定推理能力,可支撑广泛的物理 AI 应用。

6.2 Generator 评测

Cosmos 3 的 Generator 组件在一组多样化的任务上进行评测,这些任务共同度量其为物理 AI 模拟与生成多模态世界的能力。与聚焦单一模态的常规生成模型不同,Cosmos 3 在统一框架内对图像、视频、音频与动作进行联合建模,因此评测横跨图像生成、视频生成、音视频生成、迁移生成、正向与逆向动力学以及机器人策略学习。我们进一步评测了专用后训练变体,包括 `Cosmos3-Super-Text2Image`、`Cosmos3-Super-Image2Video` 与 `Cosmos3-Nano-Policy-DROID`,以考察从共享的全模态基础模型向下游适配的有效性。我们的基准评测套件综合了自动指标、人工评测(human evaluation)、领域特定的物理 AI 基准评测与真实世界机器人评测,覆盖提示词遵循(prompt adherence)、物理合理性、时序一致性、音视频同步、可控性、动作预测与任务完成等关键能力。

6.2.1 图像生成评测

我们将 Cosmos 3 的图像生成作为单帧视觉生成来评测,聚焦四条互补的轴线:宽泛的语义提示词遵循、精确的场景文字渲染、人类偏好对齐与视觉美学(aesthetic)。UniGenBench 是主要的提示词遵循指标,因为它能在测试点(testpoint)层面暴露失败;我们为 UniGenBench 增补了一个物理 AI 子集。CVTG 通过基于 OCR 的 GNED 与 PNED 分数,单独考察图像生成器的一类常见失败模式——场景文字被拼错、遗漏、模糊或重复。HPSv3 与 LAION 美学分则从整体人类偏好与视觉观感上补足这些针对性检查,合起来比任何单一聚合分数都能对文生图(Text-to-Image, T2I)质量给出更可操作的观察。对所有基准评测,我们使用 `Claude-Opus-4.7` 作为提示词改写器,把评测文本提示词转换成 §6.3.1 与 §6.3.2 所述的结构化格式,以确保其匹配生成器所偏好的训练提示词风格;指令模板见附录 B.2。所有模型均在 1024×1024 分辨率下生成并比较;对开源模型,我们遵循其文档中推荐的超参数与负向提示词。定性示例另见图 17。

> **译注** 该评测协议对所有被测模型统一使用按"生成器(即 Cosmos)偏好的训练提示词风格"改写后的结构化提示词。在此协议下各基线的得分与其原论文报告值属于不同提示词协议,不可直接跨文献比较。

图 17: `Cosmos3-Super-Text2Image` 生成的示例图像。我们的模型生成的图像既物理合理又具照片真实感(photorealism),呈现出连贯的物体几何、一致的物体–环境交互等特性。所有图像均由单次上采样得到的 JSON 提示词生成,使用 `shift=3.0`、`guidance=4.0` 与 50 步扩散步数(设置同表 21 所述)。这些结果凸显了该模型作为高效真实世界图像模拟器的潜力,可用于机器人、自动驾驶以及其他必须遵守物理定律的场景。

UniGenBench。UniGenBench (Wang et al., 2025h) 是面向文生图的统一语义评测基准。它包含 600 条提示词,横跨 5 个主主题与 20 个子主题,每条提示词按 10 个一级与 27 个二级评测准则、以"多模态大模型当判官"(MLLM-as-Judge)的框架进行评估。为更好地考察 Cosmos 3 在物理 AI 场景中的能力,我们为该基准增补了 570 条针对照片真实感物理世界场景的提示词(UniGenBench-Phys)。这些提示词覆盖六个物理世界子领域:(a) 机器人与工业、(b) 实体标识与文字、(c) 自动驾驶、(d) 建筑工地、(e) 流体动力学、(f) 医疗/临床。由 Gemini 3.1 Pro 以二值通过/不通过的方式,将每张生成图像对照每个测试点进行评判;主分数为测试点平均准确率,按全部 1,170 条提示词(即 "All")报告,并额外给出原始子集与物理子集(即 "Orig" 与 "Phys")的细分。

CVTG。CVTG (Du et al., 2025)(复杂视觉文字生成,Complex Visual Text Generation)评测模型在视觉复杂场景中准确渲染文字的能力。这一能力对真实世界环境尤其重要——标识、标签与说明文字必须清晰可读且拼写正确。生成图像先由 OCR 系统提取可见文字区域,再基于归一化编辑距离(NED)做匈牙利匹配,把预测字符串与目标字符串对齐。我们在两个提示词集合上评测:(a) CVTG-500L,一个以英文为主的子集,由 CVTG-2K 随机采样的 500 条提示词组成,同时保持对文字区域数量的覆盖大体相当;我们进一步把每条短提示词上采样为长而稠密的描述,以更好地贴合现代世界模型生成的要求。(b) CVTG-102ch,一个以中文为主的 102 条提示词集合,用于评测中文字符的准确渲染;这些提示词取自常见领域及物理落地领域,包括公共空间、户外与景区环境、数字显示屏等真实世界场景。

表 11:文生图基准评测结果。UniGenBench 分数为满足评测准则的比例;"All (1170)" 聚合了原始 600 条提示词(Orig)与 570 条物理 AI 提示词(Phys)。PNED 与 GNED 为字符级准确度指标(越高越好);CVTG-102ch 考察中文字符渲染,CVTG-500L 考察英文长提示词渲染。Aesthetic v2 与 HPSv3 为越高越好的图像级指标。 †
† Gemini 3 Pro Image 有较高概率生成大小写不符的场景文字(如 "Adventure" → "ADVENTURE")。若排除这一错误,其 CVTG-500L 分数升至 GNED = 75.97、PNED = 91.45。

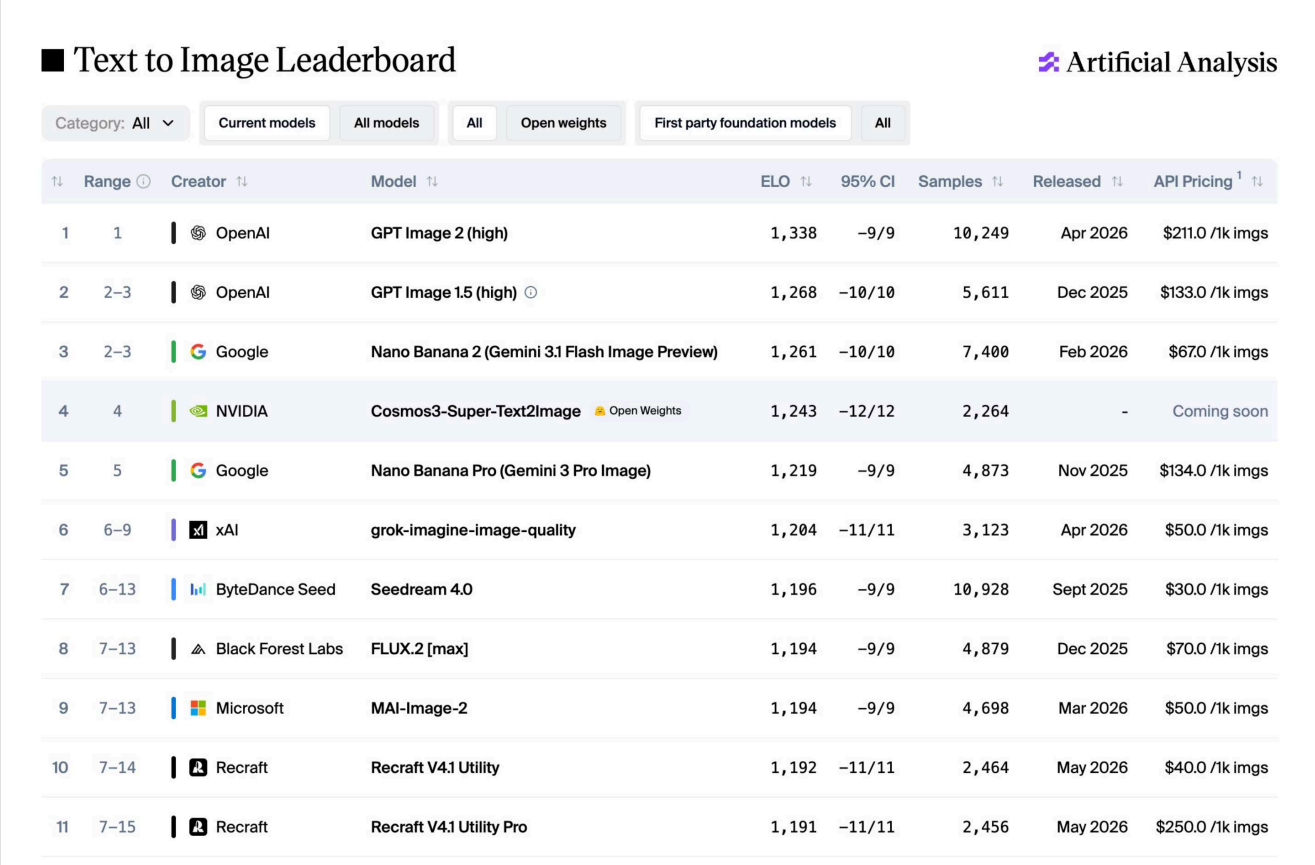
Model	类型	UniGenBench			CVTG-500L		CVTG-102ch		图像级指标	
		All (↑)	Orig (↑)	Phys (↑)	GNED (↑)	PNED (↑)	GNED (↑)	PNED (↑)	Aesv2 (↑)	HPSv3 (↑)
`Cosmos3-Super-Text2Image`	开源	**91.36**	**93.34**	**89.54**	**80.88**	89.08	32.02	41.22	5.91	11.60
`Cosmos3-Super`	开源	87.33	85.21	89.64	66.77	70.97	8.48	16.31	5.76	9.49
`Cosmos3-Nano`	开源	84.61	87.32	82.12	24.23	26.53	4.63	9.70	5.76	8.99
Gemini 3 Pro Image	闭源	90.69	92.81	89.74	59.24†	71.79†	46.00	**76.40**	5.70	11.78
FLUX.2-dev	开源	87.60	89.77	85.61	74.71	84.98	44.33	68.74	5.75	11.38
Qwen-Image-2512	开源	84.25	87.32	81.44	79.68	**90.86**	46.33	71.26	**5.92**	11.03
Hunyuan 3.0	开源	84.02	87.68	80.67	71.40	87.68	49.05	71.31	5.90	**11.93**
Z-Image-Turbo	开源	78.14	81.53	75.03	75.20	86.95	**49.18**	73.32	5.70	11.36

译注 表 11 有三处值得留意、原文正文未展开的地方。其一,Phys 列的加粗与数值不符——被加粗的 89.54(Cosmos3-Super-Text2Image)并非该列最高,Gemini 3 Pro Image 的 89.74 与 Cosmos3-Super 的 89.64 都高于它;Cosmos 的领先主要体现在 All 与 Orig 两列。其二,中文场景文字渲染(CVTG-102ch)是 Cosmos 系的明显短板:其最好成绩 32.02/41.22 远低于 Z-Image-Turbo(GNED 49.18)与 Gemini(PNED 76.40)。其三,图像级偏好指标上 Cosmos 也未领先:HPSv3 最高为 Hunyuan 3.0(11.93),Aesthetic v2 最高为 Qwen-Image-2512(5.92)。

HPSv3。HPSv3 (Ma et al., 2025) 用一个学习得到的人类偏好奖励模型来评测"感知提示词的"文生图质量。它建立在 HPDv3 之上——一个横跨合成图像与真实图像、覆盖宽广质量区间的偏好数据集。HPSv3 采用基于 VLM 的架构,并以不确定性感知的排序损失训练,产出一个以提示词为条件的分数,反映人类对语义对齐、真实感与美学质量的综合判断。我们将 HPSv3 直接应用于 Cosmos 3 的生成结果,以获得这一互补的人类偏好指标。

Aesthetic V2。Aesthetic V2 (Schuhmann and LAION, 2022) 用 LAION 美学预测器度量与提示词无关的视觉观感。该预测器用一个在 CLIP 图像嵌入之上训练的轻量模型,估计人们对一张图像的喜好程度(1–10 分)。与 HPSv3 不同,该分数不以输入提示词为条件,因此并不直接度量指令遵循或语义对齐。我们把它作为独立的图像质量信号,用来捕捉一般性的视觉吸引力与构图质量。

Artificial Analysis 文生图排行榜。为了在真实使用场景下评测 Cosmos 3,我们将专用 T2I 模型 `Cosmos3-Super-Text2Image` 连同一个智能体化调用框架(agentic harness)一起提交到 Artificial Analysis Text-to-Image 排行榜,参与众包公开投票。该模型在全部开放权重模型中排名第 1,在全部模型(开源 + 闭源)中排名第 4。提交的更多细节请见附录 B.7。



	Range	Creator	Model	ELO	95% CI	Samples	Released	API Pricing [1]
1	1	OpenAI	GPT Image 2 (high)	1,338	−9/9	10,249	Apr 2026	$211.0 /1k imgs
2	2–3	OpenAI	GPT Image 1.5 (high)	1,268	−10/10	5,611	Dec 2025	$133.0 /1k imgs
3	2–3	Google	Nano Banana 2 (Gemini 3.1 Flash Image Preview)	1,261	−10/10	7,400	Feb 2026	$67.0 /1k imgs
4	4	NVIDIA	Cosmos3-Super-Text2Image Open Weights	1,243	−12/12	2,264	-	Coming soon
5	5	Google	Nano Banana Pro (Gemini 3 Pro Image)	1,219	−9/9	4,873	Nov 2025	$134.0 /1k imgs
6	6–9	xAI	grok-imagine-image-quality	1,204	−11/11	3,123	Apr 2026	$50.0 /1k imgs
7	6–13	ByteDance Seed	Seedream 4.0	1,196	−9/9	10,928	Sept 2025	$30.0 /1k imgs
8	7–13	Black Forest Labs	FLUX.2 [max]	1,194	−9/9	4,879	Dec 2025	$70.0 /1k imgs
9	7–13	Microsoft	MAI-Image-2	1,194	−9/9	4,698	Mar 2026	$50.0 /1k imgs
10	7–14	Recraft	Recraft V4.1 Utility	1,192	−11/11	2,464	May 2026	$40.0 /1k imgs
11	7–15	Recraft	Recraft V4.1 Utility Pro	1,191	−11/11	2,456	May 2026	$250.0 /1k imgs

图 18: `Cosmos3-Super-Text2Image` 是众包竞技场排名中的第一名开放权重模型。在 Artificial Analysis Text to Image 排行榜上(日期:2026-05-28),`Cosmos3-Super-Text2Image` 在开放权重模型中排名第 1(计入闭源模型则为第 4)。

6.2.2 视频生成评测(Video Generation Evaluation)

我们通过互补的自动化与人工基准评测(benchmark)来评估 Cosmos 3 的视频生成能力。自动化基准评测提供可扩展、可复现的比较,并能捕捉广泛的质量信号与领域特定信号,但随着模型水平提升,其判别力会逐渐减弱(在领域特定的 Physical-AI 场景中尤为明显——此类指标通常对时序性与物理性的失败"视而不见")。人工评测(human evaluation)弥补了这些缺口:既能发现自动化协议系统性遗漏的长尾物理、具身与语义失败,又能把分数拉开到有意义的更宽区间,使最先进水平(SOTA)模型之间微小但真实的差异仍可被检出。我们在三个自动化基准上报告结果——PAIBench-G、RBench 与 Physics-IQ——随后是 Cosmos HUE(一个面向 Physical AI 视频生成的专用人工评测协议)与 Human World Bench(HWB,针对任务级指令下真实人体运动的定向人工评测)。对所有基准(无论自动化还是人工评测),我们都使用 `Claude-Opus-4.6` 作为提示词改写器,把文本提示词转换为 §6.3.1 与 §6.3.2 所述的结构化格式,确保其匹配训练提示词分布。

PAIBench-G。PAIBench-G 是 Physical AI Benchmark(Zhou et al., 2025c)的视频生成赛道,包含 1,044 个图像–文本提示词对,覆盖六个 Physical AI 领域:Human(299)、Autonomous Vehicle(239)、Common Sense(174)、Robotics(107)、Physics(107)与 Industry(107)。我们采用它是因为其领域覆盖广、评分均衡:每个模型获得一个 Quality Score(汇总帧一致性、运动平滑度、美学质量与视频–文本对齐等指标)和一个 Domain Score(以 VLM 为裁判(VLM-as-Judge)对各领域的物理与语义准确性做二元核验),总分对两者等权:Overall = 0.5 × Quality + 0.5 × Domain。PAIBench-G 与人工 ELO 排名的 Pearson 相关系数达 r=0.918,非常适合在训练早期阶段可靠地度量进展——那时模型尚未收敛到自动化分数开始饱和的前沿。对每个提示词,我们用 5 个随机种子生成视频,并在 720p 分辨率、16:9 宽高比、189 帧的设定下评测。PAIBench-G 原生只支持图生视频(Image-to-Video)评测,我们将其扩展到同时计算文生视频(Text-to-Video)分数[2]。表 12 报告了以 `Qwen2.5-VL-72B-Instruct` 为 VLM 裁判的文生视频与图生视频 PAIBench-G 结果;Cosmos3-Super 在两条赛道上均取得最先进水平(SOTA)的总分,成为最佳开源模型,并超过 Veo-3.1 等强力闭源模型。

[2] 我们发现公开的 PAIBench-G 图生视频排行榜结果(由 `Qwen3-VL-235B-A22B` 评判)无法复现。因此我们的内部评测(表 12)改用 `Qwen2.5-VL-72B-Instruct` 作为 Domain 分数裁判,并另行独立提交公开排行榜;在这两处,Cosmos3-Super 与 Cosmos3-Nano 的总分均分列第一、第二。

表 12:PAIBench-G 与 RBench 结果,覆盖文生视频与图生视频两种设定。PAIBench-G 评测六个 Physical AI 领域上的视觉质量与领域特定准确性,RBench 评测具身机器人场景中的任务正确性与物理合理性。**Cosmos3-Super** 在开源模型中取得 PAIBench-G T2V 与 I2V 的最高总分,**Cosmos3-Nano** 在 RBench 上领先。**加粗**表示该列第一,下划线表示第二。

Model	Type	PAIBench-G Text2Video (↑)			PAIBench-G Image2Video (↑)			RBench Image2Video (↑)
		Overall	Domain	Quality	Overall	Domain	Quality	Score
Cosmos3-Super	Open-source	**80.0**	**86.8**	73.1	**82.8**	87.3	**78.2**	58.1%
Cosmos3-Nano	Open-source	79.4	85.8	73.0	82.7	87.2	78.1	58.4%
Wan2.2-A14B	Open-source	78.0	83.2	72.8	81.3	85.3	77.3	50.7%
HunyuanVideo-1.5	Open-source	76.5	80.9	72.0	81.7	85.9	77.6	46.0%
Cosmos-Predict2.5-2B	Open-source	76.5	79.9	**73.2**	81.2	84.6	77.9	46.4%
Cosmos-Predict2.5-14B	Open-source	76.4	79.5	**73.2**	81.1	84.0	78.1	–
Wan2.1-14B	Open-source	76.4	80.1	72.7	80.2	83.6	76.8	–
Wan2.2-5B	Open-source	76.3	79.6	73.0	81.0	84.6	77.4	–
Veo-3.1	Closed-source	79.1	85.2	72.9	82.6	**87.6**	77.6	56.3%
Seedance-1.5-Pro	Closed-source	76.9	82.1	71.6	80.8	84.7	76.9	58.4%
Wan 2.6	Closed-source	78.6	85.2	72.0	81.9	85.9	77.8	**60.7%**

> **译注** 注意逐列细读:Cosmos3-Super 领先的是"总分",并非全部子分——T2V 的 Quality 子分上,其前代 Cosmos-Predict2.5(73.2)反而略高于 Cosmos3-Super(73.1);I2V 的 Domain 子分由闭源 Veo-3.1(87.6)居首。RBench 一列的全场第一是闭源的 Wan 2.6(60.7%),Cosmos3-Nano(58.4%)仅与 Seedance-1.5-Pro 并列第二,"Cosmos3-Nano 领先"仅限开源阵营。另外,本表 Domain 裁判已由公开排行榜所用的 Qwen3-VL-235B-A22B 换为 Qwen2.5-VL-72B-Instruct(见原文脚注),表中数值不能与公开排行榜直接对照。

RBench。RBench(Deng et al., 2026)在具身任务场景中评测视频生成,强调机器人–物体交互中的任务正确性与物理合理性(physical plausibility),而非单纯的感知层面真实感。PAIBench-G 广覆盖各个 Physical AI 领域,RBench 则深入机器人这一关键 Physical AI 用例,对模型能否在多样机器人形态下生成物理上连贯的操作序列做压力测试。该基准包含 650 个图像–文本评测用例,来自两个互补子集:250 个面向任务的样本对,横跨五个类别(Common Manipulation、Long-horizon Planning、Multi-entity Collaboration、Spatial Relationship 与 Visual Reasoning);以及 400 个特定具身形态的样本对,覆盖四种机器人形态(Dual-arm、Humanoid、Single-arm 与 Quadruped)。提示词取自 RoVid-X(Deng et al., 2026)——一个包含 400 万条机器人视频片段、覆盖 1,300 多种技能的精选语料库。每条视频按 Task Completion(TC,物理–语义合理性与任务遵循一致性的平均)与 Visual Quality(VQ,机器人主体稳定性与运动平滑度的带惩罚组合)两项打分,最终分数为全部 650 个用例上两者的均值。RBench 与人工判断在 25 个受评模型上的 Spearman 相关系数为 $\rho = 0.96$,能为具身生成质量提供可靠信号。对每个提示词,我们用单一随机种子生成视频,并在 720p 分辨率、16:9 宽高比、121 帧的设定下评测。表 12 报告了 Cosmos 3 在 RBench 上的图生视频结果——Cosmos 3 是其中最佳的开源模型。

如表 12 所示,Cosmos3-Super 在 T2V 与 I2V 两条赛道上均取得所有模型(含闭源)中最高的 PAIBench-G 总分,而 Cosmos3-Nano 在 RBench 上追平了第二名的成绩。两个 Cosmos 3 变体也都显著超越各自的前代 Cosmos-Predict2.5,体现了全模态(omnimodal)架构带来的增益。不过,这两个基准虽然把物理合理性纳入评分,却并未深入探查这一维度。因此我们转向 Physics-IQ,对物理遵循程度做定向评测。

表 13:Physics-IQ 基准评测结果,按条件模式分列。WMReward(BoN)表示使用 WMReward(Yuan et al., 2026)做 best-of-*N* 重排序。Physics-IQ 分数越高越好。**Cosmos3-Super** 在 I2V 与 V2V 两种模式下、无论是否使用 WMReward+BoN,均取得最先进水平的成绩。**加粗**表示列组内最优,下划线表示第二。

Model	Mode	Score (↑)
(a) 图生视频(I2V)		
Cosmos3-Super	I2V + WMReward (BoN)	**48.9**
Cosmos3-Nano	I2V + WMReward (BoN)	43.8
Sora2 (Closed-sourced)	I2V + WMReward (BoN)	46.4
Wan2.2-A14B (Open-sourced)	I2V + WMReward (BoN)	44.4
Cosmos3-Super	I2V	**43.8**
Cosmos3-Nano	I2V	40.2
Sora2 (Closed-sourced)	I2V	42.3
Wan2.2-A14B (Open-sourced)	I2V	38.3
(b) 视频生视频(V2V)		
Cosmos3-Super	V2V + WMReward (BoN)	**63.4**
Cosmos3-Nano	V2V + WMReward (BoN)	57.7
Magi-1 (Open-sourced)	V2V + WMReward (BoN)	62.6
Cosmos3-Super	V2V	**59.7**
Cosmos3-Nano	V2V	50.2
Magi-1 (Open-sourced)	V2V	56.0
Video-GPT (Open-sourced)	V2V	35.0
VideoPoet (Closed-sourced)	V2V	29.5

> **译注** 原文只强调 Super 的 SOTA,但 Cosmos3-Nano 在多处被基线反超:I2V+BoN 下 Nano(43.8)低于 Sora2(46.4)甚至 Wan2.2-A14B(44.4);V2V+BoN 下 Nano(57.7)明显低于 Magi-1(62.6)。此外,本文对 Cosmos 的评测使用了迭代式提示词上采样器生成提示词,并遵循公开排行榜协议引入 WMReward+BoN;各基线分数所对应的提示词与推理协议未必与 Cosmos 完全一致,跨模型比较需留意协议差异。

Physics-IQ。Physics-IQ(Motamed et al., 2026)评测视频生成模型是否掌握了物理规律:以真实世界的起始上下文为条件,度量生成的后续内容与实际物理结果的接近程度。不同于把物理合理性只当作众多评分成分之一的 PAIBench-G 与 RBench,Physics-IQ 将其孤立为唯一的评测轴,沿精确的空间与时间维度衡量生成运动是否与真值物理结果一致。Physics-IQ 覆盖五个物理类别——固体力学、流体动力学、光学、热力学与磁学——包含从三个固定视角拍摄的 396 个真实场景,并为文本条件模型配有成对的文字描述。它支持两种评测模式:在图生视频(I2V)模式下,模型以单张切换帧(switch frame)外加可选的文本提示词为条件,预测随后的运动;在视频生视频(V2V)续写模式下,模型以一段 3 秒条件视频外加可选的文本提示词为条件,预测接下来 5 秒的运动。生成视频沿四个互补维度(空间重叠、时间对齐、幅值加权的空间一致性、像素级误差)与真值物理续写比较,再按"真实对真实"的上界归一化为一个 0–100 的单一分数,可跨模型、跨条件模式直接比较。

我们在表 13 中以 I2V 与 V2V 两种模式评测 Cosmos3-Super,并与两种模式下领先的开源与商业基线对比。V2V 使用完整的 3 秒条件视频作为输入。我们还使用了一个提示词上采样器(prompt upsampler),沿用与 PhyT2V(Xue et al., 2025)类似的迭代求精策略,在 I2V 设定下从切换帧、在 V2V 设定下从 3 秒条件视频生成文本提示词。遵循公开 Physics-IQ 排行榜与 WMReward 推理时对齐协议(Yuan et al., 2026),我们额外为 Cosmos3-Super 在 I2V 与 V2V 上运行了 WMReward + best-of-*N*(BoN)打分。由表 13 可见,Cosmos3-Super 在 I2V 上达到最先进水平:其直接得分 43.8 超过各直接 I2V 基线,而 WMReward+BoN 进一步把分数提升到 48.9,高于表中最强的使用 WMReward+BoN 的 I2V 基线。在 V2V 上,Cosmos3-Super 同样达到最先进水平:直接得分 59.7,加 WMReward + BoN 后达 63.4,高于表中最强的使用 WMReward+BoN 的 V2V 基线。

自动化评测指标在衡量质量、提示词对齐与物理合理性上固然有价值,但即便最强的基于 VLM 的指标也会漏掉某些瑕疵(artifact)与失败案例。人工评测从两方面补充自动化指标:一是覆盖自动化协议系统性遗漏的长尾失败,二是在更宽的分数区间上提供判别力。例如,在同一提示词集合上,受评的文生视频生成器在人工评测上的分差约 10 分,而在可比的自动化 PAIBench-G 总分上仅约 4 分(见附录 F)。因此我们用两个人工评测协议补充自动化结果:面向广义 Physical AI 视频生成的 Cosmos HUE,以及专注任务级指令下真实人体运动的 Human World Bench(HWB)。

表 14:Cosmos HUE 与 Human World Bench(HWB)人工评测结果。Cosmos HUE 通过四个维度上的原子化二元核验评测广义 Physical AI 视频生成质量;HWB 通过指令遵循与物理两项通过率评测任务级指令下的真实人体运动。*Ground Truth* 对与相同提示词配对的真实视频打分,作为 Cosmos-HUE 的参照上界。**Cosmos3-Super** 在开源模型中取得最高的 HUE T2V 分数与最高的 HWB 分数。完整的分维度 HUE 排行榜见附录 F。**加粗**表示该列最优;下划线表示第二。

Model	Type	Cosmos HUE		Human World Bench
		Text-to-Video (↑)	Image-to-Video (↑)	Image-to-Video (↑)
Ground Truth	—	93.6	94.4	–
Cosmos3-Super	Open-sourced	89.3	89.6	**71.9**
Cosmos3-Nano	Open-sourced	87.6	88.6	66.9
Wan2.2-A14B	Open-sourced	88.2	88.4	60.7
HunyuanVideo-1.5	Open-sourced	86.5	85.6	54.7
Wan2.1-14B	Open-sourced	84.0	83.9	33.1
Wan2.2-5B	Open-sourced	80.8	80.4	25.4
Cosmos-Predict2.5-14B	Open-sourced	82.1	83.0	38.7
Cosmos-Predict2.5-2B	Open-sourced	81.8	82.6	32.8
Veo-3.1	Closed-sourced	**91.3**	**89.7**	67.8
Seedance-1.5-Pro	Closed-sourced	90.0	87.6	–

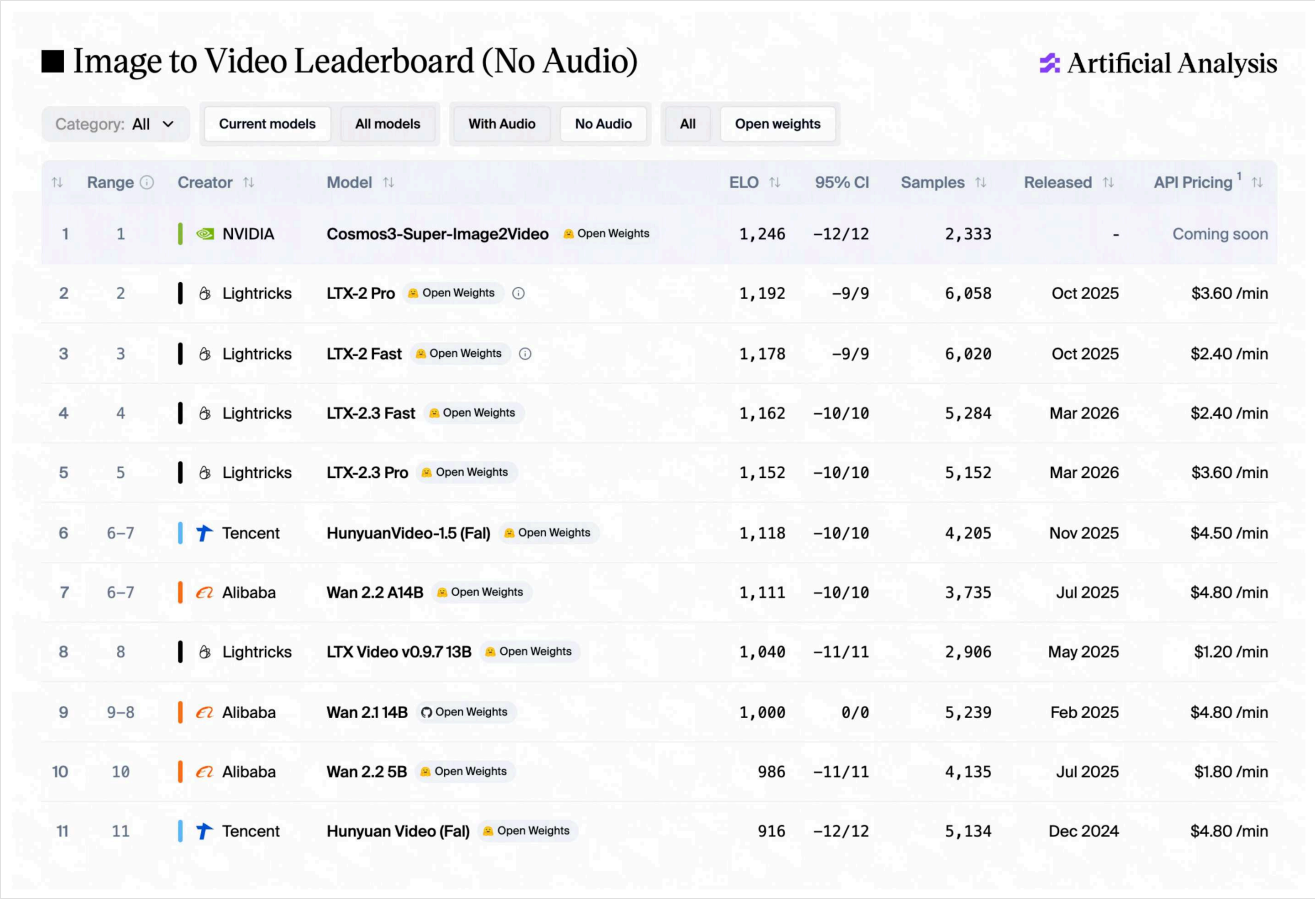
> **译注** HUE 的 T2V 与 I2V 两列最优均为闭源 Veo-3.1,Cosmos3-Super 只是开源第一(I2V 差距仅 0.1 分,T2V 差 2.0 分);HWB 一列 Seedance-1.5-Pro 缺测(—),"全场最佳"的比较范围因此受限。还需注意:Cosmos HUE 与 HWB 均为本文作者自建协议,且所有模型的提示词都先经 Claude-Opus-4.6 改写成与 Cosmos 训练分布匹配的结构化格式,该设定对 Cosmos 模型有利,分数不宜与其他文献直接对比。

Cosmos HUE。Cosmos HUE(HUman Evaluation)是一个人工打分协议,建立在与 §6.2.2 相同的 PAIBench-G 提示词集合之上。HUE 在两方面区别于以往的人工评测协议。其一,它用*原子化二元*核验取代主观的 Likert 量表打分:每条视频被分解为一组单一事实的*是 / 否 / 不确定*问题,且措辞保证"是"始终对应期望的结果,从而把标注员的任务从整体性判断转变为客观的事实核验。每个(视频, 问题)对由两名标注员独立评定,分歧升级交由质量控制审核员裁决。其二,每个提示词的问题集由一个*三层 VLM 流水线自动生成*:Domain Strategist(领域策略器)把提示词归入七个 Physical AI 领域之一,Scene Parser(场景解析器)从提示词(T2V)或成对真值参考视频的采样帧(I2V)生成结构化的场景清单,Auditor(审计器)输出最终的原子问题。这确保问题贴合实际的提示词与参考场景,而非一套固定的评分细则。三层均运行在 `GPT 5.2` 上。VLM 生成的问题集随后再经过一轮人工复核补充:审核员检查排名靠前模型的生成视频,提出针对自动化流水线尚未覆盖的失败模式的附加问题,并把这些问题并回问题库。评测集由从 PAIBench-G 采样、保持其原生领域分布的 100 个固定提示词组成,每个模型对每个提示词用 5 个随机种子生成,即每个模型 500 条视频;每个提示词的问题集(至多 20 个二元问题)应用于全部 5 条生成视频。每个问题被归入四个维度之一:"Semantic Alignment"、"Physical Laws"、"Geometric Reasoning" 与 "Visual Integrity"。分维度、分领域的 T2V 与 I2V 排行榜、正式的打分方案与可靠性估计详见附录 F。

Human World Bench(HWB)。Cosmos HUE 评测的是广义的 Physical AI 视频生成,而 Human World Bench(HWB)聚焦一个更定向、也更具挑战性的设定:任务级指令下、面向人类操作任务的第一人称(egocentric)图生视频生成。HWB 的基准视频取自 EgoVerse(Punamiya et al., 2026),共 180 个样本。HWB 采用绝对失败模式协议:标注员对每条生成视频独立评判其指令遵循与物理合理性。我们报告两个顶层通过率:*指令遵循*衡量视频是否呈现了所要求的动作与物体;*物理合理性*衡量时序连贯性与物理合理程度,包括物体动力学、接触以及手部解剖结构。HWB 分数是指令遵循与物理两项通过率的平均。

表 14 报告了 T2V 与 I2V 两个方向的 Cosmos-HUE 分数,以及 Human World Bench(HWB)结果。在 HUE T2V 上,Cosmos3-Super 以 89.3 位居开源模型之首,落后于闭源的 Veo-3.1(91.3)与 Seedance-1.5-Pro(90.0)。在 HUE I2V 上,Cosmos3-Super 再次成为最佳开源模型,并与领先的闭源生成器基本持平——仅落后 Veo-3.1 0.1 分(89.6 对 89.7)。Cosmos3-Nano 同样具备竞争力,在开源模型中 I2V 位列第二(88.5)、T2V 位列第三(87.6)。〔译注:正文此处的 88.5 与表 14 中的 88.6 不一致,疑为原文笔误。〕

在 HWB 上,Cosmos3-Super 取得 71.9,为表 14 全部受评模型中的最先进水平成绩,比表中最强的闭源基线 Veo-3.1(67.8)高 4.1 分,比最强的非 Cosmos 开源基线 Wan2.2-A14B(60.7)高 11.2 分。Cosmos3-Nano 也表现强劲,以 66.9 在开源模型中排名第二,并超过所有非 Cosmos 开源基线。两个 Cosmos 3 变体合计占据 HWB 开源前两名,展示了两个模型规模上都很强的第一人称人体运动生成能力。



	Range	Creator	Model	ELO	95% CI	Samples	Released	API Pricing [1]
1	1	NVIDIA	Cosmos3-Super-Image2Video Open Weights	1,246	−12/12	2,333	-	Coming soon
2	2	Lightricks	LTX-2 Pro Open Weights	1,192	−9/9	6,058	Oct 2025	$3.60 /min
3	3	Lightricks	LTX-2 Fast Open Weights	1,178	−9/9	6,020	Oct 2025	$2.40 /min
4	4	Lightricks	LTX-2.3 Fast Open Weights	1,162	−10/10	5,284	Mar 2026	$2.40 /min
5	5	Lightricks	LTX-2.3 Pro Open Weights	1,152	−10/10	5,152	Mar 2026	$3.60 /min
6	6–7	Tencent	HunyuanVideo-1.5 (Fal) Open Weights	1,118	−10/10	4,205	Nov 2025	$4.50 /min
7	6–7	Alibaba	Wan 2.2 A14B Open Weights	1,111	−10/10	3,735	Jul 2025	$4.80 /min
8	8	Lightricks	LTX Video v0.9.7 13B Open Weights	1,040	−11/11	2,906	May 2025	$1.20 /min
9	9–8	Alibaba	Wan 2.1 14B Open Weights	1,000	0/0	5,239	Feb 2025	$4.80 /min
10	10	Alibaba	Wan 2.2 5B Open Weights	986	−11/11	4,135	Jul 2025	$1.80 /min
11	11	Tencent	Hunyuan Video (Fal) Open Weights	916	−12/12	5,134	Dec 2024	$4.80 /min

图 19:Cosmos3-Super-Image2Video 是众包竞技场排名中最佳的开放权重模型。Cosmos3-Super-Image2Video 在 Artificial Analysis 图生视频排行榜(无音频)上位列开放权重模型第 1(含专有模型在内位列第 22)(日期:2026-05-28)。

Artificial Analysis 图生视频排行榜。为在真实世界场景下评测 Cosmos 3,我们把专用的 I2V 模型 `Cosmos3-Super-Image2Video` 提交到 Artificial Analysis 图生视频排行榜(无音频)进行众包公开投票。该模型在所有开放权重模型中排名第一,含专有模型在内则位列全球第 22。特别地,该模型与 Veo 3.1(Google DeepMind, 2025b)、Wan 2.5(Wan et al., 2025b)等专有产品不相上下,展示了出色的时序动态与视觉保真度。

> **译注** "与 Veo 3.1、Wan 2.5 不相上下"是原文的表述;同段给出的全球第 22 名意味着仍有约二十个专有模型排在其前,原文未给出与这些模型的具体 Elo 差距,该说法宜结合排名审慎理解。

6.2.3 音频生成评测(Audio Generation Evaluation)

文本到音视频生成(text-to-audiovisual generation)必须满足两条仅靠视频基准评测无法覆盖的要求:生成的音频应当包含提示词所要求的声音事件,而且这些事件应能归因于正确的视觉声源、并具有合理的时序。因此,我们用 Cosmos-SoundBench——一个专门针对音视频提示词跟随与同步性的基准评测——来评测音频生成。该指标把「语义层面的音画正确性」与「与提示词无关的音频保真度」区分开来:一个样本可能在正确的时刻出现了正确的声音,却仍带有音频伪影;也可能声学上很干净,却漏掉了被要求的声音事件。

Cosmos-SoundBench。Cosmos-SoundBench 包含 144 条评测提示词,取自 FoleyBench (Dixit et al., 2026)。这些提示词覆盖非语音的声音类别,如环境氛围声、撞击声、物体交互声、工具声、载具声、水声及其他环境音效。我们聚焦于非语音音频,因为这类信号对 Physical AI 尤其重要:接触音揭示材质与碰撞属性,工具声指示正在进行的动作,而环境氛围声提供可能只部分可见的场景上下文。

音画质量指标(Audiovisual quality metric,AVQ)。我们用一套结构化的多模态大模型评审(MLLM-as-judge)协议评测每个生成样本。评审模型只在需要相应信息的阶段才看到提示词、视频与音频,使每个评审阶段聚焦于预期的证据,而非无关的跨模态线索。

1. *提示词检查清单构建。*在检查任何生成媒体之前,由 Claude Opus 4.7、Gemini 3.1 Pro 与 GPT 5.5 组成的模型集成先抽取提示词所要求的前景声音、环境音频条件,以及对提示词至关重要的视觉证据。我们用多数投票固定下游打分所使用的检查清单。
2. *不见提示词的视觉观察。*Gemini 3.1 Pro Preview 在看不到提示词的情况下,描述可见实体、动作、场景上下文以及可能的声源。这些观察为声源归因与时间对齐提供证据,而不是作为一个独立的视觉质量分。
3. *语义音画打分。*评审模型评估被要求的声音是否出现、是否与特定声源对应、动态是否恰当,以及是否与可见或合理的动作对齐。这些检查被映射为语义音频正确性(SA)、音画对齐(audio-visual alignment,AVAlign)与提示词关键视觉支持(VisualSupport),并按如下方式组合:

$$\mathrm{SAV} = 0.60\,\mathrm{SA} + 0.30\,\mathrm{AVAlign} + 0.10\,\mathrm{VisualSupport}。$$

该阶段我们运行三次并取平均分,以缓解评审模型的波动性。

我们另外测量 audiobox-aesthetics 的 Production Quality(PQ)作为感知音频质量的代理指标,它试图量化样本级的清晰度与保真度、动态范围、频率带宽与空间化 (Tjandra et al., 2025)。

最终的音画质量分数为:

$$\text{AVQ} = 0.5\,\text{SAV} + 0.5\,\text{AQ}。$$

对每个模型,我们对每条提示词评测五个随机种子的生成结果。对 Cosmos 3 系列模型,我们用 Claude Opus 4.6 将 SoundBench 提示词改写为 §6.3.1 与 §6.3.2 所述的结构化生成格式,以匹配生成器所使用的提示词分布。

表 15 显示,最强的闭源系统在总体 AVQ 上仍保有优势,主要来自更高的感知音频质量。Seedance-1.5-Pro 取得最佳 AVQ(7.64)与 PQ(7.06),Veo-3.1 紧随其后(AVQ 7.45、AQ 6.68)。相比之下,Cosmos 3 在衡量「声音是否匹配视觉事件」的语义与对齐分量上最强:Cosmos3-Nano 取得最佳的 SAV、SA 与 AVAlign 分数,Cosmos3-Super 取得最佳的视觉支持分数。这一模式表明,Cosmos 3 的中期训练(mid-training)能有效地将声音事件锚定在生成的视频中,剩余的提升空间集中在低层的音频保真度上。图 20 展示了 Cosmos3-Nano 生成结果中视觉事件与声学事件强时间对齐的一个例子。

表 15:Cosmos-SoundBench 音画质量。我们报告总体音画质量(AVQ)、语义音画质量(SAV)及其在 §6.2.3 定义的子分数,以及 audiobox-aesthetics 的 Production Quality(PQ)。**加粗**标记每列最佳分数,下划线标记次佳。Seedance-1.5-Pro 凭借更强的 PQ 取得最高 AVQ,而 Cosmos 3 取得最强的语义音画锚定与对齐。

Model	Type	SoundBench Audiovisual Quality (↑)					
		AVQ	SAV	SA	AVAlign	Visual Sup.	PQ
Cosmos3-Super	Open-sourced	7.31	8.34	8.30	8.14	**9.18**	6.28
Cosmos3-Nano	Open-sourced	7.34	**8.35**	**8.33**	**8.16**	9.10	6.32
LTX-2.3	Open-sourced	7.10	7.80	7.86	7.58	8.12	6.39
Seedance-1.5-Pro	Closed-sourced	**7.64**	8.21	8.22	8.06	8.61	**7.06**
Veo-3.1	Closed-sourced	7.45	8.21	8.21	8.01	8.85	6.68
LTX-2.3 Pro	Closed-sourced	7.32	7.93	7.96	7.74	8.35	6.70
Wan2.6	Closed-sourced	7.23	7.90	7.99	7.54	8.45	6.55
Sora 2	Closed-sourced	6.90	7.94	7.97	7.70	8.49	5.85

> **译注** 正文公式写作 AVQ = 0.5 SAV + 0.5 AQ,但表格与叙述中该音频质量分记为 PQ(audiobox-aesthetics Production Quality),AQ 与 PQ 指同一分数,原文符号未统一。另外注意:开源的 Cosmos 3 只在语义/对齐子分数上领先,总分 AVQ 仍落后于 Seedance-1.5-Pro 与 Veo-3.1,差距主要在底层音频保真度(PQ)。

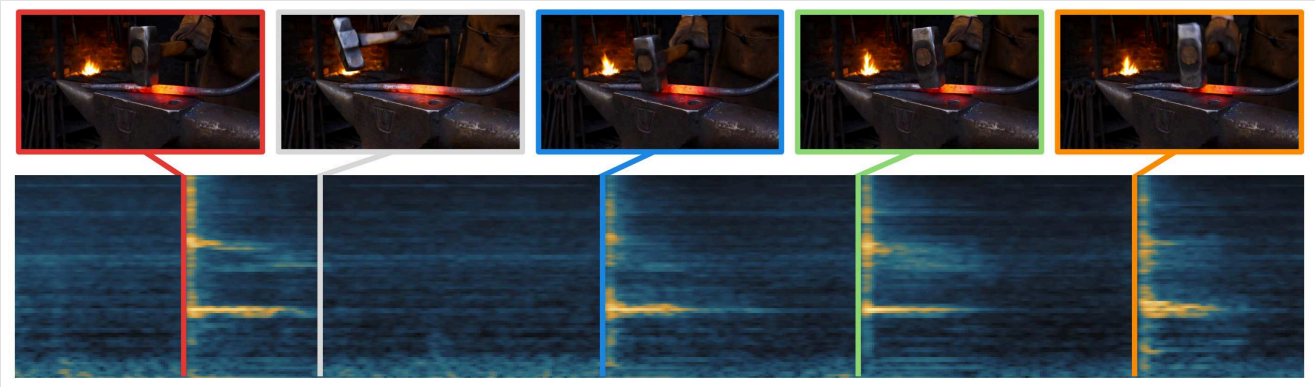


图 20:音频-视频事件对齐。从一段 Cosmos3-Nano 生成结果中选取的帧,与生成音频的频谱图配对展示。彩色边框的帧表示锤击时刻,其时间标记与频谱中的尖锐瞬态相吻合;灰色边框的帧展示两次锤击之间的非接触时刻,此处没有观察到可比的声学瞬态。这一对比提供了定性证据:声学瞬态与视觉上的撞击时刻对齐,而非与其间的运动过程对齐。

6.2.4 迁移生成评测(Transfer Generation Evaluation)

双权重无分类器引导。视频迁移(video transfer)同时以结构化文本提示词和一段控制视频为条件,而「忠实于文本描述」与「贴合结构控制」之间的最优平衡点因控制模态而异。为了独立控制这两个因素,我们采用带有两个独立权重的双权重无分类器引导(classifier-free guidance,CFG)方案:一个权重作用于控制视频,另一个作用于文本提示词。

在每个去噪步,我们对去噪器做三次求值——同时给定两种条件;只给提示词(丢弃控制视频);以及保留控制视频、但把提示词替换为一条固定的负面描述。然后组合这些预测,使控制权重从「仅提示词」的预测向「完整条件」的预测外推(增强结构控制),而文本权重则从「负面提示词」的预测向外外推(增强对文本描述的忠实度)。把这两个权重作为独立旋钮暴露出来,使我们能够针对每种模态分别调到各自的质量-保真度最佳点;我们发现这种因子化引导比标准的单一引导公式更有效。

表 16:**PAIBench-C 单控制结果**。每次生成以四种控制模态之一(深度、分割、模糊或边缘)为条件,并用对应的基于真值的指标打分:**Depth si-RMSE**,预测深度与参考深度之间的尺度不变均方根误差;**Seg. mIoU**,预测与参考分割掩码之间的平均交并比;**Blur SSIM**,预测与参考模糊图之间的结构相似度;**Edge F1**,预测与参考边缘图之间的 F1 分数。**DOVER** 是一个无参考的感知视频质量分数,对四种逐模态生成取平均。Depth si-RMSE 越低越好;DOVER、Seg. mIoU、Blur SSIM 与 Edge F1 越高越好。**加粗**表示每列最佳结果。

Model	DOVER ↑	Seg. mIoU ↑	Blur SSIM ↑	Edge F1 ↑	Depth si-RMSE ↓
Cosmos3-Super	10.14	0.71	**0.91**	**0.50**	**0.58**
Cosmos3-Nano	**10.39**	**0.72**	**0.91**	0.49	0.62
Cosmos-Transfer2.5	9.49	0.68	0.90	0.45	0.68

通用视频迁移。我们在 PAIBench-C (NVIDIA, 2025c) 上评测视频迁移。这是一个控制条件下的视频到视频基准评测,覆盖四种空间控制模态:模糊、边缘、分割与深度。该基准评测包含 600 段视频片段,横跨三个 Physical AI 领域:200 段来自 AgiBot World (Bu et al., 2025) 的机械臂操作片段、200 段来自 OpenDV (Yang et al., 2024) 的驾驶片段,以及 200 段来自 Ego-Exo-4D (Grauman et al., 2024) 的第一人称日常生活片段。

对每个样例,模型接收一条文本提示词与一段控制视频,须生成一段既跟随控制信号、又保持提示词场景语义的照片级真实视频。我们报告单控制设定(每个样例恰好一种模态)下的结果,使每种模态的贡献可以被单独衡量。质量评估的方法是:从生成视频与参考视频中重新提取相应的控制信号,并在对应空间中比较(双边滤波后的模糊 SSIM、Canny 边缘 F1、估计深度上的尺度不变 RMSE、开放词表分割掩码上的 mIoU),外加 DOVER (Wu et al., 2023a) 作为与内容无关的感知真实感度量。

我们将 Cosmos 3 与 Cosmos-Transfer2.5 基线进行比较,后者对不同模态各用一条专用的 ControlNet (Zhang et al., 2023) 分支来处理。注意 Cosmos 3 则是一个统一模型:它将控制视频与文本提示词一起放进输入序列中消费,原生支持任意模态组合,无需逐模态的 ControlNet 适配器。

表 16 显示,尽管把四条专用 ControlNet 分支收拢进了单一的统一骨干网络,Cosmos 3 仍通过其两个变体之一,在每种模态上追平或超越 Cosmos-Transfer2.5:Cosmos3-Nano 在感知质量(DOVER)与分割上领先,而 Cosmos3-Super 拿下几何要求更高的边缘与深度任务。三个模型在模糊 SSIM 上基本持平,该指标已接近其上限而趋于饱和。这些结果表明,逐模态的 ControlNet 适配器并非获得强控制保真度的前提——单一统一骨干即可原生支持全部四种空间控制,并在每项指标上以其两个规模之一击败专用适配器基线。

自动驾驶。作为 PAIBench-C 的补充,还有一个由 486 段单视角驾驶片段构成的自动驾驶(AV)专用基准评测 (NVIDIA, 2025b),用于以世界场景地图(world-scenario map)为条件的视频迁移,称为 AVBench-C。其控制输入是驾驶场景的相机视角渲染,融合了静态地图结构(车道线、道路边界、交通信号灯与交通标志)与全部场景物体(车辆、行人及其他交通参与者),包括静止与运动的。给定这份渲染的场景与一条文本提示词,模型须生成与两种输入都一致的照片级真实驾驶视频。我们用自动评测与人工评测两种方式评估结果:

- *AVBench-C 自动评测*。我们用三个基于真值的自动检查器评测每个生成结果。自车运动(egomotion)检查器用视觉里程计计算每米行驶距离的轨迹漂移;物体对应(object-correspondence)检查器将场景实体与控制输入匹配,把动态交通参与者与渲染轨迹比较、把静态结构与投影地图比较;环境(environment)检查器用一个 VLM 评审模型为提示词级的驾驶条件打分,包括天气、时段、地区与路面状态。
- *AVBench-C 人工评测*。受训标注员沿两个维度对每段生成视频按 1–3 分打分。视频质量维度衡量总体真实感:3 分表示纹理真实且交通参与者行为在时间上连贯,2 分表示轻微形变或短暂的时间卡顿,1 分表示明显不真实的动态或低质量纹理。车道线维度衡量对世界场景地图的忠实度:标注员将生成结果与期望道路布局的参考叠加图并排查看,若车道、路口与实体大多位置正确记 3 分,若少数车道缺失但多数实体位置正确记 2 分,若多数车道或路口缺失、或实体位置错误记 1 分。

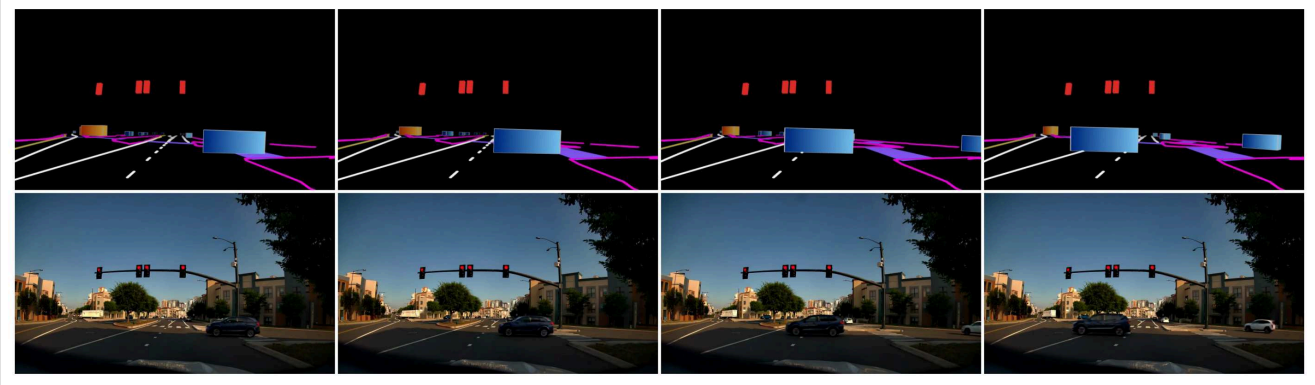


图 21:**驾驶场景视频迁移结果**。**Cosmos3-Nano** 从对应的 720p 控制视频(上排)生成视频帧(下排)。控制视频编码了高精地图元素——车道、路面标线、杆件与交通信号灯(带或不带灯态)——它们共同表达复杂的道路拓扑(包括高架桥),以及以长方体(cuboid)表示的交通参与者。每个长方体按粗粒度类别本体(如卡车、车辆、行人)着色,并通过明暗区分前后朝向。

表 17:**AVBench-C 评测分数**。我们同时报告自动与人工评测结果。自动分数由三个基于真值的检查器计算:**Ego drift**(越低越好),视觉里程计轨迹漂移;**Dyn. Obj.** 与 **Static Obj.**(越高越好),场景实体与控制输入的匹配程度;**Environment**(越高越好),提示词级驾驶条件的 VLM 评审分数。人工分数是受训标注员沿两个维度给出的 1–3 分:**Video quality** 衡量总体真实感(纹理、交通参与者行为与时间一致性),**Lane line** 衡量对世界场景地图的忠实度(车道、路口与实体的位置正确性)。两项人工指标均为越高越好。**加粗**表示每列最佳结果。

Model	Automatic evaluation				Human evaluation	
	Ego drift ↓	Dyn. Obj. ↑	Static Obj. ↑	Environment ↑	Video quality ↑	Lane line ↑
Cosmos3-Super	**0.003**	0.64	0.41	**0.90**	**2.86**	2.45
Cosmos3-Nano	**0.003**	0.67	0.41	**0.90**	2.82	**2.50**
Cosmos-Transfer2.5-AV-Singleview	0.008	0.62	**0.42**	**0.90**	2.59	2.47

> **译注** 自动评测中 Static Obj. 一列的最优其实是基线 Cosmos-Transfer2.5(0.42 vs 0.41);人工评测的 Lane line 上 Cosmos3-Super(2.45)也略低于基线(2.47),仅 Nano(2.50)领先。原文以「基本持平/差距在 ±0.05 内」带过,Cosmos 3 的确定性优势主要体现在 Video quality 与 Dyn. Obj. 上。另外原表 Dyn. Obj. 列仅加粗了 0.67,故 0.67 为该列唯一最优标记。

表 17 显示,Cosmos 3 通过其两个变体之一,在每项驾驶场景指标上追平或超越 Cosmos-Transfer2.5。自动评测方面,三个模型的自车轨迹漂移均一致地很小,Cosmos3-Nano 还在动态物体对应上领先,而静态结构与环境一致性基本持平、且已接近指标上限而饱和。人工评测强化了这一图景:Cosmos3-Super 与 Cosmos3-Nano 给出明显更高的视频质量(2.86 与 2.82,对比 2.59),而车道线忠实度在三个模型间相当(差距在 ±0.05 之内)。综合来看,这些结果表明 Cosmos 3 在保持几何与结构保真度的同时,给出肉眼可见的更高质量驾驶场景生成,并在每项指标上以其两个规模之一击败基线。图 21 展示了 Cosmos3-Nano 依据其控制输入生成驾驶场景视频的定性示例。

6.2.5 动作生成评测(Action Generation Evaluation)

我们评测动作中期训练是否赋予了 Cosmos 3 一个跨领域、跨推理模式可复用的世界-动作先验(world-action prior)。核心问题是:统一的动作中期训练能否加速对特定领域动作接口的适配;以及一个较短的后训练(post-training)阶段,能否把共享的基础模型变成与最先进的领域专用基线相当甚至更强的专用模型。表 18 报告了相机运动、自动驾驶、机器人与第一人称运动领域上的结果,覆盖前向动力学(forward dynamics)与逆向动力学(inverse dynamics)两种设定。表 19 报告了机器人策略(policy)设定的结果,评测 Cosmos 3 完成语言指定任务的能力。

表 18:跨领域前向与逆向动力学的后训练对比。FD 表示前向动力学,ID 表示逆向动力学。`PT-init` 从通用的 Cosmos 3 预训练检查点初始化,`MT-init` 从中期训练检查点初始化。领域专用基线只在其对应的应用上报告结果,而 Cosmos 3 各变体在所有被评测的动作设定上进行比较。**加粗**表示每列最佳,下划线表示次佳。

Model	Autonomous Vehicle (ID)			Camera Motion (FD)			Egocentric Motion (FD)	Robotics (FD)
	RRE (°, ↓)	RTE (m, ↓)	ATE (m, ↓)	RRE (°, ↓)	RTE (m, ↓)	ATE (m, ↓)	PSNR (↑)	PSNR (↑)
Cosmos3-Super (`MT-init`)	0.232	**0.014**	**0.90**	**0.142**	**0.026**	**0.99**	**16.19**	**26.04**
Cosmos3-Nano (`MT-init`)	**0.211**	**0.014**	0.98	0.147	0.029	1.24	16.12	25.52
Cosmos3-Super (`PT-init`)	0.284	0.018	1.32	0.293	0.036	1.82	15.34	22.69
Cosmos3-Nano (`PT-init`)	0.249	0.017	1.20	0.172	0.034	1.61	15.22	23.24
Lingbot-World	–	–	–	0.299	0.057	2.88	–	–
HY-World1.5	–	–	–	0.377	0.042	1.39	–	–
VGGT	0.596	0.768	23.46	–	–	–	–	–
DepthAnything3	0.312	0.354	9.29	–	–	–	–	–
LOME	–	–	–	–	–	–	9.36	–
Ctrl-World	–	–	–	–	–	–	–	22.99

设置。我们对每个下游领域比较两种初始化协议。第一种,预训练初始化(`PT-init`),从 Cosmos 3 预训练检查点出发,该检查点未在动作领域数据上训练过。第二种,中期训练初始化(`MT-init`),从我们的中期训练检查点出发,该检查点见过横跨多个领域与预测模式的动作数据,包括前向动力学(FD)、逆向动力学(ID)与策略(policy)。每组对比使用相同的训练配方,并固定模型规模、数据与计算预算。我们同时报告 Cosmos3-Nano 与 Cosmos3-Super 的结果。

指标。我们为每种动作接口报告相应指标。对自动驾驶逆向动力学,相对旋转误差(RRE)与相对平移误差(RTE)衡量帧间位姿精度,绝对轨迹误差(ATE)衡量全局轨迹一致性。我们也用这些指标衡量相机运动前向动力学中的相机跟随精度,即比较真值相机位姿与从生成视频中估计的相机轨迹。对机器人与第一人称前向动力学,PSNR 衡量以动作为条件的未来观测的重建质量。虽然一段合理的生成推演(rollout)不必与唯一记录下来的真值未来逐像素一致,但我们发现,在相对较短的时间跨度下,PSNR 是时间对齐、运动一致性与重建保真度的一个有用代理指标。对机器人策略,成功率(success rate)衡量成功完成的评测任务所占的百分比。

自动驾驶(ID)。我们用一个内部驾驶数据集评测 Cosmos 3 的逆向动力学能力,该数据集由 6 秒的视频片段组成,带有 10 FPS 的精确自车轨迹。在此设定下,模型直接从视频输入预测自车轨迹。我们与两个最先进的通用领域基线比较:VGGT (Wang et al., 2025b) 与 DepthAnything3 (Lin et al., 2025b)。如表 18 所示,以 `PT-init` 初始化的 Cosmos3-Nano 与 Cosmos3-Super 均超过这两个基线;采用 `MT-init` 则为 Cosmos3-Nano 带来进一步提升。值得注意的是,我们的方法在专门的驾驶数据上训练后,取得了好得多的公制尺度平移估计,而通用领域基线则受漂移误差困扰。总体而言,这些结果表明,面向 Physical AI 的视频模型只需极少的适配就能在特定的具身应用中表现出色。定性结果见图 22。

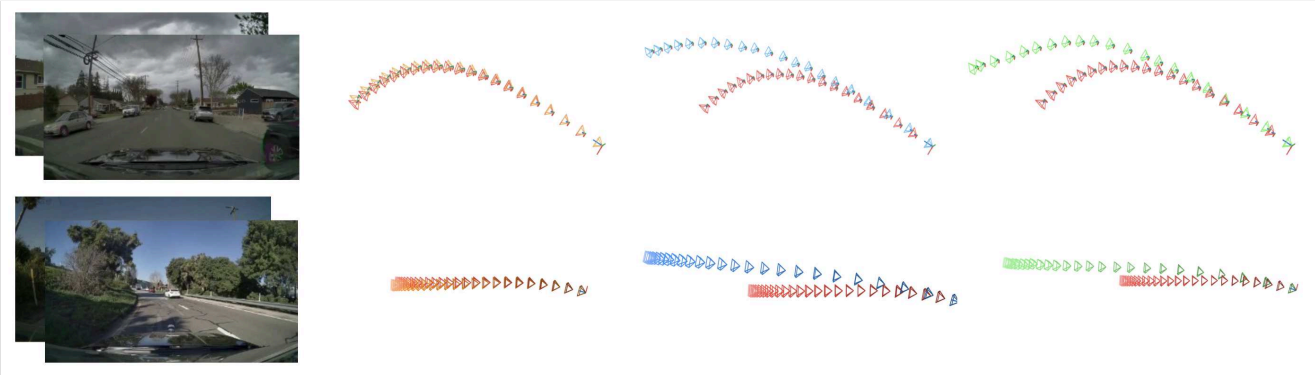


图 22:自动驾驶逆向动力学对比。我们定性比较不同方法从输入视频估计的自车轨迹,红色轨迹为真值。Cosmos3-Nano(`MT-init`)展现出估计准确的公制尺度自车位姿的能力。

相机运动(FD)。以相机为条件的视频生成可为许多应用充当世界模拟。我们让模型在给定初始图像与指定相机轨迹的情况下预测未来帧。为量化相机跟随精度,我们用 DepthAnything3 (Lin et al., 2025b) 从生成视频中估计公制尺度的相机轨迹,并与条件输入比较:若生成视频与条件轨迹高度一致,则从预测序列估计出的位姿应与输入位姿高度吻合。我们在一个内部数据集上与 Lingbot-World (Robbyant et al., 2026) 和 HY-World 1.5 (HunyuanWorld, 2025) 的双向模型对标,该数据集包含一百段 5 秒的真实视频片段,其相机运动由 DepthAnything3 (Lin et al., 2025b) 估计得到。定性结果见图 23。结果显示,采用 `PT-init` 的 Cosmos3-Nano 与 Cosmos3-Super 已能提供稳健的相机控制;`MT-init` 进一步改进结果:Cosmos3-Super 达到 0.142° RRE、0.026 m RTE 与 0.99 m ATE,在全部三项指标上超过 Lingbot-World(0.299° RRE、0.057 m RTE、2.88 m ATE)与 HY-World 1.5(0.377° RRE、0.042 m RTE、1.39 m ATE)。

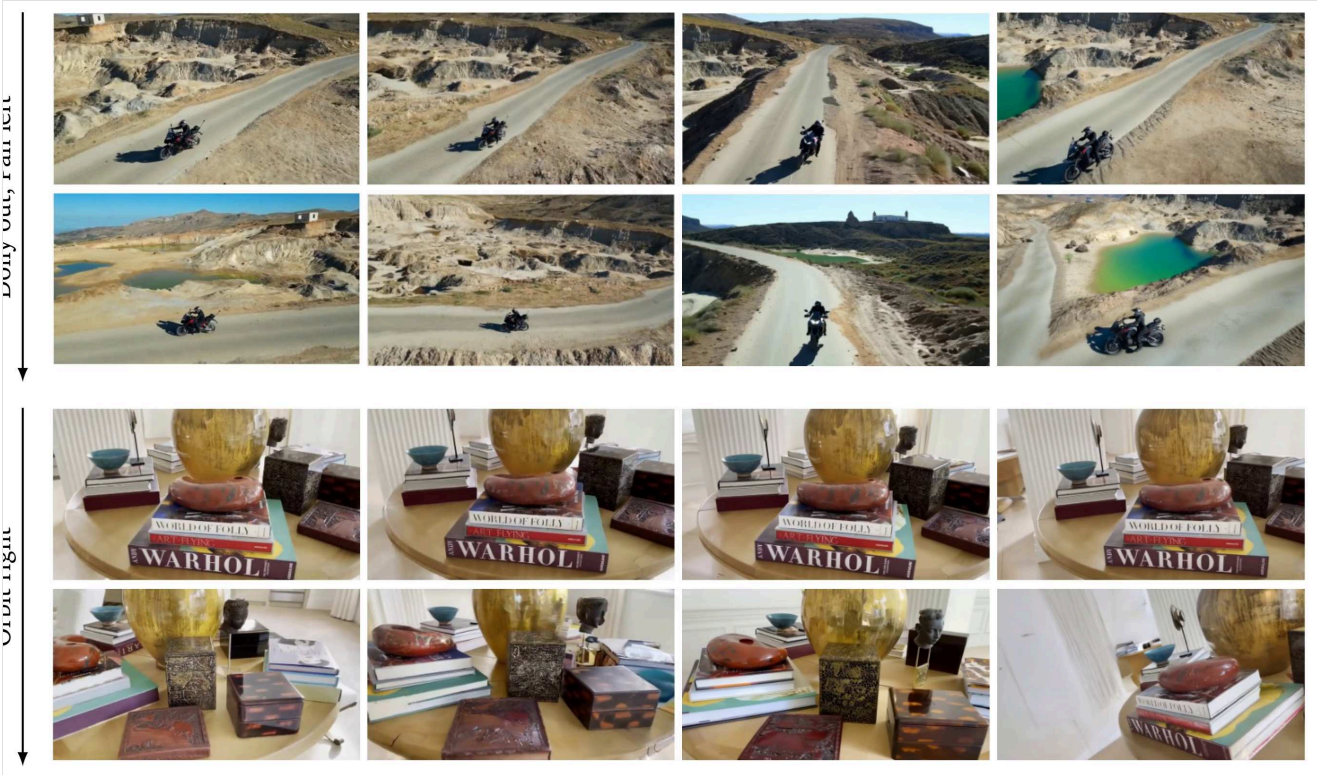


图 23:相机前向动力学对比。给定复杂的真实轨迹,Cosmos3-Nano(`MT-init`)在生成视频中忠实复现了相同的相机运动。对每个运动示例,第一行展示序列开始附近的帧,第二行展示序列结束附近的帧。向下箭头表示从头到尾的时间推进,旁边的文字标明所指令的相机运动。

第一人称运动(FD)。我们在前向动力学设定下评测 Human World Bench(HWB,见 §6.2.2)中的视频。我们使用 HWB 的动作标注——同时包含相机自运动与手部运动跟踪——并报告生成视频的 PSNR。我们把 `PT-init` 与 `MT-init` 两种检查点都适配到第一人称领域,并与 LOME (Gao et al., 2026a) 比较,后者是一个近期的、专攻第一人称手部操作数据的动作条件视频生成方法。尽管 LOME 是该设定下最接近的可用基线之一,它在 HWB 上表现不佳,PSNR 仅为 9.36 dB,很可能源于其训练数据与 HWB 基准评测之间的分布偏移。相比之下,Cosmos 3 在两种模型规模与两种初始化协议下都取得了显著更高的 PSNR:采用 `PT-init` 时,Cosmos3-Nano 达到 15.22 dB,Cosmos3-Super 达到 15.34 dB;从 `MT-init` 出发进一步把分数提升到 Cosmos3-Nano 16.12 dB、Cosmos3-Super 16.19 dB。这些结果显示 `MT-init` 带来一致的增益,表明统一的动作中期训练为第一人称前向动力学提供了明显更强的起点。总体而言,这些结果支持统一动作中期训练的核心假设:跨机器人具身形态、相机运动、自动驾驶运动与第一人称运动的联合训练,会诱导出一个可迁移的动作领域先验,使第一人称领域的下游适配收敛更快、效果更强。

机器人(FD)。机器人前向动力学解锁了「用视频模型做策略评测」等应用。我们在 DROID 数据集 (Khazatsky et al., 2024) 上开展实验,这是一个大规模真实机器人操作数据集,场景与物体多样。给定初始帧与长度为 16 的机器人末端执行器动作块(action chunk),模型预测随后的 16 帧。我们把 `PT-init` 与 `MT-init` 检查点都在 DROID 数据集上做后训练,并以 Ctrl-World (Guo et al., 2026) 为主要对比基线。定性结果见图 24,定量结果见表 18。如表 18 所示,从预训练检查点做后训练,在 Nano 规模上就已追平基线,PSNR 达到 23.24 dB,而 Ctrl-World 为 22.99 dB;从中期训练检查点做后训练则明显更强,Cosmos3-Nano 达到 25.52 dB、Cosmos3-Super 达到 26.04 dB。这些结果表明 Cosmos 3 是一个强大而灵活的前向动力学预测骨干。

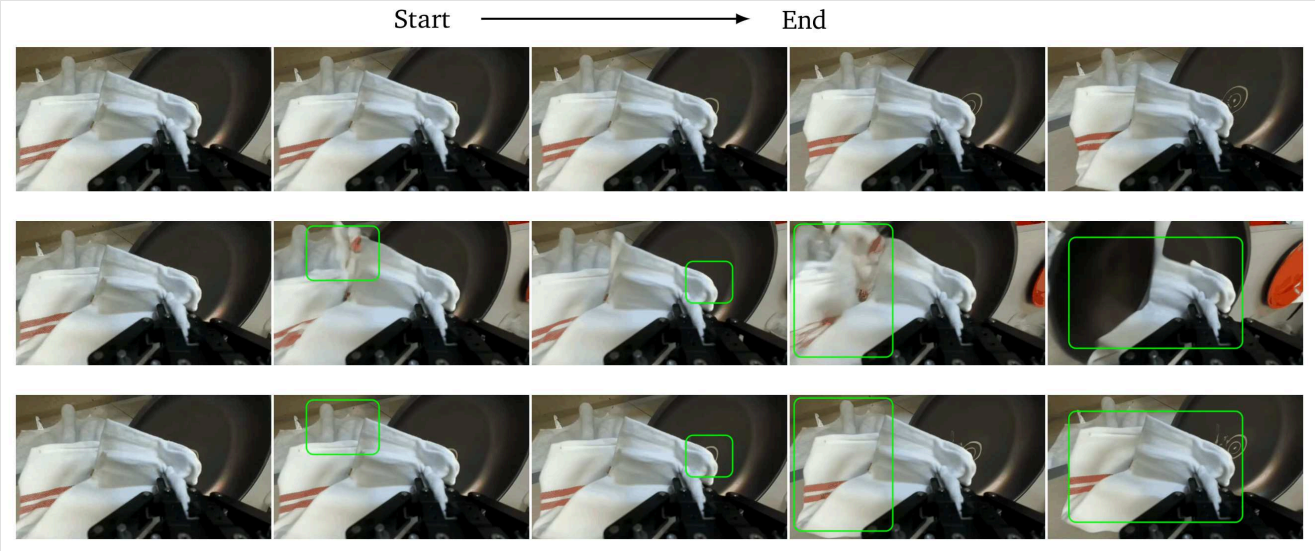


图 24:机器人前向动力学的定性对比。生成的帧紧密跟随动作指令,且机械臂与织物之间的交互比基线更真实。绿色方框标出基线输出中出现可见形变或伪影的区域,以及 Cosmos3-Nano(`MT-init`)对应的、不存在这些伪影的区域。

机器人操作(policy)。为评测策略模式,我们对 `Cosmos3-Nano-Policy-DROID` 做后训练,并将其提交到 RoboLab 仿真基准评测 (Yang et al., 2026a)、RoboArena 真实世界基准评测 (Atreya et al., 2025) 与 MolmoSpaces 仿真基准评测 (Kim et al., 2026)。如表 19、图 26 与图 27 所示,我们的模型在全部三个基准评测上取得新的最先进结果,证明 Cosmos3 全模态世界模型可以被有效地后训练为强大的机器人操作策略。

RoboLab (Yang et al., 2026a) 是一个高保真、与机器人和策略无关的仿真基准评测,包含 120 个语言条件任务,旨在从视觉、关系与流程三类能力维度评测任务通用型机器人操作策略。RoboLab 在三种指令具体程度下评测每个任务:模糊提示词(vague)、默认提示词(default)与具体提示词(specific)。这一划分检验的是对语言措辞的鲁棒性,而非单一的规范指令。在 RoboLab 上,我们的后训练策略以明显优势超过所有先前模型的任务成功率(表 19),包括 $\pi_{0.5}$ (Physical Intelligence Team, 2025) 等强大的视觉-语言-动作模型(vision-language-action,VLA)与 DreamZero (Ye et al., 2026) 等世界-动作模型(world-action model,WAM),且在所有任务指令粒度与任务难度级别上均是如此。例如在具体任务指令下,我们的策略在 120 个任务、每任务 10 次推演的设定下取得 39.7% 的平均成功率,超过 $\pi_{0.5}$ 的 28.1% 与 DreamZero 的 25.2%。此外,与直接从预训练检查点做后训练的模型 Cosmos3-Nano(`PT-init`)相比,我们的最终策略表现更好,证明了把多来源动作模态数据纳入中期训练的有效性。

表 19:Cosmos3-Nano-Policy-DROID 在 RoboLab 上创造新的最先进水平。在 RoboLab-120 上按语言具体程度与任务难度级别报告任务成功率(%)。所有策略均使用在 DROID 数据集上微调过的现成检查点。**加粗**表示每列最佳。

Model	Overall			Simple			Moderate			Complex		
	Vague	Default	Specific	Vague	Default	Specific	Vague	Default	Specific	Vague	Default	Specific
Cosmos3-Nano-Policy-DROID	**20.6**	**36.8**	**39.7**	**23.3**	**40.6**	**42.0**	**23.3**	**35.4**	**40.3**	4.1	**25.3**	**29.4**
Cosmos3-Nano (`PT-init`)	16.7	28.1	30.2	17.8	30.3	32.8	19.0	28.7	29.5	**7.1**	18.2	21.8
$\pi_{0.5}$	15.2	28.0	28.1	16.2	29.7	29.8	17.9	31.5	31.0	5.3	13.5	14.7
DreamZero	14.9	25.7	23.9	15.0	26.1	25.8	19.5	30.0	26.7	4.1	14.1	10.6
π_0-FAST	9.2	15.5	14.9	9.5	20.2	19.4	12.8	13.3	12.6	0.0	2.9	3.5
paligemma-binning	3.1	3.4	5.5	2.2	3.4	4.1	5.9	4.9	10.3	0.0	0.0	0.0
GR00T N1.6	5.4	7.2	5.3	7.2	8.8	7.5	4.9	7.9	4.1	0.0	0.0	0.0
π_0	2.8	5.0	3.5	2.8	7.2	5.3	3.8	3.6	2.1	0.0	0.0	0.0

> **译注** 两处与正文「全面领先」表述有出入。其一,Complex×Vague 一格中,最终策略仅 4.1%,反而低于 `PT-init` 变体的 7.1%(该列加粗最优)与 $\pi_{0.5}$ 的 5.3%,即「跨所有指令粒度与难度全面超越」并不严格成立。其二,正文称 DreamZero 在具体指令下为 25.2%,但表中 Overall-Specific 一列 DreamZero 为 23.9(25.7 是其 Default 分),原文数字不自洽。

RoboArena (Atreya et al., 2025) 是一个分布式的真实世界基准评测,通过众包的成对比较来评测并排名通用机器人策略,覆盖多样的任务与真实环境。任何拥有 DROID 平台的人都可以在任意环境、任意任务上,通过双盲 A/B 比较评测一对策略,最终分数由这些成对偏好聚合而成,给出每个策略的总体评分。截至 2026 年 5 月 30 日下午 2:40,我们的机器人操作策略位居排行榜首位(图 26),超过许多先前的强策略模型。我们期待收到更多来自社区的评测,以进一步验证我们的策略。在若干已测试的任务中,该策略能可靠地完成任务并良好地遵循语言指令:既能执行简单的抓取-放置任务,也能完成需要多个连续步骤的长时程任务。我们观察到它对多种未见过的物体与任务有很强的泛化能力;该策略还能容忍失败,常常在必要时重试,并对执行过程中的人为干预保持鲁棒。部分定性结果见图 25。

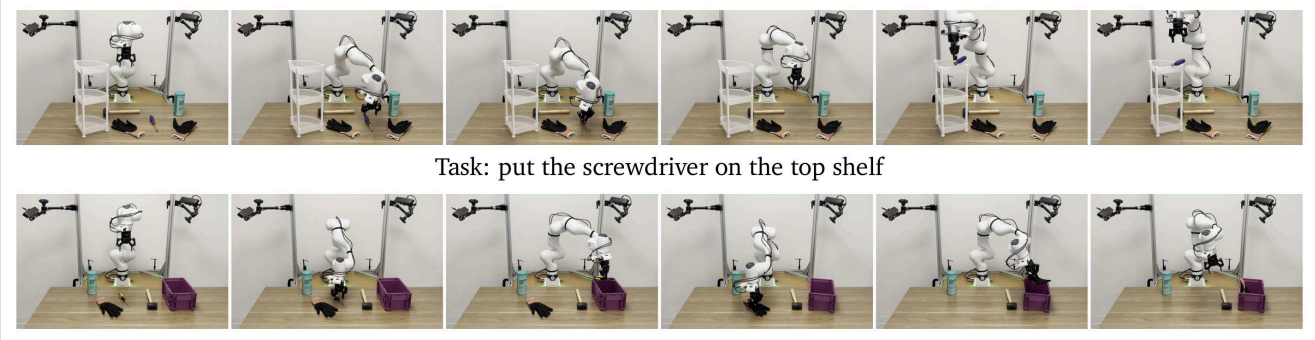


图 25:Cosmos3-Nano-Policy-DROID 的真实世界评测。我们展示实体机器人推演的快照帧。这些例子表明我们的策略已成功部署到实体机器人上,完成语言条件下的操作任务。

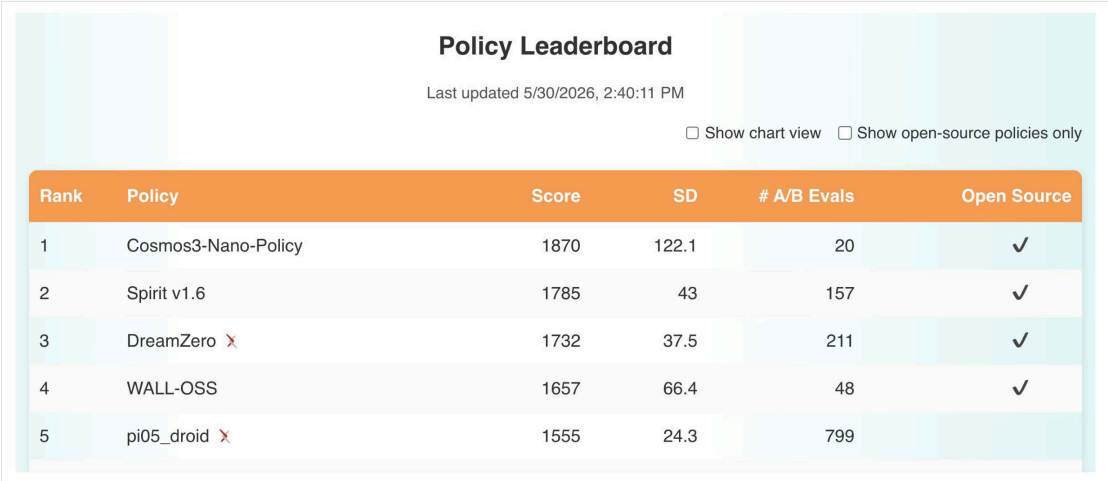


图 26:Cosmos3-Nano-Policy-DROID 曾占据 RoboArena 榜首。Cosmos3-Nano-Policy-DROID 在 RoboArena 真实世界基准评测排行榜上排名第一(日期:2026-05-30)。

> **译注** 图 26 的榜单截图显示,Cosmos3-Nano-Policy 当时仅累计 20 次 A/B 评测,远少于其后各策略(157、211、799 次),且分数标准差(122.1)明显偏大;原文亦承认「期待更多社区评测来进一步验证」。该榜首名次为特定时点快照,统计置信度有限。

MolmoSpaces (Kim et al., 2026) 是一个用于评测通用策略的仿真基准评测,侧重在系统性、受控变化下的泛化能力。我们聚焦其中的操作任务,它由两个任务集合构成:(1)MolmoSpaces Combined 与(2)MolmoBot Combined。MolmoSpaces Combined 集合包含四个基准任务(`Pick-v1`、`Pick & Place-v1`、`Open-v1` 与 `Close-v1`),MolmoBot Combined 集合包含七个基准任务(`Pick-v1.5`、`Pick-v2-classic`、`Pick-v2-filament`、`Pick-v2-RandCam`、`Pick & Place-v2`、`Pick & Place-NextTo-v2` 与 `Pick & Place-Color-v2`)。截至 2026 年 6 月 20 日,Cosmos3-Nano-Policy-DROID 在 All Combined 设定下位列排行榜第一,取得 39.0% 的 oracle 成功率(图 27),超过了 WALL-OSS-0.5 (Zhai et al., 2025) 与 TiPToP (Shen et al., 2026b) 等多个先前的强策略模型。值得注意的是,我们提交给 RoboLab 与 RoboArena 评测的是完全相同的模型与超参数,没有为 MolmoSpaces 做任何基准专属调优。

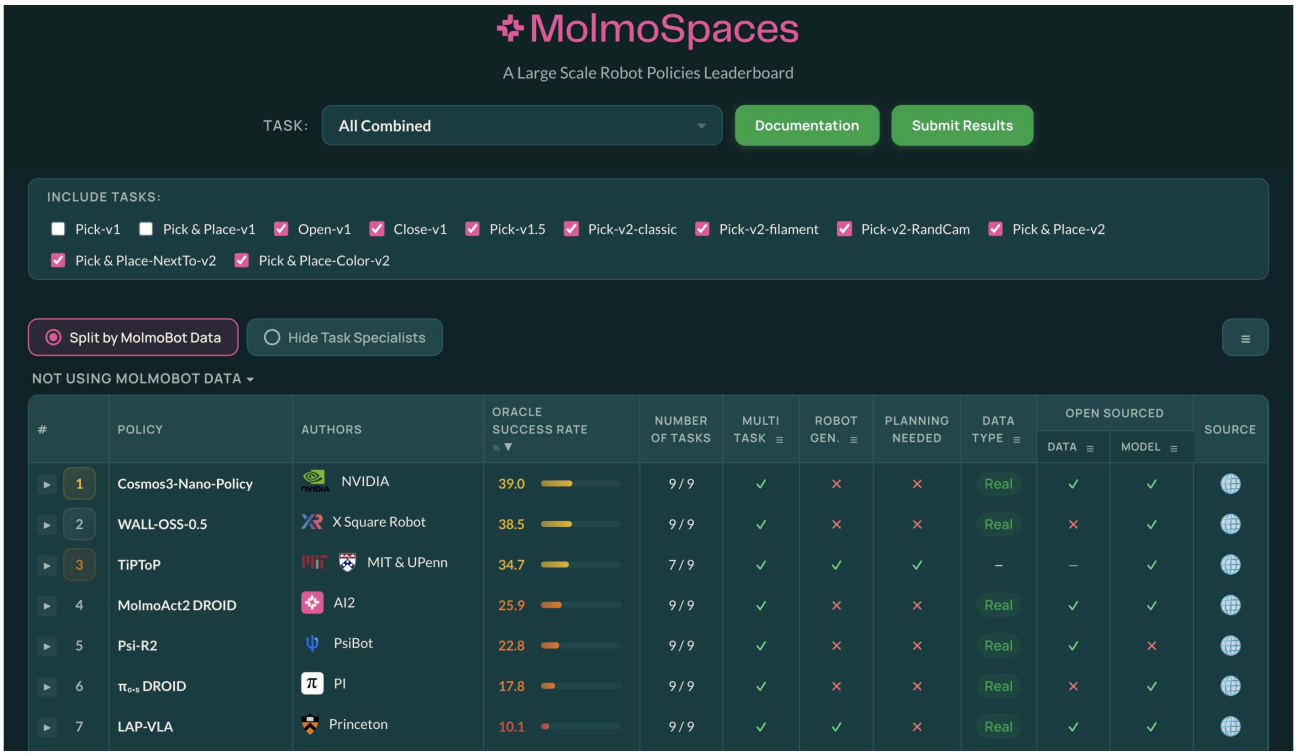


图 27:Cosmos3-Nano-Policy-DROID 在 MolmoSpaces 上排名第一。Cosmos3-Nano-Policy-DROID 是 MolmoSpaces 仿真基准评测排行榜上的最高排名模型(日期:2026-06-20)。

对新具身形态的适配。我们评测 `MT-init` 是否能让 Cosmos3-Nano 更快地适配未见过的具身形态(embodiment)与环境。我们使用 LIBERO-10 环境 (Liu et al., 2023a),观测为第三人称相机与腕部相机的多视角图像。我们在后训练迭代过程中跟踪进度,每个验证任务运行 50 次试验,即每个检查点 500 次推演。

表 20:借助 LIBERO-10 快速适配新具身形态。Cosmos3-Nano 分别从 `MT-init` 与 `PT-init` 出发的闭环成功率,每个检查点以 500 次推演评测。结果显示 `MT-init` 比 `PT-init` 适配更快。

Model	Iteration			
	500	1000	1500	2000
Cosmos3-Nano (`MT-init`)	24.6%	91.4%	95.8%	**97.4%**
Cosmos3-Nano (`PT-init`)	0.0%	73.8%	93.4%	95.2%

如表 20 所示,后训练能迅速提升两种初始化下的闭环操作成功率,但 `MT-init` 在后训练早期具有明显优势:在 500 次迭代时,Cosmos3-Nano `MT-init` 已达到 24.6% 的成功率,而 `PT-init` 仍停留在 0.0%;到 2000 次迭代时,`MT-init` 可达 97.4% 的成功率。这些结果表明 `MT-init` 能更快地适配新的具身形态与环境。

动作数据协同效应。选择哪些动作领域一起训练,是一个实际的混合配比设计问题:新增的领域可能提供有用的视觉-动作先验,也可能稀释对目标领域的监督。受近期语言领域迁移工作 (Longpre et al., 2026) 的启发——该工作度量在一种语言上训练对另一种语言是有益还是有干扰——我们为动作中期训练构建了类似的迁移图谱,覆盖自运动领域(Camera Motion、Autonomous Vehicle)、机器人操作领域(Google Robot、WidowX-250、Franka Panda Single、Franka Panda Dual、AgiBot)与人类第一人称运动(Egocentric)。定量的协同矩阵总结在图 28 与图 29 中。FD 报告 PSNR,ID 报告 MSE;对策略模式,由于推演分布是多峰的,我们报告 4 次推演中的最小 PSNR 与 MSE。在每个协同矩阵中,非对角格子使用行、列两个领域按 50/50 混合训练 4,000 次迭代,而对角格子是对应的单领域基线在 2,000 次迭代时的结果。这一设置使单领域运行与配对运行之间的逐领域训练预算相匹配。每一行固定评测领域,各列变化单领域或配对训练设置。正的增量表示跨领域迁移,负的增量表示相互干扰。

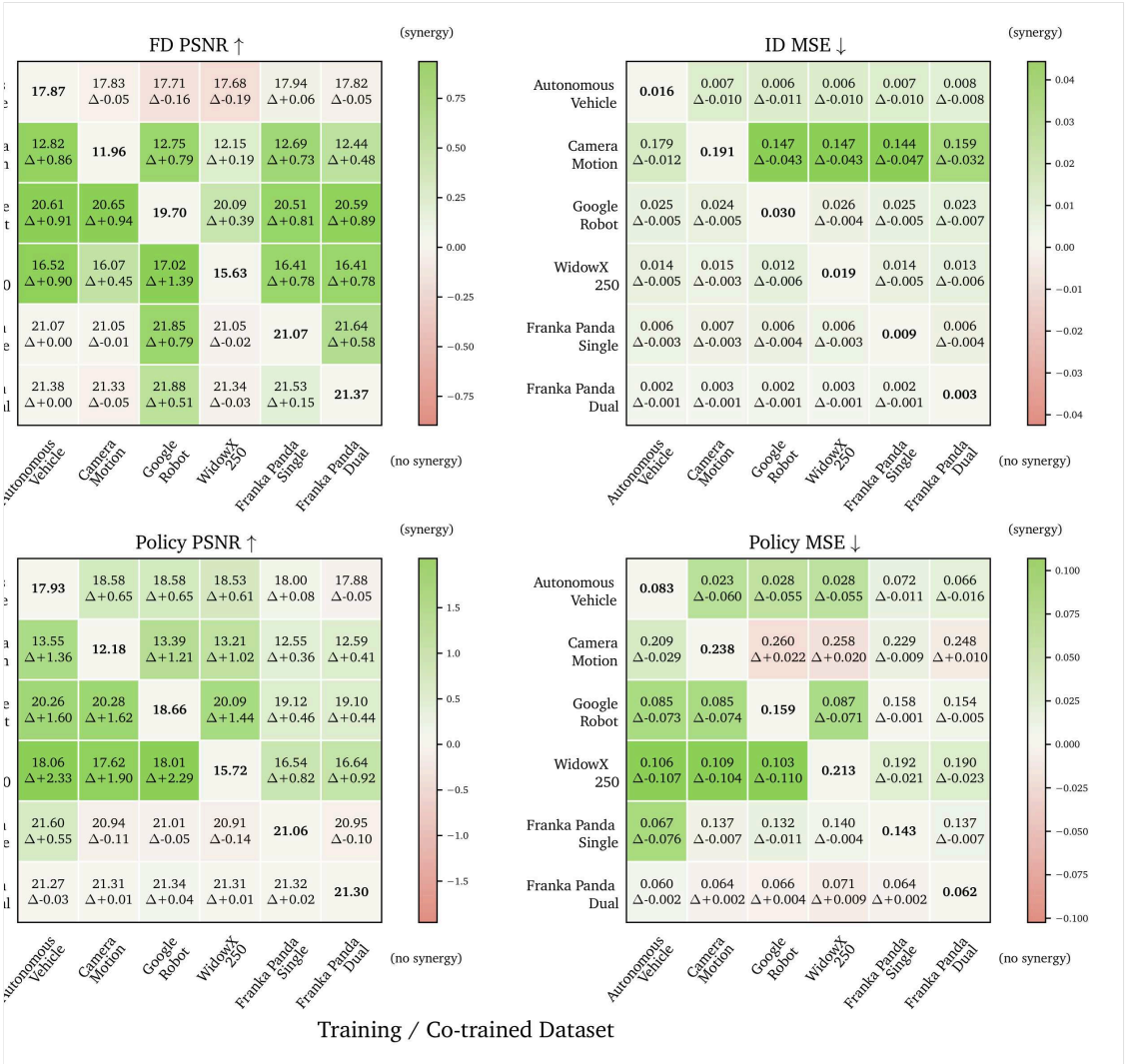


图 28:跨自运动与机器人操作领域的协同效应研究。行表示评测领域,列表示所加入的联合训练领域。对角格子是单领域基线,非对角格子使用行-列 50/50 混合、并匹配行领域的训练暴露量。每个格子报告分数及其相对行对角线的增量。绿色表示正迁移,符号方向已调整,使更高的 PSNR(↑)与更低的 MSE(↓)均为更好。

我们把主要发现总结如下:

- *相机运动受益于广泛的动作联合训练。*图 28 显示,相机运动作为表现较弱的领域之一,能从多个联合训练伙伴中获益,而不仅仅是另一个自运动数据源。AV 数据把相机 FD PSNR 从 11.96 提升到 12.82(+0.86),机器人领域同样带来正向的 FD 增益,包括 Google Robot(+0.79)、Franka Panda Single(+0.73)与 Franka Panda Dual(+0.48)。该行中每个展示的联合训练伙伴也都改善了 ID MSE。这说明相机运动的学习受益于一般性的动作条件视觉先验——如物体持存性、场景几何与运动对应线索——即使新增领域的具身形态并不相同。
- *机器人操作领域共享训练早期的先验。*图 28 的机器人领域面板显示,操作类数据集之间存在广泛的早期迁移,对单领域基线较弱或数据更异质的领域尤其明显。WidowX-250 从 Google Robot 获益显著,FD PSNR 提升 +1.39、策略 PSNR 提升 +2.29,ID 与策略 MSE 也相应改善。Google Robot 也从多个机器人联合训练伙伴中获益,FD PSNR 最高提升 +0.89,策略 PSNR 最高提升 +1.44。相比之下,本协同研究中的 Franka Panda 单臂与双臂条目对应的是很小的 RoboMIND Franka 子集 (Wu et al., 2025a),分别只有 23 小时与 4 小时数据。与图 9 汇总的大得多的动作数据源相比,这些小目标领域在单领域训练下很快饱和,因此加入更大的联合训练领域时,它们自身评测上的增益较小或喜忧参半;但作为联合训练数据源,它们仍能为其他领域提供有用的迁移。
- *人类第一人称运动有助于机器人适配。*图 29a 显示第一人称运动与 AgiBot 之间存在幅度不大但一致的迁移:第一人称数据把 AgiBot FD PSNR 从 23.07 提升到 23.20(+0.13),AgiBot 也把第一人称 FD PSNR 从 15.13 小幅提升到 15.16(+0.03)。图 29b 的预热(warmup)曲线进一步把这种迁移作为初始化效应来检验:我们先在第一人称运动上预热基础 `PT-init` 检查点,再将其适配到 AgiBot,与直接从同一 `PT-init` 检查点适配 AgiBot 对比。从第一人称预热的检查点出发,在每个测量步上都改善了 AgiBot 的 FD PSNR:5K 步时增益 +0.94,训练后期约 +1.3–1.6。综合来看,这些结果表明第一人称人类动作数据为机器人适配提供了有用的操作先验,同时随着训练推进,领域感知的采样或专门化仍然重要。

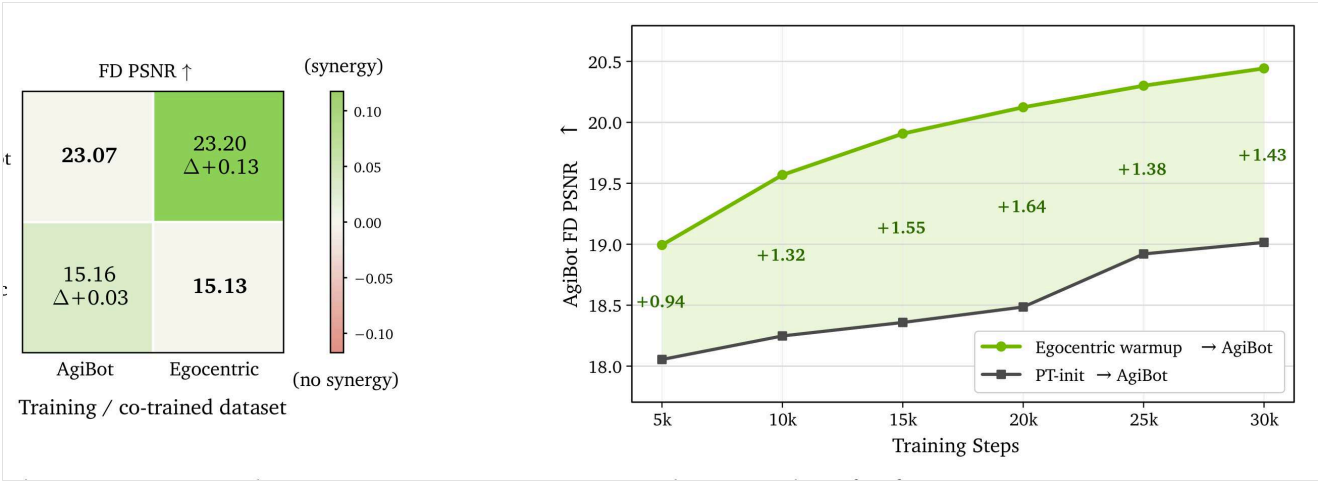


图 29:第一人称运动作为机器人适配先验。(a) 矩阵度量 AgiBot 机器人操作与第一人称(Egocentric)运动之间的成对迁移。(b) 曲线比较从第一人称预热检查点适配 AgiBot 与直接从 `PT-init` 检查点适配的效果。

结论。我们证明了 Cosmos 3 是一个跨领域、跨具身形态、跨推理模式的强大动作基础模型。单一基础模型即可被适配到相机控制的视频生成、自动驾驶逆向动力学、第一人称手部前向动力学、机器人前向动力学与机器人策略学习,同时与专门的领域基线相当甚至更强。`PT-init` 与 `MT-init` 的对比进一步表明,统一的动作中期训练不只是改进孤立的领域:它产生了一个可复用的动作领域先验,加速各类下游设定的收敛,包括对新具身形态与新环境的适配。这一点尤其值得注意,因为被评测的领域在具身形态与监督格式上差异巨大,从相机与车辆运动,到有关节的人手与机器人末端执行器。协同效应研究提供了「动作领域共享可迁移结构」的证据:跨自运动、机器人操作与第一人称运动的联合训练能改善早期适配,对单领域基线较弱的领域尤为明显。附录 E.4 与 E.5 中额外的动作专项消融实验(ablation)也强化了同一结论:FD、ID 与策略目标在联合训练下共享有用的结构,而与策略动作一同预测出的视频,与由这些相同动作驱动的模拟器推演保持对齐。总体而言,这些发现支持将 Cosmos 3 视为一个通用的世界-动作基础模型,可被高效地专门化到广泛的动作生成与控制任务。

6.3 Generator 用户指南(Generator User Guide)

基准评测结果提供了生成质量的定量评估,但要有效使用一个全模态世界模型,还需要理解它的推理接口、提示词方法与采样配置。由于 Cosmos 3 支持多样的模态与生成模式——包括图像、视频、音视频、迁移、前向动力学、逆向动力学与策略生成——输出质量不仅取决于模型权重,也取决于生成请求如何被指定与施加条件。本节提供使用 Cosmos 3 Generator 的实用指导,包括推荐的提示词策略、结构化描述格式、采样超参数,以及 Cosmos 3 Reasoner 作为提示词上采样器(prompt upsampler)的角色。这些指导反映了我们在全部评测中使用的配置,可作为在广泛的 Physical AI 应用中获得高质量、物理合理生成结果的参考。

6.3.1 生成指南(Generation Guide)

Cosmos 3 Generator 支持在很宽的推理范围内进行灵活的视觉(及音视频)生成:帧率 10–30 FPS,帧数 5 到 400 帧,分辨率涵盖 256p、480p 与 720p,以及常见宽高比(1:1、3:4、4:3、9:16、16:9)。这使得单一模型无需更换采样接口,即可服务从短预览片段到更长、更高分辨率的横屏或竖屏视频等各种用例。Cosmos 3 Generator 针对前向动力学、逆向动力学与策略三种模式进行了训练,以支持不同的动作相关应用。对动作生成,基础的 Cosmos3-Nano 与 Cosmos3-Super 模型支持在 10–30 FPS 的原生控制频率下进行动作预测,在逆向动力学与策略模式下预测时程覆盖 16–400 帧。后训练的模型则专门化到单一模式与频率——例如 `Cosmos3-Nano-Policy-DROID` 以 15 FPS 运行,预测时程为 32 步。下面我们详述使用 Cosmos 3 Generator 成功生成的关键要素,这些细节也汇总在表 21 中。

表 21:各生成器与生成模态的默认采样配置与负面提示词。我们汇总不同 Cosmos 3 Generator 模式的生成设置。各设定的负面提示词全文见附录;「Null」表示空字符串是表现最好的变体。

Model	Generation Modality	Sampling Hyperparameters	Negative Prompt
Cosmos3-Nano	Audio-Visual	`steps=50, guidance=6, shift=10, full-range CFG`	附录 B.6
Cosmos3-Super	Audio-Visual	`steps=50, guidance=6, shift=10, full-range CFG`	附录 B.6
Cosmos3-Super-Text2Image	Visual	`steps=50, guidance=4, shift=3, full-range CFG`	Null
Cosmos3-Super-Image2Video	Visual	`steps=50, guidance=6, shift=5, full-range CFG`	附录 B.3
Cosmos3-Nano	Forward/Inverse Dynamics	`steps=50, guidance=1, shift=5, full-range CFG`	Null
Cosmos3-Super	Forward/Inverse Dynamics	`steps=50, guidance=1, shift=5, full-range CFG`	Null
Cosmos3-Nano-Policy-DROID	Policy	`steps=4, guidance=3, shift=5, full-range CFG`	Null
Cosmos3-Nano	Transfer	`steps=50, guidance=3, control guidance=1.5, shift=10`	附录 B.6
Cosmos3-Super	Transfer	`steps=50, guidance=3, control guidance=1.5, shift=10`	附录 B.6

媒体规格与提示词指南。Cosmos 3 Generator 在结构化 JSON 描述(structured JSON captions)上训练,这种格式提供对场景构成的细粒度控制,涵盖主体、背景、光照、美学、摄影术语,对视频还包括时间维度的字段,如动作、状态变化、相机运动与分段级描述(完整模式定义见附录 A)。推理时,由提示词上采样器——由 Claude Opus 4.6 或 Cosmos 3 Reasoner 提供服务——把用户请求转换为同样的结构化格式,确保生成提示词与训练时所见的分布相匹配(模板指令见附录 B.1)。上采样器被指示先描述场景布局与世界状态,然后指定事件的时间推进,最后再补充音频描述。具体的上采样器指令在不同生成模式间略有差异——例如动作与迁移生成会施加针对其条件输入定制的额外任务约束(迁移生成的提示词前缀见附录 B.4,动作提示词指南见附录 B.5)。该 JSON 规格还包含显式的媒体控制项(时长、FPS、空间高宽与宽高比),使提示词解释与采样配置可检视、可复现。

负面提示词。我们通过自动化的基准评测迭代,为每个模型与生成模式分别调优负面提示词(negative prompt)。对每种配置,我们在候选模板上做消融,候选覆盖自然语言描述、关键词列表、指令式表述、针对物理一致性与身份保持的组合式扩展,以及空字符串;依据自动化基准评测分数选出表现最好的变体。基础的 Cosmos3-Nano 与 Cosmos3-Super 生成器所用的显式负面提示词见附录 B.6。对于后训练变体,我们发现空字符串负面提示词对 `Cosmos3-Super-Text2Image` 效果最好,而对 `Cosmos3-Super-Image2Video`,使用从用户提示词自动派生的负面提示词效果最佳(附录 B.3)。对动作生成模式,我们发现空字符串负面提示词效果最好。

生成采样超参数。我们对不同模态采用如下采样参数:

1. *音视频生成*。音视频生成的采样超参数在图像与视频生成的自动化基准评测(§6.2.1 与 §6.2.2)上调优。对基础的 Cosmos3-Nano 与 Cosmos3-Super 生成器,我们使用 50 步去噪、引导系数(guidance scale)6、时间偏移(time shift)10,以及全范围无分类器引导(full-range CFG)。对后训练的 `Cosmos3-Super-Text2Image` 模型,我们使用引导系数 4、时间偏移 3;对后训练的 `Cosmos3-Super-Image2Video` 模型,我们使用时间偏移 5。
2. *动作生成*。动作生成支持三种模式。对前向与逆向动力学,我们使用 50 步去噪、引导系数 1、时间偏移 5 与全范围无分类器引导;对策略模式,则切换为 4 步去噪、引导系数 3。
3. *迁移生成*。视频迁移生成的采样超参数在自动化的 PAIBench-C 基准评测(§6.2.4)上调优。我们使用 50 步去噪、文本引导系数 3、控制引导系数 1.5、时间偏移 10 与全范围无分类器引导。

6.3.2 Cosmos 3 Reasoner 作为提示词上采样器(Prompt Upsampler)

提示词上采样是 Cosmos 3 中位于生成之前的推理步骤,它把简短的用户意图翻译为一份结构化的时空场景规格。它不只是改写提示词,而是把稀疏的用户输入扩展为一种物理上有依据的控制语言,捕捉场景布局、时间演化,以及(在适用时)音频线索。用户通常只提供一句简短的自然语言请求,而高质量的图像与视频生成依赖许多很少被显式指定的细节,例如主体属性、空间布局、相机行为、光照、时间顺序、音频线索,以及分辨率、宽高比、时长与帧率等生成控制项。上采样器作为一个位于用户界面与生成器之间、遵循指令的 LLM 模块来填补这一空缺:它接收简短请求、输出结构模板,以及可选的条件输入信号(例如图生视频的起始图像),产出一份稠密的、带类型的 JSON 场景规格,供下游的图像或视频模型据此渲染。

先前的工作可分为两条线。第一条线把简短的用户查询映射为更丰富的图像生成文本提示词:DALL-E 3 使用描述性重新标注与描述上采样,以更好地匹配生成器的描述分布 (Betker et al., 2023);Prompt Expansion 显式学习把一个查询扩展为多条优化过的文生图提示词 (Datta et al., 2024);而 Promptist、BeautifulPrompt、RePrompt 等提示词适配系统训练语言模型把用户输入改写为模型偏好的提示词,同时优化图像层面的目标 (Hao et al., 2023; Cao et al., 2023; Wu et al., 2025b)。第二条线把 LLM 用作下游视觉生成器的规划器,产出物体布局、有落地依据的场景描述、逐帧提示词或多场景视频计划 (Lian et al., 2024; Feng et al., 2023; Hong et al., 2023; Lin et al., 2023)。Cosmos 3 采纳了同样的基本思想——在生成之前插入一个推理模块——但进一步改变了接口:上采样器输出的是一份带类型的多模态场景程序,而非仅仅一条改写后的提示词或某个生成器专属的布局计划,且该程序横跨图像、视频与音频条件生成。某种意义上,上采样器的工作方式是:先在抽象的语言描述状态中想象场景、考虑其时间/空间方面、并补充更多的多模态内容(如音频描述符),然后把它们翻译为与所给模板匹配的结构化输出。

我们把上采样视为一种带物理先验、面向结构化数据的复杂多模态指令遵循任务。输入可以是纯文本、图像加文本或视频加文本;输出是一份受模式约束的 JSON 控制程序,捕捉渲染器所需的语义内容、视觉属性、任务专属描述、视频任务的时间计划、视频任务的音频描述字段与生成参数。结构化输出不只是一种格式选择,它显式暴露了生成器必须施加条件的潜变量,包括实体、空间关系、动作、时序、相机行为、音频后果与生成控制项。这把提示词上采样与结构化 LLM 推理与验证的研究联系起来,后者把自然语言指令转换为在结构化字段上显式、可检查的程序 (Chegini et al., 2025)。与此同时,填写一份稠密的场景模式需要不确定性下的推理:上采样器必须从不完整的用户输入中推断可能的物体状态、时间转换、因果交互与声学结果 (Pournemat et al., 2025),从而沿创造性维度扩展,同时不与所设定现实的物理约束相矛盾。在 Cosmos 3 中,这些概率与物理先验通过模式本身来表达:模型先想象一个自洽的世界状态,再把它映射为一段时间展开,最后推导出与可见事件保持同步的音频线索。这种分离的好处在于,它把提示词理解暴露为一个可独立检视的组件,而不是与渲染模型纠缠在一起。

提示词契约(prompt contract)。推理时,上采样器接收用户描述与所选定的生成控制项;对图生视频(I2V)生成,它还接收条件图像,该图像被视为第一帧的决定性视觉证据。请求被格式化为四个带标签的块:1) `instructions` 块:介绍所有块,并定义任务级行为,例如只产出单个用代码围栏包裹的 JSON 对象、避免额外注释、把模板视为强制要求;2) `image_description` 或 `video_description` 块:承载原始语义请求;对 I2V 等带条件的请求,所附的条件内容随该文本一起提供,约束条款会指明视觉事实应如何锚定于它;3) `task_constraints` 块:承载大部分操作细节,指定字段顺序、时间格式、时长上下界、被复制的控制项(如分辨率、宽高比、时长与帧率)、对用户提供的实体与动作的保留要求,以及任务专属的锚定规则(如在 t = 0 处匹配 I2V 的第一帧);4) `output_json_template` 块:定义模式的外形——哪些键必须填充、期望哪些嵌套字段,以及场景、时间、音频、主体、相机、光照、风格与控制信息各应放在何处。换言之,约束条款描述上采样器应如何推理与复制取值,而 JSON 模板定义最终答案必须具有的形状。系统提示词是极简的「You are a helpful assistant」。完整模板见附录 B.1。使用该模板时,描述、FPS、时长、宽高比、分辨率等占位参数须先被相应填好,再传给上采样器。

7. 相关工作 Related Work

物理AI(Physical AI)近期的进展,是由若干此前彼此独立的研究领域共同推动的,其中包括世界模型(world model)、多模态理解、视频生成(video generation)、动作建模、音视频生成,以及全模态(omnimodal)基础模型。尽管这些方向各自都为感知、推理、模拟与控制贡献了重要能力,但物理AI系统最终要求这些能力在一个统一框架内协同运作。在本节中,我们回顾与 Cosmos 3 最相关的文献,重点关注世界模拟(world simulation)、多模态推理、生成式建模与具身智能方面的已有工作。我们着重说明这些研究脉络如何朝着对物理世界日益统一的表征演进,并讨论 Cosmos 3 如何延续这一轨迹——在单一全模态世界模型内对语言、图像、视频、音频与动作进行联合建模,同时服务于理解与生成。

7.1 面向物理AI的世界模型

世界模型描述世界如何演化,以及智能体的动作如何改变未来的观测。一个有用的区分是预测式隐式世界模型(predictive latent world model)与生成式世界模型(generative world model):前者学习紧凑的内部动力学以用于规划与控制,后者则把预测出的未来暴露为可供检视的多模态模拟。

预测式隐式世界模型学习紧凑的隐藏状态,其价值在于让规划与控制变得更廉价:控制器可以在一个抽象掉像素级细节的表征中进行搜索或优化。早期的隐式动力学系统,如 World Models (Ha and Schmidhuber, 2018)、Embed-to-Control (Watter et al., 2015)、PlaNet (Hafner et al., 2019) 与 Dreamer (Hafner et al., 2020),表明紧凑的预测状态足以支撑从像素出发的控制与规划。较新的 JEPA 风格模型则把预测从像素转移到隐式抽象层面,这类抽象与感知、预报和规划更为对齐 (LeCun, 2022; Assran et al., 2023; Bardes et al., 2025; Assran et al., 2025; Maes et al., 2026)。这些方法确立了内部动力学建模作为智能体核心要素的地位,但它们通常是为紧凑预测或下游控制而优化的,而非为高保真的多模态模拟。

生成式世界模型把模拟本身作为建模接口:模型以图像、视频、音频或其他有落地(grounded)的 token 形式预测可能的未来观测。通过直接暴露预测出的未来,几何、接触、时序与声音方面的错误变得可观测,而不是只能从任务损失中间接推断。Sora (OpenAI, 2024) 使这一视角在视频生成与隐式世界模拟领域受到广泛关注 (He et al., 2026)。Cosmos 世界模型系列则直接面向物理AI发展这一方向 (NVIDIA, 2025a,b,c,d)。其他领域专用系统研究操作(manipulation)、驾驶,以及具身导航与交互 (Wang et al., 2024b; Zhao et al., 2025; Liang et al., 2024a; Jang et al., 2025; Ren et al., 2025a; Zhou et al., 2024; Gao et al., 2026b)。近期的交互式系统进一步把同一方向扩展到可控的交互式环境(interactive environment)与面向智能体的推演(rollout)接口 (Bruce et al., 2024; Google DeepMind, 2024b; Ball et al., 2025; Hu et al., 2023; Yang et al., 2023; World Labs, 2025; Bahmani et al., 2025; Shen et al., 2026a; Waymo, 2026; Li et al., 2025c; Wang et al., 2025g)。Cosmos 3 建立在这一方向之上,但把世界建模视为理解与生成兼具的问题,而不仅仅是视频合成:其 Reasoner 塔解读多模态上下文并推断结构化的世界状态,而其 Generator 塔则以该上下文为条件合成未来的图像、视频、音频与动作 token。这使得一个框架就能连接场景理解、未来合成、动作推断与多模态推演,从而生成的轨迹可以以文本、观测、动作与音频为条件,而不只是以提示词为条件。

7.2 多模态理解与具身推理

多模态理解模型提供了物理智能所需的感知与推理层。Flamingo (Alayrac et al., 2022) 与 BLIP-2 (Li et al., 2023b) 帮助确立了将大语言模型与视觉输入耦合的现代范式,而 LLaVA (Liu et al., 2023b; Li et al., 2024a) 使指令微调的开源视觉-语言助手产生了广泛影响 (Zhu et al., 2024)。近期的模型家族在图像与视频上改进了规模、落地能力、时序推理、OCR 与指令跟随 (Chen et al., 2024c; Wang et al., 2025e; Dai et al., 2024; Bai et al., 2023, 2025b,c; Deitke et al., 2025; Peng et al., 2024; You et al., 2024; Zhang et al., 2024; McKinzie et al., 2024; Zhang et al., 2025a; Tschannen et al., 2025; Xiao et al., 2024a; Beyer et al., 2024; Steiner et al., 2024)。GPT-4o (OpenAI, 2024) 则体现了向全模态交互演进的趋势。这些能力使视觉-语言模型成为具身系统有用的前端,但物理智能所需要的不止是对孤立图像或片段的识别。它需要在时间上持续存在的状态、绑定到物体与智能体的空间落地、对可能交互的可供性(affordance)推理,以及在条件变化时对任务进度的追踪。这把推理的基本单元从一条描述或一个答案,转变为一个持续维护的、可供行动的场景估计。

因此,对物理AI而言,相关的挑战不仅是视觉问答,而是在对空间、时间、可供性、物体状态与任务进度进行推理的同时,维护有落地的场景状态。Cosmos-Reason1 (NVIDIA, 2025) 通过物理常识与具身思维链推理来应对这一设定。相关的机器人与第一人称视角(egocentric)工作强调空间落地、具身规划,以及人-物或机器人-物交互推理,兼用模型开发与面向任务的基准 (Sermanet et al., 2024; Wang et al., 2023; Chen et al., 2025b; Yang et al., 2025a; Chen et al., 2024a; Song et al., 2025; Zhou et al., 2025b; Yang et al., 2025c)。尽管取得了这些进展,大多数理解模型本质上仍是判别式或以语言为输出的系统:它们能够描述场景、回答问题或推断下一步意图,但通常无法在同一表征空间中生成未来观测或可执行的动作。Cosmos 3 通过让结构化的多模态理解直接为世界生成提供条件,来弥合这一割裂。

7.3 视频生成与视觉世界模拟

视频生成已从简短的文本到视频片段,发展到高分辨率、时间上连贯、遵循指令的合成。早期的扩散(diffusion)与 Transformer 系统确立了文本条件视频生成的基本配方 (Ho et al., 2022b; Singer et al., 2023; Ho et al., 2022a; Villegas et al., 2023; Blattmann et al., 2023; Bar-Tal et al., 2024; HaCohen et al., 2025; Kondratyuk et al., 2024; Girdhar et al., 2024; Yang et al., 2025e; Open-Sora Team, 2025; Genmo, 2024; Zhang et al., 2025c)。Sora (OpenAI, 2024, 2025)、Movie Gen (Polyak et al., 2024) 与 Veo 3 (DeepMind, 2025; Google DeepMind, 2025b) 反映了近期向长时程真实感、更强提示词遵循度与更高保真视觉动态的转变 (Kong et al., 2024; Wan et al., 2025a; Gao et al., 2025; Arkhipkin et al., 2025)。工业界系统也已把可控性、创作者工作流与 API 部署变成视频生成版图的核心组成部分 (Runway, 2024, 2025; Kuaishou, 2024, 2025; Luma, 2024; Luma AI, 2025; MiniMax, 2024, 2025; Pika Labs, 2025; Adobe, 2025; Amazon, 2024)。

由此形成的格局在很大程度上是围绕感知层面的视频生成来界定进展的:看上去合理的画面、平滑的运动,以及与文本或视觉条件相匹配的输出。关于视频世界模拟与物理一致性评测的工作说明了这些目标对世界建模是必要的,但对物理AI而言并不充分 (He et al., 2026; Bansal et al., 2025a; Guo et al., 2025; Motamed et al., 2026)。世界模拟施加了一个更强的契约:推演必须保持物体身份与物体恒存性,遵守时间因果性,暴露可控因素(如动作或相机运动),并且当同一场景在不同干预下被查询时保持一致。对智能体而言,只有当画面中的变化能够回溯到状态、动作与接触时,一段视频才是有用的;否则,真实感并不意味着模拟器可靠。在反复提示或闭环使用中,这一差距尤为明显——细小的不一致会累积放大,导致下游决策出错。因此,面向模拟的系统必须把合成与有落地的状态及干预语义耦合起来。在这一背景下,Cosmos 3 把视频置于更宽广的世界建模框架之内。视频不仅是一种输出模态,同时也是推理的输入、动作推断的观测流,以及与动作和控制耦合的状态轨迹。

7.4 动作建模、VLA 与世界-动作模型

动作提供了智能体与不断变化的世界状态之间的因果纽带。已有工作通常被组织为三种设定:前向动力学(forward dynamics)由状态与动作历史预测未来观测;逆动力学(inverse dynamics)从观测到的状态转移中推断背后的动作;策略(policy)模式则把观测、目标与指令直接映射为动作。这一分类法覆盖了异构的动作空间——从机器人关节指令与车辆控制,到相机、人体与第一人称佩戴者的运动——每一种都有不同的单位、频率与因果作用范围。

当动作被作为未来观测的条件对待时,前向动力学模型的形式最为直白。在驾驶领域,GAIA-1 (Hu et al., 2023)、DriveDreamer (Wang et al., 2024b; Zhao et al., 2025) 与 Cosmos-Drive-Dreams (Ren et al., 2025a) 展示了如何在自车运动、轨迹或其他控制条件下生成未来场景。在机器人与相机可控生成方面,相关系统探索了动作条件的操作视频、感知机器人形态的模拟器、相机轨迹、深度、分割以及 3D 一致的控制 (Liang et al., 2024a; Jang et al., 2025; Zhou et al., 2024; Zhu et al., 2025; Wang et al., 2025c; Gao et al., 2026b; He et al., 2024; Wang et al., 2024e; Xu et al., 2024a; Ren et al., 2025b)。这些系统表明,生成式模型可以学到有用的动作条件先验,但其中许多是针对特定领域、动作空间或具身形态(embodiment)特化的。

在实践中,逆动力学常常问的是:哪个动作解释了观测到的状态转移。VPT (Baker et al., 2022) 研究从视频中标注动作,后续工作把这一思路扩展到从观测中模仿、隐式动作发现,以及在没有显式动作标注的情况下从视频中学习 (Zhang et al., 2022; Torabi et al., 2018; Yang et al., 2019; Pavse et al., 2020; Schmidt and Jiang, 2023; Ye et al., 2025; Garrido et al., 2026)。Open X-Embodiment (Vuong et al., 2023) 与 DROID (Khazatsky et al., 2024) 等大型具身数据集,为跨具身形态、任务、视角与操作模式的学习提供了数据基底 (Walke et al., 2023; Liu et al., 2023a; Nasiriany et al., 2024; contributors, 2024; Bu et al., 2025; Grauman et al., 2024; Li et al., 2023a)。

策略模式系统从观测、目标与指令中预测动作。RT-1/RT-2 (Brohan et al., 2023b,a)、PaLM-E (Driess et al., 2023)、OpenVLA (Kim et al., 2024)、π_0 (Black et al., 2025)、Gemini Robotics (Gemini Robotics Team et al., 2025; Gemini Robotics Team, 2025) 与 GR00T N1 (Bjorck et al., 2025) 勾勒出从视觉-语言策略到通用机器人基础模型的演进脉络 (Li et al., 2024d; Octo Model Team et al., 2024; Physical Intelligence Team, 2025; Liu et al., 2024c; Zhou et al., 2025d; Zheng et al., 2026; Yang et al., 2025d; Shi et al., 2025b; Lee et al., 2025a; Wu et al., 2023b; Cheang et al., 2024)。相关的具身智能体系统利用语言或视觉-语言模型来构建空间价值图、进行动作推理与交互感知的规划 (Huang et al., 2023; Zawalski et al., 2024; Yang et al., 2025a)。世界-动作模型(world-action model)通过把视频动力学与动作表征用作先验,使策略学习更接近世界建模 (Li et al., 2025a; Ye et al., 2026; Wang et al., 2025i; Cen et al., 2025; Wang et al., 2026b; Xue et al., 2026)。Cosmos 3 把前向动力学、逆动力学与策略模式统一视为同一个多模态序列模型在视频、音频、文本与动作 token 之上的不同条件化模式。

7.5 音频与音视频生成

音频是物理世界建模的重要组成部分,因为许多事件不仅由其外观界定,也由其声音界定。AudioLDM 2 (Liu et al., 2024b)、AudioCraft (Meta AI, 2023) 与 MusicGen (Copet et al., 2023) 代表了文本条件音频与音乐生成的快速进展,后续系统进一步把这一空间扩展到语音、音效与面向创作者的产品 (Le et al., 2023; Vyas et al., 2023; Stability AI, 2024; Lee et al., 2025b; Suno, 2024; Udio, 2024; ElevenLabs, 2024)。这些模型提升了生成媒体的声学真实感,但物理世界模拟还要求声音与可见的动态保持同步。

音视频学习研究的是:声音与视觉是否对应于同一事件、同一声源或同一运动。经典脉络包括跨模态对应、声源分离、同步以及语音-唇形对齐,而 Diff-Foley (Luo et al., 2023) 及相关的视频到音频系统则专注于生成跟随可见动态的声音 (Owens and Efros, 2018; Zhao et al., 2018; Ephrat et al., 2018; Chung and Zisserman, 2016; Prajwal et al., 2020; Comunita et al., 2024; Zhang et al., 2026; Ren et al., 2025c; Gramaccioni et al., 2024; Cheng et al., 2025; Kling Team, 2025)。说话人头像、人体动画与音视频联合生成器进一步把语音、身体运动、场景动态与声音连接起来,包括 Movie Gen (Polyak et al., 2024) 与 Veo 3 (DeepMind, 2025; Microsoft Research, 2024; Tian et al., 2024; Xu et al., 2024b; Wang et al., 2025a; Xing et al., 2024; Ruan et al., 2023; Wang et al., 2024a; Liu et al., 2024a; Li et al., 2025b; HaCohen et al., 2026; Liu et al., 2025c) 中的同步音视频媒体生成。对物理模拟而言,时机与声音的身份同等重要:碰撞应当在接触发生的那一帧产生撞击声,脚步声应当匹配步态与地面材质,工具或引擎在其可见状态变化时声音也应随之改变。类似地,语音需要口型运动、说话人身份与话轮转换保持对齐,而警报声、马达声、刮擦声等环境声音,则可以为部分被遮挡的事件提供因果证据。对物理AI而言,音频应当被建模为关于事件、接触、语音、工具、引擎与环境的同步证据,而不是事后添加的装饰。Cosmos 3 把音频作为同一世界建模接口中的又一种模态,从而支持音视频联合生成、以视频为条件的音频生成,以及以音频为条件的视频生成。

7.6 兼顾理解与生成的全模态模型

一条日益壮大的研究脉络关注在单一框架中结合多模态理解与生成的全模态模型(omnimodel)。把这一目标与两个相邻的模型家族区分开来是有益的:视觉-语言模型(VLM)通常把多模态输入映射为文本,而媒体生成器则把提示词或条件映射为生成的图像、视频或音频。全模态模型旨在于一个框架内同时支持这两个方向,使感知、语言与合成能够共享表征,而不是仅靠外部流水线相互连接。

Unified-IO 2 (Lu et al., 2024) 与 Chameleon (Chameleon Team, 2024) 研究统一的多模态建模,而 GPT-4o (OpenAI, 2024) 与 Gemini (Google DeepMind, 2024a, 2025a) 则表明前沿系统正在走向原生的多模态输入与输出。其他统一或任意到任意(any-to-any)系统探索了文本、图像、音频、视频的不同组合方式,以及离散 token 化或扩散接口,其中 Qwen-Omni (Xu et al., 2025; Qwen Team, 2026a) 代表了近期的一个全模态模型家族 (Tang et al., 2023; Wu et al., 2023c; Zhan et al., 2024; Ge et al., 2024; Mizrahi et al., 2023; Xie et al., 2025; Chen et al., 2025d; Wang et al., 2024c)。

在这一更宽泛的全模态家族中,近期 MoT 风格的工作尤其相关,因为它探讨的是:一个模型如何在共享序列级上下文的同时,为不同模态或不同功能保留专门化的容量。Transfusion (Zhou et al., 2025a) 通过把文本的下一 token 预测与基于扩散的图像生成相结合,将这条线索与在混合模态序列上训练单个 Transformer 联系起来。Mixture-of-Transformers (Liang et al., 2025) 为多模态基础模型把这种专门化模式显式化,相关工作还包括改造预训练语言模型以用于多模态生成,以及研究异构的理解专家与生成专家之间的非对称桥接 (Shi et al., 2025c; Wang et al., 2025f)。BAGEL (Deng et al., 2025) 把这一方向扩展为兼顾理解与生成的统一多模态预训练,采用在大规模交错多模态数据上训练的 decoder-only 架构。近期的具身扩展工作则让同样的架构问题走向物理层面——在 VLA 与世界-动作模型中为场景理解、视觉预见、隐式动作与控制分别设置专门化通路 (Lv et al., 2025; Cai et al., 2026; Shou et al., 2026; Bi et al., 2025; MotuBrain Team et al., 2026; Li et al., 2026a; Tencent Robotics X et al., 2026)。

综合来看,这些全模态模型表明,理解与生成可以在共享的多模态框架内处理,而无需依赖彼此独立的系统。然而,现有工作仍侧重于文本-图像建模或一般性的多模态媒体生成,对物理世界动力学、动作条件生成、逆动力学与具身控制的关注较少。Cosmos 3 把全模态的思想扩展到物理AI:将用于多模态理解的自回归(autoregressive)Reasoner 塔,与以 Reasoner 表征为条件的扩散 Generator 塔配对。这一接口支持视频/文本到文本的理解、文本/视频到视频的生成、音视频生成,以及用于动作理解与预测的世界-动作建模。Cosmos 3 不仅是多模态的,而且是全功能(omni-functional)的:同一个模型可以解读世界、模拟世界如何演化、推断观测到的变化背后的动作,并生成未来的观测与动作。

8. 结论 Conclusion

我们提出了 Cosmos 3——一个面向物理AI的全模态世界模型家族。Cosmos 3 在单一架构内统一了跨语言、图像、视频、音频与动作的多模态理解与生成,降低了将视觉-语言模型、视频生成模型、世界模型与动作模型分别组合的需求。这一统一表述由模态专属编码器、结构化的 token 编排,以及将自回归推理与基于扩散的生成耦合起来的 Mixture-of-Transformers 骨干网络共同实现。配合可扩展的数据、训练、服务与评测基础设施,Cosmos 3 为开发物理AI智能体提供了坚

实的基础。在多样的理解与生成基准上,Cosmos 3 展现了强劲的结果与广泛的能力覆盖,同时仍便于通过后训练进行下游特化。我们期望 Cosmos 3 成为连接合成世界与真实世界的桥梁,为物理AI提供更好的合成数据、更好的特化模型起点,以及更好的闭环训练环境。通过开源源代码、预训练检查点与精选基准,我们希望加速通用物理AI智能体的研究,使其能够在真实世界中感知、推理、模拟与行动。

附录 A:描述文本细节 Caption Details

本附录详细介绍我们自研打标模型(captioner)的设计,以及用于图像与视频训练数据的完整结构化描述文本(caption)模式(schema)。我们训练自己的模型,而不是依赖现成的 VLM,以确保对输出结构的最大控制,并防止模式字段出现幻觉或缺失。我们不单纯依赖自由形式的自然语言描述文本,而是把标注组织成带有预定义语义字段的 JSON 对象。这一设计鼓励对视觉细节的全面覆盖,同时保持表示的一致性与可解释性。图像模式捕获静态场景属性,包括主体、布局、光照、美学、风格与空间构图;视频模式则在此表示之上扩展了时间信息,如动作、状态变化、转换、相机运动与音频线索。

A.1 打标模型

为最大化图像训练数据的细节与空间覆盖度,我们采用象限扫描(quadrant-scan)标注策略。每张图像被划分为四个象限,每个象限内的内容——连同中心区域——被独立描述。这一方法对捕获多个不同主体与复杂布局非常有效;这类情形在我们的数据集中频繁出现,却往往被标准打标方法描述不足。

对视频数据,我们引入专门针对时间动态的第二遍标注。由于运动与状态转换是视频理解的根本,这一遍标注会对所有被显式跟踪的视觉属性全面标注其时间变化。

对于最终的图像与视频打标模型,我们确定对 Qwen3-VL-8B 做 LoRA 微调能在基准指标与推理效率之间取得最佳平衡。对视频输入,我们以 8 FPS 采样帧,并将生成温度设为 0.7。我们还使用 Qwen3-VL 默认的最小与最大像素上下界,分别为 131,072 与 25,165,824。

A.2 图像模式

我们为文生图训练数据使用一种结构化描述文本表示,覆盖广泛的视觉属性,包括前景主体、场景构图、背景元素、光照、美学、艺术风格、相机视点与摄影(cinematography)属性。描述文本不依赖密集的自由形式自然语言,而是表示为带预定义语义字段的 JSON 对象。这种结构化表示在保持高精度的同时,改善了细节召回与标注一致性。

为改善空间覆盖度,打标流水线引入了基于象限的扫描机制,把每张图像划分为四个空间象限外加一个中心区域。每个区域先被独立描述,再合并进最终描述文本。这一设计对包含多个不同主体、局部交互或复杂空间布局的图像尤其有效,而这类图像在常规打标方法中往往描述不足。

我们按语义类别总结该模式,并在下方给出紧凑的 JSON 骨架。

图像描述文本模式的顶层字段(Top-level Fields in the Image Caption Schema)

字段	描述
`subjects[]`	逐主体的记录,涵盖身份、外观、空间位置、姿态、着装、表情,以及适用时的计数敏感解剖字段。人类或类人主体在可见时还额外包含人口统计与面部属性字段。
`subject_details`	无法干净地归入单个主体槽位的开放式属性或补充说明,包括有用时的自由形式细粒度面部细节。
`background_setting`	全局场景上下文、环境与背景元素。
`lighting`	光照条件、方向性、阴影与显著的光照效果。
`aesthetics`	构图、色彩组合、氛围与重复出现的视觉图案。
`cinematography`	取景、相机角度、景深、对焦行为与镜头特性。
`style_medium` / `artistic_style`	媒介、渲染风格与艺术处理。
`context`	场景的高层叙事或情境上下文。
`text_and_signage_elements[]`	任何可见文字及其外观、位置、类别与场景相关性。
`quadrant_scan`	对左上、右上、左下、右下与绝对中心区域的分区扫描。
`comprehensive_t2i_caption`	从结构化字段提炼出的自然语言总结。
`resolution` / `aspect_ratio`	用于保留尺度与布局线索的图像尺寸元数据。

`subjects` 中条目的数量与 `subject_details` 的内容随场景复杂度而变化。

人类与类人属性。对于人类或类人主体,该模式还可额外包含:

- `clothing`
- `expression`
- `gender`
- `age`
- `skin_tone_and_texture`
- 细粒度面部属性,如眼形、眼睛颜色、唇形、发色、皱纹或痣,通常通过 `appearance_details` 或 `subject_details` 表示

计数敏感属性。若某主体对应一组或一簇相似实体,该模式还可额外包含:

- `number_of_subjects`

- `number_of_arms`
- `number_of_hands`
- `number_of_fingers`
- `number_of_legs`

紧凑 JSON 骨架。为使图像标注标准化并防止遗漏关键视觉细节,我们强制使用一个严格的预定义 JSON 结构。下面的模式列出了从背景设定到美学风格的具体语义字段,用于全面捕获每张图像的静态属性。

模板框:Compact Image-specific JSON Skeleton(紧凑图像 JSON 骨架)——即图像描述文本的完整字段结构;注释标出了仅人类/类人适用与计数敏感的字段组。原文保留英文以便直接使用:

```
{
  "subjects": [
    {
      "description": "...",
      "appearance_details": "...",
      "relationship": "...",
      "location": "...",
      "relative_size": "...",
      "orientation": "...",
      "pose": "...",
      // human or human-like only
      "clothing": "...",
      "expression": "...",
      "gender": "...",
      "age": "...",
      "skin_tone_and_texture": "...",
      // group- or count-sensitive attributes
      "number_of_subjects": 0,
      "number_of_arms": 0,
      "number_of_hands": 0,
      "number_of_fingers": 0,
      "number_of_legs": 0
    },
    ...
  ],
  "background_setting": "...",
  "lighting": {
    "conditions": "...",
    "direction": "...",
    "shadows": "...",
    "illumination_effect": "..."
  },
  "aesthetics": {
    "composition": "...",
    "color_scheme": "...",
    "mood_atmosphere": "...",
    "patterns": "..."
  },
  "cinematography": {
    "framing": "...",
    "camera_angle": "...",
    "depth_of_field": "...",
    "focus": "...",
    "lens_focal_length": "..."
  },
  "style_medium": "...",
  "artistic_style": "...",
  "context": "...",
  "text_and_signage_elements": [
    {
      "text": "...",
      "category": "...",
      "appearance": "...",
      "spatial": "...",
      "context": "..."
    }
  ],
  "quadrant_scan": {
    "top_left": "...",
    "top_right": "...",
    "bottom_left": "...",
    "bottom_right": "...",
    "absolute_center": "..."
  },
  "comprehensive_t2i_caption": "...",
  "resolution": {
    "H": xx,
    "W": xx
  },
  "aspect_ratio": xx
}
```

A.3 视频模式

视频描述文本模式在图像描述文本模式的基础上扩展了捕获时间信息的字段。共享的静态属性——如主体、背景、光照、美学、风格与相机视点——遵循上文所述的图像描述文本模式。此处我们只关注视频特有的新增部分。

视频模式显式记录场景如何随时间演化。它捕获主体级动作与状态变化、全局动作时间线、时间分段、分段之间的转换、相机运动以及可选的音频描述。这些字段改善了对运动、交互与时间连续性的表示,而这些是纯图像描述文本所无法捕获的。

我们按语义类别总结视频特有的模式组成部分,并在下方给出紧凑的 JSON 骨架。

视频描述文本模式新增的顶层字段(Additional Top-level Fields in the Video Caption Schema)

字段	描述
`actions[]`	按时间顺序排列的关键视觉动作。
`segments`	按镜头、场景或视频内有意义的变化划分的不同时间分段。
`transitions`	分段之间的显著转换,或视频中的时间性变化。
`temporal_caption`	对视频中所有时间性变化的密集描述。
`resolution` / `aspect_ratio` / `duration` / `fps`	视频元数据。
`audio_description`	视频中音频内容的描述。

`actions`、`segments` 与 `transitions` 中的条目数量随视频的时间复杂度而变化。短视频或静态视频可能只包含少量时间事件,而包含多次场景切换、交互或相机运动的较长视频可能需要更细致的时间分段。

视频特有字段。与图像模式相比,视频模式引入了以下新增组成部分:

- `action`:主体级的运动或活动。
- `state_changes`:主体外观、姿态、位置或状态随时间的变化。
- `camera_motion`:相机随时间的行为,如平移(pan)、俯仰(tilt)、变焦、跟踪或手持晃动。
- `actions`:对视频中重要事件按时间索引的描述。
- `segments`:对视频主要区间的时间局部化描述。
- `transitions`:分段、镜头、场景或主体状态之间的变化。
- `temporal_caption`:对视频时间进程的整体总结。
- `duration` 与 `fps`:基本视频元数据。
- `audio_description`:对语音、音乐、音效或环境音的可选描述。

紧凑视频 JSON 骨架。图像模式捕获静态场景属性,而视频标注需要跟踪随时间的动态。为避免冗余,我们对视频数据使用一个有意省略前文已覆盖的静态视觉属性的紧凑 JSON 骨架。下面的模式严格聚焦视频特有的语义字段,如动作、转换、相机运动与音频线索。

模板框:Compact Video-specific JSON Skeleton(紧凑视频 JSON 骨架)——注意注释处标明了从图像模式继承而省略的字段;原文保留英文:

```
{
  "subjects": [
    {
      // fields inherited from the image schema are omitted here
      "action": "",
      "state_changes": ""
    },
    ...
  ],
  "cinematography": {
    // fields inherited from the image schema are omitted here
    "camera_motion": ""
  },
  "actions": [
    {
      "time": "",
      "description": ""
    },
    ...
  ],
  "text_and_signage_elements": [
    {
      // image-level text fields are inherited from the image schema
      "spatial_temporal": ""
    },
    ...
  ],
  "segments": [
    {
      "segment_index": 0,
      "time_range": "",
      "description": "",
      "key_changes": "",
      "camera": ""
    },
    ...
  ],
  "transitions": [
    "",
    ...
  ],
  "temporal_caption": "",
  "duration": xx,
  "fps": xx,
```

```
  "audio_description": ""
}
```

附录 B:Generator 默认提示词与提示词上采样模板 Default Prompts and Prompt Upsampling Templates for Generator

本附录给出推理时 Cosmos 3 Generator 在不同条件下使用的提示词上采样器(prompt upsampler)指令模板与完整的负面提示词。我们在每个小节中说明所用的上采样器及对应用例。与这些模板配套的采样配置汇总见表 21。

B.1 Cosmos 3 Reasoner 的上采样器提示词模板

下面的片段展示了将 Cosmos 3 Reasoner 作为提示词上采样器调用的指令模板:它在推理时把用户提示词转换为 Cosmos 3 Generator 期望的结构化 JSON 模式(详见 §6.3.2)。

模板框:Abridged Canonical Prompt-Template Examples(规范提示模板示例节选)——依次给出 T2V(文生视频)、T2I(文生图)、I2V(图生视频)、V2V(视频续写)四种用户消息。关键设计点:`<instructions>` 强制模型只输出恰好一个 fenced JSON 对象;`<task_constraints>` 中的 `duration`、`fps`、`aspect_ratio`、`resolution` 必须"原样照抄"——FPS 与时长在这里被硬性控制,而非交给模型自由发挥;所有时间字段一律用 M:SS 记法且不得超出 `duration`;I2V 把附带的首帧当作决定性的视觉真值、文本只承载时间/动作意图;V2V 则把条件视频当作已观测前缀的视觉与时间真值,并要求后续时间线延续其连贯性:

```
T2V user message:
<instructions>
Prompt upsampler for a text-to-video model. Produce exactly one fenced JSON
object that fully populates the template and satisfies all constraints.
</instructions>
<video_description>{description}</video_description>
<task_constraints>
1. Write scene_imagination first.
2. Write temporal_caption second as the timestamped M:SS playback timeline.
3. Write audio_description third, aligned with the visual beats when possible.
4. Copy exactly: duration="0:06", fps=24, aspect_ratio="1,1",
   resolution={"W":640,"H":640}.
5. Keep all timed fields within duration and mutually consistent.
6. Preserve the user's described subjects, actions, props, and setting.
</task_constraints>
<output_json_template>
{"scene_imagination":"...", "temporal_caption":"...", "audio_description":"...",
 "subjects":[...], "background_setting":"...", "lighting":{...},
 "aesthetics":{...}, "cinematography":{...}, "style_medium":"...",
 "artistic_style":"...", "context":"...", "actions":[...],
 "text_and_signage_elements":[...], "segments":[...], "transitions":[...],
 "resolution":"Per task constraints", "aspect_ratio":"Per task constraints",
 "duration":"Per task constraints", "fps":"Per task constraints"}
</output_json_template>

T2I user message:
<instructions>
Prompt upsampler for a text-to-image model. Produce exactly one fenced JSON
object that fully populates the template and satisfies all constraints.
</instructions>
<image_description>{description}</image_description>
<task_constraints>
1. Write scene_imagination first.
2. Write comprehensive_t2i_caption second as a dense 80-200 word image prompt.
3. Copy exactly: aspect_ratio="1,1", resolution={"W":960,"H":960}.
4. Keep subjects, background, lighting, aesthetics, and camera fields consistent.
5. Populate subject_details with 2-5 image-specific attributes.
</task_constraints>
<output_json_template>
{"scene_imagination":"...", "comprehensive_t2i_caption":"...",
 "subjects":[...], "subject_details":{...}, "background_setting":"...",
 "lighting":{...}, "aesthetics":{...}, "cinematography":{...},
 "style_medium":"...", "artistic_style":"...", "context":"...",
 "text_and_signage_elements":[...], "quadrant_scan":{...},
 "resolution":"Per task constraints",
 "aspect_ratio":"Per task constraints"}
</output_json_template>

I2V user message, with attached starting frame:
<instructions>
Prompt upsampler for an image-to-video model. Treat the attached starting frame
as definitive visual ground truth and the text as temporal/action intent.
</instructions>
<video_description>{description}</video_description>
<task_constraints>
1. Write scene_imagination first, anchoring visual facts to the image and
   temporal facts to the description.
2. Copy exactly: duration="0:20", fps=20, aspect_ratio="9,16",
   resolution={"W":480,"H":832}.
3. Use only M:SS timing and keep all timed fields within duration.
4. Ensure the first segment and earliest actions match the image at t=0.
5. Preserve concrete facts from both the image and the description.
</task_constraints>
<output_json_template>
{"scene_imagination":"...", "temporal_caption":"...", "audio_description":"...",
 "subjects":[...], "background_setting":"...", "lighting":{...},
```

```
 "aesthetics":{...}, "cinematography":{...}, "style_medium":"...",
 "artistic_style":"...", "context":"...", "actions":[...],
 "text_and_signage_elements":[...], "segments":[...], "transitions":[...],
 "resolution":"Per task constraints", "aspect_ratio":"Per task constraints",
 "duration":"Per task constraints", "fps":"Per task constraints"}
</output_json_template>

V2V user message, with attached conditioning video:
<instructions>
Prompt upsampler for a video-to-video continuation model. Treat the attached
conditioning video as definitive visual and temporal ground truth for the
observed prefix, and the text as future/action intent.
</instructions>
<video_description>{description}</video_description>
<task_constraints>
1. Write scene_imagination first, summarizing the conditioning video's state,
   subjects, motion history, and final visible configuration.
2. Write temporal_caption second as the future M:SS playback timeline after the
   conditioning video, preserving continuity with the observed prefix.
3. Write audio_description third, aligned with visible future events when possible.
4. Copy exactly: duration="0:05", fps=24, aspect_ratio="16,9",
   resolution={"W":1280,"H":720}.
5. Preserve concrete facts from the conditioning video and the description.
</task_constraints>
<output_json_template>
{"scene_imagination":"...", "temporal_caption":"...", "audio_description":"...",
 "subjects":[...], "background_setting":"...", "lighting":{...},
 "aesthetics":{...}, "cinematography":{...}, "style_medium":"...",
 "artistic_style":"...", "context":"...", "actions":[...],
 "text_and_signage_elements":[...], "segments":[...], "transitions":[...],
 "resolution":"Per task constraints", "aspect_ratio":"Per task constraints",
 "duration":"Per task constraints", "fps":"Per task constraints"}
</output_json_template>
```

B.2 Cosmos3-Super-Text2Image 的上采样器提示词模板

我们使用以下指令模板,为后训练(post-trained)的 Cosmos3-Super-Text2Image 生成上采样后的 JSON 提示词,详见 §4.2.3。

模板框:Cosmos3-Super-Text2Image Prompt Upsampler Template——关键设计点:输出必须"DENSE"(即便用户请求很简短,也要为每个字段推断出合理、场景一致的细节);只有真正不适用的字段(非人类主体的人类专属字段、无可见文字等)才允许空值;不得增删模板中的键;最终只输出用 ```` ```json ```` 代码围栏包裹的 JSON 对象(强制 JSON 输出)。占位符 `{caption_dense}` 处填入用户的视觉请求:

```
Given the user's natural-language request below, generate a dense structured JSON that fully describes the image to be produced. The JSON must strictly
follow the template provided after the request, including every top-level key and every nested sub-field.

The output is always DENSE. Even when the request is brief, you must infer plausible, scene-consistent details for every field. Do not leave fields empty
merely because the request did not mention them - the purpose of this task is to upsample a sparse request into a rich, complete annotation. Be creative
but stay grounded: your additions must be physically plausible and internally consistent with the request.

Requirements:
- For every visual field, write rich, specific content inferred from the request's scene, subjects, mood, and context.
- Empty values ("", 0, [], {}) are permitted ONLY for truly inapplicable fields:
  * Human-only subject fields (clothing, expression, gender, age, skin_tone_and_texture, facial_features, number_of_arms, number_of_legs, number_of_hands,
number_of_fingers) when the subject is non-human.
  * text_and_signage_elements = [] when no visible text or signage is present.
  * aesthetics.patterns = "" when there are no notable repeating patterns.
  * subject_details = {} when no image-specific structured attributes apply.
- Do not add keys beyond the template. Do not omit keys required by the template.

Return only the JSON object wrapped in a ```json code fence.

USER VISUAL REQUEST:
{caption_dense}

Lists (subjects, text_and_signage_elements) may contain zero or more items of the shape shown. All top-level keys must always be present in the output;
fill unused fields with "", 0, {}, or [] as appropriate.

{
  "subjects": [
    {
      "description": "full visual description of the subject",
      "appearance_details": "additional visual details (accessories, texture, distinguishing features)",
      "relationship": "how this subject relates to others or to the scene",
      "location": "where in frame (e.g., 'Center foreground', 'Top right')",
      "relative_size": "size within frame",
      "orientation": "direction subject faces relative to camera",
      "pose": "body position and posture",
      "clothing": "clothing and accessories; '' if non-human or N/A",
      "expression": "facial expression; '' if non-human or N/A",
      "gender": "one of 'Male', 'Female', 'Unknown', 'N/A'",
      "age": "age category",
      "skin_tone_and_texture": "skin tone description; '' if non-human",
      "facial_features": "notable facial features, including eye shape/color, hair color/style, lip shape, wrinkles, moles, scars, freckles, facial hair,
and other visible fine-grained facial attributes; '' if non-human or not visible",
      "number_of_subjects": "int; total in this subject group, 0 if N/A",
      "number_of_arms": "int; 2 for humans, 0 if non-human",
      "number_of_legs": "int; 2 for humans, 0 if non-human",
      "number_of_hands": "int; 2 for humans, 0 if non-human",
      "number_of_fingers": "int; 10 for humans, 0 if non-human"
    }
  ],
  "subject_details": {
    "key_name_1": "free-form image-specific attribute (keys vary by image content; {} if N/A)"
```

```
  },
  "background_setting": "full prose description of the environment and setting",
  "lighting": {
    "conditions": "type and quality of light",
    "direction": "where light comes from; 'None' for flat digital images",
    "shadows": "shadow description; 'None' for flat digital images",
    "illumination_effect": "overall effect of the lighting"
  },
  "aesthetics": {
    "composition": "framing and compositional choices",
    "color_scheme": "dominant colors and palette",
    "mood_atmosphere": "emotional atmosphere in short phrases",
    "patterns": "notable repeating visual patterns; 'None' if none"
  },
  "cinematography": {
    "framing": "shot type",
    "camera_angle": "angle (e.g., 'Eye-level', 'Low angle', 'High angle')",
    "depth_of_field": "'Shallow', 'Deep', 'Uniform focus', or 'N/A'",
    "focus": "what is in sharp focus",
    "lens_focal_length": "descriptive focal length"
  },
  "style_medium": "visual medium (e.g., 'Photography', 'Digital presentation slide', 'Screenshot')",
  "artistic_style": "genre or approach",
  "context": "scene context or use case (brief)",
  "text_and_signage_elements": [
    {
      "text": "the visible text content",
      "category": "one of 'physical_in_scene', 'ui_text', 'body_text', 'scene_sign', 'logo', 'label'",
      "appearance": "font, color, size, style",
      "spatial": "position in image",
      "context": "purpose or meaning of the text"
    }
  ],
  "quadrant_scan": {
    "top_left": "description of what appears in the top-left region",
    "top_right": "description of what appears in the top-right region",
    "bottom_left": "description of what appears in the bottom-left region",
    "bottom_right": "description of what appears in the bottom-right region",
    "absolute_center": "description of what appears at the center"
  },
  "comprehensive_t2i_caption": "a comprehensive, full-scene natural-language prose description of the image"
}
```

B.3 Cosmos3-Super-Image2Video 的上采样器提示词模板

下面的片段展示了与 Claude Opus 4.7 搭配使用的指令模板,用于为后训练的 Cosmos3-Super-Image2Video 模型生成上采样 JSON 提示词以及逐样本的上下文相关负面提示词(详见 §4.2.4 与 §6.3.1)。

模板框:Cosmos3-Super-Image2Video Prompt Upsampler Template——模板分两个阶段:阶段 1 在 `<final_prompt>` 标签内产出密集叙事式的 `temporal_caption`,其撰写规则覆盖首帧一致性、物理准确性、因果先行(先有原因后有倒影/水花等效果)、物体恒存、禁用"the video shows"之类的元表述、用单数代词避免模型渲染出多个主体、把镜头术语(probe lens/鱼眼/无人机)当作拍摄风格而非画面中的实体、默认单一连续镜头等;阶段 2 通过从默认负面提示词中"删词"来生成定制负面提示——只删除与目标视频矛盾的条目,不得新增,最终用 `<negative_prompt>` 标签包裹输出:

```
You are an expert prompt engineer for an image-to-video generative model. You are given a STARTING FRAME image (the first frame of the video) and a USER INSTRUCTION describing the desired motion or changes to animate. Your task is to produce a dense, cinematic video description that the model will use to generate the full video, together with a customized negative prompt.

Complete this task in two phases.

---
### PHASE 1: VIDEO DESCRIPTION
Write a dense, narrative caption inside `<final_prompt>` XML tags, formatted as a JSON object using this exact template:

<final_prompt>{"temporal_caption": "..."}</final_prompt>

Rules for the caption:
- The provided image is the exact starting frame - all described motion must be consistent with the starting frame.
- Opening: Establish the scene — subjects, environment, lighting — describing what is directly visible in the starting frame accurately and faithfully, noting essential elements that the motion will directly involve, the subject's orientation (e.g., "facing away", "in three-quarter profile"), and any implied ongoing motion (e.g., a cyclist leaning into a curve, water already splashing) so the video continues smoothly. Phrase it naturally as a scene description (do not say "in the starting frame", "initially shown", or similar meta-references).
- Motion: Describe the changes and actions in chronological order. Flow naturally from one action to the next. Advance time using natural conjunctions (e.g., "while," "as," "and").
- Physical Accuracy: All motion must obey gravity and reflect realistic material behavior (e.g., cloth ripples, water splashes, rigid objects resist deformation).
- Cause-and-effect: Always describe causes before their effects. Reflections, shadows, and secondary effects cannot appear on their own — the source object must first enter the frame or move into the relevant position before any reflection or shadow is described. E.g., a person must walk to the water's edge before their reflection appears on the surface; an object must strike the water before a splash erupts.
- Object Permanence: Every subject must persist throughout or have a clear reason for entering or exiting. When a new subject not present in the starting frame is introduced (e.g., an opposing team, an arriving vehicle), briefly describe their appearance (e.g., uniform color, vehicle type and color) so the generator can render them consistently, and describe a logical way for them to come into the frame (e.g., entering from a specific side of the frame, walking in through a door, or emerging from behind an existing object) rather than having them appear out of nowhere.
- Taboo Phrases: NEVER refer to the video medium itself. Avoid "the video shows...", "the scene...", "the clip...", "the frame...", "the camera shows...", "we see...".
- Perspective: Describe human body sides from the subject's own perspective (e.g., "her right hand" = the subject's right hand) to avoid ambiguity. This applies whenever a body part enters or moves in the frame: always specify whether it is the left or right (e.g., "his right hand reaches in from the lower edge"), never a bare "a hand enters the frame".
- Pronouns: Use singular pronouns ("he", "she", "him", "her", "it") or a singular noun phrase ("the person", "the rider", "the child") for single subjects. Never use "they"/"them"/"their" to refer to one person, as this can cause the model to render multiple subjects.
- Spatial Phrasing: Use spatial relationships for motion (e.g., "enters from the left", "rises above the horizon") rather than camera-centric descriptions.
- Camera: Include camera motion only if specified in the instruction; otherwise describe from a static viewpoint. Keep any described camera movement
```

```
subtle and gradual — do not exaggerate altitude loss, tilt angle, or speed beyond what is minimally implied by the instruction. Do not use the word
"transition" when describing camera motion.
- Cinematography Terms: When the instruction references a lens, camera, or filming technique (e.g., "probe lens", "macro lens", "fisheye", "drone shot",
"GoPro"), treat it as a cinematographic style describing how the footage is captured — never as a physical object visible in the scene. Mention the style
(e.g., for a probe lens: extreme close shot; for a fisheye lens: extreme wide angle fisheye view) rather than mentioning the lens or camera apparatus
itself.
- Timelapse: If the instruction implies timelapse, explicitly use the word "timelapse" in the caption and avoid exaggerating its effects.
- Cuts & Montages: Always describe a single continuous shot with no hard cuts unless the user instruction explicitly used words like "cut", "hard cut",
"jump cut", "shot change", or "montage". When multiple shots are requested without specifying an exact number, describe at most 3 shots, and dedicate the
majority of the description to the opening action before any cut. Never use phrases like "the first shot", "the opening shot", or number shots as "first",
"second", etc. — simply describe the action directly.
- Tone: Neutral, objective, descriptive. No opinions, value judgments, or inferred emotions unless physically observable.
- Length & Format: Write exactly ONE coherent paragraph of 5-8 sentences. No bullet points or lists.

USER INSTRUCTION:
"{description}"

---
### PHASE 2: NEGATIVE PROMPT
Using your final video description from Phase 1, create a customized negative prompt.

HOW IT WORKS:
A negative prompt describes exactly what a bad video looks like. Use declarative statements (e.g., "blurry faces"). Never use negative instructions like
"avoid" or "do not".

---
DEFAULT NEGATIVE PROMPT:
The video captures a series of frames showing macroblocking artifacts, chromatic aberration, high-frequency noise, and rolling shutter distortion. It
includes static with no motion, motion blur, over-saturation, shaky footage, low resolution, grainy texture, pixelated images, poorly lit areas,
underexposed and overexposed scenes, poor color balance, washed out colors, choppy sequences, jerky movements, low frame rate, bit-depth compression
artifacts, color banding, unnatural transitions, outdated special effects, fake elements, unconvincing visuals, poorly edited content, jump cuts, hard
cut, visual noise, and flickering. It features moiré patterns, edge halos, and temporal aliasing. Furthermore, the content defies common sense, generating
illogical scenarios, nonsensical entities, absurd character behaviors, and conceptual paradoxes that violate basic human reasoning and everyday reality.
The video looks like a surreal or glitchy hallucination. Overall, the video is of poor quality.
---

INSTRUCTIONS:
Delete any words from the default negative prompt that contradict your intended video. Keep most of the original wording and structure intact, and do not
add new items. Examples:
* If you want scene cuts/montages -> REMOVE "jump cuts" and "hard cut".
* If you want a motionless/static scene -> REMOVE "static with no motion".
* If you want fantasy, sci-fi, or surrealism -> REMOVE "defies common sense", "illogical scenarios", "nonsensical entities", "surreal", and related logic-
violation terms.
* If the scene has flickering light -> REMOVE "flickering".
* If it is a night-time timelapse -> REMOVE "motion blur".

Output only the final negative prompt as a single paragraph, wrapped in <negative_prompt> tags. Do not output any explanation or preamble.
```

B.4 视频迁移(transfer)的提示词前缀

对视频迁移任务,我们在用户描述文本前加一个系统提示词前缀,指示模型以视觉控制信号为条件进行生成。具体而言,在训练与推理时,以下系统提示词都会通过 chat 模板拼接在用户提供的描述文本之前:

模板框:Cosmos 3 Base Generator Video-to-Video Transfer Prompt Prefix——一句话前缀,声明可用的控制信号类型(边缘图、模糊、深度或分割):

```
You are a helpful assistant that generates images or videos following the user's instructions and control signals (edge maps, blur, depth, or
segmentation).
```

B.5 动作生成的提示词模板

对动作相关任务,模型接受对预期行为的自然语言描述。与动作和视频相关的元数据以 JSON 格式填充。该格式沿用视频生成所使用的结构化描述文本约定,同时暴露动作特有的字段:相机取景、时间范围、条件帧率、输出分辨率与宽高比。最终提示词形式如下:

模板框:Action Generation JSON Prompt Template——其中 `cinematography.framing` 描述视角(第一人称/第三人称/腕装相机/多视图拼接),`idle_frame` 标注无动作的空闲帧数,`fps` 为条件帧率:

```
{
  "cinematography": {
    "framing": "<viewpoint description>"
  },
  "actions": [
    {
      "time": "0:00-<end time>",
      "description": "<action caption as a sentence>",
      "idle_frame": "<idle frames out of total frames>"
    }
  ],
  "duration": "<integer seconds>s",
  "fps": <conditioning fps>,
  "resolution": {"H": <height>, "W": <width>},
  "aspect_ratio": "<width,height>"
}
```

`cinematography.framing` 字段用于描述输入或输出的视角,涵盖第一人称、第三人称、腕部安装(wrist-mounted)相机或拼接的多相机视图。动作条目跨越整个生成片段,其结束时间由实测时长取整得到;顶层 `duration` 截断为整数秒,以与视频 JSON-caption 格式保持一致。宽高比从一组固定的 generator 分辨率映射而

来。可选的 `idle_frame` 字段指示有多少帧不含动作(空闲帧)。对逆动力学(inverse-dynamics)样本我们省略空闲帧元数据,因为动作预测应严格跟随视频本身而非用户偏好。图 30 展示了一个多视角示例,其中拼接相机布局被直接编码进 JSON 提示词。我们使用 Claude-Opus-4.6 做提示词上采样。

Prompt

```
{
  "cinematography": {
    "framing": "This video contains concatenated views from multiple camera
    ↪  perspectives. The top row is from the wrist-mounted camera. The bottom row
    ↪  contains two horizontally concatenated third-person perspective views of
    ↪  the scene from opposite sides, with the robot visible."
  },
  "actions": [
    {
      "time": "0:00-0:02",
      "description": "Remove the cans from the tray and put them on the
      ↪  countertop.",
      "idle_frame": "4 out of 16."
    }
  ],
  "duration": "2s",
  "fps": 10.0,
  "resolution": {
    "H": 544,
    "W": 736
  },
  "aspect_ratio": "4,3"
}
```

图 30:多视角动作提示词格式化。当有多个视角可用时,我们将它们拼接为单一画布,并在结构化 JSON 提示词中附上视图布局元数据,使模型能够将每个像素区域与其对应的相机流关联起来。

B.6 Cosmos 3 Generator 负面提示词

我们对基础的 Cosmos3-Nano 与 Cosmos3-Super generator 使用以下负面提示词。它本身就是一份与正向描述文本模式同构的结构化 JSON"负面 caption":逐字段描述一段劣质视频长什么样——主体畸变、光照互相矛盾、镜头失控、时序闪烁、时间描述文本逐秒崩坏,并额外包含 `physical_realism` 字段描述物理规律全面失效的情形。以下完整保留原文,未做任何截断:

模板框:Cosmos 3 Base Generator Negative Prompt

```
"subjects": [
  {
    "description": "Blurry, poorly defined subjects with inconsistent shapes and unrealistic proportions.",
    "appearance_details": "Distorted features, visible compression artifacts, muddy textures lacking fine detail, color bleeding between elements, and
unnatural skin tones or surface textures that appear artificial or computer-generated.",
    "relationship": "Subjects appear disconnected from the environment, floating or improperly grounded in the scene without proper occlusion or spatial
coherence.",
    "location": "Subjects are poorly placed within the frame, appearing at awkward positions that violate basic compositional rules.",
    "relative_size": "Inconsistent scale relationships between subjects and the environment, with objects appearing too large or too small relative to
their surroundings.",
    "orientation": "Unnatural orientations that defy physics and spatial logic.",
    "pose": "Stiff, mannequin-like poses with unnatural joint angles and impossible limb positions that look computer-generated.",
    "action": "Incoherent motion with visible frame-to-frame discontinuities. Movement appears as a slideshow rather than smooth animation. Limbs and
appendages pop between positions without interpolation.",
    "state_changes": "Visual state transitions are abrupt and jarring. Colors shift without motivation. Surface textures flicker between different
materials randomly. Outlines shimmer and vibrate.",
    "clothing": "Clothing appears painted on with no sense of material weight or drape. Fabric textures are flat and repeat visibly.",
    "expression": "Frozen, uncanny valley expressions or expressions that change abruptly without natural transition.",
    "gender": "",
    "age": "",
    "skin_tone_and_texture": "Waxy, plastic-looking skin with visible artifacts and inconsistent texture resolution across the frame.",
    "facial_features": "Asymmetric facial features, extra fingers or limbs, teeth that appear blurry or malformed.",
    "number_of_subjects": 0,
    "number_of_arms": 0,
    "number_of_legs": 0
  },
  {
    "description": "Extremely low-quality subjects with visible rendering artifacts, broken mesh geometry, and completely unrealistic proportions
throughout.",
    "appearance_details": "Distorted features, visible compression artifacts, muddy textures lacking fine detail, color bleeding between elements, and
unnatural skin tones or surface textures that appear artificial or computer-generated.",
    "relationship": "Subjects appear disconnected from the environment, floating or improperly grounded in the scene without proper occlusion or spatial
coherence.",
    "location": "Subjects are poorly placed within the frame, appearing at awkward positions that violate basic compositional rules.",
    "relative_size": "Inconsistent scale relationships between subjects and the environment, with objects appearing too large or too small relative to
their surroundings.",
    "orientation": "Unnatural orientations that defy physics and spatial logic.",
    "pose": "Stiff, mannequin-like poses with unnatural joint angles and impossible limb positions that look computer-generated.",
    "action": "Incoherent motion with visible frame-to-frame discontinuities. Movement appears as a slideshow rather than smooth animation. Limbs and
appendages pop between positions without interpolation.",
    "state_changes": "Visual state transitions are abrupt and jarring. Colors shift without motivation. Surface textures flicker between different
materials randomly. Outlines shimmer and vibrate.",
    "clothing": "Clothing appears painted on with no sense of material weight or drape. Fabric textures are flat and repeat visibly.",
    "expression": "Frozen, uncanny valley expressions or expressions that change abruptly without natural transition.",
    "gender": "",
    "age": "",
    "skin_tone_and_texture": "Waxy, plastic-looking skin with visible artifacts and inconsistent texture resolution across the frame.",
    "facial_features": "Asymmetric facial features, extra fingers or limbs, teeth that appear blurry or malformed.",
    "number_of_subjects": 0,
    "number_of_arms": 0,
    "number_of_legs": 0
```

```
  },
  {
    "description": "Poorly generated subjects exhibiting all hallmarks of failed neural rendering -- flickering edges, inconsistent depth, and uncanny spatial relationships.",
    "appearance_details": "Distorted features, visible compression artifacts, muddy textures lacking fine detail, color bleeding between elements, and unnatural skin tones or surface textures that appear artificial or computer-generated.",
    "relationship": "Subjects appear disconnected from the environment, floating or improperly grounded in the scene without proper occlusion or spatial coherence.",
    "location": "Subjects are poorly placed within the frame, appearing at awkward positions that violate basic compositional rules.",
    "relative_size": "Inconsistent scale relationships between subjects and the environment, with objects appearing too large or too small relative to their surroundings.",
    "orientation": "Unnatural orientations that defy physics and spatial logic.",
    "pose": "Stiff, mannequin-like poses with unnatural joint angles and impossible limb positions that look computer-generated.",
    "action": "Incoherent motion with visible frame-to-frame discontinuities. Movement appears as a slideshow rather than smooth animation. Limbs and appendages pop between positions without interpolation.",
    "state_changes": "Visual state transitions are abrupt and jarring. Colors shift without motivation. Surface textures flicker between different materials randomly. Outlines shimmer and vibrate.",
    "clothing": "Clothing appears painted on with no sense of material weight or drape. Fabric textures are flat and repeat visibly.",
    "expression": "Frozen, uncanny valley expressions or expressions that change abruptly without natural transition.",
    "gender": "",
    "age": "",
    "skin_tone_and_texture": "Waxy, plastic-looking skin with visible artifacts and inconsistent texture resolution across the frame.",
    "facial_features": "Asymmetric facial features, extra fingers or limbs, teeth that appear blurry or malformed.",
    "number_of_subjects": 0,
    "number_of_arms": 0,
    "number_of_legs": 0
  }
],
"background_setting": "A poorly rendered, flat background with visible seams, repeated textures, and inconsistent depth cues. The environment lacks volumetric depth and appears as a painted backdrop rather than a three-dimensional space. Vegetation looks like flat cutouts with no volumetric depth. The background appears to have been composited from multiple source materials at different resolutions, creating visible seams and edge artifacts where elements meet. Textures swim and shift across surfaces in a way that breaks the illusion of solidity -- patterns drift laterally rather than staying anchored to the geometry they belong to. Background elements flicker in and out of existence between frames, particularly at the edges of the field of view. The rendering resolution is visibly lower for distant elements, creating a jarring transition between near and far objects. Cloud textures repeat obviously in the sky with visible tiling. Water surfaces lack proper reflection and refraction, appearing as flat animated textures. Fog and atmospheric effects pop in and out rather than smoothly transitioning. Trees and vegetation exhibit obvious LOD (level-of-detail) switching. Building facades have inconsistent window spacing and pattern repetition. The overall scene feels like a poorly assembled collage of individually rendered elements rather than a coherent whole.",
"lighting": {
  "conditions": "Harsh, flat lighting with no natural variation. The scene appears uniformly lit as if by a single overhead fluorescent light, removing all sense of depth and atmosphere.",
  "direction": "Inconsistent light sources -- shadows point in multiple contradictory directions, breaking physical plausibility.",
  "shadows": "Hard-edged, unrealistic shadows that pop in and out of existence between frames. Some objects cast no shadows while others have impossibly dark ones that don't animate smoothly with the object's motion. Shadow edges exhibit visible staircase aliasing artifacts. Shadow maps appear to have been rendered at extremely low resolution, creating blocky patterns. Self-shadowing on characters shows visible peter-panning artifacts where shadows detach from their source. Contact shadows between objects and the ground appear and disappear as objects move slightly. Shadow color is pure black with no ambient contribution, creating an unnaturally harsh contrast that flattens the image. Multiple shadow cascades have visible boundaries where resolution changes. The shadow rendering appears to be temporally unstable -- even static objects have shadows that shimmer and crawl frame to frame, breaking the illusion of a stable light source.",
  "illumination_effect": "No bounce light, no ambient occlusion, no subtle color interactions between surfaces. The scene looks like a poorly lit 3D render from the early 2000s."
},
"aesthetics": {
  "composition": "Cluttered, poorly framed composition with no clear focal point. Important elements are cut off by the frame edges. The rule of thirds is completely ignored, leading to an unbalanced and visually unpleasant arrangement.",
  "color_scheme": "Oversaturated, garish colors that clash violently. Color banding is visible in gradient areas. The overall palette feels artificial and digitally processed rather than natural.",
  "mood_atmosphere": "Unsettling, uncanny atmosphere that fails to evoke any intended emotional response. The scene feels lifeless and sterile despite attempting to portray dynamic action.",
  "patterns": "Visible tiling artifacts in textures, moiré patterns, and aliasing on edges."
},
"cinematography": {
  "camera_motion": "Extremely shaky, unstable camera with visible rolling shutter artifacts. The motion is jerky and discontinuous, causing motion sickness and making the scene impossible to follow.",
  "framing": "Poorly framed shots that cut off important elements and include unnecessary empty space.",
  "camera_angle": "Awkward, disorienting camera angles that provide no useful spatial information about the scene. The camera path exhibits visible mathematical artifacts suggesting simple interpolation between keyframes rather than natural camera operation. Camera motion is completely disconnected from the scene content -- panning away from action, dollying during dialogue, and shaking during still moments. The camera appears to pass through solid objects occasionally. Zoom is applied digitally rather than optically, revealing progressively worse resolution. Camera motion exhibits non-physical acceleration profiles -- instant starts and stops rather than smooth ease-in/ease-out. Rolling shutter simulation is applied inconsistently, present in some frames but not others. The camera occasionally exhibits impossible motion like teleporting between positions. Virtual camera stabilization creates an uncanny floating sensation disconnected from any physical camera rig.",
  "depth_of_field": "Uniform focus throughout, creating a flat, documentary-like appearance with no cinematic depth separation.",
  "focus": "Soft, out-of-focus imagery with visible chromatic aberration and lens distortion that was not corrected in post-processing.",
  "lens_focal_length": "Inappropriate focal length causing barrel distortion and unnatural perspective compression."
},
"style_medium": "Low quality compressed digital video with visible encoding artifacts",
"artistic_style": "Amateur, unpolished with inconsistent visual style",
"context": "A poorly produced video with numerous technical and artistic flaws that detract from any intended narrative or visual impact.",
"actions": [
  {
    "time": "0:00-0:08",
    "description": "Subjects attempt to move but their motion is jerky, temporally inconsistent, and physically implausible. Background elements flicker and shift between frames."
  }
],
"text_and_signage_elements": [],
"segments": [
  {
    "segment_index": 0,
    "time_range": "0:00-0:08",
    "description": "A single continuous shot suffering from severe temporal inconsistencies -- subjects that morph and deform between frames, backgrounds that shift and wobble, and rendering quality that fluctuates visibly over time. Motion blur is applied incorrectly, smearing in directions that don't match actual movement. Frame-to-frame coherence breaks down with individual pixels changing color randomly in flat areas. Texture detail level fluctuates between frames as if the rendering budget varied shot to shot. Color grading drifts over the duration with no creative motivation. Noise patterns change between frames in ways that draw attention rather than being invisible. Overall visual quality degrades progressively from start to finish.",
    "key_changes": "No meaningful progression or narrative development. Visual quality degrades over time.",
```

```
    "camera": "Unstable, poorly controlled camera work with visible mathematical interpolation artifacts."
  }
],
"transitions": [],
"temporal_caption": "The scene opens at 0.0 seconds with a poorly rendered establishing shot that immediately reveals low production quality. At 1.0
seconds, subjects begin to move but their motion is jerky and inconsistent, with limbs bending at unnatural angles and objects clipping through each
other. From 2.0 to 4.0 seconds, the camera shakes violently while the scene exhibits visible compression artifacts, color banding in the sky, and
flickering in the shadows. Between 4.0 and 6.0 seconds, temporal coherence breaks down as elements appear and disappear between frames, textures swim and
morph unnaturally, and the lighting shifts abruptly without physical cause. In the final 2 seconds, the overall visual quality deteriorates further with
increasing noise, blur, and a general loss of spatial coherence that makes the scene nearly unwatchable. Additionally, the frame rate appears inconsistent
with visible judder and stuttering throughout. Color temperature shifts randomly between warm and cool tones with no motivation. The encode quality
degrades in complex regions showing macro-blocking and mosquito noise around moving edges. Temporal noise patterns are spatially correlated, creating
swimming artifacts on flat surfaces.",
"audio_description": "",
"physical_realism": "No adherence to physical laws. Objects defy gravity, pass through solid surfaces, and change mass and momentum without cause. Fluid
dynamics, cloth simulation, and rigid body physics are all fundamentally broken. Furthermore, conservation of energy is violated as objects gain or lose
kinetic energy spontaneously. Elastic collisions produce inelastic results and vice versa. Surface friction is inconsistent -- objects slide on rough
surfaces while sticking to smooth ones. Air resistance appears to affect only some objects while others move through the atmosphere unimpeded."
}
```

B.7 Cosmos3-Super-Text2Image 的智能体式(agentic)上采样

多数顶级(闭源)文生图模型都会做某种形式的在环迭代精修和/或多模态推理(OpenAI, 2026)。Cosmos 3 模型能够接受 JSON 结构化提示词,这为细粒度的迭代式智能体式(agentic)精修打开了多种可能。我们在 Artificial Analysis 提交中对此进行了实测。在测试时,我们通过一个智能体式框架(harness)来推理 Cosmos3-Super-Text2Image:它迭代地上采样用户提示词,通过列出生成图像的缺陷/问题(如有)并给出 1 到 10 的分数来为图像打分,然后同时重写正面与负面提示词。最终输出就是这一循环中按评审(critic)总分最高的那张图像。我们最多使用 2 次重写迭代,且当分数至少为 9 分并且评审未指出严重问题时提前停止。

第一次迭代使用附录 B.2 中的 LLM 上采样模板与空的负面提示词。VLM 评审提示词模板见框 1(Box 1),LLM 正/负面重写器提示词模板见框 2(Box 2)。在我们的提交中,描述文本上采样与重写使用 GPT-5.5,评审使用 Gemini3.1-Pro。该框架在设计上是灵活的,可接受任意 LLM/VLM,包括 Cosmos 3(推理塔)本身。更多信息与脚本请见 nvidia/Cosmos3-Super-Text2Image HuggingFace 仓库。

循环结构如下:

1. 初始上采样:$(p_0, n_0 = \varnothing)$
2. 生成 $x_t \sim \text{Cosmos3}(p_t, n_t)$
3. VLM 评审:打分、视觉问题、提示词覆盖情况、重写指令
4. 若未提前停止,LLM 依据原始提示词、评审报告与历史记录重写出 (p_{t+1}, n_{t+1})

其中第 2–4 步最多重复 2 次重写,最终返回得分最高的图像。

模板框:框 1(Box 1)—— Agentic T2I Critic Prompt(智能体式文生图评审提示词)。让评审 VLM 产出穷尽的缺陷报告:先按物理、解剖、光影、深度尺度、文字渲染等清单逐项排查,再按提示词类别追加专项检查;关键设计点是强制"只返回恰好一个 JSON 对象、无 markdown 围栏、JSON 之外不得有文字",各维度打 0–10 分并输出可执行的重写指令(`improvement_directives`):

```
You are an expert image quality analyst specialising in AI-generated image evaluation.
Your job is to produce an exhaustive defect report. Be meticulous: go beyond the obvious
problems and look carefully for subtle, fleeting, or background issues too.

The following image was generated by an AI image model.
{generated_image}

The attached image was generated from this prompt:
{user_text_prompt}

Analyze this image carefully and list EVERY quality issue you observe.
For each issue give an approximate location and name the specific object or
region involved. Report each distinct occurrence separately.

To make sure you don't miss anything, mentally step through these areas before finalising
your list — but only report issues you actually see:
• Physics: gravity violations, impossible collisions, implausible trajectories
• Object deformation: morphing, melting, stretching of solid objects
• Anatomy: distorted hands, faces, fingers, limbs; wrong body proportions
• Lighting & shadows: missing shadows, inconsistent illumination
• Depth & scale: wrong spatial relationships, perspective issues, scale inconsistencies
• Text / numbers: garbled, floating, or shifting text and digits
• Visual quality: blur patches, noise, compression blocking, visual artefacts, low-resolution regions
• Colour: inconsistent coloration, bleeding, banding
• Action correctness: prompted actions are correctly displayed
• Prompt following: missing subjects, wrong objects, wrong setting, wrong action

Depending on the category of the prompt, you may also need to apply additional checks from the following list:
- Text/commercial/UI/logo checks: readable text for logos, labels, posters, billboards, product packaging, or UI. Verify exact quoted strings, spelling,
legibility, typography, placement, layout, and whether commercial/UI intent is visually clear.
- People/anatomy checks: if humans, human-like characters, body parts, portraits, or poses are present or required by the prompt, inspect faces, eyes,
hands, fingers, limbs, pose, proportions, expression, clothing coherence, and physically possible interactions.
- Fantasy/cartoon/vector/pixel-art checks: if a stylized medium is requested, judge whether stylization is intentional and clean. Penalize messy geometry,
inconsistent line language, broken vector shapes, muddy palettes, and unwanted photorealistic texture.
- Photorealistic/physical checks: if realism, physical objects, geometry, camera behavior, reflections, transparent materials, shadows, perspective,
scale, or contact matter, judge material realism, lighting physics, lens plausibility, and whether objects obey real-world physical constraints.

Return exactly one JSON object, no markdown fences and no prose outside JSON:
{
  "prompt_adherence_score": <number 0-10>,
  "visual_quality_score": <number 0-10>,
```

```
  "aesthetics_score": <number 0-10>,
  "physical_plausibility_score": <number 0-10>,
  "category_score": <number 0-10>,
  "text_rendering_score": <number 0-10 or null>,
  "photorealism_score": <number 0-10 or null>,
  "overall_score": <number 0-10>,
  "issues": [
    {
      "category": "<concise label of your choosing>",
      "description": "<what, where in frame, at what timestamp>",
      "severity": "minor" | "moderate" | "severe"
    }
  ],
  "prompt_elements": {
    "<key noun or action from the prompt>": "present" | "absent" | "partial"
  },
  "category_findings": {{"<check area>": "<concise finding>"}},
  "improvement_directives": ["<specific prompt rewrite instruction>"],
  "rationale": "<2-4 concise sentences>"
}
```

模板框:框 2(Box 2)—— Agentic T2I Joint Positive/Negative Rewriter Prompt(智能体式文生图正/负面联合重写提示词)。让 LLM 依据上一轮的 VLM 评审结果、迭代历史、上一版正向 JSON 与负面提示词,联合重写二者:关键设计点是强制"只返回合法 JSON、无 markdown",输出恰好含 `positive_prompt` 与 `negative_prompt` 两个顶层键,正向 JSON 必须保持 `resolution`/`aspect_ratio` 不变,负面提示词只用于压制具体的错误替代或伪影、不得放正向指令:

```
You are a precise text-to-image prompt engineer. Return valid JSON only, no markdown.
Jointly coordinate the positive structured prompt and generator-side negative prompt so they do not contradict each other.

Original user prompt:
{user_text_prompt}

Application-specific guidance:
Apply the following sections as one checklist program. Do not first classify the prompt.
Apply each section only when relevant to the original user prompt, previous JSON, or VLM failures.
- Text/commercial/UI/logo checks: readable text for logos, labels, posters, billboards, product packaging, or UI. Verify exact quoted strings, spelling,
legibility, typography, placement, layout, and whether commercial/UI intent is visually clear.
- People/anatomy checks: if humans, human-like characters, body parts, portraits, or poses are present or required by the prompt, inspect faces, eyes,
hands, fingers, limbs, pose, proportions, expression, clothing coherence, and physically possible interactions.
- Fantasy/cartoon/vector/pixel-art checks: if a stylized medium is requested, judge whether stylization is intentional and clean. Penalize messy geometry,
inconsistent line language, broken vector shapes, muddy palettes, and unwanted photorealistic texture.
- Photorealistic/physical checks: if realism, physical objects, geometry, camera behavior, reflections, transparent materials, shadows, perspective,
scale, or contact matter, judge material realism, lighting physics, lens plausibility, and whether objects obey real-world physical constraints.
- General scene checks: always judge object completeness, layout clarity, subject relationships, background coherence, visual appeal, and absence of
obvious AI artifacts.

Previous generated image failed or scored according to this VLM analysis:
{
  "overall_score": <number 0-10>,
  "prompt_adherence_score": <number 0-10>,
  "visual_quality_score": <number 0-10>,
  "aesthetics_score": <number 0-10>,
  "physical_plausibility_score": <number 0-10>,
  "category_score": <number 0-10>,
  "text_rendering_score": <number 0-10 or null>,
  "photorealism_score": <number 0-10 or null>,
  "issues": [
    {
      "category": "<concise issue label>",
      "description": "<what failed and where>",
      "severity": "minor" | "moderate" | "severe"
    }
  ],
  "prompt_elements": {"<prompt element>": "present" | "absent" | "partial"},
  "category_findings": {"<check area>": "<concise finding>"},
  "improvement_directives": ["<specific prompt rewrite instruction>"],
  "rationale": "<brief rationale>"
}

Iteration history summary:
[
  {
    "iteration": <integer>,
    "overall_score": <number 0-10>,
    "prompt_adherence_score": <number 0-10>,
    "category_score": <number 0-10>,
    "threshold_cleared": <boolean>
  }
]

Previous positive JSON prompt:
{previous_t2i_json_prompt}

Previous negative prompt:
{previous_negative_prompt}

Joint rewrite task:
Return a JSON object with exactly two top-level keys: "positive_prompt" and "negative_prompt".
"positive_prompt" must be a complete JSON object with exactly these top-level keys, preserving their names and types:
{schema_keys}

"positive_prompt" must keep resolution previous "resolution" and "aspect_ratio".
"negative_prompt" must be a concise generator-side negative prompt string.
Coordinate both fields: strengthen required positive constraints while using the negative prompt only to suppress concrete wrong alternatives or
artifacts.
```

```
Do not put positive instructions in negative_prompt. Do not negate content required by the original user prompt.
For exact counts, grids, text, geometry, or anatomy, explicitly block wrong alternatives when useful.
The positive "comprehensive_t2i_caption" should be direct generation guidance, not an explanation of this rewrite process.
```

附录 C:Generator 训练用合成数据集 Synthetic Dataset for Generator Training

本附录详细描述了 generator 中期训练(mid-training)所用的每一个合成数据生成(SDG)数据集、这些 SDG 数据集相对预训练(pre-training)视频数据集的分布(distribution),以及在 Cosmos 3 训练中针对这些 SDG 数据集的详细消融实验(ablation)。表 22 汇总了五个数据集的规模与所提供的模态,并附有指向 Hugging Face 数据集的 URL 链接;随后逐一给出各数据集的说明卡片。

表 22:SDG 数据集总览。RGB、深度(Depth)、实例分割(Seg)、包围盒(BBox)、物理状态(Phys)、相机参数(Cam)与描述文本(Cap)各列标示是否提供该模态。“part.”表示部分覆盖(仅限部分 clip 或特定生成器)。

数据集	Clip 数	分辨率 / FPS	RGB	深度	Seg	BBox	Phys	Cam	Cap
SDG-PhyxSim	76,489	1920×1080 / 30	√	√	√	—	√	√	√
SDG-RobotSim	208,022	不一(varies)	√	part.	part.	—	part.	√	√
SDG-DriveSim	264,000	3840×2160 / 24	√	—	—	—	—	—	√
SDG-SynHuman	236,937	1920×1080 / 30	√	√	—	—	—	√	—
SDG-Warehouse	122,952	1920×1080 / 30	√	√	√	√	—	√	—

C.1 SDG-PhyxSim

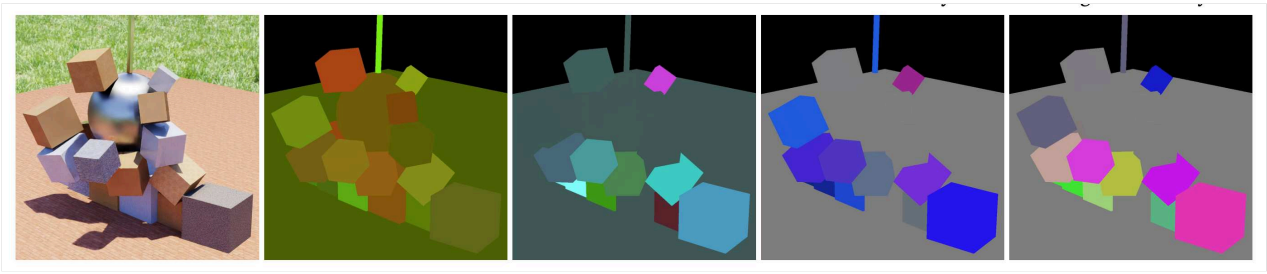


图 31:SDG-PhyxSim。`wrecking_ball` 场景撞击瞬间的单帧画面(`Corner` 相机)。从左到右:RGB、质心位移、累计旋转、线速度与角速度。

概述。SDG-PhyxSim(PhysicsAI-WorldModel-Synthetic-Physical-Interaction-Scenes)是一个大规模合成视频数据集,内容为经过物理仿真的多物体交互场景,旨在让 Cosmos 3 接触到稠密、带真值(ground truth)的刚体(rigid body)动力学——这类监督很难从真实视频中获得。每次仿真 run 都从四个固定相机视角同时采集,产出四段同步的 5–8 秒、1920×1080、30 FPS clip,每段 clip 均配有逐帧实例分割、无损度量深度,以及在渲染时直接从仿真器读取的结构化逐物体物理标注(annotation)(线速度、角速度、质心位移、累计旋转)。该发布版覆盖十个程序化(procedural)参数化的场景族(`dominoes`、`ball_mixer`、`bowling`、`billiards`、`towers`、`wrecking_ball`、`objects_falling`、`rolling_ramp_objects`、`rolling_ramp_obstruct` 与 `obstruction`),每个场景族都为考察一类特定的物理现象而设——级联撞击链、多体混合、弹道轨迹(trajectory)、受约束摆动力学、自由落体与静置、重力驱动滚动、中途偏转、物体恒存性,以及并发的多向碰撞。

仿真设置。SDG-PhyxSim 使用 `NVIDIA Isaac Sim`(NVIDIA, 2026f)基于 `PhysX` 刚体引擎生成,并用 `NVIDIA Omniverse Replicator` 采集。每个场景被编排为一个自包含的 `USD` 资产(asset):其几何、材质绑定(含物理属性)、初始位姿、运动学约束与随机化参数全部收录在单个 `.usda` 文件中,因此任何 clip 都可以仅凭其场景文件在 `Isaac Sim` 中被精确重放。仿真精度设置得很高——每个渲染帧包含 16 个 PhysX 子步,以保证碰撞求解稳定——并在采集前运行一段短暂的预热阶段(0.01 s,步长 0.001 s/步),使物体沉降到静止的初始状态。每段 clip 由 `(scene_name, scene_hash, seed)` 三元组唯一标识;整数 seed 以确定性方式控制所有随机化的场景参数(物体数量、尺寸、质量、材质、间距、初速度)。

场景。十个场景族及其目标物理现象为:

- `dominoes`:弯曲链条中的顺序动量传递;
- `ball_mixer`:旋转搅拌桨下的持续多体混合,clip 为 8 秒;
- `bowling`:定向滚动撞击瓶阵;
- `billiards`:带库边球台上的弹性多球碰撞与旋转传递;
- `towers`:堆叠积木拱在抛射物撞击下的结构性坍塌;
- `wrecking_ball`:受约束摆的动力学与对立方体堆的高冲量拆除;
- `objects_falling`:各类道具的自由落体、撞击、弹跳与静置;
- `rolling_ramp_objects`:物体沿倾斜坡道的滚动与转动惯量表现;
- `rolling_ramp_obstruct`:在坡道场景中加入静态障碍物,考察中途偏转与物体恒存性;以及
- `obstruction`:多向滚动的球穿过静态立柱阵,包含弹跳反弹以及被立柱遮挡时的物体恒存性。

对每个 seed,球体直径、初速度、搅拌桨转速(RPM)、围场尺寸、物体数量、材质分配、坡道角度与障碍物布局均被随机化。

数据集统计。SDG-PhyxSim 发布版共包含 76,489 次独立仿真 run。大多数场景产出 5 秒(150 帧)clip;`ball_mixer` 产出 8 秒(240 帧)clip。所有 clip 均以 1920×1080、30 fps 渲染,合计约 5,700 万帧 RGB。每次 run 从四个固定相机采集,相机命名随场景族而异——例如 `wrecking_ball` 场景为 `Front`、`Side`、`TopDown` 与 `Corner`。该发布版共计 1,529,752 个渲染出的 MP4 文件(RGB 及物理着色变体)、346,147 个深度视频,以及 47,073,033 张逐帧分割 PNG,总存储量约 14.9 TiB,其中以无损深度视频为主。

元数据与标注。除表 22 所列模态外,每次 run 的每个相机还提供:

- *逐物体动态物理量(NPZ)*。每相机四个文件——线速度(m/s)、角速度(deg/s)、累计旋转(rad)与质心位移(m)——每个都是按分割颜色索引的 `(objects × frames × 3)` 张量,并附逐轴取值上下界,用于归一化物理着色视频。
- *逐物体静态物理量(JSON)*。世界重力,以及每个刚体的质量、对角惯性张量、体坐标系质心、主轴四元数、静/动摩擦系数、恢复系数、密度与碰撞标志,以 USD prim 路径和分割颜色为键。
- *物理着色视频*。按相机、按物理量输出的 MP4,其中每个物体的颜色编码其瞬时物理状态(红 = X,绿 = Y,蓝 = Z;静止物体为灰色;饱和度随幅值增大而增大),并按 NPZ 中的上下界归一化,使同一 clip 内相同颜色对应相同的物理量幅值(见图 31)。
- *场景文件(USD)*。用于产出该 run 的自包含 `.usda` 文件,可在 `Isaac Sim` 中精确重放。

度量深度按相机编码为 16 位 FFV1 MKV,每相机的量化上限 $d_{\max}$ 与观测范围记录在 `depth_metadata.json` 中,因此度量深度可按 $d = (v/65535)\, d_{\max}$ 恢复(值 65535 表示无效深度)。实例分割 PNG 附带“颜色→USD prim 路径”映射以保证跨帧身份一致,逐帧相机 JSON 记录内参、外参、FOV、仰角与重力朝向。所有标注均由仿真器与 USD 场景图(scene graph)以确定性方式产出。

C.2 SDG-RobotSim

概述。PhysicalAI-WorldModel-Synthetic-Embodied-Robot-Scenes(本文简写为 SDG-RobotSim)是一个完全合成的机器人视频语料库,用于 Cosmos 训练。它旨在改善物理合理性、具身持续性(embodiment persistence)、接触理解、长时程机器人视频建模与动作条件化推理。公开的 v1.0 版包含 386,270 段 RGB MP4 clip,横跨碰撞、操作与人形运动三类任务。该发布版并不对单一平台做穷尽式建模,而是覆盖移动机器人、四足机器人、人形机器人、固定基座机械臂、双臂系统与灵巧手-臂等多种具身形态。数据集构成概览见图 32。

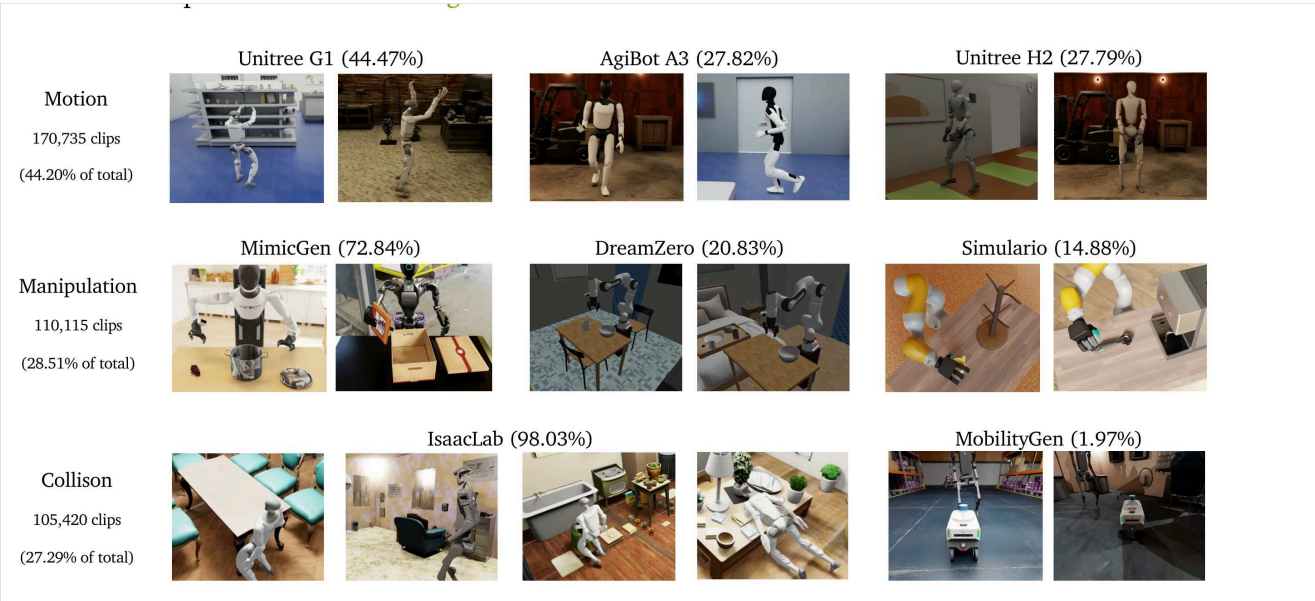


图 32:SDG-RobotSim 数据集总览。clip 被划分为三个任务类别——运动(Motion)、操作(Manipulation)与碰撞(Collision)——每个类别再按其主导的机器人具身进一步细分。

> **译注** 表 22 中 SDG-RobotSim 记为 208,022 段 clip,而本节正文的公开 v1.0 发布版为 386,270 段,两处口径不一致(表 22 可能是训练实际采用的子集,原文未作说明)。类似地,表 22 将 SDG-PhyxSim 记为 76,489,与其“仿真 run 数”一致,但每次 run 含 4 路相机 clip,按 clip 计约为 4 倍。

生成流水线。SDG-RobotSim 由围绕 `NVIDIA Isaac Sim`、`Omniverse`、`Isaac Lab` 及相关机器人数据生成系统(NVIDIA, 2026f,e)构建的、基于 USD 的仿真与渲染流水线生成。该发布版汇集了来自 `IsaacLab` 与 `MobilityGen`(NVIDIA, 2026g)的碰撞 clip,来自 `DreamZero`(Ye et al., 2026)、`MimicGen`/`DexMimicGen`(Mandlekar et al., 2023; Jiang et al., 2025)与 `Simulario`/`DextrAH`(NVIDIA, 2026d)的操作 clip,以及横跨 `SAGE`(Xia et al., 2026)与 `SceneSmith` 两类场景来源的 SOMA 人形运动 clip。生成循环定义物理接地的场景任务、随机化资产与环境、渲染同步的多相机视图、附加源自仿真器的元数据,并通过分布式渲染与索引实现规模化;数据清洗则采用基于规则的过滤、元数据检查、重复运动序列去除,以及针对可见仿真伪影的 VLM 辅助批评(critique)。

数据集统计。公开 v1.0 版包含 386,270 段 RGB MP4 clip,打包为 389 个 WebDataset 分片。运动(Motion)是最大的一族,含 170,735 段 SOMA clip(44.20%);操作(Manipulation)含 110,115 段(28.51%);碰撞(Collision)含 105,420 段(27.29%)。按来源分组,该发布版包含 170,735 段 SOMA 运动 clip、103,340 段 IsaacLab 碰撞 clip、80,162 段 MimicGen 操作 clip、22,926 段 DreamZero 操作 clip、16,384 段 Simulario 操作 clip 与 2,080 段 MobilityGen clip。SOMA 子集包含 37,709 个精选运动组,组织为 SAGE 与 SceneSmith 两个分支,每个分支均含 AgiBot A3、Unitree G1 与 Unitree H2 三个机器人家族。

元数据与标注。每段 clip 包含 RGB 视频与发布版元数据,涵盖任务族、生成器族、具身形态、场景标识符、任务文本或运动名称、相机配置、帧率、clip 长度与可用的仿真器状态。特定生成器的元数据还可能额外包含机器人位姿、关节状态、末端执行器状态、物体位姿、接触标签与任务成功标志。有条件时,描述文本

(caption)由 RGB clip 与仿真器元数据生成。

C.3 SDG-DriveSim

概述。真实世界的驾驶视频数量庞大,但存在结构性偏置:常规巡航被过采样,而对自动驾驶和世界模型压力测试最为关键的安全攸关长尾交互却被欠采样。SDG-DriveSim(PhysicsAI-WorldModel-Synthetic-Autonomous-Driving-Scenarios)是一个用 `NVIDIA Omniverse` 仿真平台生成的大规模自动驾驶场景合成视频数据集,旨在沿真实车队数据难以覆盖的两条轴线填补这一空缺。每段 clip 是围绕一辆自车(ego vehicle)及周边交通参与者的、时间上一致的多相机环视采集,并配有逐相机的 VLM 描述文本。

针对性的长尾覆盖。该数据集围绕在真实数据中明确稀少或难以采集的场景族构建——应急车辆交互、贴近绕行(nudge)停放的障碍物、相邻车道切入(cut-in)、恶劣天气下的能见度退化,以及采用非标准轨迹的行人横穿。由于场景是经由 Scenario Agent 从自然语言提示以声明式方式编排的,而非事后从驾驶日志中挖掘,我们可以按可控密度生成同一 corner case 的大量排列变体。

环境变化。每个编排出的场景都会被展开为在一天中的时段、云量、能见度、路面材质以及车辆与行人资产选择上的确定性排列组合。同一底层交互因此可以在不同环境条件下被观测到,帮助模型把场景内容与环境区分开来。

仿真设置。SDG-DriveSim 由一条构建在 `OpenUSD` 之上的智能体式场景生成流水线产出:一个由 LLM 驱动的 scenario agent 把自然语言提示转换为可运行的 USD 世界配置,随后仿真器渲染该配置并在确定性排列组合上反复重跑,从而对每条提示产出一批 clip。每个场景实例化一张地图(城市、高速、路口、环形道或测试赛道)、一种天气与时段条件、一套相机方案(camera rig)、一辆自车,以及一组可配置的、被赋予行为(行驶、跟随轨迹、变道、切入、贴近绕行、靠边停车、行人动画)的交通智能体与行人。共使用两套相机方案:一套偏重前向的 4 相机方案(120° 前广角、30° 前长焦,外加两个 70° 后侧角相机),以及一套在同样组合上扩展三个 200° 鱼眼相机(左、右、后)以实现 360° 环绕覆盖的 7 相机方案。每个编排出的场景会在时段、云量与能见度、路面材质、车辆与行人资产选择上展开为最多十个确定性排列变体。

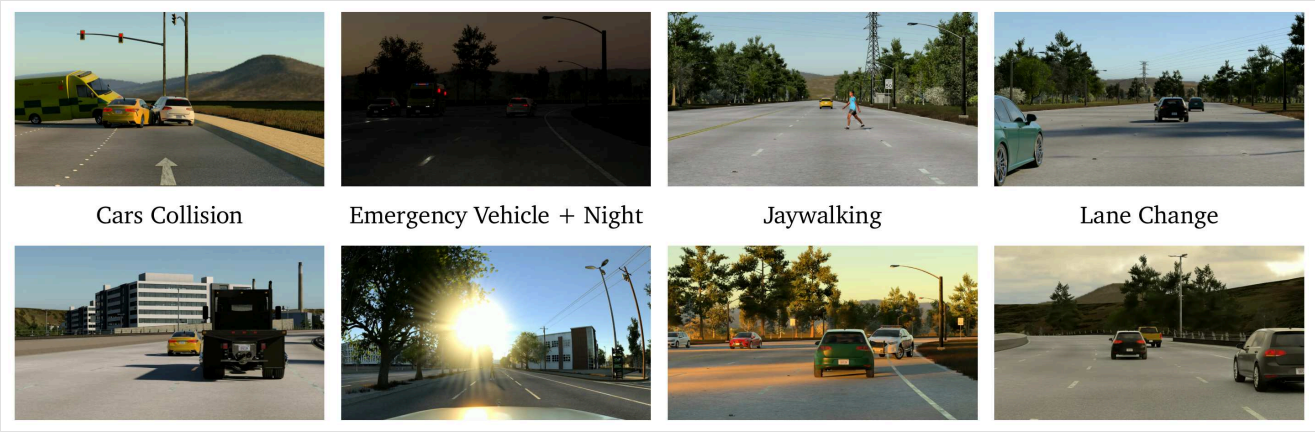


图 33:SDG-DriveSim 数据集。SDG-DriveSim 为覆盖真实世界中难以采集的长尾稀有场景而构建。图中给出数据集中八个有代表性的驾驶场景(车辆碰撞、应急车辆+夜间、乱穿马路、变道、车辆贴近绕行、行人+眩光、警车、天气变化)。

数据集统计。当前 SDG-DriveSim 发布版包含 264,000 段 clip,总计约 1,467 小时视频,以 4K(3840×2160)、24 fps 渲染,单段 clip 时长约 20 秒,约对应 1.27 亿帧 RGB。clip 分布在七个场景族:车辆切入(32.9%)、车辆-行人(21.1%)、车辆变道(12.9%)、行人(12.4%)、车辆天气退化(9.2%)、车辆贴近绕行(8.8%)与应急车辆(2.7%)。该发布版使用了 9 张不同的驾驶地图、10 类车辆资产、8 个行人资产与 3 种行人动画变体,每个场景包含 1 至 9 个交通车辆与行人。

元数据与标注。数据集按场景类别分区,含独立的 `video/` 与 `description/` 子目录。每段 clip 对环视方案中的每个相机产出一个(视频、描述文本)对:24 fps 的 H.264 编码 RGB,加上逐相机的描述文件,其中含帧率、帧数,以及一组按时间窗组织的自然语言描述(`t2w_windows` 条目形如 `(start_frame, end_frame, caption)`)。clip 级场景元数据记录 `weather`、`time_of_day`、`surface_type` 与 `region`。

C.4 SDG-SynHuman

概述。SDG-SynHuman(PhysicsAI-WorldModel-Synthetic-Digital-Human-Scenes)是一个大规模合成视频数据集,内容为渲染在多样 3D 环境中的数字人(digital human),提供真实人体视频极少具备的稠密几何监督(逐帧度量深度与相机内外参)。clip 在 60–120 秒内保持时间一致,每个场景含 1–9 个数字人,并在人物外观、动画、室内外环境、光照条件与相机轨迹上做多样化采样。逐帧相机参数由底层场景图确定性产出,因此相机位姿既可作为输入条件信号也可作为预测目标,从而支持世界模型预训练、相机运动泛化、深度感知学习与人-场景交互建模。

图 34:SDG-SynHuman 样例。从左到右:RGB、深度;分别为室外与室内视图。

仿真设置。SDG-SynHuman 由 NVIDIA 内部 SDG 流水线生成,该流水线构建在 `NeMo Agent Toolkit`、`Omniverse`、`OpenUSD` 与一个内部编排后端之上。流水线把高层场景规格转换为结构化的世界配置,定义环境、光照、数字人、动画、相机模型、相机轨迹与所需的真值输出。每个场景实例化一个 3D 环境、一种光

照条件、一台相机与多个数字人,每个角色被指派一段动画序列与场景摆位。

相机行为通过一份运动配置控制:该配置将一种主相机运动(如静止 static、跟踪 tracking、穿越飞行 fly-through、弧线/环绕 arc/orbit、第一人称 egocentric、之字形 zig-zag 或鸟瞰 bird's-eye)与可选的次级运动层(如抖动 shake、漂移 drift、呼吸式摇摆 breathing sway、荷兰角 dutch angle、变焦推拉 zoom、横移 crab 或俯仰 tilt)组合起来。时间重映射与速度曲线(包括缓入缓出以及程序化采样的自定义曲线)用于在保持 clip 时间连贯的同时改变运动节奏与加速度。数字人选择、动画、相机行为与场景构成均按场景从世界配置中采样,以在人物外观、动作、场景上下文与相机轨迹上产生广泛变化。

环境资产包括 NVIDIA 内部场景、改编自 SceneSmith 示例场景数据集(Pfaff et al., 2026)的室内场景,以及用 `The City Generator Blender` 插件(Dürr, 2026)生成的室外城市环境。渲染期间,仿真器直接从底层 USD 场景产出 RGB 视频、度量深度与逐帧相机标定,从而无需人工标注即可实现确定性的相机与几何监督。

数据集统计。最终 SDG-SynHuman 发布版包含 236,937 段 clip,合计 5,841 小时视频。clip 以 1080p、30 fps 渲染,时长介于 60 至 120 秒之间,平均约 88.8 秒,约对应 6.31 亿帧 RGB,并以相同时间分辨率生成配对的度量深度帧与相机参数。

数据集涵盖 4,050 个不同的数字人资产、8,184 段不同动画、198 个室内环境、200 个室外城市环境,以及 14 种相机运动场景,包括 static、flythrough、tracking、arc、egocentric、zig-zag、bird's eye、tilt、shake、drift、breathing sway、dutch angle、zoom 与 crab。每个场景包含一个 3D 环境、一种光照条件、一条相机轨迹与 1 至 9 个数字人,在人物外观、动画、场景上下文与相机运动上提供广泛变化。该数据集的相机运动统计汇总于表 23 与表 24。

表 23:SDG-SynHuman 中主相机运动的分布。190,670 个场景在七种主运动类型上的占比;每个场景恰好有一种主相机运动类型。可与之叠加的次级相机运动类型见表 24。

运动类型	场景数	占比	说明
static	71,006	29.97%	相机在整段序列中位置与朝向保持固定。
egocentric	33,070	13.96%	相机表现为某个角色或智能体的第一人称视点。
tracking	36,978	15.61%	相机跟随某个主体并保持构图。
flythrough	36,852	15.55%	相机沿路径向前穿越场景。
arc	37,166	15.69%	相机沿弧线绕目标运动。
zigzag	15,294	6.45%	相机前进的同时左右交替横移。
birdseye	6,571	2.77%	自上而下的俯视相机视角。
合计	190,670	100.00%	—

表 24:SDG-SynHuman 中次级相机运动的活跃时长。次级运动是可在时间上相互重叠的组合层,因此表中时长是各运动相互独立的累计值,而非某一固定总时长中的份额。

运动类型	活跃时长(小时)	说明
breathing	1,221.77	模拟人体呼吸的轻柔起伏运动。
drift	1,211.96	随时间缓慢而细微的位置漂移。
dutch_angle	1,205.72	绕前向轴旋转,产生倾斜的地平线。
shake	1,197.87	手持式的快速位置与旋转抖动。
sway	1,191.28	类摆锤的振荡运动,伴随地平线晃动。
zoom	1,184.09	不改变镜头属性的前后移动。
crab	373.72	保持视线方向不变的横向平移。
tilt	195.33	绕水平轴的上下俯仰旋转。

元数据与标注。数据集以 1,215 个 tar 分片交付,置于 `shards/` 下,每个分片打包 200 个样本。每个样本(以 UUID 为键)包含:H.264 编码的 RGB,存于 `video/uuid.mp4`;FFV1 无损 1080p 深度,存于 `depth/uuid.mkv`,并配套一份 JSON 记录深度范围、分辨率、帧率与数据类型以供度量重建;逐帧相机数据存于 `meta/uuid_camera.json`;场景级元数据存于 `metas/uuid.json`(环境、光照、智能体、相机配置、动画任务);资产级元数据存于 `description/uuid.json`(生成的资产清单、动作资产、摆位、来源出处)。所有标注均由 USD 场景图以确定性方式生成。图 34 展示了室外与室内视图的 RGB 与深度输出样例。

C.5 SDG-Warehouse

概述。SDG-Warehouse(PhysicsAI-WorldModel-Synthetic-Warehouse-Operation-Scenes)是一个在完全仿真的仓库环境中渲染的室内工业安全事件合成视频数据集。此类事件的真实监控录像稀少、难以规模化公开发布,且几乎不带物理 AI 训练所需的那种稠密真值,因此我们在仿真中生成:事件必然发生、每个参数皆可控,且每一帧都配有确定性的逐像素与逐物体标注。该发布版覆盖四个代表性场景——叉车-人员险肇(near-miss)(23%)、仓库火灾及工人疏散(36%)、叉车-货架碰撞(20%)与仓库取箱动作(21%)——共约 12.3 万段 clip(约 412 小时视频),分辨率 1920×1080、30 fps,每次仿真 run 有多个同步相机视角。

公共仿真基础设施。四个场景全部构建在 `NVIDIA Isaac Sim` 上。程序化场景组装——仓库布局、货架摆放、道具变化,以及对每个光源的色温、强度、曝光与颜色的随机化——由 `Isaac Sim Replicator Object`(IRO)负责。智能体与传感器布设——工人生成与行为、叉车摆放与导航,以及定义数据集多视角视点的相机方案——由 `Isaac Sim Replicator Agent`(IRA)负责。相机摆放是参数化的,每次 run 采样其高度、距离与俯视角;工人资产与动作从 `Isaac Sim` 的角色库中采样,以丰富人物外观与步态。每次仿真 run 以唯一随机 seed 播种,该 seed 控制所有随机化变量(场景组装、光照、智能体身份与动作、相机位姿与事件时序),因此各 run 相互独立且可复现。

标注模式。每个相机视角提供一段 H.264 RGB clip 与同步的逐帧标注:度量深度(原始值加对数归一化的着色变体);实例分割(逐像素 ID 可追溯到具体的仿真物体,附着色变体);着色分割(shaded segmentation,带基于法线着色的 3D 感知渲染);在着色分割上计算的 Canny 边缘图;对每个被跟踪智能体与道具的 2D 紧致与松

弛轴对齐包围盒,外加 3D 有向包围盒;以及逐帧相机内外参。每条标注流还会以编码视频的形式随 RGB clip 一并发布。run 级结构化元数据记录场景类型、随机 seed、资产与智能体清单、事件参数(例如叉车避让距离、火灾起火点、疏散路径点)与光照随机化。图 35 展示了每个场景一帧的各模态。

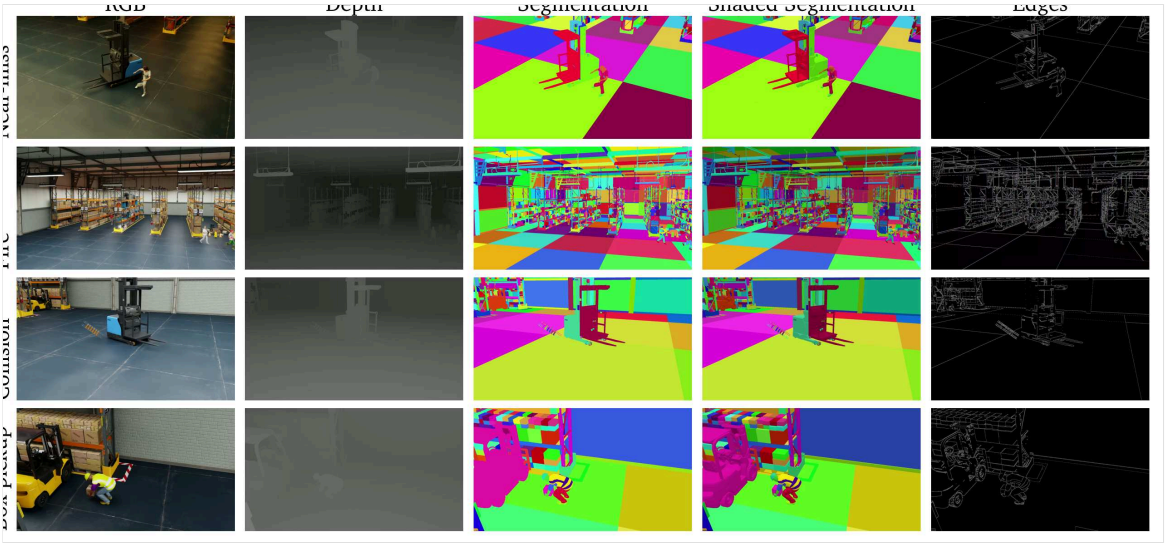


图 35:SDG-Warehouse 数据集。四个场景及其标注的示例视图(RGB clip、度量深度、实例分割、着色分割与 Canny 边缘)。

C.6 SDG 数据集的分布

为验证合成内容确实是对预训练语料的真正补充,我们在 Cosmos-Embed1 视频嵌入空间中,分析其相对预训练分布各聚类的位置。图 36 可视化了预训练与 SDG 嵌入的联合几何结构,表 25 则以“分布内自参照”为基线,量化了每个 SDG 来源到预训练流形的距离。这些结果共同表明:SDG 占据了嵌入空间中仅靠 web 规模预训练无法充分覆盖的长尾区域。

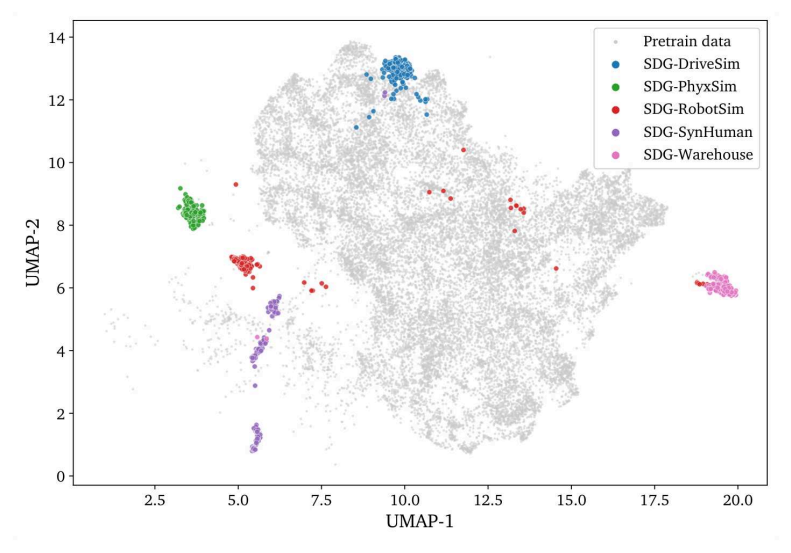


图 36:预训练与 SDG 的联合嵌入几何。对 20,000 个预训练聚类中心(灰色)与每个 SDG 来源随机采样的 200 段 clip 做 PCA→UMAP 投影。每个 SDG 来源都形成一片独特且紧密聚集的区域,与预训练主体分布仅有很窄的重叠。

表 25:SDG 到预训练流形的距离。对每个来源随机采样 10 万段 clip,对照 20,000 个聚类中心计算。*CosSim* 为 $\max_j \cos(\mathbf{g}_i, \mathbf{c}_j)$ 的均值。*MMD*^{2} 为无偏局部最大均值差异(RBF 核,中位数启发式带宽)。*#Local* 为该来源邻域所跨越的不同聚类中心数。所有 SDG 来源都远离预训练分布——它们是必要的补充,而非冗余的子集。

来源	CosSim ↑	MMD^2 ↓	#Local
预训练(参照)	**0.796**	**0.005**	19,934
SDG-DriveSim	0.667	0.119	5,577
SDG-PhyxSim	0.589	0.221	6,644
SDG-RobotSim	0.627	0.162	9,760
SDG-SynHuman	0.650	0.191	4,975
SDG-Warehouse	0.712	0.361	1,920

C.7 消融实验:SDG 数据集的影响

我们通过在每个 SDG 来源上分别单独微调、以及联合微调预训练模型(Cosmos3-Nano),研究不同合成数据生成来源如何影响视频生成质量与领域理解。所有变体均用 PAIBench-G T2V 基准(Zhou et al., 2025c)评测,该基准报告一个总体分(Overall)、一个感知质量分(Quality),以及六个领域子分:常识(Common Sense)、自动驾驶(AV)、机器人(Robot)、工业(Industry)、人体(Human)与物理(Physics)。结果汇总于表 26。

表 26:SDG 数据集消融实验,在 PAIBench-G T2V(Zhou et al., 2025c)上评测。每个模型均从同一预训练基线出发,用单一 SDG 来源的数据微调;SDG-All 混合全部五个来源。**粗体**表示相对基线有提升;每列最优分以突出样式(原文下划线)标出。括号内增减量为相对基线的变化(正为提升,负为退化)。

模型	Overall	Domain	Quality	Comm. Sense	AV	Robot	Industry	Human	Physics
基线(预训练)	79.67 ±0.00	86.87 ±0.00	72.46 ±0.00	91.89 ±0.00	70.86 ±0.00	87.37 ±0.00	88.66 ±0.00	**85.46 ±0.00**	94.58 ±0.00
+ SDG-DriveSim	**79.76** +0.09	**86.97** +0.10	**72.55** +0.09	**92.41** +0.52	70.48 −0.38	**88.26** +0.89	**88.92** +0.26	84.91 −0.55	94.58 ±0.00
+ SDG-RobotSim	79.66 −0.01	86.60 −0.27	**72.72 +0.26**	**92.22** +0.33	69.83 −1.03	**87.60** +0.23	87.44 −1.22	84.99 −0.47	94.50 −0.08
+ SDG-Warehouse	**79.74** +0.07	**87.00** +0.13	**72.48** +0.02	**92.34** +0.45	**71.15** +0.29	**87.97** +0.60	88.63 −0.03	85.01 −0.45	**94.79 +0.21**
+ SDG-PhyxSim	79.62 −0.05	86.68 −0.19	**72.56** +0.10	91.57 −0.32	69.44 −1.42	**88.17** +0.80	**89.51 +0.85**	84.77 −0.69	**94.72** +0.14
+ SDG-SynHuman	**79.79 +0.12**	**87.16 +0.29**	72.41 −0.05	**92.55 +0.66**	**71.33 +0.47**	**88.60 +1.23**	88.51 −0.15	85.08 −0.38	94.56 −0.02
+ SDG-All	**79.77** +0.10	**86.97** +0.10	**72.56** +0.10	**92.40** +0.51	**71.19** +0.33	**87.68** +0.31	**88.79** +0.13	84.99 −0.47	**94.67** +0.09

> **译注** (逐格核对后的三点提醒):(1)总分层面各来源贡献都极小——最佳单源(SynHuman)对 Overall 仅 +0.12(79.67→79.79,约 0.15%),且 RobotSim(−0.01)与 PhyxSim(−0.05)反而使 Overall 略降;SDG 的价值主要体现在个别领域子分上,而非整体分。(2)正文与表格有出入:下文称“SDG-DriveSim 在 Robot 域取得最大增益(+0.89)”,但表中 Robot 列最大增益实为 SDG-SynHuman 的 +1.23(正文随后自己也引用了该值);正文又称 SDG-All“取得最优 Quality 分(72.56)”,但表中 Quality 列最高为 SDG-RobotSim 的 72.72。(3)Human 列在所有变体中无一例外全线下降,基线 85.46 即该列最优——对我们的车载 DMS 项目,这是全表最要紧的信号:即便专为人体构建的 SynHuman 也未能提升 Human 域,作者归因于人体内容的 sim-to-real 差距尤为显著;这支持“合成人体数据须与真实高质量数据混合、放在 mid-training 阶段使用,而非单独微调”的用法。

领域特定的提升。所有 SDG 变体呈现出一个一致的模式:每个来源提升的是*不同的*领域子分,反映了各仿真器独特的内容分布。SDG-DriveSim 在 Robot 域取得最大增益(+0.89),并带来可观的 Common Sense 提升(+0.52),体现了驾驶场景中丰富的结构化动态。SDG-RobotSim 在所有来源中对感知 Quality 的提升最大(+0.26),并适度拉升 Robot 分(+0.23)。SDG-Warehouse 在 Overall、Domain、AV(+0.29)与 Physics(+0.21)上带来广泛的正向变化,同时在 Robot 上保持较强表现(+0.60)。SDG-PhyxSim 贡献了最大的单一领域增益:Industry 提升 +0.85,Robot 提升 +0.80,这得益于其物理接地的工业与操作内容。SDG-SynHuman 是表现最突出的单一来源,取得最佳总体分(79.79),并在 Domain(+0.29)、Common Sense(+0.66)、AV(+0.47)与 Robot(+1.23)上拿到最高的正向增量。以人为中心的合成内容之广度,似乎提供了一种可迁移到多个评测领域的强通用信号。

Sim-to-real 差距与领域间的此消彼长。在所有 SDG 来源中,最一致的退化模式出现在 Human 域得分上:它在每一个变体中都无一例外地下降——从 −0.38(SDG-SynHuman)到 −0.69(SDG-PhyxSim)。值得注意的是,即便是专门由合成人体场景构建的 SDG-SynHuman 也未能挽回这一分数。这提示 sim-to-real 差距在与人相关的视觉内容上尤为突出:当前的仿真器还无法以足够的保真度复现真实人体细微的外观、动作与行为特征,以使这一评测领域受益。更广泛地看,领域专精的来源也可能损害与之正交的类别——SDG-RobotSim 拉低了 AV(−1.03)与 Industry(−1.22),而 SDG-PhyxSim 对物理的侧重伴随着明显的 AV 退化(−1.42)——这凸显了必须混合多个来源,而不能依赖任何单一仿真器。

混合的 SDG-All 取得广泛而均衡的收益。混合全部 SDG 来源(SDG-All)在九项指标中的八项上取得一致的正向变化,仅 Human 仍有残余下滑(−0.47),这与上文观察到的 sim-to-real 差距一致。它取得了最优的 Quality 分(72.56),并在其他所有领域类别上获得一致提升,表明数据多样性能够抑制单一来源的偏置。基于这些结果,我们的最终模型在一个专门的中期训练阶段中,把 SDG 数据集*与真实的高质量视频一起*纳入训练。这一设计使模型既能吸收合成数据中编码的领域特定物理理解,又能保留其在预训练中从真实素材获得的感知保真度与视觉真实感。

附录 D:Cosmos3-Edge LLM 模型训练 Cosmos3-Edge LLM Model Training

Cosmos3-Edge 使用一个从零开始训练的稠密(dense)2B 主干网络。其训练遵循两阶段课程:先预训练,再监督微调(supervised fine-tuning, SFT)。预训练阶段又进一步分为基础预训练和长上下文扩展。整个训练全程使用 BF16 精度;优化器设置见下文。

基础预训练。在基础预训练阶段,我们在来自 Nemotron 预训练语料的共计 15T token 上,以 8,192 token 的序列长度从零训练 2B 的 Edge 主干。该阶段由两个子阶段组成:先在一个广覆盖的数据混合(data mixture)上做通用预训练,随后在一个更高质量的混合上做继续预训练。数据混合在训练过程中热切换(hot-swap):继续预训练从通用预训练的检查点恢复,同时保留优化器状态和学习率调度,因此只有数据混合发生变化。我们使用 AdamW,峰值学习率 1.2×10^{-3},$(\beta_1, \beta_2) = (0.9, 0.95)$,权重衰减 0.1,梯度裁剪范数 1.0。我们采用 warmup-stable-decay(WSD)学习率调度,并让数据混合的切换时点与从稳定(stable)阶段进入衰减(decay)阶段的转换对齐。分词器(tokenizer)与 NVIDIA Nemotron-3 系列模型(NVIDIA, 2025)共用。

长上下文扩展。在长上下文扩展阶段,我们将 Cosmos3-Edge 主干的上下文窗口扩展到 128K token。尽管扩展后的上下文窗口支持 128K token 的训练序列,该阶段的首要目标其实是提升部署时所用的 32K token 序列长度下的鲁棒性与质量。在上下文扩展阶段,我们以 128K 序列长度训练,并将 RoPE base 提高到 1e8。我们使用 1.2×10^{-5} 的常数学习率,长上下文阶段共训练 90B token。该阶段的数据配比为:将预训练混合降采样到 80%,其余 20% 加入长文档问答(QA)数据。

监督微调。我们使用来自 Nemotron-Cascade-2(Yang et al., 2026b)的监督微调数据,其覆盖领域广泛,包括数学、编程、科学、通用对话、指令遵循、工具使用和代码智能体(code-agent)任务。数据集总计包含约 2,600 万条 SFT 样本。我们将这些样本打包(pack)成最长 128K token 的序列,得到约 260 万条打包后的训练样本。模型在单一 SFT 阶段中以全局 batch size 32 训练。我们使用 AdamW 优化器,学习率 2×10^{-5},$(\beta_1, \beta_2) = (0.9, 0.98)$。经验上,模型能力在约 1.7 个 epoch 后达到峰值,对应 14 万个训练步。

我们在一组覆盖推理、科学、指令遵循、长上下文与通用能力的文本基准评测(benchmark)上评估 SFT 模型:HMMT25 Feb(Harvard-MIT Mathematics Tournament, 2025)、GPQA(Rein et al., 2023)、MMLU-Pro(Wang et al., 2024d)、AA-LCR(Artificial Analysis Team, 2025)、IFBench(Pyatkin et al.,

2025a)以及 Scale AI Multi-Challenge(Sirdeshmukh et al., 2025)。对比对象为 Qwen3.5-2B(Qwen Team, 2026b)——一个相同模型规模的强基线。

表 27:文本基准评测结果。在推理、科学、指令遵循与长上下文评测上对比 Cosmos3-Edge 与 Qwen3.5-2B。Cosmos3-Edge 在 HMMT25 Feb 和 GPQA 上大幅提升了数学与科学推理能力,同时在 IFBench 和 AA-LCR 上取得相当的分数。

模型	HMMT25 Feb	GPQA	MMLU Pro	AA-LCR	IFBench (prompt)	Scale AI Multi-Challenge
Qwen3.5-2B	22.9	51.6	**66.5**	**25.6**	41.3	**33.7**
Cosmos3-Edge	**76.3**	**56.4**	62.6	22.8	**43.6**	28.1

如表 27 所示,我们的文本 SFT 模型在数学推理与科学基准(如 HMMT25 Feb 和 GPQA)上大幅超越 Qwen3.5-2B;在指令遵循与长上下文评测(包括 IFBench 与 AA-LCR)上结果相当;但在 MMLU-Pro 等通用领域基准上落后于 Qwen3.5-2B。

> **译注** 按表 27 逐项核对,行文只点名了 MMLU-Pro(62.6 对 66.5)这一处落后,而 Scale AI Multi-Challenge 上的差距其实更大(28.1 对 33.7,相对差约 17%),原文未单独说明,读者宜将其一并计入「通用领域偏弱」的范畴;AA-LCR(22.8 对 25.6)被归为「相当」,也略偏乐观。

附录 E:补充消融实验 Additional Ablation Study

虽然主体实验已经证明了 Cosmos 3 在广泛的理解与生成任务上的有效性,但它们并未完全隔离出各项设计选择各自的贡献。为更好地理解模型能力背后的因素,我们开展了一系列消融实验(ablation),考察关键的架构、数据与训练决策。这些研究探讨了:Reasoner 与 Generator 在 Mixture-of-Transformers 框架内如何相互作用、音频与动作数据等多模态训练信号的影响、时间条件机制的有效性,以及学到的世界与动作表征跨领域的可迁移性。综合来看,这些分析为 Cosmos 3 作为面向 Physical AI 的统一全模态世界模型的设计原则提供了更深入的洞见。

E.1 Reasoner 如何让 Generator 受益

在本研究中,我们考察 Reasoner 如何让 Generator 模型受益。我们基于 Cosmos3-Nano 架构训练两个模型:一个以 `Qwen3-VL-8B` 作为理解塔(understanding tower),另一个使用我们的 Cosmos3-Nano Reasoner。两种情况下,Generator 塔都从零开始训练。在本消融中,我们将训练限制在 256p 与 480p 分辨率、片段长度 0–200 帧,序列长度设为 25K。沿用大规模训练的做法,我们采用图像–视频联合训练。两个模型均在 256 块 GPU 上训练 9 万次迭代。下面报告两者的 PAIBench 分数。

表 28:理解塔消融。我们在 PAIBench T2V 与 I2V 上报告两个变体的领域分与质量分,二者唯一的区别是初始化理解塔所用的预训练模型,而 Generator 塔均从零训练:(1)Cosmos 3 Reason (2)Qwen3-VL。主体一致性与背景一致性(Subject/Background consistency)为 I2V 特有指标,不适用于 T2V。用 Cosmos3 Reasoner 初始化理解塔,在 Physical AI 各领域上取得比 Qwe VL 变体更好的领域分。

评测	理解塔	汇总			领域							质量					I2V	
		Overall	Domain	Quality	C.S.	AV	Rob.	Ind.	Hum.	Phy.	Subj.	Bg.	Motion	Aesth.	Imag.	Cons.	Subj.	Bg
T2V	Cosmos3 Reasoner	**74.3**	**75.7**	**73.0**	**81.2**	**54.9**	**71.3**	**78.4**	**76.2**	**89.2**	**96.0**	**96.8**	**99.4**	**55.2**	71.4	**19.1**	–	
	Qwen-3 VL	73.3	73.7	**73.0**	80.5	52.6	66.5	77.4	74.0	88.7	95.8	96.6	**99.4**	55.0	**72.2**	19.0	–	
I2V	Cosmos3 Reasoner	**79.4**	**80.8**	**78.1**	89.0	59.4	**77.0**	**84.8**	**79.3**	**91.6**	**92.4**	**94.7**	**99.4**	**53.2**	68.7	**20.2**	**98.0**	**98**
	Qwen-3 VL	79.0	80.0	78.0	**89.7**	**59.8**	74.0	84.0	78.3	91.4	92.0	94.5	**99.4**	53.1	**69.3**	20.1	97.9	97

如表 28 所示,在理解塔中用我们的 Cosmos3-Nano Reasoner 替换 Qwen3-VL-8B 后,领域分获得一致提升,在 Physical AI 各领域尤其明显。在 T2V 上,Reasoner 将总体 Domain 分从 73.7 提升到 75.7,增益最集中在以物理为基础的类别:Robot(+4.8,66.5 → 71.3)、Physics(+0.5,88.7 → 89.2)、AV(+2.3,52.6 → 54.9)、Industry(+1.0,77.4 → 78.4)。I2V 上呈现类似模式,Domain 分从 80.0 升至 80.8,驱动因素为 Common sense(+0.7)、Industry(+0.8)、Human(+1.0)与 Physics(+0.2)的增益。这些结果表明,Reasoner 为 Physical AI 领域提供了更好的嵌入,供 Generator 学习。两个模型的质量分相当。

> **译注** 此处 I2V 的归因与表 28 数据对不上——表中 I2V 的 Common sense 是 Reasoner 89.0、Qwen3-VL 89.7,即 Reasoner 低 0.7 分而非「+0.7」;I2V Domain 分上升的真正最大驱动其实是行文未提的 Robot(74.0 → 77.0,+3.0),AV 也是微降(59.8 → 59.4)。「+0.7」疑为把降幅误写为增幅。T2V 部分的逐项数字经核对无误。

E.2 帧率控制方式的选择

我们研究两种互补的、把生成条件化到目标帧率上的机制:(i)MRoPE 帧率调制(MRoPE FPS modulation),即用目标 FPS 缩放统一 3D MRoPE 的时间轴;(ii)文本控制(Text Control),即把目标时长与 FPS 作为自然语言文本注入结构化 JSON 描述(caption)中。在本消融中,我们将训练限制在 256p 与 480p 分辨率、片段长度 0–200 帧,序列长度设为 25K。沿用大规模训练的做法,采用图像–视频联合训练。我们训练 4 个模型,各在 128 块 GPU 上迭代 13 万次,唯一变量是各机制是否启用:Base(无控制)、Text Control(仅文本)、MRoPE FPS Modulation(仅 MRoPE)、Text Control + MRoPE FPS Modulation(两者皆有)。我们构建了一个源时长与 FPS 已知的评测集,10、15、24、30 FPS(容差 ±2 FPS)每档约 100 条视频。我们以这些评测集视频的结构化描述作为提示词,在 480p(16:9)分辨率下每条提示词用 3 个随机种子生成样本,每条提示词产出一段 5 秒片段。

我们对每段片段评两项分:视频质量(Video Quality, VQ;DOVER 感知质量分(Wu et al., 2023a))与动态程度(Dynamic Degree, DD;0–1 尺度的运动存在度(Huang et al., 2024))。基于每条提示词(p)三个种子的 DD 分数,我们计算一个归一化的运动控制(motion control, MC)项:

$$MC = \mathbb{E}_p[\, var_p \,/\, (var_p + mean_p{}^2) \,] \quad (10)$$

MC 反映控制机制的鲁棒性,即在每种控制设置下、每条提示词内部的变异程度。我们将其组合成运动保真度(Motion Fidelity)分数:

$$MF = (1 - |DD - DD_{ref}|)(1 - MC) \ ,(11)$$

其中 DD_{ref} 是真实视频参考集在每个帧率档内的 DD 均值。最终的综合得分(Composite Score)是视频质量与运动保真度的乘积,反映对运动幅度的遵循程度、且不以牺牲生成视频的感知质量为代价。计算方式为 $\mathbb{E}_{\text{FPS-band}}[(\text{VQ}) \cdot \text{MF}]$,即先在每个 FPS 档内计算,再对各档取平均。

表 29 报告了四个 FPS 档的平均结果。两种机制单独使用都优于 Base;单用 MRoPE 帧率调制的增益大于文本控制(综合分 +1.12 对 +0.77)。二者结合得到最好的综合分(较 Base +1.30)。四种设置的 VQ 都落在 0.2 分的窗口内,说明增益集中在运动保真度而非视频感知质量上,即这些控制主要改善的是时间行为。基于这些结果,我们为 Cosmos 3 Generator 采用 Text Control + MRoPE FPS Modulation 配置。

表 29:Cosmos3-Nano 上的帧率控制消融。分数为四个 FPS 档(10、15、24、30)的平均。VQ 为 DOVER 视频质量分;MF 为运动保真度(0–1);Composite = $\mathbb{E}_{\text{band}}[(\text{VQ}) \cdot \text{MF}]$。加粗表示该列最优。Text Control + MRoPE FPS Modulation 是性能最好的控制设置,既能遵循参考 FPS 档的运动,又保持了感知质量。

模型	帧率控制设置	Avg. VQ (↑)	Avg. MF (↑)	Avg. Composite (↑)
Cosmos3-Nano	Base(无控制)	12.89	0.6626	8.51
Cosmos3-Nano	Text Control	12.99	0.7169	9.28
Cosmos3-Nano	MRoPE FPS Modulation	**13.03**	0.7409	9.63
Cosmos3-Nano	**Text Control + MRoPE FPS Modulation**	12.84	**0.7649**	**9.81**

E.3 预训练中的音频数据

我们考察在继续预训练中加入音频对视频指标的影响。从同一个预训练检查点出发,我们在预训练数据集上训练两个 Cosmos3-Nano 变体:一个只用视频数据,另一个用视频–音频联合数据。这些消融在 128 块 GPU 上运行 2 万次迭代,训练限制在 256p 与 480p 分辨率。

表 30 的结果显示,不带音频的继续训练在 T2V 与 I2V 分数上都有所下降,这表明视频–音频联合预训练不会损害视频生成质量,并且即便只在以视频为中心的指标上评估,也可能带来适度的收益。

表 30:预训练中引入音频数据的影响。我们报告两个 Generator 变体继续预训练后的 PAIBench T2V 与 I2V 分数,二者唯一的区别是训练中是否使用音频;视频数据在两组实验中完全相同。

评测	变体	汇总			领域						质量					I2V	
		Overall	Domain	Quality	C.S.	AV	Rob.	Ind.	Hum.	Phy.	Subj.	Bg.	Motion	Aesth.	Imag.	Cons.	Subj.	Bg.
T2V	不带音频	78.6	83.8	**73.4**	90.7	64.7	81.9	84.4	82.7	**94.8**	**95.9**	**97.1**	**99.5**	**57.3**	**70.6**	19.8	–	–
	带音频	**79.1**	**85.0**	73.2	**91.5**	**67.9**	**83.8**	**86.1**	**83.7**	93.8	95.5	96.9	**99.5**	57.1	70.2	**19.9**	–	–
I2V	不带音频	81.7	85.1	**78.4**	93.2	67.5	82.0	86.0	83.2	**95.1**	**93.3**	**95.1**	**99.5**	**55.0**	67.6	**20.4**	**98.4**	**98.3**
	带音频	**82.2**	**85.9**	**78.4**	**93.6**	**68.6**	**84.2**	**86.5**	**84.3**	94.8	92.9	94.9	99.4	**55.0**	**67.9**	**20.4**	98.2	98.2

> **译注** 逐格核对表 30 可见,「带音频」的优势几乎全部来自 Domain 侧(T2V +1.2、I2V +0.8),而 Quality 侧持平或略降(T2V 73.4 → 73.2),且 Physics 子项两个评测下都是下降的(94.8 → 93.8、95.1 → 94.8),主体/背景一致性也略降。原文「不会损害、可能有适度收益」的对冲表述与数据相符,但读者不应把它读成全面无代价的提升。

E.4 动作模式之间的协同效应

如 §2.2.2 所述、图 4 所总结,Cosmos 3 支持三种动作生成模式(action mode):前向动力学(forward dynamics, FD)、逆向动力学(inverse dynamics, ID)与视频–动作联合预测(策略,policy)。我们消融的问题是:单个联合动作模型能否在这些模式之间共享有用的结构。本实验使用 PushT 数据集与 Cosmos3-Edge。我们将三个单模式动作检查点各训练 2K 步,与一个联合 FD/ID/policy 检查点训练 6K 步相比较,从而保证每种模式获得相同的优化步数。FD 报告 PSNR,ID 报告 MSE;policy 模式报告 T 形块与目标区域的策略覆盖率(policy coverage ratio),在 50 个不同初始化上聚合,每个起点做 10 次 rollout。结果汇总于表 31。

表 31:PushT 动作模式协同效应。我们将各训练 2K 步的单模式 FD、ID、policy 检查点,与训练 6K 步的单个联合 FD/ID/policy 检查点相比较。ID MSE 越低越好;FD PSNR 与策略覆盖率越高越好。加粗表示每列中更优的值。

训练设置	FD PSNR ↑	ID MSE ↓	策略覆盖率 ↑
2K 单模式	**27.13**	1.11×10^{-3}	74.1%
6K 联合 FD/ID/policy	26.22	$\mathbf{3.09 \times 10^{-4}}$	**77.3%**

联合 FD/ID/policy 检查点在改善动作侧指标的同时,保持了相当的前向动力学质量。与单模式检查点相比,ID MSE 从 1.11×10^{-3} 降至 3.09×10^{-4},相对降低 72%;策略覆盖率从 74.1% 升至 77.3%。FD PSNR 从 27.13 降到 26.22,表明在重建保真度上存在适度折衷。总体而言,在相同的每模式优化预算下,联合检查点取得了最好的策略覆盖率与 ID 精度,这提示各动作模式之间共享着有用的结构——尽管 FD 质量从单模式特化中略有受益。

E.5 视频-动作一致性

除了 §6.2.5 报告的闭环成功率之外,我们还在 RoboLab 上评估 `Cosmos3-Nano-Policy-DROID` 联合预测的视频流与动作流之间的对齐程度(即视频-动作一致性,video-action consistency)。Cosmos3-Nano-Policy-DROID 只在 DROID 上训练,因此 RoboLab 提供了一个用于检验该一致性的留出(held-out)环境。对

每个预测出的动作块(action chunk),我们在 RoboLab 仿真器中执行同一动作块,计算模型预测视频与仿真器 rollout 结果之间的 PSNR,再对左侧与腕部两个相机视角在各动作块上聚合分数。左侧第三人称视角达到 23.19 dB,与我们的机器人前向动力学模型在 DROID 上取得的 PSNR 水平相当(表 18)。腕部(手眼式,eye-in-hand)视角更具挑战性,因为相机随末端执行器一起运动,模型必须推断新暴露出的内容;即便如此,它仍达到 17.33 dB。这些结果表明预测动作与预测视频之间具有很强的一致性。图 37 将这种一致性可视化:预测帧与从相同状态初始化的仿真器 rollout 的对应帧高度吻合。

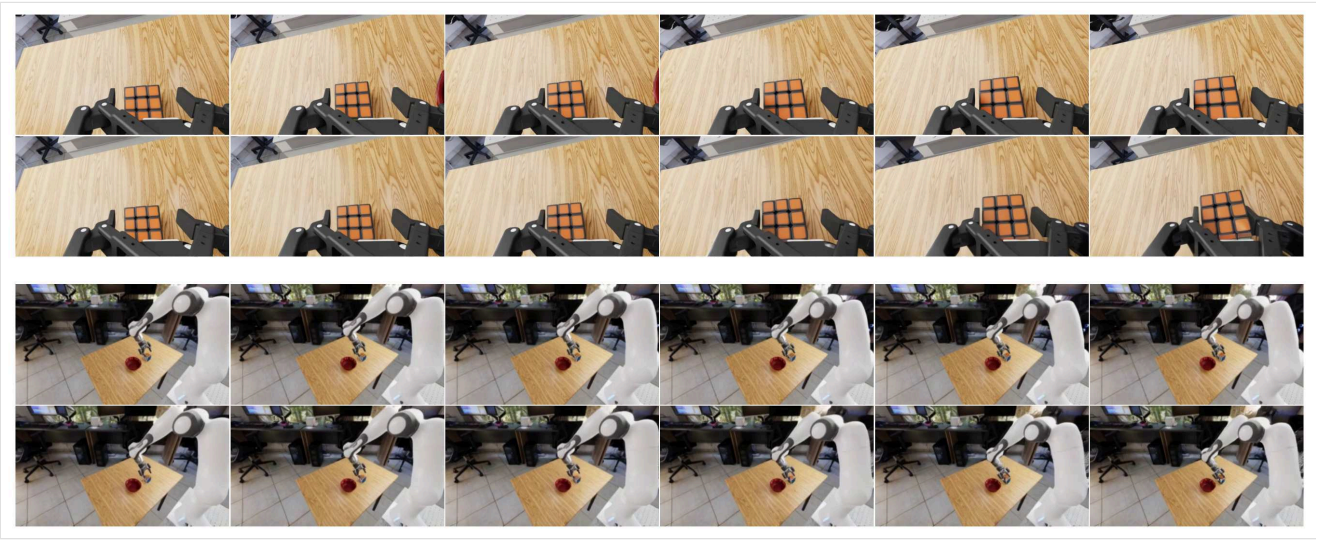


图 37:RoboLab 上预测视频与仿真器 rollout 的对比。对腕部与左侧相机,Sim Env 一行展示从相同初始状态出发、在 RoboLab 仿真器中执行预测动作块所录得的视频;Pred 一行展示 Cosmos3-Nano-Policy-DROID 与该动作块联合预测出的视频。可以观察到预测视频与仿真器 rollout 高度吻合。

附录 F:Cosmos-HumanEval 基准评测(Cosmos-HUE) Cosmos-HumanEval Benchmark (Cosmos-HUE)

Cosmos-HumanEval(Cosmos-HUE,简称 HUE)是主结果部分引入的一个视频生成人工评测(human evaluation)基准。HUE 通过一条视觉-语言模型(Vision-Language Model, VLM)流水线为每条视频生成原子化二元问题,提供以人为参照的评测信号,并沿四个 HUE 维度打分:语义对齐(Semantic Alignment)、物理规律(Physical Laws)、几何推理(Geometric Reasoning)、视觉完整性(Visual Integrity)。本附录在该概览基础上补充:正式的二元计分方案、标注协议与可靠性估计,以及分维度、分领域的文生视频(Text-to-Video, T2V)与图生视频(Image-to-Video, I2V)排行榜。

二元回答模式与计分。每个第 3 层(Layer 3)问题的回答为「是」(Yes,满足判据)、「否」(No,明确违反)或「不明确」(Unclear;例如相关区域被遮挡、运动太快无法判断,或视频根本没有呈现问题所预设的事件——比如提示词要求右转但前车始终未转弯,那么关于「平滑弯弧」的后续问题就没有明确的是或否)。*「不明确」按「否」处理*:模型不会因未能生成提示词要求的动作而得到奖励,而标注员(annotator)不自信的判断也保守地计为对模型不利。具体而言,「不明确」不计入「是」的分子、但保留在分母中,因此一条大部分回答为「不明确」的视频,不可能比拥有同样多明确「是」回答的视频得分更高。设 $\mathcal{Q}_v$ 表示对视频 v 发出的问题集合,则单视频 HUE 分数为:

$$HUE(v) = \left(\sum_{q \in \mathcal{Q}_v} \mathbb{1}[ans(v, q) = Yes] / |\mathcal{Q}_v| \right) \times 100\% \quad (12)$$

由于「是」永远是期望的结果,HUE(v) ∈ [0, 100]%。模型级分数在 V 条测试视频上聚合:

$$HUE(m) = \left(\sum_{v=1}^{V} \sum_{q \in \mathcal{Q}_v} \mathbb{1}[ans(v, q) = Yes] / \sum_{v=1}^{V} |\mathcal{Q}_v| \right) \times 100\% \text{ 。} \quad (13)$$

这一总平均对每个已回答的(视频, 问题)观测等权计入,避免了恰好被分到较少适用问题的视频造成的分数虚高。维度级分数的计算方式相同,只是把求和限制在属于各维度的问题上。

标注协议。每条视频经上述三层流水线最多获得 16 个原子化二元问题,每个(视频, 问题)对由*两名*人工标注员独立评定。若两名标注员一致,共识答案即被记录为规范(canonical)回答;若不一致,该问题升级给第三位质量控制(quality-control, QC)审核员,其答案为最终裁定。这套「双人评定 + QC 裁决」的工作流为每个(视频, 问题)对产出单一规范答案,限制了任何单个标注员的影响上限,并产生可审计的回答流,供公式 12–13 的计分公式使用。

问题设计与迭代。原子化问题的编写与打磨面向两个目标:**低方差**(同一模型在同样的提示词上重跑,得分应相近)与 **GT 饱和**(与提示词配对的真实视频应得到 100% 或接近 100% 的分数)。我们用这两条标准评估候选问题并迭代题库,修改或剔除任何不达标的条目。这一过程仍在进行中:GT 分数目前仍低于 100%(表 32 与表 33),因此题库的打磨还在继续。

可靠性与置信区间。当前的 T2V 评测池从 PAIBench-G 中采样 100 条提示词,每条提示词做 5 次随机种子生成,每条视频最多 20 个问题,因此每个模型检查点最多产生 10,000 个二元观测。将每个观测视为潜在「是」率 p = HUE(m)/100 的独立伯努利试验,则模型级 HUE 分数(以百分点计)的 95% 置信区间为 $\text{HUE(m)} \pm 100 \cdot 1.96 \sqrt{p(1-p)/N}$,其中 N 为观测数;实践中,在表 32 的分数量级上,头部模型的 CI_{95} 宽度约为 ±0.6 分。相比李克特量表(Likert-scale)打分——标注员必须综合多项观察、加以权衡、再映射到一个任意尺度上,这三步操作每一步都引入噪声——原子化二元判断降低了标注员内部方差。「真实视频 GT」一行在 T2V 提示词上得 93.6 分、在 I2V 提示词上得 94.4 分,这反映了:(i)尽管经过人工复核,采样的 PAIBench-G 配对中仍残留少量提示词–视频不匹配;(ii)自动生成的问题集并不完美,过于严苛、含糊或偏题的问题可能让真实视频也得到「否」或「不明确」。我们将 GT 到 100% 的差距视为持续打磨题库的北极星指标。

> **译注** 标注协议一段说每条视频「最多 16 个」问题,而本段计算观测上限时用的是「每条视频最多 20 个问题」(100 × 5 × 20 = 10,000),两处口径在原文中不一致,疑为题库迭代

> 过程中数字未同步。

T2V 排行榜。表 32 报告了我们的模型阵容与最强外部 T2V 基线的当前 T2V 结果。Veo-3.1 总分领先(91.3),Seedance-1.5-Pro 居次(90.0);Cosmos3-Super 以 89.3 成为最好的开源 Generator,领先 Wan2.2-A14B(88.2)与 Cosmos3-Nano(87.6)。分维度来看,情形对 Cosmos3-Super 更有利:它在 12 个坐标轴中的 9 个上是最佳开源模型,并在 AV(87.7)与 Physics(91.5)上击败了包括闭源在内的所有 Generator。最强 Generator 与真实视频(Real video GT,93.6)的差距总体约 2.3 分,为下一轮迭代留下了可观的提升空间。

I2V 排行榜。表 33 报告了平行的 I2V 评测。Veo-3.1 总分领先(89.7),Cosmos3-Super 仅落后 0.1 分(89.6);Cosmos3-Nano(88.6)险胜 Wan2.2-A14B(88.4)与 Seedance-1.5-Pro(87.6)。Cosmos3-Super 在视觉完整性(94.2)、Robotics(91.1)与 Miscellaneous(94.8)上完胜所有 Generator,并在语义对齐上与 Veo-3.1 并列第一(均为 90.3);Cosmos3-Nano 在 AV(87.6)上完胜;Wan2.2-A14B 在 Physics(91.9)上夺魁。整个 I2V 榜单在各 Generator 之间跨度约 9 分。

表 32:Cosmos HUE T2V 排行榜。分维度与分领域拆分(%;越高越好)。左块:总体 HUE 加四个维度(语义对齐、物理规律、几何推理、视觉完整性)。右块(分隔线之后):限定在每个 PAIBench-G 提示词领域内的总体 HUE(C.S.、AV、Rob.、Ind.、Hum.、Phy.、Misc. 分别缩写 Common Sense、Autonomous Vehicle、Robotics、Industry、Human、Physics、Miscellaneous)。Cosmos 3 Generator(我们的)以绿色底纹标出;Real video GT 为上界参照。加粗为该列最优,下划线为次优。Cosmos3-Super 在 12 个坐标轴中的 9 个上领跑开源阵营,并在 AV(87.7)与 Physics(91.5)上击败包括闭源在内的所有 Generator。

模型	类型	Overall	Sem. Align.	Phys. Laws	Geo. Reas.	Vis. Integ.	C.S.	AV	Rob.	Ind.	Hum.	Phy.	Misc.
Real video GT (PAI-Bench)	—	*93.6*	*95.0*	*90.5*	*94.1*	*94.8*	*92.0*	*93.7*	*94.9*	*94.4*	*93.2*	*93.1*	*93.6*
Cosmos3-Super	开源	89.3	92.4	85.4	86.6	93.7	89.5	**87.7**	88.4	89.4	88.6	**91.5**	93.0
Cosmos3-Nano	开源	87.6	91.5	83.6	84.1	92.5	89.5	87.0	86.6	85.8	86.2	87.9	94.0
Wan2.2-A14B	开源	88.2	92.1	83.4	85.6	92.6	87.8	84.9	83.7	91.1	89.9	87.5	93.4
HunyuanVideo-1.5	开源	86.5	88.4	81.8	83.6	93.6	88.3	81.0	81.3	88.0	87.9	87.4	93.3
Wan2.1-14B	开源	84.0	87.7	77.1	78.5	92.7	85.3	79.1	78.2	87.0	85.1	85.8	90.7
Wan2.2-5B	开源	80.8	86.9	73.8	73.6	89.9	83.4	73.5	76.0	84.4	81.3	82.8	87.9
Cosmos-Predict2.5-14B	开源	82.1	88.1	74.6	75.8	90.7	81.3	84.4	76.8	82.4	81.3	84.5	90.2
Cosmos-Predict2.5-2B	开源	81.8	88.1	74.0	75.1	90.6	81.5	81.8	79.5	82.7	80.3	82.7	89.8
Veo-3.1	闭源	**91.3**	**94.3**	87.7	**91.0**	**93.9**	**92.7**	85.6	**91.5**	**94.7**	90.3	90.4	**95.8**
Seedance-1.5-Pro	闭源	90.0	91.2	**88.4**	89.8	92.8	90.5	83.6	89.0	92.9	**90.7**	91.4	91.7

表 33:Cosmos HUE I2V 排行榜。分维度与分领域拆分(%;协议与表 32 相同,但带图像条件)。版式、领域缩写、底纹以及加粗/下划线约定与表 32 一致。

模型	类型	Overall	Sem. Align.	Phys. Laws	Geo. Reas.	Vis. Integ.	C.S.	AV	Rob.	Ind.	Hum.	Phy.	Misc.
Real video GT (PAI-Bench)	—	*94.4*	*94.8*	*93.2*	*95.1*	*95.4*	*94.2*	*96.0*	*94.9*	*93.4*	*92.1*	*96.3*	*98.1*
Cosmos3-Super	开源	89.6	**90.3**	87.5	87.0	**94.2**	90.7	86.2	**91.1**	89.6	87.1	91.5	**94.8**
Cosmos3-Nano	开源	88.6	89.2	86.4	86.3	93.5	90.6	**87.6**	90.6	88.0	84.7	91.0	93.9
Wan2.2-A14B	开源	88.4	88.6	86.0	85.7	93.5	90.1	84.7	84.8	92.0	86.5	**91.9**	92.5
HunyuanVideo-1.5	开源	85.6	87.1	80.8	83.0	92.5	90.0	81.7	77.2	89.6	85.2	89.5	91.8
Wan2.1-14B	开源	83.9	85.0	80.2	80.6	91.2	88.1	76.8	74.8	89.9	84.3	85.8	93.2
Wan2.2-5B	开源	80.4	83.4	74.6	75.7	89.5	86.1	74.3	68.6	88.4	79.6	84.8	90.5
Cosmos-Predict2.5-14B	开源	83.0	85.0	77.9	77.5	92.4	88.1	85.2	78.4	83.0	78.5	84.8	93.2
Cosmos-Predict2.5-2B	开源	82.6	86.1	77.3	78.2	90.2	85.6	86.0	79.0	85.6	77.1	85.1	92.4
Veo-3.1	闭源	**89.7**	**90.3**	**87.7**	**89.2**	93.2	**92.3**	86.0	88.6	**93.4**	**87.2**	91.2	94.3
Seedance-1.5-Pro	闭源	87.6	87.7	85.4	86.5	91.8	89.5	82.9	85.7	90.6	85.7	90.6	93.0

附录 G 与参考文献 Appendix G & References

附录 G(贡献者与致谢,原文 p113–117):全体作者与致谢名单,按译稿惯例不译,请查阅原文。

参考文献(原文 p118–139):文献列表保留原文不译。正文中的引用格式 (Xxx et al., 2026) 与原文一一对应,可按作者-年份在原 PDF 参考文献部分检索。

译者解读:对我们 IMS/DMS 合成数据管线意味着什么 (非原文内容)

本章面向本团队当前工程现状写作:我们用 `Cosmos-Transfer2.5-2B`(edge 结构控制 + NIR 域后训练)把 Blender 素模渲染转成真实感 NIR 座舱数据,眼部用 `eye_composite` 投影回贴保真值;Cosmos3-Nano 权重已在超算下载,迁移与否待实测(参见 docs/research/2026-07-19-Cosmos3调研.md)。以下按"边界→产物→裂缝→增量→路线→前提"展开。

一、这份报告不能给我们什么(先划边界)

- **训练本身不可复现**。预训练用 1024–4096 块 GB200、数十万亿词元(§4.2、表 8),这条线与我们无关;我们只消费权重。
- **全模态大半与 IMS 当前需求正交**。音频输入上限 0.5 秒、动作接口面向机器人/AV 具身(§2.1),座舱行为建模能否映射进去是另一个立项,本轮不并入。
- **没有 transfer 专用后训练配方**。图 8 的训练课程里,视频迁移(4M 样本)只出现在中期训练列,后训练列为空——官方自己也没有 transfer 专科检查点;in-context 结构控制是 base 模型能力。报告通篇没有给"像 Transfer2.5 `posttrain_edge_example` 那样"的域适配配方,这仍是我们最大的未知数。
- **没有红外/NIR**。全部数据与评测都在可见光域,NIR 域适配的证据要靠我们自己实测,报告帮不上。
- **表 26 的负结果不能直接外推到我们身上,但方向要记住**。该消融评的是"合成数据能否提升*生成模型*的 T2V 质量",而我们做的是反向工程("用生成模型把合成数据洗真实")。它证伪的不是我们的管线,而是"纯合成数据单独微调"这条路。

二、能直接取走的产物

- **表 21 推理参数表(§6.3.1)**。各控制模态的 steps / guidance_scale / control_guidance / flow_shift 官方推荐值,edge 档(guidance 3.0 / control 1.5 / shift 10.0)与我们在 2.5 上摸出的配方近乎平移——首次 Nano transfer 实测直接从这组值起步,不用再扫参。
- **附录 B 全套 prompt 模板**。Reasoner 上采样模板(B.1)、I2V 两阶段模板(B.3)、视频迁移系统前缀(B.4)、约 4 页的结构化负面提示词(B.6)在译稿中完整保留英文原文——我们用 Claude 充当 upsampler 时,JSON 字段契约照抄即可,这是把"prompt 需升采样成 JSON"这个迁移成本降到最低的现成材料。
- **附录 A 结构化 caption schema**。图像/视频打标的字段设计(镜头语言、人物属性、计数敏感属性)可直接借鉴到我们训练数据的打标规范。
- **表 26 的用法结论(全报告对我们最值钱的一句话)**。合成数据要*与真实高质量数据混合*、*放在 mid-training 式阶段*使用,单独用合成数据微调连 Human 域自己都救不回。如果我们后续对 Cosmos 3 做 NIR 域适配,数据配比应当是"合成 + 真实 NIR 混合",而不是纯合成 SFT。
- **附录 C.4 SDG-SynHuman 流水线细节**。数字人合成数据的相机运动配置、资产来源、逐文件元数据布局与我们的素模渲染管线同构,可逐项对照查缺(尤其元数据落盘规范)。

三、报告自身的裂缝(读表所得,正文不会告诉你)

> **正文叙述与表格数据多处不自洽,引用其结论前务必回表核对:**
>
> - 表 26 两处:正文称 DriveSim 在 Robot 域增益最大(+0.89),表中实为 SynHuman +1.23;称 SDG-All 取得最优 Quality(72.56),表中最高为 RobotSim 72.72。
> - 表 19:正文称 DreamZero specific 25.2%,表中为 23.9;Complex×Vague 一格最终策略(4.1)反被中间产物 PT-init(7.1)超越。
> - 表 8 表注 GPU 数恰为正文 2 倍(2048/4096 vs 1024/2048),疑为超级芯片与 GPU die 口径混用;表 11 Phys 列加粗值(89.54)并非该列最高(Gemini 3 Pro Image 89.74);表 14 正文 88.5 vs 表 88.6。
>
> **榜单叙事有水分:**"RoboArena 最佳策略"快照仅基于 20 次 A/B 评测(对手 157–799 次);"最佳开源 T2I/I2V"限定*开放权重*且 T2I 依赖 agentic 上采样工作流;Cosmos-HUE 上两列最优其实是闭源 Veo-3.1;Physics-IQ 被 Sora2/Magi-1 反超;基线统一用适配 Cosmos 训练分布的 Claude 提示词改写重测,跨文献不可直比。结论:Cosmos 3 是强的开源基座,但"全面 SOTA"要打折听。
>
> **官方承认且未解决:**Human 域 sim-to-real 差距(表 26 全线负增益)、长/高分辨率输出的时序不稳与动力学瑕疵(§9 局限)。*不能假设换 Cosmos 3 就自动解决我们的眼部/细节问题。*

四、我们相对报告的增量

> - **真值保持是我们有、它没有的约束**。报告的 SDG 只为生成模型供数据,不承诺像素级标注对齐;我们的管线以 gaze/姿态真值为纲(edge 控制保几何 + eye_composite 把带精确 gaze 的素模眼睛回贴),这条能力与 Cosmos 代际无关,迁移后照样成立,而且正好补上报告承认的"人体域 sim-to-real 差距"中最致命的眼部部分。
> - **我们有真实 NIR 数据可做域适配**。表 26 指出的正确用法(合成+真实混合训练)我们具备全部原料;报告没有 NIR,我们反而是在它的空白域上工作。
> - **场景窄是优势**。座舱固定机位、近景人脸、96 帧短片——报告承认的长时序不稳、开放场景动力学瑕疵对我们的影响天然受限。

五、落地路线(按报告自身证据排性价比)

1. **先做零训练对比(本周可完成,已具备全部条件)**。Nano + 我们现有 A96 edge 控制视频一次推理,参数照表 21 edge 档,prompt 用附录 B.1/B.4 模板经 Claude 升采样。对比 2.5 后训练产出:脸/皮肤/眼部三点评审。依据:in-context 控制是 base 能力(§2.2.2),16B 底座 + 更长上下文对结构保持的先验值得一次廉价验证。

2. **若 NIR 域感不足,做混合数据域适配**。用 Cosmos Framework 的 Vision SFT recipe(8×H100 配置,我们正好有),数据按表 26 结论配成"我们的真实 NIR clip + 合成素模对"混合,不做纯合成微调。前提是第 1 步证明 base 值得投入。
3. **eye_composite 保留不动**。与代际无关;第 1 步对比时同样套用,单独评"Cosmos 3 是否缓解 VAE 糊眼"(预期:32×32 空间压缩比 2.5 的 VAE 更狠——见 §2.1.1 译注——糊眼大概率仍在,回贴照旧)。
4. **观望项单独立项**。Reasoner 的物理合理性判定/2D grounding 用于合成数据 QA、音频联合合成(受 0.5s 输入限制)——都不并入本轮。

六、必须先确认的 go/no-go 前提

- **transfer 在 diffusers 的实际可用性**:主 task-pipeline 页仍标 "coming soon",modular pipeline 有 edge 示例——先跑通 `Cosmos3OmniModularPipeline` 冒烟再谈其他。
- **guardrail 必须能关**。DMS 数据全是人脸,脸部模糊 guardrail(`enable_safety_checker=False`)若在服务栈某层关不掉即 no-go。
- **Nano 16B 实测时延/显存**:121 帧@720p 单条推理成本 vs 2.5-2B,决定批量生产是否可行;Super 64B 仅在 Nano 质量不够时评估(8×H100 张量并行我们装得下,是吞吐问题不是容量问题)。
- **Framework Vision SFT 的数据格式是否兼容控制条件**(带 edge 视频的样本对),决定第 2 步配方是否存在现成路径。
- **新环境隔离**:diffusers(新)+ cosmos_framework + 可选 guardrail 是一套与 2.5 venv 分开的环境,超算上 HF 缓存必须指向大盘(`HF_HUB_CACHE=$S/hfcache`,/HOME 100G 已满)。

跨论文合成点:GazeGene(2026-07-15 精读)证明眼部几何真值合成可行,本报告表 26 证明人体域纯合成微调负收益、混真实数据是正解,两者合起来正好为我们"素模真值 + Cosmos 洗真实 + 眼部回贴"的三段式管线提供了互补证据——这条结论不在任何一篇论文里。