COMBAT: Conditional World Models for Behavioral Agent Training

Anonymous ICCV submission

Paper ID *****

Abstract

Recent advances in video generation have spurred the development of world models capable of simulating 3D-consistent environments and interactions with static objects. However, a significant limitation remains in their ability to model dynamic, reactive agents that can intelligently influence and interact with the world. To address this gap, we introduce COMBAT, a real-time, action-controlled world model trained on the complex 1v1 fighting game Tekken 3. Our work demonstrates that diffusion models can successfully simulate a dynamic opponent that reacts to player actions, learning its behavior implicitly.

Our approach utilizes a 1.2 billion parameter Diffusion Transformer, conditioned on latent representations from a deep compression autoencoder. We employ state-of-the-art techniques, including causal distillation and diffusion forcing, to achieve real-time inference. Crucially, we observe the emergence of sophisticated agent behavior by training the model solely on single-player inputs, without any explicit supervision for the opponent's policy. Unlike traditional imitation learning methods, which require complete action labels, COMBAT learns effectively from partially observed data to generate responsive behaviors for a controllable Player 1. We present an extensive study and introduce novel evaluation methods to benchmark this emergent agent behavior, establishing a strong foundation for training interactive agents within diffusion-based world models.

1. Introduction

As the fidelity of video generation methods improves with increased understanding of real-world phenomena, interactive world models trained on gameplay and real-world data have emerged [4, 5, 23]. The focus of these works remains on generating spatially and temporally consistent world simulations. Yet, in real-world scenarios, the most unpredictable components are reactive agents that can observe, plan, and influence their environment, such as in autonomous driving, navigation, and combat scenarios.

Recent works demonstrate that autoregressive diffusion models are effective at world simulation. Several advances make these models real-time through distribution matching distillation (DMD) [26–28] and diffusion forcing [14] to overcome autoregressive drift. These engineering advances have enabled neural game simulations for first-person games such as Minecraft and CS:GO [17], showcasing excellent causal understanding of actions and their effects on generated frames.

However, real-world and game environments also contain rich information about how agents (humans, NPCs, and autonomous systems) respond to environmental dynamics. Current methods could greatly benefit from learning agent behavior from this observational data, but the partial observability and unstructured nature poses significant challenges. For example, while we might observe a pedestrian changing trajectory to avoid a vehicle, the exact observations and decision processes of the human agent remain hidden.

We present COMBAT (**C**onditional world **M**odel for **B**ehavioral **A**gent **T**raining), an interactive world model that learns underlying agent behavior and movement dynamics directly from partially observed multi-agent systems. By training a world model on Tekken 3 gameplay with conditioning only on Player 1's input, we observe emergent tactical behavior in Player 2 without explicit behavioral supervision. We select Tekken 3 as it provides an ideal controlled environment with clear visual feedback, deterministic game mechanics, diverse movesets, and frame-precise timing requirements.

Our approach uses a 1.2B parameter diffusion transformer trained on 1.2M frames across 1,000 gameplay rounds. We first train a Deep Compression AutoEncoder (DCAE) to obtain highly compressed latent representations, then train the world model to generate temporally consistent gameplay sequences. COMBAT successfully learns to control Player 1 from conditioning signals, while Player 2 emerges with realistic combat behaviors including blocking, counterattacking, and combo execution. Through decoder distillation and CausVid DMD techniques, we achieve real-time generation at interactive frame rates.

We introduce novel benchmarking methods to evaluate emergent agent behavior, measuring behavioral diversity,

Figure 1. An overview of the COMBAT world model. (Top) The model is conditioned on the current state (visual frames and poses) and Player 1's control inputs to autoregressively predict subsequent frames. (Bottom) Three distinct generated trajectories showcase the model's ability to produce plausible, strategic counter-attacks from Player 2 as an emergent response to Player 1's actions, without direct supervision of the opponent's policy.

and tactical understanding. Our extensive analysis demonstrates that world models can serve as a new paradigm for learning agent behaviors from observational data, with implications for multi-agent AI systems beyond gaming.

2. Related Work

Our work is positioned at the intersection of generative world models, video diffusion architectures, and behavioral modeling. We review key advancements in these areas to contextualize our contribution.

2.1. Video Diffusion Models

The remarkable success of diffusion models in image synthesis [19, 20] has naturally inspired their extension to video generation. Early approaches adapted U-Net architectures from image models, achieving results in short-form video synthesis [3, 9]. However, the convolutional nature of U-Net presents challenges for video: it struggles to capture long-range temporal dependencies and scales poorly with sequence length, often leading to temporal incoherence.

To address these limitations, Transformer-based video models have emerged. Following Peebles et al. [18], which demonstrated that Diffusion Transformers (DiT) could surpass U-Nets in image generation with superior scaling properties, subsequent work has applied this architecture to video. Models such as W.A.L.T [10] and CogVideoX [25] show that DiT self-attention mechanisms effectively model complex spatiotemporal relationships in video data, enabling longer, more coherent sequences. Our work builds on this foundation, employing a DiT backbone tailored for action-conditioned dynamics in interactive environments.

2.2. Neural Game Engines and World Models

Recent advances demonstrate that generative models can serve as neural game engines, replacing traditional rendering and state update logic. GameGAN, Kim et al. learns to imitate 2D games from raw pixels and actions using GANs with explicit memory modules [16]. More recently, diffusion transformers have become dominant for this task.

Valevski et al. introduce GameNGen, a fully neural DOOM engine that generates frames conditioned on past frames and actions, enabling real-time simulation [23]. Alonso et al.'s DIAMOND trains diffusion-based world models achieving state-of-the-art RL performance while producing playable Counter-Strike simulations [1]. Che et al. extend this with GameGen-X, training on million-clip datasets to enable long-horizon, interactive open-world gameplay [6].

These methods validate that neural models can learn complex game dynamics from observational data. Our work adopts similar architectural foundations but introduces a novel objective: modeling emergent behavior of uncontrolled opponents that arises solely from conditioning on controllable player actions.

2.3. Multi-Modal and Behavioral World Models

While traditional world models focus on visual prediction, recent work has pushed towards greater fidelity and behavioral learning. Our work adopts joint RGB-pose representation to enforce structural consistency in character movements.

In parallel, learning agent behavior within world models has predominantly followed two paths. The first is **model-based reinforcement learning**, where an agent's policy is trained using a learned dynamics model and an extrinsic reward signal. Works like **DreamerV3** exemplify this, achieving mastery in diverse domains by learning behaviors entirely within the latent space of a world model [11]. The second path is **imitation learning**, which learns policies from expert demonstrations. Methods like **Generative Adversarial Imitation Learning (GAIL)** require explicit state-action supervision for all agents to mimic expert behavior [13].

Our approach diverges from both paradigms. We demonstrate that complex, reactive multi-agent behaviors can emerge implicitly as a property of world modeling itself, without engineered rewards and using only partially observed data where just one agent's actions are provided as a condition.

2.4. Optimization Techniques for Interactive Generation

Real-time interactive generation requires addressing both architectural efficiency and sampling speed. Recent advances in attention mechanisms include FlexAttention [8], which enables flexible attention patterns, and Longformer [2], which combines local sliding-window attention with global context. We incorporate local-global attention patterns inspired by these works to balance efficiency with temporal coverage.

For sampling efficiency, Distribution Matching Distillation (DMD) [26, 28] and diffusion forcing [14] have proven effective at reducing sampling steps while mitigating autoregressive drift. These techniques enable real-time neural simulation for complex games [5, 23]. We adapt DMD through CausVid distillation to achieve interactive frame rates while preserving behavioral quality.

The Muon optimizer [15] introduces orthogonalization into momentum-based updates, improving conditioning of weight updates and outperforming AdamW in training speed benchmarks. We incorporate Muon optimization to enhance training efficiency of our large-scale diffusion transformers.

3. Method

Our approach, **COMBAT**, learns to simulate a complex, multi-agent environment by training a generative world model on video observations. World models have shown promise in mastering diverse domains [11] and creating interactive environments [5, 24]. We extend this paradigm to a competitive fighting game, where the model must learn the opponent's behavior without explicit action labels.

3.1. Problem Formulation

We frame our task as learning a conditional video generation model that implicitly captures an opponent's policy. We select the fighting game *Tekken 3* as our environment for three key reasons:

1. **Bounded Temporal Dependency**: The game state is largely Markovian, where

$$P(s_{t+1} \mid s_{\leq t}) \approx P(s_{t+1} \mid s_{t-k:t}),$$

for a small history window k, since all relevant information is contained within recent frames.
2. **Rich Action Space**: Characters possess diverse movesets, with over 40 unique actions and complex combos, providing a challenging domain for behavior modeling.
3. **Strategic Depth**: Success requires a blend of rapid reactions and long-term tactical planning.

Formal Problem Statement: Given a dataset of partially observed multi-agent trajectories

$$D = \{(s_t, a_t^{(1)}, s_{t+1})\}_{t=1}^{T},$$

where $s_t \in \mathbb{R}^{H \times W \times 3}$ is a game frame and $a_t^{(1)} \in \{0,1\}^8$ is the observed multi-hot input for Player 1. The actions of Player 2, $a_t^{(2)}$, remain unobserved. Our objective is to learn a conditional world model

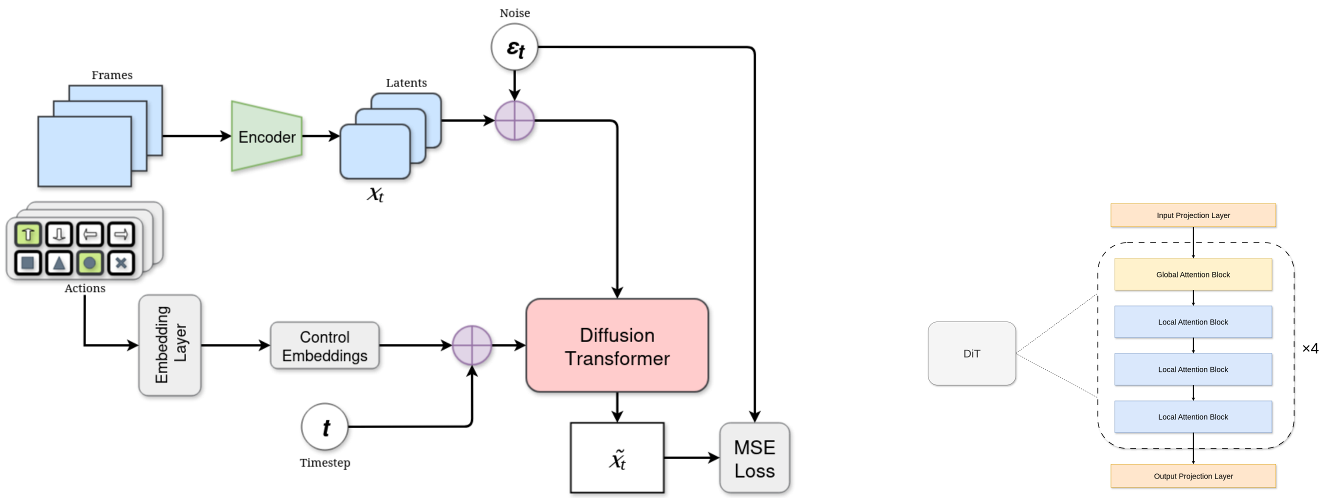
$$P_\theta(s_{t+1} \mid s_{t-k:t}, a_{t-k:t}^{(1)})$$

that can accurately predict subsequent frames.

Key Innovation: Unlike traditional imitation learning methods that require explicit action supervision for all agents [13], COMBAT is trained without Player 2's action labels. The model must infer Player 2's policy,

$$\pi^{(2)}(a_t^{(2)} \mid s_t, a_t^{(1)}),$$

as an emergent property of generating temporally consistent and plausible multi-agent interactions. This forces the world model to learn reactive and strategic opponent behavior implicitly.



(a) Training overview of COMBAT World Model.

(b) Every 4th DiT block has a global attention layer to capture long form context.

Figure 2. Architectural diagram of the COMBAT model. (a) The end-to-end training process, where a Diffusion Transformer is conditioned on action and timestep embeddings to denoise latent frame representations. (b) The internal structure of the DiT backbone, which employs a hybrid local-global attention pattern to efficiently model long-term dependencies.

3.2. Tekken 3 Gameplay Dataset

To train our model, we collected a large-scale dataset of *Tekken 3* gameplay, totaling 1,000 rounds (approximately 7 hours or 1.2 million frames). The data features a variety of characters and a balanced win–loss ratio between the two players. For each frame, captured at a resolution of $3 \times 448 \times 736$, we provide synchronized annotations, including: action inputs for both players, health and timer status, 68-point body pose coordinates, and player segmentation masks. Our data collection and annotation pipeline will be made publicly available.

3.3. Model Architecture

Our world model architecture integrates three main components:

- (1) a multi-modal variational autoencoder for high-ratio state compression,
- (2) an embedding module for player actions and diffusion timesteps, and
- (3) a Diffusion Transformer (DiT) backbone for autoregressive prediction in the latent space.

We train two versions of the model: one using only RGB latents and another using a joint visual–pose latent representation.

3.3.1. Multi-Modal Latent Encoding

To create an efficient latent representation, we first train a 340M-parameter joint RGB–pose variational autoencoder. This model learns a shared embedding space by compressing concatenated visual frames ($3 \times 448 \times 736$) and pose keypoints into a compact latent tensor of shape $128 \times 23 \times 11$. Our design is inspired by recent work in high-compression autoencoders for diffusion models [7]. To optimize for real-time performance, the 340M-parameter decoder is subsequently distilled to a 44M-parameter version by reducing its upsampling block count, which maintains high reconstruction quality at a fraction of the computational cost.

Player 1's action history, encoded as a multi-hot vector over 8 buttons, is projected into a dense embedding. This action embedding is summed with a sinusoidal time embedding for the current diffusion step, t_{emb}, to form the final conditioning vector for the DiT backbone.

3.3.2. Diffusion Transformer Backbone

The core of our generative model is a 1.2B-parameter Diffusion Transformer (DiT) [18], which learns to denoise and predict future latent frames. The architecture consists of 16 transformer blocks with a model dimension $d_{\text{model}} = 2048$ and 16 attention heads. The conditioning vector is injected into each block via an Adaptive Layer Normalization Zero (AdaLNZero) layer, and tokenization is performed using linear projection layers for spatio-temporal rasterization, bypassing conventional patch-based embeddings.

Each DiT block executes the following sequence:

$$\text{AdaLN} \rightarrow \text{Attention} \rightarrow \text{Gated Residual} \rightarrow$$
$$\text{AdaLN} \rightarrow \text{MLP} \rightarrow \text{Gated Residual}$$

To maintain computational tractability over long 128-frame sequences, we employ a hybrid attention strategy. Most layers use a frame-causal attention mask with a local sliding window of 16 frames, while every fourth layer applies global attention across the entire 128-frame context.

This structure balances long-range dependency modeling with computational efficiency. We apply Rotary Position Embeddings (RoPE) [21] across both spatial and temporal axes and utilize FlexAttention for an efficient block-sparse masking implementation.

3.4. Accelerated Inference for Real-Time Generation

Enabling real-time interaction is critical for gaming applications, but the iterative sampling process of diffusion models is computationally intensive. To overcome this, we significantly accelerate inference using two key optimizations.

First, we distill the fully trained model into a few-step sampler using Distribution Matching Distillation (DMD) [26, 27]. Specifically, we adopt the CausVid DMD framework [28] to produce a 4-step distilled model that preserves high generative fidelity while drastically reducing inference time.

Second, we further enhance speed by implementing static key-value caching, which reuses previously computed attention states across generation steps. These optimizations are applied to both the RGB and visual–pose world models.

4. Experiments

To validate our claim that a conditional world model can learn reactive agent behavior from partial observations, we conduct a series of experiments on the Tekken 3 dataset. We first detail our multi-stage training pipeline and model architectures. We then introduce our evaluation benchmarks and present results comparing our primary models and their distilled variants.

4.1. Implementation Details

Our training process is divided into three main stages: autoencoder training, world model training, and distillation for real-time inference. All models were trained on a cluster of 8× NVIDIA H200 GPUs.

Stage 1: Autoencoder Training. We first train a 340M parameter Deep Compression AutoEncoder (DCAE) to learn a compact latent representation of the game environment. The autoencoder is trained for 68,000 steps (approx. 75 hours) on our 1.2 million frame Tekken dataset. It compresses raw frames ($3 \times 448 \times 736$) into a latent space of 23×11 with 128 channels. The training objective is a combination of L2 reconstruction loss, perceptual similarity loss, and a KL divergence term to regularize the latent space. For our pose-augmented model, we use an identical architecture and training setup.

Stage 2: World Model Training. We train a 1.2B parameter autoregressive Diffusion Transformer (DiT) to function as the world model. The DiT architecture consists of 16 layers, 16 attention heads, and a model dimension of $d_{model} = 2048$. It employs a combination of local (16 frames) and global (128 frames) attention windows to capture both short-term and long-term temporal dependencies. The model is trained on video clips with a sequence length of 128 frames to predict the next latent frame conditioned on Player 1's actions. We train two distinct world models: one using latents from the RGB-only VAE and another using latents from the pose-augmented VAE.

Stage 3: Distillation for Real-Time Inference. To achieve interactive frame rates, we employ two separate distillation techniques:

- **Decoder Distillation:** We first create a lightweight VAE decoder for real-time rendering. Using student-teacher distillation, we reduce the number of upsampling blocks per stage in the decoder from four to one. This process, which took 14 hours over 50k steps, reduces the decoder's parameter count from 340M to a nimble 44M.
- **Step Distillation:** We use CausVid, a Distribution Matching Distillation (DMD) method, to drastically reduce the number of required inference steps for the world model. We distill the fully-trained DiT into a 4-step variant. This distillation process converges in 2,500 steps, utilizing a combination of a DMD loss and a critic loss. We apply this technique to both the RGB-only and the pose-augmented world models.

4.2. Evaluation Metrics and Benchmarks

Evaluating emergent agent behavior presents a fundamental challenge: how do we measure intelligence that was never explicitly supervised? Traditional video metrics assess visual fidelity, while RL metrics assume access to ground-truth actions or rewards. Since COMBAT learns behavioral patterns implicitly through world modeling, we need novel evaluation approaches that can detect tactical competence from generated gameplay alone.

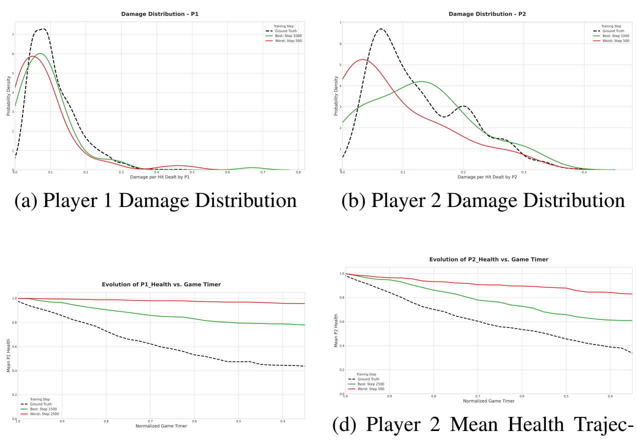
4.2.1. Standard Perceptual Metrics

To assess the perceptual quality of our generated trajectories, we employ a suite of standard metrics. Our evaluation protocol involves conditioning the models on real Player 1 action sequences extracted from a test set of 300 ground-truth videos(1-2 seconds) consisting mixed difficulty gameplays. The generated video is then compared directly against its corresponding ground-truth counterpart from which the actions were sourced. This setup provides a stringent test of the model's ability to render deterministic outcomes based on specific actions,a significantly more challenging task than unconditional video generation.

We report the Fréchet Video Distance (FVD)[22] to measure temporal coherence, the Fréchet Inception Distance (FID)[12] for per-frame visual fidelity, and LPIPS to quantify perceptual similarity. Given the high-fidelity nature of the Tekken 3 environment, characterized by rapid motion and complex visual effects, achieving strong performance on these metrics against the ground truth is a robust indicator of the model's precision and world-modeling capabilities.

Table 1. All metrics are calculated on a held-out test set of 300 video clips each with 32 frames. Lower is better for all scores.

Model	FID ↓	FVD ↓	LPIPS ↓
COMBAT: Pose	49.7	593.4	0.05
COMBAT: Non-Pose	80.9	1156.6	0.07



(a) Player 1 Damage Distribution (b) Player 2 Damage Distribution

(c) Player 1 Mean Health Trajectory (d) Player 2 Mean Health Trajectory

Figure 3. **Behavioral Consistency Metrics.** A comparison of generated gameplay (COMBAT) against the ground truth. **(a, b)** The per-frame damage distributions for Player 1 and Player 2, showing that our model learns a realistic mapping of actions to consequences. **(c, d)** The mean health trajectories over the course of a round, indicating that COMBAT captures the natural pacing of a match.

4.2.2. Behavioral Consistency Metrics

To verify that our model learns the game's intrinsic rules and pacing, we propose two metrics based on in-game health data:

- **Damage Distribution Analysis:** This metric assesses whether the consequence of individual actions is realistic. Let $H_i^{(t)}$ denote the health of player $i \in \{1, 2\}$ at frame t, and define per-frame damage as $\Delta H_i^{(t)} = \max(0, H_i^{(t-1)} - H_i^{(t)})$. We normalize by the maximum health $H_i^{\max}$ to obtain $\delta_i^{(t)} = \Delta H_i^{(t)} / H_i^{\max}$.
  The complete distribution of damage values from all generated sequences, $\{\delta_{i,\text{gen}}^{(t)}\}$, is then compared to the distribution from all ground-truth sequences, $\{\delta_{i,\text{real}}^{(t)}\}$, using the Wasserstein distance. A lower distance signifies that the model has learned a more accurate mapping from actions to their in-game consequences.
- **Health Trajectory Analysis:** This metric evaluates the overall temporal flow of the match. Define the normalized time $s = t/T$, where T is the total round duration, and let $\bar{H}^{(s)} = \frac{1}{2}\sum_i H_i^{(t)}/H_i^{\max}$ be the average normalized health at time s for a single round.
  To establish a baseline for typical match progression, we compute the **mean health trajectory** by averaging $\bar{H}^{(s)}$ across all rounds in our ground-truth test set. We do the same for our generated rounds. The similarity between these two mean trajectories is then measured using the Mean Squared Error (MSE). A lower MSE indicates that the generated gameplay, on average, exhibits a more realistic match pace.

4.3. Human Evaluation of Emergent Behavior

To assess the emergent behavior of Player 2, we conduct human evaluation based on observable action patterns in gameplay. Since Player 2 is trained without explicit supervision, emergent behavior is defined as actions that react naturally to Player 1's inputs, demonstrating plausible combat strategies such as timely punches, kicks, and defensive maneuvers.

We introduce two human-interpretable metrics: **Total Action Adherence (TAA)** and **Action Ratio Consistency (ARC)**. These metrics are based on human annotations of offensive actions observed in both ground-truth and generated gameplay sequences.

4.3.1. Total Action Adherence (TAA)

TAA measures whether the agent produces a comparable overall volume of offensive actions relative to human gameplay:

$$\text{TAA} = \frac{G_{\text{kicks}} + G_{\text{punch}}}{O_{\text{kicks}} + O_{\text{punch}}}$$

where $G.$ denotes actions performed by the generated agent, and $O.$ the actions performed in original gameplay.

A score of 1.0 indicates perfect adherence in activity level. Scores > 1.0 suggest hyperactive behavior, while scores < 1.0 indicate passive behavior.

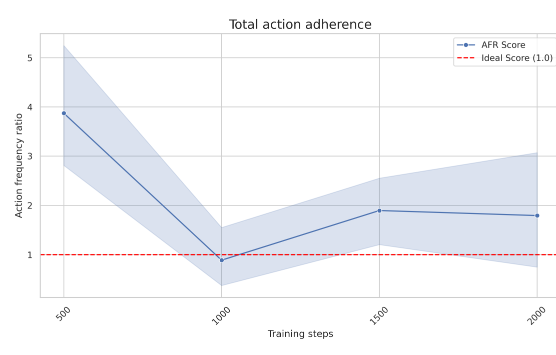


Figure 4. Total Action Adherence across training checkpoints

4.3.2. Action Ratio Consistency (ARC)

ARC evaluates whether the stylistic balance between punches and kicks aligns with the human player:

$$\text{ARC} = \frac{\frac{G_{\text{punch}}}{G_{\text{kicks}}}}{\frac{O_{\text{punch}}}{O_{\text{kicks}}}}$$

A score of 1.0 indicates identical punch-to-kick ratio as original gameplay. Scores above 1.0 reflect stronger preference for punches, while scores below 1.0 suggest heavier reliance on kicks.

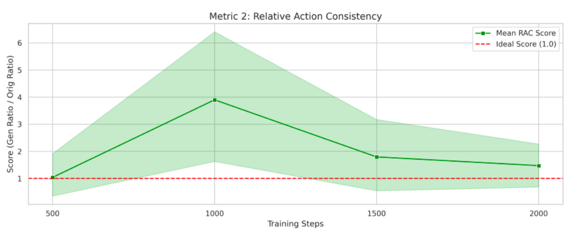


Figure 5. Action Ratio Consistency across training checkpoints

4.3.3. Results

We evaluated sequences at multiple training checkpoints. Table 2 summarizes the results:

Training Step	TAA	ARC
Ground Truth	1.00	1.00
Step 500	3.87	1.04
Step 1000	0.88	3.90
Step 1500	1.90	1.79
Step 2000	1.79	1.47

Table 2. TAA and ARC scores at different training checkpoints compared against human gameplay.

Our evaluation shows that COMBAT successfully learns emergent Player 2 behavior through distinct phases. Initially, the model is hyperactive, generating nearly four times the offensive actions of human players (TAA = 3.87), though its punch-to-kick ratio is well-aligned (ARC = 1.04). As training progresses, the model reduces hyperactivity in further steps. Beyond step 2000, performance declines, with later checkpoints showing reduced adherence to original gameplay.

By the final training stages, the model converges toward stable, human-like combat patterns. It learns to regulate activity frequency (TAA 1.8) while achieving balanced fighting style (ARC 1.5). However, overall consistency degrades noticeably. This progression from erratic behavior to stable patterns demonstrates that complex, emergent behaviors can be learned without explicit supervision.

The pose-augmented COMBAT model significantly outperforms the RGB-only variant across visual quality metrics, confirming that explicit pose information improves generation quality.

Impact of Distillation: Our 4-step distilled models, created using CausVid DMD, retain substantial visual quality while achieving 12.5× speedup. The pose-augmented 4-step model still outperforms the full RGB-only model, demonstrating efficient distillation with minimal quality trade-off.

Qualitatively, we observe intelligent behaviors including combo execution, spatial awareness, and adaptation to Player 1's patterns. These tactical responses emerge naturally from our training process without explicit behavioral supervision.

5. Conclusion

In this work, we introduce COMBAT, a conditional world model that learns complex, emergent agent behavior from partially observed gameplay. Our key finding is that by conditioning the model solely on Player 1's actions, it successfully learns a reactive, tactically coherent policy for Player 2 without any direct supervision. The model correctly associates the control inputs with the intended agent and generates plausible counter-attacks, demonstrating that intricate behaviors can arise implicitly from the objective of temporal consistency.

To foster further research in this domain, we provide an extensive analysis of emergent behavior in world models. We will also release our large-scale **Tekken 3 dataset**, complete with synchronized pose and segmentation annotations, and **open-source our pipelines** for data collection and model training.

Crucially, our approach is practical for interactive applications. Through distillation, the COMBAT world model achieves **real-time performance, operating at 85 FPS on a single NVIDIA A100 GPU**. This work represents a first step in exploring how generative world models can learn implicit agent policies, and we hope it inspires further research into multi-agent behavioral modeling in complex, interactive environments.

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. In *Thirty-eighth Conference on Neural Information Processing Systems*, 2024. 2

[2] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv:2004.05150*, 2020. 3

[3] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, Varun Jampani,

and Robin Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets, 2023. 2

[4] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024. 1

[5] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Bechtle, Feryal M. P. Behbahani, Stephanie Chan, Nicolas Manfred Otto Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando de Freitas, Satinder Singh, and Tim Rocktaschel. Genie: Generative interactive environments. *ArXiv*, abs/2402.15391, 2024. 1, 3

[6] Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[7] Junyu Chen, Han Cai, Junsong Chen, Enze Xie, Shang Yang, Haotian Tang, Muyang Li, and Song Han. Deep compression autoencoder for efficient high-resolution diffusion models. In *The Thirteenth International Conference on Learning Representations*, 2025. 4

[8] Juechu Dong, Boyuan Feng, Driss Guessous, Yanbo Liang, and Horace He. Flex attention: A programming model for generating optimized attention kernels, 2024. 3

[9] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *The Twelfth International Conference on Learning Representations*, 2024. 2

[10] Agrim Gupta, Lijun Yu, Kihyuk Sohn, Xiuye Gu, Meera Hahn, Li Fei-Fei, Irfan Essa, Lu Jiang, and José Lezama. Photorealistic video generation with diffusion models, 2023. 2

[11] Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models, 2024. 3

[12] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, Günter Klambauer, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a nash equilibrium. *CoRR*, abs/1706.08500, 2017. 5

[13] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*. Curran Associates, Inc., 2016. 3

[14] Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. *arXiv preprint arXiv:2506.08009*, 2025. 1, 3

[15] Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. 3

[16] Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, and Sanja Fidler. Learning to Simulate Dynamic Environments with GameGAN. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2

[17] Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition, 2025. 1

[18] William Peebles and Saining Xie. Scalable diffusion models with transformers. *arXiv preprint arXiv:2212.09748*, 2022. 2, 4

[19] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[20] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[21] Jianlin Su, Yu Lu, Shengfeng Pan, Ahmed Murtadha, Bo Wen, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding, 2023. 5

[22] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. FVD: A new metric for video generation, 2019. 5

[23] Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Representation Learning*, pages 73754–73776, 2025. 1, 2, 3

[24] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023. 3

[25] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, Da Yin, Yuxuan.Zhang, Weihan Wang, Yean Cheng, Bin Xu, Xiaotao Gu, Yuxiao Dong, and Jie Tang. Cogvideox: Text-to-video diffusion models with an expert transformer. In *The Thirteenth International Conference on Learning Representations*, 2025. 2

[26] Tianwei Yin, Michaël Gharbi, Taesung Park, Richard Zhang, Eli Shechtman, Fredo Durand, and William T Freeman. Improved distribution matching distillation for fast image synthesis. In *NeurIPS*, 2024. 1, 3, 5

[27] Tianwei Yin, Michaël Gharbi, Richard Zhang, Eli Shechtman, Frédo Durand, William T Freeman, and Taesung Park. One-step diffusion with distribution matching distillation. In *CVPR*, 2024. 5

[28] Tianwei Yin, Qiang Zhang, Richard Zhang, William T Freeman, Fredo Durand, Eli Shechtman, and Xun Huang. From slow bidirectional to fast autoregressive video diffusion models. 2025. 1, 3, 5

COMBAT: Conditional World Models for Behavioral Agent Training

Supplementary Material

6. Rationale

Having the supplementary compiled together with the main paper means that:

- The supplementary can back-reference sections of the main paper, for example, we can refer to Sec. 1;
- The main paper can forward reference sub-sections within the supplementary explicitly (e.g. referring to a particular experiment);
- When submitted to arXiv, the supplementary will already included at the end of the paper.

To split the supplementary pages from the main paper, you can use Preview (on macOS), Adobe Acrobat (on all OSs), as well as command line tools.