GeoWorld: Geometric World Models

Zeyu Zhang[1] Danning Li[2] Ian Reid[2] Richard Hartley[1]
[1]ANU [2]MBZUAI

https://steve-zeyu-zhang.github.io/GeoWorld

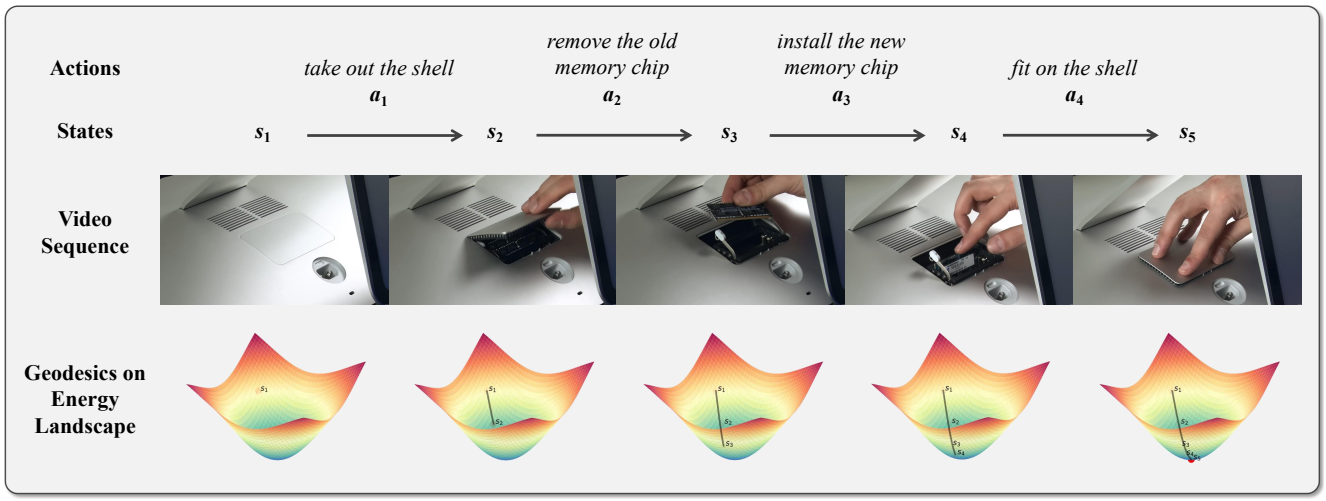


Figure 1. **Energy-based planning by GeoWorld.** The diagram shows a *Replace Memory Chip* task from the COIN dataset [71], where GeoWorld plans actions by following geodesics over a hyperbolic energy landscape rather than generating pixels.

Abstract

Energy-based predictive world models provide a powerful approach for multi-step visual planning by reasoning over latent energy landscapes rather than generating pixels. However, existing approaches face two major challenges: (i) their latent representations are typically learned in Euclidean space, neglecting the underlying geometric and hierarchical structure among states, and (ii) they struggle with long-horizon prediction, which leads to rapid degradation across extended rollouts. To address these challenges, we introduce ***GeoWorld****, a geometric world model that preserves geometric structure and hierarchical relations through a Hyperbolic JEPA, which maps latent representations from Euclidean space onto hyperbolic manifolds. We further introduce Geometric Reinforcement Learning for energy-based optimization, enabling stable multi-step planning in hyperbolic latent space. Extensive experiments on CrossTask and COIN demonstrate around* ***3%*** *SR improvement in 3-step planning and* ***2%*** *SR improvement in 4-step planning compared to the state-of-the-art V-JEPA 2.*

1. Introduction

Autoregressive (AR) next-token prediction has endowed large language models (LLMs) [79] and vision-language models (VLMs) [20, 53] with extensive world knowledge and reasoning capability, enabling them to effectively tackle complex tasks involving searching [78], reasoning [32, 33, 44], and planning [8, 35, 65, 82]. Although the success of LLMs stems from modeling within language space, which serves as a shortcut toward human level knowledge [48], they still fail to fully represent the rich information of the real world, such as its physical and geometric properties [3]. In the real world, human and biological cognition often acquire knowledge primarily through visual information rather than relying solely on language, as vision offers a higher bandwidth of information than language [62]. For example, human infants learn mainly from visual perception during the first few months before developing a language system [38], and some animals do not pos-

sess language at all [18]. Therefore, there are world models [3, 5, 46, 59, 63] that learn solely from visual input, such as videos, and perform planning with either a generative or a predictive approach. Generative world models [46, 59, 63] explicitly generate pixels or latent visual tokens that decode into pixels in order to predict only one step at a time [67]. As a result, they lack awareness of the full trajectory structure or the energy landscape over multiple steps. In contrast, predictive world models [2, 3, 5, 29] such as JPEA [41] do not generate pixels. Instead, they learn an energy landscape in latent space that measures the compatibility between current and target states. This enables multi-step hierarchical planning, where high-level reasoning minimizes energy in latent space, while lower-level modules fill in the physical details.

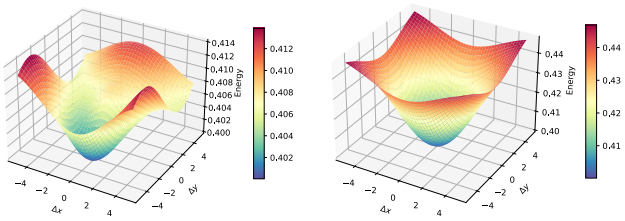
However, existing energy-based predictive world models face two significant challenges:

(1) Geometric neglect. Although predictive world models perform multi-step hierarchical planning in latent space, their representations are typically learned in a Euclidean space without preserving the underlying geometric relations among states. As a result, the learned energy landscape fails to capture meaningful geodesic distances or hierarchical embeddings between latent states [51], which weakens the model's ability to perform geometry-consistent planning over long horizons.

(2) Multi-step shortcoming. Multi-step videos are limited and expensive to acquire, so existing predictive world models are primarily trained on one-step video transitions [12, 31, 37, 40, 47, 66]. Although learning an energy landscape over entire trajectories conceptually enables long-horizon planning, their performance degrades rapidly as the planning horizon increases, exposing a weakness in modeling long-term temporal dependencies.

Our motivation is to address these problems from a *geometric* perspective. For the first challenge, a geometry-aware world model is required to preserve geometric properties when learning the energy landscape for hierarchical planning. For the second challenge, reinforcement learning (RL) has proven effective in adjusting a pretrained foundation model when its outputs are unsatisfactory in certain aspects [54, 60]. Therefore, a geometry-aware RL method is required to obtain optimal trajectories on the latent manifold, improving the model's multi-step planning capability.

Hence, we introduce the **Geometric World Model** (**GeoWorld**), a method that enhances energy-based predictive world models by preserving geometric structure and hierarchical awareness in latent space, as shown in Figure 1. To address the first challenge, we propose *Hyperbolic JEPA* (*H-JEPA*), which maps latent representations from Euclidean space $\mathbb{R}^n$ onto a hyperbolic manifold $\mathbb{H}^n$, where geodesic distances naturally encode hierarchical relations among states. By learning dynamics along hyperbolic geodesics, H-JEPA preserves latent geometry during multi-step prediction, ensuring that the learned energy landscape aligns with the underlying structure of the physical world and supports geometry-consistent planning, as shown in 2.



(a) V-JEPA 2 [3] Energy Landscape (b) **GeoWorld** Energy Landscape

Figure 2. **Energy landscape comparison for V-JEPA 2 [3] and GeoWorld.** We visualize the energy by sweeping two orthonormal tangent-space directions $(\Delta x, \Delta y)$ around a reference latent state. GeoWorlds yields a structured, curvature-aware energy landscape that better reflects geometric structure and hierarchical relations among latent states and improves energy-based planning. For more details see *Appendix* 4.

To address the second challenge, we design a *Geometric Reinforcement Learning* (*GRL*) that reformulates multi-step planning as the optimization of an energy-based value function, where lower hyperbolic energy corresponds to higher cumulative reward. GRL directly optimizes the predictor of the world model without training an additional policy or reward model. By adjusting the predictor's energy-based value representation through hyperbolic geodesics minimization and triangle inequality regularization, GRL enforces geodesic-consistent rollouts on the latent manifold, effectively improving long-horizon stability and planning performance.

To verify our method's capability on long-horizon planning, we evaluate multi-step goal-conditioned visual planning on standard benchmarks, including CrossTask [88] and COIN [71]. Our GeoWorld achieves consistent improvements over the previous state-of-the-art predictive world model V-JEPA 2, including improvements of around **3% SR** in 3-step planning and **2% SR** in 4-step planning across both datasets.

The contributions of our work can be summarized as follows:

- We introduce the **Geometric World Model** (**GeoWorld**) with a *Hyperbolic JEPA* (*H-JEPA*), which preserves geometric structure and hierarchical relations by mapping latent representations onto a hyperbolic manifold and learning dynamics along hyperbolic geodesics, resulting in a geometry-consistent energy landscape for multi-step prediction and planning.
- We propose *Geometric Reinforcement Learning* (*GRL*), an energy-based optimization framework that directly refines the predictor through hyperbolic energy mini-

mization and triangle-inequality regularization, enabling geodesic-consistent rollouts and improving long-horizon planning stability.
- We demonstrate strong performance on long-horizon goal-conditioned visual planning across CrossTask and COIN, achieving around **3%** SR improvement in 3-step planning and **2%** SR improvement in 4-step planning compared to V-JEPA 2.

2. Related Works

Video World Models There are two primary approaches for video world modeling: generative world models [46, 59, 63] and predictive world models [2, 3, 5, 29, 41]. Generative world models typically build upon autoregressive [24, 43, 77] or semi-autoregressive [17, 22, 34, 58, 72, 81] architectures that observe the visual context and explicitly generate the next frame or its latent representation. These models often incorporate an inverse dynamics module [67] trained to infer actions from consecutive observations, enabling one-step reactive control but preventing multi-step reasoning because the model lacks access to the global trajectory structure and cannot capture long-range dynamics. Moreover, generative approaches must decode visual tokens or pixels during planning, which introduces unnecessary noise and computational overhead and limits their ability to model abstract energy landscapes for hierarchical planning [59]. In contrast, predictive world models do not generate pixels. Instead, they learn an energy landscape in latent space that quantifies the compatibility between current and target states [41]. This design allows for multi-step trajectory optimization using sampling-based planners such as the cross-entropy method (CEM) [23], enabling long-horizon planning without explicit pixel decoding.

Goal-Conditioned Visual Planning Goal conditioned visual planning aims to produce a sequence of actions that achieves a given goal based on visual observations. Prior works have evolved into three independent setups depending on the modalities of the observation and the goal, which may be images, videos, or language. (1) In visual planning for assistance (VPA) [56], observations are videos and goals are described in natural language. This typically requires models built on LLMs with multimodal processing capability [16, 36, 84]. (2) In procedural planning (PP) [15], both observations and goals are specified as images without any language involved, which limits the model's ability to capture temporal information in the physical world [7, 36, 42, 50, 52, 61, 69, 74, 75, 86, 87]. (3) In visual planning with videos, both observations and goals are given as videos, which aligns more naturally with temporal dynamics in the real world and is commonly addressed using video LLMs [20, 53, 76, 79], generative world models [59], and predictive world models [3].

3. Method

3.1. Overview

We introduce **GeoWorld**, a geometric world model designed to enhance long-horizon visual planning by preserving geometric structure and hierarchical awareness in latent space. To address the limitation of Euclidean latent representations, GeoWorld incorporates *Hyperbolic JEPA* (H-JEPA), which maps encoder outputs from Euclidean space onto a hyperbolic manifold where geodesic distances naturally encode hierarchical relations among states. By learning latent dynamics along hyperbolic geodesics, H-JEPA enforces geometry-consistent transitions that better reflect the structure of real-world trajectories. To further improve stability in multi-step prediction, we develop *Geometric Reinforcement Learning* (GRL), an energy-based optimization framework that treats planning as minimizing a hyperbolic value function without training an additional policy or reward model. GRL refines the predictor through hyperbolic energy minimization and triangle-inequality regularization, encouraging geodesic-consistent rollouts and improving long-horizon temporal coherence. Leveraging energy-based planning with the Cross-Entropy Method (CEM) [23] further enables efficient trajectory optimization by searching for action sequences that follow geodesic paths in hyperbolic latent space. Together, H-JEPA and GRL form the core of GeoWorld, enabling geometry-aware multi-step planning in predictive world models, as shown in Figure 3.

For preliminaries on JEPA [41], hyperbolic geometry, and the value function in RL, see *Appendix* 1.

3.2. Hyperbolic JEPA

From a representation perspective, we aim to learn a mapping from states onto a hyperbolic space $\mathbb{H}^n$ such that the optimal plan corresponds to a geodesic in hyperbolic space. Hence, we propose *Hyperbolic JEPA* (*H-JEPA*), which models latent dynamics on the hyperbolic manifold to preserve hierarchical relations and underlying geometric coherence during multi-step planning.

We define the observation at time t as x_t. $E_\theta(\cdot)$ denotes the pretrained encoder [3], which encodes the observation x_t into the latent state s_t^x:

$$s_t^x = E_\theta(x_t) \in \mathbb{R}^n. \tag{1}$$

To effectively map the encoder output from Euclidean space $\mathbb{R}^n$ to hyperbolic space $\mathbb{H}^n$, we interpret the Euclidean embedding s_t^x as a tangent vector in the tangent space $\mathbf{T}_0\mathbb{H}^n$ at the origin. We then apply the exponential map at the origin of the Poincaré ball model $\mathbb{B}_c^n$ with curvature $K = -c$, which projects the tangent vector onto the hyperbolic manifold, as detailed in *Appendix* 1.5.1.

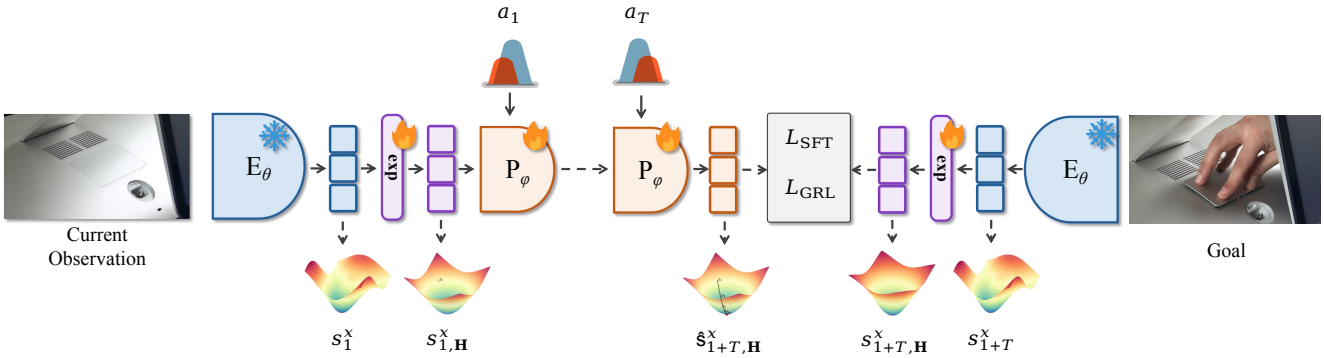


Figure 3. **Overview of GeoWorld.** Our geometric world model integrates Hyperbolic JEPA for geometry-preserving latent dynamics and Geometric Reinforcement Learning for geodesic-consistent multi-step refinement. Together with energy-based planning using CEM, GeoWorld enables stable and geometry-aware long-horizon visual planning.

Formally, the hyperbolic latent state is obtained as

$$s_{t,\mathbb{H}}^x = \exp_0(s_t^x) = \tanh\big(\sqrt{c}\|s_t^x\|\big)\frac{s_t^x}{\sqrt{c}\|s_t^x\|}, \quad s_{t,\mathbb{H}}^x \in \mathbb{B}_c^n. \tag{2}$$

Then the action-conditioned predictor $P_\phi(\cdot)$ takes a sequence of hyperbolic latent states $(s_{t,\mathbb{H}}^x)_{t=1}^T$ and a corresponding sequence of actions $(a_t)_{t=1}^T$ as input, and predicts the sequence of next-state representations $(\hat{s}_{t+1,\mathbb{H}}^x)_{t=1}^T$ over a planning horizon T:

$$(\hat{s}_{t+1,\mathbb{H}}^x)_{t=1}^T = P_\phi\big((s_{t,\mathbb{H}}^x, a_t)_{t=1}^T\big). \tag{3}$$

and θ and ϕ denote the parameters (weights) of the encoder and predictor networks, respectively.

3.3. Training Objective

The supervised training objective of H-JEPA is to learn a predictive world model that follows the geodesic path of minimum energy cost between the current and target latent states in hyperbolic space. Specifically, the model minimizes the Poincaré-ball hyperbolic distance $d_{\mathbb{H}}$ between the predicted and true latent representations, as defined in Eq. 68 of *Appendix* 1.5.1, ensuring that each transition aligns with the lowest-energy trajectory on the manifold.

The objective consists of a joint loss combining a teacher-forcing loss and a rollout loss. The teacher-forcing loss encourages accurate one-step prediction by aligning the predicted next-state representation with the ground-truth latent embedding, while the rollout loss recursively feeds the model's own predictions as inputs to enforce temporal consistency across multiple future steps.

Teacher Forcing The teacher-forcing loss trains the model to accurately perform one-step future prediction by minimizing the hyperbolic geodesic distance $d_{\mathbb{H}}$ between the predicted latent representation $\hat{s}_{t+1,\mathbb{H}}^x$ and the encoded ground-truth latent $s_{t+1,\mathbb{H}}^x$ at each time step t:

$$\mathcal{L}_{\text{TF}}(\theta,\phi) = \frac{1}{T}\sum_{t=1}^{T} d_{\mathbb{H}}\big(P_\phi\big(\exp_0\big(E_\theta(x_t)\big), a_t\big), \exp_0\big(E_\theta(x_{t+1})\big)\big) \tag{4}$$

$$= \frac{1}{T}\sum_{t=1}^{T} d_{\mathbb{H}}\big(\hat{s}_{t+1,\mathbb{H}}^x, s_{t+1,\mathbb{H}}^x\big) \tag{5}$$

$$= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\sqrt{c}}\,\text{arcosh}\Big(1 + 2c\frac{\|\hat{s}_{t+1,\mathbb{H}}^x - s_{t+1,\mathbb{H}}^x\|^2}{(1-c\|\hat{s}_{t+1,\mathbb{H}}^x\|^2)(1-c\|s_{t+1,\mathbb{H}}^x\|^2)}\Big) \tag{6}$$

Rollout The rollout loss feeds the predictor's output back as input, enabling the model to learn multi-step future prediction. In this case, we design a two-step rollout loss to enhance the model's capability for long-horizon planning:

$$\mathcal{L}_{\text{rollout}}(\theta,\phi) = \frac{1}{T}\sum_{t=1}^{T} d_{\mathbb{H}}\big(P_\phi\big(\exp_0\big(E_\theta(x_t)\big), a_t, a_{t+1}\big), \exp_0\big(E_\theta(x_{t+2})\big)\big) \tag{7}$$

$$= \frac{1}{T}\sum_{t=1}^{T} d_{\mathbb{H}}\big(\hat{s}_{t+2,\mathbb{H}}^x, s_{t+2,\mathbb{H}}^x\big) \tag{8}$$

$$= \frac{1}{T}\sum_{t=1}^{T} \frac{1}{\sqrt{c}}\,\text{arcosh}\Big(1 + 2c\frac{\|\hat{s}_{t+2,\mathbb{H}}^x - s_{t+2,\mathbb{H}}^x\|^2}{(1-c\|\hat{s}_{t+2,\mathbb{H}}^x\|^2)(1-c\|s_{t+2,\mathbb{H}}^x\|^2)}\Big), \tag{9}$$

Total Loss Hence, the total loss in the supervised stage is defined as

$$\mathcal{L}_{\text{SFT}}(\theta, \phi) = \lambda \mathcal{L}_{\text{TF}}(\theta, \phi) + (1 - \lambda) \mathcal{L}_{\text{rollout}}(\theta, \phi), \quad (10)$$

where λ is a loss weighting hyperparameter.

Together, these two components train the predictor to learn smooth, geodesically consistent trajectories that capture both short-term accuracy and long-horizon stability within the hyperbolic latent space.

3.4. Geometric Reinforcement Learning

We propose a *Geometric Reinforcement Learning* (*GRL*) approach that improves the predictor in multi-step planning by adjusting its energy-based value representation, aligning lower energy with higher expected reward.

Energy Cost Given a frozen encoder E and a trainable predictor P_ϕ, we define the energy cost of moving from state $s^x_{t,\mathbb{H}}$ to state $s^x_{t+1,\mathbb{H}}$ as

$$c_t(s^x_{t,\mathbb{H}}, s^x_{t+1,\mathbb{H}}) = d_{\mathbb{H}}\big(P_\phi\big(\exp_0\big(E(x_t)\big), a_t\big), \exp_0\big(E(x_{t+1})\big)\big) \quad (11)$$

$$= d_{\mathbb{H}}(\hat{s}^x_{t+1,\mathbb{H}}, s^x_{t+1,\mathbb{H}}). \quad (12)$$

which ideally indicates that we aim to minimize the energy cost of moving from state $s^x_{t,\mathbb{H}}$ to state $s^x_{t+1,\mathbb{H}}$, which is identical to minimizing the geodesic distance between the predicted state $\hat{s}^x_{t+1,\mathbb{H}}$ and the target state $s^x_{t+1,\mathbb{H}}$.

Reward We then define the reward as the negative energy cost of moving between states:

$$r_t(s^x_{t,\mathbb{H}}, a_t, s^x_{t+1,\mathbb{H}}) = -c_t(s^x_{t,\mathbb{H}}, s^x_{t+1,\mathbb{H}}). \quad (13)$$

Path Value Function As mentioned in *Appendix 1.6*, the value function V is a mathematical object that quantifies the amount of energy required for an agent to reach optimality from a given state to a target state, where lower energy corresponds to a higher expected cumulative reward.

Hence, the path value function between the current and goal latent states, given a planning horizon T, is defined as the expected cumulative reward:

$$V(s^x_{1,\mathbb{H}}, s^x_{1+T,\mathbb{H}}) = \mathbb{E}_{a_{1:T}\sim\phi}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t(s^x_{t,\mathbb{H}}, a_t, s^x_{t+1,\mathbb{H}})\right], \quad (14)$$

where $\gamma \in [0, 1)$ is the discount factor.

Our objective is to maximize the total reward (i.e., maximize the return) such that P_ϕ follows the geodesics. Therefore, the optimal path value function maximizes the expected cumulative reward:

$$V^*(s^x_{1,\mathbb{H}}, s^x_{1+T,\mathbb{H}}) = \max_\phi \mathbb{E}_{a_{1:T}\sim\phi}\left[\sum_{t=1}^{T} \gamma^{t-1} r_t(s^x_{t,\mathbb{H}}, a_t, s^x_{t+1,\mathbb{H}})\right].$$

$$= \min_\phi \mathbb{E}_{a_{1:T}\sim\phi}\left[\sum_{t=1}^{T} \gamma^{t-1} d_{\mathbb{H}}(\hat{s}^x_{t+1,\mathbb{H}}, s^x_{t+1,\mathbb{H}})\right]. \quad (15)$$

which is equivalent to minimizing the total hyperbolic distance between the predicted and target states.

Triangle Inequality Regularization The hyperbolic geodesic distance $d_{\mathbb{H}}$ satisfies the triangle inequality. Therefore, for any consecutive triplet in the predictor's rollouts:

$$d_{\mathbb{H}}(\hat{s}^x_{t,\mathbb{H}}, \hat{s}^x_{t+2,\mathbb{H}}) \le d_{\mathbb{H}}(\hat{s}^x_{t,\mathbb{H}}, \hat{s}^x_{t+1,\mathbb{H}}) + d_{\mathbb{H}}(\hat{s}^x_{t+1,\mathbb{H}}, \hat{s}^x_{t+2,\mathbb{H}}). \quad (16)$$

This indicates that minimizing the sum of consecutive step distances encourages the predicted trajectory to align with the geodesic path. Hence, we introduce a regularization term:

$$\mathcal{L}_\Delta = \frac{1}{T-2} \sum_{t=1}^{T-2} \Big[d_{\mathbb{H}}(\hat{s}_t, \hat{s}_{t+2}) - d_{\mathbb{H}}(\hat{s}_t, \hat{s}_{t+1}) - d_{\mathbb{H}}(\hat{s}_{t+1}, \hat{s}_{t+2})\Big]_+. \quad (17)$$

This term enforces multi-step rollout consistency by encouraging predicted trajectories to satisfy hyperbolic geodesic properties.

Total Loss Hence, the total loss in Geometric Reinforcement Learning can be expressed as

$$\mathcal{L}_{\text{GRL}}(\phi) = \mathbb{E}_{a_{1:T}\sim\phi}\left[\sum_{t=1}^{T} \gamma^{t-1} d_{\mathbb{H}}(\hat{s}^x_{t+1,\mathbb{H}}, s^x_{t+1,\mathbb{H}})\right] + \beta \mathcal{L}_\Delta. \quad (18)$$

where β is the regularization factor.

3.5. Energy-Based Planning

We then perform energy-based planning after training, with the frozen encoder E and predictor P. The predictor serves as a world model, capable of predicting how latent representations evolve when an action sequence is applied. During planning, we search for an optimal action sequence that follows the geodesic path between the current and goal latent states, effectively minimizing a goal-conditioned energy cost defined in the hyperbolic latent space.

Given the current observation x_1, the future target x_{1+T}, and the planning horizon T, we encode the current and goal observations as

$$s^x_{1,\mathbb{H}} = \exp_0(E(x_1)), \qquad s^x_{1+T,\mathbb{H}} = \exp_0(E(x_{1+T})). \tag{19}$$

We then define the energy cost function C based on the Poincaré geodesic distance, which measures the hyperbolic energy between the predicted and goal latent states over the planning horizon:

$$C((\hat{a}_t)_{t=1}^T; s^x_{1,\mathbb{H}}, s^x_{1+T,\mathbb{H}}) = d_{\mathbb{H}}\big(P((\hat{a}_t)_{t=1}^T; s^x_{1,\mathbb{H}}),\ s^x_{1+T,\mathbb{H}}\big)\,, \tag{20}$$

Hence, the optimal action sequence $(a_t^*)_{t=1}^T$ is obtained by minimizing this hyperbolic energy cost:

$$(a_t^*)_{t=1}^T = \arg\min_{(\hat{a}_t)_{t=1}^T} d_{\mathbb{H}}\big(P((\hat{a}_t)_{t=1}^T; s^x_{1,\mathbb{H}}),\ s^x_{1+T,\mathbb{H}}\big)\,. \tag{21}$$

The optimization is performed with the Cross-Entropy Method (CEM) [23], as detailed in Algorithm 1 of *Appendix* 1.3.3

$$(a_t^*)_{t=1}^T = \mathrm{CEM}(x_1,\ x_{1+T},\ P(\cdot),\ E(\cdot),\ T,\ N,\ K,\ I,\ \mu_0,\ \Sigma_0) \tag{22}$$

where x_1 is the current observation, x_{1+T} is the goal observation, P_ϕ is the predictor, $E(\cdot)$ is the encoder, T is the planning horizon, N is the number of samples, K is the number of elites, I is the number of iterations, and (μ_0, Σ_0) denote the initial mean and covariance of the action distribution.

4. Experiments

4.1. Benchmarks and Evaluation Metrics

Benchmarks For evaluating our world model's capability in multi-step goal-conditioned planning, we adapt two standard goal-conditioned visual planning datasets, CrossTask [88] and COIN [71], which contain diverse fine-grained action labels and timestamps of human daily activities.

CrossTask consists of 4.7K videos across 83 tasks, covering 105 actions, with an average of 8 actions per video. The total duration is 375h.

COIN consists of 11,287 videos across 180 tasks, covering 778 actions, with an average of 3.9 actions per video. The total duration is 476h.

Metrics Following previous works in goal-conditioned visual planning [7], we adopt three metrics for evaluation: (1) Success Rate (SR) computes whether the predicted action sequence exactly matches the ground truth sequence. (2) Mean Accuracy (mAcc) computes the average accuracy of the predicted actions at each time step. (3) Mean Intersection over Union (mIoU) quantifies the overlap between the predicted procedure and the ground truth.

4.2. Baseline and Evaluation Protocol

We follow previous works [3, 7, 15, 59] and evaluate goal-conditioned visual planning in two setups based on the modality of the observation and the target, as discussed in Section 2. For *procedural planning* [15], both observations and goals are specified as images, which is more aligned with the traditional visual planning setup. For *visual planning with videos* [59], both observations and goals are specified as video clips, which more faithfully reflect the temporal–spatial information in the real world.

For both setups, evaluation is conducted over a planning horizon T, where the model outputs a sequence of T actions given the observation and the goal.

In both setups, we include three categories of baselines. *LLM-based* methods leverage LLMs or VLMs for reasoning and planning [20, 36, 42, 52, 53, 76, 79]. *Generative (world) models* explicitly generate pixels or latent visual tokens that decode into pixels for planning [7, 15, 50, 59, 61, 75, 86, 87]. *Predictive (world) models* predict a sequence of actions without relying on pixel generation [1, 3, 27, 68, 69, 74].

There are two extra baselines in the procedural planning setup. *Random* randomly selects an action from all actions and serves as the empirical lower bound of performance [15]. The *Retrieval-Based* approach retrieves the nearest neighbor by minimizing the visual feature distance within the training dataset, and the action sequence associated with the retrieved neighbor is then used as the plan [87].

Besides, for both V-JEPA 2 [3] and our GeoWorld, we adopt frozen encoders, while for VideoWorld [59] we perform full finetuning. For general VLMs [20, 53, 76, 79] in visual planning with videos, all evaluations are conducted in a zero-shot setting. For more details on the baselines, see *Appendix* 3.

4.3. Implementation Details

For a fair comparison, both the V-JEPA 2 [3] baseline and our GeoWorld adopt frozen encoders pretrained on VideoMix22M. The exponential map $\exp_0(\cdot)$ is implemented and trained as a differentiable hyperbolic projection layer, where the curvature c is treated as a learnable parameter [14]. The predictor network $P_\phi(\cdot)$ is a $\sim$300M-parameter transformer with 24 layers, 16 heads, a 1024-dimensional hidden size, and GELU activations.

We conduct a two-stage training procedure for both V-JEPA 2 and our GeoWorld, consisting of supervised post-training followed by geometric reinforcement learning.

In the supervised post-training stage, both V-JEPA 2 and

Table 1. **Goal-conditioned visual planning with images on CrossTask [88] and COIN [71] datasets.** We evaluate multi-step planning over a horizon T under the *procedural planning* setup [15], where both observations and goals are specified as images.

Method	CrossTask Dataset [88]						COIN Dataset [71]					
	T=3			T=4			T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU	SR	mAcc	mIoU	SR	mAcc	mIoU
Random [15]	0.01	0.94	1.66	0.01	1.83	1.66	0.01	0.01	2.47	0.01	0.01	2.32
Retrieval-Based [87]	8.05	23.30	32.06	3.95	22.22	36.97	–	–	–	–	–	–
LLM-Based												
LFP [42]	30.55	59.59	76.86	15.97	50.70	75.30	30.64	54.72	76.86	15.97	50.70	75.30
VidAssist (zero-shot) [36]	14.60	52.60	68.38	9.89	40.85	70.35	18.44	50.63	75.64	9.07	42.72	80.83
VidAssist [36]	28.85	58.12	75.36	15.45	51.51	72.61	29.20	54.76	78.02	20.78	49.07	78.93
SCHEMA [52]	38.93	63.80	79.82	24.50	58.48	76.48	32.09	49.84	83.83	22.02	45.33	83.47
Generative (World) Models												
DDN [15]	12.18	31.29	47.48	5.97	27.10	48.46	13.90	20.19	64.78	11.13	17.71	68.06
Int-MGAIL [7]	17.03	44.66	58.08	9.47	37.16	57.24	–	–	–	–	–	–
Ext-MGAIL [7]	21.27	49.46	61.70	16.41	43.05	60.93	–	–	–	–	–	–
P^3IV [86]	23.34	49.96	73.89	13.40	44.16	70.01	15.40	21.67	76.31	11.32	18.85	70.53
PDPP [75]	37.20	64.67	66.57	21.48	57.82	65.13	21.33	45.62	51.82	14.41	44.10	51.39
KEPP [50]	38.12	64.74	67.15	24.15	59.05	66.64	20.25	39.87	51.72	15.63	39.53	53.27
ActionDiffusion [61]	37.79	65.38	67.45	22.43	59.42	66.04	24.00	45.42	54.29	18.04	44.54	56.23
MTID [87]	40.45	67.19	69.17	24.76	60.69	67.67	30.44	51.70	59.74	22.74	49.90	61.25
Predictive (World) Models												
WLTDO [27]	1.87	21.64	31.70	0.77	17.92	26.43	–	–	–	–	–	–
UAAA [1]	2.15	20.21	30.87	0.98	19.86	27.09	–	–	–	–	–	–
UPN [68]	2.89	24.39	31.56	1.19	21.59	27.85	–	–	–	–	–	–
PlaTe [69]	16.00	36.17	65.91	14.00	35.29	55.36	–	–	–	–	–	–
E3P [74]	26.40	53.02	74.05	16.49	48.00	70.16	19.57	31.42	84.95	13.59	26.72	84.72
V-JEPA 2 ViT-L [3]	43.33	68.63	67.84	27.53	63.80	65.45	32.10	54.25	61.18	20.86	52.61	64.33
V-JEPA 2 ViT-H [3]	44.07	70.18	68.32	28.75	64.71	66.82	32.76	55.37	61.57	22.60	53.19	65.74
V-JEPA 2 ViT-g [3]	44.84	71.62	68.87	30.03	65.04	67.93	33.42	56.29	63.31	23.04	54.47	66.13
V-JEPA 2 ViT-g_{384} [3]	45.58	72.74	69.42	31.36	65.45	69.21	34.08	57.20	64.53	23.43	55.58	66.57
GeoWorld ViT-L (Ours)	43.89	68.96	82.93	27.64	64.35	79.43	33.42	57.26	88.03	24.96	52.92	85.26
GeoWorld ViT-H (Ours)	45.33	70.84	84.70	29.19	65.47	80.16	34.08	58.70	88.42	26.24	53.66	87.17
GeoWorld ViT-g (Ours)	46.25	71.95	85.44	30.63	66.02	81.82	34.41	60.47	89.00	27.46	54.55	88.20
GeoWorld ViT-g_{384} (Ours)	**47.47**	**73.69**	**86.55**	**31.48**	**67.30**	**82.48**	**34.85**	**61.86**	**89.88**	**27.79**	**55.97**	**88.61**

GeoWorld are trained with the AdamW optimizer [45] using a warmup–constant–decay learning rate schedule and a constant weight decay of 0.04. We linearly warm up the learning rate from 7.5×10^{-5} to 4.25×10^{-4} over 4500 iterations, hold it constant for 85,500 iterations, and then decay it to 0 over the final 4500 iterations, with a batch size of 256.

For geometric reinforcement learning, we keep the same AdamW optimizer and weight decay as in the supervised post-training stage, but adopt a smaller learning rate and a shorter schedule due to the higher variance of the RL objective. Specifically, we linearly warm up the learning rate from 5.0×10^{-5} to 2.0×10^{-4} over 2,000 iterations, hold it constant for 18,000 iterations, and then linearly decay it to 0 over the final 5,000 iterations, with a batch size of 128. Unless otherwise specified, we set the discount factor to $\gamma = 0.99$ and the triangle-inequality regularization weight to $\beta = 0.1$.

For energy-based planning with CEM [23], we adopt a sample size of $N = 800$, an elite set size of $K = 80$, and $I = 10$ refinement iterations.

The entire training is conducted on 4 nodes, each equipped with 8 NVIDIA H100 GPUs, 48-core Intel Xeon Platinum 8469C CPUs, and 230 GB of RAM. We use only a single H100 GPU for inference.

4.4. Main Results

As shown in Table 1 and 2, GeoWorld consistently improves multi-step goal-conditioned visual planning across both CrossTask and COIN. Under the procedural planning setup, GeoWorld yields notable gains over prior predictive world models, especially in long-horizon settings, achieving higher SR, mAcc, and mIoU for both T=3 and T=4. In the video-based planning setup, GeoWorld continues to outperform V-JEPA 2 across all model scales, with the ViT-g_{384} variant achieving the best overall results and surpass-

Table 2. **Goal-conditioned visual planning with videos on CrossTask [88] and COIN [71] datasets.** We evaluate multi-step planning over a horizon T under the *visual planning with videos* [59] setup, where both observations and goals are specified as video clips.

Method	CrossTask Dataset [88]						COIN Dataset [71]					
	T=3			T=4			T=3			T=4		
	SR	mAcc	mIoU	SR	mAcc	mIoU	SR	mAcc	mIoU	SR	mAcc	mIoU
LLM-Based												
InternVL3.5-241B [76]	44.03	70.01	84.41	27.65	63.54	80.13	36.54	57.22	89.02	25.46	55.30	88.22
Qwen3-VL-Max [79]	45.47	70.93	86.18	28.76	62.91	81.51	37.56	57.80	90.46	26.17	57.13	87.56
Gemini 2.5 Pro [20]	48.91	73.82	90.30	31.53	60.58	84.56	42.07	61.02	92.94	30.20	60.13	84.82
GPT-5 [53]	50.03	72.38	91.18	30.20	64.48	82.15	43.84	64.67	91.12	32.64	56.84	86.38
Generative (World) Models												
VideoWorld [59]	41.59	66.11	82.64	25.50	60.26	76.85	34.88	54.71	85.58	23.74	51.27	85.33
Predictive (World) Models												
V-JEPA 2 ViT-L [3]	43.36	69.55	84.75	28.86	64.34	78.40	36.10	56.70	87.02	25.29	53.30	87.21
V-JEPA 2 ViT-H [3]	46.02	71.98	87.29	32.18	67.23	80.84	39.42	59.42	89.44	27.38	56.07	90.20
V-JEPA 2 ViT-g [3]	48.13	73.42	89.62	33.46	69.26	82.61	40.97	61.86	90.77	29.60	57.73	92.23
V-JEPA 2 ViT-g_{384} [3]	50.16	74.86	91.73	35.01	70.24	85.05	42.74	64.08	91.88	31.63	59.28	94.51
GeoWorld ViT-L (Ours)	44.80	70.54	86.30	30.63	65.46	79.73	37.76	58.14	88.00	26.40	54.52	88.98
GeoWorld ViT-H (Ours)	47.79	74.42	88.84	34.51	68.89	82.95	40.40	60.97	91.66	28.82	58.10	91.48
GeoWorld ViT-g (Ours)	49.23	76.64	90.61	35.49	71.00	84.50	42.84	62.63	93.69	29.93	60.65	92.81
GeoWorld ViT-g_{384} (Ours)	**51.71**	**77.30**	**92.95**	**37.04**	**71.35**	**87.04**	**45.29**	**65.52**	**93.91**	**33.29**	**61.56**	**95.84**

Table 3. **Long horzion planning** on CrossTask [88].

Method	Successful Rate (SR, %)			
	T=3	T=4	T=5	T=6
Procedural Planning (PP)				
Random [15]	0.01	0.01	0.01	0.01
Retrieval-Based [87]	8.05	3.95	2.40	1.10
DDN [15]	12.18	5.97	3.10	1.20
P^3IV [86]	23.34	13.40	7.21	4.40
E3P [74]	26.40	16.49	8.96	5.76
PDPP [75]	37.20	21.48	13.45	8.41
KEPP [50]	38.12	24.15	14.20	9.27
SCHEMA [52]	38.93	24.50	14.75	10.53
MTID [87]	40.45	24.76	15.26	10.30
V-JEPA 2 ViT-L [3]	43.33	27.53	16.94	11.55
GeoWorld ViT-L (Ours)	**43.89**	**27.64**	**17.38**	**12.37**
Visual Planning with Videos				
VideoWorld [59]	41.59	25.50	15.36	10.97
InternVL3.5-241B [76]	44.03	27.65	17.31	12.44
Qwen3-VL-Max [79]	45.47	28.76	17.95	13.20
Gemini 2.5 Pro [20]	48.91	31.53	20.08	15.93
GPT-5 [53]	50.03	30.20	21.46	16.07
V-JEPA 2 ViT-g_{384} [3]	50.16	35.01	23.17	16.88
GeoWorld ViT-g_{384} (Ours)	**51.71**	**37.04**	**24.83**	**18.26**

ing strong LLM-based planners. These improvements highlight the effectiveness of geometry-aware latent dynamics and geometric reinforcement learning in enhancing long-horizon stability and planning accuracy.

For ablation study, please refer to *Appendix* 5.

4.5. Long-Horizon Planning

Table 3 highlights GeoWorld's strength in long-horizon planning. As the horizon increases from $T = 3$ to $T = 6$, the performance of existing predictive and generative world models consistently degrades due to accumulated geometric drift in Euclidean latent space. In contrast, GeoWorld maintains higher stability and achieves the best Success Rate across all horizons.

5. Conclusion

We introduced **GeoWorld**, a geometric world model designed to improve long-horizon visual planning by preserving geometric structure and hierarchical relations in latent space. Through *Hyperbolic JEPA*, GeoWorld maps Euclidean latent representations onto a hyperbolic manifold, enabling geodesic-aware latent dynamics that produce a more structured and physically meaningful energy landscape. Building on this representation, *Geometric Reinforcement Learning* refines the predictor via hyperbolic energy optimization and triangle-inequality regularization, yielding geodesic-consistent rollouts and reducing error accumulation across extended horizons. Extensive experiments on CrossTask and COIN demonstrate that GeoWorld consistently improves long-horizon performance over strong predictive world models such as V-JEPA 2, achieving higher success rates across $T = 3$ to $T = 6$ planning. These results highlight the importance of incorporating geometric principles into predictive world models and reinforce the value of geometry-aware reinforcement learning for stable and effective multi-step planning.

References

[1] Yazan Abu Farha and Juergen Gall. Uncertainty-aware anticipation of activities. In *Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops*, pages 0–0, 2019. 6, 7, 10

[2] Mahmoud Assran, Quentin Duval, Ishan Misra, Piotr Bojanowski, Pascal Vincent, Michael Rabbat, Yann LeCun, and Nicolas Ballas. Self-supervised learning from images with a joint-embedding predictive architecture. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15619–15629, 2023. 2, 3, 1, 4

[3] Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025. 1, 2, 3, 6, 7, 8, 4, 10, 12, 13

[4] Adrien Bardes, Jean Ponce, and Yann Lecun. Vicreg: Variance-invariance-covariance regularization for self-supervised learning. In *ICLR 2022-International Conference on Learning Representations*, 2022. 2

[5] Adrien Bardes, Quentin Garrido, Jean Ponce, Xinlei Chen, Michael Rabbat, Yann LeCun, Mido Assran, and Nicolas Ballas. Revisiting feature prediction for learning visual representations from video. *Transactions on Machine Learning Research*, 2024. 2, 3, 1, 4

[6] Richard Bellman. Dynamic programming. *science*, 153 (3731):34–37, 1966. 8

[7] Jing Bi, Jiebo Luo, and Chenliang Xu. Procedure planning in instructional videos via contextual modeling and model-based policy learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15611–15620, 2021. 3, 6, 7, 9

[8] Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, Lachy Groom, Karol Hausman, Brian Ichter, et al. π_0: A Vision–Language–Action Flow Model for General Robot Control. *arXiv preprint arXiv:2410.24164*, 2024. 1

[9] Rinu Boney, Juho Kannala, and Alexander Ilin. Regularizing model-based planning with energy-based models. In *Conference on Robot Learning*, pages 182–191. PMLR, 2020. 1

[10] David Brandfonbrener, Ofir Nachum, and Joan Bruna. Inverse dynamics pretraining learns good representations for multitask imitation. *Advances in Neural Information Processing Systems*, 36:66953–66978, 2023. 4

[11] Martin R Bridson and André Haefliger. *Metric spaces of non-positive curvature*. Springer Science & Business Media, 2013. 5

[12] Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the ieee conference on computer vision and pattern recognition*, pages 961–970, 2015. 2

[13] Edoardo Cetin, Benjamin Paul Chamberlain, Michael M Bronstein, and Jonathan J Hunt. Hyperbolic deep reinforcement learning. In *The Eleventh International Conference on Learning Representations*, 2023. 5, 7

[14] Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019. 6, 11

[15] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer, 2020. 3, 6, 7, 8, 9

[16] Delong Chen, Theo Moutakanni, Willy Chung, Yejin Bang, Ziwei Ji, Allen Bolourchi, and Pascale Fung. Planning with reasoning using vision language world model. *arXiv preprint arXiv:2509.02722*, 2025. 3

[17] Guibin Chen, Dixuan Lin, Jiangping Yang, Chunze Lin, Junchen Zhu, Mingyuan Fan, Hao Zhang, Sheng Chen, Zheng Chen, Chengcheng Ma, et al. Skyreels-v2: Infinite-length film generative model. *arXiv preprint arXiv:2504.13074*, 2025. 3

[18] Dorothy L Cheney and Robert M Seyfarth. Why animals don't have language. *Tanner lectures on human values*, 19: 173–210, 1998. 2

[19] Krzysztof Chris Ciesielski, Alexandre Xavier Falcão, and Paulo AV Miranda. Path-value functions for which dijkstra's algorithm returns optimal mapping. *Journal of Mathematical Imaging and Vision*, 60(7):1025–1036, 2018. 8

[20] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 1, 3, 6, 8, 9

[21] Balázs Csanád Csáji and László Monostori. Value function based reinforcement learning in changing markovian environments. *Journal of Machine Learning Research*, 9(8), 2008. 8

[22] Justin Cui, Jie Wu, Ming Li, Tao Yang, Xiaojie Li, Rui Wang, Andrew Bai, Yuanhao Ban, and Cho-Jui Hsieh. Self-forcing++: Towards minute-scale high-quality video generation. *arXiv preprint arXiv:2510.02283*, 2025. 3

[23] Pieter-Tjerk De Boer, Dirk P Kroese, Shie Mannor, and Reuven Y Rubinstein. A tutorial on the cross-entropy method. *Annals of operations research*, 134(1):19–67, 2005. 3, 6, 7, 4

[24] Haoge Deng, Ting Pan, Haiwen Diao, Zhengxiong Luo, Yufeng Cui, Huchuan Lu, Shiguang Shan, Yonggang Qi, and Xinlong Wang. Autoregressive video generation without vector quantization. *arXiv preprint arXiv:2412.14169*, 2024. 3

[25] Karan Desai, Maximilian Nickel, Tanmay Rajpurohit, Justin Johnson, and Shanmukha Ramakrishna Vedantam. Hyperbolic image-text representations. In *International Conference on Machine Learning*, pages 7694–7731. PMLR, 2023. 5, 11

[26] Yilun Du, Shuang Li, Joshua Tenenbaum, and Igor Mordatch. Learning iterative reasoning through energy minimization. In *International Conference on Machine Learning*, pages 5570–5582. PMLR, 2022. 1

[27] Kiana Ehsani, Hessam Bagherinezhad, Joseph Redmon, Roozbeh Mottaghi, and Ali Farhadi. Who let the dogs out?

modeling dog behavior from visual data. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4051–4060, 2018. 6, 7, 10

[28] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018. 5, 7

[29] Quentin Garrido, Mahmoud Assran, Nicolas Ballas, Adrien Bardes, Laurent Najman, and Yann LeCun. Learning and leveraging world models in visual representation learning. *arXiv preprint arXiv:2403.00504*, 2024. 2, 3, 4

[30] Songwei Ge, Shlok Mishra, Simon Kornblith, Chun-Liang Li, and David Jacobs. Hyperbolic contrastive learning for visual representations beyond objects. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6840–6849, 2023. 5

[31] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzynska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, et al. The" something something" video database for learning and evaluating visual common sense. In *Proceedings of the IEEE international conference on computer vision*, pages 5842–5850, 2017. 2

[32] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 1

[33] Ting Huang, Zeyu Zhang, and Hao Tang. 3d-r1: Enhancing reasoning in 3d vlms for unified scene understanding. *arXiv preprint arXiv:2507.23478*, 2025. 1

[34] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 3

[35] Physical Intelligence, Kevin Black, Noah Brown, James Darpinian, Karan Dhabalia, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, Niccolo Fusai, et al. $\pi_{0.5}$: A Vision–Language–Action Model with Open-World Generalization. *arXiv preprint arXiv:2504.16054*, 2025. 1

[36] Md Mohaiminul Islam, Tushar Nagarajan, Huiyu Wang, Fu-Jen Chu, Kris Kitani, Gedas Bertasius, and Xitong Yang. Propose, assess, search: Harnessing llms for goal-oriented planning in instructional videos. In *European Conference on Computer Vision*, pages 436–452. Springer, 2024. 3, 6, 7, 9

[37] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 2

[38] Philip J Kellman, ME Arterberry, W Damon, RM Lerner, D Kuhn, RS Siegler, et al. Infant visual perception. 2006. 1

[39] Alexander Khazatsky, Karl Pertsch, Suraj Nair, Ashwin Balakrishna, Sudeep Dasari, Siddharth Karamcheti, Soroush Nasiriany, Mohan Kumar Srirama, Lawrence Yunliang Chen, Kirsty Ellis, et al. Droid: A large-scale in-the-wild robot manipulation dataset. In *RSS 2024 Workshop: Data Generation for Robotics*. 3

[40] Hildegard Kuehne, Hueihan Jhuang, Estíbaliz Garrote, Tomaso Poggio, and Thomas Serre. Hmdb: a large video database for human motion recognition. In *2011 International conference on computer vision*, pages 2556–2563. IEEE, 2011. 2

[41] Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022. 2, 3, 1, 4, 5, 13

[42] Jiateng Liu, Sha Li, Zhenhailong Wang, Manling Li, and Heng Ji. A language-first approach for procedure planning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1941–1954, 2023. 3, 6, 7, 9

[43] Jinlai Liu, Jian Han, Bin Yan, Hui Wu, Fengda Zhu, Xing Wang, Yi Jiang, Bingyue Peng, and Zehuan Yuan. Infinitystar: Unified spacetime autoregressive modeling for visual generation. *arXiv preprint arXiv:2511.04675*, 2025. 3

[44] Qingxiang Liu, Ting Huang, Zeyu Zhang, and Hao Tang. Nav-r1: Reasoning and navigation in embodied scenes. *arXiv preprint arXiv:2509.10884*, 2025. 1

[45] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 7

[46] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos. 2023. 2, 3, 4

[47] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 2

[48] Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023. 1

[49] Sho Mitsuhashi and Shin Ishii. Triangle inequality for inverse optimal control. *IEEE Access*, 11:119187–119199, 2023. 8

[50] Kumaranage Ravindu Yasas Nagasinghe, Honglu Zhou, Malitha Gunawardhana, Martin Renqiang Min, Daniel Harari, and Muhammad Haris Khan. Why not use your textbook? knowledge-enhanced procedure planning of instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18816–18826, 2024. 3, 6, 7, 8, 9

[51] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017. 2, 5

[52] Yulei Niu, Wenliang Guo, Long Chen, Xudong Lin, and Shih-Fu Chang. Schema: State changes matter for procedure planning in instructional videos. *arXiv preprint arXiv:2403.01599*, 2024. 3, 6, 7, 8, 9

[53] OpenAI. Gpt-5 system card, version 1.0, 2025-08-13. 2025. https://cdn.openai.com/gpt-5-system-card.pdf. 1, 3, 6, 8, 9

[54] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language

models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022. 2

[55] Avik Pal, Max van Spengler, Guido Maria D'Amely di Melendugno, Alessandro Flaborea, Fabio Galasso, and Pascal Mettes. Compositional entailment learning for hyperbolic vision-language models. *arXiv preprint arXiv:2410.06912*, 2024. 5

[56] Dhruvesh Patel, Hamid Eghbalzadeh, Nitin Kamra, Michael Louis Iuzzolino, Unnat Jain, and Ruta Desai. Pretrained language models as visual planners for human assistance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15302–15314, 2023. 3

[57] Silviu Pitis, Harris Chan, Kiarash Jamali, and Jimmy Ba. An inductive bias for distances: Neural nets that respect the triangle inequality. In *International Conference on Learning Representations*, 2020. 8

[58] Sucheng Ren, Chen Chen, Zhenbang Wang, Liangchen Song, Xiangxin Zhu, Alan Yuille, Yinfei Yang, and Jiasen Lu. Autoregressive video generation beyond next frames prediction. *arXiv preprint arXiv:2509.24081*, 2025. 3

[59] Zhongwei Ren, Yunchao Wei, Xun Guo, Yao Zhao, Bingyi Kang, Jiashi Feng, and Xiaojie Jin. Videoworld: Exploring knowledge learning from unlabeled videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29029–29039, 2025. 2, 3, 6, 8, 4, 9

[60] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 2

[61] Lei Shi, Paul Bürkner, and Andreas Bulling. Actiondiffusion: An action-aware diffusion model for procedure planning in instructional videos. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8816–8825. IEEE, 2025. 3, 6, 7, 9

[62] Mariano Sigman and Stanislas Dehaene. Brain mechanisms of serial and parallel processing during dual-task performance. *Journal of Neuroscience*, 28(30):7585–7598, 2008. 1

[63] Himanshu Gaurav Singh, Antonio Loquercio, Carmelo Sferrazza, Jane Wu, Haozhi Qi, Pieter Abbeel, and Jitendra Malik. Hand-object interaction pretraining from videos. In *2025 IEEE International Conference on Robotics and Automation (ICRA)*, pages 3352–3360. IEEE, 2025. 2, 3, 4

[64] Geri Skenderi, Hang Li, Jiliang Tang, and Marco Cristani. Graph-level representation learning with joint-embedding predictive architectures. *Transactions on Machine Learning Research*, 2025. 9

[65] Zirui Song, Guangxian Ouyang, Mingzhe Li, Yuheng Ji, Chenxi Wang, Zixiang Xu, Zeyu Zhang, Xiaoqing Zhang, Qian Jiang, Zhenhao Chen, et al. Maniplvm-r1: Reinforcement learning for reasoning in embodied manipulation with large vision-language models. *arXiv preprint arXiv:2505.16517*, 2025. 1

[66] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012. 2

[67] Mark W Spong and Romeo Ortega. On adaptive inverse dynamics control of rigid robots. *IEEE Transactions on Automatic Control*, 35(1):92–95, 2002. 2, 3, 4

[68] Aravind Srinivas, Allan Jabri, Pieter Abbeel, Sergey Levine, and Chelsea Finn. Universal planning networks: Learning generalizable representations for visuomotor control. In *International conference on machine learning*, pages 4732–4741. PMLR, 2018. 6, 7, 10

[69] Jiankai Sun, De-An Huang, Bo Lu, Yun-Hui Liu, Bolei Zhou, and Animesh Garg. Plate: Visually-grounded planning with transformers in procedural tasks. *IEEE Robotics and Automation Letters*, 7(2):4924–4930, 2022. 3, 6, 7, 10

[70] Richard S Sutton, Andrew G Barto, et al. *Reinforcement learning: An introduction*. MIT press Cambridge, 1998. 8

[71] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 1, 2, 6, 7, 8, 10, 13

[72] Hansi Teng, Hongyu Jia, Lei Sun, Lingzhi Li, Maolin Li, Mingqiu Tang, Shuai Han, Tianning Zhang, WQ Zhang, Weifeng Luo, et al. Magi-1: Autoregressive video generation at scale. *arXiv preprint arXiv:2505.13211*, 2025. 3

[73] Yang Tian, Sizhe Yang, Jia Zeng, Ping Wang, Dahua Lin, Hao Dong, and Jiangmiao Pang. Predictive inverse dynamics models are scalable learners for robotic manipulation. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[74] An-Lan Wang, Kun-Yu Lin, Jia-Run Du, Jingke Meng, and Wei-Shi Zheng. Event-guided procedure planning from instructional videos with text supervision. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13565–13575, 2023. 3, 6, 7, 8, 10

[75] Hanlin Wang, Yilu Wu, Sheng Guo, and Limin Wang. Pdpp: Projected diffusion for procedure planning in instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14836–14845, 2023. 3, 6, 7, 8, 9

[76] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 3, 6, 8, 9

[77] Yuqing Wang, Tianwei Xiong, Daquan Zhou, Zhijie Lin, Yang Zhao, Bingyi Kang, Jiashi Feng, and Xihui Liu. Loong: Generating minute-level long videos with autoregressive language models. *arXiv preprint arXiv:2410.02757*, 2024. 3

[78] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022. 1

[79] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 1, 3, 6, 8, 9

[80] Chao Yang, Xiaojian Ma, Wenbing Huang, Fuchun Sun, Huaping Liu, Junzhou Huang, and Chuang Gan. Imitation learning from observations by minimizing inverse dynamics disagreement. *Advances in neural information processing systems*, 32, 2019. 4

[81] Shuai Yang, Wei Huang, Ruihang Chu, Yicheng Xiao, Yuyang Zhao, Xianbang Wang, Muyang Li, Enze Xie, Yingcong Chen, Yao Lu, et al. Longlive: Real-time interactive long video generation. *arXiv preprint arXiv:2509.22622*, 2025. 3

[82] Angen Ye, Zeyu Zhang, Boyuan Wang, Xiaofeng Wang, Dapeng Zhang, and Zheng Zhu. Vla-r1: Enhancing reasoning in vision-language-action models. *arXiv preprint arXiv:2510.01623*, 2025. 1

[83] Yun Yue, Fangzhou Lin, Kazunori D Yamada, and Ziming Zhang. Hyperbolic contrastive learning. *arXiv preprint arXiv:2302.01409*, 2023. 5

[84] Ce Zhang, Yale Song, Ruta Desai, Michael Louis Iuzzolino, Joseph Tighe, Gedas Bertasius, and Satwik Kottur. Enhancing visual planning with auxiliary tasks and multi-token prediction. *arXiv preprint arXiv:2507.15130*, 2025. 3

[85] Zeyu Zhang, Yiran Wang, Danning Li, Dong Gong, Ian Reid, and Richard Hartley. Flashmo: Geometric interpolants and frequency-aware sparsity for scalable efficient motion generation. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. 5

[86] He Zhao, Isma Hadji, Nikita Dvornik, Konstantinos G Derpanis, Richard P Wildes, and Allan D Jepson. P3iv: Probabilistic procedure planning from instructional videos with weak supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2938–2948, 2022. 3, 6, 7, 8, 9

[87] Yufan Zhou, Zhaobo Qi, Lingshuai Lin, Junqi Jing, Tingting Chai, Beichen Zhang, Shuhui Wang, and Weigang Zhang. Masked temporal interpolation diffusion for procedure planning in instructional videos. *arXiv preprint arXiv:2507.03393*, 2025. 3, 6, 7, 8, 9, 13

[88] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 2, 6, 7, 8, 11, 12, 13