Interactive Video Generation With Metadata Conditioning

James Baker
jbaker15@umbc.edu

1. Problem

Interactive Video Generation (10) is a form of video/image generation where each frame is conditioned on user input. This can be used to essentially encode a video game engine into an image generation model (8). The training sequences are created from videos of games being played by humans (3) or Reinforcement Learning agents (1; 8).

2. Methodology

Our approach consists of multiple models:

- Variational Autoencoder (6) trained on image frames from gameplay sequences
- Metadata prediction network g_ϕ trained to predict metadata at time t given metadata at t-1, t-2 and latent embedding frame at time t
- Latent diffusion (7) model, f_θ trained in two stages: 1) LoRA training on gameplay frames, the loss function is $\epsilon - f_\theta(Z_k + \epsilon)$, where $Z_k = VAE(X_k)$ is an embedded image 2) Training to denoise given past frames and metadata, the loss function is $\epsilon - f_\theta(Z_k + \epsilon | Z_{k-1:k-K}, m_k, a_{k-1})$, where m_k is the metadata for the image at time k, and a_{k-1} is the action taken at the prior stage

At inference, given Z_0, m_0, a_0, we will predict Z_1. We will then either choose an action or used some trained RL agent to choose $a_1|Z_1$. Then we will predict $m_1 = g_\phi(m_0, Z_0)$

3. Novelty

what has not been done is to use metadata from the video games being played, similar to (5). For example, it is possible to extract player location, camera location, score, etc. from video game simulators. This metadata could be used to condition the next frame prediction. An auxiliary model would be necessary to predict the metadata as it evolves in response to user action (for example, camera coordinates would likely change in response to a player moving their sprite). The auxiliary model would likely to some sort of RNN or transformer. This work would be novel in its use of game-specific metadata for the diffusion model, thus allowing it to learn a better world model, than previous works, which only conditioned the next frame on player actions and past frames.

4. Outcomes

Given a predefined sequence of frame, action pairs generated the same way as the training data, our approach should outperform a baseline that is ignorant of m_k. We can compare the MSE between the predicted and actual frames. Even if our method does not succeed, we will have at least compiled a large dataset of gameplay videos that future researchers may be able to use.

5. Milestones

These milestones are optimistic, but exact time estimates are tricky.

- 10/6 Use https://github.com/Farama-Foundation/stable-retro to train an RL agent to play *Sonic The Hedgehog*. Partition this into Training Set A, Training Set B and Testing
- 10/13 Train a few candidate sequential models g_ϕ on sequences in Training Set A. We could use LSTM (4), GRU (2) or Transformer (9)
- 10/20 Train a VAE based on https://github.com/huggingface/diffusers/blob/main/src/diffusers/models/autoencoders/autoencoder_kl.py to encode/embed images in Training Set A
- 10/27 Train LORA weights for a UNet based off of https://huggingface.co/SimianLuo/LCM_Dreamshaper_v7 on Training Set A, using the VAE from the prior step
- 11/3, 11/10 Add modules to UNet to condition on past frames + actions
- 11/17 Write up results?

References

[1] Eloi Alonso, Adam Jelley, Vincent Micheli, Anssi Kanervisto, Amos Storkey, Tim Pearce, and François Fleuret. Diffusion for world modeling: Visual details matter in atari, 2024. 1

[2] Kyunghyun Cho, Bart van Merrienboer, Dzmitry Bahdanau, and Yoshua Bengio. On the properties of neural machine translation: Encoder-decoder approaches, 2014. 1

[3] Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft, 2025. 1

[4] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. 1

[5] Samar Khanna, Patrick Liu, Linqi Zhou, Chenlin Meng, Robin Rombach, Marshall Burke, David Lobell, and Stefano Ermon. Diffusionsat: A generative foundation model for satellite imagery, 2024. 1

[6] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022. 1

[7] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 1

[8] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines, 2025. 1

[9] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023. 1

[10] Jiwen Yu, Yiran Qin, Haoxuan Che, Quande Liu, Xintao Wang, Pengfei Wan, Di Zhang, Kun Gai, Hao Chen, and Xihui Liu. A survey of interactive generative video, 2025. 1