ALWM: Action Language World Model

Bing Lin
Linma Research 455178083[at]qq[dot]com

Abstract

In this paper, actions and language are closely bound together. Action sequences could be coded and act as fundamental building block of language. Action language describes the world. Thus, the world has been mapped into action space. Language Image Natural Modeling Architecture (LINMA) contributes laws to construct the mapping engineering: law of head, law of limb posture, law of hand shape, law of motion trajectory. Language token could be action token. Action language could directly drive humanoid robot to perform actions.

1 Introduction

The world model is the core underlying technology connecting artificial general intelligence (AGI) and embodied intelligence. Fundamentally, it functions as an internal digital simulator of the physical world built inside an AI agent. By learning causal correlations among environmental observations, agent actions and state transitions, a world model enables off-line counterfactual imagination, future state prediction and action planning without constant interaction with the real environment. It fills the critical capability gap of large language models, which only process textual semantics, and static vision models limited to frame recognition.

Evolving from a simple auxiliary reinforcement learning module to an integrated technical ecosystem combining video generation, neural physical simulation and robot motion planning, world models have diverged into two dominant paradigms and spawned a large set of derivative architectures, auxiliary algorithms and engineering optimization techniques. This paper systematically sorts out its complete evolutionary timeline, mainstream architectures, core derivative technologies, practical bottlenecks and industrial development trends.

1.1 Four Stages of World Model Evolution

Stage 1: Theoretical Origin & Initial Prototype (1990–2018) The core idea of world models originates from cognitive science: humans rely on internal mental models to foresee action outcomes. Before deep learning prevailed, research depended on manually coded physical engines and Markov Decision Processes (MDPs), suffering from weak generalization and limited to simple games and fixed mechanical control.

In 2018, Ha & Schmidhuber proposed the original end-to-end deep learning architecture World Models, marking the birth of modern world models [1]. It consists of a VAE visual encoder, recurrent MDN-RNN dynamics module and reinforcement learning controller. Raw pixels are compressed into low-dimensional latent vectors, and the recurrent network predicts latent state evolution conditioned on actions. The prototype validated the feasibility of training policies purely within imagined latent space. However, pixel-level reconstruction brought heavy computational overhead, and the system lacked abstract causal reasoning. Concurrently, MuZero pioneered a divergent path by discarding image reconstruction and directly modeling states, rewards and value functions via self-play [2]. It proved that abstract latent dynamics prediction outperforms pure visual reproduction, laying the foundation for the latent-space world model lineage.

Stage 2: Latent-Space Dynamics Optimization (2019–2022) This phase focused on improving computational efficiency and long-sequence prediction stability. PlaNet first proposed the Recurrent State-Space Model (RSSM) to realize end-to-end latent planning from pixel input [3]. Later the Dreamer series (DreamerV1/V2/V3) further improved latent imagina-

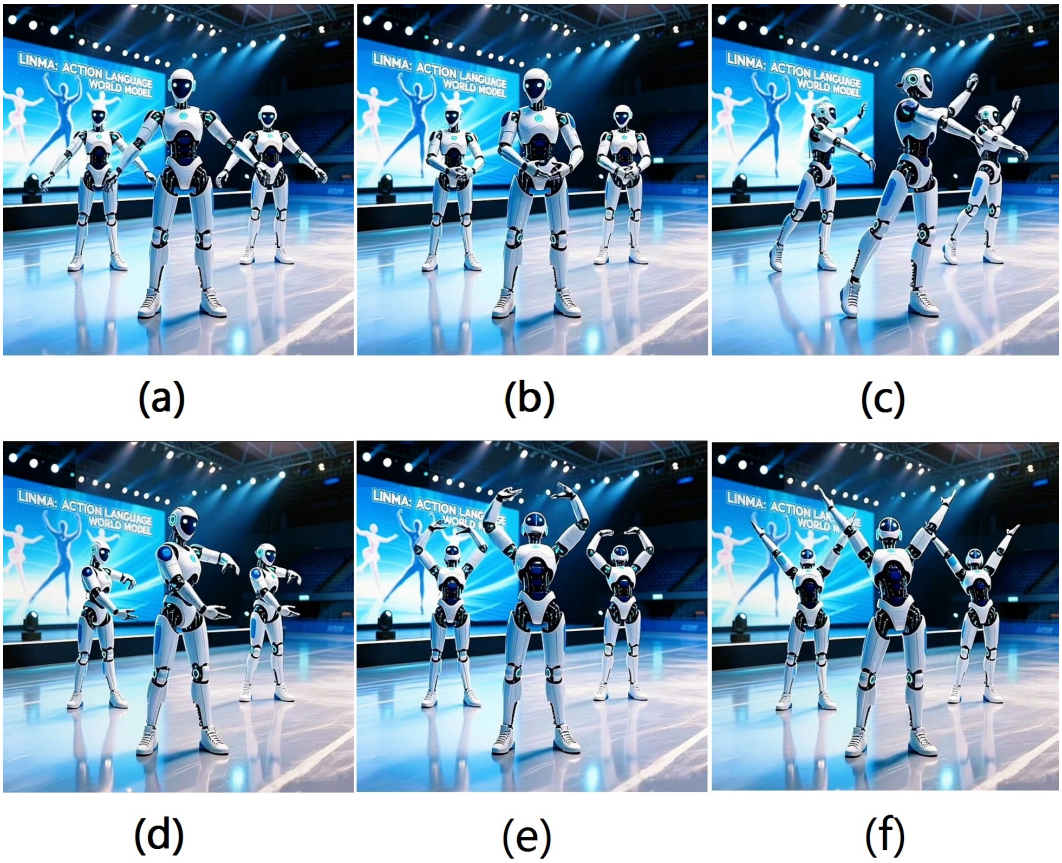

Figure 1: Action demonstration: (a) Action of STanding. (b) Action of holding a STone. (c) Action of holding a STICK. (d) Action of holding a chiLd. (e) Action of Crab holding claws. (f) Action of Carrying.

tion training efficiency for cross-domain tasks [4]. Multiple improved variants emerged: HRSSM adopted dual-branch latent alignment; DreamerPro replaced heavy decoders with prototype representations to suppress visual noise.

Nevertheless, world models remained auxiliary tools for reinforcement learning, confined to game AI and simulated robot training. Most models only fitted temporal statistical patterns rather than inherent physical rules such as gravity and collision, leading to frequent physically implausible prediction drift.

Stage 3: Generative Boom & Rise of Video-Based World Models (2023–2024) The maturity of diffusion models and large temporal Transformers ignited the generative world model wave. OpenAI released the Sora video world simulator as a representative work of generative world models [8]. DeepMind proposed Genie to generate interactive environments from a single static image [9]. NVIDIA launched the Cosmos physical world foundation model oriented to robot physical simulation [10].

In parallel, LeCun put forward the top-level theoretical framework of autonomous machine intelligence and proposed the JEPA non-generative prediction paradigm [5]. Meta further extended the framework to video prediction and released V-JEPA [6]. Since then, the field has split into two parallel technical routes: generative video world models optimized for visual rendering, and non-generative latent prediction models oriented toward robotic reasoning and control.

Stage 4: Embodied Intelligence Integration Era (2025–Present) Meta upgraded V-JEPA to V-JEPA2 specially optimized for robot perception and physical planning [7]. The industry focus has shifted toward physically interactive real-world deployment, prioritizing sim-to-real transfer, physical consistency and action con-

trollability rather than visual fidelity. Technical fusion between the two paradigms accelerates: generative video models generate synthetic training data, while latent dynamics models conduct pre-action rehearsal for robots. Lightweight world models are gradually deployed for dexterous hand manipulation and robotic calligraphy tasks.

1.2 Two Mainstream Architectures & Rich Derivative Technologies

1.2.1 Generative Video World Model Family

Core Architecture Temporal Transformer / Spacetime Diffusion (DiT) + video tokenizer.

Representative Works Sora [8], Cosmos Predict [10], WorldDreamer, Seedance.

Key Derivative & Improved Technologies

- Action-conditioned video diffusion; spacetime patch prediction; flow-guided video generation for mitigating motion distortion.
- Masked visual token prediction: adopted by WorldDreamer, borrowing the masked modeling paradigm from large language models to enhance physical rule learning.
- Video-to-motion re-targeting: transfers human motion sequences extracted from videos to humanoid robots and dexterous hands, supporting motion imitation for performing robots.

Advantages: High visual fidelity, intuitive output for digital human animation and performance robot material generation. Defects: Prone to visual hallucinations, object penetration and physically inconsistent motion; Kang et al. analyzed the essential gap between video generation and real physical modeling [15].

1.2.2 Non-Generative Latent Dynamic World Model Family

Core Architecture Visual encoder + latent dynamics predictor + planning module.

Representative Works DreamerV3 [4], V-JEPA2 [7], PlaNet [3].

Key Derivative Variants

- RSSM extended series: HRSSM, Disentangled World Model (DisWM), which decouples semantic information from texture features to improve cross-domain generalization.
- Hierarchical world models (H-WM): separates high-level semantic planning and low-level motion execution, suitable for long-horizon continuous robot tasks.
- Object-centric latent world models: SlotFormer [11], SAVi++ [12], C-SWM. The model decomposes scenes into independent object embeddings, supporting combinatorial reasoning and individual object dynamics prediction.

1.3 Emerging Cross-Domain Derivative Branches

1.3.1 3D Geometric World Models (Spatial Intelligence Route)

3D Gaussian Splatting technology becomes the core rendering foundation of modern 3D world models [17]. DriveDreamer builds 4D world models for autonomous driving to solve the view drift problem of 2D video prediction [18]. Typical implementations also include Marble, OccWorld and BEVWorld.

1.3.2 Physics-Informed & Differentiable Neural Simulation

OrbiSim integrates differentiable contact mechanics to build a physical enhanced world model for flexible object manipulation such as brush hairs [13]. Hybrid neural-analytical simulators combine MuJoCo, Genesis physical engines with learned dynamics modules.

1.3.3 World-Action Model (WAM) & VLA Fusion Architecture

It is the fastest-growing derivative direction for robotic manipulation. It takes pre-trained world model backbones as motion priors to jointly predict scene evolution and robot actions, reducing dependence on massive real-world robot interaction data.

1.3.4 Event-Based & Lightweight World Models

TD-MPC2 proposes a lightweight world model framework for continuous real-time control of robotic arms and dexterous hands [14]. Event-camera world models and Mamba-based world models further optimize edge computing efficiency.

1.3.5 Causal & Counterfactual World Models

Different from standard statistical correlation fitting, causal world models support intervention reasoning and counterfactual imagination. World-of-Thought mechanism realizes latent-space chain-of-thought task decomposition without text prompts.

1.4 Core Technical Challenges

First, long-horizon prediction drift. Autoregressive inference accumulates errors continuously, leading to collapse of physical consistency after dozens of steps.

Second, poor modeling capability for flexible nonlinear objects. Contact dynamics and elastic deformation of brush hairs, fabrics and cables cannot be fully captured by pure data-driven models — this is the core obstacle preventing robotic calligraphy from reproducing subtle brushstroke techniques such as "square outline with rounded inner transition".

Third, the persistent sim-to-real gap. Strategies trained in virtual simulators degrade sharply on physical hardware due to mechanical clearances, friction uncertainty and sensor noise. The D4RL dataset is widely used as a standard benchmark to evaluate such reinforcement learning control performance [20].

Fourth, trade-off between computational cost and fidelity. High-fidelity generative models demand enormous computing resources, while lightweight latent models lack intuitive visual output for debugging.

1.5 Conclusion & Future Trends

Bandaru systematically summarized the overall framework of latent dynamics and generative simulation world models [19]. Shang et al. published a comprehensive survey focusing on world models applied to embodied AI [16].

The evolutionary trajectory of world models can be summarized as several major shifts: from pixel reconstruction toward abstract geometric/latent representation; from hand-coded physical rules toward data-and-physics hybrid modeling; from offline video generation toward interactive embodied simulation; from independent prediction modules toward unified world-action policy frameworks.

In industrial segmentation, different technical lines will occupy differentiated scenarios: generative video world models continue supporting motion imitation, micro-motion rendering for performing humanoid robots and digital humans. Latent dynamics, 3D geometric and physics-augmented world models dominate precision manipulation tasks, autonomous driving and industrial robotics. In the long run, next-generation unified world models will integrate 3D spatial perception, differentiable physical engines and Vision-Language-Action frameworks, balancing visual generation quality and physical rigor. Equipped with internal imagination and counterfactual deduction capabilities, world models enable AI agents to perceive, reason and interact with the physical environment, forming foundational infrastructure for advanced embodied intelligence.

2 Background

The goal of this paper is to construct a concise, effective, and lightweight language image natural modeling architecture (LINMA).

The idea of this paper is to establish a direct correspondence mapping relationship between visual images and textual elements.

Before proposing a specific solution, let's first examine the relationship between sign language, semaphore, and written language.

Sign Language: Utilizes the configuration of finger shapes to represent language text letters and build words.

Semaphore: Involves moving both arms, placing them in various positions and angles around the body, using combinations of arm postures and angles to express text letters. On a clock face, there are hour and minute hands, and the different positions and angles of these two hands indicate different times. In semaphore, the arms function like the two hands of a clock, utilizing position and angle to represent different language letters.

3 Model Architecture

Principle of the work:

Make full, comprehensive and integrated use of the morphology, shape, and positional relationship of the main and terminal limbs, including arms, palms, fingers, legs, feet, mouth, tongue, etc.

Analyze and summarize the placement of the arms, simplify and abstract them into symbols with similar shapes to depict the postures of the arms.

Analyze and summarize the placement forms of the hands, including the direction of the palm of

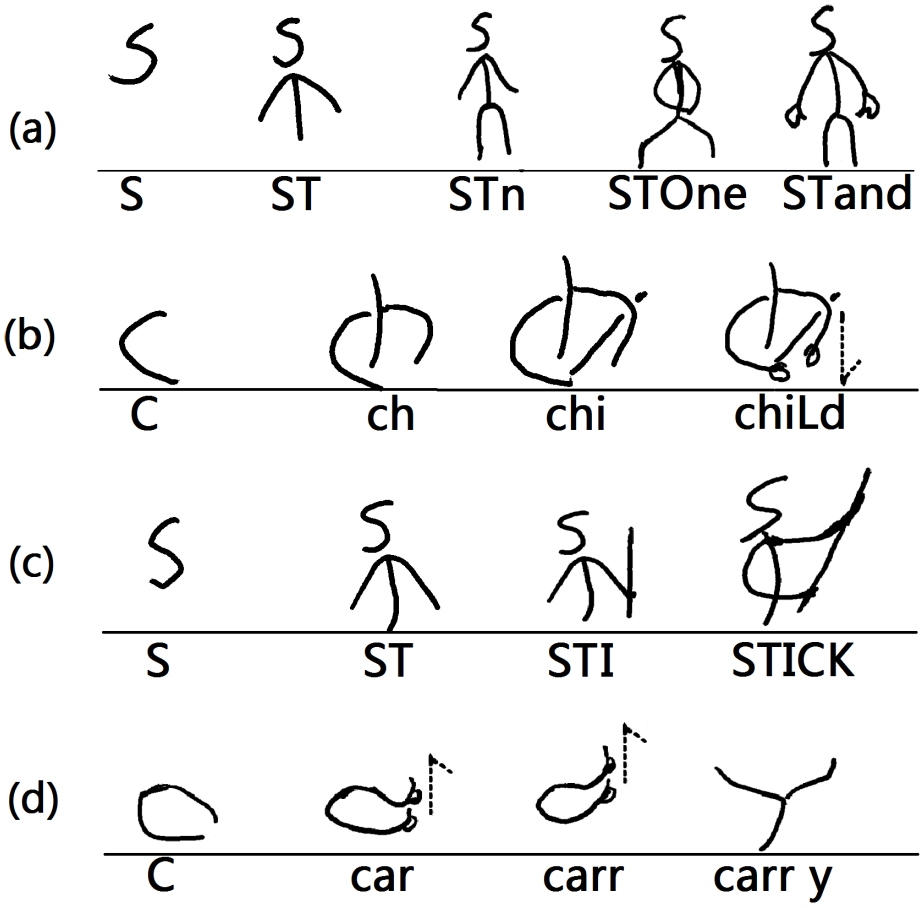

Figure 2: LINMA Thinking Model: symbolization of action sequence. (a) Just placing STn vertically, a depiction of human body basic structure is presented, which could be used further for deriving various actions, such as: Holding a STOne, STanding upright, etc. (b) Depiction of Holding a chiLd. (c) Depiction of Holding a STICK. (d) Depiction of Carrying items. [22]

the hand, the vertical relationship between the palm and the wrist, simplify and abstract them into symbols with similar forms to depict the postures of the hand shapes.

Analyze, summarize, and classify the movement trajectories of body parts during the process of shape changing, and describe them in symbolic format based on the shape characteristics of the trajectories.

A specific posture action sequence has a specific function and completes a specific function. It can be fixed, programmed, and streamlined, and can be continuously optimized until it is finalized.

The target object of the expression of the meaning of a specific modeling action posture can be the function of the modeling posture itself, or can refer to the interactive objects involved in the modeling action.

As for visual restoration, based on static graphic images, the modeling postures before and after are inferred to fill in and form a continuous and complete modeling action sequence.

Various limb organs are the basic components of the bodies of human and animal species, and they are the basic tools for animal species to interact and communicate with the natural world. External limbs and internal organs have different characteristics: under normal circumstances, internal organs are invisible and do not have the characteristics of visual presentation, while external limbs are organs that can be visually presented. Because they can be visually presented,

they have the function of generating visual information. The main and basic functions of animal limbs and organs are to support and move the body, to hunt for food, and so on. Visual information presentation is an additional function, which can play a role in demonstration, communication and exchange. Whether a species can fully develop and utilize these additional information functions is an important aspect to judge the accumulation and evolution of species intelligence. By comprehensively understanding and mastering the essence of these visual information carriers, human beings have evolved into higher form of intelligence, which enables them to surpass other species.

3.1 Body, Parts, Limbs

The arms and ten fingers are important limb organs of humans and similar species, such as monkeys, apes and other quadrumana. The relationship between the two arms is complicated and varied, and the placement and shape are also variable. They can accomplish a wide variety of functional tasks, and have evolved to the extent of being used at will.

It is necessary to analyze and describe various placements and shapes in detail.

When humans explore and understand things in the world, they need to establish an interactive relationship with the things, so that they can truly understand and master such things. Only by establishing a relationship with the body parts that humans can control, and this relationship is able to be expressed, reproduced, and understood, can the things be truly mastered.

Parts such as the arms and ten fingers are the body limb organs of humans themselves, which can be freely controlled and are the most familiar and controllable resource tools.

The limb shapes, postures, gestures, and movement patterns of humans and animals are visually perceivable and can be transformed into expressions by copying and depicting. Symbolic expression is achieved, and symbols and actions can be mapped and transformed into each other. Symbolic text is obtained by copying the movement posture, and the movement posture can be restored from the symbolic text.

In order to complete specific target tasks, humans or animal species need to construct a series of subtle movement patterns and shapes, and perform shape switch transitions to achieve a specific task. Over time, a binding relationship is formed between specific tasks and specific movement shape transition sequences. The movement shape sequences gradually become fixed, programmed, and process-oriented.

Different individuals will get different results when observing and summarizing the movement postures from different angles. In the process of evolution, the schemes that are the simplest, most abstract, and most expressive will gain more recognition. Compatibility should be maintained as much as possible for existing symbolic schemes.

3.2 Labor Scenes and Actions

Analyze and examine several common life labor scenes: holding a stone, holding a child, holding a stick, and standing. From these basic scenes, look at the details of the body postures.

Figure 1 demonstrates common life actions played by robots.

Figure 2 depicts examples of mapping conversion between textual symbol and shape, limb posture, motion trajectory.

Firstly, let's look at the basic movements of human arms. When the human body is standing, the arms naturally hang down and are idle. When it is necessary to work, the arms are generally placed in front of the abdomen to interact with some objects.

The state of the arms naturally hanging down and slightly stretched can be depicted by T, which is similar to the shape of merging "Λ"(two arms) and "I"(trunk) together, with "I" covered by "Λ", (or similar to "/|\"), as shown in Figure 2 (a) (c).

The arms are naturally bent in front of the body, and the shape of the arms is a semi-enclosed circular posture, which is depicted by C, also named as C shape, as shown in Figure 2 (d).

Standing, holding a stone, lifting it up, and carrying it away, if you look at this series of actions carefully, you will inevitably see the basic modeling postures such as T shape and C shape, as shown in Figure 2 (a) (d).

3.2.1 Holding a Stone

"Holding a stone" motion posture decomposition involves the following set of gesture motion sequences, as shown in Figure 2 (a):

Firstly, the main body of the action should stand in front of the stone.

To depict a standing posture:

s: Depicts the head;

T: Depicts both arms located on both sides of the thighs, similar to the shape of merging "Λ"(two

arms) and "I"(trunk) together, with "I" covered by "Λ", (or similar to "/|\");

Then, both arms start to change the action:

O: Depicts both arms enclosing into a circular shape (to hold the stone);

Then describe the actions of the legs:

n: Depicts the legs;

e: Describes the legs spreading forward and sideways;

Among these symbols, "sTn" depicts the head, the trunk with both arms, and the legs respectively. Simply placing "T" below "s", and "n" below "T" , a standing posture of humans or other animals is depicted;

In this way, the action modeling sequence of "holding a stone" is obtained: "sTone". The action posture can also be restored from this symbol sequence.

3.2.2 Carrying Items

Motion decomposition of "carrying items", as shown in Figure 2 (d):

c: Depicts the bending of both arms to form a semi-circular shape;

a: Depicts the extension of the thumbs to stabilize the object being carried;

r: Describes the upward movement and trajectory of both arms;

r: Describes the further upward movement and trajectory of both arms;

y: Depicts both arms being raised above the head;

In this way, the carrying motion sequence is obtained: "carry". The motion posture can also be restored from this symbol sequence.

3.2.3 Holding a Child

For another example, when it comes to the concept of child, humans originally did not know how to express the concept of a child. If they just point to the child, squeaking and screaming, it is not easy to identify and understand. However, the interactive relationship that adults can have with a child is reflected in the ten fingers and both arms, which must be the action of holding a child. The actions that all mankind do to a child are the same, and they are born the same. The standard posture for holding a child with both arms is that one hand poses around the back, the other hand supports the hips, one hand is higher, and the other hand is lower, so that the child can remain in a sitting position.

How to depict the action of "holding a child"?

The basic shape is, as shown in Figure 2 (b):

The lower arm forms the C shape;

The higher arm forms the h shape, with elbow lifted;

ch is combined into the posture of holding a child with both arms;

Then, continue to add related action elements:

i: Depicts the small body of the child being held, (which can be represented by the little finger);

L: Describes the arms shaking and sliding downward;

d: Depicts the shape of the hand, with the palm lower than the wrist and the back of the hand facing forward;

Combined together, it is "chiLd". Among them, c, h, L and d these four symbols are the basic elements of limb shape and movement postures, being combined to form a relatively complex action sequence. And "i" represents the operated thing or object. After the action sequence is fixed and stylized, the action sequence of holding a child can be used to refer to the child. The action of holding a child has an extremely high repetition rate and recognition rate. Using this action sequence to represent a child has become a kind of thinking model.

The above-mentioned action of holding a child can be further expanded. By shaking the arms up and down, such a sequence is obtained: "chiLdren", in which "L" describes the arms sinking downward, and "r" describes the arms lifting upward. Thus, holding a child and shaking up-down is completed, and interaction with the child is achieved through this action.

3.2.4 Holding a Stick

Holding a stick is the most primitive and basic posture of human beings. Through mastering the use and swinging of sticks, human wisdom has accumulated, evolved, and upgraded.

The series of actions involved in "holding a stick" comprise, as shown in Figure 2 (c):

Maintaining a standing posture:

ST: The upper body is depicted in a posture with T placed below S.

S: Depicts the head;

T: Depicts the trunk with both arms located on both sides of the thighs, similar to the shape of merging "Λ"(two arms) and "I"(trunk) together, with "I" covered by "Λ", (or similar to "/|\");

Indicating the presence of an object (i.e. stick) in the hand:

i or I: Depicts an object being held in hand, (which can be represented by a wildcard or a small finger to signify being held);

Changing and forming a specific pose with both hands:

CK: Depicts the positions and shapes of the arms. One arm is bent and positioned in front of the abdomen, forming a "C" shape with the single arm. The other arm extends forward and upward, forming the upper single-arm shape of a "K".

CK: Alternatively, both arms can first be arranged in a "C" shape and then transformed into a "K" shape, achieving the same result and effect.

Thus, the action sequence for "holding a stick" is derived as "STICK".

This represents the standard posture of standing with both hands gripping a single stick. Using the posture of gripping a stick to represent the object stick aligns with the aforementioned thinking model.

These concepts: sTone, chiLd, STICK, are established by demonstrating their corresponding actions: "holding a stone", "holding a child", "holding a stick".

3.2.5 Standing Upright

In the early evolution of human beings, it took millions of years to evolve from walking on all fours to standing upright. How to express the action of standing upright?

Break down the posture of human standing, as shown in Figure 2 (a):

s: Depicts the head;

T: Depicts the trunk with both arms located on both sides of the thighs, similar to the shape of merging "Λ"(two arms) and "I"(trunk) together, with "I" covered by "Λ", (or similar to "/|\").;

Further refine the depiction of the small extremities:

a: Depicts the thumb extending;

d: Depicts the hand with the palm below the wrist and the back of the hand facing forward;

Describe the supporting base of the body, namely the legs:

n: Depicts the shape of the legs.

Among these symbols, "sTn" depicts the head, the trunk with both arms, and the legs respectively. Simply placing "T" below "s", and "n" below "T" , the human standing posture is depicted;

Thus, we can derive a sequence of poses for standing: "sTand". And vice versa, the action posture can also be reconstructed from this symbol sequence.

In this paper, the pair of upper case letter and lower case letter of the same shape, such as C c, K k, O o, P p, S s, U u, V v, W w, and so on, depict the same objects.

3.3 Motion Trajectory

The various poses of limbs or body parts such as arms, hands, legs, and feet can be mutually switched and changed. The process of change involves the movement and transformation of spatial positions, including: up, down, left, right, inward, outward, etc.

In this paper, the directional vector arrows are simplified and used to indicate the direction and trajectory of the movements of body parts.

r (Γ) is used to describe upward movement (simplified form of an upward arrow);

L is used to describe downward movement (simplified form of a downward arrow);

O is used to describe movements towards each other forming a circular shape;

For example, using the morphological symbols including C, r (Γ), L and O, can derive these combination sequence: "Cr (Γ)" "CL" "CO" "Cro (CΓO)" "CLO" "Cor (COΓ)" "COL", and the following related action forms can be described, (paying attention to the order of actions over time):

Both arms bend in front of the body to form a semi-circular shape, then move upwards, depicted by using the combination sequence "Cr (Γ)";

Both arms bend in front of the body to form a semi-circular shape, then move downwards, depicted by using the combination sequence "CL";

Both arms bend in front of the body to form a semi-circular shape, then both hands move towards each other to form a closed circular shape, depicted by using the combination sequence "CO";

Both arms bend in front of the body to form a semi-circular shape, then move upwards, then

both hands move towards each other to form a closed circular shape, depicted by using the combination sequence "Cro (CΓO)";

Both arms bend in front of the body to form a semi-circular shape, then move downwards, then both hands move towards each other to form a closed circular shape, depicted by using the combination sequence "CLO";

Both arms bend in front of the body to form a semi-circular shape, then both hands move towards each other to form a closed circular shape, then move upwards, depicted by using the combination sequence "Cor (COΓ)";

Both arms bend in front of the body to form a semi-circular shape, then both hands move towards each other to form a closed circular shape, then move downwards, depicted by using the combination sequence "COL";

3.4 Advantages and Effects

Language image natural model architecture (LINMA) aims to

build a natural corresponding mapping conversion system mechanism architecture between language text and visual graphics, images, video images;

analyze and summarize the main limbs modeling, terminal small limb modeling, and motion trajectories involved in modeling transformation;

use, express or understand specific meanings and intentions based on the sequence of modeling and motion trajectory.

This work conforms to the laws of nature and is simple, efficient, flexible, easy to use, highly extensible, widely applicable and highly versatile.

3.5 Application Scenarios

Some example embodiment application systems of this innovation:

Morphological Recorder: It captures data by employing visual cameras or wearable key node sensor, recognizes the forms, shapes and movements of body parts, and converts them into textual symbols.

Virtual Digital Human: According to input text symbols, it converts and generates a sequence of actions comprising forms, shapes, postures, gestures of body parts, and renders into realistic animations by using physics-based animation techniques or machine learning-driven motion synthesis; controls its virtual arms, hands, and body parts to demonstrate or virtually restore the original meaning or function carried by the input text symbols, through a display device.

Humanoid Robot: According to the input text symbols, it converts and generates a sequence of actions comprising forms, shapes, postures, gestures of body parts, and translates these actions into motor commands for the robot's actuators, controls its artificial arms, hands, and body parts to restore, demonstrate or archive the original meaning or function carried by the input text symbols.

4 Linma's Law of Thinking

4.1 Linma's First Law (Law of Head)

S is to depict the head, especially the side view head shape, or head-spine, or long hair, or being on the head.

4.2 Linma's Second Law (Law of Limb Posture)

T is to depict the shape of trunk and both arms, with both arms located on the sides of the body, extending obliquely downward, and both hands suspended in the outer areas of both thighs; there are some included angles between the arms and the trunk, basically similar to the shape of merging "Λ"(two arms) and "I"(trunk) together, with "I" covered by "Λ", (or similar to "/I\");

C is to depict the shape of the both arms, with both hands at the same height in front of the abdomen, and the arms are bent and curled into a semicircle (not closed into a circle), forming a C shape;

K is to depict the shape of trunk and both arms, with the arms positioned in front of the body, extending forward-upper and forward-lower respectively, maintaining a certain angle between the arms to create the K shape;

h is to depict the shape of trunk and a single arm, with the elbow raised upward, which can be flush with or higher than the abdomen;

W is to depict a double-arm shape, with the arms positioned on either side of the body, elbows bent, one arm in a V shape, and the both arms combined to form a W shape, with both hands at or above shoulder level;

U is to depict a double-arm shape, with the arms raised parallel upward or extended forward-upward to create the U shape;

V is to depict a double-arm shape, with the arms raised obliquely upward to present an V shape;

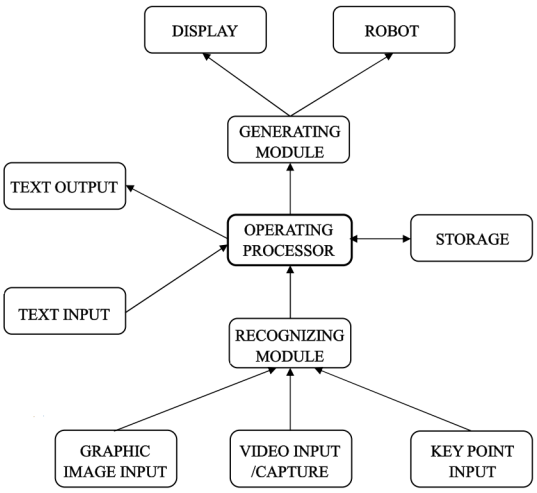

Figure 3: LINMA - example system architecture [21][22].

O is to depict a double-arm shape, with both arms bent into a closed circular shape;

J is to depict a single-arm shape, with the arm raised upward and the hand above the head;

y is to depict the shape of trunk and both arms, with both arms raised upward and extended to the sides, showing one hand at the same level or slightly higher than the other;

X is to depict a double-arm shape, with the arms in front of the body and the forearms crossed;

Z is to depict a double-arm shape, with the arms in front of the body and both forearms parallel up and down, right arm extended to the left, left arm extended to the right;

using asymmetrically combined double-arm shapes; wherein the aforementioned C T W K y and some other shapes are symmetrical double-arm shapes; when necessary, they can be simplified to single-arm shapes; one arm maintains the shape in the above shape, and the other arm is posed in another shape, comprising: Ch, Th, Wh, CT, TW;

CK is to depict a double-arm combination shape, with one arm posed in the shape of the C shape, and the other arm posed in the upper arm shape of the K shape;

n is to depict the two legs, especially both thighs;

g is to depict feet, or knee-shin-foot, or indicate the area under the feet;

ng is to depict the combination of parts: legs and feet.

4.3 Linma's Third Law (Law of Hand Shape)

(the front/forward in the following text refers to the direction in which the body stands upright and looks forward, and the opposite direction to the front is the back/backward);

p is to depict the hand shape, with the palm over the wrist and the palm facing forwards;

q is to depict the hand shape, with the palm over the wrist and the palm facing backwards;

d is to depict the hand shape, with the palm under the wrist and the palm facing backwards;

b is to depict the hand shape, with the palm under the wrist and the palm facing forwards;

(The aforementioned p q d b, comprising the finger shape being determined based on the function of the action;)

F is to depict a hand shape, with the palm over wrist and the fingers in a loose grip, fingertips pointing forward, thumb separated from the other fingers, thumb positioned below other fingers;

O is to depict a hand shape, with the fingers wrapped into a circular shape;

F is to depict a foot shape;

p is to depict a foot shape;

m is to depict the fingers, especially the middle three fingers, with the three fingertips pointing downward;

n is to depict two non-thumb fingers with the two fingertips pointing downward;

a is to depict the thumb, extended;

i is to depict the small finger, extended;

i or I is to refer to an object or thing, refer to the object involved in the action as a wildcard;

V is to depict the tongue, in a bending shape;

W is to depict the tongue, in a bending shape;

C is to depict the mouth, in a side-view opening shape;

O is to depict the mouth, in an opening shape;

O is to depict round objects;

D is to depict semicircular objects;

l is to depict line-shaped objects and things, also comprising straight legs and arms;

V,W is to depict fluctuation, turbulence, vibration.

4.4 Linma's Fouth Law (Law of Motion Trajectory)

r (Γ) is to describe moving upward;

L is to describe moving downward;

O is to describe closing inward, or circular motion, or moving backward, or moving leftward (note: it can have multiple meanings depending on the scene, the same below);

e is to describe separating outward, or parabola, or moving forward, or moving rightward;

5 System Architecture

An example system architecture is shown in Figure 3 [21][22], for the professional's reference. We'll describe the architecture in detail in other papers.

6 Conclusion

Actions and language are closely bound together in this paper. Action sequences could be coded and act as fundamental building block of language. Action language describes the world. Thus, the world has been mapped into action space. Language Image Natural Modeling Architecture (LINMA) contributes laws to construct the mapping engineering: law of head, law of limb posture, law of hand shape, law of motion trajectory. Language token could be action token. Action language could directly drive humanoid robot to perform actions.

In this paper, Language Image Natural Modeling Architecture (LINMA) is proposed, based on at least millions years of evolution and compression of interactive intelligence along with the real spatial world. It is interaction that bridges human and the real world.

In fact, the evolution of interactive intelligence has been driven by limbs of human or animals. Thus, interactive action based depiction of limbs could be critical component of human intelligence.

We propose LINMA's pattern of limbs, illustrating various shapes, gestures, postures and motion trajectories. Symbolization of these patterns can provide language building blocks. Arms, hands and fingers have played fundamental role in construction of human civilization. They are deserved to be depicted as a visible carrier of intelligence. Thus a very straightforward means is available for human being to explore the nature of intelligence.

Actually, our hands hold the secrets of language intelligence. It couldn't be simpler and more powerful.

By integrating image, graphics and video, depiction or/and symbolization of action sequence will achieve multimodal language.

By animating the depiction of action sequence, multimodal language could be achieved.

As a simple simulation of the real world, multimodal language may be the Pearl on the crown of language intelligence.

Multimodal language could serve as an action data set to empower wearable devices, virtual digital human and humanoid robot with embodied intelligence.

Through comprehensive integration of the intelligence chain involving visualization, interaction, depiction, simulation, symbolization, animation, compression, and optimization, etc., we may achieve ultimate universal language intelligence?

Acknowledgements I am grateful to Professor Liwei Chen (former chairman of Chinese Information Processing Society of China) for his supervision and inspiration in my early day's research in computational linguistics.

This paper is composed based on patent appplication documents [21][22].

References

[1] Ha David, Schmidhuber Jürgen Recurrent World Models Facilitate Policy Evolution *Advances in Neural Information Processing Systems (NeurIPS)*, 2018 1

[2] Schrittwieser Julian, Antonoglou Ioannis, Hubert Thomas, et al. Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model *Nature*, 2020, 588(7839): 377-382 1

[3] Hafner Danijar, Lillicrap Timothy, Fischer Ian, et al. Learning Latent Dynamics for Planning from Pixels *International Conference on Machine Learning (ICML)*, 2019: 2555-2565 1, 3

[4] Hafner Danijar, Lillicrap Timothy, Pasukonis Jan, et al. Mastering Diverse Domains through World Models *arXiv preprint arXiv:2301.04104*, 2023 2, 3

[5] LeCun Yann A Path Towards Autonomous Machine Intelligence *arXiv preprint arXiv:2206.02792*, 2022 2

[6] Bardes Adrien, Assran Mahmoud, Pinto Lerrel, et al. V-JEPA: Latent Video Prediction for Visual Representation Learning *arXiv preprint arXiv:2404.08471*, 2024 2

[7] Assran Mahmoud, Bardes Adrien, Fan Deyu, et al. V-JEPA2: Self-Supervised Video Models Enable Understanding, Prediction and Planning *arXiv preprint arXiv:2506.09985*, 2025 2, 3

[8] OpenAI Research Team Video Generation Models as World Simulators *OpenAI Technical Report*, 2024 2, 3

[9] Bruce Jake, Dennis Michael, Edwards Tom, et al. Genie: Generating Interactive Environments from a Single Image *arXiv preprint arXiv:2402.15391*, 2024 2

[10] NVIDIA Research Group Cosmos World Foundation Model Platform for Physical AI *arXiv preprint arXiv:2501.03575*, 2025 2, 3

[11] Wu Ziyi, Dvornik Nikita, Greff Klaus, et al. SlotFormer: Unsupervised Visual Dynamics Simulation with Object-Centric Models *International Conference on Learning Representations (ICLR)*, 2023 3

[12] Elsayed Gamaleldin F., Mahendran Aravindh, Van Steenkiste Sjoerd, et al. SAVi++: Towards End-to-End Object-Centric Learning from Real-World Videos *Advances in Neural Information Processing Systems (NeurIPS)*, 2022, 35: 29307-29320 3

[13] Li Jiahao, Huang Jiawei, Gong Jun, Wang Hao OrbiSim: World Models as Differentiable Physics Engines for Embodied Intelligence *arXiv preprint arXiv:2605.16395*, 2026 3

[14] Hansen Nicklas, Su Hao, Wang Xiaolong TD-MPC2: Scalable, Robust World Models for Continuous Control *arXiv preprint arXiv:2310.16828*, 2023 3

[15] Kang Boran, Yue Yutong, Lu Ruibo, et al. How Far Is Video Generation from World Model: A Physical Law Perspective *arXiv preprint arXiv:2411.02385*, 2024 3

[16] Shang Xinzhe, Mou Xiangyu, Lin Junfeng, Gao Hongsheng A Comprehensive Survey on World Models for Embodied AI *arXiv preprint arXiv:2510.16732*, 2025 4

[17] Kerbl Bernhard, Kopanas Georgios, Leimkühler Thomas, Drettakis George 3D Gaussian Splatting for Real-Time Radiance Field Rendering *ACM Transactions on Graphics*, 2023, 42(4): 1-14 3

[18] Wang Xinyu, Zhu Zirui, Huang Gaojie, et al. DriveDreamer: Towards Real-World Drive World Models for Autonomous Driving *European Conference on Computer Vision (ECCV)*, 2024: 372-389 3

[19] Bandaru Rohan Understanding World Models: Latent Dynamics, Generative Simulation and Embodied Planning *arXiv preprint arXiv:2411.14499*, 2025 4

[20] Fu Justin, Kumar Aviral, Nachum Ofir, et al. D4RL: Datasets for Deep Data-Driven Reinforcement Learning *arXiv preprint arXiv:2004.07219*, 2020 4

[21] Bing Lin. Methods and Devices for Language Image Natural Modeling Architecture, 2024. *CN Patent App. 2024102187162*. 10, 11

[22] Bing Lin. Methods and Devices for Language Image Natural Modeling Architecture, 2024. *US Patent App. 18/922,460*. 5, 10, 11