GS-World: An Engine-driven Learning Paradigm for Pursuing Embodied Intelligence using World Models of Generative Simulation

Guiliang Liu, Yueci Deng, Zhen Liu, Kui Jia*
School of Data Science, The Chinese University of Hong Kong, Shenzhen
liuguiliang@cuhk.edu.cn, yuecideng@link.cuhk.edu.cn,
{zhenliu,kuijia}@cuhk.edu.cn

Abstract

As a pivotal direction driving artificial intelligence into the physical world, embodied intelligence is drawing great research attention across academia and industries; yet the scaling law driving the success of many large AI models in the past decade has not been observed in pursuing embodied intelligence, partly due to the scarcity of multi-modal, heterogeneous, and physics-related data required for learning it. In this perspective paper, we analyze the reasons behind and derive an *efficiency law* that is more demanded in this context. To meet the law, we first propose *world models of generative simulation (GS-World)* that are expected to model and predict the world dynamics in a perfectly physics-accurate manner, by generative learning of physics simulation, including the 3D assets, the environments, and the physical rules governing their dynamic interactions. Based on such a GS-World engine, we propose an efficient, engine-driven learning paradigm for pursuing embodied intelligence, which is characterized by an automated pipeline of data generation and streaming, training of vision-language-action models, model verification, and model deployment, termed as *Engine-driven, Sim2Real VLA*. Task-oriented embodiments can also be optimized backwardly, given the differentiable nature of the learning pipeline. We will showcase the paradigm by releasing a prototype engine of GS-World and automatically trained Sim2Real VLAs. We call for the collective community contributions to this promising umbrella of research fields.

1 Introduction

To advance the robotic learning research, in this perspective paper, we first present an *efficiency law*, which is an empirical power law derived from the scaling law [1, 2]; it characterizes how model performance is related to the efficiency of data generation. Give potential ways to improve the data generation efficiency, we favor those based on generative world models, and propose *world models of generative simulation (GS-World)* (cf. an illustration in Fig.2) that learn to generate all the internal controlling ingredients required for modeling and predicting action dynamics in a physics-accurate manner, including the 3D assets, the environment, and the physical rules governing their dynamic interactions; GS-World generates a simulation world strictly adhering to physical laws, where robotic agents can act, interact with the environment, and themselves evolve for task-optimal embodiments. We give potential technical approaches for learning a GS-World.

To meet the efficiency law, we also propose an efficient, engine-driven learning paradigm for pursuing embodied intelligence, based on the proposed engine of GS-World. The paradigm is characterized by an automated pipeline of data generation and streaming, training of vision-language-action (VLA) model, model verification, and model deployment, which we term as *Engine-driven, Sim2Real VLA*

*Corresponding author: Kui Jia, kuijia@cuhk.edu.cn.

(cf. Fig.1 for an illustration); task-oriented embodiments can also be optimized backwardly, given the differentiable nature of the learning pipeline. GS-World can optionally absorb knowledge and priors from real-world observations. By playing in such a generative simulation world, robotic skills can be discovered and learned in either analytic or generative manners, where a rich set of scene-, object-, and affordance-level attributes can be automatically computed. This gives the paradigm an opportunity to learn attribute-enhanced VLAs; our Sim2Real VLA models favor the neural design of a high-level, VLM-based planner and a low-level actor of policy model, which are bridged by the automatically computed mid-level attributes. A suite of online and continual learning algorithms connect these pipeline modules and learn embodied intelligence with data streamed on the fly. In summary, our proposed paradigm has nice properties, including 1) automation and efficiency, 2) scalability and elasticity, 3) physics accuracy, 4) robustness to disturbance, 5) support of various embodiments via hardware calibration, 6) sim2Real transferability, and 7) evolution of embodiments; details are given in Section 2.1.

We will showcase the paradigm by releasing a prototype engine of GS-World and automatically trained Sim2Real VLAs. The remainder of this paper goes as follows. Section 2 presents an efficiency law which motivates us to propose the paradigm of Engine-driven Sim2Real VLA in Section 2.1. Details of GS-World is presented in Section 3, and in Section 4, we present how robotic skills can be discovered in GS-World. Section 5 and Section 6 give the architecture of Sim2Real VLAs and the framework for designing the robot's morphological structure.

2 An *Efficiency Law* for Pursuing Embodied Intelligence and A Proposed Learning Paradigm

In the past decade, the fascinating progress of generative models [3, 4, 5] is largely driven by increasing the sizes of neural models and the corresponding budget of compute for training. Empirical analyses on progress of LLMs also establish the *scaling law* [1, 2], which is an empirical power law characterizing the relations between model performance and the size of model, the size of dataset, and the amount of compute used for training; given a fixed compute budget, the scaling law also suggests optimal allocations among these factors, e.g., to allocate the compute budget more on training larger models on modest amounts of data with early stops, which are less influenced by other training practices (e.g., choices of architectures and training algorithms).

These insights are obtained under the regime of unlimited training data. However, subsequent research identifies violation of the neural scaling law when either the training data are insufficient or they are of low quality [6, 7]. This matters for learning generative models beyond LLMs. For learning embodied intelligence, especially, it requires multi-modal, heterogenous data including various sensory inputs (e.g., visual, force-torque, and/or tactile ones), proprioceptive states of robots, in addition to the language commands. Data collection, cleaning, and calibration for training robotic models (e.g., vision-language-action models [8]) are much more time-consuming and costly than training LLMs, since text data are massively available on the Internet and new data are daily added by Internet users, even though analysis also projects that publicly available, human-generated text data would be consumed to an end very soon [9].

How efficiently data are generated indeed matters for learning generative models. To better characterize its importance, we introduce the notion of *data-generation rate*, denoted as r_D, which precisely measures how many tokens are generated per time unit — for simplicity, we here do not differentiate among difference choices of tokenizers and tokens for data of different modalities. In the context of scaling law [1, 2], the following empirical law can be derived for an allowed amount of time t_D

$$l(r_D) = (c_D \cdot r_D \cdot t_D)^{-\alpha_D}, \tag{1}$$

where α_D and c_D are model- and task-dependent constants that establish a power law w.r.t. the loss l. We term equation (1) as *efficiency law*, which states that the model performance, measured by l, improves with a higher data-generation rate r_D, given a fixed amount of allowed time. We note that (1) is empirically established only when $r_D > r_D^{\min}$, i.e., when enough data are generated for training models. When considering data generation together with the model size, measured by its number P of parameters, we can derive the following empirical relation

$$l(r_D, P) = \left((c_P \cdot P)^{-\alpha_P/\alpha_D} + (c_D \cdot r_D \cdot t_D)^{-1}\right)^{\alpha_D}, \tag{2}$$

where α_P and c_P are also model- and task-dependent constants. One may refer to [1, 2] for how the original neural scaling laws are established.

The efficiency law suggests that r_D plays a decisive role in learning better models. Given a fixed amount of time allowed to generate the data, a lower r_D would make the learning under the data-scarcity regime, and correspondingly, advanced techniques, such as mode reusing [6] and data pruning [10], should be used in order for the model performance to obey the scaling law. Conversely, when higher enough r_D is possible, data amounts would not be an issue, and (2) will have a degenerate version that tells a power relation of $l(P) = (c_P \cdot P)^{-\alpha_P}$, i.e., larger models are able to be used to have better performance.

Given the multi-modal, heterogenous data nature of robotic learning, possible ways to increase r_D include 1) investment on building up data factories where robotic data can be obtained by tele-operation, 2) invention of new business models that can amortize the costly data collection with cheap and colectively efficient crowd-sourcing manners, 3) relying on generative models including world models, and 4) mixing of the above manners. While other manners are useful, in this paper, we favor the third manner and propose a learning paradigm, with an automated data generation and model training pipeline, powered by a proposed world model of generative simulation (GS-World).

2.1 The Paradigm of Engine-driven Sim2Real VLA using GS-World

Challenges indicated by the efficiency law suggest that, if we aim for continuously advancing the levels of machine intelligence [11] towards AGI, we must resolve the issue of data scarcity *in a more efficient manner*. Fortunately, the blessing from many practical applications of high value is that what we really need is those agents that are generally knowledgable while being specialized in certain fields (i.e., specialized generalists [12]). When such specialized generalists can be developed across various application field, it would probably pave a more efficient way towards achieving AGI. While this philosophy applies to large foundation models of various kinds, it is in particular illuminating for learning those of embodied intelligence. In fact, different from generative foundation models of LLMs and multimodal LLMs that are usually deployed on clouds serving for more general purposes, embodied AI models (e.g., VLAs) would be deployed at end sites for certain application scenarios requiring bounded generalization, and be better calibrated to hardware configurations of the embodiments for lower costs.

In this perspective paper, we propose a new paradigm of learning embodied intelligence that addresses the curse brought by efficiency law, termed *Engine-driven Sim2Real VLA*. Fig. 1 gives the illustration. The proposed paradigm is anchored on a novel engine termed *Generative Simulation of World Models (GS-World)*, which pursues generative learning of physics simulation that simulates a world complying with physical laws, where robotic agents can act, interact with the environment, and themselves evolve for task-optimal embodiments; details of GS-World are given in Section 3.

The forward pipeline of engine-driven sim2real VLA goes by streaming into GS-World few seeding data of heterogeneous, real-world priors, including those related to semantics, scene structures, object articulations, physical properties, action dynamics, and/or to hardware calibrations, if necessary. GS-World itself is a trained foundation model that, given a natural language description (optionally the aforementioned real-world priors) of a robotic task, generates all the ingredients necessary for learning embodied intelligence in a perfectly physics-accurate manner, including 3D assets of various forms (e.g., rigid, non-rigid, soft bodies, and/or fluids), and their task-plausible layout, dynamics, and physical properties; Fig. 2 illustrates different implementations of GS-World. Robotic skills are autonomously discovered by enabling the robot to act within and interact with the generated environment. This process is structured through LLM–based subtask decomposition and the automated design of objectives and reward functions, which, in turn, drive skill acquisition via reinforcement learning or motion-planning algorithms [13, 14, 15]. The discovered robotic skills are represented as trajectory data to train/fine-tune a VLA model. Our VLA architecture stacks a low-level actor on top of a high-level planner, and is featured by the bridging intermediate representations of scene-, object-, and affordance-level attributes — these attributes can be automatically computed given the core engine of GS-World. The trained VLA policy is verified in the same engine before deploying it into the real-world scenario. Note that the VLA model may also be pre-trained by leveraging online, action-relevant videos or those generated by *2D video generation [4, 16]*. The whole forward pipeline goes automatically in an online, data-streaming fashion, which enables quite a few nice properties to be discussed shortly.

Given the differentiable nature of the GS-World engine, the backward pipeline can optimize, for any specified task, the VLA policy, the GS-World foundation, and even the robotic embodiment itself, by

back-propagating error signals from policy verification; details are given in Section 6. Our proposed paradigm of Engine-driven Sim2Real VLA has the following nice properties for pursuing embodied intelligence.

- **Automation and Efficiency.** The proposed paradigm is featured by its automated pipeline rolling out seed-feeding of real-world priors, generation of the simulation environment, robotic skill discovery, trajectory data streaming, and VLA policy learning and verification. Such an online, data-streaming fashion pushes through the bottleneck in the existing data-driven learning paradigm, where data collection itself is of low efficiency and costly, especially for the collection of robotic data via tele-operation . Given enough compute budget, the proposed paradgim can effectively addresses the issue brought by efficiency law and scale up the learning of embodied intelligence.
- **Scalability and Elasticity.** In the preceding paragraphs, we elaborate on how the proposed engine-driven learning paradigm works for a specified robotic task. Since the learning pipeline is automated and efficient, assuming enough capacity of the VLA model and use of effective learning algorithms, it is possible to scale up the learning for as many and diverse robotic tasks as demanded, until achieving a generalist robot policy. From an economical perspective of high-value applications, we might instead favor specialized generalists [12], and the proposed paradigm is the very one that supports elastic learning of robotic skills tailored for special applications.
- **Accuracy of Physics.** The proposed paradigm is anchored on GS-World, the engine of the world model that enforces physics accuracy by learning to generate the underlying controlling factors of the world, including those physical properties governing action dynamics and how states of the world would be changed by these actions. Even when some real-world priors optionally fed into GS-World is of less physics accuracy, the engine would generate a world of perfect physics accuracy that closely matches the priors. Robotic skills are subsequently discovered and verified in this physics-accurate world. This is in contrast to world models of video generation [4, 16] that can only be used either to generate data for pre-training of VLA models or to filter obviously wrong policies.
- **Robustness to Disturbance.** To learn a policy that is robust against environmental changes and disturbances, the policy must be learned in an environment that is subject to as many changes and disturbances as possible. GS-World generates state changes of the environment efficiently, which supports policy verification such that a less mature policy can be updated to become a more robust one. In contrast, robust policy learning in real-world environments is costly, less generalizable across tasks, and brittle to changes in environmental conditions.
- **X-embodiment via Hardware Calibration.** The promise to be deployed on multiple robots is one of the properties that characterize a generalist policy. Existing learning paradigms aim for this property by collecting tele-operated data from as many and diverse robots as possible [17]. We argue that this is both prohibitively costly and unnecessary, since new robots are continuously designed for emerging demands, and it is less possible to cover all the configuration spaces of robots. In this paper, we argue for what may appear to be a reverse approach: the precise calibration of the simulation engine, its generated data, and the trained policy to the specific robot and its associated sensors. The rationale behind this perspective is that, because GS-World enables automated and efficient data generation, the demand for relearning policies for each new hardware embodiment is substantially reduced.
- **Sim2Real Transferability.** The GS-World framework naturally supports robust Sim2Real transfer by representing both simulated and real-world environments in an affordance-driven latent space. Instead of aligning raw sensory streams or complex continuous trajectories, GS-World projects object dynamics and robot interactions into compact affordance attributes that are semantically consistent across domains. This representation mitigates the discrepancies caused by sensor noise, physical dynamics, and unmodeled environmental factors, thereby narrowing the Sim2Real gap. More importantly, the automated construction of these affordance labels in simulation (via privileged object-level annotations) provides abundant supervision to train affordance extractors, which can later be deployed to parse real-world observations without additional manual labeling. Consequently, GS-World enables reliable transfer of policies and skills from synthetic rollouts to real-world deployments, ensuring both sample efficiency in simulation and robustness in execution.
- **Evolution of Embodiments.** GS-World enables the evolutionary co-design of robotic morphology and control policy within a unified Sim2Real framework. Instead of fixing robot structures as static priors, the backward pipeline adaptively searches, evaluates, and refines embodiments in response to task requirements and environmental dynamics. By representing robotic morphology in

graph-based latent variables and integrating them into VLA learning, GS-World permits structure-aware policy optimization. Moreover, strategic exploration methods replace inefficient random search in morphology space, accelerating convergence toward practical robot configurations that balance dexterity, mobility, and robustness. Through this iterative evolution of bodies and behaviors, GS-World cultivates adaptable embodied agents whose physical structures grow hand-in-hand with their cognitive policies, improving transferability to real-world deployment.

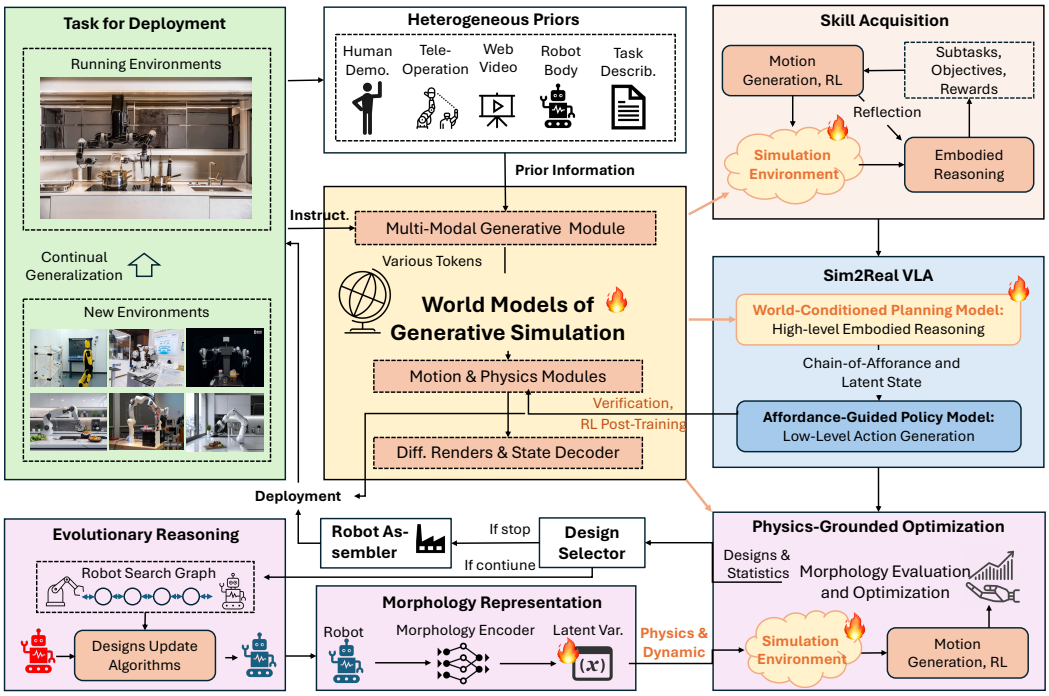

Figure 1: The proposed engine-driven learning paradigm for pursuing embodied intelligence, where the engine is based on the proposed world models of generative simulation (GS-World) (cf. Figure 2). The pipeline connects several key components: GS-world for constructing the learning environments (Section 3), embodied reasoning for skill acquisition (Section 4), a dual-system VLA model for Sim2Real transfer (Section 5), and dedicated modules for designing the robot's morphological structure (Section 6). In this paradigm, orange arrows indicate utilizing of the GS-world and dark arrows denote the working flow.

3 World Models of Generative Simulation

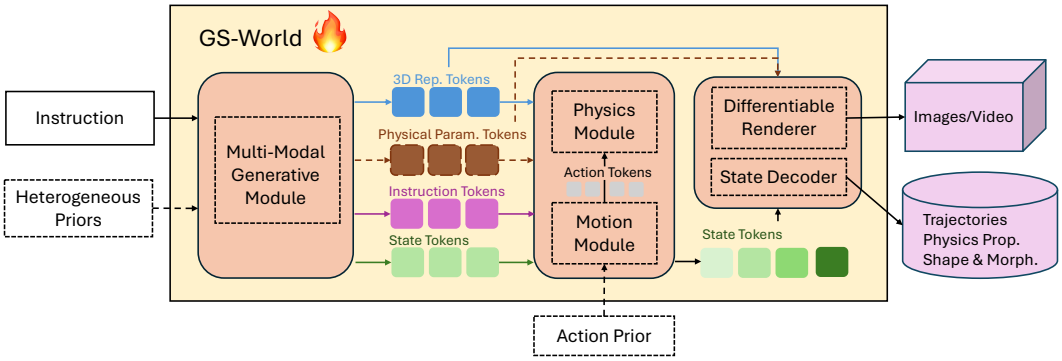

Figure 2: An illustration of world models of generative simulation (GS-World).

In the literature, there exists no precise definition of world models. They generally refer to those models capturing common sense knowledge of the world, enabling simulation of world dynamics by modeling its internal working mechanism, and being able to be used for the prediction of future world states. World models are explicitly stated in [18] as a neural, generative modelling approach emulating our humans' cognitive system; it learns latent representations of the world in an unsupervised manner, which are used to support policy learning of actors via reinforcement learning in such latent world representations. In [19], LeCun argues for a configurable world model towards learning autonomous machine intelligence, which can be trained to predict future states of the world, where training is based on a joint embedding predictive architecture (JEPA) framework via self-supervised learning.

Ingredients of world models are reflected in recent auto-regressive or diffusion-based generative foundation models such as LLMs [20, 21], multimodal LLMs [22], and video generation models [4, 16, 23]. For example, several LLMs demonstrate the presence of intuitive world knowledge (e.g., spatio-temporal understanding and reasoning, similar to human cognitive map [24]), which is useful for them to make predictions about physical conditions [20, 21]. Starting from the prominent work of Sora [4], video generation models advance rapidly to be able to generate long-horizon videos containing visual dynamics of various environments; to a large extent, these models are considered as world models since they can generate temporally consistent and visually plausible results, and some of them also capture causal effects and physical intuitions. Video generation models are subsequently extended to integrate other modalities [25, 23], support interactive generation [23, 26, 27], and be action-controllable [28, 29]; these capabilities can be obtained either in a training-free manner [27] or by collective training on foundational language and video generation models using diverse sets of videos with rich dynamics [23, 26, 28, 29]. In particular, the capability to generate videos of action consequences enables them to be used for dynamic decision-making and policy learning [30, 28]. For a specified goal of policy learning, sequential action sequences can be generated in a roll-out manner, which can be used either directly for extraction of actions [31] or as a reward signal to learn the policy. As such, world foundation models [16] are even pre-trained on massive amounts of action related videos, such as driving sequences, hand motion and manipulation, navigation, and natural dynamics, which can then be fine-tuned for various downstream application scenarios of policy learning, refinement, and policy verification.

As such, these video generation models show the promise as a scalable approach for learning embodied intelligence. While this might be useful in some cases for learning coarse actions such as driving behaviors and robot navigation, learning of many robotic tasks involving finer manipulations requires generation and simulation that should strictly adhere to the physical laws. Due to the in-distribution nature of the diffusion-based generative learning in the raw signal spaces (e.g., the RGB color space), video generation models might produce visually plausible results, such as shadow effects, surface reflection of light, although not precisely accurate, there are many other physical properties that are the causing factors of world dynamics but not directly visual observations, such as surface friction, mass distribution of different objects, articulation of non-rigid objects, deformation of soft bodies, fluid and thermal dynamics, among others. For these physical phenomena, these methods may produce counterfactual results violating physical rules. In fact, as analyzed in [32], even by restricting the analysis in the scope of classical mechanics laws in the 2D space, video generation models fail to adhere to these laws consistently; the analyses show that its behaviors of generalization are confined within the distribution of training data, and its generation of video dynamics relies more on retrieval from training dynamics than on governance by physical laws. Similar analysis is given in [33] by establishing the Physics-IQ benchmark for video generation methods.

These analyses indicate that statistical generative learning for the objective of visual realism does not translate directly as learning of correct physical principles. As such, quite a number of existing methods aim for improving physical correctness into the generation, by explicitly injecting physical constraints into the learning and generation process. For example, physical plausibility is improved in [34] by expanding a video generation model with depth and normal channels; integration of physics simulators with diffusion-based video generation is also used for simulation of rigid-body dynamics [35] and action-conditioned object dynamics [36]; generation of 3D object assets with compatible physical properties (e.g., absolute scale, material, affordance, and kinematics) is also presented in [37].

In this paper, we argue that many physical AGI tasks, including embodied intelligence, must be learned in strictly physics-aware manners, since, unlike LLM-based agents, physical agents are deployed in a real world governed by physical laws. Consequently, world models that support such

learning must evolve beyond producing merely visually plausible video sequences; instead, they must capture the underlying mechanisms of physical dynamics and enable their precise simulation, whether explicitly or implicitly. Such models should be capable of computing the internal states of simulated entities (e.g., acceleration changes under external forces and the resulting velocity and position obtained through time integration, all of which are observable and learnable by the agent), reasoning about changes in externally unobserved or occluded elements, maintaining long-term consistency across extended horizons, and supporting the saving and restoration of world states to facilitate accurate simulation and planning. Toward this goal, we categorize the proposed potential approaches as follows.

Explicit Enforcement of Physical Constraints. Perhaps the most straightforward manner is to enforce physical constraints into existing statistically learned generative models in a hard and explicit manner. For example, consistent video generation across multiple viewpoints can be induced by enforcing epipolar geometries [38]; objects, other entities, and their relational dynamics contained in videos can be generated by preserving their sharp boundaries and relations of occlusion, with alteration of only their appearance. A shortcoming of such a manner is that physically-correct properties are enforced in a one-by-one, ad-hoc manner. We note that existing generative learning methods perform far from satisfactory even for the very basic physics-accurate requirements of spatio-temporal consistency [39, 40].

Compatible Learning with Physics Simulators. A list of physics simulators and engines [41, 42, 43] exist that implement mathematical modeling and numerical simulation of physical processes, such as rigid-body dynamics with collision and joint constraints, elastic and plastic deformations of soft bodies, and interactive fluid dynamics capturing splashing effect. Generative world models that strictly follow physical rules can be learned relying on these simulators, for example, by learning to generate 3D assets and learning to predict hyper-parameters controlling physical simulations via differentiable physics engines, together with statistical learning for rich variations of visual appearance. By doing it this way, consistency of 3D shapes, motions adhering to space gravity, physics-correct scene configurations, external forces, and simulations of causal dynamics can be achieved automatically. More precisely, such a hybrid approach gives a world model that is internally governed by the simulator affiliated with the predicted hyper-parameters, which are differentiable and learnable, and is by nature interactively controllable with instructions of different formats, whose past, current, and future states of world dynamics can be precisely computed from the simulator as well; images or videos from arbitrary viewpoints are merely projections of these simulation dynamics via differentiable rendering (eg, 3D-GS). A few recent methods [36, 35, 37, 44] pursue this direction and more research is advocated for a truly physics-accurate world model.

Learning with a Unified Neural Representation. While off-the-shelf physics simulators guarantee strict adherence of generated world dynamics to the physical laws, the scope and flexibility of such generated dynamics are bounded by the laws specified by the simulators. To increase flexibility, taking simulation of rigid-body dynamics as the simplest example, neural differential equations (NDEs) [45] can be used to model motions, providing the opportunity to learn the laws of motion dynamics and making them adaptive to various real-world phenomena, while still benefiting from strict physical constraints. This approach can be extended to other physical simulations as well: for soft-body dynamics, NDEs can learn complex material constitutive behaviors or update operators in finite element analysis (FEM) or material point method (MPM) frameworks; for fluid dynamics, NDEs can model governing dynamics such as turbulence or viscous effects that are difficult to capture explicitly, while particle-based frameworks like smoothed-particle hydrodynamics (SPH) can be enhanced by neural operators that approximate local particle interactions; more generally, a single neural representation can encode and learn the governing dynamics of diverse physical systems, including rigid bodies, soft bodies, and fluids, capturing unknown forces, contact interactions, and simulation hyper-parameters in an end-to-end, differentiable manner. Taken together, this leads to a unified neural representation of 3D assets and their governing laws, since the 3D assets themselves can also be represented [46] and learned by neural generative models. In fact, differentiable equations specified in physics simulators often fail to precisely capture real-world dynamics, and the way to make it learnable has the additional benefit of closing the simulation gap. [2]

[2]We note that various neural simulation methods have been proposed in the literature, covering from rigid/non-rigid bodies [47], articulated objects [48], to soft bodies [49], and fluids [50]; these methods are typically tailored to specific types of simulation methods, without consideration in the context of generative world models.

Both the above second and third approaches (i.e., learning the simulator hyper-parameters or the laws themselves) can be implemented as learning from real-world observations (e.g., video observations) using self-supervised or reinforcement learning methods. Since the learned models would be universal ones capturing internal mechanisms of world dynamics, supporting physics-accurate computation of world states, and are based on generative simulation, we term them as *World Models of Generative Simulation (GS-World)*, Fig.2 gives the illustration. In this paper, we use GS-World as a generative engine in our proposed engine-driven learning pipeline. We note that such a GS-World also enables convenient learning of both VLMs and language models compatible with embodied tasks, similar to what have been done in [51].

4 Skill Acquisition via Embodied Reasoning with the World Model

The VLA model is designed to acquire robotic skills across diverse tasks and environments. In particular, the skill acquisition can be built upon the underlying *world models* of generative simulation (GS-World), which provides a physics-grounded generative simulation of the environment, enabling bodies and objects to interact in a causally consistent manner. Rather than relying solely on supervised demonstrations or task-specific rewards, the proposed pipeline employs *embodied reasoning* within the simulated world as guided by the world model to automate the process of discovering and refining skills. These learned skills constitute the fundamental building blocks for complex, long-horizon tasks, forming a closed loop between reasoning, action, and simulation.

In robotic practice, completing a goal-directed task usually requires executing a structured chain of sub-level objectives [52]. For instance, the task of brewing a cup of coffee may require: (1) recognizing and grasping a cup, (2) positioning it under the coffee machine, (3) pressing the brew button, and (4) serving the completed drink. To achieve robust learning over such long-horizon activities, tasks are decomposed into a series of *atomic subtasks*, which are then solved independently using physically realistic simulation under the guidance of the world model. The following paragraphs describe the main components of this pipeline.

Chain-of-Affordance Reasoning for Task Decomposition. In the task decomposition, complex actions are divided into atomic units by inferring the causal and physical dependencies between objects, agents, and the environment. To achieve this, we favor a physically grounded reasoning mechanism termed *Chain-of-Affordance (CoA)* reasoning [53], which extends the idea of modular Chain-of-Thought reasoning into an embodied context. Rather than operating purely over symbolic text, CoA reasoning leverages the internal state of the world model to infer a structured sequence of sub-goals connected through object- and action-level affordances.

The decomposition relies on three key categories of affordances represented in the simulated environment. 1) *Object affordances* identify manipulable entities in the world, describing their location, geometry, material, and articulation; under simulation, these affordances are precisely derived from the latent world state. 2) *Manipulation affordances* define allowable interactions, specifying the modes of contact or control available to the robot's embodiment (e.g., grasp, press, turn). These affordances are physically validated through the world model's generative dynamics, ensuring realistic and consistent results [54, 55]. 3) *Spatial affordances* describe the relational topology among objects, pathways, and manipulation zones in 3D space, reflecting navigable regions or feasible object placements. Figure 4 illustrates the examples of affordance in the task of water pouring.

Figure 3: Examples of (1) object affordance (left), (2) manipulation affordance (middle), and (3) spatial affordance (right).

Together, these affordances allow the world model to perform grounded Chain-of-Affordance reasoning, wherein task decomposition is both semantically interpretable and physically compliant.

Affordance attributes can be derived by (i) distilling them from large-scale foundation models [56], (ii) retrieving relevant information from affordance memory banks [57], or (iii) extracting and encoding them from demonstration trajectories [58]. Since the world model maintains full physical parameterization of the environment, this pipeline ensures that decomposed subtasks are dynamically feasible and consistent with real-world affordances.

Task-Level Reasoning for Oracle Design. Once tasks are decomposed into atomic components, the system must determine whether and when each subtask has been achieved. The evaluation of completion relies on an *oracle function*, which quantitatively measures progress based on internal states simulated by the world model. The oracle serves as a bridge between physical simulation and learning, providing reward signals or structured feedback for each atomic skill.

In the context of world-model-based skill acquisition, the oracle function $\mathcal{O}(s_t) : \mathbb{R}^n \to \mathbb{R}$ compares simulated states s_t to desired conditions derived from affordance constraints. When used for motion planning [59], such functions can explicitly encode physical metrics such as contact distances or joint configurations; when used in reinforcement learning [60], they function as differentiable black-box evaluators producing scalar reward feedback. To automate their creation, we adapt large language models as oracle generators [61], prompting them with structured task descriptions, affordance attributes, and environmental conditions present in the world model. The generated functions can be directly executed within the simulation to compute dynamic reward signals such as object proximity, stability, or adherence to target kinematics. As the world model provides access to intrinsic state variables (e.g., poses, velocities, forces), the oracles computed therein remain physically precise and differentiable, forming an essential component of the learning process.

Automatic Skill Learning for Atomic Tasks. Following the generation of subtasks and their oracle functions, automatic skill discovery is performed within the simulated world. A skill is defined as a learnable policy that maximizes the oracle-associated reward given a specific subtask configuration. Through generative simulation under the world model, policies are trained under physically accurate rollouts, ensuring both realism and control consistency across skills.

Two major categories of skills are learned: *Mobility Skills*, which concern with locomotion and spatial navigation, and *Manipulation Skills*, which concern with physical interaction and object handling. For mobility-oriented skills, the world model provides predictive affordance maps that indicate feasible navigation paths and spatial boundaries [14]. Wheeled agents employ path-planning solutions guided by oracle rewards, while legged or humanoid agents adopt reinforcement learning strategies to optimize cumulative returns $\mathbb{E}_\pi[\sum_t r_t]$, where r_t is obtained from simulation feedback. The resulting joint-angle trajectories are stabilized by controllers such as proportional–derivative (PD) schemes that ensure physically consistent torque control. For manipulation tasks, simpler operations such as grasping or pick-and-place are executed using inverse kinematics guided by the oracle-defined grasp points, whereas more complex object interactions leverage reinforcement learning in the world model's simulation engine. To align simulated movements with human-like preferences, imitation regularization is introduced through loss minimization on policy divergence, $\mathcal{D}_{KL}(\pi_\theta \| \pi_{\theta^*})$, relative to reference policies derived from demonstrations.

Search-Based Reasoning for Self-Reflection. In long-horizon or composite tasks, the success of each skill strongly depends on the correctness of preceding subtasks and their underlying affordance or oracle definitions. When the learning or motion generation of a subtask fails, embodied reasoning with the world model enables introspection and correction. This reflective ability is realized by modeling the process of skill acquisition as a structured search problem, in which the agent dynamically explores alternative task decompositions, oracle formulations, or control strategies when inconsistencies arise.

More specifically, the entire reasoning and learning process can be organized as a *Monte Carlo Tree Search (MCTS)* structure [62]. Each node in the tree represents a configuration regarding environmental setup, affordance composition, or oracle design. Trial rollouts, powered by the world model's simulation, are

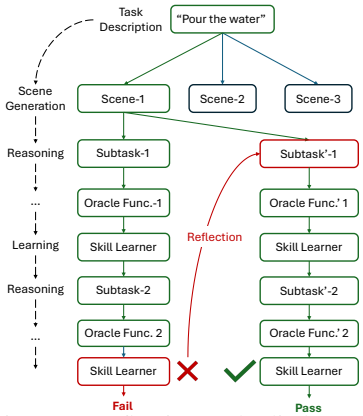

Figure 4: Reflective Embodied Reasoning through Tree Search.

performed from each node to estimate the long-term success
rate, which serves as the reward signal. The search process follows an Upper Confidence Bound (UCB) criterion [63] for balancing exploration and exploitation, enabling efficient identification of error sources or suboptimal design decisions. Once detected, problematic components are reconsidered and regenerated, either by adjusting affordances or redefining oracle logic, after which learning resumes from the refined configuration. This *reflective embodied reasoning* process results in self-improving skill structures that progressively enhance scalability and reliability over successive learning cycles. Figure 4 illustrates an example of reflective embodied reasoning.

Taken together, the proposed skill acquisition pipeline, grounded in the generative simulation of the world model, forms an integrated system for scalable embodied intelligence. Through chain-of-affordance reasoning, physically faithful oracle design, reinforcement-based skill optimization, and reflective search, this approach learns reusable, environment-consistent skills that generalize across diverse embodiments and tasks. Critically, since learning occurs within a world model that enforces the laws of physics, the resulting skills correspond not merely to visual plausibility but to physically realizable, causally coherent strategies. This establishes a principled framework for developing embodied agents whose learned behaviors arise from—and adhere to—the underlying mechanisms of world dynamics, advancing the path toward physics-grounded AGI.

5 Sim2Real Transferable VLA Models Built upon World Models

In pursuit of embodied intelligence, vision-language-action models represent a class of multimodal foundation policies that integrate visual perception, linguistic understanding, and motor control into a unified reasoning–action framework [8]. Unlike conventional vision-language models that operate solely on symbolic or statistical correlations, VLA models must interpret the dynamic, physically governed nature of the world. When trained purely on visual or linguistic correlations, they often lack awareness of underlying causal mechanisms, resulting in weak generalization and poor transfer to real-world environments. To overcome this limitation, we propose grounding VLA models on the proposed GS-World, thereby unifying multimodal reasoning and physically consistent control within a shared world-simulation framework.

The central goal of this design is to construct a *world-aware VLA system* capable of learning within simulation while maintaining zero-shot transfer to real-world environments. In our formulation, the GS-World serves as a generative physics prior that encodes differentiable dynamics of embodied interactions by generating latent representations of geometry, contact, and action consequence through generative simulation. These representations are reused across both the planning and acting modules of a dual-level VLA structure, ensuring that the entire reasoning–control loop remains consistent with the fundamental physical constraints defined by the world model. In effect, the VLA becomes an extension of the GS-World: the planner conducts physics-aware reasoning over latent simulation states, while the actor performs control actions constrained by those same latent dynamics, jointly forming a closed perception–action loop that is compliant with both semantics and physics.

An ideal GS-World-based VLA model is designed to fulfill several desiderata: 1) *Physics-aware Consistency:* predictions and actions comply with simulated and learnable world dynamics, preserving causal coherence; 2) *Unified Representational Space:* both reasoning and execution share latent parameters derived from world simulation, linking perceptual semantics to physical outcomes; 3) *Zero-Shot Sim2Real Transferability:* embodied policies learned in simulation are executable in the real world through domain-invariant physical embeddings; 4) *Affordance Interpretability:* intermediate reasoning relies on human-understandable affordances, promoting transparency, transfer, and modular skill reuse.

Architecture of the Sim2Real VLA Model Our proposed architecture follows the bi-system (dual-level) design commonly adopted in cognitive robotics frameworks [64, 65, 66, 67], where the high-level **planning model** (System 2) and the low-level **actor model** (System 1) are jointly coupled through the latent simulation states inherited from the GS-World. Both systems interact over shared physical priors derived from the generative simulation, including mass distribution, contact constraints, frictional forces, and motion dynamics. This shared foundation guarantees that semantic reasoning and control behaviors remain physically coherent and causally meaningful.

The **planning model**, referred to as the *World-Conditioned Planning Model (WCP)*, performs task understanding, reasoning, and decomposition through simulated rollouts within the world model's differentiable latent space. Given a sequence of observations $o_{t-H:t}$, proprioceptive input p_t, and language goal l_t, the WCP predicts a structured sequence of *affordance chains*, $\mathcal{A}_t = \{a_t^1, \ldots, a_t^K\}$, where each element corresponds to a feasible object–action relation (see Section 4). Distinct from traditional symbolic or chain-of-thought reasoning [52], WCP conducts *simulation-infused reasoning*: it evolves potential future states $\hat{s}_{t+k}$ using the GS-World's learned differential dynamics $\dot{s} = f_\theta(s, a)$, directly predicting which object interactions satisfy physical constraints and task goals. Parameters of the state-transition network, including its temporal attention mechanisms and physics-based encoders, are reused from GS-World's internal simulator module, ensuring that reasoning remains grounded in physically faithful representations of motion and causation. The planner's outputs, including affordance attributes (shape, grasp pose, torque axis, or spatial relation) and textual subgoal annotations, serve as physically consistent intermediate representations passed to the actor.

The **actor model**, named *Affordance-Guided Execution Policy (AGEP)*, generates low-level robot actions a_t based on current perception, proprioceptive signals, and the affordances supplied by the planner. Specifically, the policy can be formulated as:

$$a_t = \pi(o_t, p_t, \mathbf{z}_t^{world}, \mathcal{A}_t),$$

where the proprioceptive state p_t provides the internal state of the robot's body, complementing the external sensory observation o_t (e.g., camera images), and $\mathbf{z}_t^{world}$ denotes the shared latent state vector projected from the GS-World. The SGEP employs a generative diffusion or transformer-based decoder [68, 69] augmented with physical priors to produce motor actions that are dynamically stable, ensuring torque consistency and adherence to realistic motion constraints. While the WCP updates affordances and high-level goals at a lower frequency, the SGEP executes reactive, high-frequency control aligned with the simulated dynamics. This asynchronous coupling allows the VLA to blend deliberate planning with responsive execution, representing a hallmark of embodied reasoning consistent with the laws of physics.

Learning Paradigm for Sim2Real Adaptation The training of GS-World-based VLA models operates entirely within the simulated world environment, drawing on the skill acquisition pipeline described in Section 4. By leveraging the generative simulation engine of GS-World, the system has access to privileged internal information, including contact forces, potential energies, and dynamic coefficients of object interactions. These physical signals act as dense supervision for both planning and control modules, yielding a physically grounded self-supervised learning process.

Each learned policy is optimized through a hybrid loss:

$$\mathcal{L}_{\text{VLA-GS}} = \mathcal{L}_{\text{dyn}} + \lambda_{\text{aff}}\mathcal{L}_{\text{aff}} + \lambda_{\text{align}}\mathcal{L}_{\text{latent-align}},$$

where $\mathcal{L}_{\text{dyn}}$ enforces simulation-consistent predictions of next-state dynamics, $\mathcal{L}_{\text{aff}}$ optimizes matching between predicted and simulated affordances, and $\mathcal{L}_{\text{latent-align}}$ encourages domain invariance between simulated and real-world latent states. Thus, affordances serve as both the unit of reasoning and the supervisory signal bridging simulation and real execution.

A key to Sim2Real transfer lies in the *affordance-driven latent alignment* enabled by the GS-World. The shared latent representation $\mathbf{z}^{world}$ encodes object–robot interactions abstracted from raw visual data by capturing geometric correspondence, contact mechanics, and actuation features that are largely invariant to environmental differences such as lighting or texture. Based on such knowledge, affordances form a physically interpretable basis linking simulated and real environments. During deployment, sensory inputs from the real robot are projected through the GS-World encoder, situating them within the same simulation manifold used during training. As long as the inferred affordance attributes $\mathcal{A}_t = \{a_t^1, \ldots, a_t^K\}$ remain consistent (e.g., grasp surface normals, reachable regions, or constraint boundaries), the policy executes identically in the real world without additional fine-tuning or domain randomization. This shared physics-centric abstraction not only bridges the simulation–reality divide but also allows the system to engage in reflective reasoning similar to the skill refinement process described in Section 4, where failed interactions trigger self-consistent adjustments of affordance predictions or action policies.

Toward Engine-driven Learning for VLA Model. Since our learning paradigm operates entirely within a simulated world (i.e., GS-World), it can proceed in a fully automated and parallelized manner, allowing the VLA model to be fine-tuned simultaneously across multiple robots, environments, and

manipulation tasks (as illustrated in Figure 5, right). This engine-driven learning paradigm leverages a generative simulation engine to autonomously produce diverse, physics-consistent experiences for continuous model training. In contrast, traditional approaches to training VLAs remain bounded by the availability and diversity of offline datasets, such as real-world collections [70, 17] and synthetic reconstructions [13, 71], which remain orders of magnitude smaller and less varied than the web-scale corpora used to train LLMs or VLMs (see Figure 5, left).

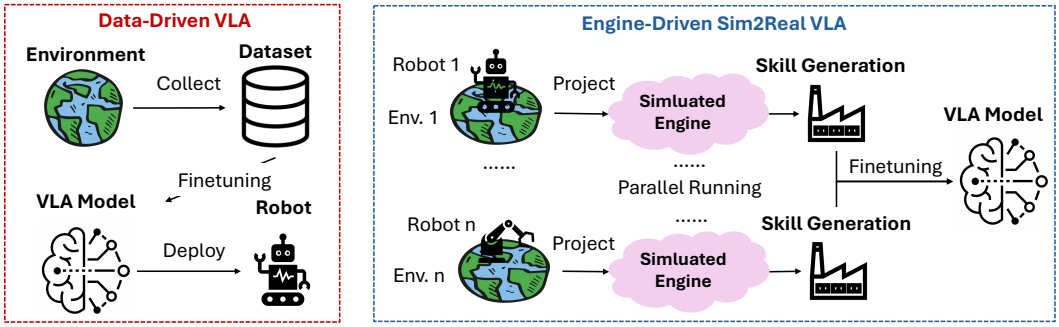

Figure 5: The data-driven VLA (left) relies on manually collected or web-crawled offline datasets to train the VLA model. In contrast, the engine-driven Sim2Real VLA (right) simulates the target robot and its environment within a physics-based engine. This simulation allows for parallelized and automated generation of robot manipulation skills, which can be used to train or fine-tune the VLAs.

The proposed shift from data-driven to engine-driven learning redefines the nature of scalability in embodied AI. Rather than passively consuming static offline data, the learning process is continuously fueled by a physics-grounded data engine that generates rich, task-relevant skill trajectories on demand. These trajectories encapsulate realistic sensorimotor experiences, affordance interactions, and environmental dynamics—serving as ever-expanding supervision for embodied model improvement. In effect, the simulation engine functions as a self-sufficient data generator, autonomously creating and streaming training data while ensuring fidelity to physical laws.

As illustrated in Figure 1, this engine-driven pipeline for Sim2Real VLA learning consists of several key components: 1) automated generation of task-specific simulation worlds via the GS-World engine, 2) continuous skill acquisition and adaptation through embodied reasoning, 3) dynamic data streaming and model fine-tuning across parallel environments, and 4) integrated verification and deployment loops for Sim2Real transfer. Together, these components form an end-to-end, engine-driven learning system capable of scaling embodied intelligence beyond the limitations of static datasets—enabling perpetual, self-improving learning cycles that combine physical realism, automation, and efficiency.

6 Morphological Co-Design through Physics-Grounded World Model

Most existing generalist robotic frameworks, including VLA models, treat the robot body as a fixed input and optimize only the control policy defined over it. Consequently, the search for optimal robot structures relies heavily on handcrafted design and trial-based engineering. However, just as biological organisms evolve their body morphologies to achieve adaptive fitness in specific environments, embodied agents should co-evolve their physical configurations in accordance with environmental dynamics and task objectives [72]. The world model provides a natural substrate for this co-design process: as it simulates physically consistent interactions between the robot and the environment, structural modifications can be continuously evaluated by analyzing how changes in morphology influence dynamics, stability, and task success. In this sense, morphological co-design within a world model generalizes the concept of evolutionary optimization into a differentiable, physics-aware framework for embodied learning.

In this section, we introduce a physics-grounded framework for **morphological co-design**, where the robot's structural parameters are optimized jointly with its control policies within a world model. This establishes a closed-loop design of *body–brain co-evolution*, ensuring that both morphology and behavior emerge under consistent physical principles simulated by the world model.

Co-Design Formulation with World-Model Coupling. Let ξ denote the morphological parameters of the robot (e.g., link lengths, joint configurations, mass distribution, or actuation capacity), and π_θ represent the control policy with learnable parameters θ. The unified optimization objective is expressed as:

$$\max_{\xi,\theta} \mathcal{J}(\pi_\theta, \mathcal{M}_\xi) = \max_{\xi,\theta} \mathbb{E}_{\pi_\theta,(\mathcal{T},\mu_0)\sim\mathcal{M}_\xi} \left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) \right] \tag{3}$$

where $\mathcal{M}_\xi$ refers to the world model parameterized by ξ, r is the reward function measuring task progress, and π_θ denotes the policy model (e.g., the VLA model in Section 5). Unlike traditional simulators, $\mathcal{M}_\xi$ enables bidirectional gradient propagation: morphological parameters are embedded directly within the differentiable world simulation, allowing learning signals to flow from behavioral outcomes back to structural representations. As a result, both body and policy parameters can be updated jointly or alternately, yielding morphology–policy pairs that are dynamically compatible, causally coherent, and task-specialized. Figure 1 (bottom) provides an overview of this joint optimization pipeline, which consists mainly of the following steps.

1) World-Model-Driven Evolutionary Reasoning. While gradient-based optimization provides local fine-tuning of morphology, global exploration of the design space remains crucial for discovering novel and structurally diverse configurations. The world model enables an embodied evolutionary reasoning mechanism in which simulated populations of robot morphologies are evaluated under consistent physical dynamics. Analogous to natural evolution, candidate morphologies undergo virtual reproduction, mutation, and selection within simulation rollouts, where fitness signals are derived directly from oracle evaluations built upon the world model. This process is efficiently organized as a Monte Carlo Tree Search (MCTS) structure [63], where each node represents a distinct morphology–policy pair, and rollout trials simulate their performance across varied physical conditions. Guided by an upper-confidence criterion and assisted by affordance-conditioned priors, this search identifies design topologies yielding maximal generalization across task classes, effectively turning the world model into a self-contained evolutionary simulator for robotic design.

2) Morphology Representation under Differentiable Dynamics. To enable structured optimization, robot morphologies are represented as parameterized graphs $\mathcal{G}_\xi = (\mathcal{V}, \mathcal{E})$, where nodes correspond to limbs and actuators, and edges represent joint connections with annotated physical attributes (e.g., torque limits, damping coefficients, material stiffness). Each morphological graph is embedded into a latent structural manifold via a graph neural encoder [73], producing differentiable embeddings that interface with the policy model and the world model alike. In simulation, these structural embeddings are bound to dynamic parameters within the physics simulator, such as inertia matrices or deformation tensors, enabling the world model to continuously recompute physical dynamics as morphology changes. The representation thus bridges geometric configuration and dynamic simulation, ensuring that morphology updates yield physically meaningful outcomes.

3) Physics-Grounded Optimization via GS-World. Within the paradigm of GS-World [16, 35, 37], morphology optimization is driven not merely by visual plausibility but by rigorous physical computation. Each candidate morphology is instantiated within the GS-World, where forward dynamics are computed via differentiable physics solvers [42, 43]. During rollout, world states, including contact forces, energy consumption, and joint stress, are continuously monitored to assess structural feasibility. Importantly, gradients of performance metrics (e.g., manipulation success rate or gait stability) with respect to morphology are automatically propagated through the GS-World simulation. This allows robot morphology to evolve naturally toward designs that minimize energy cost, maximize control stability, or satisfy specified task constraints. As GS-World captures accurate causal relations between control and dynamics, learned morphologies inherently obey real-world physics, improving their transferability in Sim2Real scenarios.

The proposed morphology–policy co-design on a physics-grounded world model can be interpreted as an instantiation of embodied evolution within a controlled generative framework. By embedding morphology into the latent physical space of the world model, and allowing physical consistency, affordance reasoning, and simulated evolution to jointly drive optimization, one obtains a scalable route toward the autonomous synthesis of both bodies and behaviors. In essence, the same generative simulation engine that supports skill acquisition for control (Section 4) is now extended to drive self-optimization of embodiment, thereby closing the feedback loop between structure and intelligence. As the GS-World continues to improve its physical fidelity and differentiable simulation capability, the boundary between morphology learning and policy learning will gradually dissolve, leading to

an integrated paradigm of **physics-grounded robotic co-design**, where embodied agents evolve holistically under the same laws that govern the physical world.

7 Conclusion

In this paper, we present GS-World, as well as the engine-driven Sim2Real VLA paradigm, which forms a blueprint for achieving scalable, physics-grounded embodied intelligence. By shifting from static, data-driven training to a generative, engine-driven simulation loop, the proposed system directly addresses the core limitations of current robotic learning—data scarcity, inefficiency, and weak physical grounding. Through the introduction of the efficiency law, the work articulates a quantitative foundation linking model performance to data-generation efficiency, thus motivating the use of generative simulation engines as perpetual sources of multimodal, physically accurate experiences.

Under this paradigm, GS-World serves as a universal engine that produces coherent 3D environments, dynamic interactions, and physically faithful trajectories in a differentiable manner. As a result, skill acquisition, policy learning, model verification, and embodiment evolution can operate continuously in a closed-loop process. The integration of Chain-of-Affordance reasoning, reflective embodied learning, and morphology–policy co-design enables robots not only to act intelligently but also to evolve their structures and strategies through self-supervised adaptation.

Ultimately, GS-World redefines the pathway toward physical AGI by unifying generative modeling, physical simulation, and embodied reasoning within a shared computational framework. This engine-driven approach transforms simulation into a foundational engine of knowledge, allowing embodied agents to autonomously learn, test, and improve in diverse virtual worlds before seamlessly transferring their intelligence to the real world.

References

[1] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[2] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[3] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training. 2018.

[4] OpenAI. Sora: creating video from text, 2024.

[5] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[6] Peihao Wang, Rameswar Panda, and Zhangyang Wang. Data efficient neural scaling law via model reusing. In *Proceedings of the 40th International Conference on Machine Learning*, pages 36193–36204, 2023.

[7] Zhengyu Chen, Siqi Wang, Teng Xiao, Yudong Wang, Shiqi Chen, Xunliang Cai, Junxian He, and Jingang Wang. Sub-scaling laws: on the role of data density and training strategies in llms. *arXiv preprint arXiv:2507.10613*, July 2025.

[8] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision-language-action models for embodied ai. *arXiv preprint arXiv:2405.14093*, 2024.

[9] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. Will we run out of data? limits of llm scaling based on human-generated data. *arXiv preprint arXiv:2211.04325*, 2022.

[10] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *arXiv preprint arXiv:2206.14486*, 2022.

[11] Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Position: levels of AGI for operationalizing progress on the path to AGI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36308–36321, July 2024.

[12] Kaiyan Zhang, Biqing Qi, and Bowen Zhou. Towards building specialized generalist ai with system 1 and system 2 fusion. *arXiv preprint arXiv:2205.11487*, 2024.

[13] Ajay Mandlekar, Soroush Nasiriany, Bowen Wen, Iretiayo Akinola, Yashraj Narang, Linxi Fan, Yuke Zhu, and Dieter Fox. Mimicgen: a data generation system for scalable robot learning using human demonstrations. In *Conference on Robot Learning (CoRL)*, 2023.

[14] Yufei Wang, Zhou Xian, Feng Chen, Tsun-Hsuan Wang, Yian Wang, Katerina Fragkiadaki, Zackory Erickson, David Held, and Chuang Gan. Robogen: towards unleashing infinite data for automated robot learning via generative simulation. In *International Conference on Machine Learning (ICML)*, 2024.

[15] Soroush Nasiriany, Abhiram Maddukuri, Lance Zhang, Adeet Parikh, Aaron Lo, Abhishek Joshi, Ajay Mandlekar, and Yuke Zhu. Robocasa: large-scale simulation of everyday tasks for generalist robots. *arXiv preprint arXiv:2406.02523*, 2024.

[16] NVIDIA, Alisson Azzolini, Hannah Brandon, Prithvijit Chattopadhyay, Huayu Chen, et al. Cosmos-reason1: from physical common sense to embodied reasoning. *arXiv preprint arXiv:2503.15558*, 2025.

[17] Abby O'Neill, Abdul Rehman, Abhiram Maddukuri, Abhishek Gupta, Abhishek Padalkar, Abraham Lee, Acorn Pooley, Agrim Gupta, Ajay Mandlekar, Ajinkya Jain, et al. Open x-embodiment: robotic learning datasets and rt-x models: open x-embodiment collaboration. In *IEEE International Conference on Robotics and Automation (ICRA)*, pages 6892–6903. IEEE, 2024.

[18] David Ha and Jurgen Schmidhuber. World models. *CoRR*, abs/1803.10122, 2018.

[19] Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 2022.

[20] Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

[21] Rohin Manvi, Samar Khanna, Gengchen Mai, Marshall Burke, David Lobell, and Stefano Ermon. Geollm: extracting geospatial knowledge from large language models. *arXiv preprint arXiv:2310.06213*, 2023.

[22] Yuxing Long, Xiaoqi Li, Wenzhe Cai, and Hao Dong. Discuss before moving: visual language navigation via multi-expert discussions. *arXiv preprint arXiv:2309.11382*, 2023.

[23] Jake Bruce, Michael D. Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C. Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: generative interactive environments. In *Proceedings of the 41st International Conference on Machine Learning*, pages 4603–4623, 2024.

[24] Edward C. Tolman. Cognitive maps in rats and men. *Psychological Review*, 55(4):189–208, 1948.

[25] Sudipta Paul, Amit K Roy-Chowdhury, and Anoop Cherian. Avlen: audio-visual-language embodied navigation in 3d environments. *arXiv preprint arXiv:2210.07940*, 2022.

[26] Jialong Wu, Shaofeng Yin, Ningya Feng, Xu He, Dong Li, Jianye Hao, and Mingsheng Long. ivideogpt: interactive videogpts are scalable world models. In *Advances in Neural Information Processing Systems*, 2024.

[27] Yash Jain, Anshul Nasery, Vibhav Vineet, and Harkirat Behl. Peekaboo: interactive video generation via masked-diffusion. *arXiv preprint arXiv:2312.07509*, 2023.

[28] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

[29] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, Zhengzhong Liu, Eric Xing, and Zhiting Hu. Pandora: towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, June 2024.

[30] Alejandro Escontrela, Ademi Adeniji, Wilson Yan, Ajay Jain, Xue Bin Peng, Ken Goldberg, Youngwoon Lee, Danijar Hafner, and Pieter Abbeel. Video prediction models as rewards for reinforcement learning. *Neural Information Processing Systems*, 2023.

[31] Huayi Zhou, Ruixiang Wang, Yunxin Tai, Yueci Deng, Guiliang Liu, and Kui Jia. You only teach once: learn one-shot bimanual robotic manipulation from video demonstrations. *arXiv preprint arXiv:2501.14208*, 2025.

[32] Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model? – a physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

[33] Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models understand physical principles? *arXiv preprint arXiv:2501.09038*, 2025.

[34] Haoyu Zhen, Qiao Sun, Hongxin Zhang, Junyan Li, Siyuan Zhou, Yilun Du, and Chuang Gan. Tesseract: learning 4d embodied world models. *arXiv preprint arXiv:2504.20995*, 2025.

[35] Shaowei Liu, Zhongzheng Ren, Saurabh Gupta, and Shenlong Wang. Physgen: rigid-body physics-grounded image-to-video generation. In *European Conference on Computer Vision (ECCV)*, 2024.

[36] Tianyuan Zhang, Hong-Xing Yu, Rundi Wu, Brandon Y. Feng, Changxi Zheng, Noah Snavely, Jiajun Wu, and William T. Freeman. Physdreamer: physics-based interaction with 3d objects via video generation. In *European Conference on Computer Vision (ECCV)*, 2024.

[37] Ziang Cao, Zhaoxi Chen, Liang Pan, and Ziwei Liu. Physx-3d: physical-grounded 3d asset generation. *arXiv preprint arXiv:2507.12465*, 2025.

[38] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. New York, NY, USA, 2003.

[39] Zhengfei Kuang, Shengqu Cai, Hao He, Yinghao Xu, Hongsheng Li, Leonidas Guibas, and Gordon Wetzstein. Collaborative video diffusion: consistent multi-video generation with camera control. In *arXiv preprint arXiv:2402.00000*, 2024.

[40] Mohammad Asim, Christopher Wewer, Thomas Wimmer, Bernt Schiele, and Jan Eric Lenssen. Met3r: measuring multi-view consistency in generated images. *arXiv preprint arXiv:2501.06336*, 2025.

[41] NVIDIA. Physx physics engine, 2025.

[42] Emanuel Todorov, Tom Erez, and Yuval Tassa. Mujoco: a physics engine for model-based control. In *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 5026–5033. IEEE, 2012.

[43] Genesis Authors. Genesis: a generative and universal physics engine for robotics and beyond, December 2024.

[44] Hanxiao Jiang, Hao-Yu Hsu, Kaifeng Zhang, Hsin-Ni Yu, Shenlong Wang, and Yunzhu Li. Phystwin: physics-informed reconstruction and simulation of deformable objects from videos. *International Conference on Computer Vision (ICCV)*, 2025.

[45] Patrick Kidger. On neural differential equations. *arXiv preprint arXiv:2202.02435*, 2022.

[46] Jiabao Lei and Kui Jia. Analytic marching: an analytic meshing solution from deep implicit surface networks. In *International Conference on Machine Learning (ICML)*, July 2020.

[47] Samuel Pfrommer, Mathew Halm, and Michael Posa. Contactnets: learning discontinuous contact dynamics with smooth, implicit representations. *arXiv preprint arXiv:2009.11193*, 2020.

[48] Jie Xu, Eric Heiden, Iretiayo Akinola, Dieter Fox, Miles Macklin, and Yashraj Narang. Neural robot dynamics. In *9th Annual Conference on Robot Learning*, 2025.

[49] Hugo Bertiche, Meysam Madadi, and Sergio Escalera. Neural cloth simulation. *ACM Transactions on Graphics*, 41(6):1–14, 2022.

[50] L'ubor Ladický, SoHyeon Jeong, Barbara Solenthaler, Marc Pollefeys, and Markus Gross. Data-driven fluid simulations using regression forests. *ACM Transactions on Graphics*, 34(6):1–14, October 2015.

[51] Mohit Shridhar, Xingdi Yuan, Marc-Alexandre Côté, Yonatan Bisk, Adam Trischler, and Matthew Hausknecht. Alfworld: aligning text and embodied environments for interactive learning. In *International Conference on Learning Representations (ICLR)*, 2021.

[52] Michal Zawalski, William Chen, Karl Pertsch, Oier Mees, Chelsea Finn, and Sergey Levine. Robotic control via embodied chain-of-thought reasoning. In *Conference on Robot Learning (CoRL)*, volume 270, pages 3157–3181, 2024.

[53] Jinming Li, Yichen Zhu, Zhibin Tang, Junjie Wen, Minjie Zhu, Xiaoyu Liu, Chengmeng Li, Ran Cheng, Yaxin Peng, and Feifei Feng. Improving vision-language-action models via chain-of-affordance. *CoRR*, abs/2412.20451, 2024.

[54] Wentao Yuan, Jiafei Duan, Valts Blukis, Wilbert Pumacay, Ranjay Krishna, Adithyavairavan Murali, Arsalan Mousavian, and Dieter Fox. Robopoint: a vision-language model for spatial affordance prediction in robotics. In *Conference on Robot Learning (CoRL)*, volume 270, pages 4005–4020, 2024.

[55] Weikang Wan, Haoran Geng, Yun Liu, Zikang Shan, Yaodong Yang, Li Yi, and He Wang. Unidexgrasp++: improving dexterous grasping policy learning via geometry-aware curriculum and iterative generalist-specialist learning. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 3868–3879, 2023.

[56] Yihe Tang, Wenlong Huang, Yingke Wang, Chengshu Li, Roy Yuan, Ruohan Zhang, Jiajun Wu, and Li Fei-Fei. Uad: unsupervised affordance distillation for generalization in robotic manipulation. *arXiv preprint arXiv:2506.09284*, 2025.

[57] Shijie Wu, Yihang Zhu, Yunao Huang, Kaizhen Zhu, Jiayuan Gu, Jingyi Yu, Ye Shi, and Jingya Wang. Afforddp: generalizable diffusion policy with transferable affordance. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6971–6980, 2025.

[58] Teli Ma, Jia Zheng, Zifan Wang, Ziyao Gao, Jiaming Zhou, and Junwei Liang. Glover++: unleashing the potential of affordance learning from human behaviors for robotic manipulation. *arXiv preprint arXiv:2505.11865*, 2025.

[59] Huihui Guo, Fan Wu, Yunchuan Qin, Ruihui Li, Keqin Li, and Kenli Li. Recent trends in task and motion planning for robotics: a survey. *ACM Computing Surveys*, 55(13s):1–36, 2023.

[60] Bharat Singh, Rajesh Kumar, and Vinay Pratap Singh. Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55(2):945–990, 2022.

[61] Yecheng Jason Ma, William Liang, Guanzhi Wang, De-An Huang, Osbert Bastani, Dinesh Jayaraman, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Eureka: human-level reward design via coding large language models. In *The Twelfth International Conference on Learning Representations (ICLR)*, 2024.

[62] Maciej Świechowski, Konrad Godlewski, Bartosz Sawicki, and Jacek Mańdziuk. Monte carlo tree search: a review of recent modifications and applications. *Artificial Intelligence Review*, 56(3):2497–2562, 2023.

[63] Jing Yu Koh, Stephen McAleer, Daniel Fried, and Ruslan Salakhutdinov. Tree search for language model agents. *CoRR*, abs/2407.01476, 2024.

[64] Johan Bjorck, Fernando Castañeda, Nikita Cherniadev, Xingye Da, Runyu Ding, Linxi Fan, Yu Fang, Dieter Fox, Fengyuan Hu, Spencer Huang, et al. Gr00t n1: an open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.

[65] Qingwen Bu, Jisong Cai, Li Chen, Xiuqi Cui, Yan Ding, Siyuan Feng, Shenyuan Gao, Xindong He, Xu Huang, Shu Jiang, et al. Agibot world colosseo: a large-scale manipulation platform for scalable and intelligent embodied systems. *arXiv preprint arXiv:2503.06669*, 2025.

[66] Haoming Song, Delin Qu, Yuanqi Yao, Qizhi Chen, Qi Lv, Yiwen Tang, Modi Shi, Guanghui Ren, Maoqing Yao, Bin Zhao, et al. Hume: introducing system-2 thinking in visual-language-action model. *arXiv preprint arXiv:2505.21432*, 2025.

[67] Qingwen Bu, Hongyang Li, Li Chen, Jisong Cai, Jia Zeng, Heming Cui, Maoqing Yao, and Yu Qiao. Towards synergistic, generalized, and efficient dual-system for robotic manipulation. *arXiv preprint arXiv:2410.08001*, 2024.

[68] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4172–4182, 2023.

[69] Tony Z. Zhao, Vikash Kumar, Sergey Levine, and Chelsea Finn. Learning fine-grained bimanual manipulation with low-cost hardware. In *Robotics: Science and Systems (RSS)*, 2023.

[70] Homer Rich Walke, Kevin Black, Tony Z. Zhao, Quan Vuong, Chongyi Zheng, Philippe Hansen-Estruch, Andre Wang He, Vivek Myers, Moo Jin Kim, Max Du, Abraham Lee, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: a dataset for robot learning at scale. In *Conference on Robot Learning (CoRL)*, volume 229, pages 1723–1736, 2023.

[71] Zhenyu Jiang, Yuqi Xie, Kevin Lin, Zhenjia Xu, Weikang Wan, Ajay Mandlekar, Linxi Fan, and Yuke Zhu. Dexmimicgen: automated data generation for bimanual dexterous manipulation via imitation learning. In *IEEE International Conference on Robotics and Automation (ICRA)*, 2025.

[72] Agrim Gupta, Silvio Savarese, Surya Ganguli, and Li Fei-Fei. Embodied intelligence via learning and evolution. *Nature Communications*, 12(1):5721, 2021.

[73] Carmelo Sferrazza, Dun-Ming Huang, Fangchen Liu, Jongmin Lee, and Pieter Abbeel. Body transformer: leveraging robot embodiment for policy learning. In *Conference on Robot Learning (CoRL)*, volume 270, pages 3407–3424, 2024.