Controllability Matters: Interpreting World Models from a Controllability Perspective

Zicheng Duan

Australian Institute for Machine Learning (AIML)
The University of Adelaide

Visual Generative Models: Past, Present, and Future, DICTA 2025 Workshop

World Models? What and How?

There is no single definition. Current research approaches ”World Models” from diverse angles:

- **The 3D Reconstruction Perspective:**
  Constructing explicit, spatially consistent 3D environments (e.g., 3DGS, Point Clouds) to support camera navigation.

- **The Generative Perspective (Next-Token Prediction):**
  “The Magical Scaling Law.”

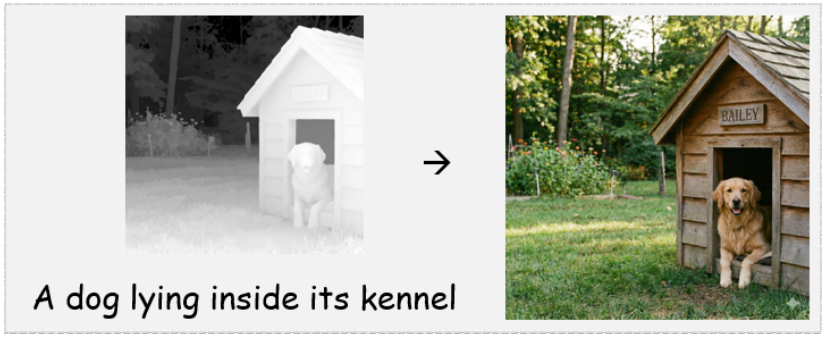
Core Goal: To build a digital simulator that mirrors the laws of reality.

One Concept Above All: Controllable Generation.

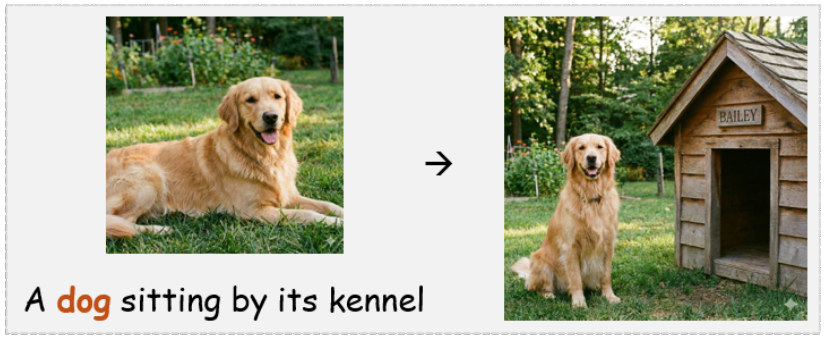
Why Controllable Generation?

Human creators use control signals to anchor their specific thoughts into the generated content, e.g., I want a dog sitting on the left side of the kennel.

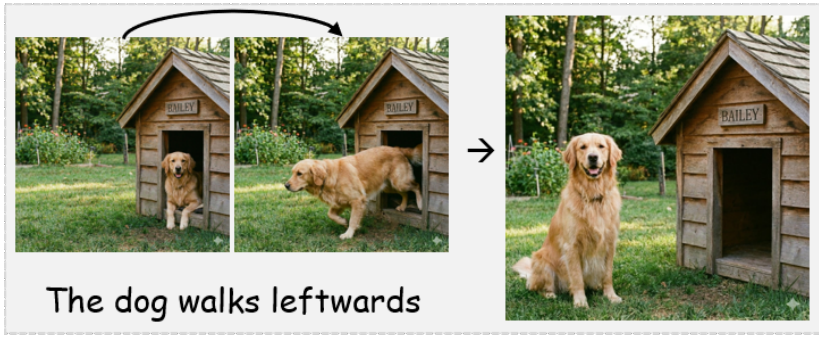
1. Semantic Control: Class, Text, Style



2. Spatial Control: Edges, Depth, Layout



3. Subject Control: Object Appearances, Identity.



4. Temporal Control: Motion & Events over Time.

1.2 The Ultimate Goal: From Control to World Simulation

- We apply **Semantic Controls** $\rightarrow$ define **Context & Content**.
  i.e., What the world contains.
- We apply **Spatial Controls** $\rightarrow$ enforce **Geometric Consistency**.
  i.e., The world has a valid structure.
- We apply **Subject Controls** $\rightarrow$ enforce **Object Permanence**.
  i.e., Entities keep a consistent identity.
- We apply **Temporal Controls** $\rightarrow$ shape **Temporal Evolution**.
  i.e., Events evolve coherently, and we can steer when and how fast they happen.

Defining "World Models" from a Controllability Perspective

When a generative model follows these controls, it moves from a *content generator* to a **World Simulator** with **physically plausible, temporally coherent evolution**.

2. Our Attempts

To move closer to **Physically Plausible World Models**, we specifically investigated **Subject** and **Temporal** controls through our two recent attempts:

- **Subject Controls** (**EZIGen**) $\rightarrow$ enforce **Object Permanence**.
  i.e., Specific subjects (e.g., pets) appear across diverse scenes without identity loss.
  Paper: EZIGen: Enhancing zero-shot personalized image generation... (BMVC 2025)

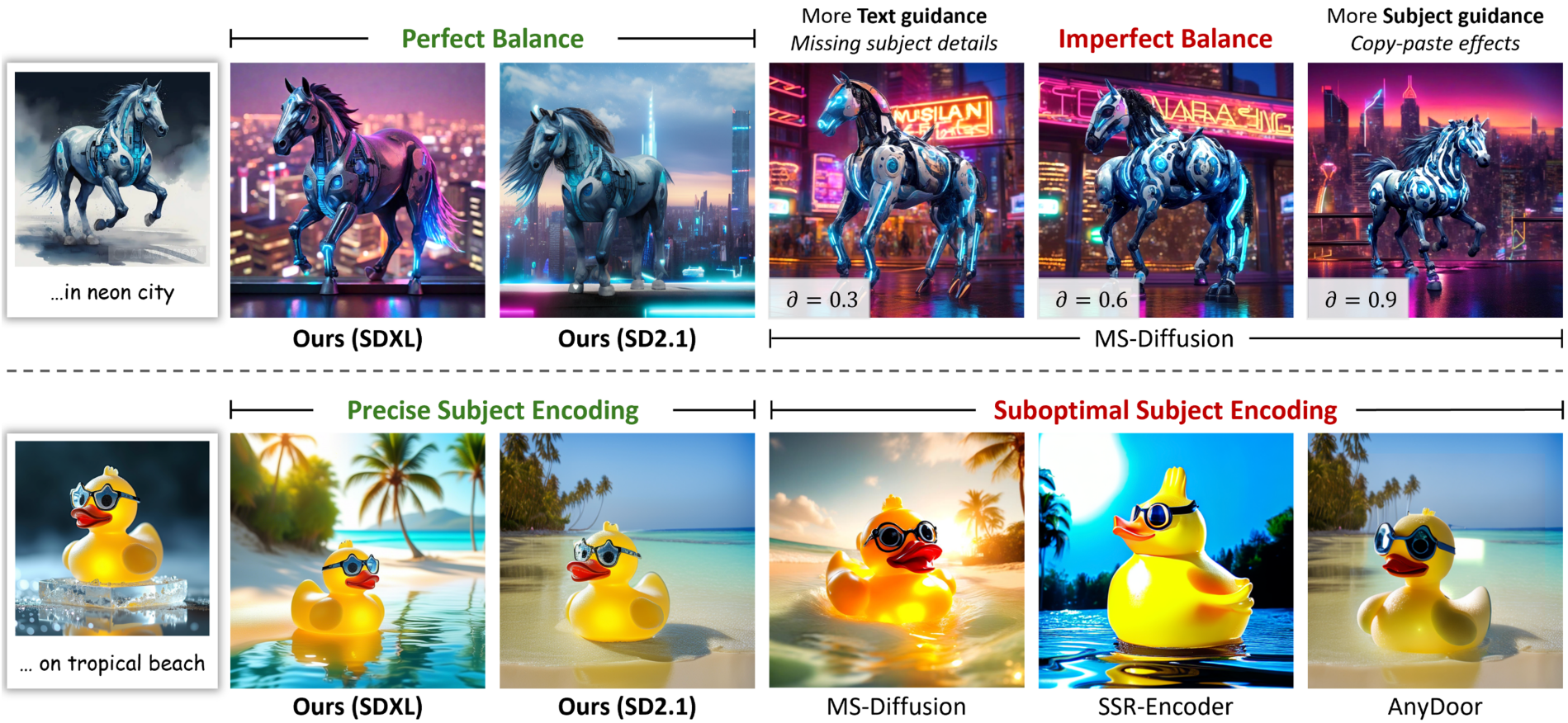
- **Temporal Controls** (**MVAA**) $\rightarrow$ enforce **Rhythmic Causality**.
  i.e., Motion dynamics follow the musical beat, as in real-world performance.
  Paper: Let Your Video Listen to Your Music!... (ACM MM 2025)

2.1.1 Subject Control: EZIGen (Motivation)

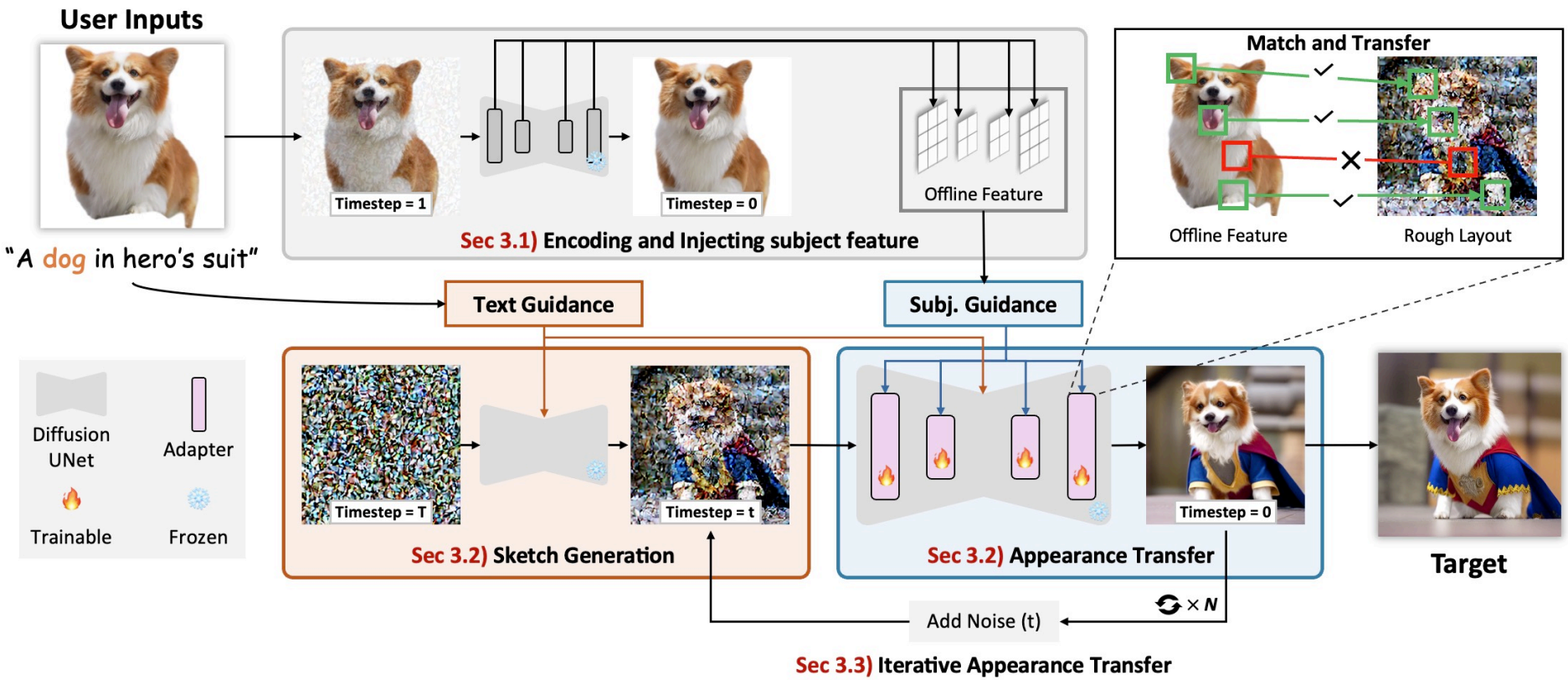
Definition: Generate content according to a *text prompt* and a *subject image(s)*.

Current Challenges:

- Subject appearance cannot be fully maintained.
- Text prompts conflict with subject features (trade-off).



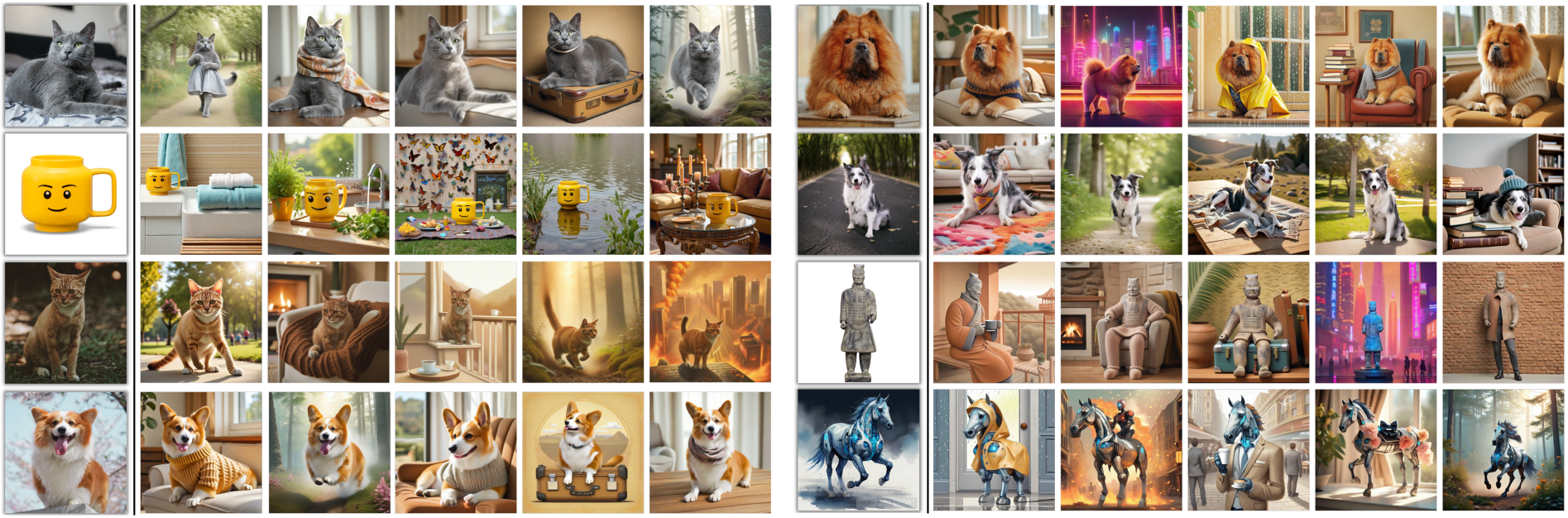
2.1.2 Subject Control: EZIGen (Method)



Our Solution:

- **Precise Encoding:** Use the Diffusion UNet itself to encode the subject for better quality and smaller domain gap.
- **Decoupled Guidance:** Separate generation into text-dominant and subject-dominant stages, iteratively refining for quality.

2.1.3 Subject Control: EZIGen (Results)



Conclusion

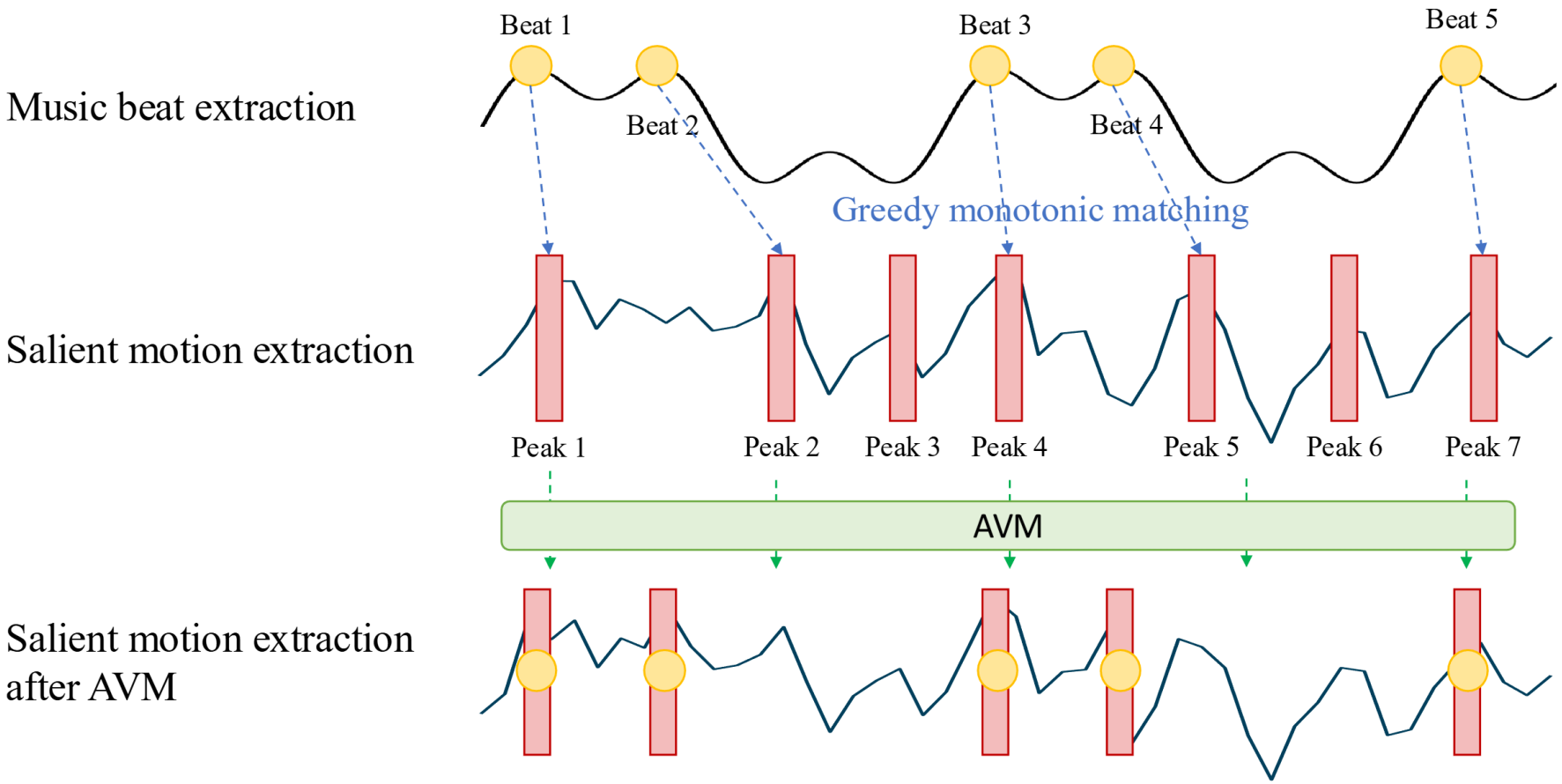
EZIGen seamlessly injects personalized subjects into any text-defined scenario, populating the generated world with **consistent characters**.

2.2.1 Audio/Motion Control: MVAA (Method)

Our Goal: Let musical beats *control* the generated video.

Our Solution:

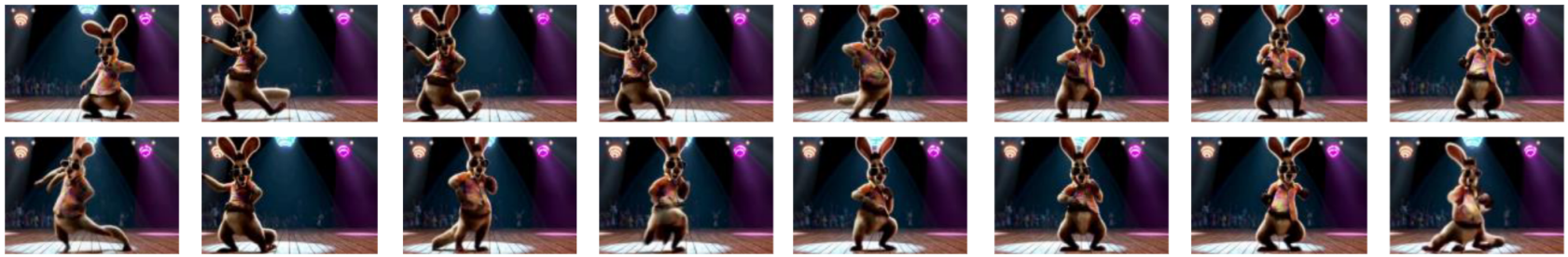
- **Beat-to-Motion:** Extract and match musical beats and salient motion keyframes.
- **Frame Inpainting:** Generate frames between beat-aligned keyframes.



2.2.2 Audio/Motion Control: MVAA (Results)

Ori Video

Our MVAA



Music: Birds of a Feather. Prompt: In a cozy kitchen, a golden retriever wearing a white chef's hat and a blue apron stands at the table, holding a sharp knife and skillfully slicing fresh tomatoes...

Ori Video

Our MVAA

Music: Rocket Man. Prompt: In a cozy kitchen, a golden retriever wearing a white chef's hat and a blue apron stands at the table, holding a sharp knife and skillfully slicing fresh tomatoes...

Conclusion

MVAA lets the video "listen" to the music—aligning motion with rhythm and mimicking **real-world causality**, where actors move in response to a soundtrack.

3. The Transition: From Passive Observation to Active Interaction

Current State:

- Existing generative models aim at simulating **Physically Plausible** worlds with controlled **Scenes**, **Subjects**, and **Dynamics**.
  e.g., Sora2 & Wan2.5 (Video-based), Marbel (3D-based)

This forms what we have at present.

However, we are essentially **Directing a Movie** (Passive).

⇓ *Transition*

We want to **Play the Game** (Active)!

What's Missing?

To complete the world, we need to enable **Embodied Interaction**, allowing us to navigate and act within the generated reality.

4. Future Paradigm: Action-Conditioned World Simulation

Definition

The model should accept **User Actions** as control and **interact** with the world.

Three Emerging Action-Interaction Paradigms:

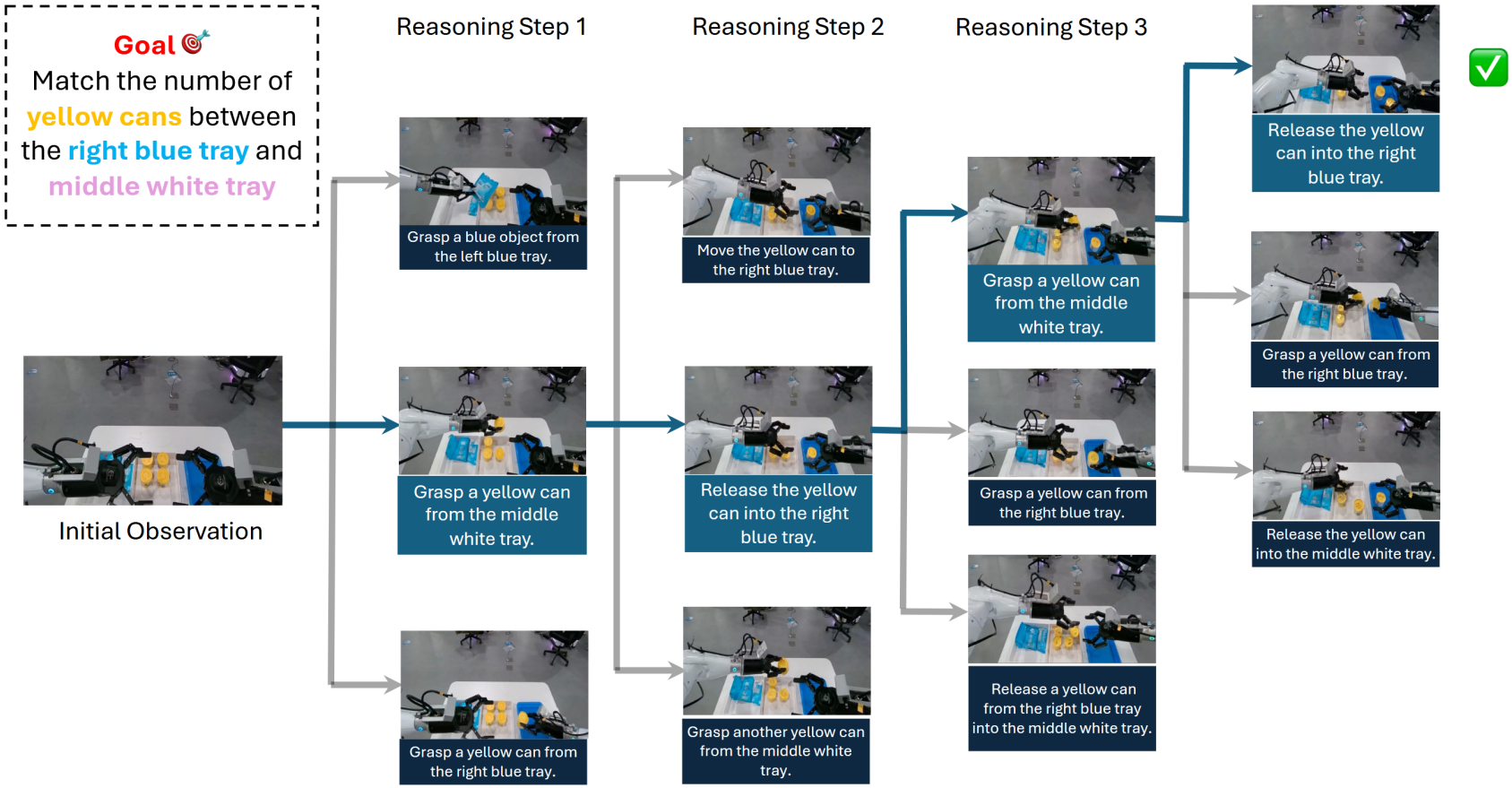
- **Embodied Manipulation:**
  Imagining the consequences of physical actions (e.g., robot arm manipulation).
  e.g., PEVA, RynnVLA-002

- **Camera Action:**
  Treating the camera as the "body" to explore generated 3D scenes.
  e.g., Camera Control II, Navigation World Model

- **Keyboard Control:**
  Controlling an avatar for game-like interaction.
  e.g., Hunyuan-GameCraft, Genie3

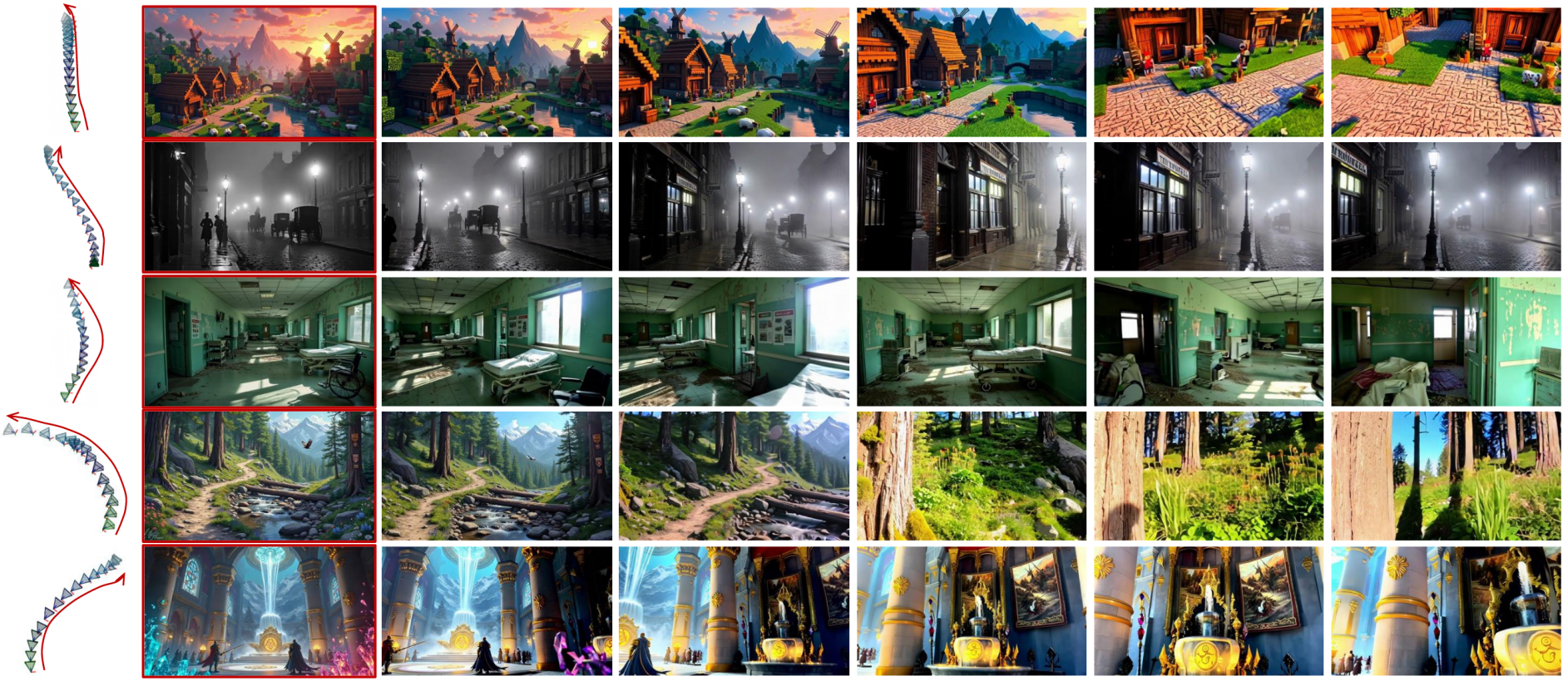
4.1 Embodied Manipulation Control

- **Concept:** Imagine the consequences of physical actions.
- **Input:** End-effector trajectories or gripper commands.
- **Goal:** Predict what will happen and support decision-making.



4.2 Camera Action Control

- **Concept:** Regard a moving camera itself as the ”Body” (The Observer).
- **Input:** Camera Trajectory.
- **Goal:** To actively **explore** the generated (3D) environment.



4.3 Keyboard Control (Playable Video)

- **Concept:** Controlling a specific avatar/actor within the scene.
- **Input:** Discrete signals (WASD keys, Joystick).
- **Goal: Playable Generation**. Controlling movements (jump, run, turn) just like in video games (e.g., Hunyuan-GameCraft).

5. Conclusion: Where will Controllable Generation Lead Us?

Past

Visual Content Controls

(Generating Pixels)

↓

Present

Consistent Simulation

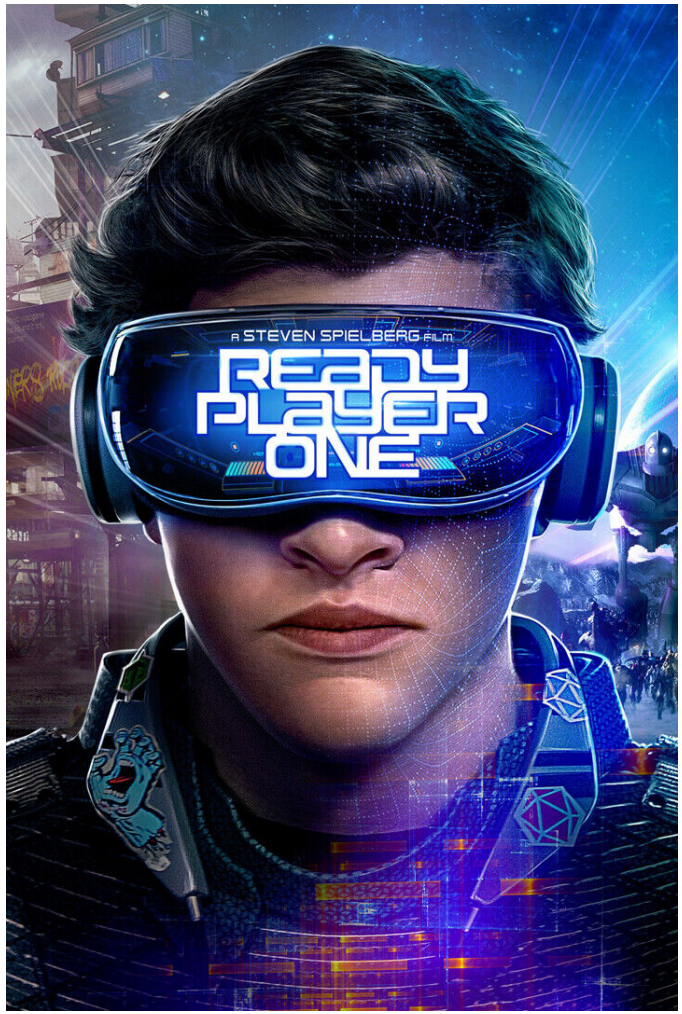
(Building the World)

↓

Future

Embodied Interaction

(Living in the World)



Final Thought

We are seeking **Free-form Embodied Interaction** in the generated worlds.

Thank you!