导航世界模型

阿米尔·巴尔1 Gaoyue Zhou2 丹尼·陈3 特雷弗·达雷尔3 扬·勒库恩1,2
Meta 上的 1FAIR 2纽约大学 3. 伯克利人工智能研究中心

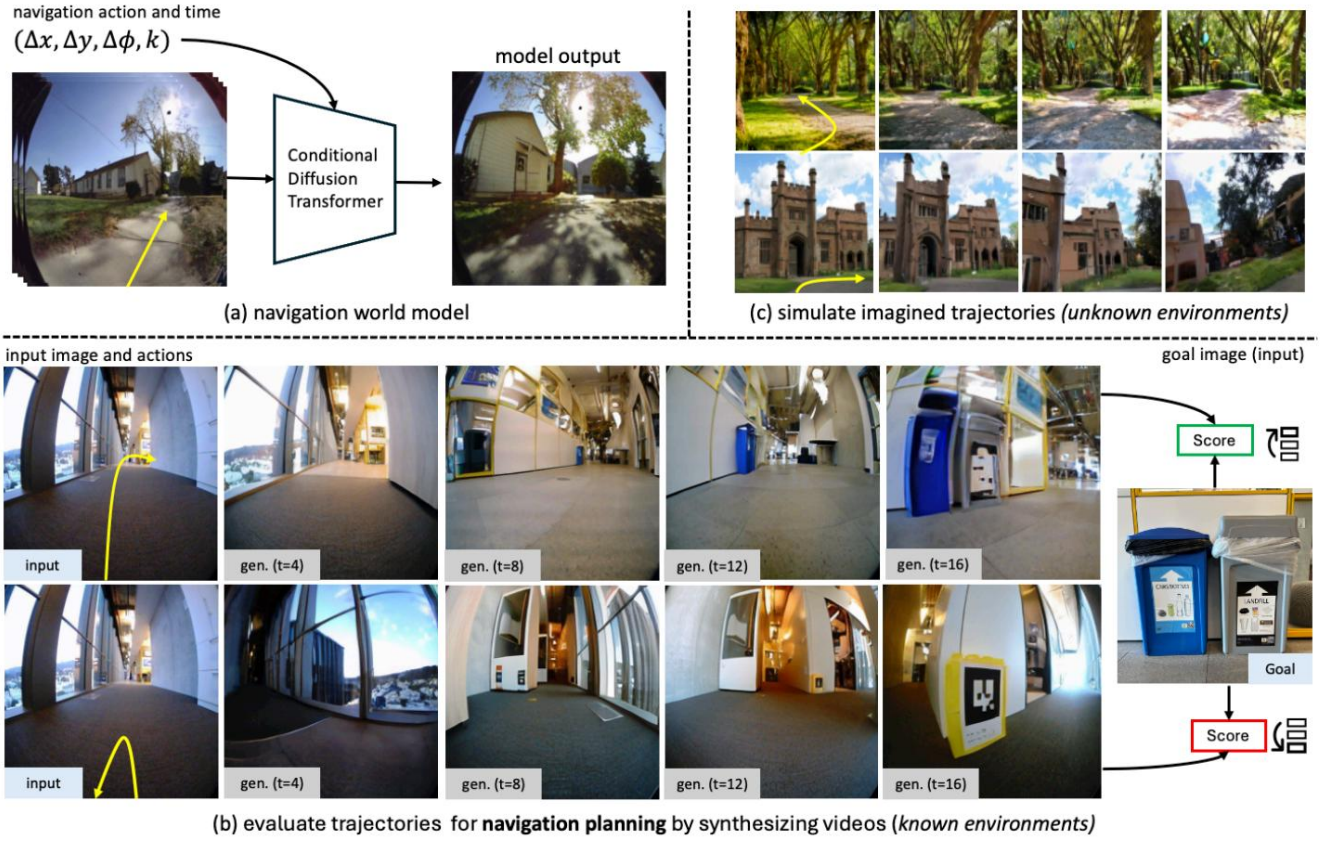

图 1. 我们根据机器人的视频片段及其相关的导航动作训练导航世界模型 (NWM) (a)。之后
训练过程中,NWM 可以通过合成视频并评估最终帧与目标 (b) 的相似度来评估轨迹。我们使用
NWM 可以从零开始规划或对专家进行排序的导航轨迹,从而提高下游视觉导航的性能。在未知领域
环境中,NWM 可以从单幅图像中模拟想象的轨迹（c）。在上面所有的例子中,模型的输入都是第一个
图像和动作,然后模型会自回归地合成未来的观察结果。点击图像即可在浏览器中查看示例。

抽象的

导航是具有视觉运动能力的智能体的一项基本技能。我们引入了导航世界模型
（NWM），一种可控视频生成模型,可根据过去的观察预测未来的视觉观察
和导航操作。捕捉复杂环境
动态,NWM 采用条件扩散变换器 (CDiT),对人类和机器人代理的多种自我中心视频进行训练,并按比例放大

高达 10 亿个参数。在熟悉的环境中,NWM
可以通过模拟来规划导航轨迹，
评估他们是否达到了预期目标。不像
具有固定行为的监督导航策略,NWM
可以在规划过程中动态地纳入约束。
实验证明了其在从零开始规划轨迹或通过对采样的轨迹进行排序方面的有效性
外部政策。此外,NWM 利用其
学习视觉先验来想象不熟悉的轨迹
环境,使其成为一种灵活的
是下一代导航系统的强大工具1 。

1项目页面： https://amirbar.net/nwm

1. 简介

导航是任何具有视觉的生物的一项基本技能,它通过允许智能体

寻找食物、庇护所,并躲避捕食者。为了
成功导航环境,智能代理主要
依靠视觉,使他们能够构建表象
周围环境来评估距离并捕捉环境中的地标,所有这些对于规划导航路线都很有用。

当人类智能体进行规划时,他们通常会考虑约束条件和反事实条件来设想未来的轨迹。另一方面,目前最先进的机器人导航策略[53, 55]是“硬编码”的,训练完成后,很难引入新的约束条件(例如,“不”

左转”。当前监督视觉导航模型的另一个局限性是它们无法动态分配更多计算资源来解决难题。我们的目标是设计一种能够缓解这些问题的新模型。

问题。

在这项工作中,我们提出了一个导航世界模型
(NWM),经过训练可以预测
基于过去帧表示的视频帧和
动作(见图1(a))。NWM 在视频上进行训练
从各种渠道收集的镜头和导航动作
机器人代理。训练完成后,NWM 可以用于规划新的
通过模拟潜在导航来规划导航轨迹
计划并验证它们是否达到目标(见图1(b))。为了评估其导航技能,我们测试了NWM
在已知环境中,评估其规划新事物的能力
轨迹可以独立地或通过对外部
导航策略。在规划设置中,我们使用 NWM
模型预测控制(MPC)框架,优化
使 NWM 能够达到目标的动作序列
目标。在排名设置中,我们假设可以访问现有的导航策略,例如NoMaD [55],它允许
我们对轨迹进行采样,使用 NWM 进行模拟,并且
甄选最佳。我们的 NWM 达到了最先进的
与现有方法相结合时具有独立的性能和有竞争力的结果。

NWM 在概念上类似于最近基于扩散的
基于模型的离线强化学习的世界模型,例如 DIAMOND [1]和 GameNGen [66]。然而,
与这些模型不同,NWM 是在广泛的

范围的环境和实施例,利用来自机器人和人类代理的导航数据的多样性。

这使我们能够训练大型扩散变压器模型
能够根据模型大小和数据进行有效扩展
适应多种环境。我们的方法也
与新视图合成(NVS)方法的相似之处
NeRF [40]、 Zero-1-2-3 [38]和GDC [67],其中
我们汲取灵感。然而,与 NVS 方法不同,我们的
目标是训练一个单一的模型,用于跨不同的导航
环境和模型时间动态来自自然

视频,无需依赖 3D 先验。

为了学习 NWM,我们提出了一种新的条件扩散变换器 (CDiT),用于预测下一张图片
给定过去的图像状态和动作作为上下文,状态。与 DiT [44]不同,CDiT 的计算复杂度是线性的
相对于上下文帧的数量,它可以缩放
对于在不同环境和实施例中训练的多达 1B 个参数的模型来说,效果显著,所需的计算量减少了 4 倍
与标准 DiT 相比,FLOP 性能更佳
未来的预测结果。

在未知环境中,我们的结果表明 NWM
接受无标记、无行动、无奖励训练的好处
来自 Ego4D 的视频数据。定性地,我们观察到改进的
单幅图像上的视频预测和生成性能(见图1(c))。定量分析显示,使用额外的未标记数据,NWM 在以下情况下可以产生更准确的预测:

在保留的 Stanford Go [24]数据集上进行评估。

我们的贡献如下。我们引入了一个导航世界模型 (NWM),并提出了一种新颖的条件扩散变换器 (CDiT),它可以有效地扩展到 1B 参数,同时显著降低

与标准 DiT 相比,计算要求更高。我们
使用来自不同机器人代理的视频片段和导航动作来训练 CDiT,通过独立或与外部导航一起模拟导航计划来实现规划

策略,实现了最先进的视觉导航性能。最后,通过在无动作和无奖励的视频数据(例如 Ego4D)上训练 NWM,我们展示了改进的

看不见的环境中的视频预测和生成性能。

2.相关工作

目标条件视觉导航是
需要感知和规划技能的机器人[8,
13, 15, 41, 43, 51, 55]. 给定上下文图像和
图像指定导航目标,目标条件视觉导航模型[51,55]旨在生成可行的

如果环境已知,则选择通向目标的路径;否则,则探索目标。最近的视觉导航方法包括
NoMaD [55]通过行为克隆训练扩散策略
以及时间距离目标,以便在条件设置中遵循目标,或在非条件设置中探索新环境。先前的方法,如主动神经

SLAM [8]使用神经 SLAM 和分析规划器来规划 3D 环境中的轨迹,而其他

像[9]这样的方法可以通过强化学习来学习策略。在这里,我们展示了世界模型可以使用探索性
数据来规划或改进现有的导航政策。

与学习政策不同,世界的目标
模型[19]是为了模拟环境,例如给定
当前状态和动作来预测下一个状态和相关的奖励。先前的研究表明,联合

学习策略和世界模型可以改进样本

在 Atari [1, 20, 21]、模拟机器人环境[50]甚至应用于现实世界机器人[71] 上都表现出了很高的效率。
最近， [22]提出使用单一世界模型
通过引入动作和任务来跨任务共享
嵌入，而[37, 73]提出用
语言， [6]提出学习潜在动作。世界
模型也在游戏模拟的背景下进行了探索。DIAMOND [1]和 GameNGen [66]
建议使用
扩散模型来学习计算机游戏的游戏引擎
比如雅达利和 Doom。我们的工作灵感来源于这些作品，
我们的目标是学习一个可以在许多环境中共享的通用扩散视频变换器，并且

导航的不同实施例。

在计算机视觉领域，视频生成一直是一个
站立挑战[3, 4, 17, 29, 32, 62, 74]。最近，
文本到视频的合成已经取得了巨大的进步，例如 Sora [5]和 MovieGen [45]
等方法。过去
提出的控制视频合成的工作给出了结构化
动作对象类别[61]或动作图[2]。
视频生成模型以前在强化学习中被用作奖励[10]、预训练方法[59]，用于

模拟和规划操纵动作[11, 35]以及
用于在室内环境中生成路径[26, 31]。有趣的是，扩散模型[28, 54]对视频

生成[69]和预测[36]等任务，而且
视图合成[7, 46, 63]。不同的是，我们使用条件
扩散变压器模拟轨迹进行规划
没有明确的 3D 表示或先验。

3. 导航世界模型

3.1. 配方

接下来，我们来描述我们的 NWM 公式。直观地说，NWM 是一个接收

世界（例如，图像观察）和导航动作（描述移动到哪里以及如何旋转）。该模型
然后产生下一个关于
代理人的观点。

我们获得了一个以自我为中心的视频数据集以及
代理导航动作 $D = \{(x_0, a_0, \ldots, \text{是图像}， a_i = (u, \quad) \text{是} \quad x_T \quad , a_T)\}_{n \quad i=1,}$
使得 $x_i \in \mathbb{R}^{\text{高}\times\text{宽}\times 3}$
导航命令由平移参数 $u \in \mathbb{R}$ 给出
控制前进/后退和右/左的变化
运动，以及控制偏航变化的 $\quad \in \mathbb{R}$
旋转角度。[2]

导航动作 a_i 可以被完全观察到（如
栖息地[49]），例如，朝着墙壁前进会
根据物理原理触发环境响应，
这将导致代理留在原地，而在
其他环境中的导航操作可以近似

[2]这可以自然地扩展到三维，即 $u \in \mathbb{R}3$ 和 $\theta \in \mathbb{R}3$ 定义偏航角、俯仰角和滚转角。为简单起见，我们假设在平坦表面上以固定的俯仰和横滚进行导航。

根据代理位置的变化进行匹配。

我们的目标是学习一个世界模型 F，即从先前的潜在观察 s_τ 和动作 a_τ 到

未来潜在状态表示 s_{t+1}：

$$\text{是} = \text{enc}_\theta(x_i) \qquad s_{\tau+1} \quad F_\theta(s_{\tau+1} \mid s_\tau \quad , \text{在}) \qquad (1)$$

其中 $s_\tau = (s_\tau, \ldots, s_\tau - m)$ 是通过预训练的 VAE [4] 编码的过去 m 个视觉观测值。使用 VAE 有
使用压缩潜伏期的好处是，允许
将预测解码回像素空间以进行可视化。

由于这种表述的简单性，它可以自然地跨环境共享，并轻松扩展到

更复杂的动作空间，例如控制机械臂。
与[20]不同，我们的目标是训练一个单一的世界模型
跨环境和实施例，无需使用任务
或像[22]中那样的动作嵌入。

公式1中的公式模拟了动作，但
不允许控制时间动态。我们扩展
此公式具有时间平移输入 $k \in [T_{min}, T_{max}]$，设置 $a_\tau = (u, \quad, k)$，因此现在 a_τ 指定时间变化
k，用于确定模型应该执行多少步
进入未来（或过去）。因此，给定当前状态
英石，我们可以随机选择一个时间移位 k，并使用相应的时间移位视频帧作为我们的下一个状态 $s_{\tau+1}$。
导航操作可以近似为
从时间 τ 到 $m = \tau + k - 1$ 的总和：

$$u_{\tau \to m} = \sum_{t=\tau}^{*} \text{出} \quad \tau \to m = \sum_{t=\tau}^{*} t \text{模} 2\pi \quad (2)$$

这个公式可以学习导航动作，
还有环境的时间动态。在实践中，
我们允许时间偏移最多 ± 16 秒。

可能出现的一个挑战是动作和时间的纠缠。例如，如果到达特定位置总是发生在特定时间，模型可能会学习

只依赖时间而忽略后续行动，
反之亦然。实际上，数据可能包含自然
反事实 比如在不同时间到达同一区域。为了鼓励这些自然的反事实，我们

在训练过程中，为每个状态采样多个目标。我们
在第 4 节中进一步探讨这种方法。

3.2. 扩散变压器作为世界模型

如上一节所述，我们将 F_θ 设计为
随机映射，以便模拟随机环境。这是通过条件扩散实现的

变压器（CDiT）模型，如下所述。
条件扩散变压器架构。
我们使用的架构是一个时间自回归变换器模型，利用高效的 CDiT 块（见图 2），该模型在输入序列上应用 $\times N$ 次

具有输入动作条件的潜在因素。

CDiT 通过以下方式实现时间高效的自回归建模
仅将注意力限制在第一个注意力块中
来自正在去噪的目标帧的标记。
为了对过去帧中的标记进行条件处理,我们结合了
交叉注意力层,这样来自
当前目标关注的是过去帧中的标记,这些标记是
用作键和值。然后,交叉注意力机制使用跳过连接层将这些表示情境化。

以导航动作 a ∈ R 为条件 3 , 我们首先
将每个标量映射到 R d 3.通过提取正弦余弦特征,
然后应用 2 层 MLP,并将它们连接成
单个向量ψa ∈ R d 。我们遵循类似的流程
将时间移位 k ∈ R 映射到ψk ∈ R d 和扩散
时间步长 t ∈ R 到ψk ∈ R d 。最后,我们将所有嵌入相加
转化为用于调节的单个向量:

$$\xi = \psi a + \psi k + \psi t \tag{3}$$

然后将 ξ 馈送到 AdaLN [72]模块以生成尺度
和调节层归一化[34]输出的移位系数,以及注意层的输出。为了在未标记数据上
进行训练,我们简单地省略了显式

计算 ξ 时的导航动作（见公式3）。
另一种方法是直接使用 DiT [44],然而,对完整输入应用 DiT 计算成本很
高。n 表示每个输入的输入 token 数量

帧,m 表示帧数,d 表示 token 维度。缩放多头注意力层[68] 的复杂度主要由
注意力项 O(m2n2d) 决定,它

与上下文长度呈二次函数关系。相比之下,我们的 CDiT
块由交叉注意层复杂性决定
O(mn2d),它与上下文呈线性关系,允许我们使用更长的上下文大小。我们分析
这两个
第 4节中的设计选择。CDiT 类似于原始的
Transformer Block [68],无需对上下文标记应用昂贵的自注意力。

扩散训练。在前向过程中,加入噪声
根据随机选择的
时间步长 t ∈ {1, . . . , T}。噪声状态 s (吨) 可以解τ+1
罚款金额: (吨) t+1 $= \sqrt{\alpha t} s\tau+1+\sqrt{1-\alpha t}$, 其中 N (0, I)
是高斯噪声， {αt}是噪声计划控制- （t）
方差。随着 t 的增加,s τ+1收敛于纯
噪声。逆过程试图恢复原始噪声
最终状态表示sτ+1来自噪声版本 s t+1,
以上下文sτ 、当前动作aτ和扩散时间步长 t 为条件。我们定义Fθ(sτ+1|和
sτ) ,即以 θ 为参数的噪声神经网络模型。我们遵循与 DiT 在 ,t)作为de-
[44] 相同的噪声方案和超参数。

训练目标。模型的训练目标是最小化
清洁目标和预测目标之间的均方,旨在学习去噪过程:

Lsimple = Esτ+1,aτ ,sτ , ,t ∥ sτ+1 − Fθ(s (吨) τ+1|sτ , 在 ,t)∥ 2 2.

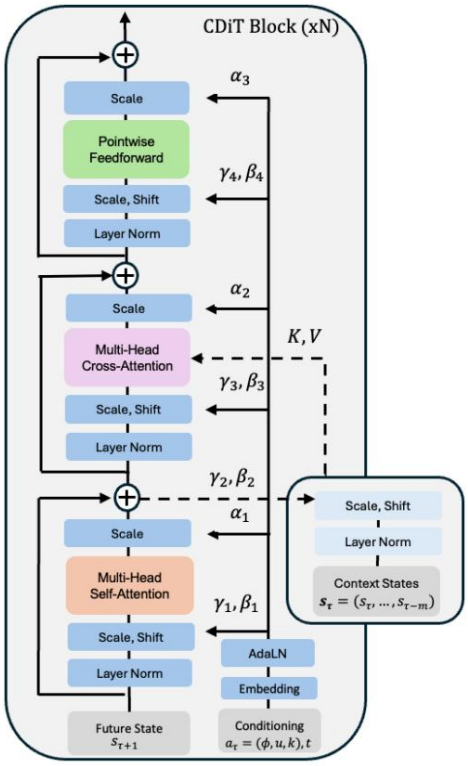

图 2.条件扩散变压器 (CDiT) 块。
该块的复杂度与帧数呈线性关系。

在这个目标中,时间步长 t 被随机采样为
确保模型能够学习对不同程度的损坏帧进行去噪。通过最小化这种损失,模型

学习从噪声版本 s 中重建sτ+1 (吨) t+1, 和-
基于上下文sτ和动作aτ ,从而能够
生成现实的未来框架。根据[44],我们
还预测噪声的协方差矩阵并监督
它与变分下界损失Lvlb有关[42]。

3.3. 使用世界模型进行导航规划

现在我们来描述如何使用经过训练的 NWM 来
规划导航轨迹。直观地说,如果我们的世界模型
熟悉一个环境,我们可以用它来模拟
导航轨迹,并选择到达
目标。在未知、分布不均的环境中,
学期规划可能依赖于想象力。
形式上,给定潜在编码s0和导航
目标 ，我们寻找一系列动作(a0, ..., aT −1)
最大化达到 s 的可能性 。令 S(sT , s)
表示达到状态 s 的非标准化分数
给定初始条件s0， sT的动作 a =
(a0, . . . , aT −1),并且通过自回归滚动 NWM 获得状态 s = (s1, . . . sT):s
Fθ(· |s0, a)。
我们定义能量函数 E(s0, a0, . . ., aT −1, sT),
这样最小化能量就相当于最大化
非标准化感知相似性得分及后续

对状态和行动的潜在限制：

$$E(s_0, a_0, ..., a_{T-1}, s_T) = -S(s_T, s_{T-1}) + \sum_{\tau=0} I(a_\tau \in A/有效) + \sum_{\tau=0} I(s_\tau \in S / safe) \quad (4)$$

相似度计算方法是：使用预训练的 VAE 解码器[4]将 s 和sT解码为像素，然后测量感知相似度[14, 75]。诸如“永不左转后右转”之类的约束可以通过将aτ限制在有效动作集合 Avalid 中来编码；而“永不探索悬崖边缘”则可以通过确保状态sτ位于Ssafe 中来编码。I (·) 表示指示函数，如果违反任何动作或状态约束，则会施加较大的惩罚。

那么问题就简化为寻找最小化的动作
模拟该能量函数：

$$愤怒的矿井_{a_0,...,a_{T-1}} \quad E_s [E(s_0, a_0, ..., a_{T-1}, s_T)] \quad (5)$$

该目标可以重新表述为模型预测控制 (MPC) 问题，我们使用交叉熵方法[48] 对其进行优化。交叉熵方法是一种简单的无导数、基于种群的优化方法，最近与世界模型一起用于规划[77]。附录 7 中提供了交叉熵方法的概述和完整的优化技术细节。

导航轨迹排序。假设我们有一个现有的导航策略 Π(a|s0, s),我们可以使用 NWM 对采样轨迹进行排序。这里我们使用 NoMaD [55] ，这是一种最先进的机器人导航策略。

为了对轨迹进行排序，我们从 Π 中抽取多个样本，并选择能量最低的样本，如公式 5 所示。

4.实验与结果

我们描述了实验设置和设计选择，并将 NWM 与之前的方法进行了比较。更多结果包含在补充材料中。

4.1. 实验设置

数据集。对于所有机器人数据集（SCAND [30]、 Tartan-Drive [60]、 RECON [52]和 HuRoN [27]），我们都可以访问机器人的位置和旋转，从而可以推断出相对于当前位置的相对动作（参见公式2）。

为了标准化不同智能体的步长，我们将智能体在帧间移动的距离除以其平均步长（以米为单位），以确保不同智能体的动作空间相似。我们进一步滤除后向移动，效仿NoMaD [55]。此外，我们使用未标记的Ego4D [18]视频，其中我们唯一考虑的动作是时间偏移。SCAND提供在不同环境下符合社交规范的导航视频片段，TartanDrive专注于越野驾驶，RECON涵盖开放世界导航，HuRoN捕捉社交互动。我们训练

未标记的 Ego4D 视频和 GO Stanford [24]作为未知的评估环境。完整详情请参阅附录8.1。

评估指标。我们使用绝对轨迹误差 (ATE) 评估预测的导航轨迹的准确性，并使用相对姿态误差 (RPE) 评估姿态一致性[57]。为了检查世界模型预测与地面真实图像在语义上的相似程度，我们应用了 LPIPS [76]和 DreamSim [14] ，通过比较深度特征来测量感知相似度，并使用 PSNR 来评估像素级质量。对于图像和视频合成质量，我们使用 FID [23]和 FVD [64]来评估生成的数据分布。更多详情，请参阅附录 8.1 。

基线。我们考虑以下所有基线。

- DIAMOND [1]是一个基于 UNet [47]架构的扩散世界模型。我们根据其公开代码，在离线强化学习环境中使用 DIAMOND。该扩散模型经过训练，可在 56x56 分辨率下进行自回归预测，并使用上采样器进行 224x224 分辨率预测。为了对连续动作进行条件化，我们使用了线性嵌入层。· GNM [53]是一个通用的目标条件化导航策略，基于机器人导航数据集进行训练，并配备全连接轨迹预测网络。

  GNM 在多个数据集上进行训练，包括 SCAND、TartanDrive、GO Stanford 和 RECON。

- NoMaD [55]使用扩散策略扩展了 GNM，用于预测机器人探索和视觉导航的轨迹。NoMaD 使用与 GNM 和 HuRoN 相同的数据集进行训练。

实现细节。在默认的实验设置中，我们使用具有 1B 个参数的 CDiT-XL，上下文为 4 帧，总批次大小为 1024，并包含 4 个不同的导航目标，最终总批次大小为 4096。我们使用稳定扩散[4] VAE 标记器，类似于 DiT [44]。

我们使用 AdamW [39]优化器，学习率为 8e − 5。训练结束后，我们从每个模型中采样 5 次，以报告平均值和标准差结果。XL 尺寸的模型在 8 台 H100 机器上训练，每台机器配备 8 个 GPU。除非另有说明，我们使用与 DiT-*/2 模型相同的设置。

4.2. 消融

模型评估基于已知环境 RECON 上的验证集轨迹，进行单步 4 秒未来预测。我们通过测量 LPIPS、DreamSim 和 PSNR 来评估模型相对于真实帧的性能。图3 提供了定性示例。

模型大小与 CDiT。我们将 CDiT（参见3.2 节)与标准 DiT（其中所有上下文 token 都作为输入)进行比较。我们假设，对于导航已知环境，模型的容量是最重要的，图5 中的结果表明 CDiT 确实表现良好。

图 3.在已知环境中跟踪轨迹。我们包含不同模型的定性视频生成比较
遵循地面真实轨迹。点击图片即可在浏览器中播放视频片段。

消融 1 2 #	点数 ↓	梦幻模拟战 ↓	峰值信噪比↑
目标 4	0.312 ± 0.001	0.098 ± 0.001	15.044 ± 0.031
	0.305±0.000	0.096±0.001	15.154±0.017
	0.296 ±0.002	0.091 ±0.001	15.331 ±0.027
1	0.304±0.001	0.097±0.001	15.223±0.033
2 #上下文 4	0.302±0.001	0.095±0.000	15.274±0.027
	0.296 ±0.002	0.091 ±0.001	15.331 ±0.027
仅时间作用	0.760±0.001	0.783±0.000	7.839±0.017
仅动作+时间	0.318±0.002	0.100±0.000	14.858±0.055
0.295 ±0.002	0.091 ±0.001	15.343 ±0.060	

表 1.每个样本数、上下文的预测目标消融
大小,以及动作和时间条件的使用。我们在 RECON 上报告未来 4 秒的预测结果。

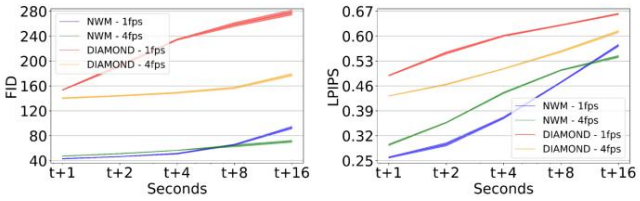

图 4. NWM生成精度和质量比较
并以 1 和 4 FPS 的 DIAMOND 作为时间函数,在 RECON 数据集上生成长达 16 秒的视频。

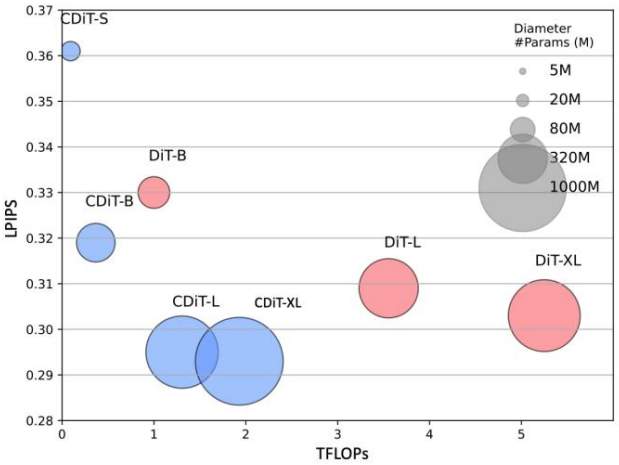

图 5. CDiT 与 DiT。衡量模型预测 4 的准确性
在 RECON 上预测未来几秒。我们将 LPIPS 报告为一个函数
Tera FLOPs,越低越好。

模型钻石 NWM（我们的）	
FVD ↓ 762.734 ± 3.361	200.969 ±5.629

图6.视频合成质量比较。16秒
在 RECON 上以 4 FPS 生成的视频。

对于包含多达 1B 个参数的模型来说,效果会更好,同时消耗
少于 2 倍 FLOP。令人惊讶的是,即使使用相同数量的
参数（例如,CDiT-L 与 DiT-XL 相比）,CDiT 是
速度提高 4 倍,性能更佳。

目标数量。我们训练模型时,目标数量是可变的
在给定固定上下文的情况下,改变目标状态的数量
目标从 1 到 4。每个目标都是随机选择的
当前状态周围±16秒的窗口。结果

表1中报告的结果表明,使用 4 个目标可以显著提高所有指标的预测性能。

上下文大小。我们训练模型时会改变
条件框架从1到4（见表1）。不出所料,更多的上下文有帮助,而短上下文模型

经常“迷失方向”,导致预测错误。

时间和动作调节。我们用以下方法训练模型：
时间和动作调节,并测试每个

输入有助于预测性能（我们包括
结果见表1。我们发现,使用
时间只会导致表现不佳,而不是调节
准时也会导致绩效略有下降。这
确认两种输入对模型都有益。

4.3. 视频预测与合成

我们评估模型对真实动作的遵循程度以及对未来状态的预测能力。该模型是
经过条件
在第一个图像和上下文帧上,然后自回归
使用真实动作预测下一个状态,输入
支持每个预测。我们将预测与地面进行比较
1、2、4、8 和 16 秒的真实图像,报告 FID
以及 RECON 数据集上的 LPIPS。图4显示了与 DIAMOND 在 4 FPS 和 1

FPS,表明 NWM 预测明显更
比 DIAMOND 更精确。最初,NWM 1 FPS 变量

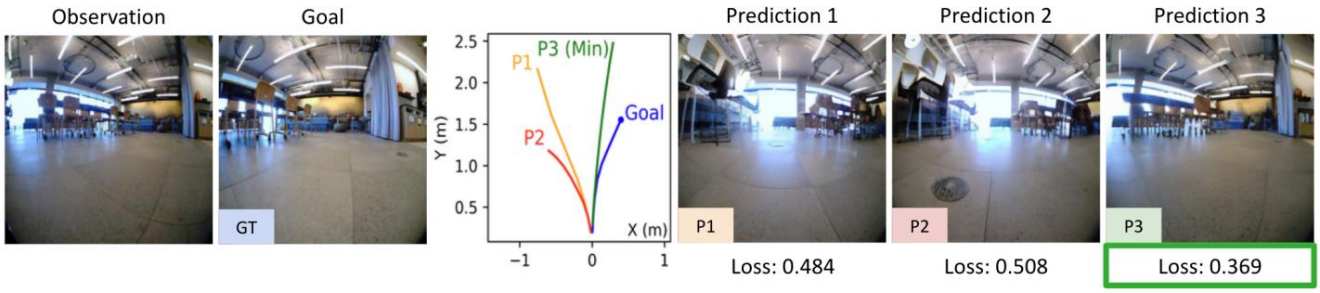

图 7.使用 NWM 对外部策略的轨迹进行排序。为了从观察图像导航到目标,我们采样
从 NoMaD [55] 中提取轨迹,使用 NWM 模拟每条轨迹,对其进行评分（参见公式4)并进行排序。使用 NWM
我们可以准确地选择更接近真实轨迹的轨迹。点击图片可在浏览器中播放示例。

模型	ATE ↓ RPE ↓
GNM 1.87±0.00 0.73±0.00 平均差值 1.93±0.04 0.52±0.00	
NWM + NoMaD (×16) 1.83 ± 0.03 0.50 ± 0.01	
NWM + NoMaD (×32) 1.78 ± 0.03 0.48 ± 0.01	
NWM（规划） 1.13 ±0.02 0.35 ±0.01	

表 2.目标条件视觉导航。ATE和 RPE
RECON 的结果,预测了 2 秒的轨迹。NWM
与以前相比,所有指标均取得了更好的结果
接近NoMaD [55]和GNM [53]。

模型	相对δu ↓ 相对δ ↓
先前进	+0.36±0.01+0.61±0.02
左到右优先 −0.03 ± 0.01 +0.20 ± 0.01	
直线然后向前 +0.08 ± 0.01 +0.22 ± 0.01	

表 3.带导航约束的规划。我们提出
在三个行动约束下使用 NWM 进行规划的结果，
报告最终位置（δu)和偏航（δ)的差异
相对于无约束基线。所有约束均满足，
证明 NWM 可以有效地遵守这些规定。

蚂蚁表现更好,但 8 秒后预测效果变差
由于累积错误和上下文丢失以及 4 FPS
变得更加优越。参见图3中的定性示例。

生成质量。为了评估视频质量,我们以 4 FPS 的速度对视频进行 16
秒的自回归预测,以生成视频,同时以真实动作为条件。我们

然后评估使用 FVD 生成的视频的质量，
与 DIAMOND [1] 相比。图6中的结果表明,NWM 输出的视频质量更
高。

4.4 使用导航世界模型进行规划

接下来,我们来描述测量
我们可以使用 NWM 进行导航吗?我们包含完整的
实验的技术细节见附录8.2。

独立规划。我们证明 NWM 可以
有效地独立用于目标条件导航。我们根据过去的观察结果和目标图
像进行条件调整，
并使用交叉熵方法找到一条轨迹
最小化最后预测图像的 LPIPS 相似度
到目标图像（见公式5）。为了对动作序列进行排序,我们执行 NWM
并测量
最后状态和目标 3 次以获得平均分数。
我们生成长度为 8 的轨迹,时间偏移为
k = 0.25。我们在表2中评估了模型性能。
我们发现使用 NWM 进行规划可以带来竞争力
采用最先进的政策取得成果。

约束规划。世界模型允许规划
在限制条件下 例如要求直线运动
或单次转弯。我们展示了 NWM 支持约束感知规划。在前向优先中，
智能体向前移动
走 5 步,然后转 3 步。先左后右,然后转
向前走之前要走三步。先直走,再向前走。
它直线移动 3 步,然后向前。约束是
通过将特定操作归零来强制执行;例如,在左右
首先,前 3 步的前进运动为零,然后由独立规划优化剩余步骤。我们
报告
与无约束规划相比,最终位置和偏航角的差异。结果（表3)表明,
NWM规划在约束条件下有效,但性能略有下降

滴（见图9 中的示例）。

使用导航世界模型进行排名。NWM
可以增强目标条件导航中现有的导航策略。基于过去的观测结果和
目标图像,我们以 NoMaD 为条件,采样 n ∈ {16, 32} 条轨迹,每条
轨迹的长度为 8,并使用 NWM 自回归跟踪动作来评估它们。最后,
我们对

通过测量LPIPS与目标图像的相似度来计算每条轨迹的最终预测（见
图7）。我们报告ATE
和 RPE 在所有域内数据集上（表2)并发现
基于 NWM 的轨迹排序提高了导航性能,更多的样本可以产生更好
的结果。

4.5. 推广至未知环境

在这里,我们尝试添加未标记的数据,并询问
NWM 是否可以在新环境中做出预测
运用想象力。在这个实验中,我们训练一个模型
在所有域内数据集上,以及未标记的子集上

图 8.导航未知环境。NWM以单幅图像为条件，自回归地预测下一个状态
给出相关操作（标记为黄色）。点击图像即可在浏览器中播放视频片段。

数据	未知环境（Go Stanford） lpips ↓ dreamsim ↓ psnr ↑	已知环境（侦察） lpips ↓ dreamsim ↓ psnr ↑
域内数据	0.658 ± 0.002 0.478 ± 0.001 11.031 ± 0.036	0.295 ±0.002 0.091 ±0.001 15.343 ±0.060
+ Ego4D（未标记）	0.652 ±0.003 0.464 ±0.003 11.083 ±0.064	0.368 ± 0.003 0.138 ± 0.002 14.072 ± 0.075

表 4.使用额外的未标记数据进行训练可提高在未知环境中的表现。报告未知环境中的结果
环境（Go Stanford)和已知环境（RECON）。通过评估未来 4 秒的情况来报告结果。

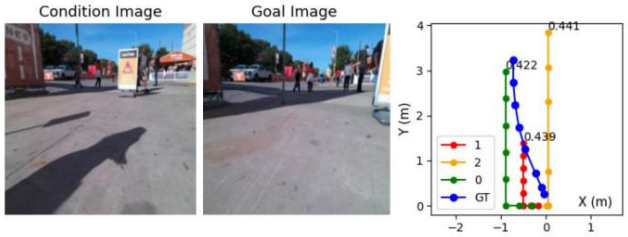

图 9.使用 NWM 进行约束规划。我们将
在左移约束下使用 NWM 规划的轨迹
或者先右后左，然后向前运动。规划目标
是为了达到与地面相同的最终位置和方向
真实（GT)轨迹。图中显示的是拟议轨迹的成本
0、1 和 2，其中轨迹 0（绿色)实现成本最低。

来自 Ego4D 的视频，我们只能访问时移动作。我们训练了一个 CDiT-XL 模型，并在

Go Stanford 数据集以及其他随机图像。我们在表4 中报告了
结果，发现在未标记图像上进行训练
数据可以显著提高视频预测的准确性
所有指标，包括提高发电质量。我们
包括图 8 中的定性示例。与域内（图3)相比，该模型崩溃得更快，并
且如预期的那样
在产生对想象环境的遍历时，会产生幻觉路径。

5. 限制

我们发现了许多限制。首先，当应用于
分布数据，模型往往会慢慢失去背景
并生成与训练数据相似的下一个状态，
在图像生成中观察到的现象，以及
被称为模式崩溃[56, 58]。我们包括这样一个
图10中的示例。其次，虽然模型可以规划，
它很难模拟行人运动等时间动态（尽管在某些情况下确实如此）。
这两个限制很可能通过更长的上下文和更多

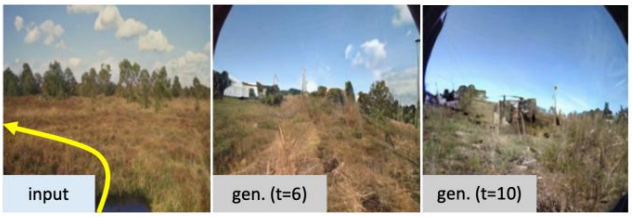

图 10.局限性和失败案例。在未知环境中，一种常见的失败案例是模式
崩溃，即模型
输出逐渐与训练数据更加相似。点击
在图像上播放视频片段。

训练数据。此外，该模型目前使用 3
DoF 导航操作，但扩展到 6 DoF 导航
甚至可能更多（比如控制机器人的关节
arm)也是可能的，我们将此留待以后的工作。

6.讨论

我们提出的导航世界模型（NWM)提供了
可扩展的、数据驱动的学习世界模型的方法
视觉导航；然而，我们还不完全确定
什么样的表征能够实现这一点，因为我们的 NWM 并没有明确地利
用环境的结构化地图。
这个想法是从自我中心的角度预测下一帧
视角可以驱动非中心表征的出现[65]。最终，我们的方法将学习

视频、视觉导航和基于模型的规划和
可能为自我监督系统打开大门
不仅可以感知，还可以制定计划来指导行动。

致谢。我们感谢Noriaki Hirose的
帮助 HuRoN 数据集并分享他的见解，
致 Manan Tomar、David Fan、Sonia Joseph、Angjoo
Kanazawa、Ethan Weber、Nicolas Ballas 和匿名评论者的有益
讨论和反馈。

参考

[1] 埃洛伊·阿隆索、亚当·杰利、文森特·米切利、安西·卡纳维斯托、阿莫斯·斯托基、蒂姆·皮尔斯和弗朗索瓦·弗勒雷。
世界建模的扩散：视觉细节在雅达利中很重要。
在第三十八届神经信息处理系统会议上。2、3、5、7

[2] 阿米尔·巴尔、罗伊·赫齐格、王晓龙、安娜·罗尔巴赫、
Gal Chechik、Trevor Darrell 和 Amir Globerson。基于动作图的合成视频合成。国际
机器学习会议，第 662-673 页。PMLR，
2021. 3

[3] 奥马尔·巴尔-塔尔、希拉·谢弗、奥马尔·托夫、查尔斯·赫尔曼、罗尼·佩斯、施兰·扎达、阿里尔·埃弗拉特、俊花
Hur、Guanghui Liu、Amit Raj 等人。Lumiere：用于视频生成的时空扩散模型。arXiv 预印本
arXiv:2401.12945, 2024. 3

[4] 安德烈亚斯·布拉特曼、蒂姆·多克霍恩、苏米斯·库拉尔、丹尼尔
门捷列维奇、马切伊·基利安、多米尼克·洛伦茨、亚姆·莱维、
Zion English、Vikram Voleti、Adam Letts 等。视频稳定
扩散：将潜在视频扩散模型扩展到大
数据集。arXiv 预印本 arXiv:2311.15127, 2023. 3, 5

[5] 蒂姆·布鲁克斯、比尔·皮布尔斯、康纳·霍姆斯、威尔·德普
郭宇飞、李静、大卫·施努尔、乔·泰勒、特洛伊·卢曼、埃里克·卢曼等。视频生成模型作为世界
模拟器， 2024.3

[6] 杰克·布鲁斯、迈克尔·D·丹尼斯、阿什利·爱德华兹、杰克
帕克-霍尔德、史玉格、爱德华·休斯、马修·赖、
Aditi Mavalankar、Richie Steigerwald、Chris Apps 等。Ge-nie：生成式交互环境。
2024 年第四十一届国际机器学习大会。3

[7] Eric R. Chan、Koki Nagano、Matthew A. Chan、Alexander W.
伯格曼、朴正俊、阿克塞尔·利维、米卡·艾塔拉、沙利尼
德梅洛、泰罗·卡拉斯和戈登·韦茨斯坦。生成式
利用 3D 感知扩散模型进行新颖的视图合成。在
IEEE/CVF 国际会议论文集
计算机视觉（ICCV），第 4217-4229 页，2023 年。3

[8] 德文德拉·辛格·查普特、迪拉杰·甘地、索拉布·古普塔、
Abhinav Gupta 和 Ruslan Salakhutdinov。《学习利用主动神经冲击进行探索》。国际会议
学习表征。2

[9] Tao Chen、Saurabh Gupta 和 Abhinav Gupta。《学习》
导航探索策略。国际学习表征会议。2

[10] 亚历杭德罗·埃斯孔特雷拉、阿德米·阿德尼吉、威尔逊·严、阿贾伊
Jain、Xue Bin Peng、Ken Goldberg、Youngwoon Lee、Dani-jar Hafner 和
Pieter Abbeel。视频预测模型如下：
强化学习的奖励。神经信息处理系统进展， 36，2024。3

[11] Chelsea Finn 和 Sergey Levine。深度视觉预见
规划机器人运动。2017 年 IEEE 国际机器人与自动化会议（ICRA），第 2786-2793 页。
IEEE， 2017.3

[12] Elias Frantar、Saleh Ashkboos、Torsten Hoefler 和 Dan
阿利斯塔尔。 Gptq：准确的训练后量化
用于生成预训练的 Transformer。arXiv 预印本
arXiv:2210.17323, 2022. 3

[13] J Frey、M Mattamala、N Chebrolu、C Cadena、M Fallon 和
M Hutter。野外视觉导航的快速可穿越性估计。《机器人：科学与系统学报》，19，
2023. 2, 3

[14] Stephanie Fu、Netanel Tamir、Shobhita Sundaram、Lucy
Chai、Richard Zhang、Tali Dekel 和 Phillip Isola。《梦境模拟：学习人类视觉相似性的新维度》
使用合成数据。神经信息处理系统进展，36，2024. 5, 1

[15] Zipeng Fu, Ashish Kumar, Ananye Agarwal, Haozhi Qi, Ji-tendra Malik 和
Deepak Pathak. 视觉与前躯体感觉的耦合在腿式机器人导航中的应用. 论文集
IEEE/CVF 计算机视觉与模式识别会议，第 17273-17283 页，2022 年。2

[16] 高俊宇，姚宣，徐长生.在线视觉语言导航的快慢测试时间自适应。
在第41届国际机器学习会议论文集，第14902-14919页。PMLR，2024年。3

[17] Rohit Girdhar、Mannat Singh、Andrew Brown、Quentin Duval、Samaneh Azadi、Sai Saketh Rambhatla、Akbar Shah、Xi
Yin, Devi Parikh 和 Ishan Misra。Emu 视频：通过显式图像条件分解文本到视频的生成。
arXiv 预印本 arXiv:2311.10709, 2023. 3

[18] 克里斯汀·格劳曼、安德鲁·韦斯特伯里、尤金·伯恩
扎克瑞·查维斯、安东尼诺·福纳里、罗希特·吉达尔、杰克逊
Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d:
3000小时的自我中心视频带你环游世界。摘自IEEE/CVF计算机视觉会议论文集
和模式识别，第 18995-19012 页，2022年。5，2

[19] David Ha 和 Jurgen Schmidhuber。世界模型。预印本 arXiv
arXiv:1803.10122，2018 年。2

[20] Danijar Hafner、Timothy Lillicrap、Jimmy Ba 和 Mohammad Norouzi。《梦境控制：潜在想象下的学习行为》。国际学习大会
陈述，. 3

[21] 丹尼贾·哈夫纳 (Danijar Hafner)、蒂莫西·P·利利克拉普 (Timothy P Lillicrap)、穆罕默德·诺鲁齐 (Mohammad Norouzi)，
和 Jimmy Ba。通过离散世界模型掌握 Atari。
在国际学习表征会议上，。
3

[22] Nicklas Hansen, Hao Su, and Xiaolong Wang. Td-mpc2:
可扩展且强大的世界模型，用于连续控制。在
第十二届国际学习表征会议。3

[23] 马丁·赫塞尔、休伯特·拉姆绍尔、托马斯·翁特蒂纳，
Bernhard Nessler 和 Sepp Hochreiter。Gans 受训于
两个时间尺度更新规则收敛到局部纳什均衡。神经信息处理系统的进展，
30，2017.5，1

[24] Noriaki Hirose、Amir Sadeghian、Marynel Vazquez、Patrick Goebel 和 Silvio Savarese。戈内特：半监督
深度学习方法用于可通行性评估。2018 年
IEEE/RSJ国际智能机器人会议
与系统（IROS），第 3044-3051 页。IEEE，2018 年。2 , 5

[25] Noriaki Hirose、Amir Sadeghian、Fei Xia、Roberto Mart´ n-Mart´ n
和 Silvio Savarese。Vunet：动态场景视图
使用 RGB 相机进行可穿越性估计的综合。
IEEE 机器人与自动化快报，2019 年2

[26] 广濑纪明、夏飞、Mart´ n Roberto-Mart´ n、Amir
Sadeghian 和 Silvio Savarese。深度视觉 mpc 策略
导航学习。IEEE 机器人与自动化快报，4(4):3184–3191，2019。3

[27] 广濑典明、Dhruv Shah、Ajay Sridhar 和 Sergey
Levine。Sacson：可扩展的社交导航自主控制。IEEE机器人与自动化快报，2023年，
5，1，
2

[28] Jonathan Ho, Ajay Jain, 和 Pieter Abbeel. 去噪扩散概率模型。神经信息学进展

处理系统，33：6840–6851，2020 年。3

[29] Jonathan Ho、William Chan、Chitwan Saharia、Jay Whang、
高瑞琪、Alexey Gritsenko、Diederik P Kingma、Ben
普尔、穆罕默德·诺鲁兹、大卫·J·弗利特等人。图像
视频：使用扩散模型生成高清视频。arXiv 预印本 arXiv:2210.02303，2022. 3

[30] Haresh Karnan、Anirudh Nair、Xuesu Xiao、Garrett War-nell、Soren Pirk、
Alexander Toshev、Justin Hart、Joydeep Biswas 和 Peter Stone。符合社会规范的导航
数据集 (scand)：大规模演示数据集
社交导航。IEEE 机器人与自动化快报，7
(4):11807–11814, 2022.5 , 1

[31] Jing Yu Koh, Honglak Lee, Yinfei Yang, Jason Baldridge,
以及 Peter Anderson。《Pathdreamer：室内装饰的世界典范》
导航。在 IEEE/CVF 国际会议论文集
计算机视觉会议，第 14738–14748 页，2021 年。
3

[32] Dan Kondratyuk, Lijun Yu, Xiuye Gu, Jose Lezama,
Jonathan Huang、Grant Schindler、Rachel Hornung、Vigh-nesh Birodkar、
Jimmy Yan、Ming-Chang Chiu 等。
Videopoet：用于零样本视频生成的大型语言模型。第 41 届国际机器学习大会

学习。3

[33] Alex Krizhevsky、Ilya Sutskever 和 Geoffrey E Hinton。
基于深度卷积神经网络的图像分类。神经信息处理系统的进展，

25，2012.1

[34] 吉米·雷·巴、杰米·瑞安·基罗斯、杰弗里·E·欣顿。
层归一化。ArXiv 电子出版物，arXiv-1607 页，
2016年4月

[35] 梁俊邦、刘若诗、Ege Ozguroglu、Sruthi Sud-hakar、Achal Dave、Pavel
Tokmakov、Shuran Song 和 Carl
Vondrick。Dreamitate：通过视频生成进行真实世界视觉运动策略学习，2024
年。3

[36] 韩林、Tushar Nagarajan、Nicolas Ballas、Mido Assran、
Mojtaba Komeili、Mohit Bansal 和 Koustuv Sinha。视频编辑：
用于程序性视频表示学习的潜在预测架构， 2024.3

[37] 林洁西、杜雨晴、奥利维亚·沃特金斯、丹尼尔·哈夫纳、彼得
Abbeel、Dan Klein 和 Anca Dragan。学习建模
有语言的世界， 2024.3

[38] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tok-makov, Sergey
Zakharov 和 Carl Vondrick. Zero-1-to-3：零样本单图像到三维物体。载于

IEEE/CVF 国际计算机视觉会议，
第 9298–9309 页，2023 年。2

[39] I Loshchilov. 解耦权重衰减正则化。arXiv
预印本 arXiv:1711.05101, 2017. 5

[40] 本·米尔登霍尔、普拉图·P·斯里尼瓦桑、马修·坦西克，
乔纳森·T·巴伦 (Jonathan T Barron)、拉维·拉马莫蒂 (Ravi Ramamoorthi) 和吴仁 (Ren Ng)。削弱：
将场景表示为用于视图合成的神经辐射场。ACM通讯，65(1):99–106，2021年。

2

[41] 彼得·米罗夫斯基、拉兹万·帕斯卡努、法比奥·维奥拉、休伯特·索耶、
安迪·巴拉德、安德里亚·巴尼诺、米沙·德尼尔、罗斯·戈罗辛
Laurent Sifre、Koray Kavukcuoglu 等人。《学习在复杂环境中导航》。国际会
议
学习表征，2022 年2

[42] Alexander Quinn Nichol 和 Prafulla Dhariwal。改进
去噪扩散概率模型。在会议论文集
第 38 届国际机器学习大会，
第 8162–8171 页。PMLR，2021 年。4

[43] Deepak Pathak、Parsa Mahmoudih、Guanghao Luo、Pulkit
Agrawal、Dian Chen、Yide Shentu、Evan Shelhamer、Jiten-dra Malik、
Alexei A Efros 和 Trevor Darrell。Zero-shot
视觉模仿。在 IEEE 会议论文集上
计算机视觉和模式识别研讨会，页面
2050–2053 年，2018 年. 2

[44] William Peebles 和 Saining Xie。基于 Transformer 的可扩展扩散模型。IEEE/
CVF 国际计算机视觉会议 (ICCV) 论文集，第 35-36 页

4195–4205，2023.2，4，5

[45] 亚当·波利亚克、阿米特·佐哈尔、安德鲁·布朗、安德罗斯·詹德拉、
Animesh Sinha、Ann Lee、Apoorv Vyas、Bowen Shi、Chih-Yao Ma、Ching-
Yao Chuang 等。电影一代：演员阵容
媒体基础模型。arXiv 预印本 arXiv:2410.13720，
2024. 3

[46] Ben Poole、Ajay Jain、Jonathan T Barron 和 Ben Mildenhall。Dreamfusion：
利用 2D 扩散将文本转为 3D。在
第十一届国际学习表征会议。3

[47] Olaf Ronneberger、Philipp Fischer 和 Thomas Brox。U-net：用于生物医学图像
分割的卷积网络。在医学图像计算和计算机辅助

干预 MICCAI 2015：第 18 届国际会议，
德国慕尼黑，2015年10月5日至9日，会议记录，第三部分
18，第234-241页。Springer，2015年。5

[48] Reuven Y Rubinstein.计算机模拟的优化
具有罕见事件的模型。欧洲作战
研究，99(1):89–112, 1997.5 , 1

[49] 马诺利斯·萨瓦、阿布舍克·卡迪安、亚历山大·马克西梅茨、
赵伊利、Erik Wijmans、Bhavana Jain、Julian Straub、Jia
刘、Vladlen Koltun、Jitendra Malik 等。栖息地：A
具身人工智能研究平台。在
IEEE/CVF 国际计算机视觉会议，
第 9339–9347 页，2019 年。3

[50] Younggyo Seo, Danijar Hafner, Hao Liu, Fangchen Liu,
Stephen James、Kimin Lee 和 Pieter Abbeel。《蒙面》
视觉控制的世界模型。在机器人会议上
学习，第 1332-1344 页。PMLR，2023 年。3

[51] Dhruv Shah、Ajay Sridhar、Nitish Dashora、Kyle Stachow-icz、Kevin
Black、Noriaki Hirose 和 Sergey Levine。Vint：
视觉导航的基础模型。第七届年度
机器人学习会议。2

[52] 德鲁夫·沙阿、本杰明·艾森巴赫、格雷戈里·卡恩、尼古拉斯
Rhinehart 和 Sergey Levine。快速探索开放

具有潜在目标模型的世界导航。arXiv 预印本
arXiv:2104.05859, 2021. 5, 1, 2

[53] Dhruv Shah、Ajay Sridhar、Arjun Bhorkar、Noriaki Hirose、
以及 Sergey Levine。Gnm：一种通用导航模型
驱动任何机器人。2023 年 IEEE 国际会议上
机器人与自动化（ICRA），第 7226-7233 页。IEEE，
2023. 2, 5, 7, 3

[54] 贾沙·索尔-迪克斯坦、埃里克·韦斯、尼鲁·马赫斯瓦拉纳坦、
以及 Surya Ganguli。深度无监督学习，使用
非平衡热力学。国际机器学习会议，第2256-2265页。PMLR，2015年。

3

[55] Ajay Sridhar、Dhruv Shah、Catherine Glossop 和 Sergey
Levine。Nomad：用于导航和探索的目标掩蔽扩散策略。2024 年 IEEE 国
际机器人与自动化会议 (ICRA)，第 63-70 页。

IEEE，2024. 2, 5, 7, 1, 3

[56] 阿卡什·斯里瓦斯塔瓦、拉扎尔·瓦尔科夫、克里斯·拉塞尔、迈克尔·U
Gutmann 和 Charles Sutton。Veegan：利用隐式变分学习减少 GAN 中
的模式崩溃。进展
神经信息处理系统， 30，2017.8

[57] Jurgen Sturm、Wolfram Burgard 和 Daniel Cremers。使用 tum rgb-d
基准评估自我运动和运动结构方法。在“运动结构与运动”研讨会论文集
上。
IEEE/RJS 国际智能机器人系统会议 (IROS) 上的机器人彩色深度相机融
合，
第6页，2012年， 5,1

[58] Hoang Thanh-Tung 和 Truyen Tran. GAN 中的灾难性遗忘和模式崩溃。
2020 年国际联合
神经网络会议（ijcnn），第 1-10 页。IEEE，
2020年8月

[59] 马南·托马尔、菲利普·汉森-埃斯特鲁奇、菲利普·巴赫曼、
Alex Lamb、John Langford、Matthew E. Taylor 和 Sergey
Levine。视频占用模型，2024年。3

[60] Samuel Triest、Matthew Sivaprakasam、Sean J Wang、Wen-shan
Wang、Aaron M Johnson 和 Sebastian Scherer。Tar-tandrive：用于
学习越野动力学模型的大规模数据集。2022 年国际机器人大会

与自动化（ICRA），第2546-2552页。IEEE，2022. 5，
1、2

[61] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan
Kautz。MoCoGAN：分解动作和内容
视频生成。在 IEEE 计算机视觉会议上
和模式识别（CVPR），第 1526-1535 页，2018 年。3

[62] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan
Kautz. Mocogan: 分解动作和内容
视频生成。在 IEEE 会议论文集上
计算机视觉和模式识别，第 1526-1535 页，
2018年3月

[63] Joseph Tung, Gene Chou, Ruojin Cai, Guandao Yang, Kai
张、戈登·韦茨斯坦、巴拉特·哈里哈兰和诺亚
Snavely。Megascenes：大规模场景级视图合成。
在计算机视觉 - ECCV 2024，第 197-214 页，Cham，
2025.施普林格·自然瑞士。3

[64] 托马斯·翁特蒂纳 (Thomas Unterthiner)、Sjoerd van Steenkiste、卡罗尔·库拉赫 (Karol Kurach)，
拉斐尔·马里尼尔、马尔辛·米哈尔斯基和西尔万·盖利。
Fvd：视频生成的新指标。2019. 5, 1

[65]贝尼尼奥·乌里亚、博尔哈·伊巴尔兹、安德里亚·巴尼诺、维尼修斯·赞巴尔
迪、达尚·库马兰、德米斯·哈萨比斯、卡斯韦尔·巴里、
以及查尔斯·布伦德尔。一个从自我中心到他者中心的模型
理解哺乳动物的大脑。bioRxiv，2022 年8 月

[66] 丹尼·瓦列夫斯基、亚尼夫·利维坦、莫阿布·阿拉尔和什洛米
Fruchter.扩散模型是实时游戏引擎。
arXiv 预印本 arXiv:2408.14837, 2024. 2, 3

[67]Basile Van Hoorick、Rundi Wu、Ege Ozguroglu、Kyle Sargent、
Ruoshi Liu、Pavel Tokmakov、Achal Dave、Changxi
郑和 Carl Vondrick。生成式摄影车：超强单目动态新视角合成。2024. 2

[68] Vaswani A. 注意力就是你所需要的一切。《神经科学进展》
信息处理系统， 2017.4

[69] Vikram Voleti、Alexia Jolicoeur-Martineau 和 Chris Pal。
用于预测的 Mcvd 掩蔽条件视频扩散，
生成和插值。神经信息处理系统进展，35:23371–23385，2022年。3

[70] Fu-Yun Wang, Zhaoyang Huang, Alexander Bergman,
Dazhong Shen, Peng Gao, Michael Lingelbach, Keqiang
Sun, Weikang Bian, Guanglu Song, Yu Liu, et al. Phased
一致性模型。神经信息处理系统进展，37:83951–84009，2024年。3

[71] 菲利普·吴、亚历杭德罗·埃斯孔特雷拉、丹尼尔·哈夫纳、彼得
Abbeel 和 Ken Goldberg。《白日梦想家：世界模特》
物理机器人学习。在机器人学习会议上，
第2226–2240页。PMLR，2023年。3

[72] Jingjing Xu, Xu Sun, Zhiyuan Zhang, Guangxiang Zhao, and
林俊阳.理解和改进层归一化，2019. 4

[73] Sherry Yang、Yilun Du、Seyed Kamyar Seyed Ghasemipour、
乔纳森·汤普森、莱斯利·帕克·凯尔布林、戴尔·舒尔曼斯和彼得·阿贝尔。交
互式现实世界学习
模拟器。在第十二届国际会议上
学习表征. 3

[74] Lijun Yu, Yong Cheng, Kihyuk Sohn, Jose Lezama, Han Zhang,
Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang,
Yuan Hao, Irfan Essa, et al. 马格维特：
蒙版生成视频转换器。
IEEE/CVF计算机视觉与模式会议
Recognition，第 10459–10469 页，2023 年。3

[75] 张理查德、菲利普·伊索拉、阿列克谢·埃弗罗斯、伊莱·谢赫特曼、
和Oliver Wang。深度
特征作为感知指标。CVPR，2018. 5, 1

[76] Richard Zhang、Phillip Isola、Alexei A Efros、Eli Shecht-man 和 Oliver
Wang。
深度特征作为感知指标。在
IEEE 计算机视觉和模式识别会议，第 586-595 页，2018 年。5

[77] 周高跃，潘恒凯，Yann LeCun，Lerrel Pinto。
Dino-wm：基于预训练视觉特征的世界模型实现零样本规划，2024. 5

导航世界模型

补充材料

附录的结构如下:我们开始
通过在第 7 节中描述如何通过独立规划来规划导航轨迹,然后在第 8 节中包括
更多的实验和结果。

7.独立计划优化

如第 3.3 节所述,我们使用预训练的 NWM 来
独立规划目标条件导航轨迹
优化方程5。在这里,我们提供了有关
使用交叉熵方法[48]进行优化,
所使用的超参数。完整的独立导航规划结果在第8.2 节中给出。

我们使用交叉熵来优化轨迹
方法,一种无梯度随机优化技术
用于连续优化问题。该方法迭代更新概率分布,以提高生成更好解的可能性。在无约束

独立规划场景中,我们假设轨迹是
直线并仅优化其端点,表示为
三个变量:单个平移 u 和偏航旋转 。
然后我们将这个元组映射到八个均匀分布的增量步骤,
在最后一步应用偏航旋转。时间间隔
步骤之间的时间间隔固定为 k = 0.25 秒。主要步骤
我们的优化流程如下:

·初始化:定义高斯分布
平均值 $\mu = (\mu\Delta x, \mu\Delta y, \mu\Phi)$和方差 $\Sigma =$
对数 $(\sigma^2_{\Delta x}, p^2_{\Delta y, \phi} p2)$在解空间上。
·采样:通过以下方式生成 N = 120 个候选解决方案
从当前高斯分布中采样。
·评估:通过使用 NWM 进行模拟并测量模拟输出和输入目标图像之间的 LPIPS
分数来评估每个候选解决方案。由于

NWM 是随机的,我们评估每个候选解决方案
M 次,取平均值得到最终得分。
·选择:选择表现最佳的解决方案的子集
根据 LPIPS 分数进行评估。
·更新:调整分布的参数,以增加生成类似解决方案的概率

表现最好的分布。此步骤最小化了旧分布和更新分布之间的交叉熵。

·迭代:重复采样、评估、选择和
更新步骤直到达到停止标准(例如收敛
或迭代限制)得到满足。

为了简单起见,我们只运行一次迭代的优化过程,我们发现这对于短期

规划了两秒钟,但还有进一步的改进
可能通过更多迭代来实现。当导航约束
应用时,部分轨迹被清零以尊重
这些约束。例如,在“前向优先”场景中,前五个平移动作为 u = (Δx, 0)

步骤,最后三步为 u = (0, Δy)。

8.实验与结果

8.1.实验研究

我们详细说明所使用的指标和数据集。
评估指标。我们描述评估指标
用于评估预测的导航轨迹和我们的 NWM 生成的图像的质量。

对于视觉导航性能,绝对轨迹误差(ATE) 通过计算估计轨迹与地面真实轨迹中对应点之间的欧氏距离来衡量轨迹估计的整体精度。相对姿态误差(RPE) 评估

通过计算误差来保证连续姿态的一致性
它们之间的相对变换[57]。
更严格地评估世界上的语义
模型输出,我们使用学习感知图像块
相似性(LPIPS)和 DreamSim [14],通过比较神经网络的深度特征来评估感知相似性
网络[75]。LPIPS尤其使用 AlexNet [33]来
关注人类对结构差异的感知。此外,我们使用峰值信噪比 (PSNR) 来

通过测量最大像素值与误差的比率来量化生成图像的像素级质量,其中

值越高,表示质量越好。
为了研究图像和视频合成质量,我们使用 ,
Frechet 初始距离 (FID) 和 Frechet 视频距离 (FVD),它们比较了

真实和生成的图像或视频。降低 FID 和 FVD
分数越高,视觉质量就越高[23, 64]。
数据集。对于所有机器人数据集,我们可以访问
机器人的位置和旋转,我们用它来推断
动作作为位置和旋转的增量。我们移除
所有向后移动在 No-MaD [55] 之后都可能出现抖动,从而将数据拆分为
SCAND [30]、 TartanDrive [60]、 RECON [52]的前向行走段,

以及 HuRoN [27]。我们还利用未标记的 Ego4D 视频,
我们只使用时间移位作为动作。接下来我们描述
每个单独的数据集。

· SCAND [30]是一个机器人数据集,包含社交
使用轮式机器人进行兼容导航演示
Clearpath Jackal 和一只腿状的 Boston Dynamics Spot。
SCAND 在室内和室外都有演示
德克萨斯大学奥斯汀分校的设置。数据集包含 8.7 小时,
138 条轨迹,25 英里的数据,我们使用相应的相机姿态。我们使用 484 个视频片段

	未知环境	已知环境			
数据	前往斯坦福	侦察	休罗恩	斯堪的纳维亚	格子呢大道
域内数据	0.658±0.002	0.295 ±0.002	0.250 ±0.003	0.403±0.002	0.414 ±0.001
+ Ego4D （未标记）	0.652 ±0.003	0.368±0.003	0.377±0.002	0.398 ±0.001	0.430±0.000

表 5.使用额外的未标记数据进行训练可提高在未知环境中的表现。报告未知环境中的结果
环境（Go Stanford)和已知环境（RECON）。通过评估未来 4 秒的 LPIPS 报告结果。

训练集和 121 个视频片段用于测试。用于
培训和评估。

- TartanDrive [60]是一个户外越野驾驶数据集
使用改装的 Yamaha Viking ATV 在匹兹堡收集的数据。该数据集包含 5 小
时的 630 条轨迹。我们使用 1,000 个视频片段进行训练,并使用 251

用于测试的视频片段。
- RECON [52]是一个使用 Clearpath Jackal UGV 平台收集的户外机器人数
据集。该数据集包含 9 个开放世界环境中的 40 小时数据。我们

使用 9,468 个视频片段进行训练,并使用 2,367 个视频
用于测试的片段。用于训练和评估。
- HuRoN [27]是一个机器人数据集,包含加州大学伯克利分校在室内环境中使
用 Roomba 机器人进行的社交互动。该数据集包含超过 75

在 5 个不同的环境中进行了数小时的训练,其中涉及 4,000 次人机
交互。我们使用 2,451 个视频片段进行训练,并
613 个视频片段用于测试。用于训练和
评估。
- GO Stanford [24, 25],一个机器人数据集,捕捉
两个不同的遥控机器人的鱼眼视频片段,
收集了至少 27 栋不同的斯坦福建筑
约25小时的视频片段。由于图像分辨率较低,我们仅将其用于域外评估。

- Ego4D [18]是一个大规模的以自我为中心的数据集,包含 74 个地点的 3670
个小时的数据。Ego4D 涵盖了各种场景,例如艺术与手工艺、烹饪、建筑、清洁
与洗衣以及杂货。

购物。我们只使用涉及视觉导航的视频,例如“杂货店购物”和“慢跑”。

我们总共使用了超过 908 小时的 1619 个视频用于训练。仅用于无标签训练。

培训。我们使用的视频来自以下
Ego4D 场景：“滑板/踏板车”、“轮滑”、“足球”、“参加节日或集市”、
“园丁”、“迷你高尔夫”、“骑摩托车”、“打高尔夫”

“骑自行车/慢跑”、“在街上行走”、“步行
狗/宠物”、“室内导航（步行）”、“在
户外用品店”、“衣服/其他购物”、“玩
宠物”、“室内购物”、“户外锻炼”、“农民”、“骑自行车”、“采花”、
“参加
体育赛事（观看和参与）”、“无人机飞行”、“参加讲座/课程”、“徒步
旅行”、“篮球”
“园艺”、“雪橇”、“去公园”。

视觉导航评估集。我们的主要发现是

构建视觉导航评估集的一个关键问题是,向前运动非常普遍,如果不仔细考虑,它
可能会主导评估数据。为了创建

不同的评估集,我们根据潜在评估轨迹的预测能力对它们进行排序,

继续前进。对于每个数据集,我们选择 100 个最不可预测的例子,并使用它们

以供评估。

时间预测评估集。预测未来
k 秒后的帧比估计更具挑战性
轨迹,因为它需要预测代理的轨迹及其在像素空间中的方向。因此,我们不

施加额外的多样性约束。对于每个数据集,
我们随机选择500个测试预测示例。

8.2. 实验和结果

在其他未标记数据上进行训练。我们在表5中列出了其他已知环境的结果,以及

图11.我们发现,在已知环境中,模型
仅使用领域内数据进行训练的模型往往表现更好,可能是因为它们更适合领域内

分布。唯一的例外是 SCAND 数据集,
存在动态物体（例如行走的人)的地方。
在这种情况下,添加未标记的数据可能有助于通过提供额外的多样化示例来提
高性能。
已知环境。我们在以下位置添加了使用 NWM 跟踪轨迹的附加可视化结果：

已知环境RECON（图12）、SCAND（图13）、HuRoN（图14)和Tartan Drive
（图15）。
此外,我们在表6 中还包括了 DIA-MOND 和 NWM 的完整 FVD 比较。

数据集	钻石	NWM（我们的）
侦察	762.734±3.361	200.969 ±5.629
HuRoN	881.981±11.601	276.932 ±4.346
TartanDrive	2289.687 ± 6.991	494.247 ± 14.433
斯堪的纳维亚	1945.085±8.449	401.699 ±11.216

表6.视频合成质量比较。16秒
以 4 FPS 生成的视频,报告 FVD（越低越好）。

规划（排名）。表 7 显示了所有域内数据集的完整目标条件导航结果。

与 NoMaD 相比,我们观察到持续的改进
当使用 NWM 从 16 条轨迹池中进行选择时,
从 32 个更大的池子中进行选择时,可以获得进一步的收益。

模型	侦察		休罗恩		格子呢		斯堪的纳维亚	
	吃	实时传输	吃	实时传输	吃	实时传输	吃	实时传输
前进	1.92±0.00	0.54±0.00	4.14±0.00	1.05±0.00	5.75±0.00	1.19±0.00	2.97±0.00	0.62±0.00
GNM	1.87±0.00	0.73±0.00	3.71±0.00	1.00±0.00	6.65±0.00	1.62±0.00	2.12±0.00	0.61±0.00
平均径	1.95±0.05	0.53±0.01	3.73±0.04	0.96±0.01	6.32±0.03	1.31±0.01	2.24±0.03	0.49±0.01
NWM + NoMaD (×16)	1.88 ± 0.03	0.51 ± 0.01	3.73 ± 0.05	0.95 ± 0.01	6.26 ± 0.06	1.30 ± 0.01	2.18 ± 0.05	0.48 ± 0.01
NWM + NoMaD (×32)	1.79 ± 0.02	0.49 ± 0.00	3.68 ± 0.03	0.95 ± 0.01	6.25 ± 0.05	1.29 ± 0.01	2.19 ± 0.03	0.47 ± 0.01
NWM（仅）	1.13 ±0.02	0.35 ±0.01	4.12±0.03	0.96±0.01	5.63 ±0.06	1.18 ±0.01	1.28 ±0.02	0.33 ±0.01

表 7.目标条件视觉导航。ATE和 RPE 在所有领域数据集上的结果,预测最多 2 条轨迹
秒。与之前的方法 NoMaD [55]和 GNM [53] 相比,NWM 在所有指标上都取得了更好的结果。

对于 Tartan Drive,我们注意到数据集主要由向前运动主导,这反映在与“向前”基线的比较结果中，“向前”基线是一个预测模型，

始终选择仅向前运动。

独立规划。对于独立规划,我们运行
第7节中概述的优化程序,共 1 步,并且
对每条轨迹进行 3 次评估。对于所有数据集,我们
初始化μΔy和μ为 0,σ $^{2}_{\Delta y}$ 和 s $^{2}_{\phi}$ 为 0.1。我们
使用不同的（μΔx，σ2 Δx)在每个数据集上：(-0.1,0.02）
对于 RECON,对于 TartanDrive,为 (0.5, 0.07),对于 (-0.25, 0.04)
SCAND,HuRoN 为（-0.33, 0.03）。我们包括完整的
独立导航规划结果见表7。我们发现
在独立环境中使用规划比其他方法表现更好,尤其是以前的

硬编码策略。

现实世界的适用性。部署的关键瓶颈
现实世界机器人技术中的 NWM 关键在于推理速度。我们评估了提升 NWM 效率的方法,并测量了它们的
对运行时间的影响。我们专注于使用带有生成策略的 NWM（第3.3 节)对
32 条四秒轨迹进行排名。
由于轨迹评估是可并行的,我们分析
模拟单个轨迹的运行时间。我们发现现有的解决方案已经可以实现实时应用

NWM 为 2-10HZ（表8）。

NWM	+时间跳跃	+蒸馏。	+量化。4位
30.3±0.2	14.7±0.1		0.4±0.10.1（估计值[12]）

表 8. NVIDIA RTX 6000 Ada 卡上的运行时间（秒）。

推理时间可以通过组合每个
相邻的动作对（通过等式2），然后仅模拟 8
未来状态而不是 16 个（“时间跳跃”），这并不
降低导航性能。减少扩散
通过模型蒸馏将去噪步骤从 250 减少到 6 [70]
进一步加快推理速度,同时减少视觉质量损失。3
综合起来,这两个想法可以使 NWM 运行
实时。量化到4位,我们还没有探索过，
可以在不影响性能的情况下实现 4 倍的速度提升[12]。

3使用来自https:// 的DiT 蒸馏实现
github.com/hao-ai-lab/FastVideo

CDiT-L	背景	2 仅行动	目标	2	0.656	0.654	我们的	我们的+TTA
					0.655	0.661	0.652	0.650

表 9. 未知环境下的结果（“Go Stanford”）。报告了 4 秒未来预测的 lpips。值越低越好。

测试时适应。测试时适应已证明
改进视觉导航[13, 16]。使用世界模型进行规划和测试时自适应之间是什么关系?我们假设这两个想法是正交的,并且

包括测试时自适应结果。我们考虑一种简化的自适应方法,通过对 NWM 进行 2k 步微调
在未知环境中的轨迹上。我们证明
这种调整改善了该环境中的轨迹模拟（参见表9 中的“我们的+TTA”），其中我们还包括了额外的基线和消融。

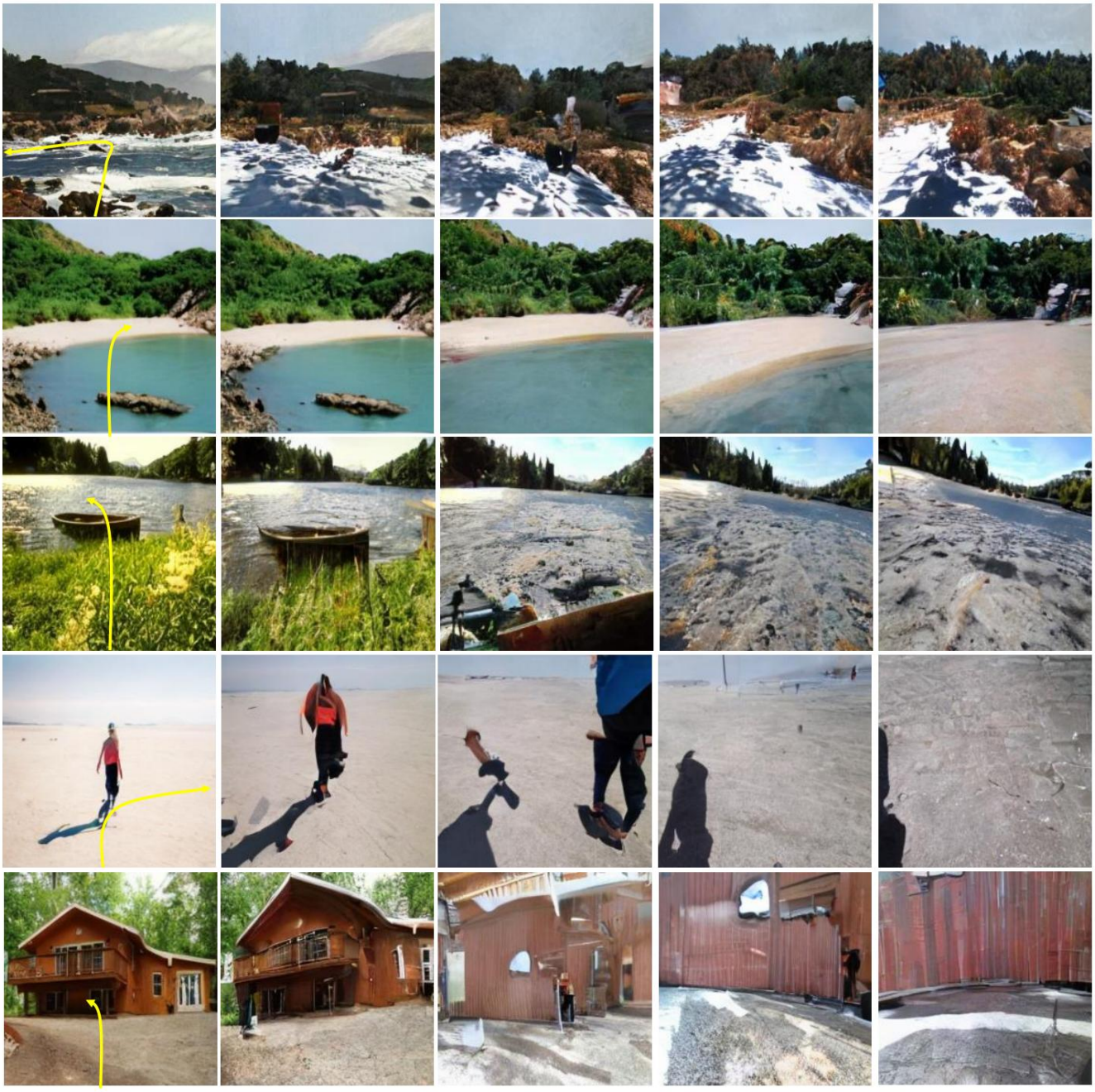

图 11.未知环境导航。NWM以单幅图像为条件，并根据相关动作（黄色标记）自回归预测下一个状态，最长 4 秒，帧率为 4 FPS。我们绘制了 1、2、3 和 4 秒后的生成结果。

图 12. RECON 上的视频生成示例。NWM以单张首图和一条真实轨迹为条件，以 4 FPS 的帧率自回归预测接下来最多 16 秒的视频。我们以每秒 1 帧的频率绘制了 2 到 16 秒的生成结果。

图 13. SCAND 上的视频生成示例。NWM以单张首图和一条真实轨迹为条件，以 4 FPS 的帧率自回归预测接下来最多 16 秒的视频。我们以每秒 1 帧的频率绘制了 2 到 16 秒的生成结果。

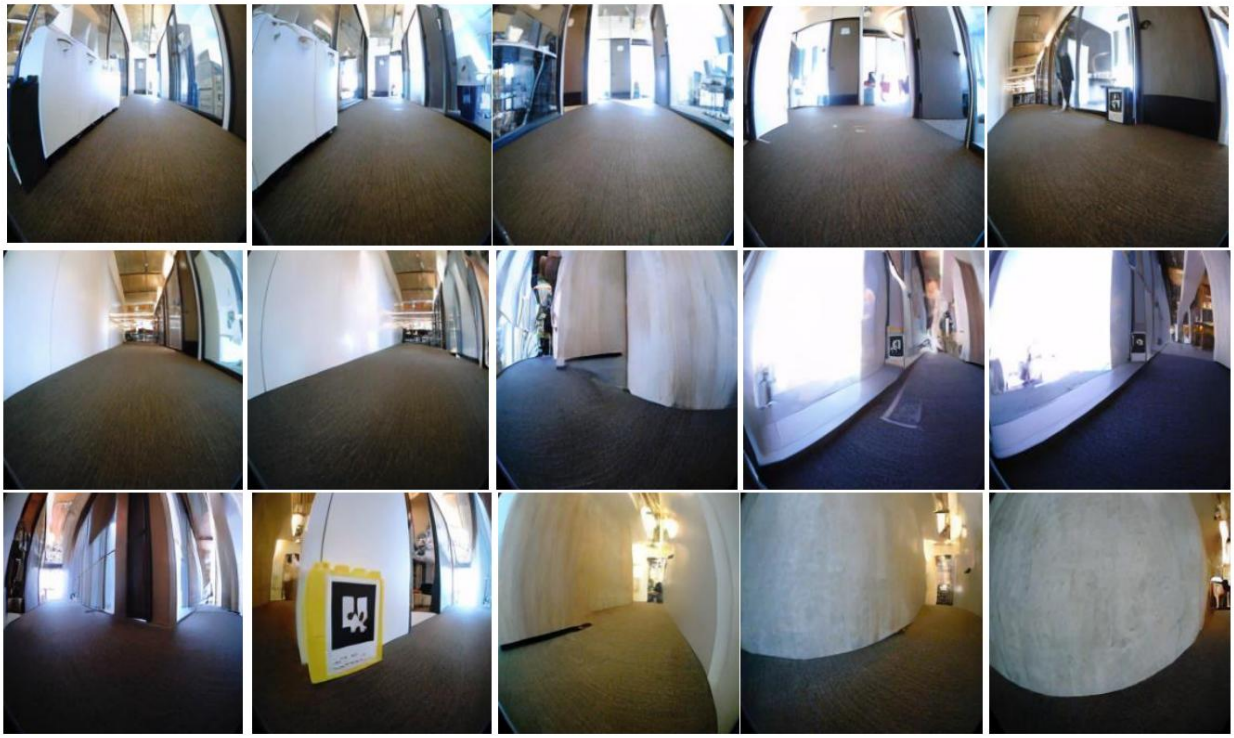

图 14. HuRoN 上的视频生成示例。NWM以单张首图和一条真实轨迹为条件，以 4 FPS 的帧率自回归预测接下来最多 16 秒的视频。我们以每秒 1 帧的频率绘制了 2 到 16 秒的生成结果。

图 15. Tartan Drive 上的视频生成示例。NWM以单张首图和一条真实轨迹为条件，以 4 FPS 的帧率自回归预测接下来最多 16 秒的视频。我们以每秒 1 帧的频率绘制了 2 到 16 秒的生成结果。