STRUCT Group Paper Reading

An Introduction to World Model and Beyond

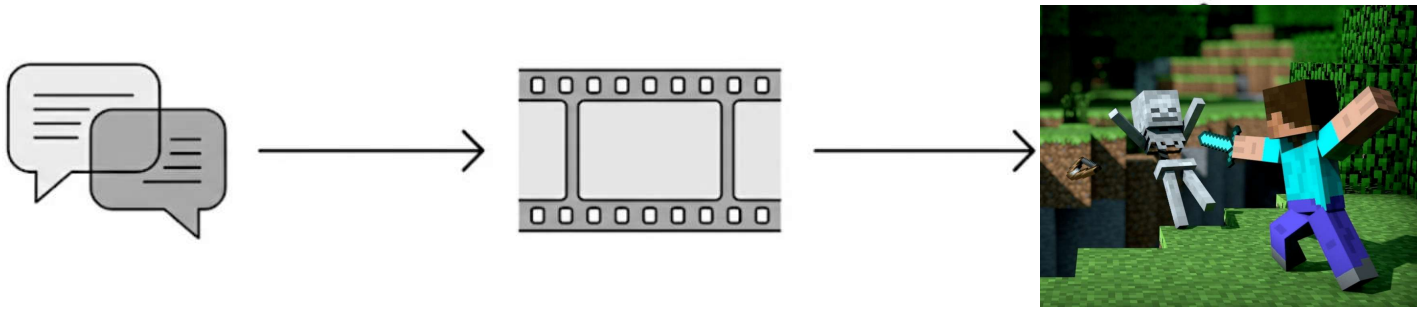
Day 2
World Model Study Group @ STRUCT

Presented by ZhangJiahang, KuangHaowei, GaoWenshuo
2026.2.2

Content

- Generative World Models
  - Overview
  - The Frontiers of World Models in Industry
  - New Theories and Architectures in Academia
- Generation Improvement with Understanding Model
  - Prompt Enhancement
  - Chain-of-Thought for Generation

Generative World Models



LLMs
Predicting the Next Word

Video Gen
Predicting the Next Pixel
Passive / Hallucination-Prone

World Models
Predicting the Next State
Interactive / Consistent

- LLMs alone are insufficient to achieve AGI
- Li Fei-Fei: AGI is incomplete without **spatial intelligence**
- Yann LeCun: No need to generate pixels, generate **states**

Generative World Models

Base Stone: Video Generation Models

ADDING:

1. Long-term Memory
   - —3D modeling
   - —Compressed Feature/Representation

2. Understanding the World
   - —Physics understanding
   - —Realism
   - —Future Prediction & Decision

3. Interactivity
   - —WASD

Generative World Models

The Researchers

Google DeepMind (Genie 3)

Meta (V-JEPA 2)

Focus on Theory & Realism.

The Builders

NVIDIA (Cosmos)

Focus on Robotics & Physical AI.

The Creators

World Labs (Marble)

Runway (GWM-1)

Focus on 3D Assets & Creative Tools.

The Open Frontier

Tencent (HY-World 1.5)

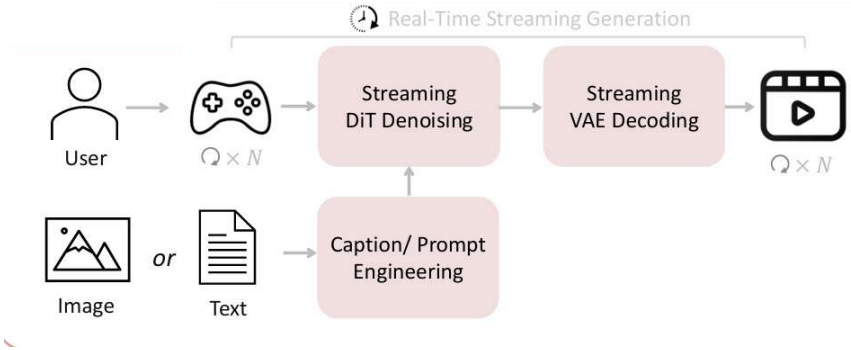
Odyssey

LingBot

Focus on Open Source & Speed.

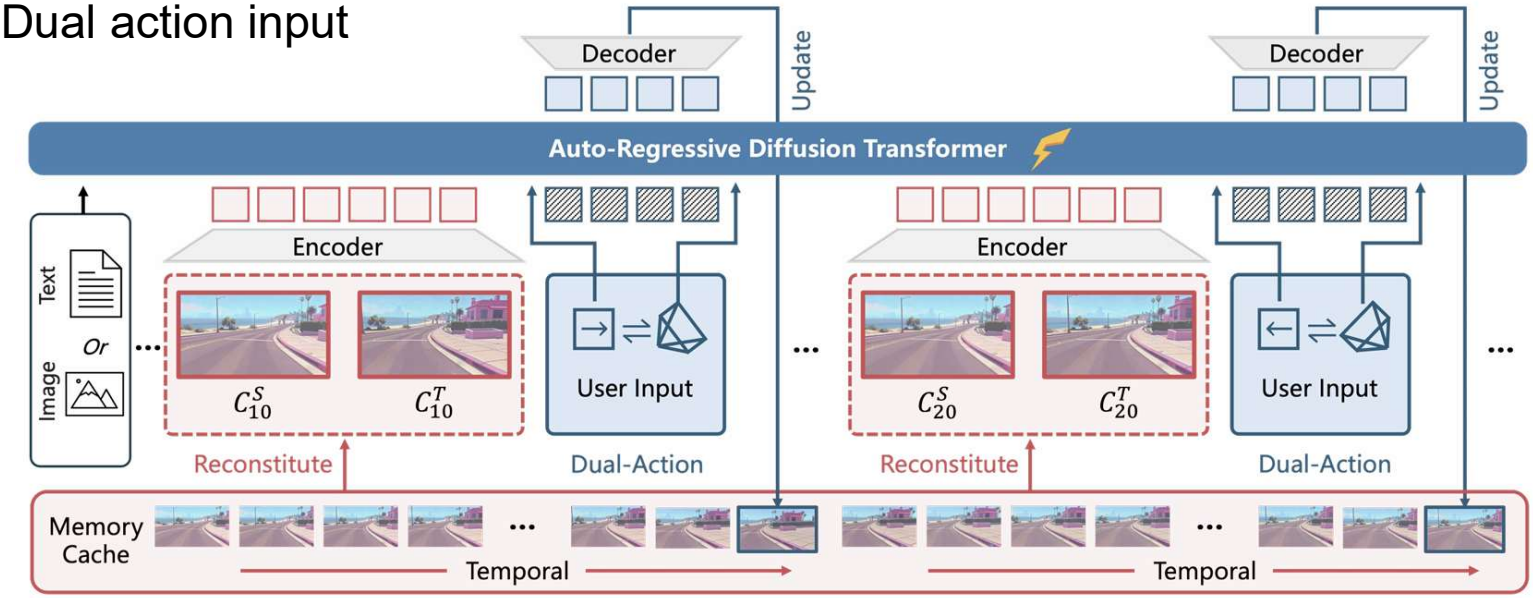
HY-World 1.5 Tencent

- Extending Video Model
  - Input:
    - Current action
    - Last chunk
  - Output:
    - Current chunk



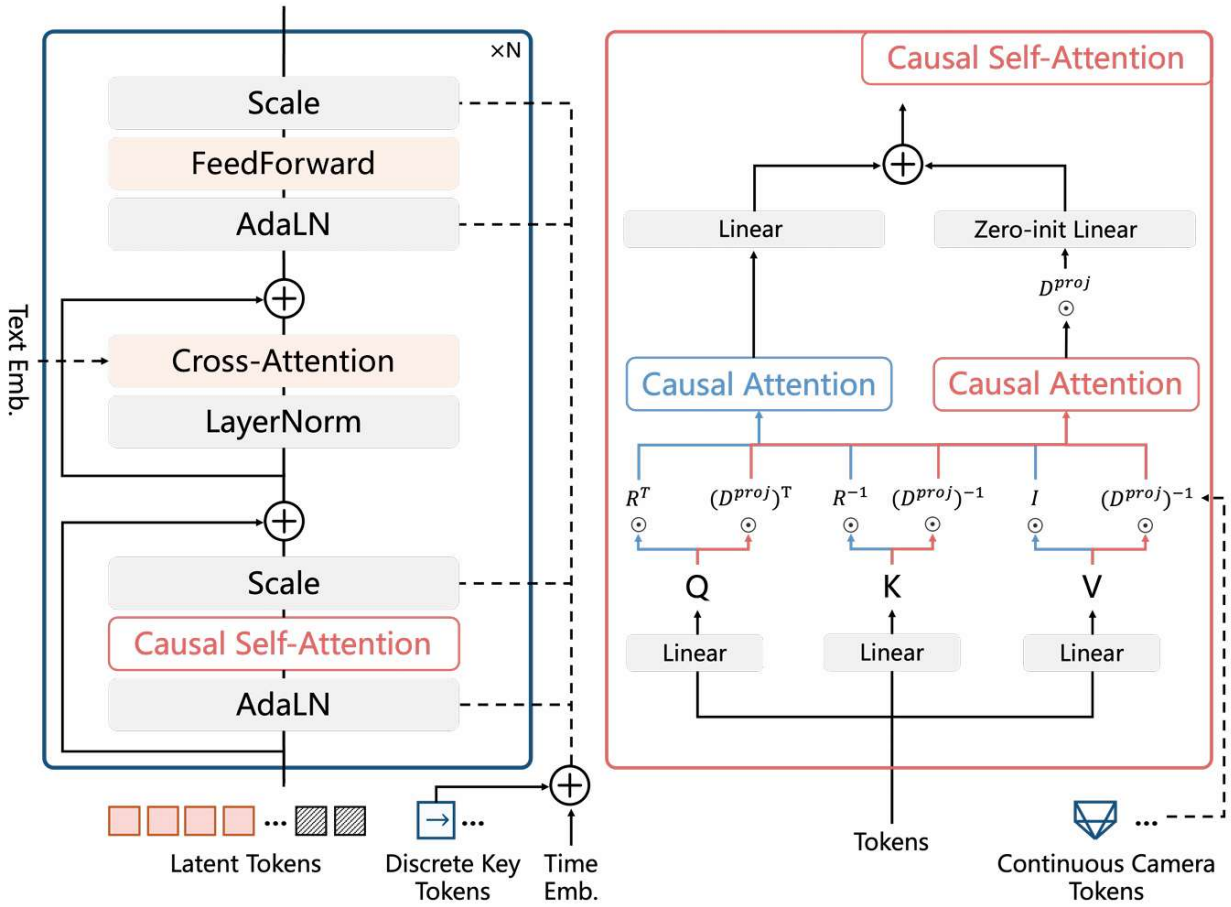
HY-World 1.5

- Extending Video Model
  - AR DiT architecture
  - Dual action input



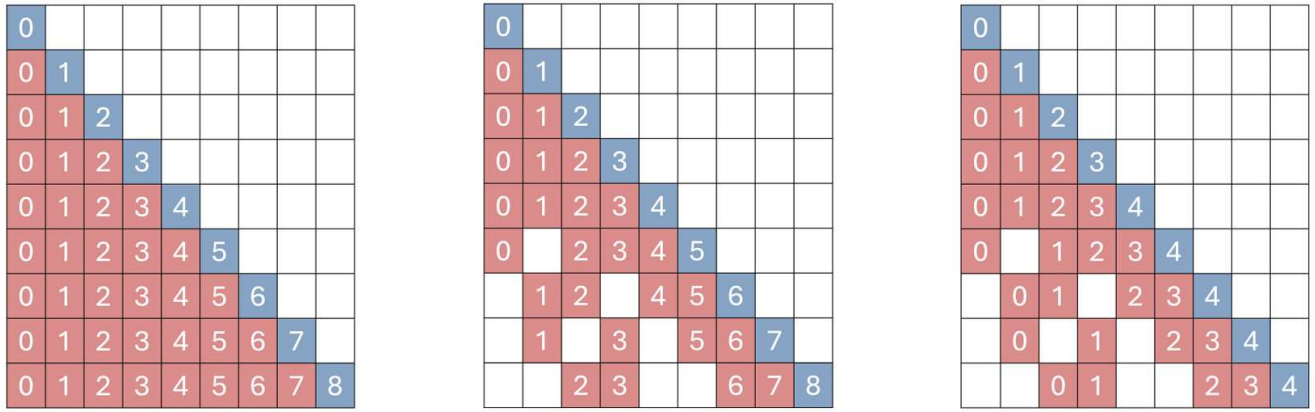
HY-World 1.5

- Extending Video Model
  - AR DiT architecture
  - Dual action input



HY-World 1.5

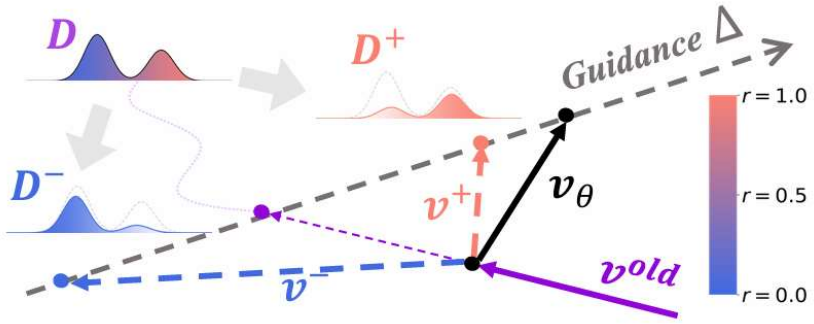
- Long-Term Memory
  - Causal self-attention
  - Find the most geometry-relative chunks in the past
  - Pull them closer in RoPE



(a) Full context (b) Absolute indices (c) Relative indices

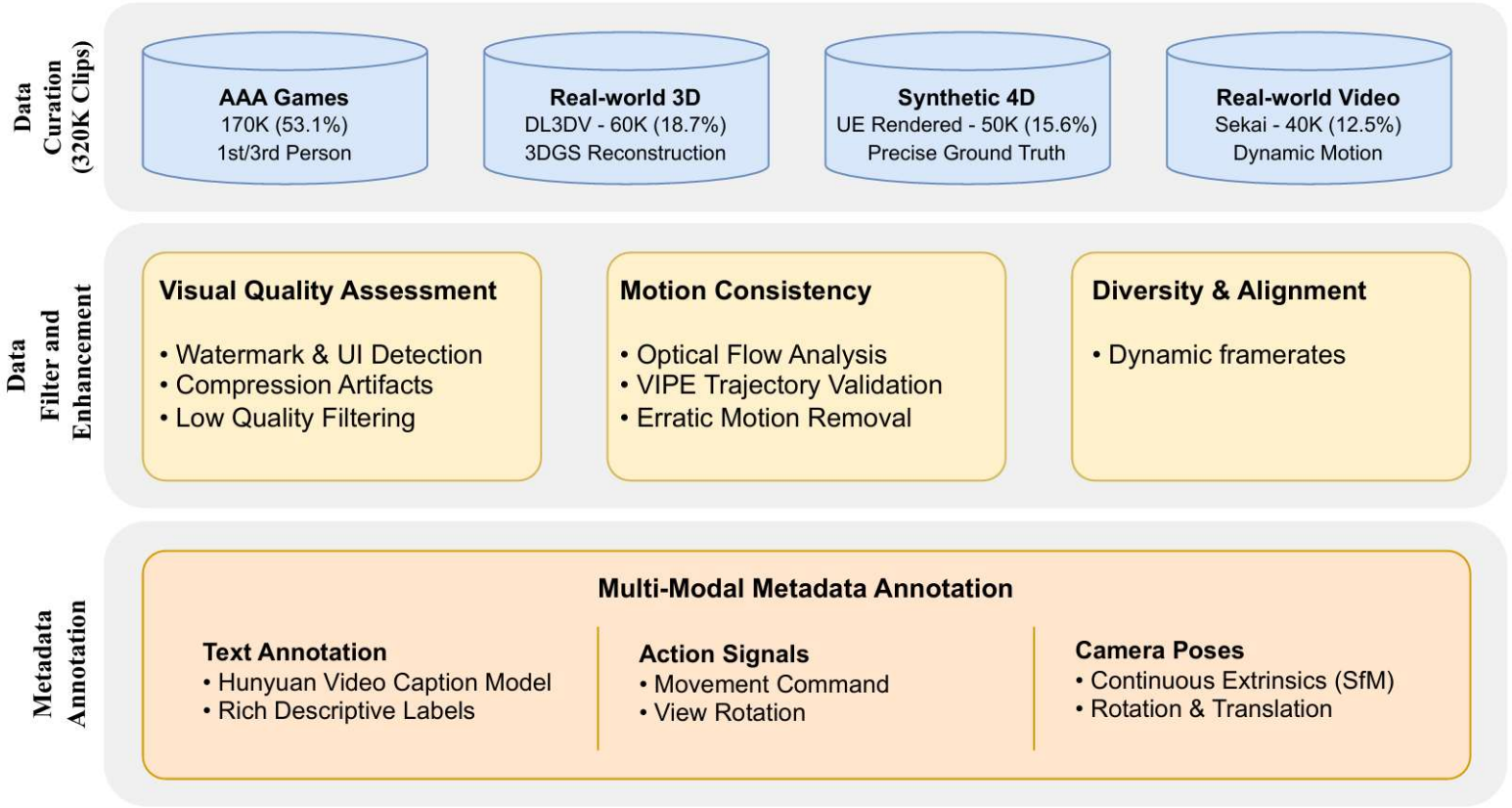
HY-World 1.5

- Reinforcement Learning
  - Rewards:
    - Following score: video ↔ action
    - Visual quality score
  - Algorithm:
    - DiffusionNFT



HY-World 1.5

- Data



HY-World 1.5

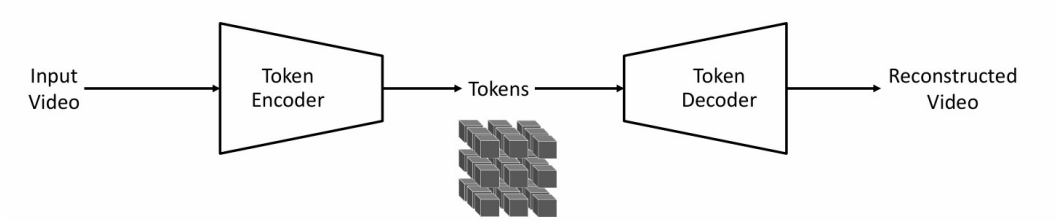
- Conclusion
  - Open-source
  - 24 FPS streaming
  - Keyboard & mouse control

Cosmos

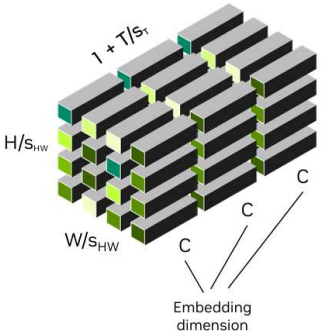
- World Models for Physical AI
- Components:
  - Tokenizer
  - Diffusion model
  - AR model
- Similar to common video models in architectures

Agarwal, et al. Cosmos world foundation model platform for physical ai. arXiv:2501.03575, 2025.

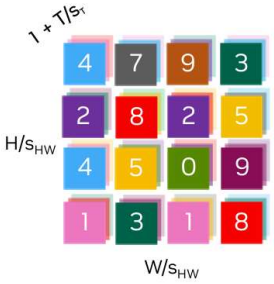
Cosmos



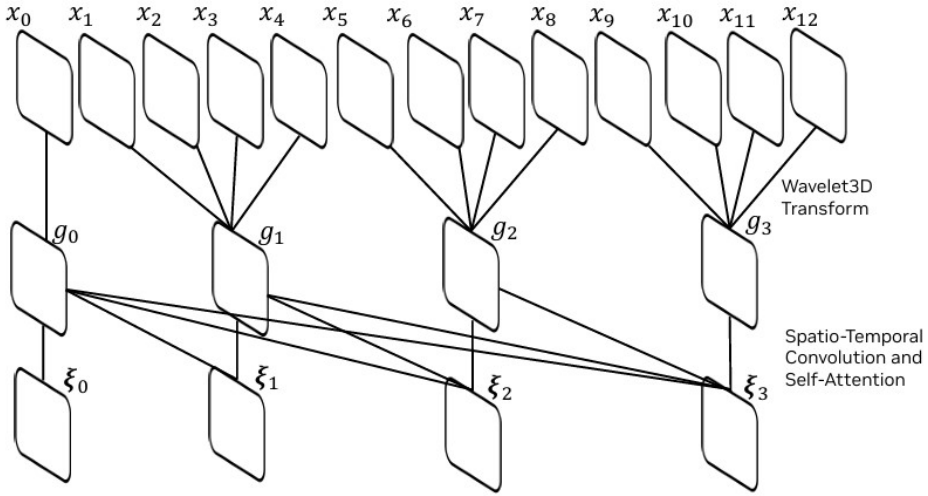
- Tokenizer:
  - Continuous version for diffusion
  - Discrete version for AR
  - Causal self-attention



Continuous

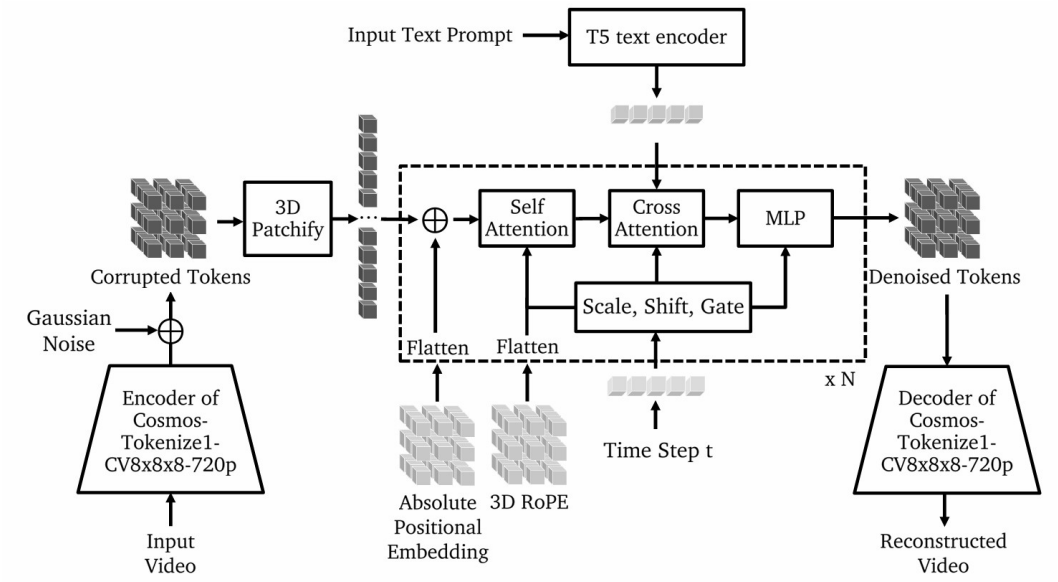


Discrete



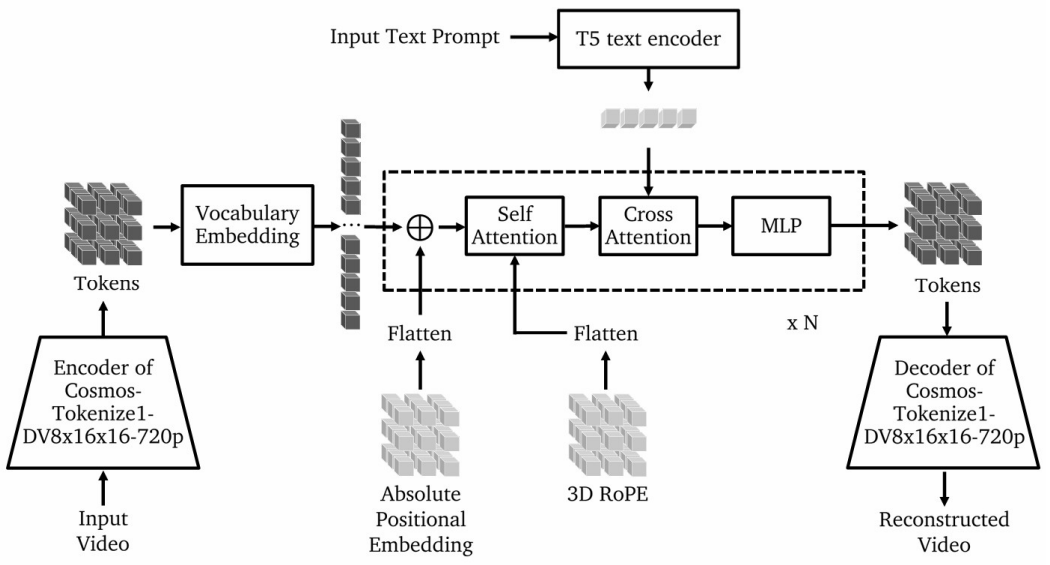
Cosmos

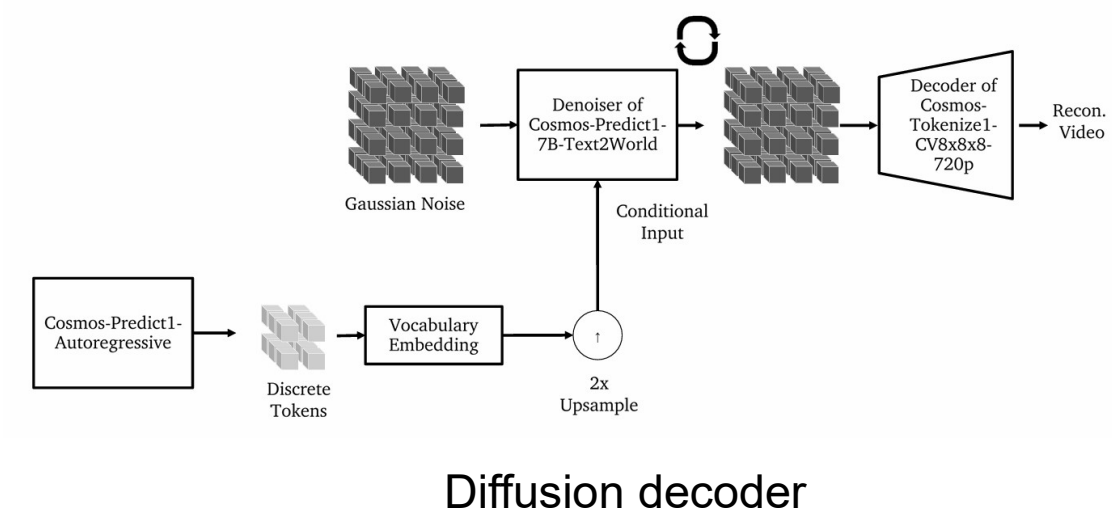
- Diffusion Model
  - Conditioned on past frames: concatenation in temporal dimension



Cosmos

- AR Model
  - Discrete tokens: decode using diffusion decoder





Diffusion decoder

Cosmos

- Pre-training dataset

1. Driving (11%),
2. Hand motion and object manipulation (16%),
3. Human motion and activity (10%),
4. Spatial awareness and navigation (16%),
5. First person point-of-view (8%),
6. Nature dynamics (20%),
7. Dynamic camera movements (8%),
8. Synthetically rendered (4%), and
9. Others (7%).

Cosmos

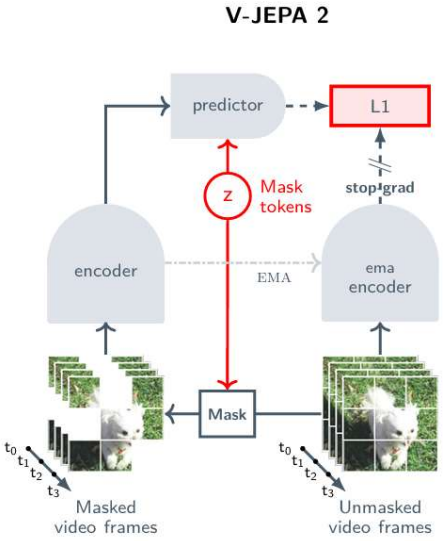
- Post-training for Downstream Application
  - Instruction-based finetuning for robots
  - Action-based finetuning for autonomous driving

Marble

- Gaussian-Splatting Based
  - 3D static scene generation / editing
  - Exportable assets

V-JEPA 2

- DiT: Reconstructing pixels
- V-JEPA 2: Predicting semantics in masked area



Yann LeCun:
No need to generate pixels, generate **states**

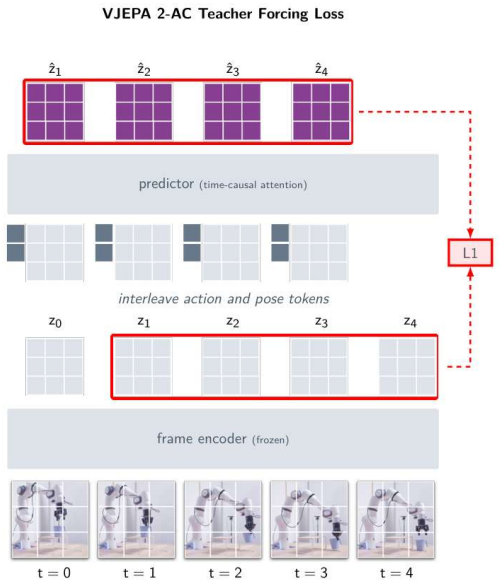
$$\text{minimize}_{\theta,\phi,\Delta_y} \quad \|P_\phi(\Delta_y, E_\theta(x)) - \text{sg}(E_{\overline{\theta}}(y))\|_1$$

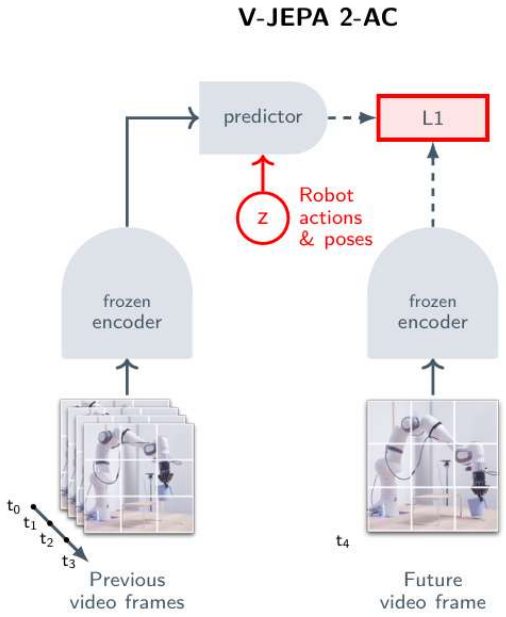
V-JEPA 2

- V-JEPA 2-AC: Conditioning on actions
  - Teacher forcing loss: autoregressive learning

$$\mathcal{L}_{\text{teacher-forcing}}(\phi) := \frac{1}{T}\sum_{k=1}^{T}\|\hat{z}_{k+1} - z_{k+1}\|_1 = \frac{1}{T}\sum_{k=1}^{T}\left\|P_\phi\left((a_t, s_t, E(x_t))_{t\leq k}\right) - E(x_{k+1})\right\|_1$$

  - Limitation: error accumulation



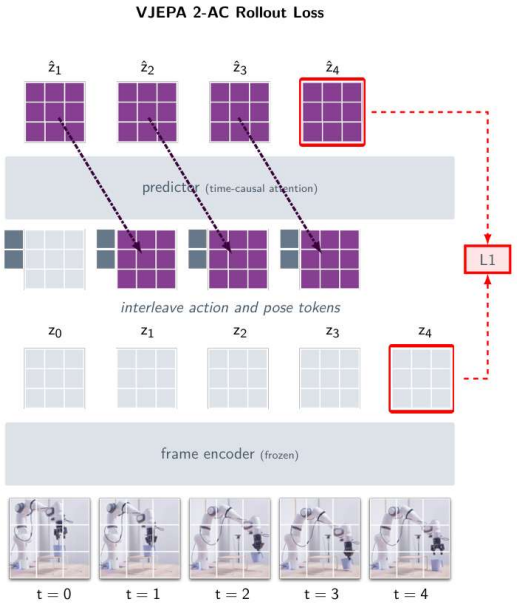


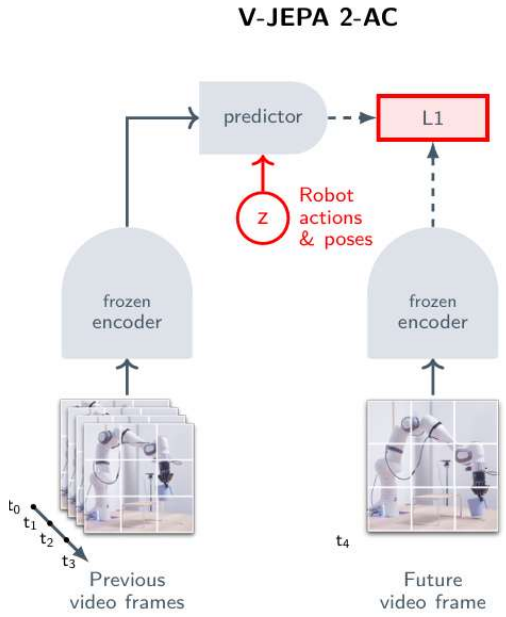
V-JEPA 2

- V-JEPA 2-AC: Conditioning on actions
  - Rollout loss: consecutively predict T times

$$\mathcal{L}_{\text{rollout}}(\phi) := \|P_\phi(a_{1:T}, s_1, z_1) - z_{T+1}\|_1$$

  - Gradient propagates through time steps (BPTT)

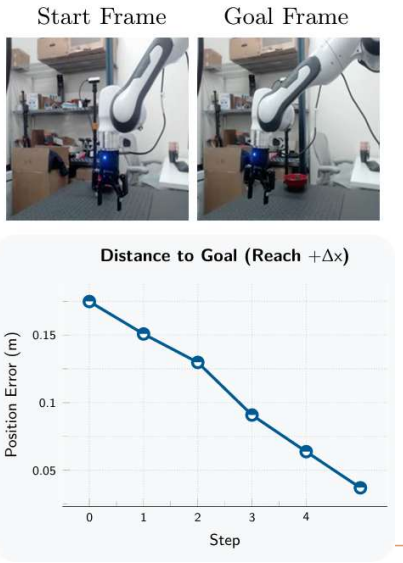


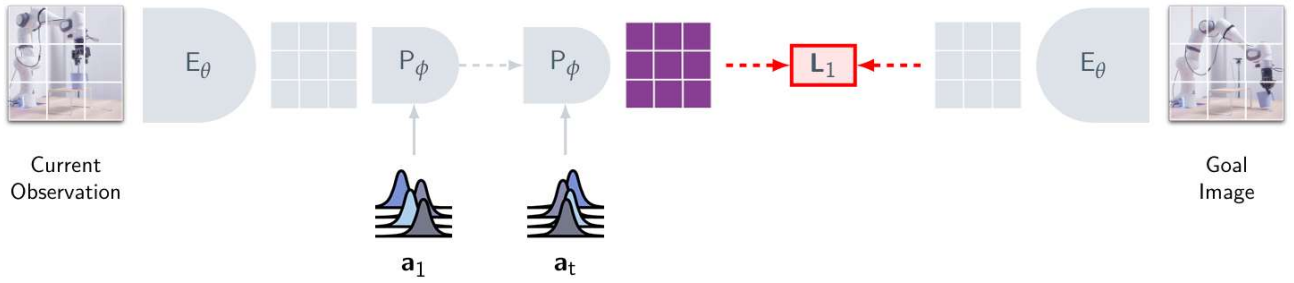


V-JEPA 2

- V-JEPA 2-AC: Guidance on Robot Action Planning

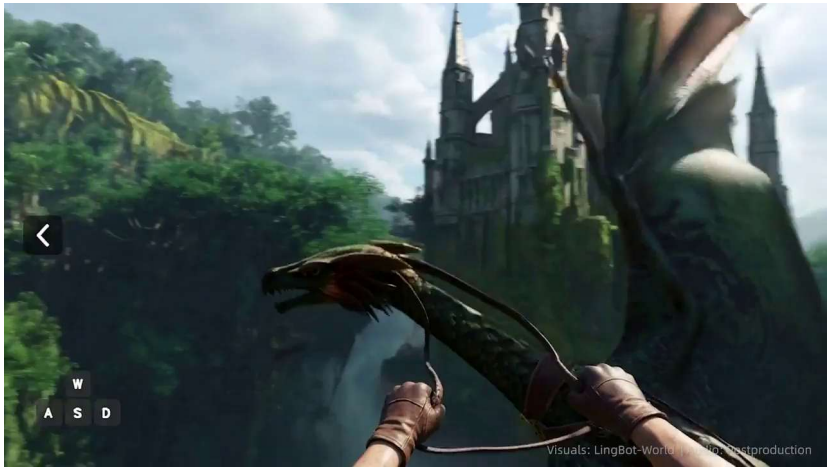
$$\mathcal{E}(\hat{a}_{1:T};\ z_k, s_k, z_g) := \|P(\hat{a}_{1:T}; s_k, z_k) - z_g\|_1$$

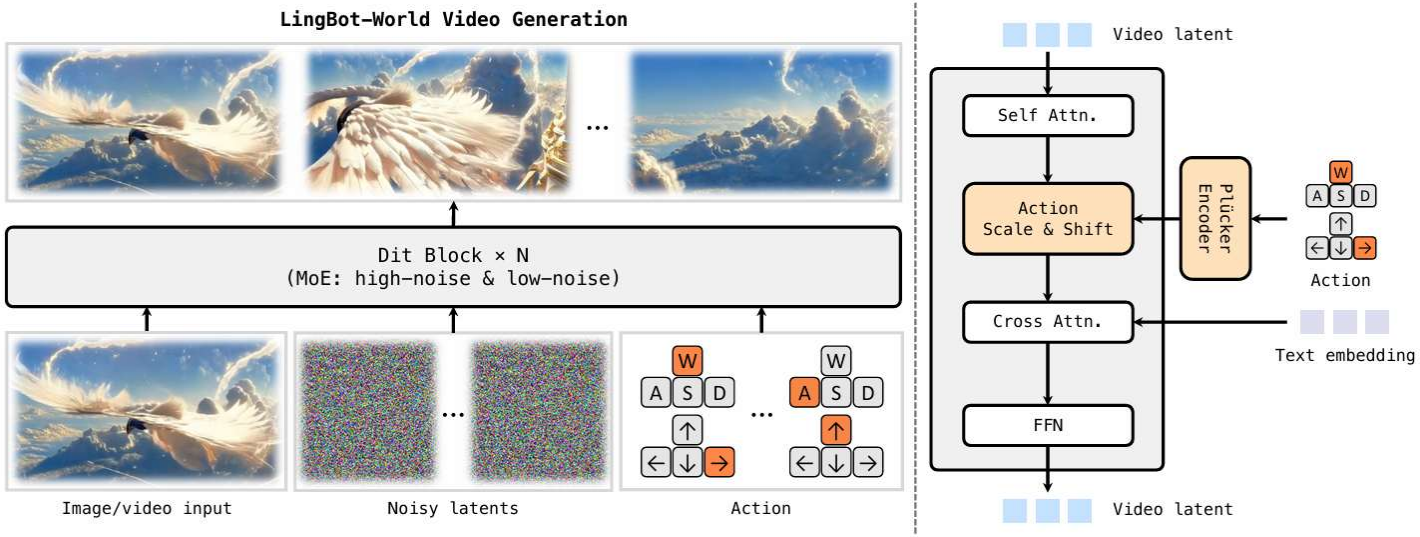




LingBot-World

- Extending Video Model (Wan2.2)
  - MoE: high-noise expert & low-noise expert





Genie 3

- State-of-the-art World Model
  - 720p @ 24 FPS
  - Learned physics
  - Minutes of consistency
  - *！蚂蚁后面5s有飞鸟*

World Models: The Future

- **True real-time AAA quality:** Real-time 4K @ 60+ FPS photorealism
- **Extended persistence:** Long-horizon world state (hours to days)
- **Complexity:** Multi-agent interactions & complex NPC ecosystems
- **Accessibility:** Widespread public access as compute costs decrease

- **Not just games:** Crucial for achieving autonomous driving and AGI
- **Social impact:** Experiences for disabilities, education, and therapy

We are no longer just watching the video. We are stepping inside.

Generative World Models

Base Stone: Video Generation Models

ADDING:

1. Long-term Memory
   - —3D modeling
   - —Compressed Feature/Representation

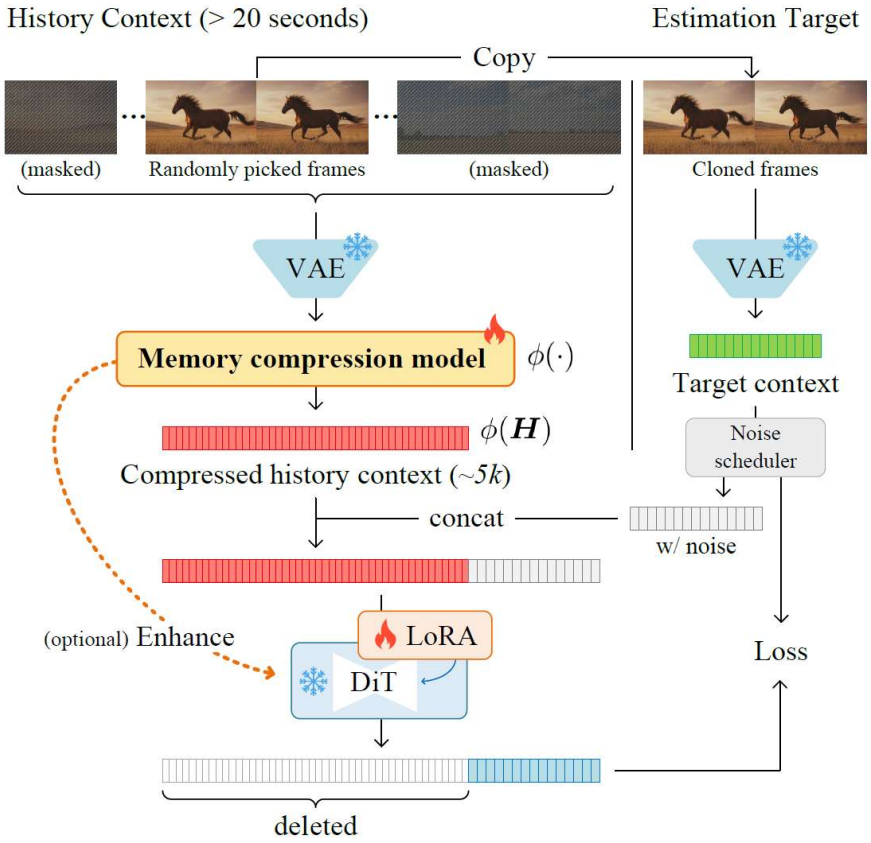
2. Understanding the World
   - —Physics understanding
   - —Realism
   - —Future Prediction & Decision

3. Interactivity
   - —WASD

Enough for Industry! What can we do for WMs in Academia?

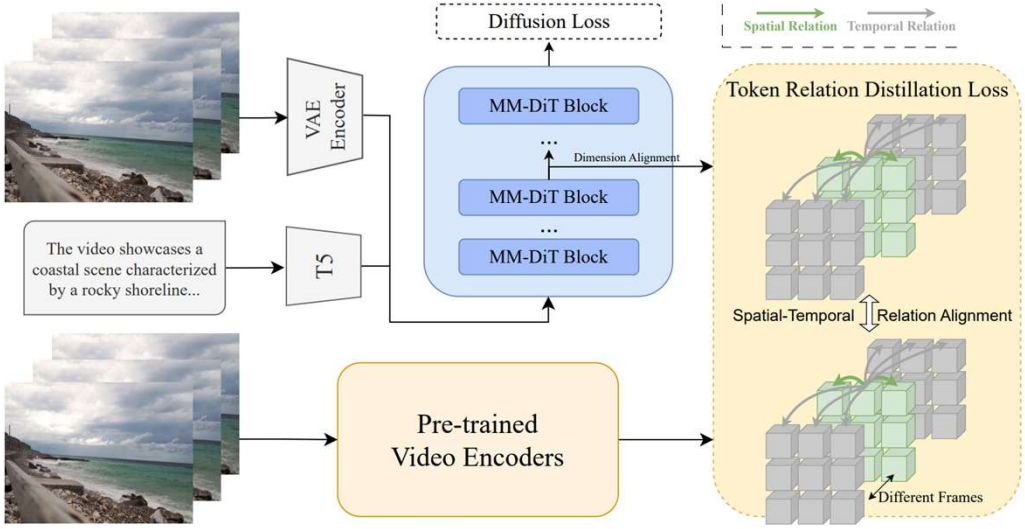
Long-term Memory Compression

- **Compression Model:**
  - 20s video → ~5k token context
  - Training:
    - Mask out
    - Use the Context to denoise / reconstruct a clear frame from a random time point



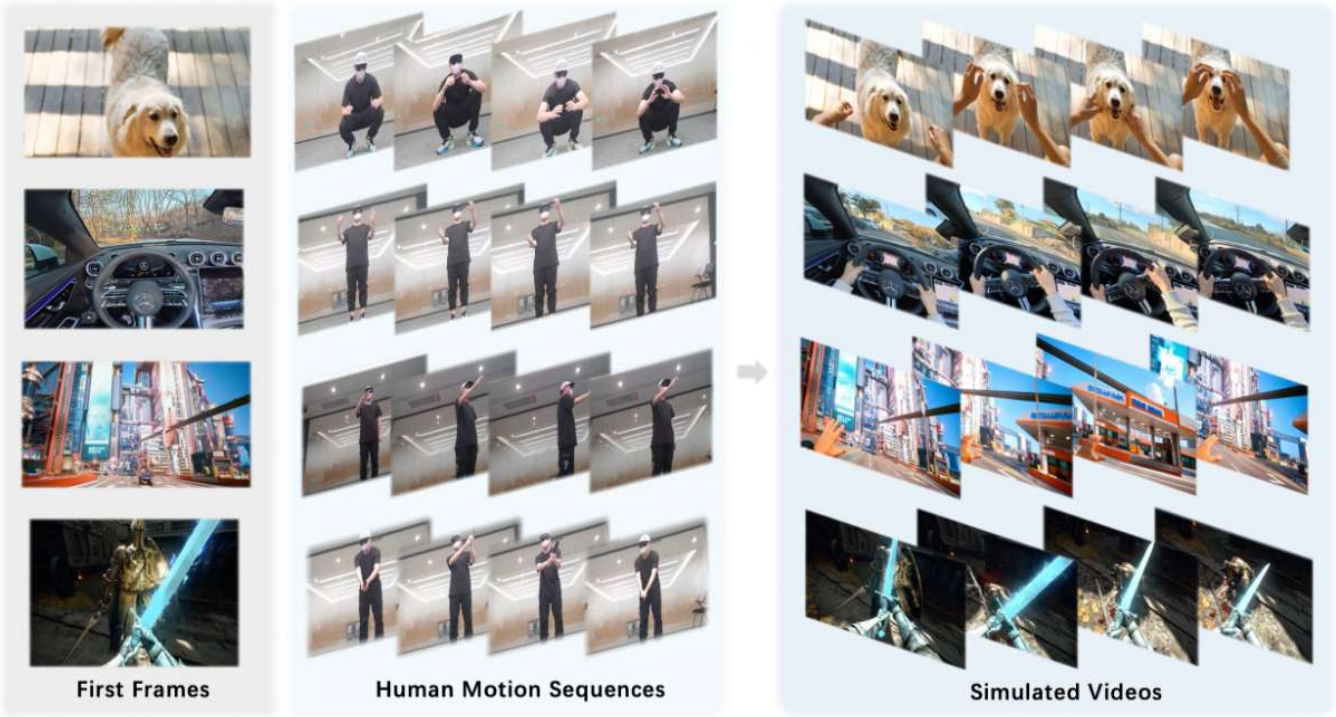
Enhanced Physical Understanding

- **VideoREPA:**
  - Guiding VDMs using VFMs
  - Student: VDM (Video Diffusion Model)
  - Teacher: VFM (Video Foundation Model)
  - Aligning relationships between tokens
  - Spatial: $$y_{\text{spatial}}^{d,i,j} = \frac{\mathbf{y_v^{d,i}} \cdot \mathbf{y_v^{d,j}}}{\|\mathbf{y_v^{d,i}}\| \|\mathbf{y_v^{d,j}}\|}$$
  - Temporal: $$y_{\text{temp}}^{d,i,j,e} = \frac{\mathbf{y_v^{d,i}} \cdot \mathbf{y_v^{e,j}}}{\|\mathbf{y_v^{d,i}}\| \|\mathbf{y_v^{e,j}}\|}$$



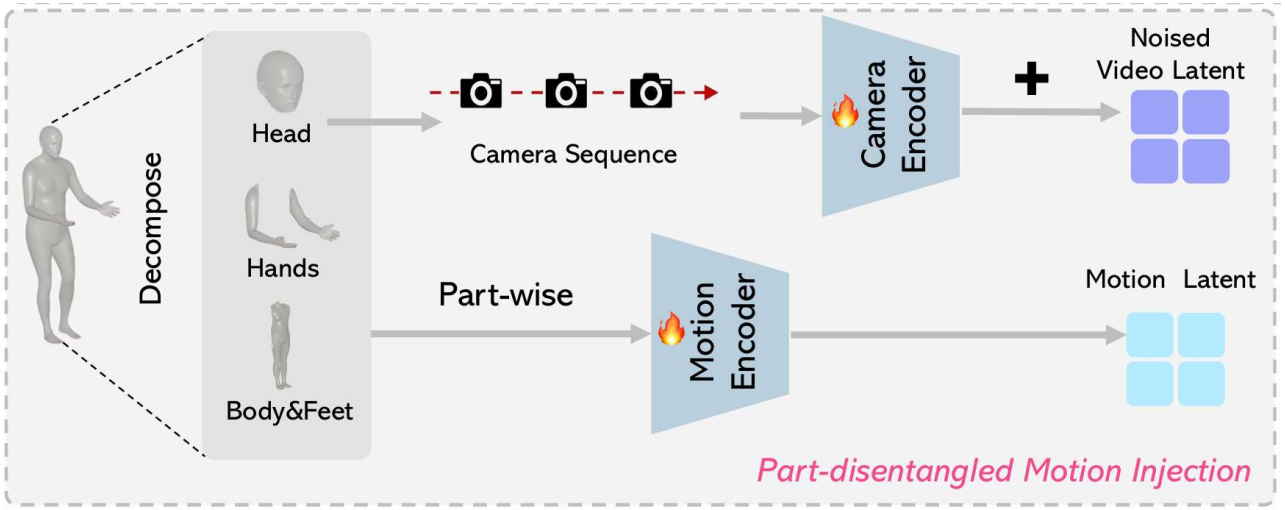
Extended Interactivity

- **PlayerOne:**
  - VR
  - Human motion
  - Rather than keyboard & mouse



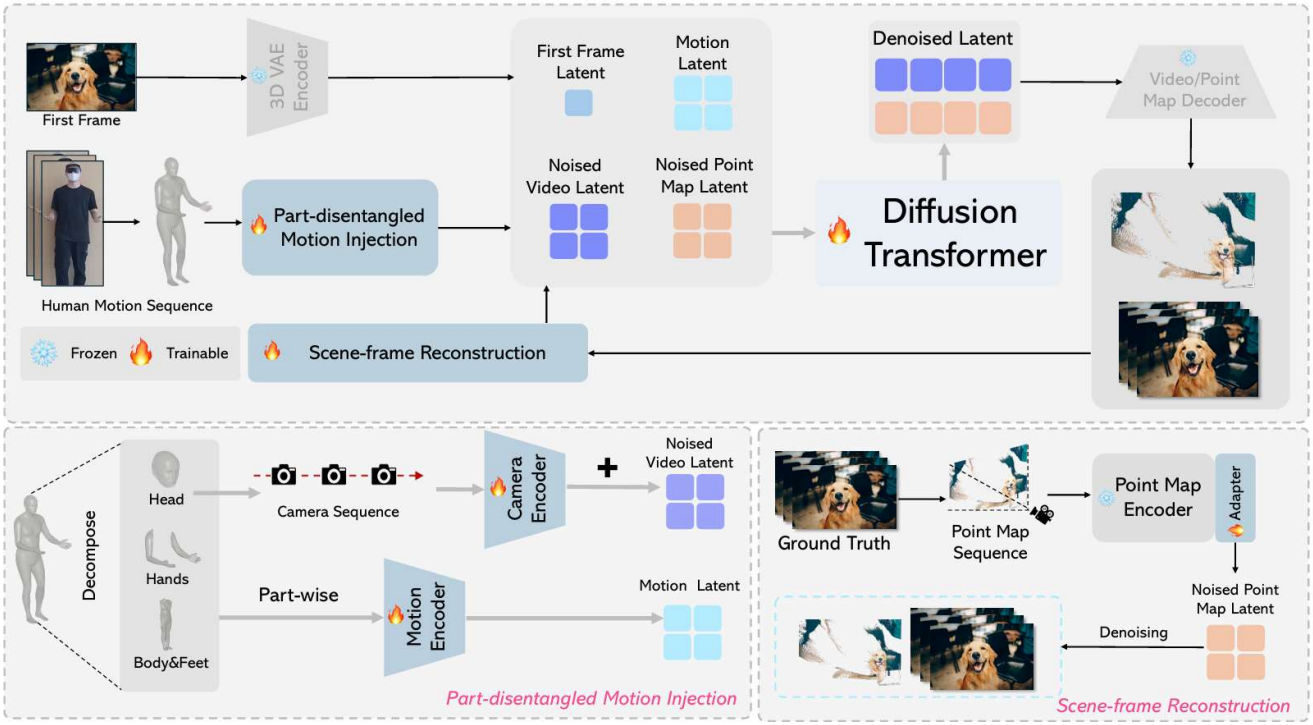
Extended Interactivity

- **PlayerOne:**
  - Human motion decomposites into 3 parts:
  - **Head** → camera params
  - **Hands** & **Feet** → motions



Extended Interactivity

- **PlayerOne:**
  - Diffusion objective: video + point cloud (only in training, cut off in inference)



Generation with Understanding

- Core conflict for generative models
  - Semantic drift
  - Slow convergence
  - Lack of logical consistency in complex tasks
- Solution
  - Leveraging "Understanding Models" provides a roadmap for generators
    - Feature Alignment
    - Prompt Enhancement
      - CoT for Generation

Prompt Enhancement

- Problem of user inputs
  - Intent Sparsity
  - Biased Description
  - Ambiguity
- Prompt Enhancement
  - Using a LLM to transform vague, underspecified user inputs into rich, high-fidelity instructions that generative models can better execute.

Prompt Enhancement

- Prompt Enhancement operational modes
  - Expansion
    - Add style descrption, details ...
  - Alignment
    - Transform informal language into professional terms
  - Negative Prompting
    - Add some negative feature

CoT for Generation

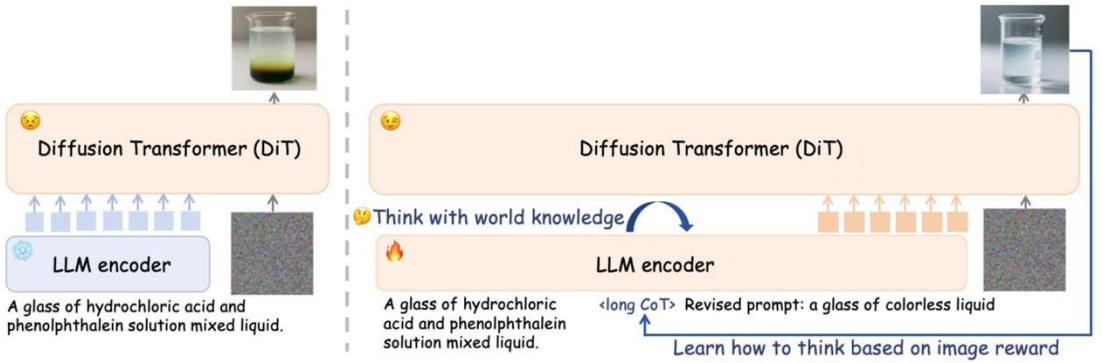
- Chain of Thought （CoT）
  - Decomposing complex generation into verifiable logical steps
  - Transform "implicit understanding" into "explicit logic"
- CoT for Generation
  - Prompt-based CoT
  - Multimodal CoT
  - CoT with Planning&Evaluation
  - CoT for Video Generation

Prompt CoT for Generation

- Think-Then-Generate

Think-Then-Generate: Reasoning-Aware Text-to-Image Diffusion with LLM Encoders

Siqi Kou[1*], Jiachun Jin[1*], Zetong Zhou[1*], Ye Ma[2], Yugang Wang[1], Quan Chen[2], Peng Jiang[2], Xiao Yang[3], Jun Zhu[3], Kai Yu[1], Zhijie Deng[1†]
[1]Shanghai Jiao Tong University [2]Kuaishou Technology [3]Tsinghua University



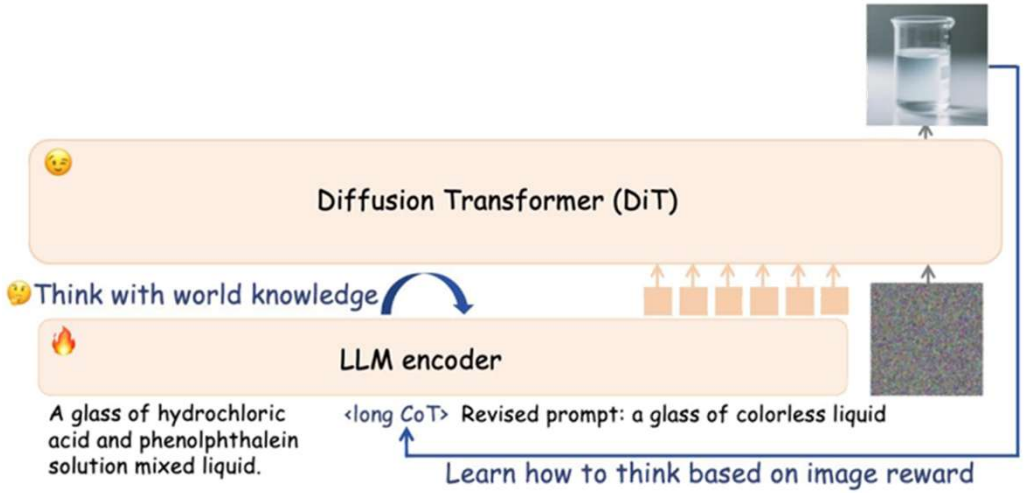
Kuaishou, Think-Then-Generate: Reasoning-Aware Text-to-Image Diffusion with LLM Encoders, Arxiv 2026

Prompt CoT for Generation

- Inference
  - Raw Prompt → [Chain-of-Thought Reasoning] → Refined Descriptive Prompt
  - Example
    - Raw Prompt: "Generate Einstein's favorite instrument"
    - CoT: <think>Einstein was known to play the violin... This suggests creating an image of a classic wooden violin...</think>
    - Refined Prompt: A classic wooden violin, rich brown color, elegant craftsmanship...

Prompt CoT for Generation

- Training
  - Supervised Fine-Tuning
    - Tune Qwen2.5-VL with data generated by Gemini 2.5
  - Dual-GRPO
    - Generate J inference token list z_j
    - Generate K images for each z_j
    - Reward for both LLM & DiT



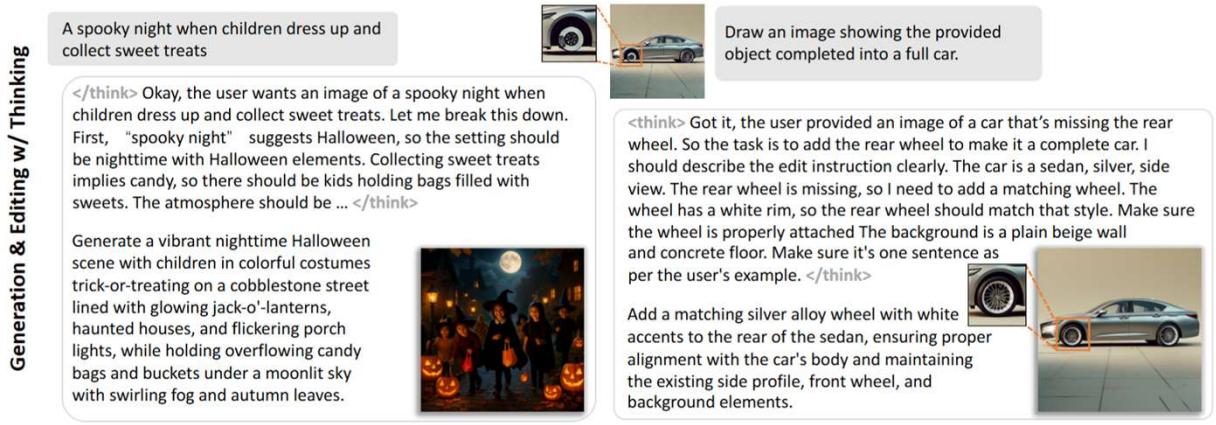
Multimodal CoT for Generation

- ThinkGen

ThinkGen: Generalized Thinking for Visual Generation

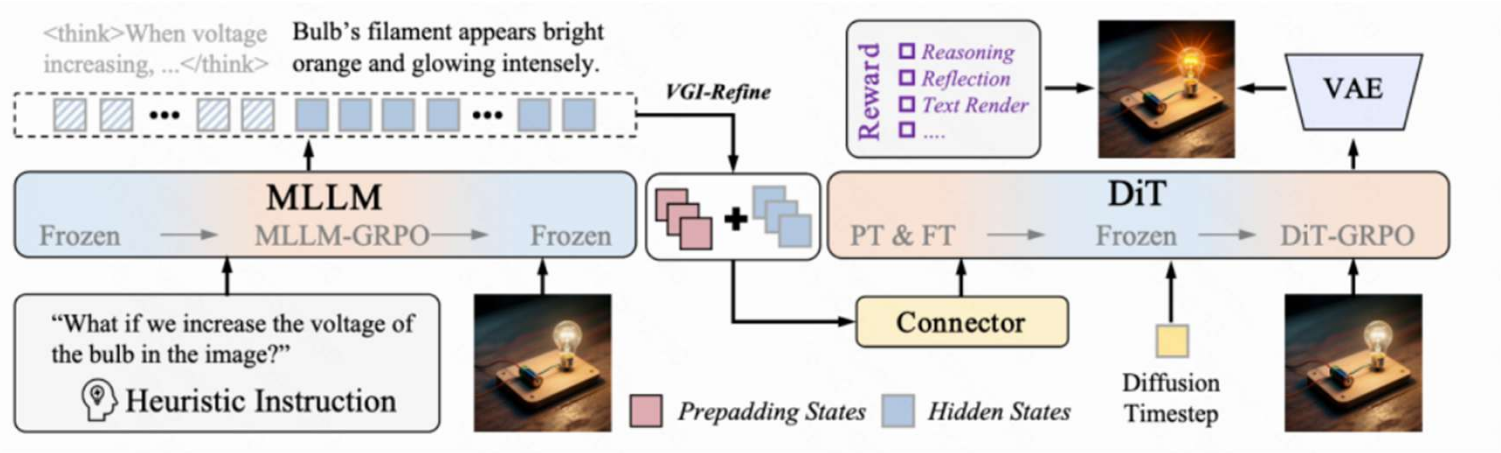
Siyu Jiao[1*] Yiheng Lin[1*] Yujie Zhong[2†] Qi She[2] Wei Zhou[2] Xiaohan Lan[2]
Zilong Huang[2] Fei Yu[2] Yingchen Yu[2] Yunqing Zhao[2] Yao Zhao[1] Yunchao Wei[1†]

[1] Beijing Jiaotong University [2] Bytedance



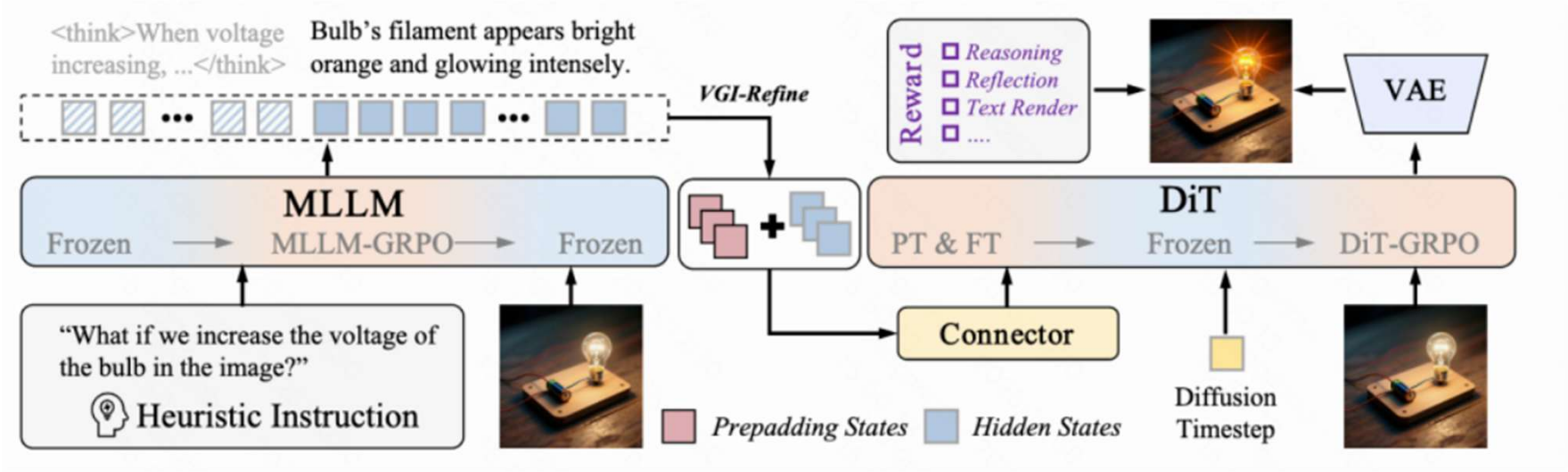
Multimodal CoT for Generation

- Overview
  - Introduce MLLM into the CoT generation process
    - Directly inject img into CoT process needs too much context length
    - Introduce VGI-Refine Module



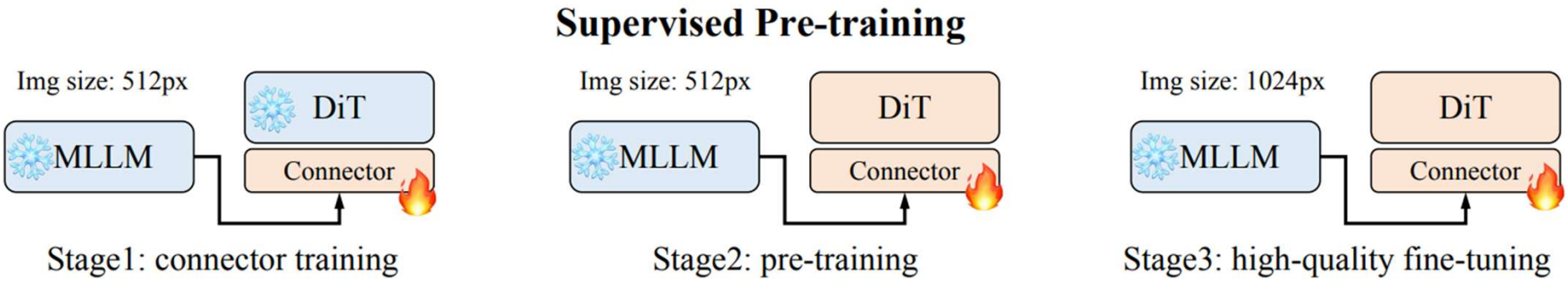
Multimodal CoT for Generation

- VGI-Refine
  - Extract Hidden States
  - Padding with Prepadding States



Multimodal CoT for Generation

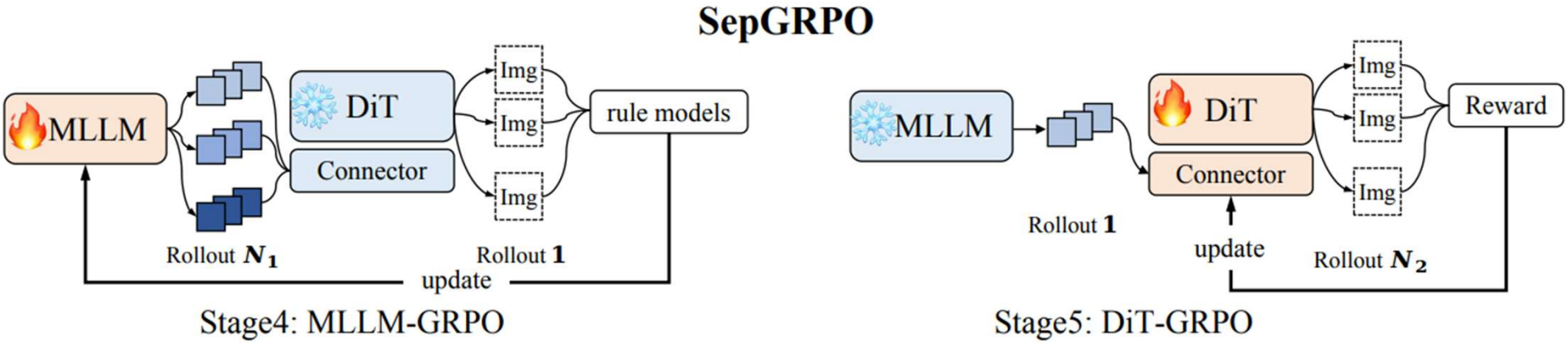
- Training
  - SFT
    - Training connector module
    - Training connector&DiT
    - Training with larger dataset



ByteDance, ThinkGen: Generalized Thinking for Visual Generation, Arxiv 2025

Multimodal CoT for Generation

- Training
  - GRPO
    - MLLM-GRPO
    - DiT-GRPO



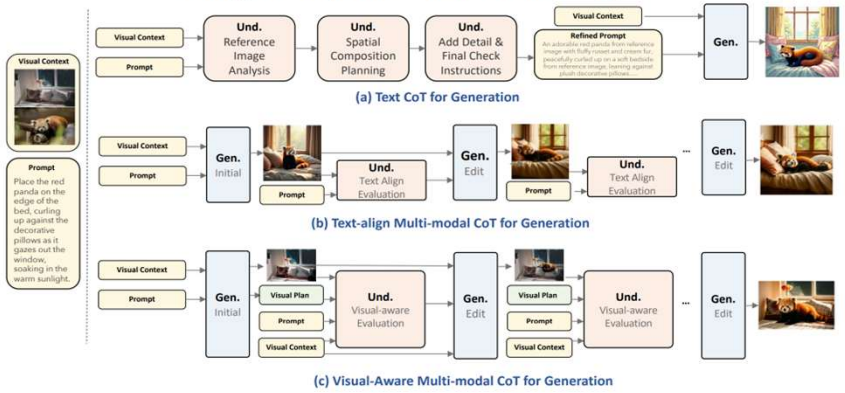
ByteDance, ThinkGen: Generalized Thinking for Visual Generation, Arxiv 2025

CoT for Editing with Planning&Evaluation

- Visual-Aware CoT

Visual-Aware CoT: Achieving High-Fidelity Visual Consistency in Unified Models

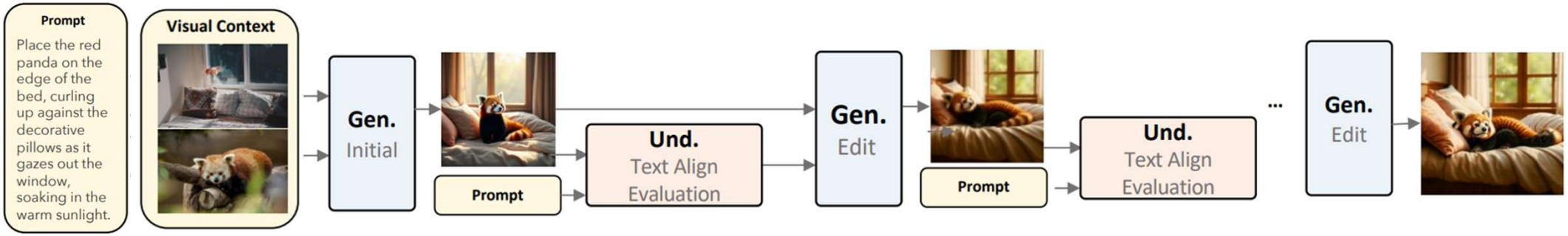
Zixuan Ye[1]* Quande Liu[2]† Cong Wei[2] Yuanxing Zhang[2] Xintao Wang[2]
Pengfei Wan[2] Kun Gai[2] Wenhan Luo[1]†
[1]The Hong Kong University of Science and Technology
[2]Kling Team, Kuaishou Technology



Kling, Visual-Aware CoT: Achieving High-Fidelity Visual Consistency in Unified Models, Arxiv 2025

CoT for Editing with Planning&Evaluation

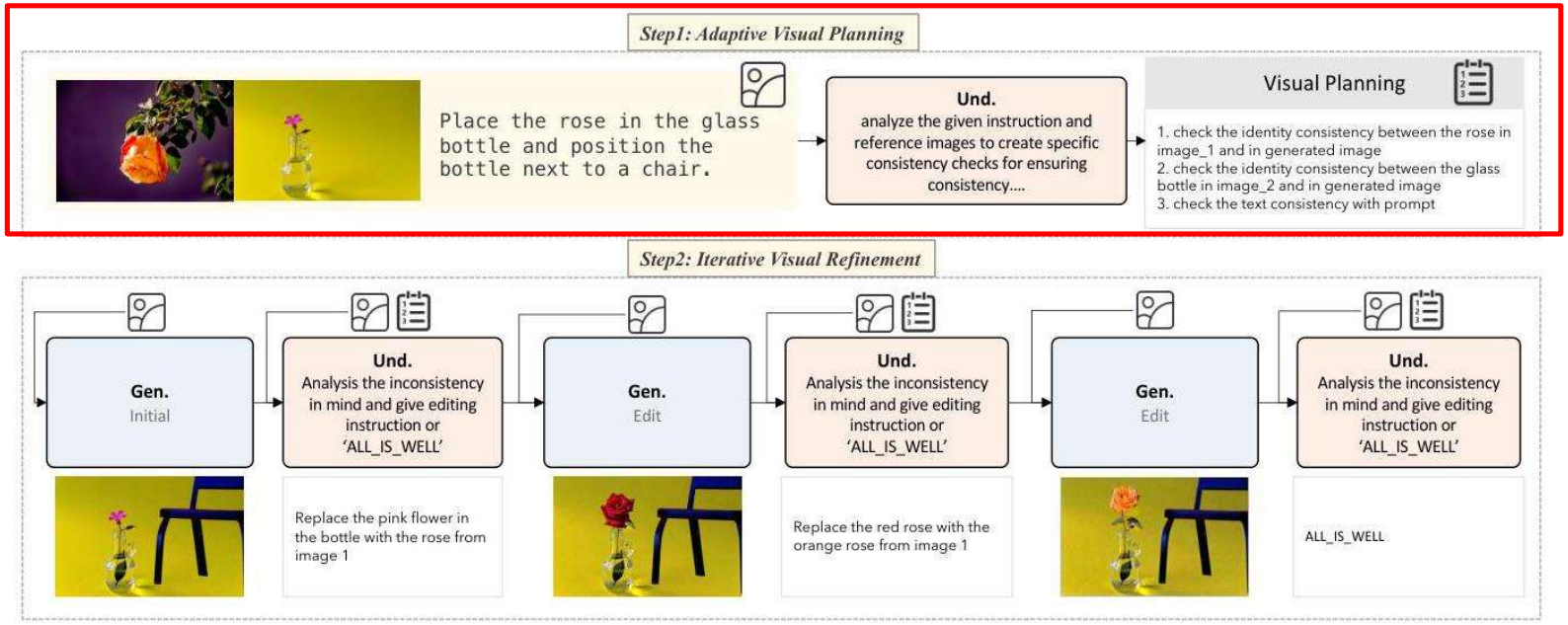
- Motivation
  - Current multimodal CoT is difficult to maintain the fidelity to the visual context
- Solution
  - Introduce the evaluation to maintain the fidelity with original visual context



Kling, Visual-Aware CoT: Achieving High-Fidelity Visual Consistency in Unified Models, Arxiv 2025

CoT for Editing with Planning&Evaluation

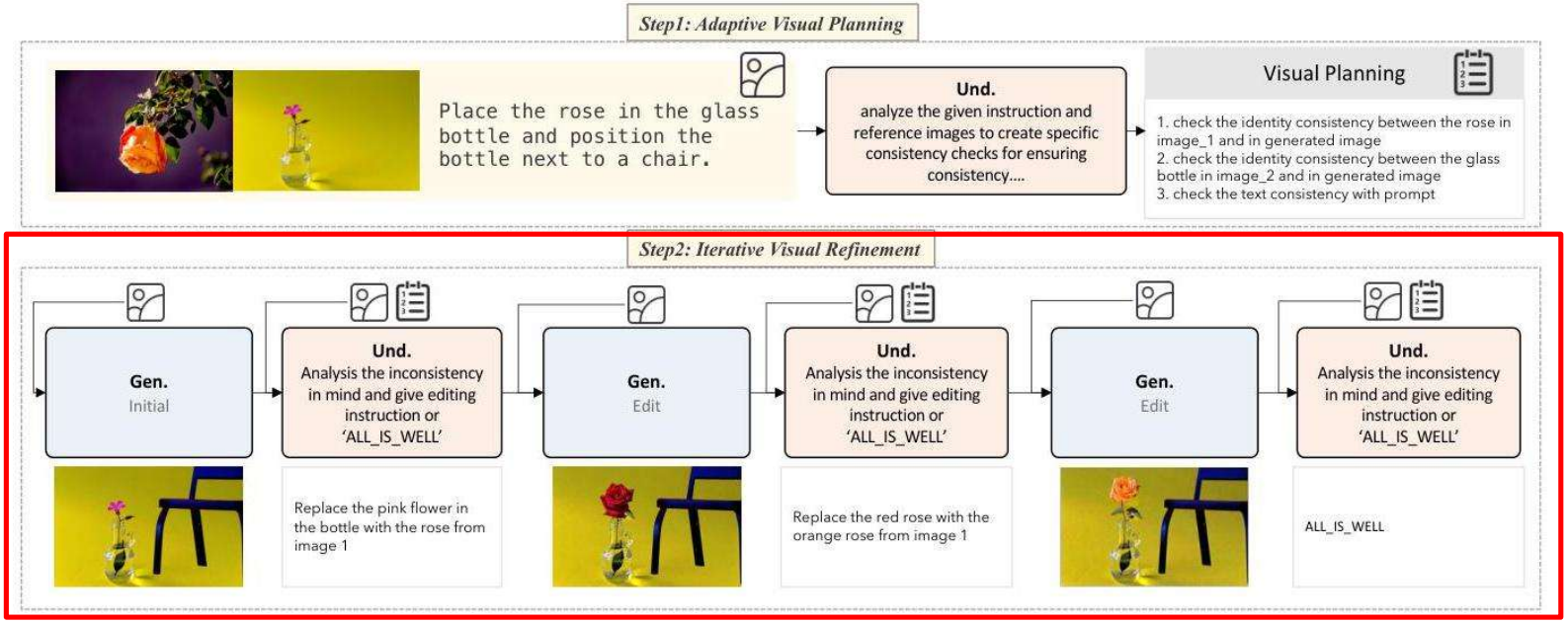
- Inference
  - Step 1： Generate a visual planning list for fidelity evaluation



Kling, Visual-Aware CoT: Achieving High-Fidelity Visual Consistency in Unified Models, Arxiv 2025

CoT for Editing with Planning&Evaluation

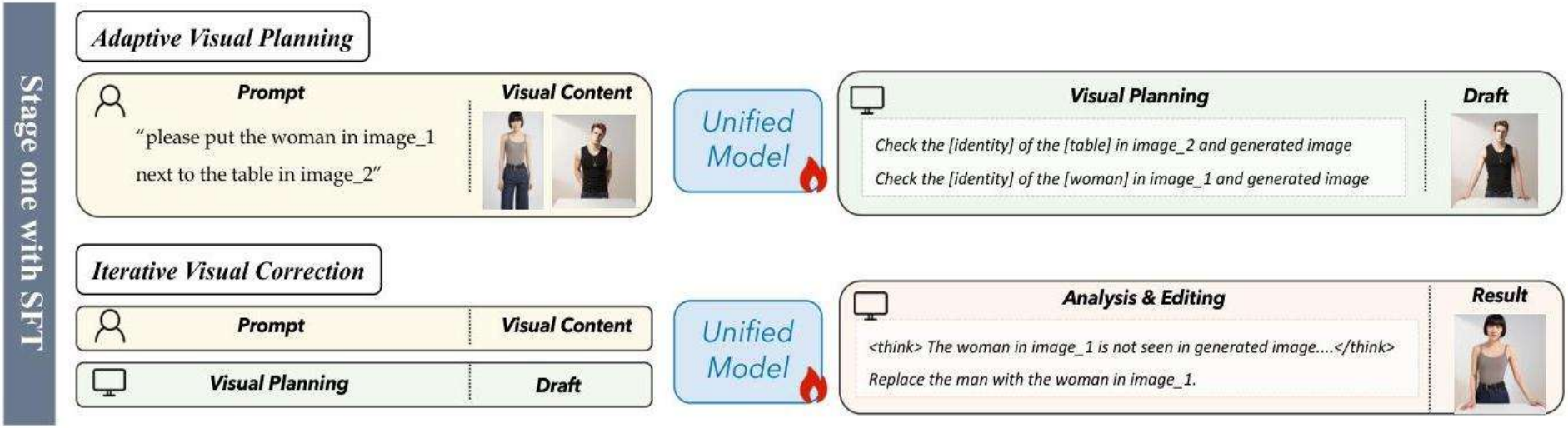
- Inference
  - Step 2：Iterative Visual Refinement



Kling, Visual-Aware CoT: Achieving High-Fidelity Visual Consistency in Unified Models, Arxiv 2025

CoT for Editing with Planning&Evaluation

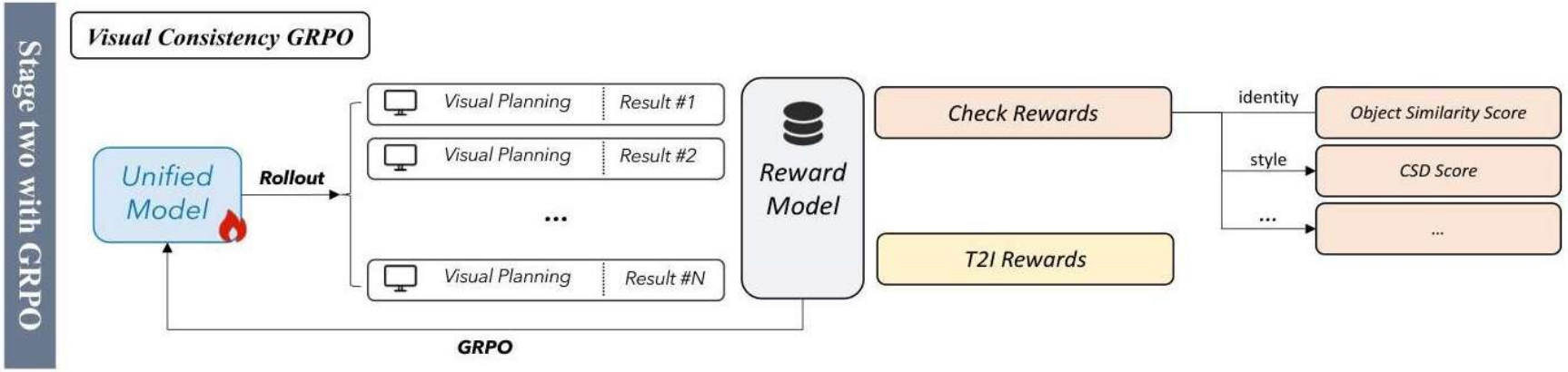
- Training
  - Stage 1： Supervised Fine-Tuning
    - Generate dataset with Gemini and training BAGEL



Kling, Visual-Aware CoT: Achieving High-Fidelity Visual Consistency in Unified Models, Arxiv 2025

CoT for Editing with Planning&Evaluation

- Training
  - Stage 2：GRPO
    - Object Position: GroundingDINO
    - Object Similarity: DINO
    - Style Similarity: CSD-Score



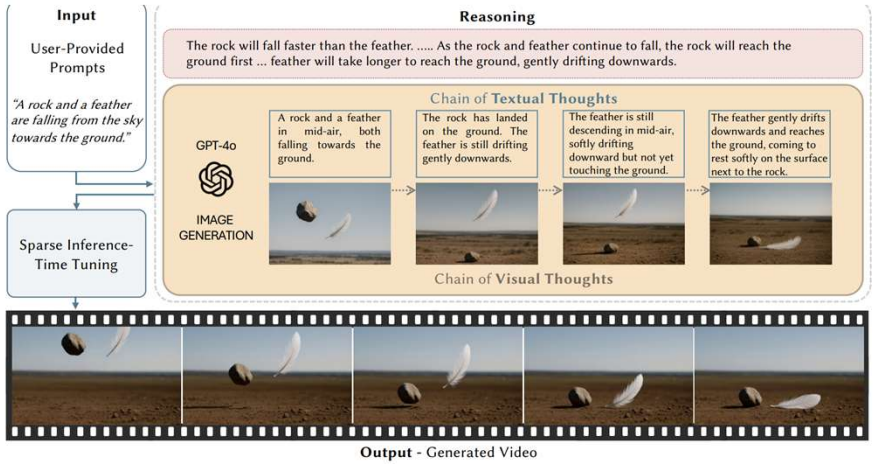
Kling, Visual-Aware CoT: Achieving High-Fidelity Visual Consistency in Unified Models, Arxiv 2025

CoT for Video Generation

- VChain

VChain: Chain-of-Visual-Thought for Reasoning in Video Generation

Ziqi Huang, Ning Yu✉†, Gordon Chen, Haonan Qiu, Paul Debevec, Ziwei Liu✉

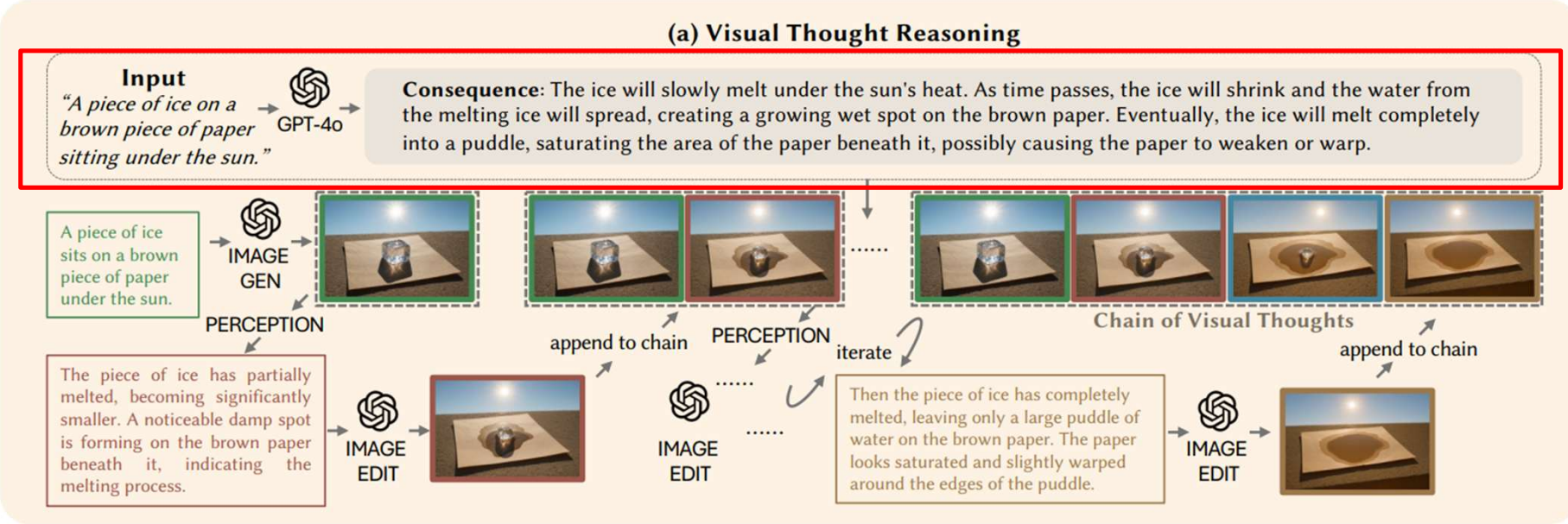


CoT for Video Generation

- Motivation
  - Directly generation videos is difficult to fully conform to the laws of world
  - Text tokens of LLMs is difficult to transform into visual sequences
- Solution
  - Introduce Visual CoT with Unified LLM&Diffusion
  - Decompose the complex state into sparse keyframes

CoT for Video Generation

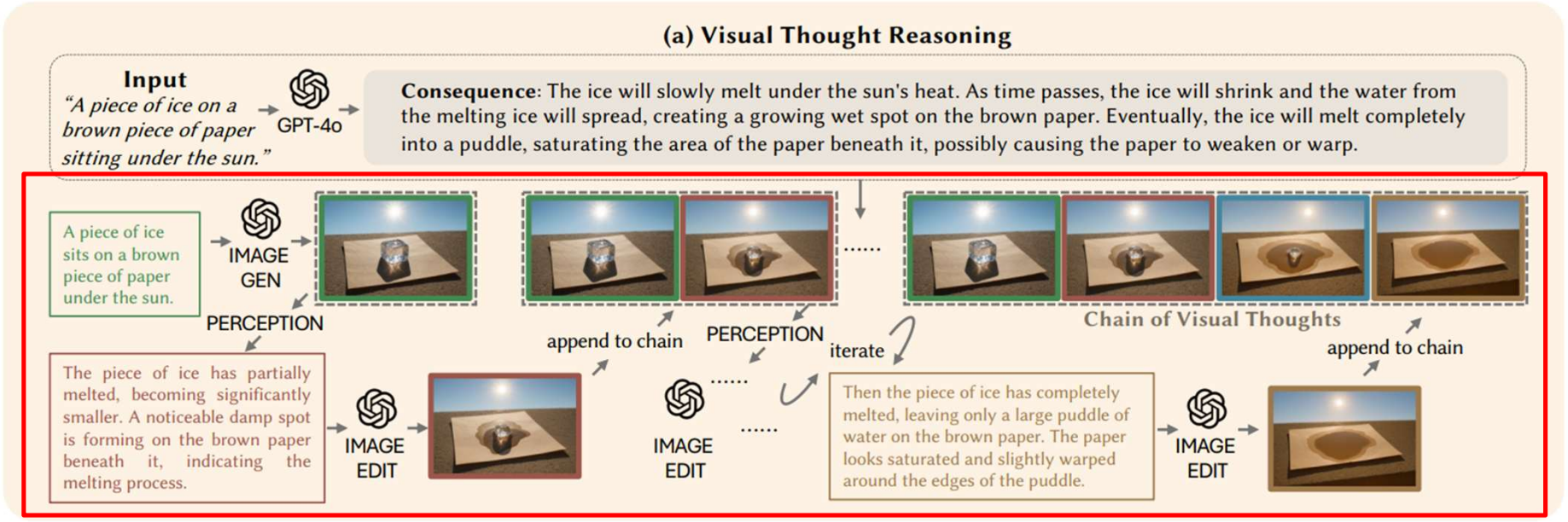
- Visual Thought Reasoning
  - Step 1: Create the text description of consequence



NTU, VChain: Chain-of-Visual-Thought for Reasoning in Video Generation, Arxiv 2025

CoT for Video Generation

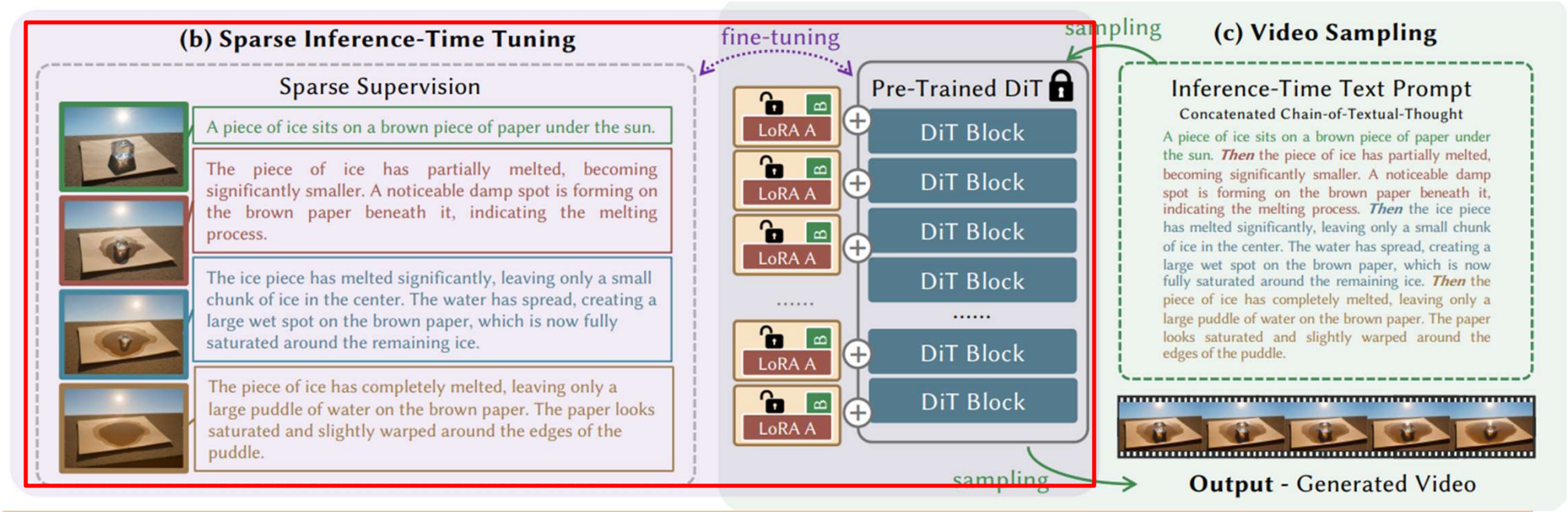
- Visual Thought Reasoning
  - Step 2: Generate keyframes with image editing guided by consequence



NTU, VChain: Chain-of-Visual-Thought for Reasoning in Video Generation, Arxiv 2025

CoT for Video Generation

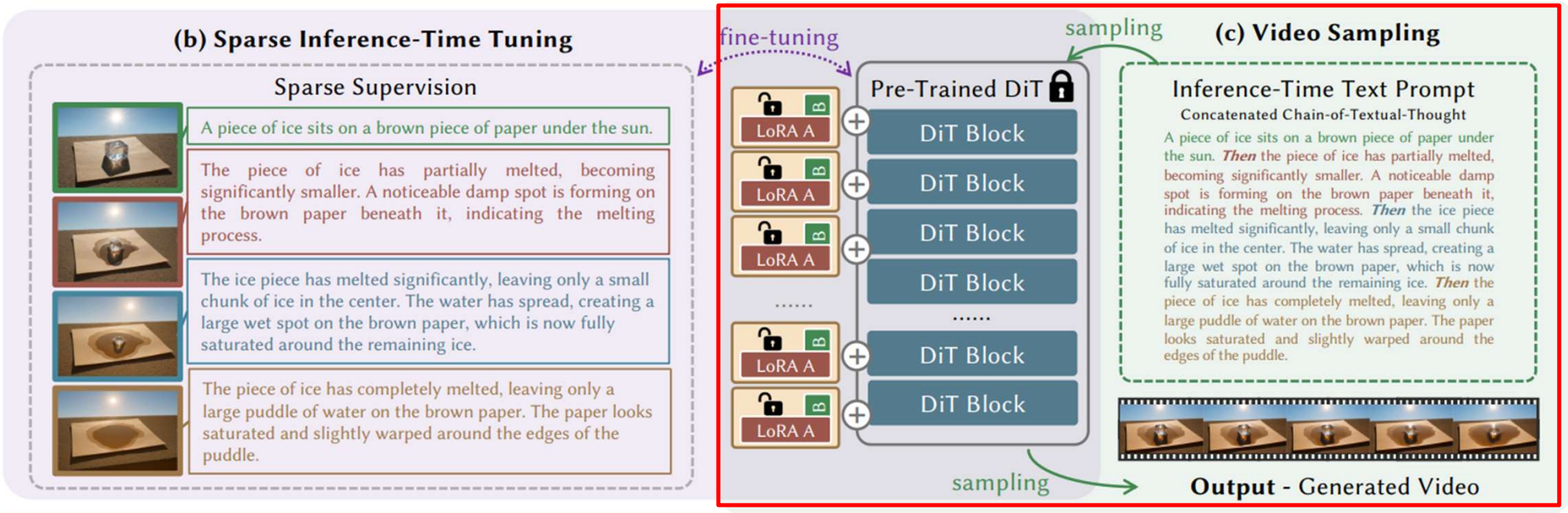
- Visual Thought Reasoning
  - Step 3: Finetuning the DiT with LoRA



NTU, VChain: Chain-of-Visual-Thought for Reasoning in Video Generation, Arxiv 2025

CoT for Video Generation

- Visual Thought Reasoning
  - Step 4: Generate the video with fully prompt



NTU, VChain: Chain-of-Visual-Thought for Reasoning in Video Generation, Arxiv 2025

Generation with Understanding

- Conclusion: Why does CoT work?
  - Understanding models offload intent comprehension from the generator
  - Enhance inference compute
  - Error isolation via decomposition

Thanks!