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.”

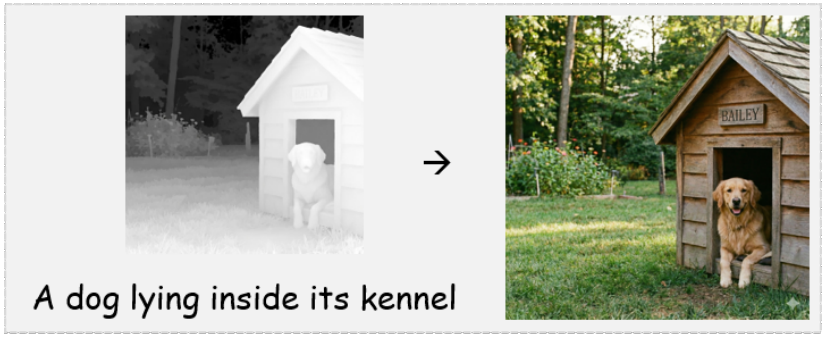
Core Goal: To build a digital simulator that mirrors the laws of reality.

One Concept Above All: Controllable Generation.

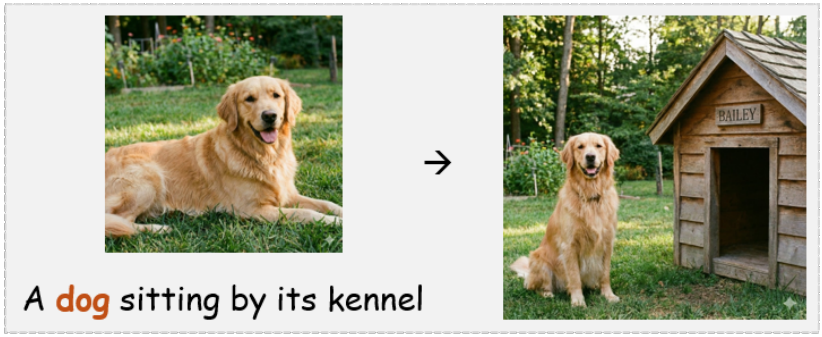
Why Controllable Generation?

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.

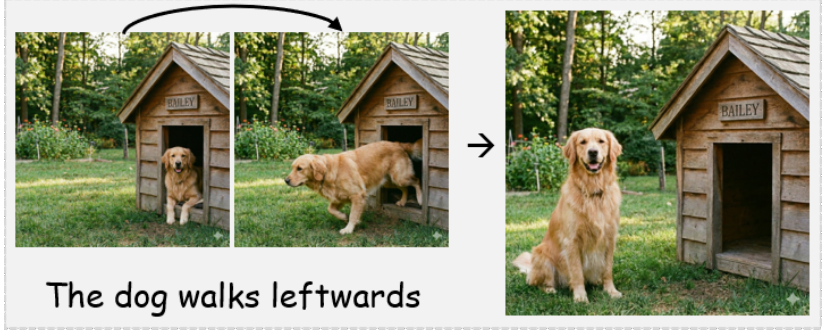
1. Semantic Control: Class, Text, Style



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** ! define **Context & Content**.
i.e., What the world contains.

We apply **Spatial Controls** ! enforce **Geometric Consistency**.
i.e., The world has a valid structure.

We apply **Subject Controls** ! enforce **Object Permanence**.
i.e., Entities keep a consistent identity.

We apply **Temporal Controls** ! 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**) ! 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**) ! 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)

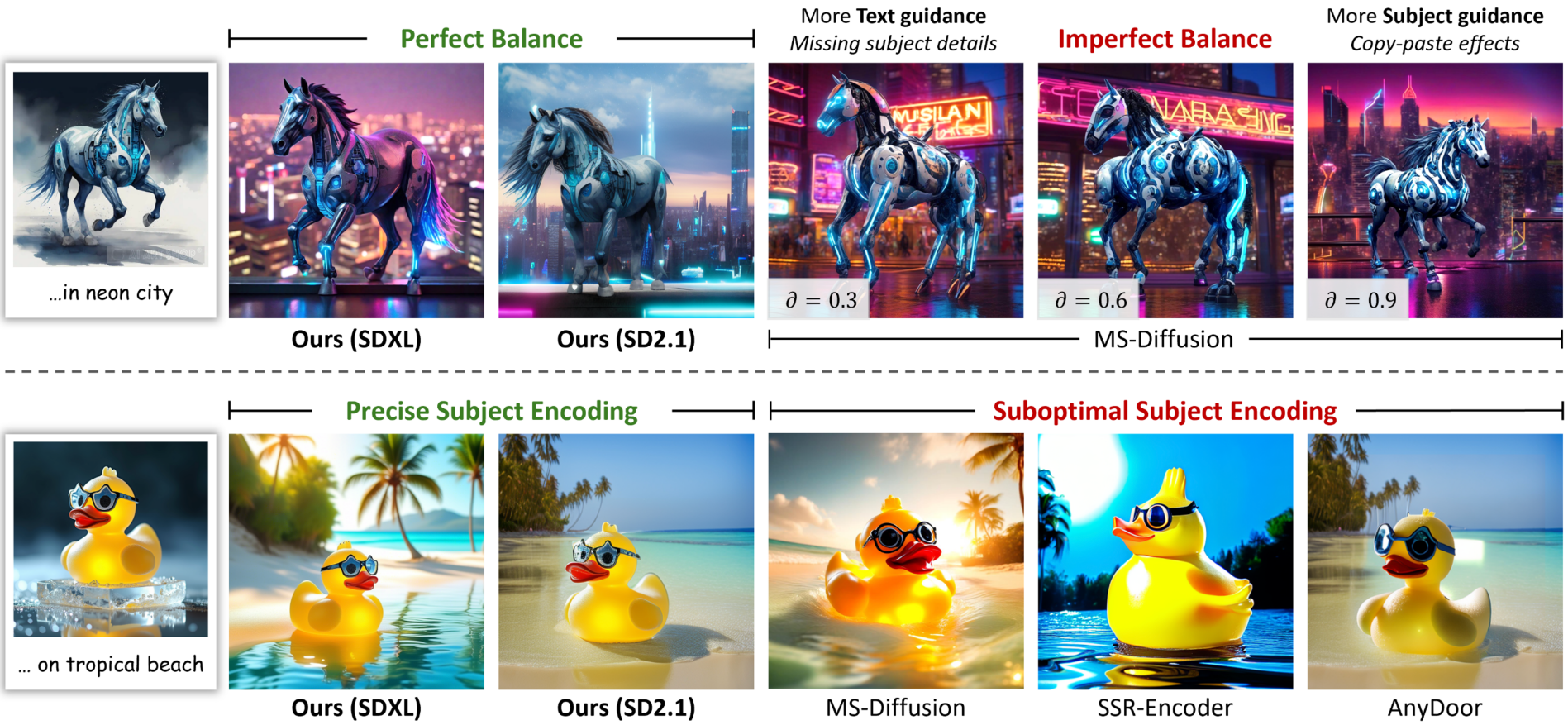
2.1.1 Subject Control: EZIGen (Motivation)

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

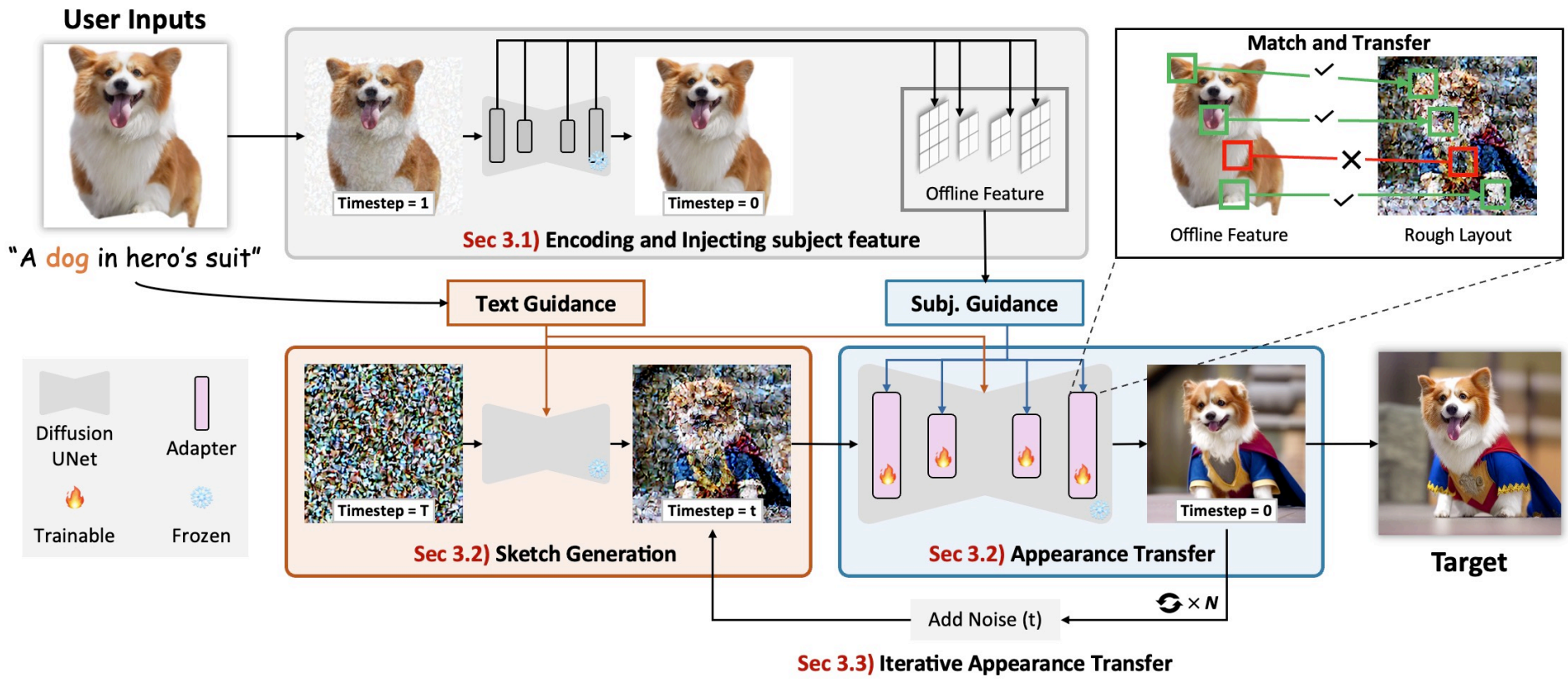
Current Challenges:

Subject appearance cannot be fully maintained.

Text prompts conflict with subject features (trade-off).



2.1.2 Subject Control: EZIGen (Method)

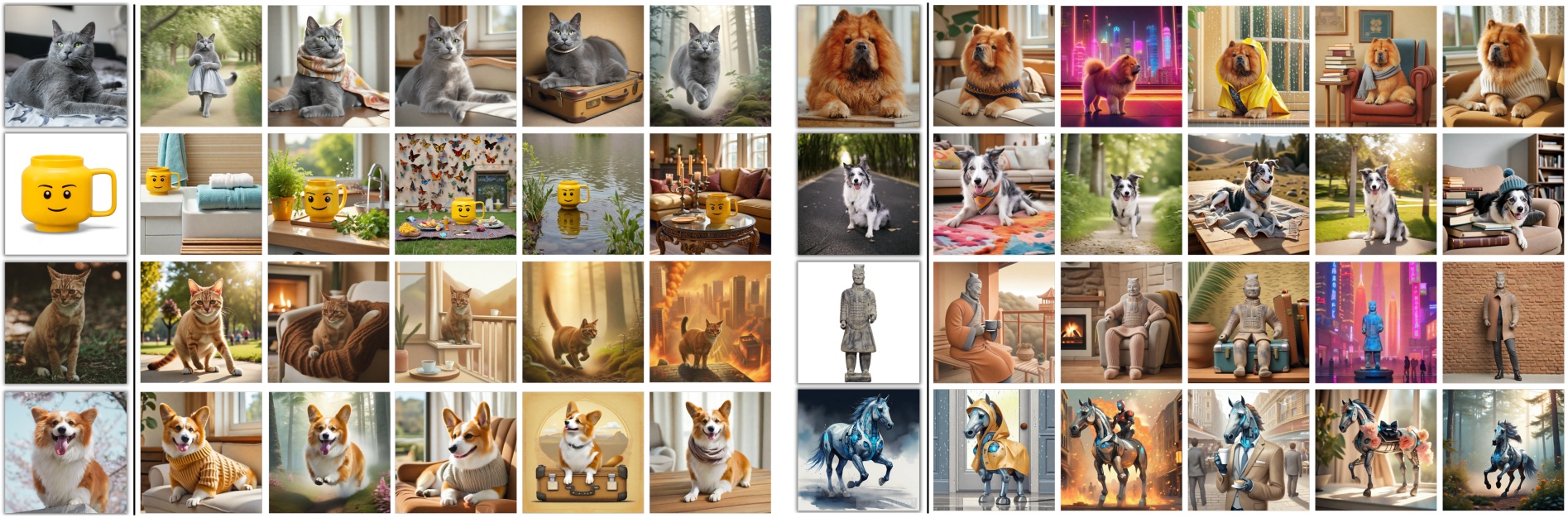


Our Solution:

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)



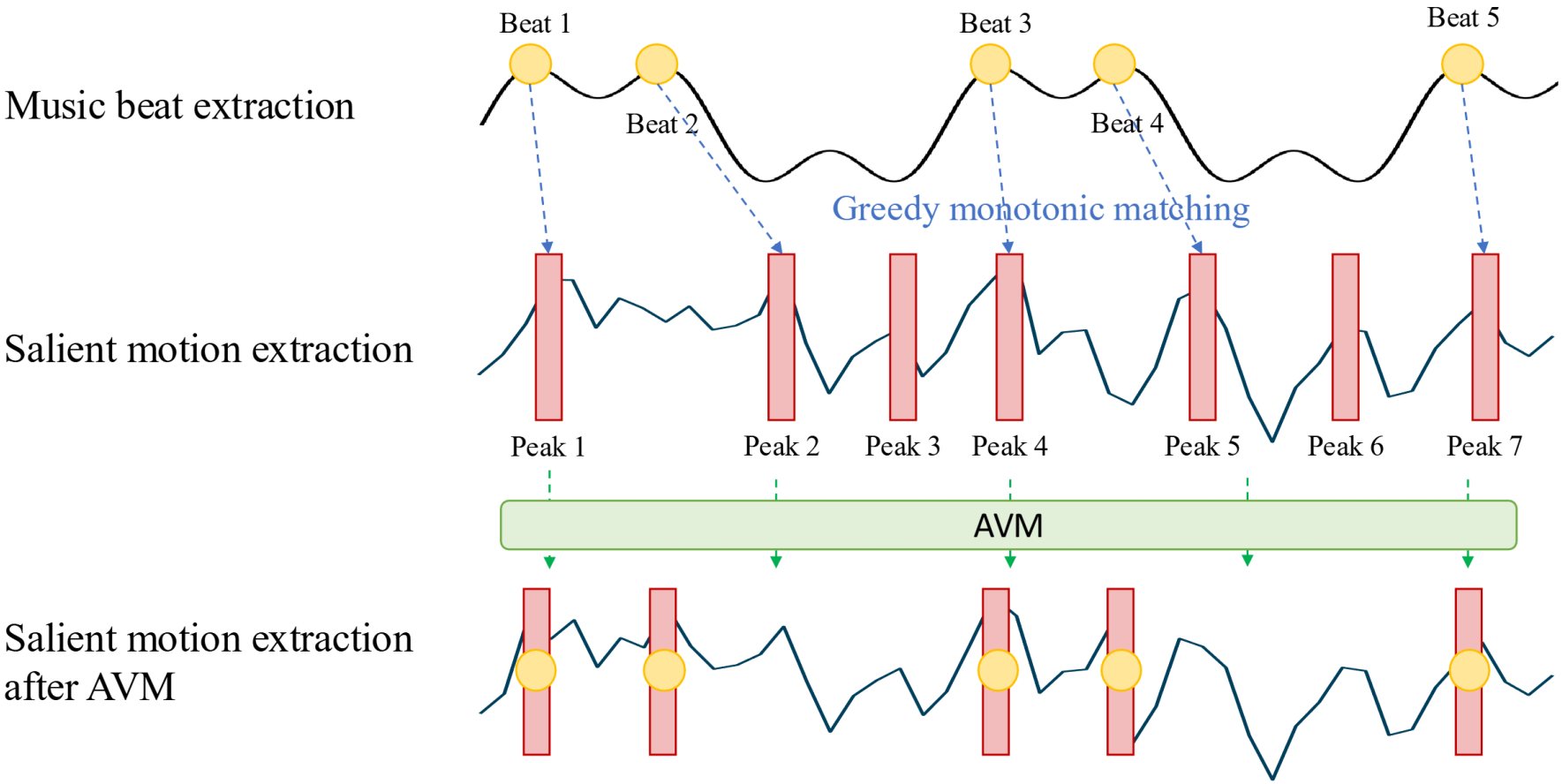
Conclusion

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.