Language Image Natural Modeling Architecture (LINMA)

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

Abstract

In this paper, Language Image Natural Modeling Architecture (LINMA) is proposed, based on research over 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. LINMA language could serve as action dataset to empower wearable devices, virtual digital human and humanoid robot with embodied intelligence.

1 Introduction

In the 2020s, large language model (LLM) technology advanced rapidly, with ChatGPT leading a wave of explosive AI transformation and demonstrating significant technical potential in the field of artificial intelligence [1].

Multimodal AIGC (Artificial Intelligence Generated Content) technology aims to integrate and process various forms of information such as language text, graphics, images, audio, and video. Large language models play a central role in this integration, serving as the core for processing text and interfacing with modules that handle visual imagery, video, and audio information, thus building a complex multimodal information processing system [16].

Through big data and large-scale deep learning, these systems acquire extensive parameters, enabling them to handle a wide range of complex scenarios with intelligence [2] [4–13]. The entire learning and processing process requires high-performance hardware, with large models, big data, and massive scales necessitating the handling, scanning, and computation of vast amounts of information at a high-intensity computing power costs.

Large language models based on Transformer architectures[14][15] operate by extensively scanning existing human knowledge corpora to build a database of word association probabilities. Based on computed probability values, they predict and generate sequences of words. The advantage is that after training, it can cope with intellectual challenges in various aspects; however, a potential issue is that they may not fully understand the underlying logic of content, leading to occasional low-level errors and intellectual vulnerabilities [3].

Spatial intelligence is the ability to understand and manipulate visual and spatial information. It involves skills such as spatial reasoning, visualization, and orientation. People with high spatial intelligence are typically good at understanding visual representations and spatial relationships between objects, and be able to visualize objects in their mind's eye. This type of intelligence is important in fields such as architecture, engineering, art, and design.

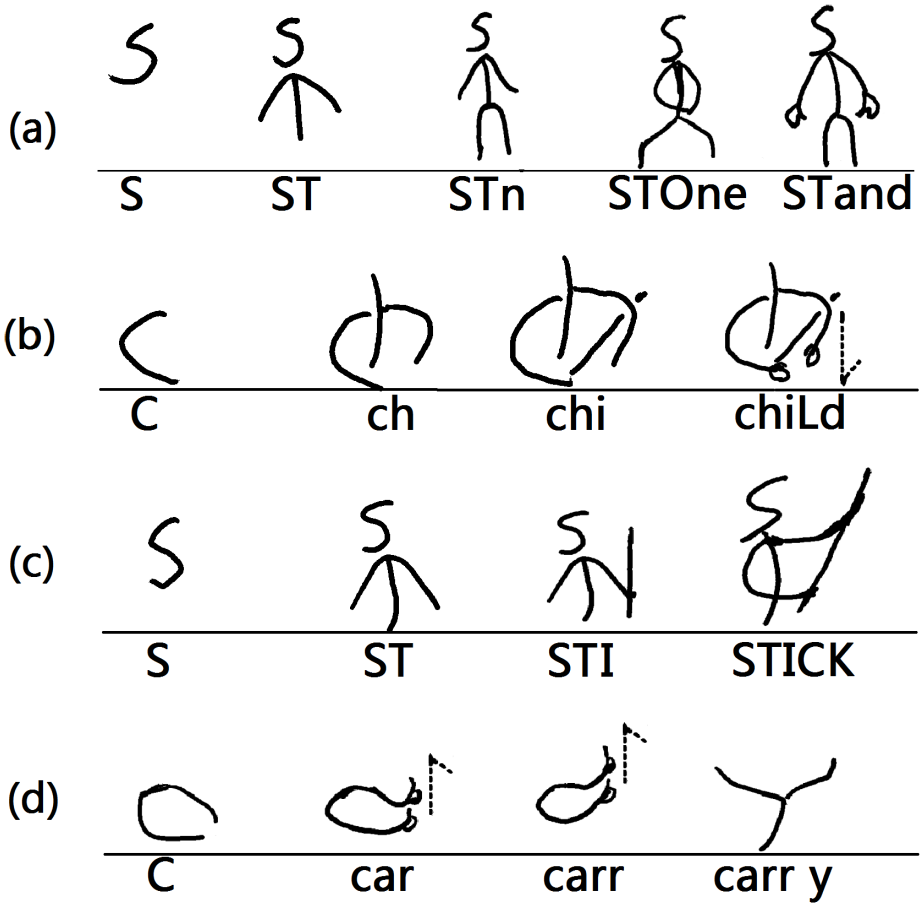

Figure 1: LINMA Thinking Model: symbolization of action sequence. (a) Just placing STn vertically, a depiction of human body basic structure is presented, which could be used further for deriving various actions, such as: Holding a STOne, STanding upright, etc. (b) Depiction of Holding a chiLd. (c) Depiction of Holding a STICK. (d) Depiction of Carrying items. [18]

Visual information contains a large volume of data, and various kinds of information are mixed, but only a portion of it is relevant or of interest. The challenge lies in, how to extract and obtain the information we care about from visual graphics, images, and video imagery? How to represent it? To what level of abstraction?

Intelligence compressing is to reduce the size of complexity of intelligence information while trying to retain its essential characteristics and value.

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 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 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 1 (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 1 (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 1 (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 1 (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 1 (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 1 (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 1 (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 1 (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;

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;)

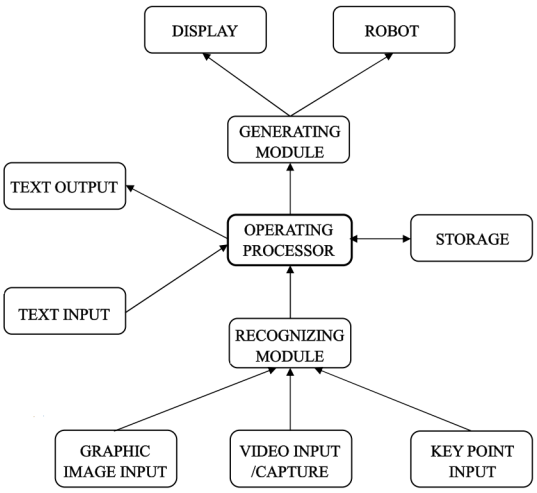

Figure 2: LINMA - example system architecture [17][18].

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 2 [17][18], for the professional's reference. We'll describe the architecture in detail in other papers.

6 Conclusion

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 [17][18].

References

[1] OpenAI. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774[cs.CL]*, 2023. 1

[2] LeCun Yann, Bengio Yoshua, Hinton Geoffrey. Deep learning. *Nature 521, 436-444(2015)* https://doi.org/10.1038/nature14539. 1

[3] Mirzadeh, I. et al. GSM-Symbolic: understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024. 1

[4] LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. (1990a). Handwritten digit recognition with a back-propagation network. In Touretzky, D. S., editor, *Advances in Neural Information Processing Systems 2*, pages 396–404. Morgan Kaufmann. 1

[5] LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] J. Deng, R. Socher, L.J. Li, K. Li, L. Fei-Fei. Imagenet: A large-scale hierarchical image database. *In 2009 IEEE conference on computer vision and pattern recognition,* pp.248-255. IEEE, 2009.

[8] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *In Advances in Neural Information Processing Systems*, 2012.

[9] Tomáš Mikolov, Kai Chen, Greg D. Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

[10] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112, 2014.

[11] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *NIPS*, 2014.

[12] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu and Kaiming He. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500, 2017. 1

[14] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan

N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017. 1

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 1

[16] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning Transferable Visual Models From Natural Language Supervision. *arXiv preprint arXiv:2103.00020*, 2021. 1

[17] Bing Lin. Methods and Devices for Language Image Natural Modeling Architecture, 2024. *CN Patent App. 2024102187162*. 8, 9

[18] Bing Lin. Methods and Devices for Language Image Natural Modeling Architecture, 2024. *US Patent App. 18/922,460*. 2, 8, 9