HESAI | KOSMO

空间智能白皮书

从数字世界到物理世界

构建AI与物理世界的3D桥梁



CONTENTS目录

执行摘要 02

01 空间智能是什么 04

1.1 空间智能的定义 05

1.2 物理AI向空间智能的演进 05

02 空间智能：物理AI演进的必经之路 07

2.1 空间智能：人类智能与人工智能演进的底层基础与当前AI发展的根本局限 08

03 空间智能产业：三维空间数据是发展瓶颈 12

3.1 空间智能产业发展现状 13

3.2 当前产业发展的瓶颈：三维空间数据的匮乏 21

04 Kosmo：空间智能的基础设施 23

4.1 Kosmo：三维世界数据引擎 24

4.2 Kosmo如何解决问题 27

05 小结 43

参考文献 46

© 2026 上海禾赛科技有限公司

HESAI | KOSMO

EXECUTIVE SUMMARY

执行摘要



过去十年，AI完成了从“读懂和生成语言”到“初步理解并推演世界”的第一次大跨越——走向物理AI。World Labs联合创始人兼首席执行官（拟任 AMD 执行副总裁兼首席科学家）李飞飞将下一阶段的跃迁概括为“从Words到Worlds[1]”——AI将从处理文本、图像和代码等数字信息，走向能够感知、理解、推理并与三维世界交互的空间智能。今天基于海量文本、图像和视频训练的大模型，已经在数字世界中展现出强大的语言理解、内容生成和模拟能力；但当物理AI真正走出屏幕，进入家庭、商业、道路和工厂时，它面对的将不再是“生成一个合理答案”，而是一个完全不同的问题——“如何在真实世界里做出正确动作”。

我们认为，当前空间智能相关能力正沿多个赛道落地，孕育独立的商业价值池，主要包括世界模型与生成式AI、具身智能、影视娱乐、文旅文博与泛测绘。首先，**在世界模型与生成式AI及具身智能赛道**，空间智能为机器人预训练、后训练及效果验证、世界模型与生成式AI的搭建，提供真实三维空间数据，让AI在执行物理动作前，就能理解并模拟、预测现实世界。摩根士丹利预测，2050年人形机器人市场规模将达到约7.5万亿美元[2]。其次，**在影视娱乐赛道**，空间智能正落地院线影视、AI生成影片以及广告、音乐、游戏等内容生产环节。通过构建稳定三维场景与空间关系，该技术能够提升跨镜头画面一致性、镜头可控性与数字资产复用效率，全面提升消费者的媒体体验，并有望赋能规模约2.8万亿美元的全球媒体与娱乐产业[3]。**在文旅文博与泛测绘赛道**，空间智能已应用于沉浸式体验打造、文化遗产三维数字化、场馆数字孪生与智慧运营，实现建筑、文物与场馆的高精度存档、动态更新与反复复用，并催化出立体导航的全新应用。借助真实空间数字化、3D重建与数字资产建设能力，空间智能有望推动文博机构由单点数字化应用向系统化数字资产建设与长期运营转型，提升文化遗产数据的持续积累、跨场景复用与管理效率。典型案例如大英博物馆。其于2023年宣布，计划在五年内新增或完善约240万条藏品线上记录[4]；其后发布的年报显示，整个藏品记录与数字化计划的预算为1000万英镑[5]。

KEY JUDGMENT

当前产业的核心矛盾非常突出：支撑行业生态基础的三维空间数据仍极度稀缺，行业普遍陷入“模型超前、数据供给滞后”的状态。物理AI进步的核心限制从来不只是电力、算力、存储或模型，而是高质量的三维空间数据。

这正是本白皮书的核心判断：空间智能，正是构建AI与物理世界的3D桥梁。谁能持续稳定，且规模化地提供高质量、多模态的三维空间数据，谁就掌握了下一阶段空间智能的话语权，谁就有机会撬动数万亿美元级的新市场。

© 2026 上海禾赛科技有限公司

HESAI | KOSMO

CHAPTER 01

空间智能是什么

01



1.1 空间智能的定义

纵观人类文明史：波利尼西亚航海家依靠星辰、海浪横渡大洋；外科医生借助影像定位病灶、规划手术路径；普通人在路口判断来车，决定何时过街。人类无论是在开拓未知世界、完成复杂专业任务，还是处理最日常的生活场景，都依赖同一种底层能力：在大脑中感知空间、解析位置关系、推演环境变化并据此行动——这就是空间智能。它并非独立的高阶技能，而是人类依靠 “感知-理解-推理-交互” 闭环与物理世界交互的基础能力，持续拓展人类行动边界，支撑文明演进。

学术界与产业界对空间智能给出了多种定义。李飞飞将其概括为“连接感知、想象和行动的核心能力”， 并提出世界模型应能根据输入的行动预测世界的后续状态[1]，将预测视为从感知通向行动的关键环节；斯坦福大学以人为本人工智能研究院（Stanford HAI）将其定义为理解和推理三维物理世界的能力，涵盖物体间的空间关系、运动与交互，以及对物理环境的感知、解释和导航[6]；英伟达创始人兼首席执行官黄仁勋则指出，下一波AI应理解物理规律，并能够在现实世界中工作[7]。以上的观点共同勾勒出空间智能从感知、理解、推理到交互的完整能力链条。

尽管表述各异，但各方共识清晰：**空间智能是物理AI走向现实物理世界的必要能力，感知、理解、推理、交互是空间智能的核心要素。**

> **KEY DEFINITION**
>
> 综合上述权威观点，本白皮书将空间智能定义为：智能系统基于真实世界的空间信息，完成感知-理解-推理-交互的综合能力；其目标是在空间中形成可预测、可行动的认知闭环。

1.2 物理AI向空间智能的演进

空间智能的概念最初源于心理学领域，由霍华德·加德纳（Howard Gardner）在其多元智能理论中提出[8]，用以描述人类感知、理解和操作三维空间的能力。1950年，艾伦·图灵在其开创性论文《计算机器与智能》中发出了“机器能思考吗？”这一根本追问[9]，由此开启了将人类“原生智能”赋予机器的漫长征程。在此背景下，空间智能的概念逐渐被引入并深植于物理AI领域。

AI向空间智能演进，表现为各种模型的演进，实际上是由底层“原材料”和“加工方法”的升级同步驱动。而底层的“原材料”升级，即为数据升维——每一次数据维度的增加，实际上都在给物理AI增加一种“理解现实世界”的可能性。

模型范式	核心突破	主要信息结构 / 表征维度	与前一阶段的关系
LLM Large Language Model 大语言模型	理解与生成语言	一维序列：由离散符号（词元）构成的线性语言表征。	—
MLLM Multimodal Large Language Model 多模态大模型	感知并理解世界	二维视觉及时序信息：图像表征二维空间特征，视频在此基础上引入时间维度。	从单一语言模态扩展至视觉等多模态感知。
VLA Vision-Language-Action Model 视觉-语言-动作模型	在物理世界中行动	三维空间及时间：由视觉感知、语言认知和物理动作共同构成的多模态时空表征。	在多模态感知基础上，进一步引入行动维度与物理交互。
WM World Model 世界模型	预测并模拟世界	三维空间及时间：对动态环境状态及其随时间的演化进行表征，并可模拟行动对未来状态的影响。	与VLA并非严格线性阶段关系，其核心扩展在于对环境动态及行动后果进行预测与推演。

图01 不同AI模型范式的核心能力与表征结构比较

© 2026 上海禾赛科技有限公司

上图直观展示了数据维度拓展给AI带来的跃迁：

文本时代——AI学会“理解语言”。大语言模型的训练依赖海量互联网文本。它以“预测下一个词元”为核心训练目标，在海量文本中学习语言结构、语义关系与知识关联，并由此获得语言理解、内容生成和一定的推理能力。这一阶段的AI主要生活在“文字世界”里，通过语言符号来理解和描述现实。它知道“猫”这个词，却从未“见过”猫。数据维度是单一的：只有文本，一维符号。

多模态时代——AI学会“感知世界”。当训练数据从纯文本进一步扩展到图像、视频等视觉信息后，AI开始获得直接感知现实世界的能力。图像为模型增加了二维结构，使其能够理解物体的形状、位置和场景关系；视频进一步引入时间信息，使模型能够观察物体和场景如何连续变化。与此同时，跨模态对齐与融合使语言、图像和视频能够被映射到共通语义空间中——模型开始能够把文字里的“猫”和画面里的“猫”对应起来。数据由一维语言符号扩展到二维视觉及时间信息，AI因而从“听别人描述世界”，迈向“自己看见世界”。但这一阶段的理解仍主要停留在视觉表征层面：模型可以识别“杯子放在桌上”，却并不天然理解杯子在真实三维空间中的距离、质量、碰撞关系以及受到外力后会如何运动。

行动时代——让AI学会“与世界交互”。VLA模型在视觉和语言理解的基础上进一步引入动作输出，使AI从“认识环境”走向“改变环境”。训练数据不再只有文本、图像或视频，还加入机器人示教、遥操作轨迹、机器人自身状态、控制指令以及动作后的环境反馈。在信息结构上，模型面对的不再只是二维画面，而是真实三维物理空间中随时间展开的“状态—动作—反馈”轨迹。模型需要根据当前感知和任务指令生成动作，并根据动作执行后的新状态继续调整下一步决策——伸手去摸猫。摸到猫炸毛，立刻缩回手；发现猫蹭你，就顺势挠它下巴。因此，这一阶段真正增加的不只是“动作”这一种输出，而是三维空间和时间中的物理交互。AI从“看懂世界”进一步走向“进入世界并与之交互”。

空间智能时代——AI学会“预测世界变化”。世界模型进一步要求AI不仅识别当前环境、生成下一步动作，还要学习世界状态如何随时间演化，以及不同动作会对未来状态产生什么影响。相较多模态大模型，世界模型的训练信息不再只是描述“看到了什么”，而是进一步加入深度、三维点云、物体与自身位姿等空间状态信息，以建立对真实三维环境的内部表征；相较VLA，世界模型还需要学习“当前状态—动作—未来状态”之间的变化关系，从而预测不同动作将如何改变后续世界——在摸猫之前会先预演：“摸肚子可能会被抓伤，摸下巴猫会打呼噜。”于是做出摸下巴的选择。数据层面，AI所处理的信息进一步走向“三维空间 + 时间”的动态世界状态；能力层面，则从单步感知与行动升级为对物理世界演化规律的持续建模。

数据维度的每一次跃迁，不仅是技术指标的提升，更是AI理解现实世界能力的质变。从读懂文字，到看懂图片，到理解空间关系，与现实交互并预演未来——**AI正在完成从“数字世界的思考者”到“物理世界的参与者”的根本转变。**

KEY JUDGMENT

这意味着，谁能率先构建起从一维符号到多维时空、从感知数据到交互与状态演化数据的多层次数据体系，谁就更有可能在空间智能竞争中占据数据基础设施的先发位置。

© 2026 上海禾赛科技有限公司

HESAI | KOSMO

CHAPTER 02

空间智能：物理AI演进的必经之路

02



2.1 空间智能：人类智能与人工智能演进的底层基础与当前AI发展的根本局限

第一章中，我们界定了何为“空间智能”及其发展至今的演变史。那为什么空间智能对物理AI至关重要？本章将从两条相互印证的路径展开：一条路径追溯人类智能的进化源头，揭示空间智能是构成高级认知能力的底层基础；另一条路径则从当前AI的技术瓶颈出发，观察其面临的缺陷。两条路径交汇于同一结论——**空间智能是物理AI发展不可绕行的战略高地**。

2.1.1 从进化与认知发展的视角：空间智能是认知“脚手架”



图02 AI的演进和人类的演进异曲同工

智能诞生于与世界的互动。

从进化视角审视，早在语言与文明诞生之前，生物便已通过感知光线、纹理、温度等原始信息与环境互动。这种生存本能驱动神经系统不断演化，形成解读世界、协调行动的复杂机制，最终塑造了人类“感知-理解-推理-交互”的完整认知闭环。

在个体认知发展层面，瑞士心理学家皮亚杰指出：儿童通过抓取、触碰、移动和观察，逐步建立对外部空间的表征，并在此基础上发展出更复杂的语言、记忆和抽象思维[10]。**这意味着空间能力不是高级智能成熟之后附加的技能，而是高级智能得以发育的认知根基。**

借鉴人类进化路径，AI领域正达成相同的共识：物理AI的演进若要实现真正意义上的通用智能，必须复现人类智能的底层基础——空间智能。1958年，麦卡锡主张智能体需要以常识推理支持行动[11]，而物理世界中最重要的常识之一即为空间相关常识。到2025年，《从语言到世界：空间智能是AI的下一个前沿》一文中点明：要理解AI的下一站，必先回溯人类智能的源头[1]。文中指出，“感知与行动构成了智能进化的核心循环，而空间智能正是这一循环的产物，是人类认知的‘脚手架’——它不仅支撑着我们与物理世界的日常互动，更是想象力、创造力与文明进步的底层动力”。构建对空间和物理世界的理解能力——空间智能，是物理AI发展更复杂智能的重要基础。

© 2026 上海禾赛科技有限公司

2.1.2 从当前AI的根本局限倒推：无法绕过的“空间盲区”

如果说上一节从人类智能的进化与认知发展出发，说明空间能力是高级智能的底层基础，那么反向观察当前AI的发展，也会得到同样的结论：当AI从文本、图片等数字信息走向真实世界时，普遍暴露出感知不稳定、理解不深入、交互不可靠三类问题。这类问题并非孤立的模型缺陷，本质是空间智能 “感知-理解-推理-交互”的能力缺少完整闭环，成为AI落地真实世界最底层的瓶颈。

第一，当前AI已经能够识别画面，却仍难形成稳定、连续的空间感知。真实空间不会因为观察视角改变而改变，但模型往往难以把不同位置、不同时间看到的画面整合为同一个环境。李飞飞、谢赛宁等提出的VSI-Bench（Visual-Spatial Intelligence Benchmark）是一项专门评估多模态大语言模型三维视觉空间智能的视频基准测试。该测试以连续视频为输入，通过空间问答任务评估模型表现。结果显示，尽管现有多模态大模型已具备初步空间意识，但整体表现仍明显落后于人类：在完整测试集上，表现最佳的Gemini-1.5 Pro综合平均得分为45.4%；在用于人机对比的400题子集（VSI-Bench tiny）上，人类综合平均得分达到79.2%，而Gemini-1.5 Pro仅为48.8%。进一步的错误分析显示，约71%的受分析错误可归因于空间推理缺陷，涉及距离、大小、方向等空间关系判断，以及自我视角与环境视角之间的转换[12]。值得注意的是，VSI-Bench的输入形式是连续视频，而非单张图片——这意味着AI的问题并非“没有看到足够多的画面”，而是感知未能沉淀为稳定的空间认知——这正是空间智能在感知层未能闭合的表现。感知一旦无法输出稳定的环境模型，闭环的起点就无法真正建立。

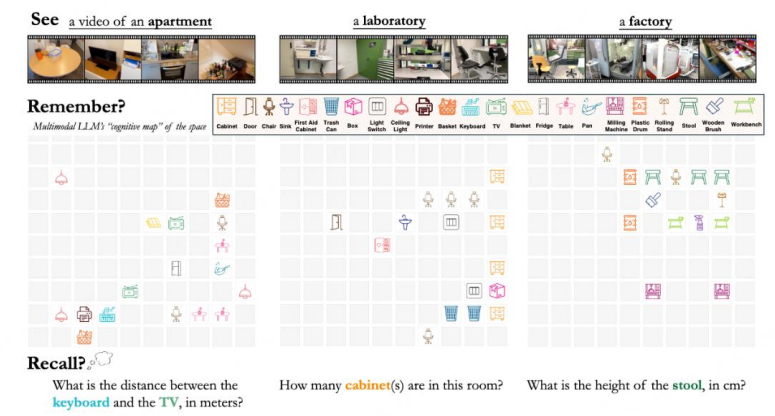


图03 VSI-Bench的评估流程与测试结果。左侧为基准测试的任务范式：模型首先观看连续的第一人称视角视频（See），尝试记忆物体及其空间位置（Remember），随后回答诸如距离、数量、高度等跨视角空间问题（Recall）（原文Figure 1）[12]

第二，当前AI已经能够识别物体，却仍难理解物体之间的空间关系与物理约束。SPHERE（Spatial Perception and Hierarchical Evaluation of REasoning，用于系统评估视觉语言模型空间推理能力的分层框架）按照基础空间能力、多能力组合和物理情境推理逐级测试视觉语言模型，结果显示，模型可以处理部分简单方向关系，但在距离与邻近判断、自我视角与第三方视角转换，以及需要结合现实情境的空间逻辑上存在显著短板——GPT-4o与人类之间存在25个百分点的准确率差距[13]。另一项覆盖13个主流视觉语言模型的SRBench（Spatial Reasoning Bench，用于系统评估视觉语言模型空间推理能力的综合基准测试套件）研究发现，模型在空间关系、方向导航、心理旋转和空间可视化等任务上的平均表现接近随机水平[14]。机器能够回答“这是什么”，并不代表它已经理解“它在哪里、与周围对象是什么关系、能否通行或操作”。这种从对象识别到空间理解之间的断层，是空间智能在理解与推理层未能闭合的表现。没有空间理解与推理，AI就无法从“识别物体”走向“理解世界”。

© 2026 上海禾赛科技有限公司

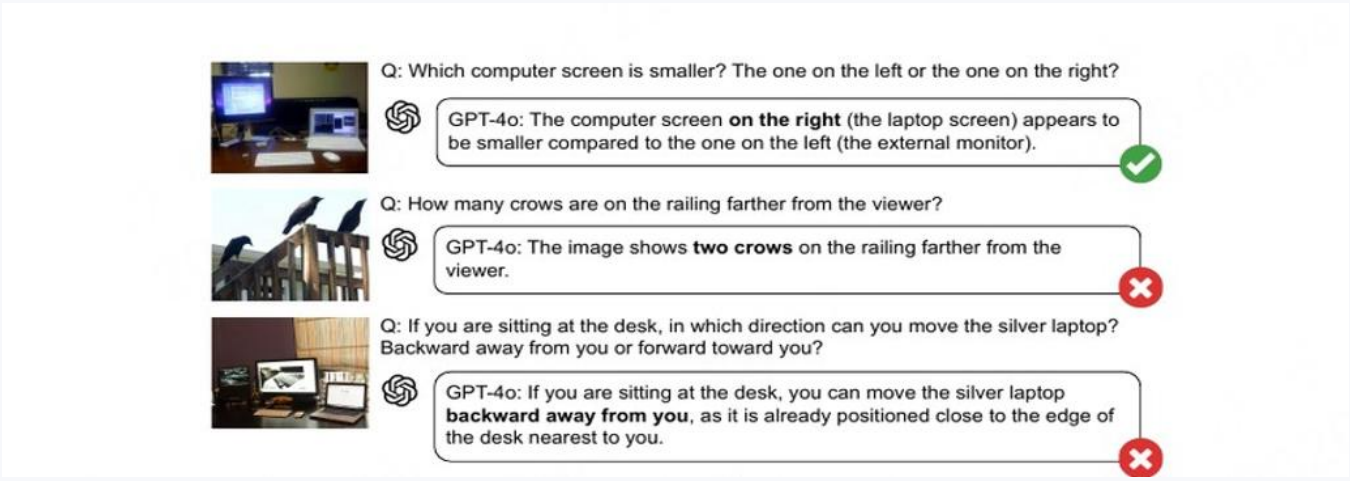


图04 GPT-4o在空间计数、视角转换和物理约束推理上的典型失败案例（原文Figure 1）[13]

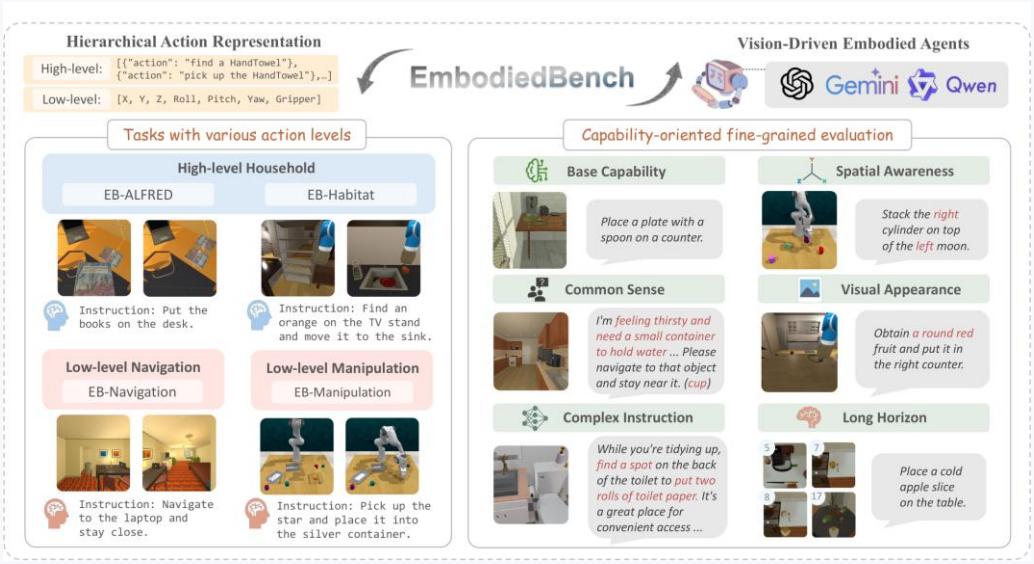
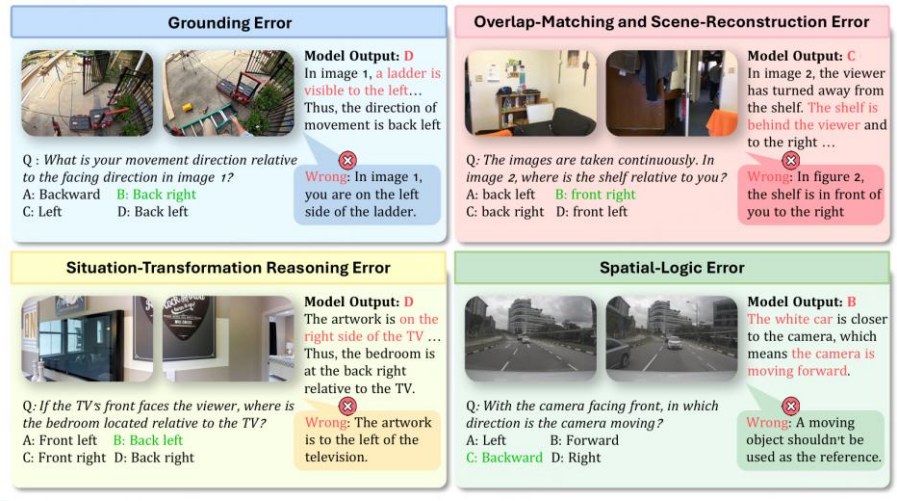


图05 EmbodiedBench四类测试环境及其对应的任务层次（原文Figure 1）[15]

第三，当前AI即使能够描述行动，也仍难在真实环境中持续、可靠地完成交互。真实任务不是生成一次答案，而是“观察环境—采取动作—接收反馈—修正下一步”的连续循环。EmbodiedBench（一个用于系统评估多模态大语言模型在视觉驱动具身任务中表现的综合性基准）在四类环境中设置了1,128项视觉驱动具身任务，覆盖家庭任务、导航和操作等不同层级；研究显示，多模态大模型在高层语义任务上的表现相对较强，其中Claude-3.5-Sonnet在EB-ALFRED和EB-Habitat上的平均准确率分别为64.0%和68.0%；但在低层任务中，即使是表现领先的GPT-4o，在EB-Navigation和EB-Manipulation上的平均得分也仅为57.7%和28.9%[15]。这表明，模型虽然能够理解指令并生成操作步骤，但在根据环境反馈持续修正感知、推理与行动规划方面仍存在明显不足，反映出当前AI在空间智能交互层面的局限。

这三类问题还会在复杂任务中相互放大。MMSI-Bench（Multi-Image Spatial Intelligence，专门评估多模态大模型多图像空间智能的视觉问答基准）从超过12万张候选图像中筛选出1,000道必须跨图像完成的真实空间推理题，并对多模态大模型进行评估。在主流的37个大模型中，GPT-5达到41.9%，而人类水平为97.2%[16]；更值得关注的是，错误分析揭示了四种主导性失败模式：接地错误、重叠匹配与场景重建错误、情境转换推理错误和空间逻辑错误，其中重叠匹配与场景重建错误在所有模型中都占据了最大比例——识别跨图像的同一物体并隐式重建场景布局，是所有模型最薄弱的环节。这些错误并非孤立存在于某一环节：接地错误意味着感知未能建立稳定的空间锚点，重叠匹配与场景重建错误意味着理解无法将多视角观察整合为统一的空间结构，情境转换与空间逻辑错误则意味着推理无法在空间关系变化时保持一致性。这些问题在同一条推理链上依次触发、相互叠加。**当前AI共同出现的“感知漂移、理解断层、交互失效”，本质上是同一条“感知-理解-推理-交互”的空间智能链没有闭合。**

© 2026 上海禾赛科技有限公司



Models	Positional Relationship						Attribute		Motion		MSR	Avg.
	Cam.–Cam.	Obj.–Obj.	Reg.–Reg.	Cam.–Obj.	Obj.–Reg.	Cam.–Reg.	Meas.	Appr.	Cam.	Obj.	–	
Proprietary												
GPT-5	43.0	35.1	32.1	48.8	42.4	51.8	60.9	**36.4**	32.4	36.8	**42.0**	**41.9**
o3	**45.2**	**39.4**	37.0	44.2	47.1	**62.6**	54.7	28.8	31.1	32.9	34.9	41.0
GPT-4.5	34.4	29.8	**39.5**	**51.2**	47.1	55.4	39.1	33.3	**41.9**	**40.8**	36.4	40.3
GPT-4.1	36.6	26.6	27.2	29.1	36.5	27.7	37.5	24.2	36.5	32.9	28.8	30.9
GPT-4o	34.4	24.5	23.5	19.8	37.6	27.7	32.8	31.8	35.1	36.8	30.8	30.3
Gemini-2.5-Pro	39.7	31.9	**39.5**	45.3	35.2	43.3	51.5	21.2	36.4	30.2	34.3	36.9
Claude-3.7-Sonnet	35.5	29.8	29.6	31.4	32.9	34.9	37.5	25.8	23.0	31.6	25.8	30.2
dots.vlm 1	33.3	31.9	38.3	27.9	34.1	42.2	**62.5**	**36.4**	23.0	25.0	30.8	34.1
GLM-4.5V-thinking	38.7	35.1	29.6	30.2	40.0	39.8	45.3	28.8	25.7	32.9	30.3	33.8
Doubao-1.5-pro	25.8	33.0	33.3	33.7	**49.4**	39.8	37.5	31.8	27.0	26.3	29.8	33.0
Seed1.5-VL	32.2	30.8	25.9	23.2	38.8	32.5	39.0	21.2	36.4	25.0	26.2	29.7
Open-source												
InternVL3-78B	34.4	23.4	32.1	12.8	37.6	26.5	37.5	19.7	**28.4**	31.6	29.3	28.5
InternVL2.5-78B	23.7	22.3	**39.5**	29.1	31.8	**42.2**	35.9	19.7	17.6	26.3	27.3	28.5
Qwen2.5-VL-72B	25.8	**34.0**	34.6	23.3	34.1	36.1	**45.3**	27.3	27.0	30.3	27.3	**30.7**
LLaVA-OneVision-72B	**43.0**	31.9	33.3	30.2	37.6	38.6	28.1	19.7	13.5	32.9	15.7	28.4
InternVL3-38B	21.5	20.2	33.3	23.3	35.3	25.3	39.1	21.2	16.2	31.6	25.8	26.3
InternVL2.5-38B	18.3	22.3	35.8	22.1	**38.8**	34.9	37.5	25.8	14.9	38.2	25.3	27.9
Qwen2.5-VL-32B	24.7	26.6	29.6	22.1	32.9	31.3	31.2	24.2	18.9	35.5	27.8	27.7
InternVL2.5-26B	24.7	19.1	29.6	33.7	31.8	37.3	35.9	30.3	10.8	31.6	26.8	28.0
NVILA-15B	30.1	39.4	28.4	36.0	**38.8**	20.5	29.7	**31.8**	18.9	35.5	27.8	30.5
InternVL3-14B	19.4	24.5	24.7	23.3	37.6	24.1	31.2	22.7	24.3	31.6	29.3	26.8
Llama-3.2-11B-Vision	25.8	30.8	32.0	25.6	21.2	25.9	20.3	19.7	25.6	28.9	19.2	25.4
InternVL3-9B	18.3	25.5	32.1	29.1	31.8	22.9	29.7	24.2	16.2	38.2	26.8	26.7
InternVL3-8B	25.8	31.9	37.0	25.6	35.3	28.9	23.4	24.2	16.2	32.9	14.6	25.7
InternVL2.5-8B	32.3	27.7	29.6	32.6	24.7	32.5	26.6	27.3	16.2	31.6	**30.3**	28.7
NVILA-8B	17.2	29.8	24.7	30.2	22.4	34.9	34.4	25.8	25.7	34.2	29.8	28.1
Qwen2.5-VL-7B	24.7	24.5	24.7	25.6	29.4	26.5	25.0	18.2	20.3	**39.5**	25.8	25.9
LLaVA-OneVision-7B	20.4	33.0	29.6	29.1	25.9	30.1	29.7	25.8	18.9	34.2	11.6	24.5
InternVL2.5-4B	31.2	23.4	21.0	31.4	34.1	25.3	23.4	24.2	13.5	31.6	26.8	26.3
Qwen2.5-VL-3B	26.9	27.7	30.9	29.1	28.2	34.9	31.2	16.7	17.6	27.6	23.2	26.5
InternVL3-2B	26.9	25.5	29.6	31.4	28.2	27.7	26.6	22.7	12.2	23.7	23.7	25.3
InternVL2.5-2B	28.0	27.7	24.7	**37.2**	29.4	36.1	43.8	15.2	21.6	31.6	26.8	29.0
InternVL3-1B	24.7	35.1	22.2	30.2	29.4	30.1	32.8	28.8	17.6	19.7	26.3	27.0
InternVL2.5-1B	23.7	26.6	24.7	25.6	31.8	25.3	31.2	30.3	17.6	25.0	26.3	26.1
DeepSeek-VL2	23.7	31.9	22.2	36.0	30.6	22.9	28.1	15.2	**28.4**	26.3	28.3	27.1
DeepSeek-VL2-Small	24.7	28.7	18.5	33.7	**38.8**	27.7	28.1	33.3	24.3	25.0	29.8	28.6
DeepSeek-VL2-Tiny	29.0	27.7	21.0	23.3	17.6	31.3	14.1	24.2	14.9	25.0	27.3	24.0
Baseline												
Blind GPT-4o	20.2	17.0	29.6	13.9	29.4	19.2	21.8	12.1	20.2	29.0	20.2	22.7
Random Guessing	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0	25.0
Human Level	95.7	98.9	97.5	94.2	98.8	96.4	95.3	98.5	98.6	98.7	97.0	97.2

图06-07 MMSI-Bench中多模态大模型空间推理的四类典型错误示意图，与在空间推理任务中的表现对比（原文Figure 6&Table 3）[16]

因此，真正的问题是AI还没有学会像生物一样，在与世界的互动中建立并维护一个可更新、可推理的空间。文本、图片和语言可以描述世界，却无法替代机器亲自去经历世界——去观察、去行动、去从每一次反馈中修正自己对空间的判断。空间智能的本质，从来不是某一项技术的堆叠，而是把感知、理解、推理与交互连成一个能够自我迭代的闭环：让机器不只是处理关于世界的信息，而是真正进入世界、理解世界，并在世界中持续行动。没有这个闭环，AI就永远停留在“谈论世界”的阶段；有了它，AI才真正拥有了从数字世界走向真实世界的入场券。

KEY JUDGMENT

由此，第二章的两条论证路径最终交汇：从人的认知发展看，空间智能是高级智能形成的“脚手架”；从当前AI的共同困境看，缺少空间智能又是其无法可靠进入物理世界的根本限制。空间智能不是现有AI能力之外的一项附加功能，而是物理AI发展不可绕过的底层能力。

© 2026 上海禾赛科技有限公司

HESAI | KOSMO

CHAPTER 03

空间智能产业：3D空间数据是发展瓶颈

03



3.1 空间智能产业发展现状



图08 空间智能产业主要应用场景

空间智能对于下游应用场景的影响，本质上是由于空间智能扩展了AI感知、理解、推理和与现实世界交互的能力边界。一方面提升了三维资产的生产和应用能力，**打破了供给端的瓶颈**，从而促进了相关产业的发展。比如为世界模型与生成式AI、具身智能、影视娱乐、游戏、泛测绘等原本高度依赖三维空间数据的领域提供大规模、低成本的生产资料，使其快速发展；另一方面则催生出了更丰富的沉浸式、可交互产品形态，**从需求端拓展了产业发展的更多可能**。比如为文旅文博、广告和音乐等行业带来新的产品与用户体验。

世界模型、生成式AI及具身智能：AI正在真正进入物理世界

世界模型是实现空间智能的重要技术路径。与此同时，三维空间数据、空间表征、仿真与合成数据等空间智能基础技术的发展，也为世界模型及生成式AI的训练和演进提供了更丰富的数据基础。

我们可以把物理AI比作深入陌生丛林的探险家。世界模型就是他脑中的动态地图：不仅记录山川地貌与障碍物，还能预判环境变化，推演行进中可能遇到的状况。依靠这张地图，探险家能够感知、理解周围环境，并与环境交互，也就是空间智能；反过来，探险家持续观测周围环境，不断积累新的观察和认知，又持续完善、丰富这张地图——即空间智能让世界模型愈发精准，推动其优化与迭代。

随空间智能发展，世界模型正加速从技术研究走向产品化与产业应用。根据“A Functional Taxonomy of World Models”一文，世界模型可以分为三个层次：渲染器（Renderer）、模拟器（Simulator）和规划器（Planner）[17]。当前，以图像和视频生成为代表的渲染器已实现商业化；能够表征和推演世界状态的模拟器正处于技术快速突破和早期产业应用阶段；能够基于环境和目标自主生成行动的规划器仍总体处于技术探索期。

渲染器解决“世界长成什么样子”。Google视频生成模型Google Veo能够依据文本或参考图片生成视频。该模型现已通过Google Cloud提供标准化调用能力与清晰计费规则，例如Veo 3.1按输出规格计费，生成1080p视频及音频的价格为0.40美元/秒 [18]。面向影视、广告与创意制作场景的Runway Gen-4系列图像及视频生成模型，已开放开发者API，按视频时长或图片数量计费，其中Gen-4 Turbo的收费标准为0.05美元/秒 [19]。

© 2026 上海禾赛科技有限公司

模拟器解决“如果这样做，世界会变成什么样？”。以NVIDIA Cosmos为代表的模拟器型世界模型，已落地机器人与自动驾驶的实际研发工作，用于合成数据生成、场景仿真与模型训练。机器人领域，Agility Robotics、Figure等企业借助Cosmos[20]生成大规模、可调控的合成训练数据与仿真环境，让机器人在虚拟空间反复练习抓取、行走、操作等动作，再将习得能力迁移至现实世界。自动驾驶领域，Uber等企业利用Cosmos扩充驾驶场景库，重点生成现实中成本高昂、样本稀少或具备危险性的长尾场景，用于自动驾驶系统的训练与验证。World Labs最新推出的Atlas，则从单纯的3D场景生成，进一步拓展至世界重建与模拟，旨在理解环境的动态变化，为机器人仿真与现实场景训练提供环境支撑。2026年9月，Atlas已对部分合作伙伴开放早期访问权限[21]。

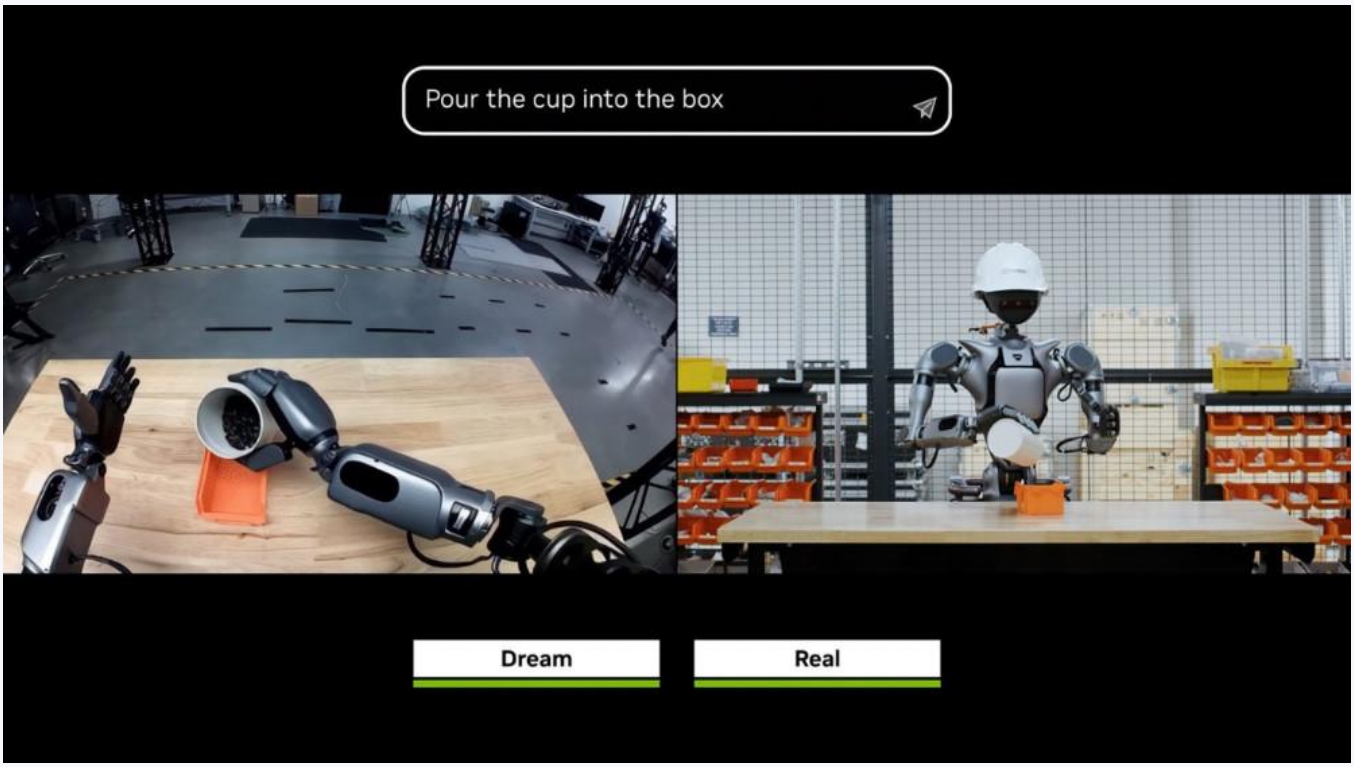


图09 Cosmos先让机器人在虚拟世界中反复练习和预测动作结果（左侧Dream），再把学到的策略迁移到真实世界执行（右侧Real）[22]

规划器解决“为达成目标，在众多可能性中，下一步应当如何行动？”。这一层要求模型不仅理解当前环境，还能推演不同行动对应的结果，并据此完成决策。以下棋为例：规划器落子后棋盘状态改变，模型会重新推演、再次规划，最终判断“想要取胜，这一步应当走马，而非走车”。目前该方向整体仍处于早期探索阶段。

从渲染器到模拟器再到规划器，这背后正是空间智能在不断推动世界模型的空间理解与推理能力的提升，推动世界模型从“生成逼真的画面”走向“建立稳定的三维世界认知”。过去，世界模型难以始终保持不同视角下物体位置、几何结构与空间关系的一致性。而随着空间智能增强，世界模型开始能够将不同视角的观察整合为统一的三维空间表征，并在此基础上开展更复杂的环境模拟。例如，前文提到的World Labs Atlas，在官方演示中最初仅凭一张花园照片生成周围场景；随着相邻小屋、主屋的照片陆续加入，模型能够不断完善对同一场景的空间理解，更准确地重建建筑之间的空间布局 [21]。这意味着，**空间智能不仅帮助世界模型建立稳定的三维空间认知，还使其能够在此基础上支持更复杂的环境仿真，为未来预测行动后果与自主规划奠定基础。**

当前市场对于世界模型相关企业的关注持续升温，资本投入一年内增长至近5倍：外部融资规模由2024年的14亿美元增至2025年的69亿美元[23]。总体来看，世界模型是与空间智能关联最紧密的应用方向之一，随着世界模型在内容生成、环境仿真、机器人训练等领域的应用不断深化，对空间智能的需求也将持续增长，进一步拓展商业化空间。

© 2026 上海禾赛科技有限公司

与世界模型侧重环境状态预测和行动后果不同，**空间智能助力生成式AI发展的成果，体现在如何将创意快速转化为可交付的数字资产**。开发者开始借助大模型搭建三维场景、编排镜头、连接专业软件，让AI从"出一张图"走向"参与一套制作流程"。

2026年9月，OpenAI推出GPT-6 Astra。在借助工具重建物体三维几何结构的评测中，Astra的几何重合度达到95.9%，显著高于上一代GPT-5.6 Sol的83.3%[24]——**空间智能的进步使模型的空间理解能力与推理能力不断提升，帮助生成式AI从单纯追求"画面逼真"，进一步走向"结构准确、空间可控"**。更值得关注的是Astra在实际三维创作流程中的表现。OpenAI公开案例中，用户提出住宅设计需求，Astra便通过Blender脚本构建建筑、家具，还能根据需求呈现不同的材质与打光效果，甚至可以具体到不同机位和角度；随后将场景导入Unreal Engine 5，实现可自由漫游的互动空间[25]。这一成果背后正是空间智能的作用。这说明生成式AI正从输出视觉效果，走向搭建可编辑的三维资产和完整制作流程。不过，当前实现此类成果仍需专业软件、人工审阅与反复迭代，模型在几何和物理约束上仍有缺陷。

当前，生成式AI已通过API、订阅服务及企业软件进入设计、研发和内容生产流程。Gartner于2026年9月预测，全球生成式AI模型支出2026年约283亿美元，2027年将增至约516亿美元[26]。资本市场也持续追捧前沿模型企业：据2026年10月报道，OpenAI新一轮融资洽谈的目标估值约为1.25万亿美元[27]；Anthropic则于2026年5月完成650亿美元融资，投后估值达9650亿美元[28]。

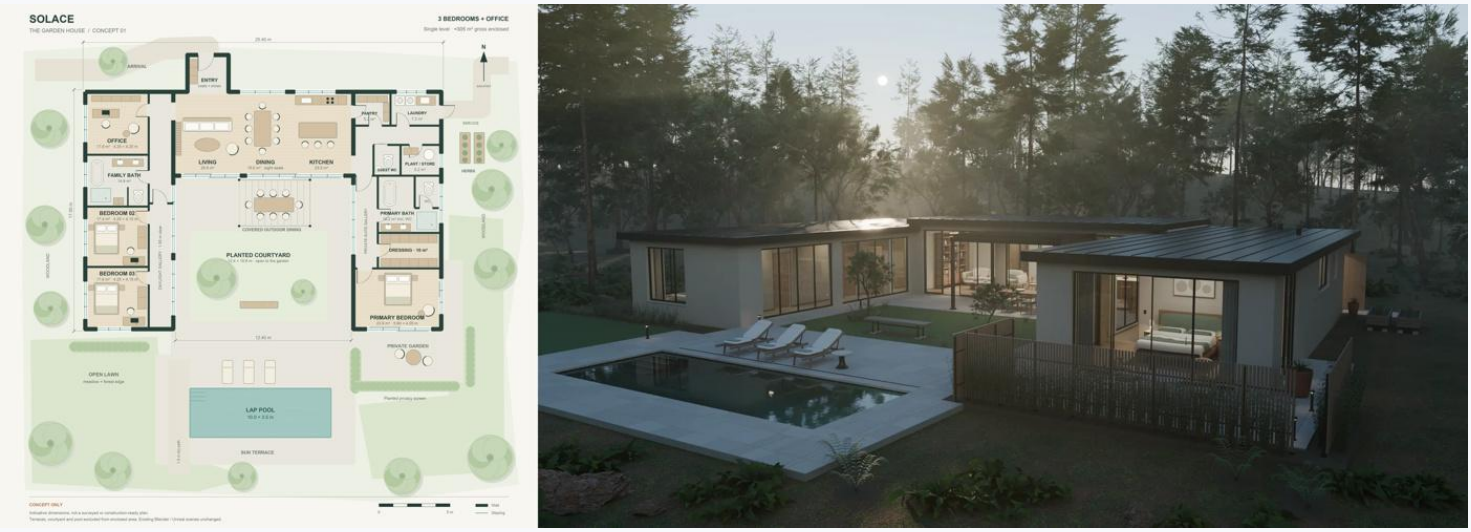


图10 GPT-6 Astra根据住宅设计需求生成建筑平面图（左）及对应的三维场景（右）[25]

具身智能是空间智能在物理世界中最直观、最典型的应用方向。依托空间智能，具身智能才能搭建起对物理世界"感知-理解-推理-交互"的完整闭环，才能实现在物理世界中的真实部署。其物理载体呈现多元化特征，不仅以人形机器人形式出现，轮式移动操作机器人、机械臂及四足机器人等都属于具身智能的范畴。**如今具身智能产业商业化关键节点已至，但规模化落地仍待跨越。**

国际数据公司（IDC）统计，2025年全球人形机器人出货量约1.8万台[29]；谷歌、英伟达、Agility Robotics、Figure AI、宇树科技等海内外头部机器人企业，已陆续迭代出机器人产品，全面切入工业制造、仓储物流、家庭服务、商业陪护等多元场景，可独立完成物料搬运、精密分拣、简单人机协作、居家辅助等标准化与半标准化作业，标志着具身智能从实验室技术研发正式走向实景落地应用阶段。

© 2026 上海禾赛科技有限公司



图11 具身智能落地应用示意图

但行业整体距离规模化复制尚有明显距离。许多应用场景并非真正创造生产价值的成熟生产场景，上述1.8万台人形机器人出货中，超过85%仍集中在表演/导览、教育/科研等展示及验证场景[29]。其余落地场景也主要集中在汽车制造、仓储物流等结构化场景中的少数固定任务，往往需要针对客户场地、流程和既有系统进行适配，跨场景快速复制能力仍待验证。并且，具身智能设备的安全认证体系、事故责任界定机制、行业监管规范也尚未完全成熟。

但空间智能的进步，正在明显加快具身智能的能力迭代。过去的具身智能往往需要针对固定物体、固定工位和固定动作分别编程或训练；随着空间智能增强，具身智能开始能够理解物体之间的空间关系、适应环境变化，并将已有能力迁移至此前未见过的物体、场景和任务。例如，Google DeepMind于2025年推出Gemini Robotics，将Gemini的世界理解能力进一步延伸至物理空间，其在综合泛化测试中的表现较此前先进VLA模型的表现提升超过一倍[30]，并能够处理新物体、新指令和新环境；数月后推出的On-Device版本，仅需50-100个示例即可快速适配新任务[31] 。NVIDIA的迭代速度同样明显：GROOT N1于2025年3月发布，仅三个月后即升级至N1.5；在相同1,000个遥操作样本训练下，部分真实机器人任务成功率由N1的44.0%提升至N1.5的98.8%[32]。

市场也对具身智能未来发展非常乐观：摩根士丹利判断，全球人形机器人的规模化部署预计将在2030年代中后期明显提速：全球人形机器人保有量将由2030年的约200.0万台增至2040年的约1.4亿台，并于2050年达到约10.2亿台；同期，全球人形机器人收入预计达到约7.5万亿美元[2]。

© 2026 上海禾赛科技有限公司

影视娱乐：三维场景，从“一次性制作成本”转向 “可复用生产资料”与媒体体验升级

影视和游戏的核心生产资料，本来就是“空间”。二者都需要先构筑一方世界，或真实，或虚构；再在其中安放人物，调度镜头，铺展故事。空间智能真正改变的，正是这方“世界”被生产、被操控和被复用的方式。因此，在影视行业，空间智能首先影响的是供给端：让前景的人物与背景的环境解耦，分开拍摄、自由合成。过去需要美术、建模、布景、特效团队大量人工制作的场景，可以更多由AI完成重建、生成和编辑；导演还可以在统一三维空间中自由调整机位、镜头和场景，本质上是把AI从“生成一段视频”升级为“帮助制作一部电影的虚拟摄影棚”。这也推动AI生成影视从单镜头、单片段生成，进一步走向场景统一、镜头可控、资产可复用的完整制作流程。在游戏行业，影响更进一步：供给端可以自动生成地图、建筑、角色和关卡，降低3D资产生产成本。AI不再只是帮开发者“做游戏素材”，而开始参与“运行一个活的游戏世界”。

相较于具身智能，空间智能在影视娱乐领域商业化落地更快、成果也更为显性，核心在于它的落地门槛相对更低。影视游戏优先追求视觉真实、空间自洽与可用的交互体验，即便模型尚未完全掌握复杂的物理规律，也能够投入场景生成、虚拟拍摄、游戏世界搭建等实际生产工作。比如，工业光魔（ILM）在漫威剧集《铁心》中，采用NeRF技术从无人机拍摄的芝加哥实景2D素材重建了高保真3D环境。传统方式需要完整CG建模整座城市，在剧集制作周期内资源与时间成本不可行。NeRF方案以远超传统技术的效率完成了视效镜头制作，并因此获得HPA（好莱坞专业协会）技术创新奖。

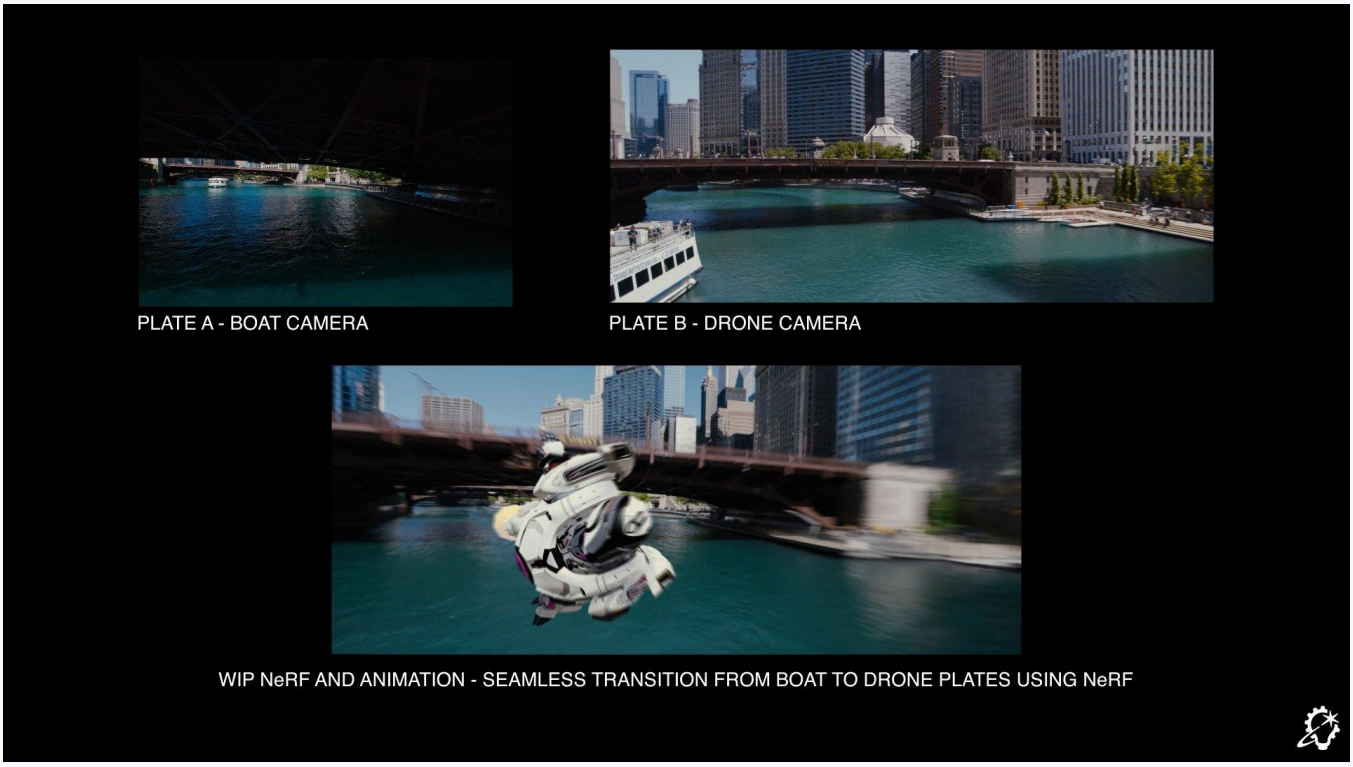


图12 左上PlateA：真实相机在船上、桥下拍到的低机位画面；右上PlateB：无人机从更高位置拍到同一片河道和城市建筑。下方WIP NeRF + Animation：把两组真实素材重建成一个连续的NeRF三维场景后，虚拟镜头可以从“船上机位”平滑飞到“无人机机位”，中间那些现实中没有真正拍到的视角由NeRF根据三维场景进行合成，再叠加CG角色/物体[33]

© 2026 上海禾赛科技有限公司

"3D导演台"也是近期空间智能在影视娱乐领域中具体的商业化落地方向。它让创作者能像在真实片场一样，先在三维空间里搭好场景、摆好人物、设好机位，再让AI按既定调度去生成画面。目前已有多个平台推出了相关产品并投入商业运营。



图13 3D导演台工作流示意图

游戏领域，World Labs Marble ×《Gaussian Mansion》[34]中，开发者把二维概念图直接交给Marble生成可探索的3D世界，再导入Unreal Engine 5制作成可玩的射击游戏，把原本需要建模、贴图、灯光团队完成的"概念图→3D场景"流程大幅压缩。更进一步，World Labs与Rosebud AI已演示从一句文字或一张图片生成3D世界，再直接加入规则和多人交互变成游戏，说明空间智能开始从"生成素材"走向"生成可玩的世界"。

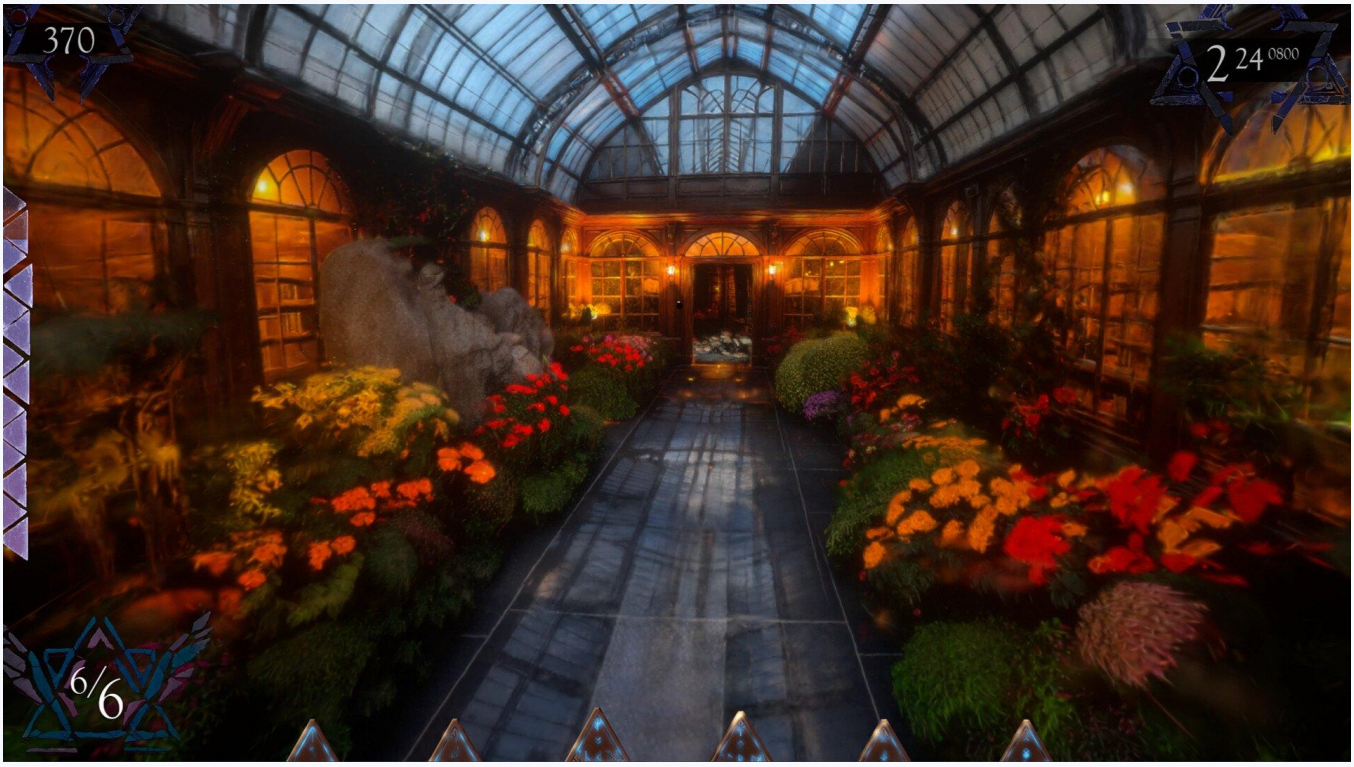


图14 José Tijerín使用World Labs Marble，把二维概念图/全景图生成可探索的3D环境，再导入Unreal Engine 5做成可玩的游戏《Gaussian Mansion》[34]

© 2026 上海禾赛科技有限公司

音乐与广告则更明显体现“需求创造”——内容从被观看、被聆听，进一步变成可以进入、探索和互动的空间体验。如《堡垒之夜》（Fortnite）Soundwave系列虚拟演唱会[35]：音乐不再仅仅是单向“播放”的音频，而是被封装在可供行走、探索、互动的三维世界之中。玩家可以和好友一同进入为艺人量身打造的专属空间；Epic还配套提供创作工具，支持音乐、灯光、场景乃至游戏机制随玩家行为实时动态调整。2026年10月，韩国知名音乐制作人、SM Entertainment创始人李秀满在推出新企划 A2O 的男团 A2O SOUL 和女团 A2O LITE 时，首次公开展示了应用于两支团体 MV 的4D高斯泼溅（4D Gaussian Splatting，4DGS）技术[36]。该技术通过多机位同步采集，将艺人与表演场景重建为包含时间维度的三维空间，使制作团队能够在拍摄完成后重新选择虚拟摄影机的视角，生成传统拍摄方式难以实现的镜头。这并非单纯提升音乐制作效率，而是将 “聆听音乐”这件事，拓展为“置身音乐世界”的沉浸式体验。

广告领域，Roblox沉浸式广告同样印证这一趋势：**品牌广告已经跳出二维视频、平面广告牌的形态，延伸至用户可进入、可互动的3D品牌空间。**Roblox在2026年推出全新广告形态，能够直接将品牌的视频创意转化为沉浸式三维环境，e.l.f.、山姆会员店、环球影业等品牌均已参与测试[37]。广告的逻辑由此转变，从“向用户讲述品牌故事”，升级为“让用户走进品牌构建的世界”。

在影视娱乐领域，空间智能有望成为产业智能化升级的重要技术支撑。根据美国商务部国际贸易管理局（ITA）发布的行业概览，全球媒体与娱乐市场规模约为2.8万亿美元[3]。随着三维场景重建、生成与交互技术不断成熟，空间智能有望推动影视、游戏、广告等领域降低内容制作成本、提升数字资产复用效率，并催生沉浸式、可交互的新产品形态，成为推动内容产业升级、拓展商业价值的重要驱动力。

文旅文博与泛测绘：把真实世界变成数字资产

空间智能在文旅文博领域已形成三条较清晰的商业化路径：**体验端，**通过LBE/XR、AR导览、空间定位和虚实融合提升游客体验；**保护/研究端，**通过3D重建、数字孪生及环境感知实现文物数字化存档、状态监测与风险预警；**运营治理端，**通过空间感知、客流预测和数字孪生平台实现客流调度、安全管理和场馆智慧运营。

Google Arts & Culture与CyArk已通过LiDAR激光扫描、摄影测量、无人机航拍和高精度摄影，持续将全球文化遗产转成可测量、可浏览、可下载的3D数字资产[38]。



图15 印度埃勒凡塔石窟（Elephanta Caves）三维虚拟游览界面 [38]

© 2026 上海禾赛科技有限公司

故宫博物院（中国最大的古代文化艺术博物馆）更是将空间智能在文旅文博方面的应用展现到了极致：在体验端，2005—2008年，故宫就开始对养心殿区域采用“三维扫描+人工建模”进行三维数据采集，构建养心殿建筑及文物的数字“骨架”；2015年养心殿研究性保护启动后，又通过全景拍摄和摄影测量进一步记录建筑、室内环境及文物的颜色、纹理等表面信息，为数字模型补充更加真实的“皮肤”。2026年，故宫进一步建设数字孪生应用服务平台，将文物本体、时空环境等4类36项多源多模态数据[39]接入统一的“故宫数据底座”，并将养心殿的数字空间与GIS、气象等现实环境数据进行关联，构建了一个实时同步的数字平行世界，现实中的气象、时间与空间位置被精准映射至虚拟场景，用户既可在其中沉浸式漫游，亦可将其作为立体化教学空间。

其价值也由“数字展示”进一步延伸至“管理与保护”：一方面，平台可以实时监测客流热力、模拟客流变化并辅助制定调控方案；另一方面，高精度三维模型和持续积累的环境、文物数据，可以作为遗址预防性保护、学术研究和未来修缮决策的重要数字依据。



图16 故宫博物院数字孪生平台，实时监测客流热力、模拟客流变化并辅助制定调控方案[39]

KEY JUDGMENT

综上，空间智能正在成为万亿美元级既有产业的新基础能力，率先赋能世界模型与生成式AI、具身智能、影视娱乐、文旅文博与泛测绘等领域。真实世界正在被转化为机器可以理解、内容产业可以生产、消费者可以进入的数字空间。

© 2026 上海禾赛科技有限公司

3.2 当前产业发展的瓶颈：三维空间数据的匮乏

> **The biggest problem is data. In the case of language models, all the English and all the language that we have on the internet... And so, agentic systems, robotic systems, physical AI—the data is the hardest problem.**
>
> 最大的问题是数据。对于语言模型而言，我们拥有互联网上积累的英语及其他语言数据……而对于智能体系统、机器人系统和物理AI而言，数据是最棘手的问题。
>
> **—— 黄仁勋，NVIDIA创始人兼首席执行官**
>
> NVIDIA，GTC Taipei 2026 Keynote，2026年6月1日

> **What SceniX team is doing is trying to solve this extremely difficult problem in robotics, which is the lack of data. The lack of data in training, the lack of data in evaluation.**
>
> SceniX正在尝试解决机器人领域一个极为棘手的问题，数据匮乏。训练数据不足，评估数据不足。
>
> **—— 李飞飞，World Labs联合创始人兼首席执行官（拟任 AMD 执行副总裁兼首席科学家）**
>
> a16z，Fei-Fei Li on Spatial Intelligence and Robotics，2026年7月28日

> **数据的质量和数量直接决定了AI的能力，而目前人形机器人的相关数据非常稀缺。**
>
> **—— 王兴兴，宇树科技创始人兼首席执行官**
>
> 亚布力中国企业家论坛，主题演讲，2026年3月17日

© 2026 上海禾赛科技有限公司

我们认为，不同行业，对3D空间数据的要求并不相同，从数据数量和质量两个维度出发，呈现如下的分布状态：

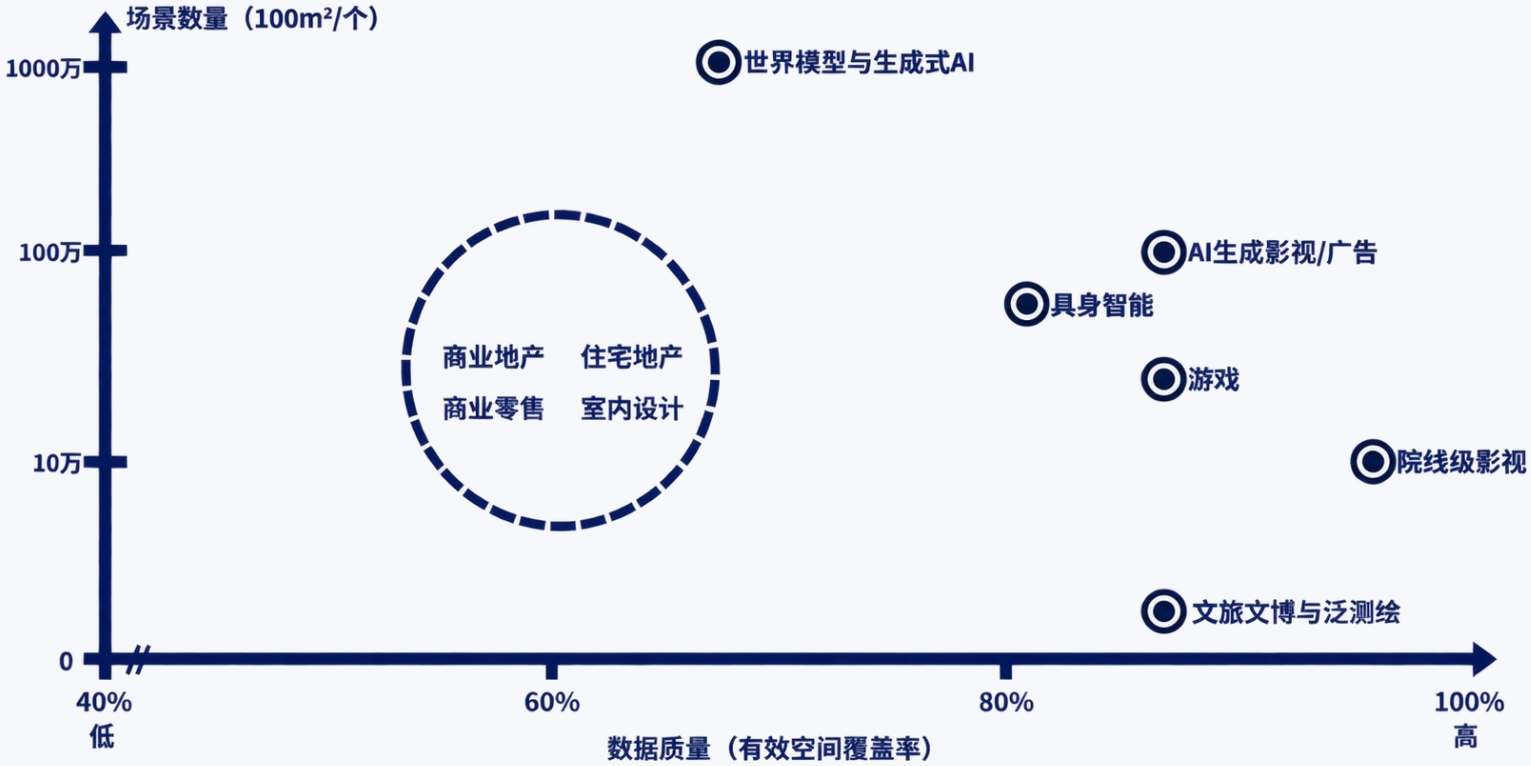


图17 空间数据的质量 × 规模需求图谱

世界模型与生成式AI：核心目标是获得更强的泛化能力，需要覆盖大量、多样化的真实场景来学习通用空间规律。因此对场景规模和覆盖广度要求最高，而对单个场景达到影视级精度的要求相对较低，属于“**高规模要求**”。

具身智能：机器人不仅要识别空间，还要在其中导航、抓取和完成任务，因此数据既要覆盖足够多的环境，也要具备可靠的几何、语义和物理属性。相较基础模型，其单场景可用性要求明显更高，属于“**质量与规模并重**”。

影视娱乐：最终内容直接面向消费者，对视觉真实感、几何一致性和多视角稳定性要求较高。其中，院线级影视面向院线、长视频等高规格制作，对单场景细节和交付质量要求最高；AI生成影视的生产频率和场景数量更大，对单个场景极致精修的要求略低，但更强调规模化生产和快速周转。

文旅文博与泛测绘：核心是对特定文物、古建或景区进行高保真数字化，因此场景数量相对有限，但对几何精度、纹理细节、真实性和长期保存价值要求较高，属于“**低规模、高质量**”应用。

KEY JUDGMENT

当空间智能进入规模化落地，行业真正需要的，不是单点工具，而是能够以全栈能力持续、标准化生产高质量数据，并连接上下游形成生态的平台型玩家。禾赛推出的三维世界数据引擎Kosmo，正是为此而生。

© 2026 上海禾赛科技有限公司

HESAI | KOSMO

CHAPTER 04

Kosmo：空间智能的基础设施

04



4.1 Kosmo：三维世界数据引擎

Kosmo是一个三维世界数据引擎，包含AI空间相机Kosmo One和AI驱动的云平台Kosmo World。Kosmo One将真实世界轻松数字化，Kosmo World将其转化为高质量的三维空间数据，既为具身智能和世界模型及生成式AI提供其最稀缺的真实世界数据“燃料”，也让每个人都能真实记录三维世界，轻松创作身临其境的体验。



图18 Kosmo包含AI空间相机Kosmo One和AI驱动云平台Kosmo World，最终生成超高清3D智能模型

Kosmo One是全球首款 8K 超高清 AI 空间相机[1]，以毫米级几何精度真实还原三维世界，让高质量空间建模像拍照一样简单。

- **极致画质：**Kosmo One 集成超高空间解析力激光雷达、8K 超高清影像系统、一英寸大底 CMOS 与高精度 IMU，微秒级同步采集空间几何、色彩与纹理信息；融合 3DGS 重建技术，实现毫米级几何精度，影像级还原真实世界的每一处细节。
  - **8K影像级画质[2]，细节更丰富：**Kosmo 搭载 8K 超高清影像系统，采用 1 英寸大底 CMOS 与 8K 鱼眼相机，实现丰富细节捕捉与 360° 全景覆盖。独家光学系统设计，确保全视场照度均匀，有效抑制鬼影与眩光；自研面向 3DGS 的 ISP（Image Signal Processor）系统，实现多视角间色彩与曝光一致，为高质量三维空间重建奠定基础。
  - **毫米级精度，几何更真实：**禾赛自研高空间解析力激光雷达，具备 3.5 mm 超高测距精度与无限加密扫描技术，高效采集高精度三维几何信息。结合高精度 IMU 对设备运动与位姿的持续追踪，并与激光雷达、影像系统实现微秒级时空同步，精准还原墙面、边缘与物体轮廓，真实重建空间结构与物体间的几何关系。

© 2026 上海禾赛科技有限公司

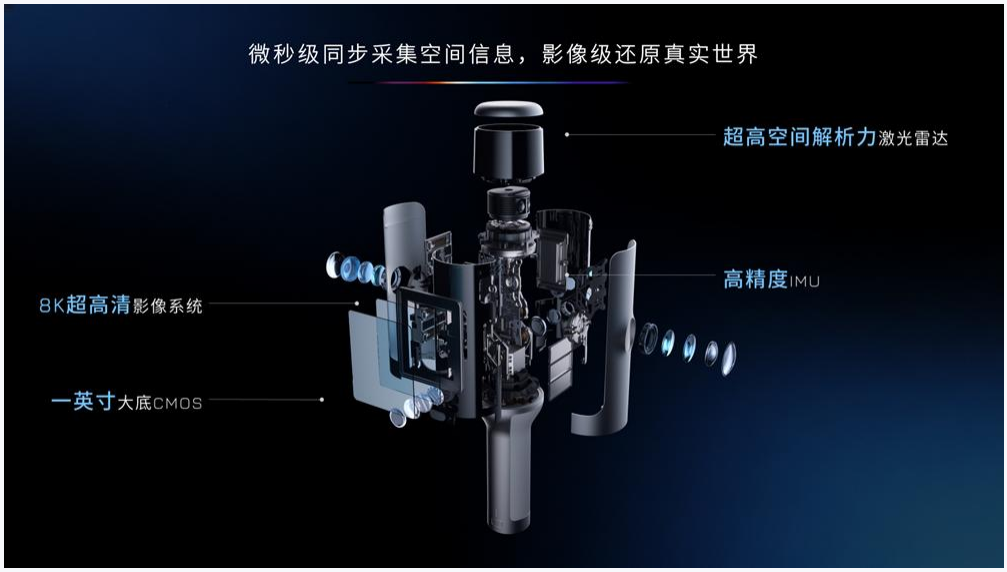


图19 Kosmo One爆炸图

- **易于使用：**Kosmo One 具备行业首创的即开即扫、智能指引、一键续扫功能，小白也能轻松上手；配合 Kosmo World 云平台处理，模型生成速度最高可达行业 5 倍[4]，让人人都能轻松创作专业级三维内容。
  - **即开即扫，启动更快捷：**无需复杂的设备连接、软件配置与参数设置，开机即可迅速进入扫描状态，让空间采集真正做到拿起即扫、随时开拍。
  - **智能指引，操作更简单：**实时可视化呈现已采集与待补区域，智能引导采集路径与覆盖范围，无需专业培训，也能完成高质量采集。
  - **一键续扫，采集更灵活：**支持不同时空的数据智能融合，无需一次完成整场采集，中途可随时暂停，回来后一键续扫，轻松补全完整三维空间。
  - **快速采集，采集更高效：**15分钟即可完成100㎡空间采集[3]，大幅缩短现场作业时间，让高质量三维空间采集更高效。



图20 一站式3D内容创作与管理平台Kosmo World示意图

Kosmo World是一站式 3D 内容创作与管理平台，集 3D 资产与云服务于一体，支持 3D 模型的生成、编辑与分享，让三维内容从生成到应用更简单高效。它的背后，是 Kosmo 三维世界数据引擎的技术底座——**Kosmo Spatial Core（KSC），**一套 AI 原生的 3D 空间重建与理解系统：通过智能补全和 3D 语义分割，将真实世界转化为 AI 可理解、可交互、可训练的超高清三维智能空间。

© 2026 上海禾赛科技有限公司

- **未扫到的AI生成补全：**KSC一次重建即可同步生成 3DGS、mesh、depth 三种形态的三维数据，在视觉、交互和空间几何上都做到精准，画质最高可达 8K。原生 AIGC 能力则让空间更完整：未被扫描的区域自动补全，材质与纹理智能识别、重建，多视角信息无缝融合，最终生成完整、连贯、可自由漫游的三维场景。
- **空间万物皆可识别与交互：**借助 AI 驱动的 3D 语义分割技术，场景中的每一个物体都能被独立识别与定义，并转化为可交互、可编辑的三维实体 —— 它们不再只是静态三维模型，而是具备语义信息与操作属性、可被物理 AI 理解的空间对象。真实世界由此转化为面向具身智能训练的 Sim-ready 数据，帮助机器人在高还原度的数字空间中完成运动规划、交互与操作训练，进一步缩短从仿真到现实的距离。



图21 Kosmo能够精准识别空间中的物体并做3D语义分割

- **快速生成，让空间重建更流畅：**依托 KSC 的强大能力，Kosmo World 云平台的模型生成速度最高可达行业 5 倍[4]。传统三维重建动辄数天的等待，被压缩到数小时。

Kosmo致力于构建开放的空间智能生态，以开放真实世界空间数据为核心，连接世界模型与生成式AI开发、具身智能训练与3D内容创作，打通从空间采集、数据生成到AI应用与媒体升维的完整链路。

- **开放数据：**Kosmo开放从相机参数、原始图像、激光雷达、IMU等原始传感器数据，到3DGS、mesh等三维空间资产的全栈数据。开发者可自由调用、处理与开发，让真实空间数据成为开放的底层能力。
- **开放生态：**开放与主流创作和仿真平台的连接能力，通过Pixar's RenderMan、NVIDIA Isaac Sim、Unreal Engine等插件，连接影视娱乐、游戏、动画美术与机器人仿真等工作流，让真实空间数据跨平台流转，持续用于物理AI训练与3D内容创作。

CONCLUSION

Kosmo 打通了从真实空间采集、AI 三维重建到空间理解与内容创作的完整链路，让真实世界快速转化为高精度、可编辑、可交互的三维数字资产。无论是具身智能训练、世界模型和生成式AI开发，还是影视娱乐、文旅文博和泛测绘，Kosmo 都让真实空间成为可复用、可创造的数字资源。让真实世界不止被记录，更能被理解、被创造、被赋予无限可能。

© 2026 上海禾赛科技有限公司

4.2 Kosmo如何解决问题

Kosmo从象限“右上角”切入，先攻克“大规模、高质量”这一行业最难的数据供给问题，聚焦高价值、高壁垒、强数据刚需的应用场景。

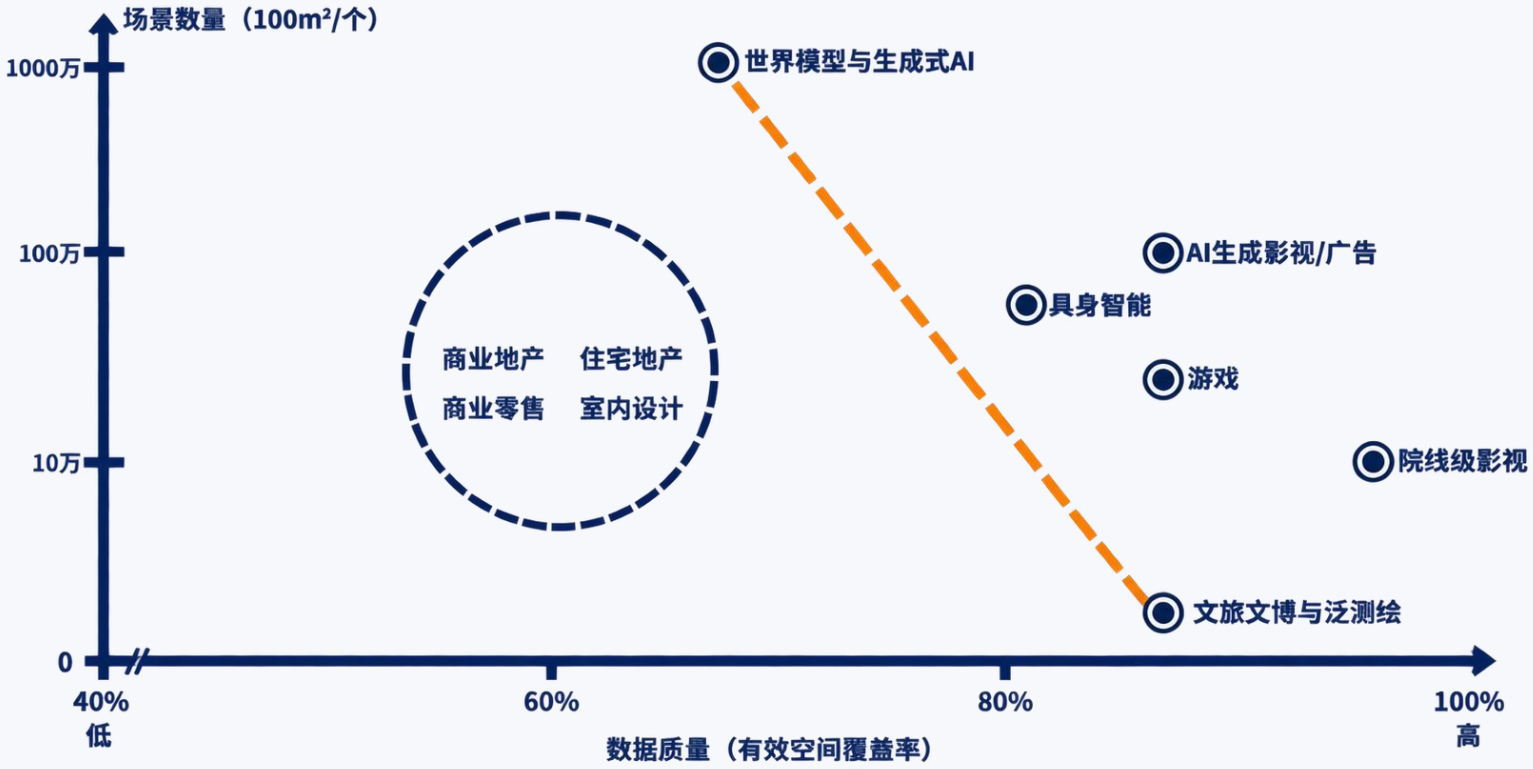


图22 禾赛从大规模、高质量的“右上角”切入市场

目前，Kosmo的核心应用主要分为两类：Sim-ready，服务于世界模型与生成式AI、具身智能；以及Production-ready，服务于影视娱乐、文旅文博与泛测绘。

4.2.1 Sim-Ready

4.2.1.1 世界模型与生成式AI

世界模型

世界模型要继续从二维内容向三维空间演进，缺的已经不只是更多图片和视频，而是能够告诉模型“世界如何在三维空间中成立”的数据。传统2D视频能够提供丰富的颜色、纹理和动态信息，但对于物体真实的空间位置、距离、深度以及不同视角之间的对应关系，只能依靠模型推测。这也是为什么当前前沿研究不断将深度、几何、相机位姿引入视频生成模型：GEN3C通过3D点云和深度信息提升世界一致性[40]，World-consistent Video Diffusion则直接引入XYZ三维坐标监督[41]。

World Labs最新发布的Atlas[21]，将这个趋势展示得更加直观。

© 2026 上海禾赛科技有限公司

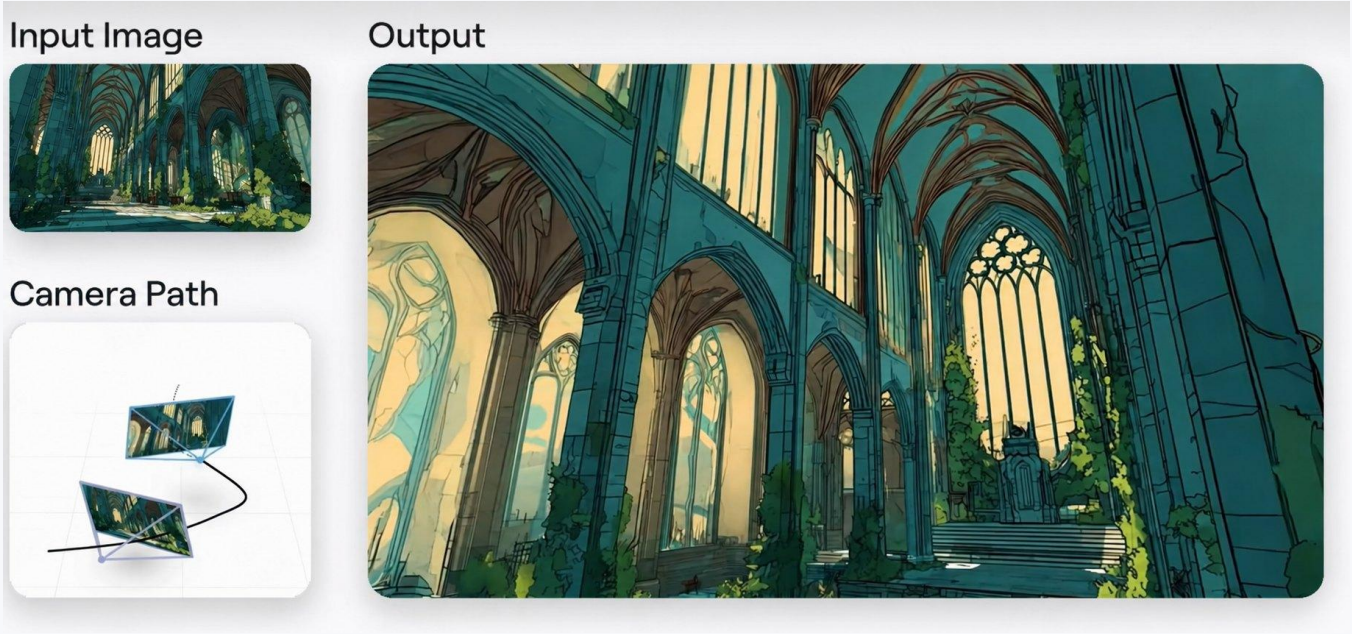


图23 Atlas可以基于一张或多张参考图，在任意指定的相机位置和角度生成新视角，同时保持与输入图像内容和几何结构的一致[21]

为什么Atlas呈现出来的效果如此惊艳？因为它生成的世界能够保持很强的三维空间一致性和几何一致性：就像你在真实房间里走动，不管走到哪里，家具的大小、位置和遮挡关系都不会乱变。这背后依赖的是世界模型对空间结构和观测关系的建模能力。输入越少，Atlas越依赖学习得到的世界先验（就像人脑见过无数真实场景，天然知道桌子是硬的、墙不会突然变形，这套从大量数据里学来的“常识”就是习得世界先验）。

所以，Atlas的成功真正说明的是：世界模型越向三维演进，其能力就越依赖能够清晰表达空间关系的三维数据。世界模型要获得更强的镜头控制、多视角一致性和几何稳定性，就需要越来越多具有准确位姿、深度、几何结构和跨视角对应关系的真实3D数据。训练数据的规模越大、质量越高、场景覆盖越丰富，模型学到的空间规律和世界先验就越充分，从而能够在更少的输入条件下生成更一致、更合理的空间。**而这恰恰是Kosmo的核心能力：规模化生产具备准确位姿、深度、几何结构与跨视角对应关系的高质量真实三维数据。**Kosmo每重建一个真实空间，就能够裂变出这个空间中万条以上的训练轨迹，且每一条轨迹都高度贴近真实环境与真实交互，每一帧都有精确的相机位姿及深度信息，帮助世界模型更高效地学习和迭代，推动三维生成结果持续逼近真实世界。

一次真实世界重建，衍生万条以上可复用训练数据。

© 2026 上海禾赛科技有限公司

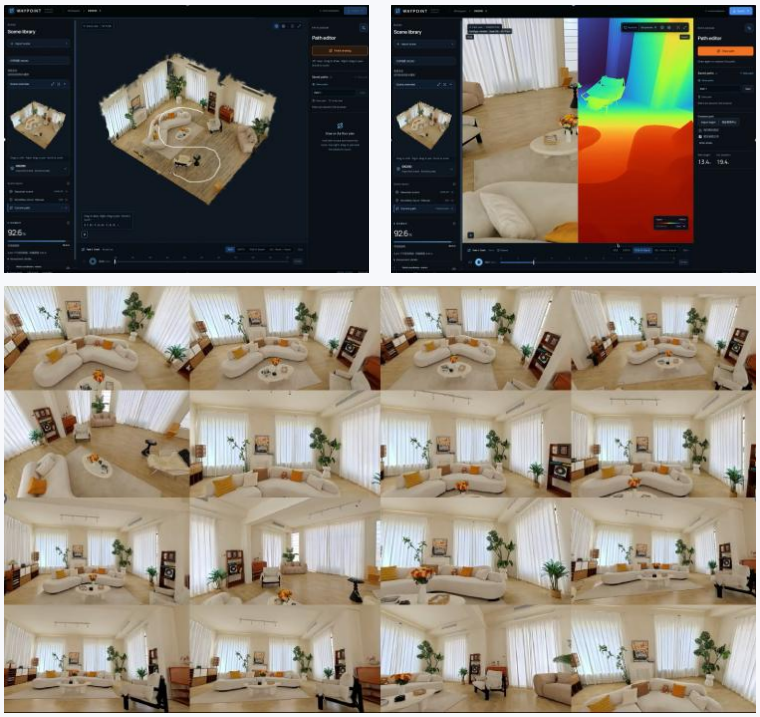


图24 在Kosmo重建的真实场景中，可以裂变出万条以上轨迹，每一条轨迹都高度贴近真实环境与真实交互，每一帧都有精确的相机位姿及深度信息

更重要的是，这类实测数据作为现实校验依据的作用，难以被纯仿真数据完全替代：仿真中的几何与观测由预设规则产生，内部一致并不等于与真实世界一致，仍需要独立的实测参考来识别和校正偏差。因此，Kosmo所提供的价值，不仅是增加训练数据，更是为世界模型提供可追溯的现实依据；谁能持续交付高质量、丰富场景数据，谁就有望成为世界模型公司重点采购与合作的对象。当一个真实空间能够被多个模型、多个版本反复用于训练与校验，Kosmo的商业空间也将从一次设备销售，延伸到持续积累和供给真实世界数据资产。

生成式AI

随着生成式AI快速发展，人们只需输入一段文字或几张图片，就能生成看起来相当逼真的三维场景。这自然带来一个问题：既然AI已经能够直接生成三维世界，为什么还需要使用专业设备重建真实世界？

答案在于，**生成一个看起来合理的世界，与准确还原一个真实存在的世界，是两件完全不同的事。**

以一间真实客厅为例。生成式AI能够根据照片推测家具的形状、颜色和布局，生成一间视觉上相似的客厅。但如果进一步观察，沙发的位置可能发生偏移，椅子的形状可能有所变化，甚至出现原本不存在的物体。对于创意表达，这些差异未必重要；但对于需要精准空间关系的机器人训练、真实场景重建和专业影视制作，这些差异可能直接影响结果的可用性。

下方最左侧的图，是 GPT-6 Astra 参考 20 张照片生成的三维场景。GPT-6 Astra虽然能够识别并生成楼梯、海报、座椅、边桌等主要元素，但与最右侧Kosmo重建的真实场景相比，仍存在明显的失真。

首先是物体形态和材质失真。 真实场景中的座椅拥有金属框架，表面覆盖着蓬松的浅色毛绒织物，靠背上放置着黑色靠枕；而GPT-6 Astra生成的座椅不仅造型、比例和支撑结构发生变化，毛绒材质的细节也基本丢失。**其次是物体位置与空间关系失真。** 真实场景中，黑色圆桌位于座椅左侧，二者具有明确的相对位置关系；而生成结果中的座椅明显偏向左侧，与圆桌的位置关系发生改变，甚至出现视觉上的重叠。此外，海报、楼梯与周围墙体之间的空间布局也存在偏差，右侧原本连通其他房间的空间结构未得到准确还原。

© 2026 上海禾赛科技有限公司

这说明，仅依赖多张图片输入，生成式模型虽然能够“生成一个看起来合理的房间”，但仍难以稳定、准确地还原真实世界中的几何关系与物理细节。

下方中间的图进一步展示了Kosmo真实三维数据对生成式AI的赋能效果。**在引入Kosmo的3DGS模型作为参考后，GPT-6 Astra生成的场景在空间结构、物体位置及场景布局上明显更接近真实世界**。具体而言，原本被错误放置在房间左侧、朝向发生偏移的座椅，回到了海报正前方，与左侧黑色圆形边桌的相对位置基本吻合；楼梯、墙体等建筑结构也更加接近真实布局，原本错误生成的窗户不再出现。尽管座椅的形态、毛绒材质，以及房间右侧与其他空间的连通关系仍未得到完全准确的还原，但模型对真实场景整体空间关系的重建能力已得到明显改善。



图25 GPT-6 Astra在不同输入条件下的三维场景重建效果对比：20张参考照片（左）、Kosmo 3DGS模型辅助生成（中）与Kosmo真实场景重建（右）

这揭示了两条技术路线之间的重要关系：**生成式AI擅长依据已有信息创造可能的世界，而Kosmo擅长高保真记录真实存在的世界。二者并非替代关系，而是互补关系。**

Kosmo One 负责高精度采集真实空间，Kosmo World 则将采集结果转化为具有几何、深度、位姿和多视角关系的三维资产。这些数据既可以直接用于数字内容生产，也可以作为生成模型训练、评测和条件生成的现实参考，使模型不仅“会生成”，而且“更懂真实世界”。

更重要的是，随着 Astra 一类生成式AI能力持续跃升，高质量真实空间数据的价值并不会被削弱，反而会被进一步放大。原因在于，大模型越强，其对高质量输入和高质量先验的利用能力就越强；模型越希望生成稳定、可控、符合真实空间规律的内容，就越需要来自真实世界的精确三维数据作为支撑。也就是说，Astra 的进步在不断放大 Kosmo 的价值：**更强的生成式AI，需要更强的现实世界数据底座。**Kosmo 所提供的，不只是三维重建结果，更是一种能够被大模型持续吸收、调用和放大的高质量空间数据基础设施。

因此，当生成式AI从“生成看起来像的世界”进一步走向“生成结构正确、关系可信、能够被下游应用直接使用的世界”时，Kosmo的重要性只会进一步提升。

CONCLUSION

无论是世界模型对真实世界的理解与模拟，还是生成式AI对三维场景的准确还原，都离不开高质量的真实空间数据。Kosmo通过规模化重建真实场景，提供具备精确几何、深度、位姿及空间关系的高质量三维数据，帮助模型突破空间一致性与真实性瓶颈。Kosmo，就是世界模型与生成式AI进一步实现空间智能的重要基础设施；禾赛，致力于成为所有世界模型与生成式AI公司的数据供应商。

© 2026 上海禾赛科技有限公司

4.2.1.2 具身智能

具身智能行业长期面临数据供给的“不可能三角”：**高质量、大规模与低成本，三者难以同时兼得**。

当前，具身智能训练数据主要来自四条路径：互联网二维数据、第一人称视角数据（Egocentric）、真机遥操作数据（Teleoperation）和仿真数据（Simulation）。不同的数据来源各有优势，却也面临着不同的局限。

首先，互联网上拥有海量图片和视频，获取成本相对较低，但这些数据主要记录了世界的视觉表象，普遍缺乏准确的深度、位姿、三维空间关系及动作反馈信息。它们能够帮助机器人认识物体、理解场景，却难以直接教会机器人如何在真实三维空间中执行动作，因而难以独立满足具身智能训练对空间信息与物理交互的要求。

相比之下，**第一人称视角数据和真机遥操作数据虽然能够记录更加真实的交互过程，但都难以摆脱对人工采集的依赖**。前者通常需要人类佩戴相机或相关传感设备执行任务，能够记录真实操作过程，但往往缺少与机器人本体直接对应的精确控制信号；后者则需要操作人员控制机器人完成任务，能够获得真实的状态、动作与反馈信息，却受限于人力、设备和操作时间。两种方式都需要持续投入人工资源，难以在保持数据质量的同时大规模降低采集成本。

由此，**仿真成为具身智能突破数据规模与成本约束最具潜力的技术路径之一。**与依赖人工逐次采集不同，机器人可以在数字环境中反复执行任务，自动生成大量训练轨迹，并通过调整场景、物体和任务条件，获得更加丰富的数据。一旦训练环境建立，机器人便可以在其中持续练习，不必每一次都依赖真实设备与人工操作，从而具备更强的规模化潜力。

然而，仿真自身也面临一个根本问题：**虚拟世界与真实世界之间存在差距。**传统仿真环境通常依赖人工建模或程序化生成，物体的外观、材质、光影和空间结构未必能够准确还原现实；与此同时，物理引擎也难以完全模拟真实世界中复杂的摩擦、碰撞和形变。因此，机器人即使在仿真环境中完成了大量训练，进入真实世界后仍可能因为环境差异而出现识别错误、定位偏差或操作失败。这就是具身智能领域长期存在的**Sim-to-Real Gap**。

同时，**仿真数据还存在长尾场景缺失的缺陷。**真实世界是长尾分布的：99%的时间里，机器看到的是晴天、白天、正常道路、正常行人；但决定系统安全性的，恰恰是那1%的极端场景——暴雨中的施工路段、深夜无路灯的乡村公路、突然窜出的儿童、倒地的电线杆。合成数据基于已有真实数据的分布进行采样，本质上是对“常识”的复述而非对“例外”的发现。越是高风险、低频率的corner case，仿真数据越无能为力。这种“长尾失明”是结构性的，不是靠增加算力或改进模型就能解决。此外，当生成模型反复使用前代模型生成的场景或视频训练后续模型，且真实数据与校验不足时，有可能出现“模型坍塌”风险，2024年，牛津大学等机构联合在《Nature》上发表了一篇论文。该研究发现：当AI模型在多代训练中反复使用自己或同类模型的输出作为训练数据时，模型会出现不可逆的退化现象——输出分布逐渐坍缩到少数模式，多样性持续下降，最终失去泛化能力[42]。

© 2026 上海禾赛科技有限公司



图26 用合成数据训练AI，就像让它在自己越来越窄的圈子里“近亲繁殖”，最终与真实世界脱节

CONCLUSION

综上，破解“不可能三角”的关键逐渐清晰：如果能够将真实世界高质量地引入仿真环境，同时保留仿真可编辑、可扩展、可反复训练的优势，就有机会同时改善训练数据的质量、规模与成本。

这正是Kosmo所提供的解决路径。

以机器人在门店中执行任务为例。假设我们希望训练机器人从操作台上拿起一个杯子，再将其放到指定位置。要让机器人学会这项任务，首先需要一个训练环境：机器人必须看清柜台、杯子和周围障碍物，判断杯子的位置与距离，并在执行动作时正确处理接触和碰撞。

传统方式往往需要专业人员测量门店尺寸，逐一建立柜台、桌椅和物品的三维模型，再将它们导入仿真系统。这不仅耗时，而且人工搭建的环境难以完整还原真实门店的材质、纹理与光影。**Kosmo则从真实世界出发，将高质量训练环境的构建，从大量依赖人工建模的过程，转变为以真实采集和自动化重建为基础的数据生产流程。**

首先，通过Kosmo One扫描真实门店，再由Kosmo World完成三维重建，即可获得3DGS、mesh和Depth三类数据。三者并非三份重复的模型，而是在帮助机器人从不同角度理解同一个世界。

3DGS让机器人“看得真”。 它能够高保真还原门店的色彩、纹理、材质和视觉细节。对于依靠摄像头识别环境的机器人而言，仿真画面越接近真实环境，训练时看到的视觉信息就越有参考价值。Kosmo还能够针对不同机器人相机的成像特性进行视觉适配，使仿真中的画面更接近机器人在真实门店中实际获得的视觉输入。

Mesh让机器人拥有可交互的几何环境。 它提供柜台、桌椅和杯子等物体的三维形状与空间结构，使仿真系统能够进一步建立碰撞模型，并配置可供机器人操作的对象。例如，机器人可以据此判断机械臂是否会碰到桌沿、夹爪应该从哪个方向接近杯子。

© 2026 上海禾赛科技有限公司

Depth则让机器人“知道有多远”。 它提供场景中的距离与深度信息，为机器人学习物体定位、空间距离判断和运动路径规划提供依据。例如，机器人不仅需要识别出一个杯子，还需要知道它在三维空间中的具体位置，才能规划机械臂的运动。

更重要的是，**这三类数据来自同一个真实场景，彼此保持空间对齐。这意味着，机器人在画面中看到的杯子、在几何模型中能够接触的杯子，以及深度信息所描述的杯子，指向的是同一个位置。** Kosmo由此为机器人仿真提供了视觉、几何和空间信息相互一致的环境基础，有助于缩小仿真与现实之间的感知差距。

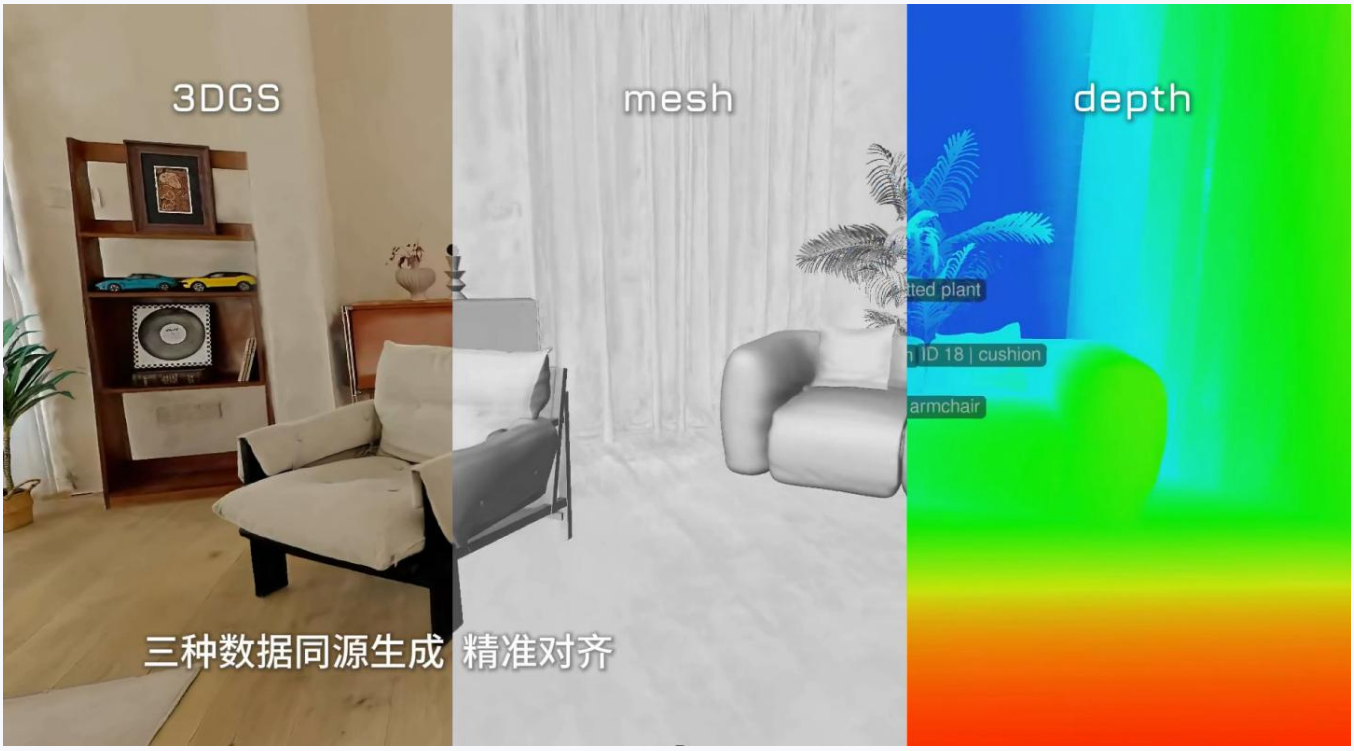


图27 Kosmo同源生成3DGS、mesh、depth三种数据，精准对齐

但这还只是机器人训练的第一步。**Kosmo负责提供真实、可编辑的环境，而让机器人真正学会动作，还需要将这些环境数据与机器人训练系统结合。**

当门店的三维资产接入仿真平台后，训练人员可以加入机器人的数字模型，配置关节运动、物体碰撞和必要的物理参数，再设定“找到杯子—伸手抓取—移动杯子—放到指定位置”的任务。机器人随后可以利用真实操作示教或强化学习等方法，在这个数字门店中反复练习，并根据每次动作的结果持续优化策略。

这时，Kosmo的另一个重要价值开始体现：**它不仅能够还原真实门店，还能让同一个门店衍生出更多不同的训练环境。**

借助Kosmo World的物体识别、语义分割和场景编辑能力，原本固定的桌椅、杯子等物体可以被独立处理。在此基础上，结合仿真工具，训练人员可以把杯子从桌子左侧移到右侧，调整柜台上的物品摆放，增加临时障碍物，或改变机器人的观察视角。通过这样的场景增广，一个真实门店便可以衍生出海量不同布局、不同视觉条件的训练环境。

机器人不再只是反复练习“从固定位置拿起同一个杯子”，而是需要适应不同位置、不同遮挡和不同环境条件下的任务。训练系统可以在这些变化的场景中重新生成、执行和验证动作，积累更加多样化的训练经历，从而帮助机器人提升对环境变化的适应能力。

© 2026 上海禾赛科技有限公司

这也正是Kosmo在数据规模上的重要价值：无需为每一种训练情况都重新采集真实场景，就能在已有高质量环境资产的基础上，持续扩展训练条件。



图28-29 Kosmo生成场景可编辑、可交互，且可以自由切换场景元素，衍生出更多不同的训练环境

与此同时，**Kosmo提供的高质量三维环境数据，还可以与真机遥操作、第一人称视角等方式采集的真实操作数据相结合，进一步放大已有数据的价值**。真实操作数据记录了任务执行过程中的动作与反馈，而Kosmo则为这些数据补充更完整的三维环境与空间关系。通过数据对齐、场景编辑和视觉增广，有限的真实操作样本可以在更多视角、光照和环境条件下得到复用，并结合仿真验证用于训练，从而提高真实数采数据的利用效率，降低对重复人工采集的依赖。

由此可见，Kosmo破解“不可能三角”的逻辑并不是简单提升某一项技术指标，而是改变高质量真实环境数据的生产和使用方式：以真实世界采集保障环境的视觉与几何质量，以场景编辑和仿真增广扩大训练环境的多样性，再通过自动化重建与资产复用降低新增环境的生产成本。

© 2026 上海禾赛科技有限公司



图30 Kosmo 将真实世界转化为 Sim-ready 的三维空间数据

CONCLUSION

因此，在具身智能场景，Kosmo不只是提供三维场景数据，而是为具身智能交付一套源自真实世界、可训练、可交互、可复现评测的完整数据底座。算法决定机器人的上限，而数据底座，决定具身智能能不能真正走出实验室，实现产业级规模化落地。

© 2026 上海禾赛科技有限公司

4.2.2 Production-Ready

Production-ready主要包括影视娱乐与文旅文博。两类行业的生产目标不同：影视娱乐强调真实空间能否被取景、运镜、编辑和反复调用，文旅文博更强调对真实场景的高保真记录、线上展示与长期保存；但底层需求高度一致——都需要把现实世界转化为可以直接进入生产流程的高质量3D空间资产。Kosmo在其中扮演的角色，不是单纯完成一次扫描，而是让真实世界从“被记录的对象”变成“可持续使用的数字资产”。

4.2.2.1 影视娱乐

影视娱乐作为production-ready最具代表性的应用场景，现正处于爆发式增长阶段。

AI已经快速进入影视生产：中国市场首先表现为AI短剧、AI漫剧的大规模增长，2026年第二季度国内主流平台上线微短剧23.9万部，其中AI剧15.3万部，占比约64%[43]；海外成熟影视工业也在加快引入AI，Netflix已有约300部作品在制作过程中使用生成式AI，覆盖概念设计、预可视化和后期等环节[44]。AI正在显著降低内容生产门槛，但产能增加之后，行业的核心问题也开始从“能不能生成”转向“能不能稳定地生产精品内容”。

真实，成为进入影视生产的第一道门槛。与机器人等场景相比，影视内容最终直接面对观众，人眼会天然判断画面是否真实。几何关系是否准确、色彩是否自然、纹理是否细腻、材质与光影是否可信，都会直接决定一个数字场景能否真正进入影视生产。因此，对影视娱乐而言，数据最核心的要求首先不是“数量多”，而是足够真。只有在真实性成立之后，才有进一步讨论意义。

而当前AI生成无法跨越真实瓶颈的原因，主要集中在**一致性、可控性和经济性**：

一致性。一个精品作品不是一个漂亮的五秒镜头，而是几十、上百个镜头共同构成同一个世界。场景在不同视角、不同机位下必须保持几何和空间关系的一致。

可控性。影视制作本质上是导演对画面的控制。机位在哪里、摄影机如何运动、人物与环境如何构图，都需要被精确设计。而纯Prompt驱动仍带有明显概率性，要求越精确，越可能出现重复生成和“抽卡”。

经济性。因此真正应该比较的，不是“生成一秒视频多少钱”，而是：得到一秒最终可用成片，需要多少钱。反复生成、一致性修复、镜头返工，最终都会进入制作成本。公开制作案例也反映了这一问题。2026年，《华尔街日报》报道，95分钟AI电影《Hell Grind》的制作支出约为50万美元，其中约80%用于算力[45]；该项目的公开制作披露还显示，前25分钟经历了16,181次视频生成，最终形成253个镜头，生成与入选数量之比约为64：1[46]。这并不代表所有未入选结果都是技术失败，却说明精品制作的成本不仅来自单次生成，还来自大量尝试、筛选与调整。

但Kosmo的价值正是围绕“真实”展开：

第一，几何一致与多视角一致。与传统生成式视频逐帧或逐片段生成画面不同，Kosmo基于真实世界的三维数据进行高保真重建，保留场景中物体的尺度、位置、距离及遮挡关系，建立稳定、统一的三维空间。无论摄影机进行高速推进、360°环绕、升降还是跟拍，镜头始终围绕同一个三维场景展开，物体不会因视角变化而被重新生成，建筑比例、物体形态及相对位置得以保持一致，从而有效避免传统生成式视频中常见的物体漂移、空间变形和跨镜头不连贯等问题。换言之，镜头可以自由移动，但镜头背后的世界始终保持一致。

© 2026 上海禾赛科技有限公司



图31 Kosmo重建世界与真实世界具有高度的几何一致性，因此无论视角如何变化，物体形态与相对位置等均保持一致

第二，照片级真实感。在真实几何基础上，Kosmo进一步保留高分辨率纹理、色彩和环境细节。在典型4K显示设备、约2米观看距离下，其重建环境已经能够达到接近肉眼难以察觉明显重建差异的视觉效果，使数字环境不仅“结构正确”，也足以进入高规格内容生产。



图32-34 Kosmo极致空间画质，远超同类方案

© 2026 上海禾赛科技有限公司

跨过“真实”的第一道门槛后，Kosmo通过两条路径赋能影视行业：第一条，是把真实世界直接转化为可拍摄、可编辑、可复用的三维数字资产，进入影视生产流程；第二条，是把真实世界转化为高质量训练数据，为AI视频生成模型提供理解三维世界的“燃料”。

➢ **第一条路径：直接提供真实三维资产，进入影视生产全流程。**

- **勘景预演阶段：把主创团队从现实世界拉进数字空间。**

  传统流程需要导演、摄影、美术和制片团队多次前往现场勘景，并围绕真实场地反复确认机位、构图和布景方案。Kosmo可以提前完成场景数字化，让主创团队直接在三维空间中完成选景、机位设计和镜头预演，把“必须去现场才能做的判断”提前到数字世界中完成，让“勘景重、沟通慢、依赖现场”成为历史。



图35 Kosmo数字勘景与虚拟拍摄预演

- **拍摄阶段：实现真实质感与数字创作的革命性融合。**

  传统影视拍摄往往需要演员和剧组前往真实场景，而前后景分离拍摄则将人物表演与背景环境分开制作，再融合为完整画面。简单来说，就是“演员在影棚里演，背景在数字世界里拍”。 演员在摄影棚、绿幕或大型LED屏幕前完成表演，而背景则由提前采集、重建的三维场景提供。例如，演员无需真正站在繁忙的街道上，也可以在摄影棚中完成街头追逐的拍摄。Kosmo通过高保真三维重建，将真实街道、建筑等场景转化为影像级数字背景，并支持在虚拟拍摄系统中灵活调整机位、镜头视角与光影效果，让真实人物与数字环境自然融合。这显著减少了剧组实地取景、设备搬运和场景搭建的成本，万平米的欧洲街景实景搭建需要 2 个月，扫描、重建加编辑只需 1 周，成本节省最高达 80%[5]；还能突破天气、时间、场地及安全等现实限制，让导演拥有更自由、更高效、更可控的创作空间。

© 2026 上海禾赛科技有限公司

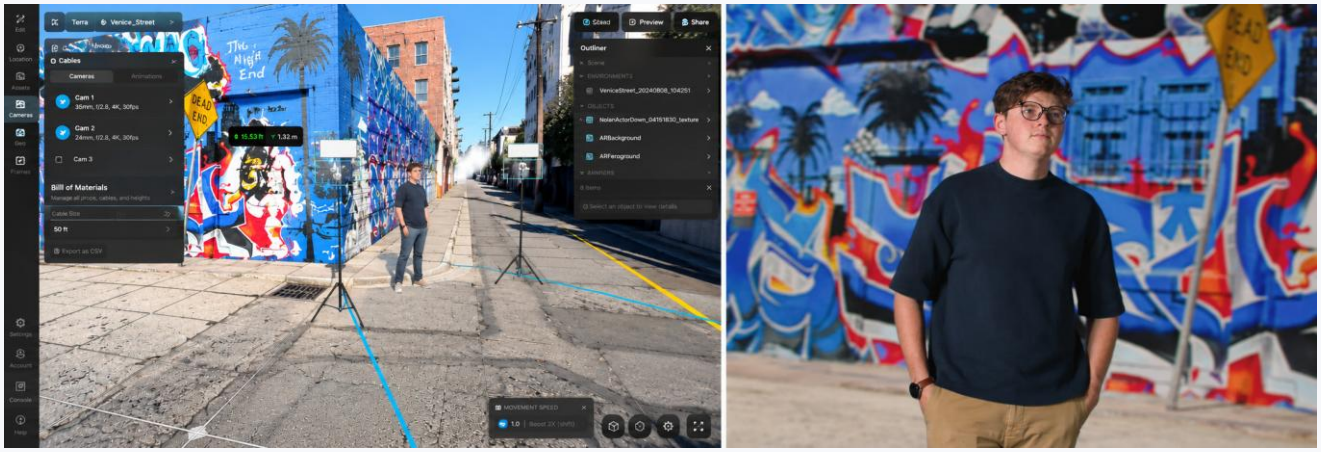


图36 Kosmo赋能影视前后景分离虚拟拍摄示意图：左侧为基于Kosmo真实街景三维重建的虚拟拍摄场景，可进行机位、灯光及人物走位的灵活调整；右侧为将真实人物前景与数字背景融合后的最终拍摄画面

- **后期阶段：打破“拍完即锁定”的行业魔咒。**

  传统影视后期制作中，添加光影、改变天气或延展实景，往往需要技术美术团队花费大量时间重新建立三维模型，再经过复杂的渲染流程才能看到最终效果。一旦导演希望调整光照或特效，团队还需要反复修改、渲染，耗费大量时间与成本。Kosmo正在改变这一流程。 通过将真实场景转化为高质量3DGS数字资产，影视团队可以减少从零重建场景的工作量，并将真实世界数据直接接入专业渲染管线，实现光影调整、环境延展与风格化创作。例如，Pixar’s RenderMan可将Kosmo生成的3DGS数据接入其渲染管线，使艺术家能够在专业渲染环境中对真实场景进行编辑、渲染与风格化处理，并灵活提取其中的三维元素作为创作素材。Pixar还发布了官方教程[6]，向全球创作者展示如何在即将推出的新版本RenderMan中使用Kosmo。这意味着，真实空间不再只是影视拍摄的背景，也正在成为后期制作中可直接调用、灵活编辑的数字资产。由此，Kosmo不仅让真实世界更高效地融入影视后期制作流程，也正在推动3DGS从场景重建技术走向专业影视创作工具，Kosmo 正在成为影视级 3DGS 的标准定义者。

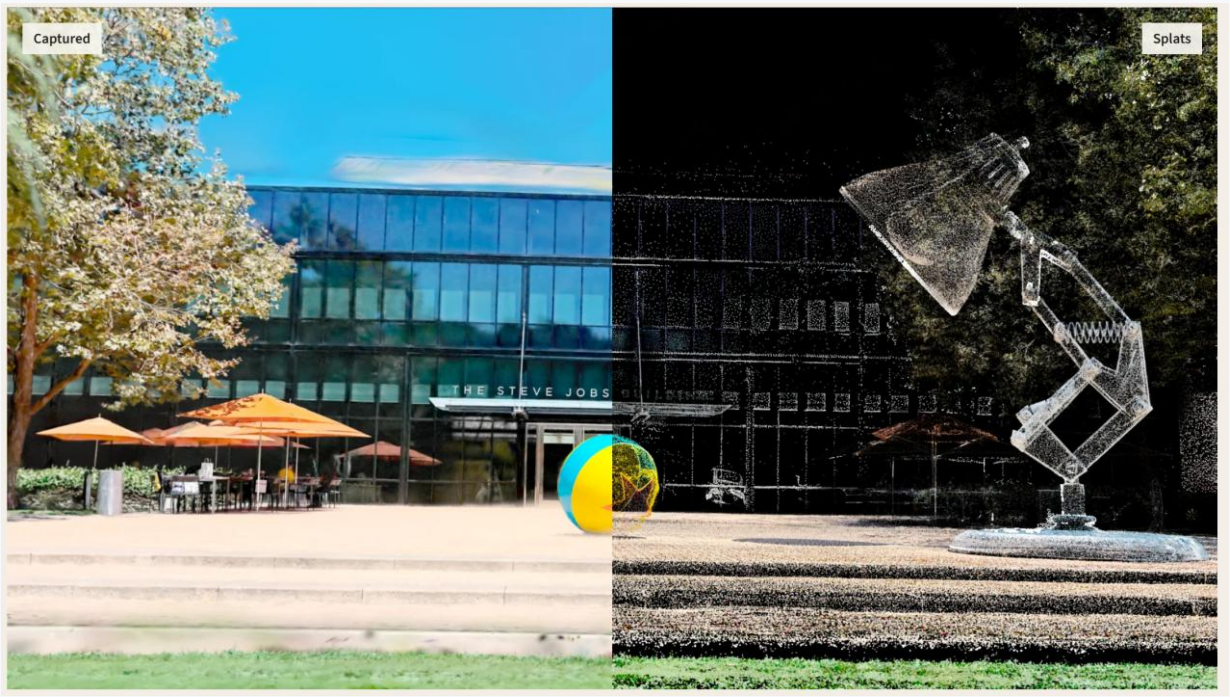


图37 用 Kosmo 扫描的皮克斯广场，再用 RenderMan 渲染的效果[47]

© 2026 上海禾赛科技有限公司

➢ **第二条路径：为视频生成模型提供真实世界“燃料”**

Kosmo的另一条价值路径，是服务视频生成模型。今天的视频生成模型已经越来越擅长生成“好看”的画面，但“好看”并不等于“真实”。尤其在长镜头、多视角和复杂空间中，仍容易出现物体变形、位置漂移、空间关系不稳定等问题。其背后的一个重要原因，是模型过去主要从二维图片和视频中学习“画面长什么样”，但对于一个真实三维世界应该如何成立，获得的约束仍然有限。

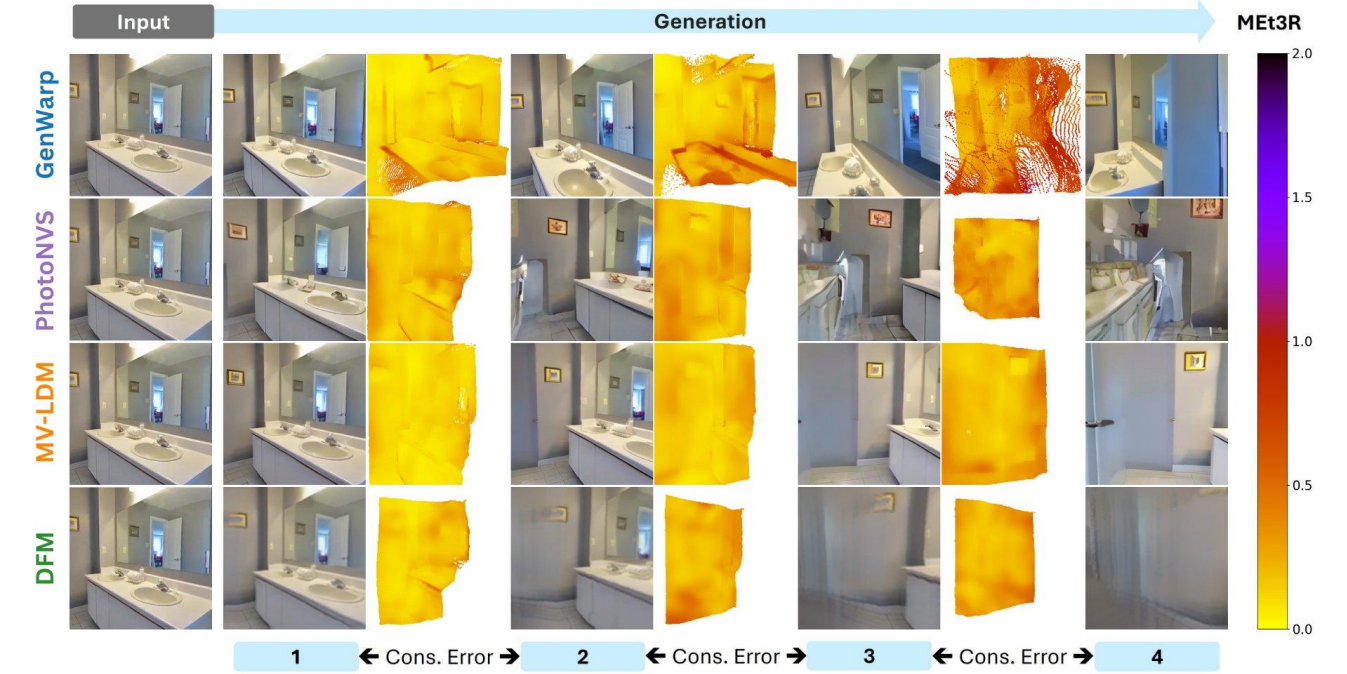


图38 最左边Input是真实输入图——一个卫生间；往右的1 → 2 → 3 → 4为相机不断换机位后，AI连续生成的几个新视角。每两张生成图中间的黄橙色图，为Consistency Error（一致性误差）。右侧色条就是MEt3R指标：越黄越接近0，表示两个视角能被同一个3D结构解释；越红、紫、黑，误差越大，表示两个视角里的几何关系对不上。白色区域一般是无法有效匹配/没有共同可见区域[48]

Kosmo提供的数据来自真实空间，并同时包含精准几何、真实色彩、高质量纹理、多视角关系、深度和空间结构信息。这些数据不仅告诉模型“一个场景看起来是什么样”，还能够帮助模型理解“不同视角为什么仍然是同一个场景”“物体之间应该保持怎样的距离和遮挡关系”“摄影机移动之后空间结构应该如何保持稳定”。

因此，Kosmo可以成为生成模型训练、微调和条件输入的高质量真实世界数据源，在提升模型训练效果的同时，有效降低内容生成成本。例如，借助Kosmo确定人物与环境的空间关系，以及相机位置和运动路径后，获得可用镜头所需的“抽卡”次数可从几十次降至几次，对应生成环节的费用可节省约90%[7]。

CONCLUSION

Kosmo通过两条路径助力影视娱乐发展：第一条路径，为影视行业输出可直接投入生产、高保真的数字实景世界，实现真实场景的数字化复用；第二条路径，驱动视频生成模型构建具备空间逻辑的虚拟世界，从根源解决AI视频画面跳变、镜头失控的行业痛点。两条技术路径殊途同归，直击影视工业化的核心瓶颈：解放创作者的生产力，大幅压缩前期置景与后期返工成本，让创意构想高效落地为可用的影视镜头。

© 2026 上海禾赛科技有限公司

4.2.2.2 文旅文博与泛测绘

文旅文博与影视娱乐相比，其应用目前更轻量，主要集中在虚拟漫游、数字展陈、景区与文化遗产数字化等场景。文化遗产是人类最珍贵也最脆弱的资产：风化侵蚀、自然灾害、城市变迁，都可能让千年文明以不可逆的方式消逝。传统文物数字化往往需要“专业扫描设备+大量摄影+人工点云处理+三维建模”，巴黎圣母院火灾后的数字模型制作动用了12台激光扫描仪、7名工程师、4.6万张照片[49]，耗时超过一年，说明传统高精度方法非常依赖设备、人工和后处理。这正是传统数字化手段的缩影：依赖专业团队的摄影测量与人工建模，周期长、成本高，长期以来只有极少数顶级机构负担得起，绝大多数遗产从未被完整记录；即便是已经完成数字化的，成果也大多是躺在硬盘里的静态档案，既无法走近公众，也难以支撑再创作。

Kosmo通过“LiDAR精确测距+高清影像采集+AI算法+3DGS自动重建”，将原本重人工、重设备的文化遗产数字化流程压缩为一次采集、自动成模，并形成可同时用于保护修复、数字存档和沉浸展示的3D资产。



图39 目前Kosmo 已与世界遗产数字化先锋TRIC 展开合作，以成都杜甫草堂为蓝本创作沉浸式数字艺术作品

CONCLUSION

在文旅文博与泛测绘方面，Kosmo以高精度三维重建能力，大幅降低文化遗产数字化的时间与人力成本。一个人、一台设备，即可以8K画质与毫米级几何精度完整记录一处遗产。

4.2.2.3 立体导航

Kosmo的核心能力，是把真实商业空间数字化为可识别、可理解、可持续更新的三维世界数据，让位置服务从“知道一个地点在哪里”，进一步升级为“理解这个地点里面到底是什么样”。传统二维地图主要解决门店级定位，但真实消费决策往往发生在更细的空间尺度：店内怎么走、商品在哪里、某个座位视野如何、一个展览或门店到底值不值得专程前往。Kosmo将门店、楼层、货架、商品陈列和空间关系完整数字化，使服务能力从“导航到一家店”继续延伸到“导航到店内空间”，甚至“找到一件具体商品”——立体导航。

© 2026 上海禾赛科技有限公司

对于目标明确的用户，Kosmo解决的是“信息不够精准”。比如购买演唱会门票时，网页上的座位图看起来位置很好，但真正到了现场，才发现前方有立柱或设备遮挡，距离和视角也与想象不同。传统平面图告诉用户的是“座位在哪里”，但很难回答“坐在那里究竟能看到什么”。当真实场馆被三维数字化后，用户可以在购买前直接了解座位与舞台、遮挡物和周围空间的真实关系，把原本只能到现场才知道的信息，提前到消费决策之前。

对于目标模糊的用户，Kosmo解决的是“体验不够真实”。现实中，人们常常因为一家热门门店、一个展览或一个商业空间专门安排时间，甚至绕很远的路前往，但真正到达后，却发现空间、环境和氛围与线上宣传存在明显落差。图片和短视频只能展示被选择过的有限视角，而Kosmo可以把真实空间完整搬到线上，让用户在出发前就能自由查看和感受真实环境，再决定是否值得专程前往。消费决策由此从“看介绍、靠想象”变成“先看真实空间，再做决定”。

对商家而言，这套能力的价值也不止于展示。门店一旦被数字化，就从一次性的线下经营场所变成可以长期保存、持续更新和反复使用的数字资产；进一步地，楼层、货架、商品陈列和空间利用都可以被量化和分析，为动线设计、陈列优化和商品运营提供新的数据基础。在空间商业化领域，Kosmo已与中国领先的数字地图及导航平台高德达成战略合作，首批三维重建的重美术馆、偏锋画廊已上线高德扫街榜“飞行街景”。未来，更多真实场景将经由 Kosmo 接入高德这个近 10 亿月活的国民级应用，让三维世界数据服务每一次出行与消费决策。

CONCLUSION

在立体导航方面，Kosmo将传统的“二维位置数据”升级为可理解、可交互、可分析的三维空间数据：对消费者，减少“买完才发现不合适”和“到了才发现不值得”；对商家，则让真实空间第一次成为可以持续运营的数据资产。

© 2026 上海禾赛科技有限公司

HESAI | KOSMO

CHAPTER 05

小结

05



过去十年，AI正从处理数字信息走向真实物理世界。空间智能将感知-理解-推理-交互连接成完整闭环，使智能体不仅能够识别物体，还能理解位置、距离、结构和物理关系，预测环境变化并采取行动。这标志着AI正从Words迈向Worlds，从屏幕中的思考者成长为现实世界的参与者。

本白皮书从空间智能的定义与演进出发，分析了空间智能之于AI发展的重要性，并梳理了空间智能在世界模型与生成式AI、具身智能、影视娱乐、文旅文博与泛测绘等领域的产业机遇。它们虽然面向不同市场，却共同依赖一种关键生产资料：三维空间数据。世界模型与生成式AI需要借此学习真实世界物理规律，机器人需要在能够规模化的真实、可控且可交互的环境中完成训练与验证，内容产业则需要高质量、稳定、可复用的三维资产。空间智能竞争的核心，正在延伸至三维数据的质量提升、规模化生产与持续供给。

面对时代的变化与新需求，禾赛也在不断更新战略定位。如今，我们正站在一个全新的起点：禾赛正拓展为面向物理AI与机器人的全栈基础设施平台：激光雷达帮助机器“感知”世界，三维世界数据引擎Kosmo帮助机器“理解”世界，机器人运动引擎（机器人动力模组）帮助机器与世界“交互”。三层能力共同构成从感知到交互的完整闭环。

在这一闭环中，三维世界数据引擎Kosmo，作为“理解”层的核心，正为空间智能领域的三维数据供给不足的问题提供系统性方案。AI空间相机Kosmo One负责高效、准确地采集真实空间，AI驱动的云平台Kosmo World则将其转化为具有真实几何、高清纹理、准确深度、语义信息和跨视角一致性的三维空间资产。Sim-ready方面，这些资产既能为世界模型与生成式AI提供真实世界“燃料”，也能用于具身智能的训练与验证；Production-ready方面，还可进入影视娱乐、文旅文博与泛测绘等生产流程，成为可编辑、可交互、可持续复用的数字资产。空间智能的核心竞争力，正逐步转至三维数据的质量提升、规模化生产与持续供给。

更重要的是，Kosmo以开放数据和开放生态推动空间智能规模化发展：开放从原始传感器信息到3DGS、mesh等三维资产的全栈数据，并连接Pixar’s RenderMan、NVIDIA Isaac Sim、 Unreal Engine等创作与仿真平台，让真实空间数据能够跨平台流转，在模型训练、机器人仿真和三维内容创作中不断产生新价值。

当空间能够被规模化记录、被智能体理解并被持续复用，空间智能将从前沿技术真正成长为AI的底层能力。禾赛正在构建让智能体感知、理解并作用于世界的全栈基础设施；Kosmo则以真实、开放、可规模化的三维数据，为智能体走出实验室、走进千行百业铺就坚实底座。从Words到Worlds，是AI能力的一次历史性升维，更是一场真实世界数据基础设施的重建。而禾赛与Kosmo，正是这场变革中不可或缺的关键一环。

© 2026 上海禾赛科技有限公司

注：

1. 截至2026年10月，基于公开产品信息，Kosmo One为全球首款集成8K影像系统的AI空间相机。
2. 8K指摄像头支持的最高拍摄分辨率，用户最终导出的文件最高分辨率为6K。
3. 以普通室内房间为例，约15分钟可完成100㎡空间扫描；实际采集时间会因空间结构、面积及采集要求而有所不同。
4. 与行业内同类3DGS三维重建产品相比，在相似硬件、数据集和质量要求下，模型生成速度最高可达约5倍；实际结果因场景、数据规模及参数设置而异。
5. 以典型场景测算：1万㎡欧洲街景由实景搭建约2个月缩短至Kosmo扫描、重建与编辑约1周；200㎡室内场景由传统CG建模约14人天缩短至约2人天。相比实景租赁与搭建，综合成本最高可降低80%。实际结果因场景及制作要求而异。
6. Pixar's RenderMan 官方教程请见：https://renderman.pixar.com/gaussian-splatting-the-pixar-campus
7. 通过真实场景数字化资产与生成式技术结合，提高场景空间结构及镜头生成的一致性，有望减少为获得符合要求的镜头而进行的重复生成与筛选。在特定测试场景下，可用镜头所需的生成尝试次数有望由数十次减少至数次，相关生成费用最高有望降低约90%。实际效果因模型能力、场景复杂度、镜头要求及生成参数而异。

© 2026 上海禾赛科技有限公司

HESAI | KOSMO

REFERENCES

参考文献



[1] F.-F. Li, "From Words to Worlds: Spatial Intelligence is AI's Next Frontier," Dr. Fei-Fei Li, Nov. 10, 2025. Accessed: Sep. 24, 2026. [Online]. Available: https://drfeifei.substack.com/p/from-words-to-worlds-spatial-intelligence

[2] A. Jonas, S. Zhong, W. Tackett, C. Chai, and D. Roser, "Humanoid Horizons: Putting the Robots to Work," Morgan Stanley Research, Oct. 6, 2026, pp. 26–30.

[3] U.S. Department of Commerce, International Trade Administration, "Media & Entertainment," Industry Overview, Aug. 2024. Accessed: Oct. 31, 2026. [Online]. Available: https://www.trade.gov/media-entertainment

[4] The British Museum, British Museum Sets Out Plans to Digitise Fully the Collection, Oct. 18, 2023. Accessed: Sep. 24, 2026. [Online]. Available: https://www.britishmuseum.org/sites/default/files/2023-10/British_Museum_sets_out_plans_to_digitise_fully_the_collection.pdf

[5] The British Museum, Report and Accounts for the Year Ended 31 March 2025, 2025. Accessed: Sep. 24, 2026. [Online]. Available: https://www.britishmuseum.org/sites/default/files/2025-07/Annual_Report_and_Accounts_2024-25.pdf

[6] Stanford Institute for Human-Centered Artificial Intelligence, "What is Spatial Intelligence?" Stanford HAI. Accessed: Sep. 24, 2026. [Online]. Available: https://hai.stanford.edu/ai-definitions/what-is-spatial-intelligence

[7] B. Caulfield, "'Accelerate Everything,' NVIDIA CEO Says Ahead of COMPUTEX," NVIDIA Blog, Jun. 2, 2024. Accessed: Sep. 24, 2026. [Online]. Available: https://blogs.nvidia.com/blog/computex-2024-jensen-huang/

[8] H. Gardner, Frames of Mind: The Theory of Multiple Intelligences. New York, NY, USA: Basic Books, 1983. Accessed: Oct. 31, 2026. [Online]. Available: https://www.pz.harvard.edu/resources/frames-of-mind-the-theory-of-multiple-intelligences

[9] A. M. Turing, "Computing machinery and intelligence," Mind, vol. 59, no. 236, pp. 433–460, 1950, doi: 10.1093/mind/LIX.236.433.

[10] J. Piaget, The Construction of Reality in the Child, M. Cook, Trans. New York, NY, USA: Basic Books, 1954.

[11] J. McCarthy, "Programs with Common Sense," presented at the Symp. on the Mechanization of Thought Processes, Teddington, U.K., 1958. Accessed: Oct. 31, 2026. [Online]. Available: https://www-formal.stanford.edu/jmc/mcc59.html

[12] J. Yang et al., "Thinking in space: How multimodal large language models see, remember, and recall spaces," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 10632–10643, doi: 10.1109/CVPR52734.2025.00994. Accessed: Oct. 31, 2026. [Online]. Available: https://github.com/vision-x-nyu/thinking-in-space

© 2026 上海禾赛科技有限公司

[13] W. Zhang et al., "SPHERE: Unveiling spatial blind spots in vision-language models through hierarchical evaluation," in Proc. 63rd Annu. Meeting Assoc. Computational Linguistics, Vienna, Austria, 2025, pp. 11591–11609, doi: 10.18653/v1/2025.acl-long.568. Accessed: Oct. 31, 2026. [Online]. Available: https://aclanthology.org/2025.acl-long.568/

[14] I. Stogiannidis, S. McDonagh, and S. A. Tsaftaris, "Mind the Gap: Benchmarking Spatial Reasoning in Vision-Language Models," arXiv:2503.19707, 2025. Accessed: Oct. 31, 2026. [Online]. Available: https://arxiv.org/abs/2503.19707

[15] R. Yang et al., "EmbodiedBench: Comprehensive benchmarking multi-modal large language models for vision-driven embodied agents," in Proc. Int. Conf. Machine Learning (ICML), 2025. Accessed: Oct. 31, 2026. [Online]. Available: https://proceedings.mlr.press/v267/yang25f.html

[16] S. Yang, R. Xu, Y. Xie, S. Yang, M. Li, J. Lin, C. Zhu, X. Chen, H. Duan, X. Yue, D. Lin, T. Wang, and J. Pang, "MMSI-Bench: A Benchmark for Multi-Image Spatial Intelligence," in Proc. International Conference on Learning Representations (ICLR), 2026. Accessed: Oct. 31, 2026. [Online]. Available: https://arxiv.org/abs/2505.23764

[17] World Labs, "A Functional Taxonomy of World Models," Jun. 3, 2026. Accessed: Oct. 31, 2026. [Online]. Available: https://www.worldlabs.ai/blog/taxonomy-of-world-models

[18] Google Cloud, "Agent Platform Pricing," Google Cloud. Accessed: Nov. 2, 2026. [Online]. Available: https://cloud.google.com/model-garden/pricing

[19] Runway, "Billing," Runway API. Accessed: Sep. 24, 2026. [Online]. Available: https://docs.dev.runwayml.com/usage/billing/

[20] NVIDIA, "NVIDIA Announces Major Release of Cosmos World Foundation Models and Physical AI Data Tools," Mar. 18, 2025. Accessed: Oct. 31, 2026. [Online]. Available: https://nvidianews.nvidia.com/news/nvidia-announces-major-release-of-cosmos-world-foundation-models-and-physical-ai-data-tools

[21] World Labs, "Atlas: A World Model for Spatial Intelligence," Sep. 1, 2026. Accessed: Oct. 31, 2026. [Online]. Available: https://www.worldlabs.ai/blog/atlas

[22] NVIDIA, "NVIDIA Cosmos," NVIDIA. Accessed: Oct. 8, 2026. [Online]. Available: https://www.nvidia.cn/ai/cosmos/

[23] CB Insights, "The physical AI models market map: Behind the arms race to control robot intelligence," CB Insights Research, 2026. Accessed: Oct. 8, 2026. [Online]. Available: https://www.cbinsights.com/research/the-physical-ai-models-market-map/

[24] OpenAI, "GPT-6 Astra: A new generation of intelligence," Sep. 3, 2026. Accessed: Oct. 27, 2026. [Online]. Available: https://openai.com/index/gpt-6-astra/

[25] T. Ricouard, "Architectural visualization with Astra," OpenAI Developers, Sep. 4, 2026. Accessed: Oct. 27, 2026. [Online]. Available: https://developers.openai.com/blog/architectural-visualization-with-astra

© 2026 上海禾赛科技有限公司

[26] Gartner, "Gartner Forecasts Worldwide AI Spending to Grow 49.5% in 2026," Sep. 16, 2026. Accessed: Oct. 27, 2026. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2026-09-16-gartner-forecasts-worldwide-ai-spending-to-grow-49-point-5-percent-in-2026

[27] S. Ghaffary and R. Torrence, "OpenAI Shifts Value to $1.25 Trillion as Funding Talks Finalize," Bloomberg News, Oct. 22, 2026. Accessed: Oct. 30, 2026. [Online]. Available: https://www.bloomberg.com/news/articles/2026-10-22/openai-shifts-value-to-1-25-trillion-as-funding-talks-finalize

[28] Anthropic, "Anthropic raises $65B in Series H funding at $965B post-money valuation," May 28, 2026. Accessed: Oct. 30, 2026. [Online]. Available: https://www.anthropic.com/news/series-h

[29] IDC, "Humanoid Robotics Commercialization Trends 2026," 2026. Accessed: Oct. 31, 2026. [Online]. Available: https://www.idc.com/resource-center/blog/humanoid-robotics-commercialization-2026/

[30] Google DeepMind, "Gemini Robotics brings AI into the physical world," Mar. 12, 2025. Accessed: Oct. 31, 2026. [Online]. Available: https://deepmind.google/blog/gemini-robotics-brings-ai-into-the-physical-world/

[31] Google DeepMind, "Gemini Robotics On-Device brings AI to local robotic devices," Jun. 24, 2025. Accessed: Oct. 31, 2026. [Online]. Available: https://deepmind.google/blog/gemini-robotics-on-device-brings-ai-to-local-robotic-devices/

[32] NVIDIA Research, "GR00T N1.5," 2025. Accessed: Oct. 31, 2026. [Online]. Available: https://research.nvidia.com/labs/gear/gr00t-n1_5/

[33] L. O. Seastrom, "A New Kind of World-Building: Behind ILM's Award-Winning Partnership with Nerfstudio," Industrial Light & Magic, Mar. 4, 2026. Accessed: Oct. 8, 2026. [Online]. Available: https://www.ilm.com/ilm-nerfstudio-ironheart-hpa-award-interview/

[34] World Labs, "Gaussian Mansion: Creating Playable Worlds with Marble,". Accessed: Sep. 24, 2026. [Online]. Available: https://www.worldlabs.ai/case-studies/3-gaussian-mansion

[35] Epic Games, "Fortnite Brings Global Artists and Music to the Metaverse With the Soundwave Series," Sep. 28, 2021. Accessed: Oct. 31, 2026. [Online]. Available: https://www.epicgames.com/site/news/fortnite-brings-global-artists-and-music-to-the-metaverse-with-the-soundwave-series

[36] SBS World News, "Lee Soo-man Returns to Korean Scene, Unveils A2O: 'I've Spent 16 Years Pondering the Harmony of Art and Technology'," Oct. 12, 2026. Accessed: Oct. 30, 2026. [Online]. Available: https://news.sbs.co.kr/english/article.do?news_id=N1008793666

[37] Roblox, "Roblox Expands Advertising Platform as Essential Channel for the Next Generations," Jan. 2026. Accessed: Oct. 31, 2026. [Online]. Available: https://about.roblox.com/newsroom/2026/01/roblox-expands-advertising-platform

© 2026 上海禾赛科技有限公司

[38] CyArk, "Elephanta Caves," Tapestry by CyArk. Accessed: Oct. 9, 2026. [Online]. Available: https://tapestry.cyark.org/content/elephanta

[39] 故宫博物院，"故宫博物院'十四五'数字孪生成果亮相第九届数字中国建设峰会," Apr. 30, 2026. [in Chinese]. Accessed: Oct. 9, 2026. [Online]. Available: https://www.dpm.org.cn/classify_detail/378876.html

[40] NVIDIA Research, "GEN3C: 3D-Informed World-Consistent Video Generation with Precise Camera Control," 2025. Accessed: Oct. 31, 2026. [Online]. Available: https://research.nvidia.com/labs/toronto-ai/GEN3C/

[41] Q. Zhang, S. Zhai, M. A. B. Martin, K. Miao, A. Toshev, J. Susskind, and J. Gu, "World-consistent Video Diffusion with Explicit 3D Modeling," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025, pp. 21685–21695. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/html/Zhang_World-consistent_Video_Diffusion_with_Explicit_3D_Modeling_CVPR_2025_paper.html

[42] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, and Y. Gal, "AI models collapse when trained on recursively generated data," Nature, vol. 631, no. 8022, pp. 755–759, Jul. 2024, doi: 10.1038/s41586-024-07566-y.

[43] J. Shen (沈杰群)，"《微短剧创作指引（2026年Q2）》发布：AI剧发展趋稳、真人剧明显回暖," 中国青年报 (China Youth Daily), Jul. 24, 2026. [in Chinese]. Accessed: Oct. 27, 2026. [Online]. Available: https://finance.sina.com.cn/jjxw/2026-07-24/doc-iniiwzsf4883782.shtml

[44] Netflix, Inc., "Q2 2026 Letter to Shareholders," Jul. 16, 2026. Accessed: Oct. 27, 2026. [Online]. Available: https://www.sec.gov/Archives/edgar/data/1065280/000106528026000211/ex991_q226.htm

[45] I. Bousquette, "This film cost $500,000 to make. $400,000 was AI compute costs," The Wall Street Journal, May 21, 2026. Accessed: Sep. 30, 2026. [Online]. Available: https://www.wsj.com/cio-journal/this-cannes-film-cost-500-000-to-make-400-000-was-ai-compute-costs-a823b08d

[46] D. Abbatescianni, "In pictures: Higgsfield unveils fully AI-generated feature 'Hell Grind' in Cannes," Screen Daily, May 19, 2026. Accessed: Sep. 30, 2026. [Online]. Available: https://www.screendaily.com/news/in-pictures-higgsfield-unveils-fully-ai-generated-feature-hell-grind-in-cannes/5216871.article

[47] D. Sisson, "A Brush with Reality," Dylan Sisson. Accessed: Oct. 30, 2026. [Online]. Available: https://www.dylansisson.com/stories/kosmo/

[48] M. Asim et al., "MEt3R: Measuring Multi-View Consistency in Generated Images," in Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR), 2025. [Online]. Available: https://openaccess.thecvf.com/content/CVPR2025/html/Asim_MET3R_Measuring_Multi-View_Consistency_in_Generated_Images_CVPR_2025_paper.html

[49] Autodesk, "Notre-Dame de Paris Cathedral Restoration,". Accessed: Sep. 24, 2026. [Online]. Available: https://www.autodesk.com/campaigns/make-anything/notre-dame

© 2026 上海禾赛科技有限公司

HESAI | KOSMO

CONTACT

联系我们

资本市场部：ir@hesaitech.com

市场部：pr@kosmoai.com

Kosmo销售团队：sales@kosmoai.com



版权所有 © 2026 上海禾赛科技有限公司。保留所有权利。

版权与知识产权

本白皮书由上海禾赛科技有限公司（以下简称“我们”）发布。除另有说明或属于第三方的内容外，本白皮书中的文字、图表、图片及其他材料的相关知识产权归我们所有。除法律允许或另有授权外，未经相关权利人事先书面许可，不得复制、改编、转载或将本白皮书的全部或部分内容用于商业用途。依法引用时，应注明来源，保持内容原意，不得作误导性删改。本白皮书所涉及的第三方商标、名称及其他知识产权归各自权利人所有；相关引用不当然表示该第三方与我们存在合作、认可或背书关系。本白皮书不构成对任何专利、商标、软件、数据或其他知识产权的许可。

市场与行业数据

本白皮书旨在分享空间智能领域的技术观点、行业研究及Kosmo的产品与应用理念，仅供信息参考。除另有注明外，文中观点基于发布之日可获得的信息作出，代表我们当时的认识与判断。本白皮书引用的第三方研究、行业数据及市场预测，应结合其原始来源、统计口径、适用范围、时间及假设理解。除明确说明外，我们并未对全部第三方信息进行独立核验。市场预测及本白皮书对行业趋势、应用前景和商业机会的分析存在不确定性，实际市场规模、增长情况及应用发展可能与相关预测或判断存在重大差异。所述行业规模及市场空间不等同于 Kosmo 能够取得的市场份额、订单或收入。

前瞻性陈述与安全港声明

本白皮书包含前瞻性陈述。对于符合适用条件的前瞻性陈述，拟援引美国《1995年私人证券诉讼改革法》项下的“安全港”保护。这些前瞻性陈述可通过“将”“预期”“预计”“目标”“未来”“计划”“相信”“潜在”“有望”“致力于”等词语或其他类似表述识别，但并非所有前瞻性陈述均包含这些词语。本白皮书中的前瞻性陈述包括有关空间智能及相关应用领域的发展前景、Kosmo 的应用拓展及预期效益、开放生态建设，以及在空间智能领域的业务目标和发展方向等陈述。该等陈述仅反映作出时的预期、判断及假设，不构成对未来结果的保证。前瞻性陈述涉及固有风险及不确定性。多项因素可能导致实际结果与相关陈述明示或暗示的结果存在重大差异，包括但不限于：空间智能、世界模型与生成式AI、具身智能及相关应用领域的发展速度和市场需求；三维采集、重建及人工智能技术的发展与实际应用效果；Kosmo 满足具体应用场景要求的能力；与创作及仿真平台的适配情况；客户及开发者的采用程度；合作生态建设及相关业务的商业化进展；以及与上述事项有关的假设。更多风险信息请参阅 Hesai Group 向美国证券交易委员会（SEC）提交的文件，以及其在香港联合交易所有限公司披露的相关报告及公告。读者不应过度依赖前瞻性陈述。除适用法律法规、监管要求或上市规则另有要求外，我们不承担因新信息、未来事件或其他原因而更新本白皮书中任何前瞻性陈述的义务。

产品性能与数据使用

本白皮书涉及的产品性能、内部测试结果及效率、成本比较，应结合相应的测试条件、比较对象、计算口径和应用场景理解。实际效果可能因采集环境、场景复杂程度、处理条件及具体工作流程而异，不代表在所有场景下均可取得相同结果。本白皮书中有关“开放数据”“开放生态”及跨平台使用的描述，不表示相关数据或第三方内容可不受限制地使用；实际使用仍应遵守适用法律、相关协议及授权范围。

非投资建议、非要约及责任限制

本白皮书不构成投资建议或招股文件，亦不构成在任何司法管辖区出售 Hesai Group 股份、美国存托股份或其他证券的要约，或购买、认购该等证券的要约招揽或邀请。本白皮书未考虑任何读者的具体投资目标、财务状况或需求，不能替代 Hesai Group 依法披露的信息。读者应结合自身情况及相关公开披露文件独立判断，不应仅依据本白皮书作出投资或其他重大商业决策。在适用法律允许的范围内，我们不对本白皮书中的一般性信息适用于读者特定目的作出明示或默示保证。本声明不排除或限制依法不得排除或限制的责任，也不减损我们或 Hesai Group 各自依法应承担的信息披露义务、证券法律责任或有效合同项下的义务。