World Model & Embodied AI

Overview of Robot Control with World Model and VLA

Shang-Ching Liu

University of Hamburg
Faculty of Mathematics, Informatics and Natural Sciences
Department of Informatics
Technical Aspects of Multimodal Systems

July 2, 2025

Outline

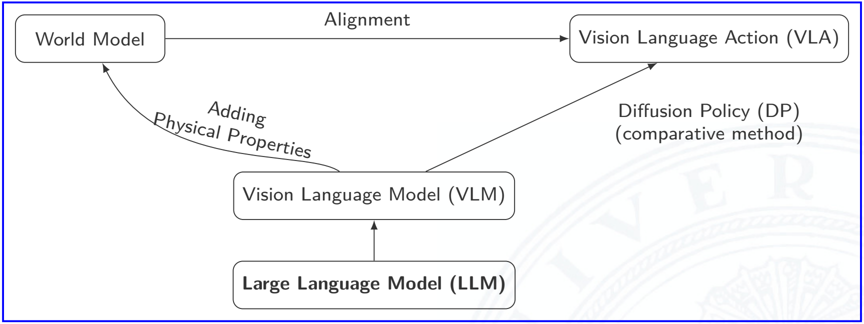
π0.5: VLA Model for Open-World Tasks

- **Key Capabilities of π0.5:**
  - **Open-World Generalization:** Executes tasks in unknown environments (e.g., different home scenes).
  - **Multimodal Training:** Joint optimization of images, language, and action trajectories in an end-to-end manner.
  - **Task Planning Ability:** Automatically decomposes complex instructions and generates action sequences.
- **Experimental Performance:**
  - ✔ Successfully completes multi-step complex tasks like cleaning kitchens and wiping surfaces.
  - ✔ Adapts flexibly to real-world changes in layout and target objects.
  - ✖ Has difficulty opening unfamiliar drawers or cabinets.
  - ✖ Currently handles only relatively simple prompts: e.g., repeatedly opening and closing drawers in long item-cleanup tasks.

▸ Click Here to Watch the Video

Relationship Between LLM and Robotics Actions

Foundation Models



Regression & Diffusion

One is predict the step by step action sequence, the other is to generate the whole action sequence in one step.

From LLM to MLLM

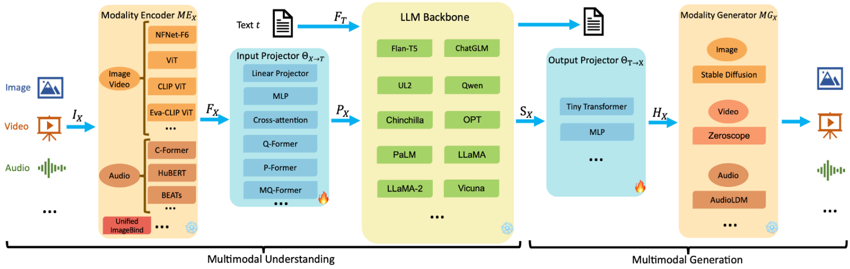
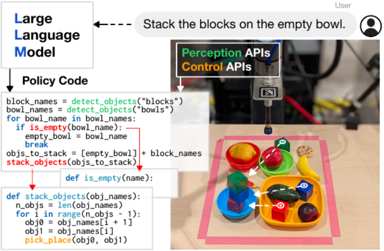


Figure: *Source:* "Mm-llms: Overview Architecture in MLLM" [8]

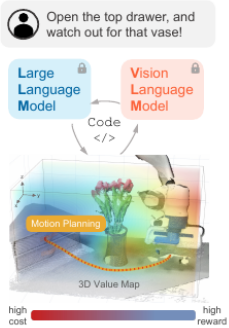
- **Providing Fundamental Functional Modules (APIs):** Clearly defined interfaces including Perception APIs and Control APIs.
- **High-Level Planning:** LLMs treat these APIs as available tools and use natural language to generate instruction flows or policies to accomplish tasks.



Source: Code as Policies (CaP) [5]

Extending Vision-Language Models (Generalization-Enhanced): VoxPoser

- **Vision-Language Model (VLM) as Backbone:** Equipped with **zero-shot generalization ability**, capable of understanding and handling relative spatial relationships such as "above", "below", "high", and "low".
- **Voxel Affordance-Based Spatial Representation:** Identifies key anchor locations in 3D space through voxel analysis, enhancing generalization and reliability in task execution.



Source: VoxPoser [4]

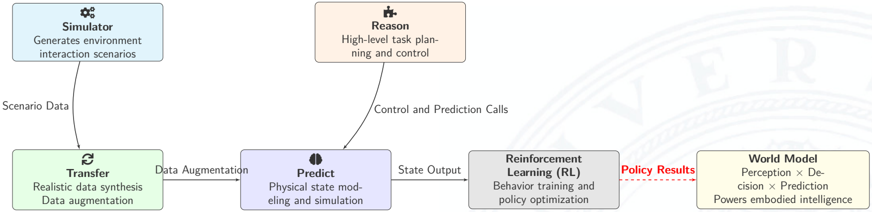
▸ Click Here to Watch the Demo Video

World Models (Enhanced with Physical Knowledge): Cosmos

- **Cosmos [10]** is a world model framework proposed by **NVIDIA**, consisting of three sub-models:
  - **Cosmos-Predict1:** A collection of general-purpose world foundation models used for modeling and predicting the physical world, with the ability to fine-tune for specific applications.
  - **Cosmos-Transfer1:** Helps bridge the perception gap between simulation and real-world environments by generating more realistic synthetic data, supporting more effective training of the Predict model.
  - **Cosmos-Reason1:** Incorporates physical attribute training data in the third stage of fine-tuning to enable deeper physical commonsense reasoning, generating embodied decisions and natural language explanations.

World Models (Enhanced with Physical Knowledge): Cosmos (cont.)

Synergy of the three: A comprehensive world modeling system for embodied intelligence



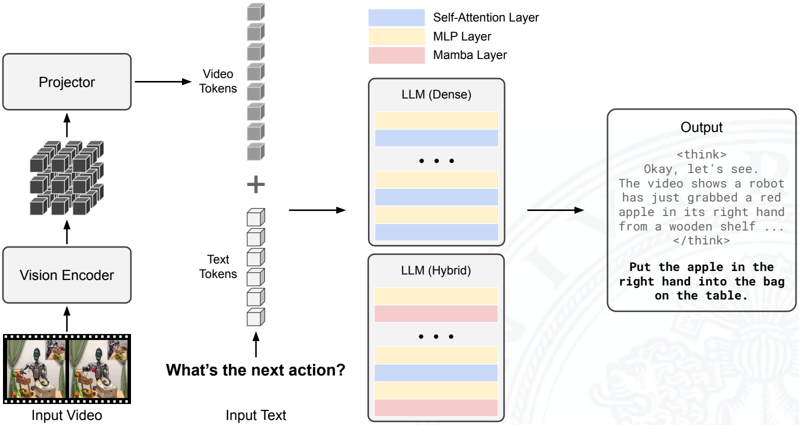


Figure: Cosmos-Reason1 Architecture Diagram [12]

Transfer Model (Cosmos Transfer1)

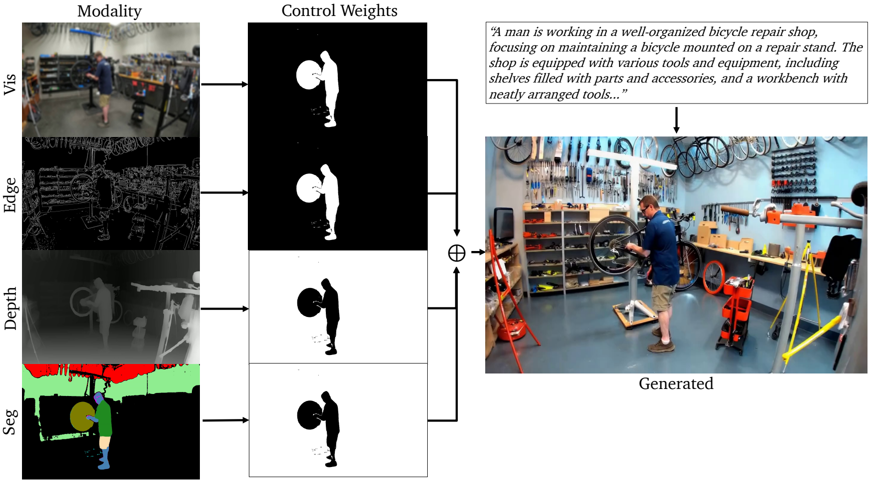


Figure: Concept Diagram of Cosmos Transfer1 [11]

Data Generation with Cosmos Transfer1

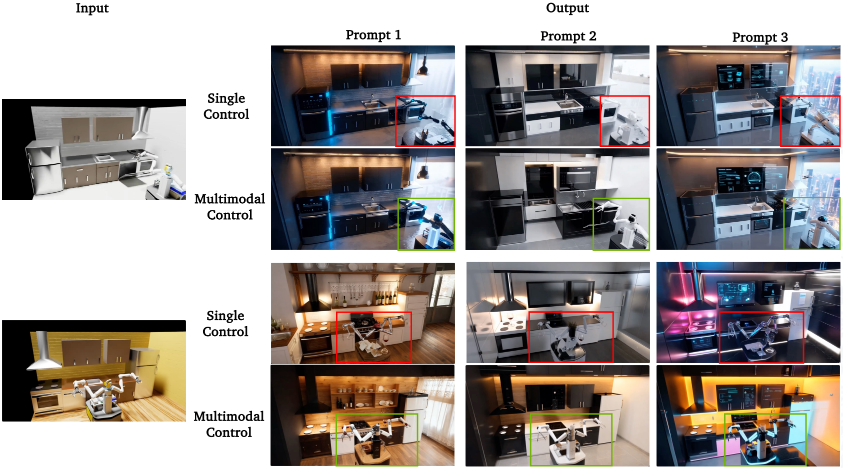


Figure: Workflow of Synthetic Data Generation Using Cosmos Transfer1 [11]

Prediction Model (Cosmos Predict1)

Overview of Inputs and Outputs in Cosmos Predict1 [10]

Animation: Simulated Output Sequence of the Prediction Model

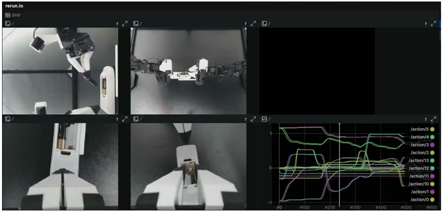
- Used for **RLAIF (Reinforcement Learning with AI Feedback)**.
- Provides environment simulation and feedback signals to help the reasoning model explore as many future paths as possible while filtering out infeasible options.

Interim Summary

- The evolution of Multimodal Large Language Models (MLLMs) has expanded generalization capabilities: the more modalities, the stronger the generalization.
- We can leverage the successful experiences of MLLMs to build world models required for embodied intelligence.
- A world model integrates:
  - High-level physical planning engines (for abstract decision-making and task decomposition);
  - Low-level reasoning and state prediction modules;
  - Scheduled model-based methods to support robots in executing long-horizon, complex tasks.

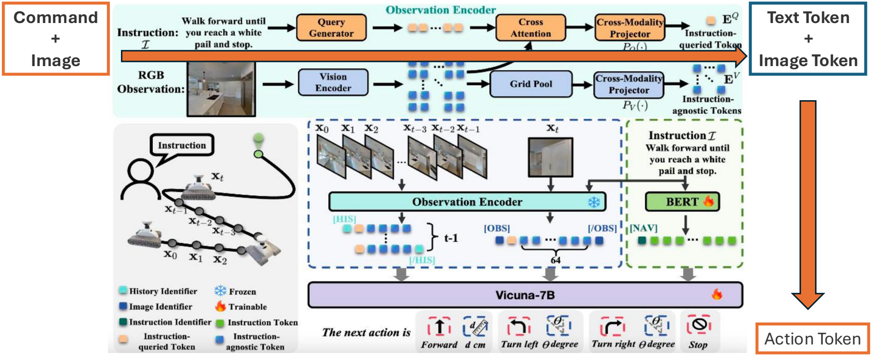
VLA

- **VLA (Vision-Language-Action)** is an extended form of Multimodal Large Language Models (MLLMs).
- Input: Multi-view visual scenes + instruction-based language descriptions
  Output: Rotation angles (in radians) for each joint servo.
- In robotic manipulation tasks, the VLA framework has been widely adopted:
  - **RDT** from Tsinghua University [6] (Robotics Diffusion Transformer)
  - **GR00T** from NVIDIA [13] (Generalist Robot)
  - The π series models from Physical Intelligence [15]



Overview

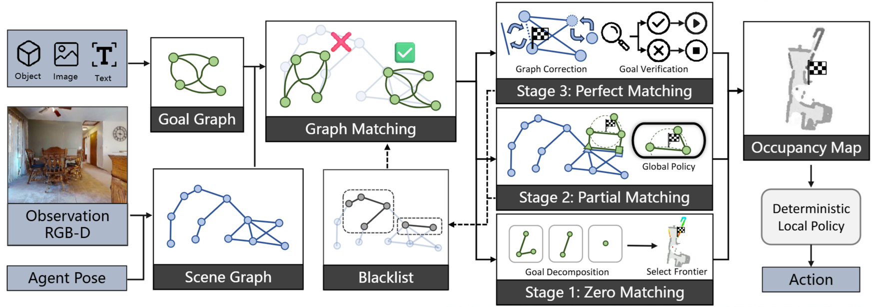
It integrates multimodal inputs to guide agents navigating through complex indoor environments. [9]



[Zhang et al. ArXiv 2024]

UniGoal Method

- **UniGoal** uses scene graphs as additional prior knowledge to improve navigation performance. [14]
  - Click Here to Watch Demo Video



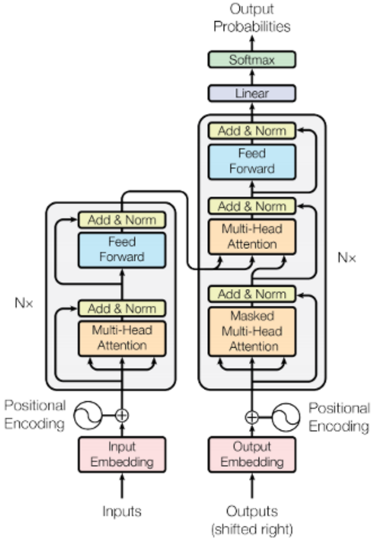
[Hang Yin et al. CVPR 2025]

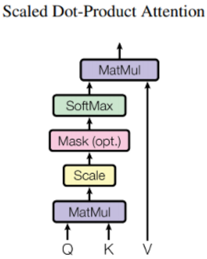
Core Idea

- Models relationships between words in a sequence using attention mechanisms.
- Fully based on attention — no RNNs or CNNs.

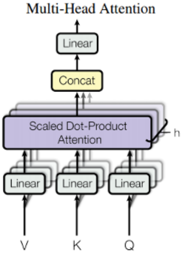
- **Basic Components**:
  - **Encoder**: Understands the input content.
  - **Decoder**: Generates the output.
  - The two are connected through the attention mechanism.

Transformer Architecture [1]





Self-Attention

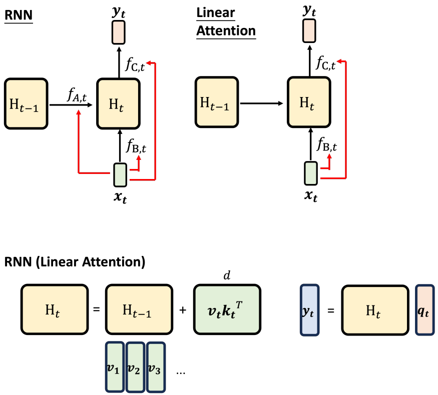


Multi-Head Attention

Understanding the Self-Attention Mechanism

- **Tokenizing:** Converts input text into tokens (numerical representations).
- **Self-Attention Task:**
  - Use input to formulate a query (Q).
  - Compare the query with keys (K) to measure relationships among words.
  - Apply a mask to exclude padding or future tokens (if decoding).
  - Normalize using SoftMax to compute attention weights.
  - Multiply attention weights with values (V) to obtain new contextualized embeddings.
- **Multi-Head Attention:** Combines multiple attention heads to learn different aspects of the input context.

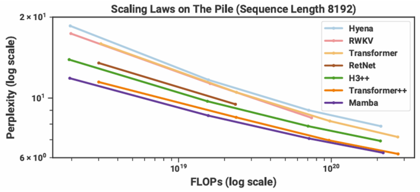
Introduction to Linear Attention [2]



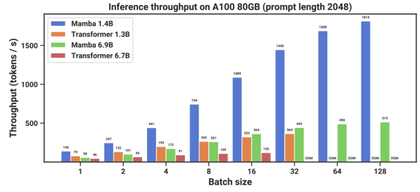
- Inspired by sequence processing in RNNs.
- Reduces from $O(N^2)$ to $O(N)$.
- **Advantages**: Efficient and suitable for modeling long sequences.
- **Limitations**: Lacks reflection (reverse context integration), which restricts performance.
- **Further Development**:
  - In 2023, Mamba [3] was proposed, combining state space models to address limitations.

Mamba and Mamba2

- Mamba reintroduces the reflection mechanism on top of linear attention.
- Mamba2 further addresses efficiency bottlenecks in parallel training.
- The Mamba series significantly outperforms traditional Transformers in terms of speed.
- It also surpasses Transformers in performance across multiple tasks.



Model Speed Comparison



Model Accuracy Comparison

Loss Function and Gradient:

$$L_t(H) = \frac{1}{2}\|H\mathbf{k}_t - \mathbf{v}_t\|^2, \quad \nabla L_t(H_{t-1}) = (H_{t-1}\mathbf{k}_t - \mathbf{v}_t)\mathbf{k}_t^\top$$

Update Derivation:

$$\begin{aligned}
\text{Start:}\quad H_t &= H_{t-1} + \mathbf{v}_t\mathbf{k}_t^\top \\
\text{Rewrite:}\quad H_t &= H_{t-1} - \mathbf{v}_{t,\text{old}}\mathbf{k}_t^\top + \mathbf{v}_t\mathbf{k}_t^\top \\
\text{with:}\quad \mathbf{v}_{t,\text{old}} &= H_{t-1}\mathbf{k}_t \\
\text{Add LR:}\quad H_t &= H_{t-1} - \beta_t\mathbf{v}_{t,\text{old}}\mathbf{k}_t^\top + \beta_t\mathbf{v}_t\mathbf{k}_t^\top \\
\text{Substitute:}\quad H_t &= H_{t-1} - \beta_t H_{t-1}\mathbf{k}_t\mathbf{k}_t^\top + \beta_t\mathbf{v}_t\mathbf{k}_t^\top \\
\textbf{Final:}\quad H_t &= H_{t-1} - \beta_t(H_{t-1}\mathbf{k}_t - \mathbf{v}_t)\mathbf{k}_t^\top
\end{aligned}$$

Gradient Descent Structure:

$$\underbrace{H_t}_{\text{updated}} = \underbrace{H_{t-1}}_{\text{old}} - \underbrace{\beta_t}_{\text{LR}} \cdot \underbrace{(H_{t-1}\mathbf{k}_t - \mathbf{v}_t)\mathbf{k}_t^\top}_{\text{gradient}}$$

Goal: Improve H so that projection of $\mathbf{k}_t$ approximates $\mathbf{v}_t$ better.

Liquid Neural Networks (LNN)

- Inspired by the Reservoir Computing architecture.

Advantage

Most low-weight neurons can self-suppress under input variation and are excluded from computation, improving energy efficiency.

- **Limitation:** Scalability and performance optimization of the network remain active research challenges.

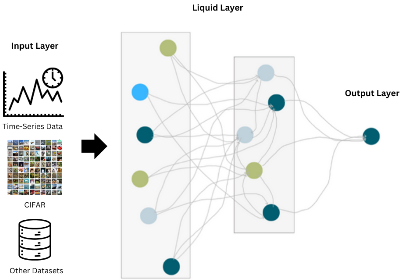


Figure: **Figure: Schematic Diagram of the LNN Architecture**

Summary: Comparison of Transformer and Its Successors

Model	**Complexity**	**Capability**	**Efficiency**	**Performance**
Transformer	$O(N^2)$	Moderate	Medium	Baseline
Linear Attention	$O(N)$	Stronger	High	Close to Transformer
Mamba	$O(N)$	Strong	Very High	Often Outperforms Transformer
TTT	$O(N)$	Strong	Very High	Outperforms Mamba
LNN	$O(N)$ (Dynamic)	Very Strong	Extremely High	Leads in some tasks

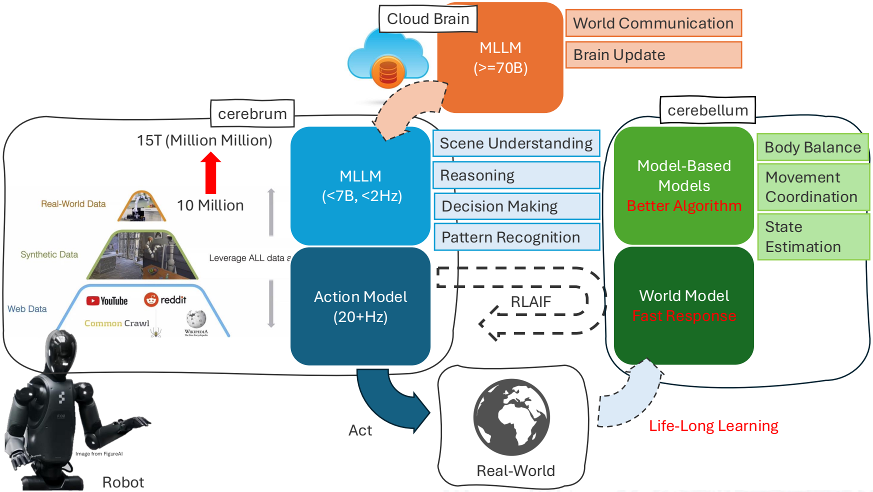
Meta

DeepSeek

MiniMax

Liquid AI

Challenges Faced by Robots



Reference

[1] Ashish Vaswani et al. "Attention is all you need". In: *Advances in neural information processing systems* 30 (2017).

[2] Angelos Katharopoulos et al. "Transformers are rnns: Fast autoregressive transformers with linear attention". In: *International conference on machine learning*. PMLR. 2020, pp. 5156–5165.

[3] Albert Gu and Tri Dao. "Mamba: Linear-time sequence modeling with selective state spaces". In: *arXiv preprint arXiv:2312.00752* (2023).

[4] Wenlong Huang et al. "Voxposer: Composable 3d value maps for robotic manipulation with language models". In: *arXiv preprint arXiv:2307.05973* (2023).

Reference (cont.)

[5] Jacky Liang et al. "Code as policies: Language model programs for embodied control". In: *2023 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE. 2023, pp. 9493–9500.

[6] Songming Liu et al. "Rdt-1b: a diffusion foundation model for bimanual manipulation". In: *arXiv preprint arXiv:2410.07864* (2024).

[7] Songlin Yang et al. "Parallelizing linear transformers with the delta rule over sequence length". In: *arXiv preprint arXiv:2406.06484* (2024).

[8] Duzhen Zhang et al. "Mm-llms: Recent advances in multimodal large language models". In: *arXiv preprint arXiv:2401.13601* (2024).

Reference (cont.)

[9] Jiazhao Zhang et al. "Navid: Video-based vlm plans the next step for vision-and-language navigation". In: *arXiv preprint arXiv:2402.15852* (2024).

[10] Niket Agarwal et al. "Cosmos world foundation model platform for physical ai". In: *arXiv preprint arXiv:2501.03575* (2025).

[11] Hassan Abu Alhaija et al. "Cosmos-Transfer1: Conditional world generation with adaptive multimodal control". In: *arXiv preprint arXiv:2503.14492* (2025).

[12] Alisson Azzolini et al. "Cosmos-reason1: From physical common sense to embodied reasoning". In: *arXiv preprint arXiv:2503.15558* (2025).

[13] J Bjorck Nvidia et al. "Gr00t n1: An open foundation model for generalist humanoid robots". In: *ArXiv, abs/2503.14734* (2025).

Reference (cont.)

[14] Hang Yin et al. "Unigoal: Towards universal zero-shot goal-oriented navigation". In: *arXiv preprint arXiv:2503.10630* (2025).

[15] Kevin Black et al. "π0: A vision-language-action flow model for general robot control, 2024". In: *URL https://arxiv.org/abs/2410.24164* ().