LEARNING TO ACT AND EDIT IN VIDEO WORLDS

Anonymous authors
Paper under double-blind review



Figure 1: Cela-World-180K unifies multi-source video data spanning Action, Navigation, and Editing categories. This dataset allows Cela-World-5B to unify navigation, character actions, and event-level editing across diverse scenes.

ABSTRACT

Interactive video world models have made substantial progress in generating explorable environments with navigation control. However, extending such control to diverse character actions and environmental events is hindered by fragmented supervision across heterogeneous video sources. To bridge this gap, we develop a unified data pipeline that transforms real-world videos, gameplay video, edited videos and synthetic videos into a temporally aligned dataset for camera control, character actions, and environmental events. **The resulting dataset contains 180,162 clips, covering 91 action types across 14 visual domains.** Using this dataset, we train an autoregressive world model initialized from a pretrained bidirectional video generator. Experiments show that the resulting model (i) achieves strong instruction following across everyday interactions, gameplay actions, visual effects, and scene-level edits while preserving spatial navigation fidelity, (ii) exhibits zero-shot compositional generalization across unseen instruction combinations alongside precise camera-actor disentanglement, and (iii) reveals the distinct roles of different data sources, yielding a reproducible, resource-efficient recipe for training semantically controllable video world models. We will release the code and dataset.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

Interactive video world models turn generative video models into environments whose future can be steered during generation. Conditioned on camera trajectories or keyboard inputs, these models synthesize continuous visual rollouts as users explore interactive scenes (Wang et al., 2026c; Zhu et al., 2026a; Zhao et al., 2026a). While navigation provides a clean test of spatial coherence and long-horizon consistency, full interactive world modeling requires fine-grained, in-scene control. Executing precise interactions, such as manipulating objects, operating vehicles, or triggering environmental events, requires that a model selectively modify specific entities at specified timestamps while maintaining global context. Incorporating these semantic controls elevates world models from passive visual simulators to active substrates for embodied AI , interactive gaming, and controllable video synthesis.

Recent works demonstrate this shift by introducing language-guided character actions, and prompt-driven events (Xiong et al., 2026; Gao et al., 2026b; Team et al., 2026a;b). However, scaling these capabilities reveals a fundamental data bottleneck, as effective supervision remains scattered across heterogeneous sources. Camera trajectories provide broad spatial coverage (Li et al., 2026b), egocentric videos capture close-up manipulations (Xiong et al., 2026), gameplay recordings supply discrete control signals (Guo et al., 2026), and edited videos isolate localized state changes. Compounding this heterogeneity, these sources feature highly fragmented supervision, ranging from continuous navigation trajectories and discrete keypresses to noisy global captions, while dense temporal annotations remain critically scarce. Unifying these disparate, weakly annotated sources into a coherent training target demands a standardized data pipeline that filters noisy inputs and structures multimodal controls into temporally grounded labels, explicitly disentangling *what* changed, *when* it occurred, and *which* entity (*camera, character, or environment*) governed the transition.

To address this challenge, we develop a unified data curation pipeline alongside an accompanying dataset tailored for interactive video world modeling. Our pipeline ingests heterogeneous camera-motion sequences, egocentric interactions, third-person gameplay recordings, edited videos, and synthetic data systematically structuring their supervision across three distinct control axes: camera motion, character action, and environmental event. By filtering low-quality segments and assigning temporally grounded control annotations, the pipeline converts unaligned raw inputs into a standardized training dataset. The resulting dataset comprises 180,162 clips spanning 91 action types and 14 visual domains, establishing a solid foundation for unified world model training.

Leveraging this dataset, we train Cela-World-5B, a chunk-autoregressive world model initialized from Wan2.2 TI2V-5B (Wan et al., 2025), with geometric camera conditioning and chunk-aligned semantic prompts. On the WBench Navi split (Ying et al., 2026), our model achieves an overall score of 76.9, exceeding the Wan 2.2 baseline by 7.8 points and reaching 81.6 in interaction adherence. On the Full split, it improves the overall score from 61.1 to 69.8. Our data analysis further demonstrates that action priors learned from gameplay effectively transfer to real-world scenarios. Furthermore, with only minimal data post-training, our model successfully generalizes and composes diverse action and editing skills across varied scenes, viewpoints, and subjects. Together, these findings show how temporally aligned, multi-source supervision strengthens semantic control while preserving navigation fidelity.

In summary , our contributions can be summarized as follows:

❶ **A Unified Data Pipeline and Dataset**: We present a standardized data curation pipeline that cleans, filters, and annotates heterogeneous multi-modal inputs into a unified training corpus across three distinct control axes (camera motion, character action, and environmental event).

❷ **Model Training and Validation**: We train an autoregressive world model initialized from Wan 5B, validating both our end-to-end training methodology and the efficacy of our dataset design in fostering robust world-modeling capabilities.

❸ **Empirical Data Insights**: We systematically analyze the impact of different data compositions on world-model training, demonstrating how multi-source temporal supervision jointly enhances semantic control, action responsiveness, and navigation fidelity.

2 Related Work

Interactive World Models and Control Interfaces. Early video world models established interactive generation in constrained games using learned latent actions or game-native commands (Bruce et al., 2024; Valevski et al., 2025; Yu et al., 2025; Guo et al., 2025). As video generators matured, this paradigm expanded toward open-domain navigation, with increasingly precise camera control (Huang et al., 2026a; Zhao et al., 2026c; Zhang et al., 2026), real-time autoregressive streaming (He et al., 2026; Zhao et al., 2026a), and memory mechanisms for longer, more consistent rollouts (Sun et al., 2026; Qian et al., 2026; Gao et al., 2026a). These systems are supported by a common causal-generation stack that converts bidirectional video priors into causal or chunk-autoregressive models through ODE or consistency distillation, distribution matching, and training on model-generated histories (Chen et al., 2024; Huang et al., 2026c; Yin et al., 2025; Zhu et al., 2026b; Zhao et al., 2026b; Huang et al., 2025b). Together with frame-aligned geometric and action conditioning (He et al., 2025; Li et al., 2026a; Taghavi et al., 2026), this stack provides a mature basis for online navigation and control.

Recent architectures expand this interface toward natural language instructions, enabling real-time steering via prompts (Xiong et al., 2026; Gao et al., 2026b; Team et al., 2026b;a; Zhu et al., 2026c; Chen et al., 2026). However, existing approaches lack a unified, open end-to-end data and training pipeline, leaving the relationship between multi-source data compositions, supervision granularities, and the resulting control capabilities unexamined. To bridge this gap, we build an integrated pipeline as a controlled testbed to study how data strategies shape controllability.

Data Curation and Temporally Grounded Supervision. Training video world models requires balancing spatial coverage with control-aligned annotations. Early datasets focused predominantly on global clip captions and camera trajectories, offering geometric diversity but lacking temporal alignment for fine-grained entity behaviors (Grauman et al., 2022; 2024; Li et al., 2026b; Wang et al., 2026b; Huang et al., 2026b). To incorporate explicit temporal boundaries, subsequent data efforts leveraged instrumented game engines, physics simulators, or edited video pairs to align actions and events with discrete time spans (Wu et al., 2025; Guo et al., 2026; Li et al., 2026c; Hu et al., 2026; Wang et al., 2026a; Shen et al., 2026). Modern interactive systems synergize egocentric, gameplay, synthetic, and real-world video into multi-source mixtures with multi-track or segment-level temporal labels (Xiong et al., 2026; Gao et al., 2026b; Team et al., 2026b;a). Nevertheless, existing data recipes remain tightly coupled with specific model backbones and scheduling heuristics, obscuring the isolated impact of source composition and label granularity. We address this issue by constructing a standardized, temporally grounded representation across heterogeneous sources, systematically isolating how dataset mixtures, supervision density, and scheduling strategies determine navigation, action, and event control.

3 Data Curation and Representation

To train an interactive video world model capable of supporting fine-grained, multi-axis controls, we develop a standardized five-stage data construction pipeline (Fig. 2). Specifically, we first present *Source Collection, Filtering and Mining* (§3.1), elaborate on *Action Localization & Clip Extraction*(§3.2) together with *Multi-level Captioning* (§3.3), and describe *Dataset Curation & Verification* (§3.4). Finally, we summarize the resulting *Dataset Statistics* (§3.5).

3.1 Source collection, Filtering and Mining

To encompass diverse visual scenes, agent interactions, and environmental dynamics, we curate three complementary video sources.

- **Real-world videos** from SpatialVID (Wang et al., 2025) provide authentic visual appearances, complex illumination, and realistic camera trajectories.
- **Gameplay recordings** from OpenPixel2Play (Yue et al., 2026) and Gaming-500-Hours (Markov AI, 2026) provide frame-aligned keyboard and mouse actions for fine-grained interactive control, while ABot-World-Explorer (Jiang et al., 2026), EmbodiedWorld-200K (Xiaokunfeng2022, 2026), OmniWorld-Game (Zhou et al., 2025), and Sekai-Game-Drone (Li et al., 2025) contribute long-horizon continuous navigation and diverse viewpoints.

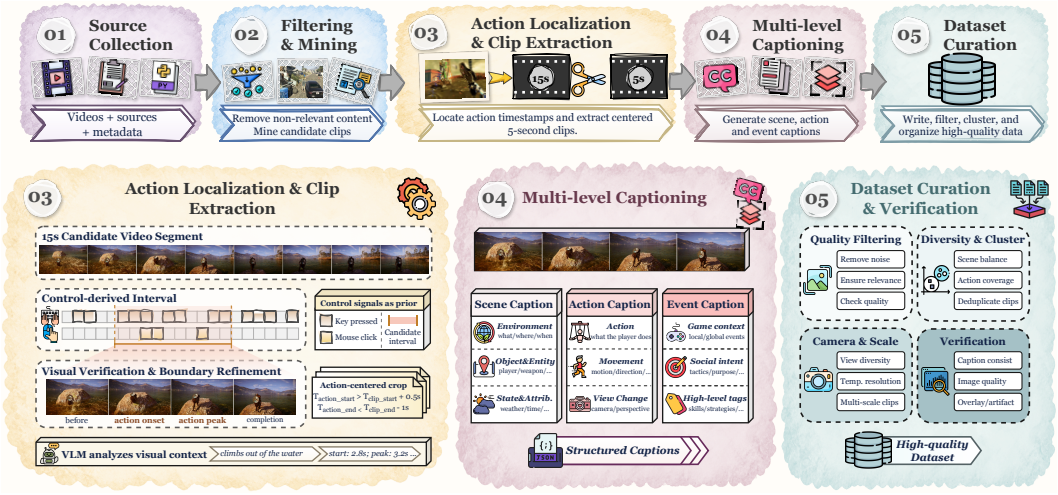

Figure 2: **Data construction pipeline.** Multi-source videos undergo content filtering and candidate segment mining. When available, gameplay input logs supply coarse action intervals. A VLM verifies action visibility and refines frame boundaries to extract action-centered five-second clips. Next, a dedicated captioning VLM annotates multi-level descriptions covering scene entities, subject movements, camera dynamics, and environmental events. Source-specific quality filtering, caption validation, diversity rebalancing, and deduplication finalize the dataset.

- **Synthetic videos** generated with MiniMax H3 (MiniMax, 2026) supplement the corpus with localized visual effects, dynamic object appearances, and environmental transitions.

Filtering and candidate mining. For gameplay recordings with synchronized keyboard and mouse logs, we identify potential player interactions by parsing key presses, releases, and mouse clicks into temporal event sequences. By filtering out idle intervals and menu operations, we estimate coarse action intervals. Because these inputs serve only as approximate temporal cues, we extract 15-second candidate segments around the estimated intervals to prevent action truncation and provide sufficient contextual buffer for subsequent visual verification. Conversely, for exploration and real-world videos lacking control telemetry, candidate segments are derived directly from source-provided splits or caption-aligned timestamps. Synthetic videos directly inherit their native action/event annotations and generation metadata without additional inference.

3.2 Action Localization and Clip Extraction

For the 15-second gameplay candidates mined in §3.1, we localize atomic interactions and extract action-centered 5-second clips through two steps:

(1) Action Localization. A Vision-Language Model (VLM) inspects timestamped frames within each candidate segment to verify the interaction against coarse control priors. Compensating for input latency and recovery lag, the VLM determines the exact visual boundaries—including preparation, onset, execution peak, and completion—while identifying boundary uncertainty and excluding truncated actions.

(2) Clip extraction. Guided by the refined boundaries, we extract an 81-frame, 5-second clip ($T_{\text{clip}} = 5\,\text{s}$) centered on the target action. To preserve complete physical transitions, the crop enforces $\geq 0.5\,\text{s}$ of preceding context and $\geq 1.0\,\text{s}$ of post-action recovery margin. Targets lacking feasible margins are pruned, and localized timestamps are projected onto the output clip's timeline.

3.3 Multi-Level Captioning

To disentangle static environment appearance from chunk-specific interaction dynamics, a dedicated VLM synthesizes structured captions across three complementary semantic axes: **(1) Scene captions** are grounded on the initial frame of each temporal chunk to describe the persistent spa-

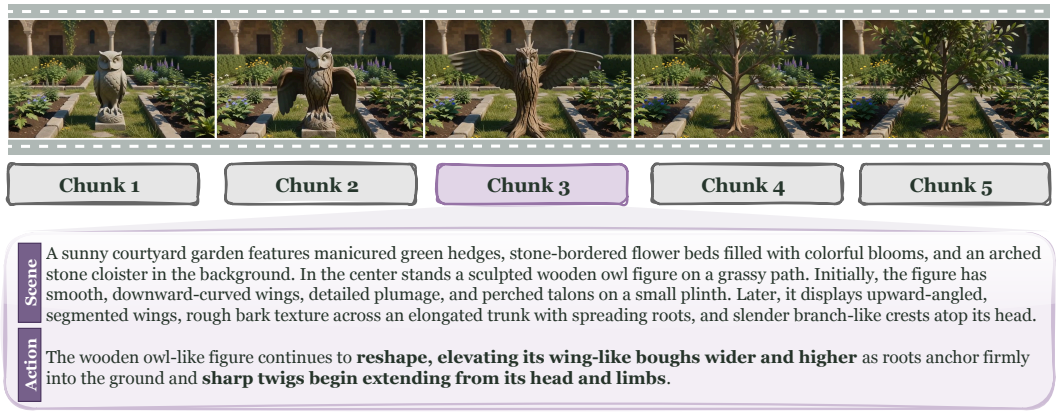

Figure 3: **Example of a chunk prompt.** For a representative video chunk, the scene caption details persistent environmental context and object attributes, whereas the action caption explicitly isolates local state transitions and motion dynamics. Disentangling invariant background context from chunk-specific action dynamics provides precise temporal supervision for localized control. Since this chunk contains no event, its prompt includes only scene and action captions.

tial layout, terrain, objects, and environmental attributes (e.g., lighting and weather). **(2) Action captions** examine sampled frame sequences to capture fine-grained subject behavior, locomotion direction, interaction targets, and viewpoint shifts. **(3) Event captions** characterize high-level tactical contexts, localized visual effects, and dynamic state transitions. We associate localized actions with their intersecting chunks; multi-chunk actions retain local phase descriptions while preserving temporal continuity via instance identifiers.

3.4 Dataset Curation and Verification

Following captioning, we apply rigorous quality control and multi-modal validation to construct the final training data: **(1) Visual quality filtering:** clips are screened to eliminate exposure anomalies, severe letterboxing, UI/subtitle overlays, and shot breaks. **(2) Annotation and consistency check:** we verify chunk ordering, temporal bounds, and cross-modal consistency to prune clips with action hallucinations or temporal misalignment. **(3) Diversity rebalancing and deduplication:** perceptual hashing and trajectory clustering are applied to rebalance scene and action coverage while removing visually redundant clips. **(4) Camera geometry validation:** poses and intrinsics are verified for numerical validity and aligned with video frames, adjusting for spatial crops and supplementing missing trajectories via ViPE (Huang et al., 2025a) under explicit provenance flags.

3.5 Dataset Statistics

Cela-World-180K contains 180,162 clips covering 91 action categories across 14 visual domains, including recordings from 50 different games and both first- and third-person viewpoints. As shown in Fig. 4, the dataset features diverse actions, ranging from movement and posture to combat, equipment use, and object interaction. It combines game recordings, real-world videos, and synthetic videos, with gameplay spanning open-world simulation, adventure, competitive combat, and cooperative action. Navigation, action, and event content are represented across both viewpoints. Scene settings encompass indoor, urban, natural, industrial, and coastal environments, while environmental events include weather and lighting changes, fire, smoke, and changes in object appearance, material, and structure.

4 World-Model Training

We train a video world model to validate the end-to-end viability of our constructed dataset and its temporally aligned controls. The training follows a three-stage pipeline: bidirectional adaptation, block-autoregressive conversion, and few-step causal distillation.

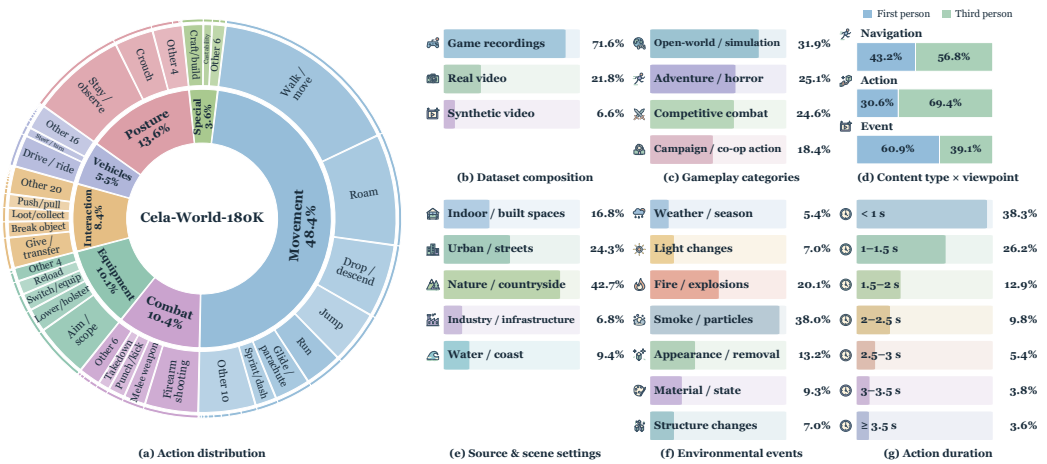

Figure 4: Dataset statistics of Cela-World-180K. (a) Hierarchical distribution of action-label assignments. (b) Video-source composition. (c) Gameplay categories among named-game clips. (d) Viewpoint distributions for navigation, actions, and environmental events. (e) Scene categories. (f) Environmental-change types. (g) Action-duration distribution.

4.1 Model Architecture

Video backbone. We initialize our model from Wan2.2 TI2V-5B, a video DiT pre-trained for image-to-video synthesis (Wan et al., 2025). An 81-frame video sequence $x \in \mathbb{R}^{81\times3\times H\times W}$ is compressed by a temporal VAE into 21 latent frames $z \in \mathbb{R}^{21\times C\times h\times w}$. We anchor the first latent $L_0 \in \mathbb{R}^{C\times h\times w}$ as the clean initial visual condition and partition the remaining 20 target latents into five uniform chunks:

$$z = [L_0;\, z_1;\, z_2;\, z_3;\, z_4;\, z_5], \quad \text{where } z_i = L_{4i-3:4i} \in \mathbb{R}^{4\times C\times h\times w}. \tag{1}$$

Chunk-aligned text conditioning. Both during training and inference, each temporal chunk z_i ($i \in \{1,\dots,5\}$) is conditioned on the chunk prompt y_i that exclusively describes character actions or environmental events within the specific window. Text embeddings are extracted from a frozen T5 encoder and modulate the video transformer through cross-attention, maintaining a strict chunk-wise mapping to prevent temporal text leakage.

Geometric camera conditioning. Camera trajectories provide frame-aligned geometric guidance through two complementary branches. In every transformer block, a parallel UCPE camera-attention branch (Zhang et al., 2026) applies ray-local transforms to its own queries, keys, and values, and adds its output to the main self-attention output. Frame-wise Plücker raymaps are packed within each VAE temporal stride, embedded at the latent resolution, and added through per-block projections after self-attention in the early blocks. Both branches operate alongside text conditioning in the shared backbone.

4.2 Training Strategy

Stage I: Bidirectional Data Adaptation We first fine-tune the pretrained video DiT under full bidirectional attention to establish stable control capabilities before applying causal masking. Training follows a progressive curriculum that starts with camera trajectory warm-up and smoothly transitions to full joint supervision.

Stage II: Chunk-autoregressive teacher. We then apply the block-causal temporal mask to construct a chunk-autoregressive teacher generator. Under teacher forcing, each target chunk z_i independently samples a noise level $t_i \sim U(0,1)$ and attends to the clean initial frame L_0, clean preceding visual chunks $z_{<i}$, and its own noisy latents $z_i^{t_i}$. We optimize the conditional flow-matching

loss across all target chunks:

$$\mathcal{L}_{\mathrm{AR}} = \mathbb{E}_{t_i,z}\left[\sum_{i=1}^{K} w(t_i)\left\|v_\theta(z_i^{t_i} \mid L_0, z_{<i}, u_{\le i}) - v_i^\star\right\|_2^2\right], \tag{2}$$

where $v_i^\star$ is the target velocity and $w(t_i)$ denotes the loss weighting. This stage converts the bidirectional interface into a causal streaming model while preserving parallel training efficiency.

Stage III: Few-step causal distillation. To enable real-time interactive rollout, we distill the multi-step teacher into a four-step causal student generator. We sample probability-flow ODE trajectories from the frozen teacher and train the student network s_ϕ to directly predict the trajectory endpoints $\hat{z}_i^0$ under identical causal histories and control inputs:

$$\mathcal{L}_{\mathrm{distill}} = \mathbb{E}\left[\sum_{i=1}^{K}\left\|s_\phi(z_i^t \mid L_0, z_{<i}, u_{\le i}) - \hat{z}_i^0\right\|_2^2\right]. \tag{3}$$

By matching ODE endpoints across discrete time steps, this formulation compresses sampling to four forward passes per chunk while preserving interactive controllability.

5 Experiments

5.1 Quantitative Results

Table 1: **WBench Navi split.** Models in the top section are strong baselines with parameter counts exceeding 5B. Models in the middle section are all built upon the same 5B backbone. Rows highlighted in light gray denote evaluations on the Navi split, while rows in light purple report our model's performance on the full split. Parameter counts are based on official releases (— denotes not reported). Scores are on a 0–100 scale, with higher values indicating better performance.

Model	Params	Avg.	Quality	Setting	Interaction	Consistency	Physical
HY-World 1.5	8.3B	78.1	78.1	72.2	86.8	86.9	66.3
LingBot-World	14B	78.5	78.9	72.6	80.1	**89.9**	71.2
Kling 3.0	—	79.0	81.4	**91.0**	69.4	83.7	69.3
LingBot-World v2	14B	79.4	81.8	76.8	82.8	86.5	69.1
Alaya-EVOKE	14B	80.8	**83.8**	83.8	78.6	86.9	**72.2**
JoyAI-Echo-1.5	22B	**81.0**	82.8	77.5	**87.9**	88.3	72.1
DreamX-World	5B	75.0	77.5	**80.8**	78.6	74.9	63.3
YUME 1.5	5B	73.3	77.6	72.4	71.4	80.1	**65.2**
Wan 2.2	5B	69.1	72.2	69.3	65.4	86.4	52.3
Cela-World-5B (ours)	5B	**76.9**	**78.9**	75.2	**81.6**	**88.4**	60.5
Wan 2.2	5B	61.1	72.1	67.3	28.3	85.8	51.8
Cela-World-5B (ours)	5B	69.8	77.6	74.7	49.2	87.1	60.1

Baselines. We evaluate our model against eight recent baselines and our base model on the Navi split of WBench (Ying et al., 2026), which assesses video quality, setting adherence, interaction adherence, temporal consistency, and physical fidelity. Furthermore, to evaluate complex actions and editing capabilities on the full split, where most baselines lack support for control, we compare our method against our base model.

Analysis. Table 1 summarizes the WBench results. We select eight recent models to cover both leading large-scale systems and representative methods built on 5B backbones, and include Wan 2.2 as our base model. On the Navi split, Cela-World-5B achieves an average score of 76.9, outperforming all other 5B models. Compared with Wan 2.2, our model improves the average score by 7.8, with gains across all five dimensions. Cela-World-5B also reaches 81.6 in interaction adherence, exceeding several substantially larger baselines, while its consistency score of 88.4 ranks second among all compared models. On the Full split, our model improves over Wan 2.2 from 61.1 to 69.8. Overall, these results show that our training pipeline substantially strengthens controllability.

5.2 Data Composition and Specialized Action Priors

Our dataset comprises three complementary streams tailored for distinct generation controls: camera navigation, multi-perspective actions, and event editing sequences. Navigation data establishes spatial geometry priors, enabling the model to synthesize temporally coherent scenes under large viewpoint shifts without physical or semantic degradation. Consequently, we initialize training with a camera navigation warm-up phase prior to introducing action and editing streams. Evaluated on a 500-case test benchmark, this warm-up protocol reduces the mean movement direction error from 19.09° to 12.13°.

Action sequences supply foundational motion priors that enable rapid adaptation to novel interaction combinations. As illustrated in Figure 5, training on FPS gameplay embeds robust weapon-switching priors into the backbone. For tool interactions, fine-tuning on as few as ten synthetic real-world instances allows the model to generalize weapon-switching dynamics to arbitrary tool transitions.

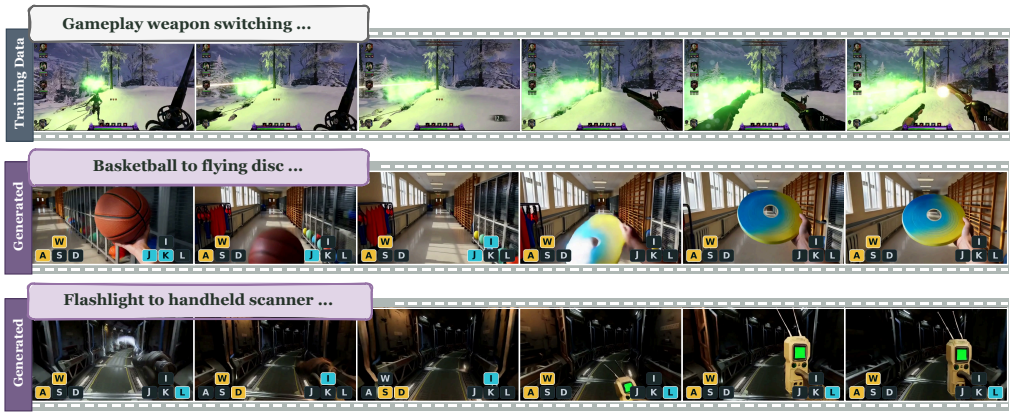

Figure 5: **Action prior generalization.** The top row shows gameplay training data capturing first-person weapon switching. Leveraging these learned dynamics, the model generalizes to diverse real-world object and tool transitions, such as swapping a basketball for a flying disc (middle) or replacing a flashlight with a handheld scanner (bottom).

Integrating action and editing data further unlocks the video generator's capabilities in reasoning and conditional generation. By conducting post-training with just 500 samples, our model successfully generalizes and composes diverse editing and generation skills across various scenes, viewpoints, and subjects, as shown in Figure 6.

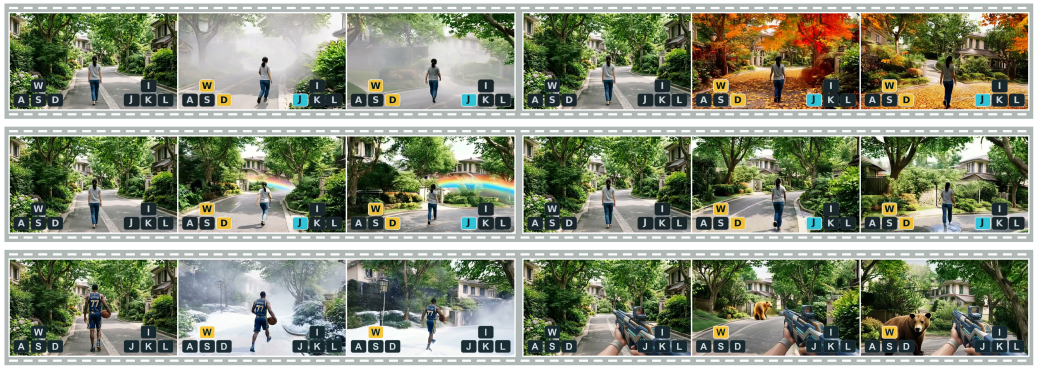

Figure 6: **Visualization of compositional generalization.**Our model seamlessly composes diverse editing and action generation capabilities across varied scenes, viewpoints, and subjects.

5.3 QUALITATIVE RESULTS.

As shown in Figure 7, our model generates spatiotemporally coherent video streams under diverse scene editing and composition controls. Across global environmental shifts such as sandstorms, material level dynamics including crystal transformations, and physical action edits such as object levitation, the model maintains frame-level temporal continuity without breaking background geometric consistency. Furthermore, complex composition setups demonstrate that multi-axis control inputs smoothly integrate localized action execution with background scene evolution.

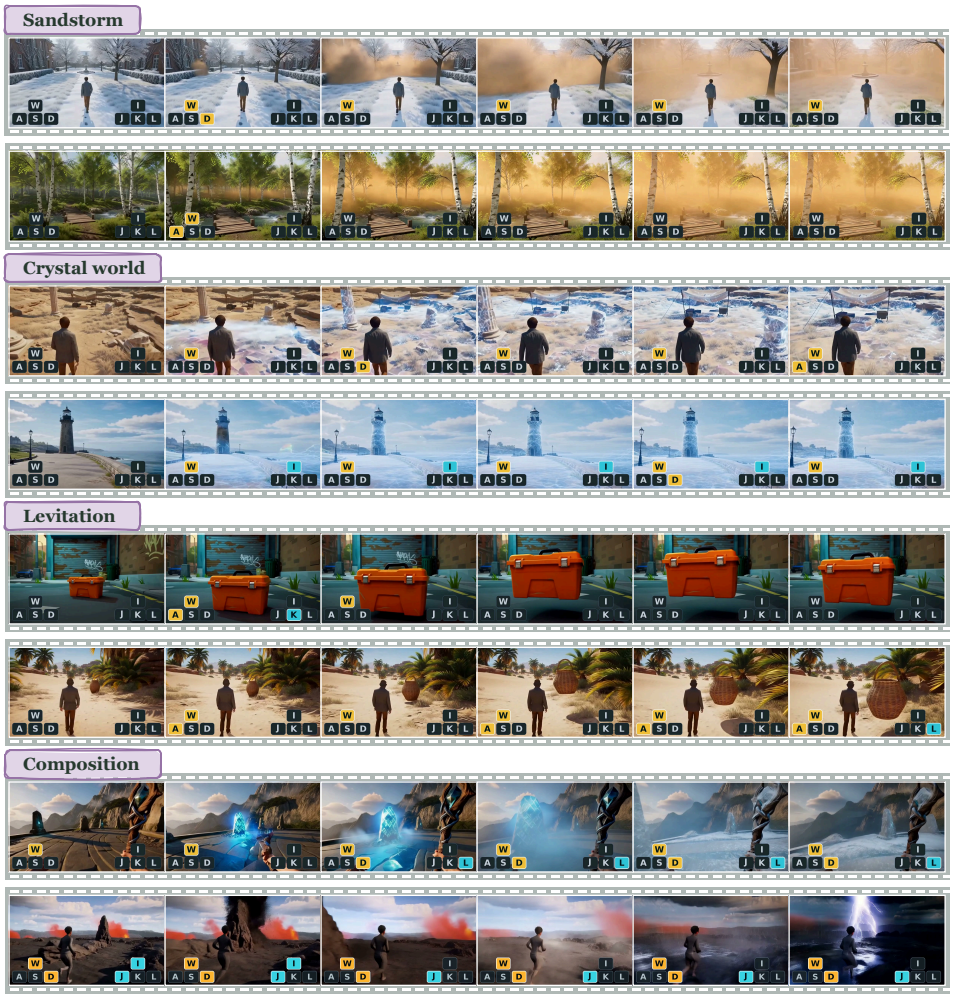

Figure 7: **Visualization of Cela-World-5B.** Our model supports diverse scenes along with both global scene editing and local object editing.

6 CONCLUSION

We introduced a unified data pipeline for learning interactive video world models from heterogeneous video sources. The pipeline converts diverse supervision into temporally aligned controls over camera motion, character actions, and environmental events, enabling a single autoregressive model to learn navigation, interaction, and editing within a common framework. Our experiments demonstrate strong instruction following, zero-shot compositional generalization, and effective cross-domain transfer, while our data ablations reveal how different sources contribute complementary capabilities. These results suggest that carefully structured multi-source supervision is a practical path toward scalable and semantically controllable video world models.

AI Use Statement. In preparing this manuscript, the authors used generative language models only as writing aids, to correct grammar, tighten phrasing, and improve readability. The models were not used to form research ideas, design or carry out experiments, develop proofs, collect or analyze data, or determine the scientific conclusions. All model-suggested edits were examined by the authors before inclusion, and the authors accept full accountability for the methods, results, and claims reported in this paper.

REFERENCES

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, Yusuf Aytar, Sarah Maria Elisabeth Bechtle, Feryal Behbahani, Stephanie C.Y. Chan, Nicolas Heess, Lucy Gonzalez, Simon Osindero, Sherjil Ozair, Scott Reed, Jingwei Zhang, Konrad Zolna, Jeff Clune, Nando De Freitas, Satinder Singh, and Tim Rocktäschel. Genie: Generative interactive environments. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 4603–4623. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/bruce24a.html.

Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. In *Advances in Neural Information Processing Systems*, volume 37, pp. 24081–24125, 2024. doi: 10.52202/079017-0759. URL https://doi.org/10.52202/079017-0759.

Mingyang Chen, Shengdong Chen, Xiaoxiao Fu, Bosheng Gong, Haoyuan Guo, Bowen Li, Jiawen Li, Kejun Li, Tianpeng Li, Yin Liu, Haoze Sun, Zeyang Tian, Meng Wang, Xinmiao Wu, Jiangqiao Yan, and Zining Zhao. Zing-0.5: Toward playable worlds with real-time joint action and text control. *arXiv preprint arXiv:2609.17909*, 2026.

Xinhang Gao, Junlin Guan, Shuhan Luo, Wenzhuo Li, Guanghuan Tan, and Jiacheng Wang. Memcam: Memory-augmented camera control for consistent video generation. *arXiv preprint arXiv:2603.26193*, 2026a.

Zelin Gao, Qiuyu Wang, Jiapeng Zhu, Jingye Chen, Zichen Liu, Qingyan Bai, Jiahao Wang, Yufeng Yuan, Hanlin Wang, Yichong Lu, Ka Leong Cheng, Haojie Zhang, Jian Gao, Tianrui Feng, Yuzheng Liu, Yao Yao, Yinghao Xu, Xing Zhu, Yujun Shen, and Hao Ouyang. Infinite worlds with versatile interactions. *arXiv preprint arXiv:2607.07534*, 2026b.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu, Abrham Gebreselasie, Cristina González, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jáchym Kolář, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbeláez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18995–19012, June 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, Eugene Byrne, Zach Chavis, Joya Chen, Feng Cheng, Fu-Jen Chu, Sean Crane, Avijit Dasgupta, Jing Dong, Maria Escobar, Cristhian Forigua, Abrham Gebreselasie, Sanjay Haresh, Jing Huang, Md Mohaiminul

Islam, Suyog Jain, Rawal Khirodkar, Devansh Kukreja, Kevin J Liang, Jia-Wei Liu, Sagnik Majumder, Yongsen Mao, Miguel Martin, Effrosyni Mavroudi, Tushar Nagarajan, Francesco Ragusa, Santhosh Kumar Ramakrishnan, Luigi Seminara, Arjun Somayazulu, Yale Song, Shan Su, Zihui Xue, Edward Zhang, Jinxu Zhang, Angela Castillo, Changan Chen, Xinzhu Fu, Ryosuke Furuta, Cristina Gonzalez, Prince Gupta, Jiabo Hu, Yifei Huang, Yiming Huang, Weslie Khoo, Anush Kumar, Robert Kuo, Sach Lakhavani, Miao Liu, Mi Luo, Zhengyi Luo, Brighid Meredith, Austin Miller, Oluwatumininu Oguntola, Xiaqing Pan, Penny Peng, Shraman Pramanick, Merey Ramazanova, Fiona Ryan, Wei Shan, Kiran Somasundaram, Chenan Song, Audrey Southerland, Masatoshi Tateno, Huiyu Wang, Yuchen Wang, Takuma Yagi, Mingfei Yan, Xitong Yang, Zecheng Yu, Shengxin Cindy Zha, Chen Zhao, Ziwei Zhao, Zhifan Zhu, Jeff Zhuo, Pablo Arbelaez, Gedas Bertasius, Dima Damen, Jakob Engel, Giovanni Maria Farinella, Antonino Furnari, Bernard Ghanem, Judy Hoffman, C.V. Jawahar, Richard Newcombe, Hyun Soo Park, James M. Rehg, Yoichi Sato, Manolis Savva, Jianbo Shi, Mike Zheng Shou, and Michael Wray. Ego-exo4d: Understanding skilled human activity from first- and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19383–19400, June 2024.

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. *arXiv preprint arXiv:2504.08388*, 2025.

Rongjin Guo, Dong Liang, Yuhao Liu, Fang Liu, Tianyu Huang, Gerhard P Hancke, and Rynson WH Lau. Egocs-400k: An egocentric gameplay dataset for world models. *arXiv preprint arXiv:2606.18180*, 2026.

Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2025.

Xianglong He, Chunli Peng, Zexiang Liu, Boyang Wang, Yifan Zhang, Qi Cui, Fei Kang, Biao Jiang, Mengyin An, Yangyang Ren, Baixin Xu, Hao-Xiang Guo, Kaixiong Gong, Size Wu, Wei Li, Xuchen Song, Yang Liu, Yangguang Li, and Yahui Zhou. Matrix-game 2.0: An open-source real-time and streaming interactive world model. *arXiv preprint arXiv:2508.13009*, 2026.

Bin Hu, Yanwen Ma, Jiehui Huang, Ziliang Zhang, Haoning Wu, Ruicheng Zhang, Yaokun Li, Zijun Wang, Yuechen Zhang, Chun-Mei Tseng, Hanhui Li, Shengju Qian, Jun Zhou, Kaipeng Zhang, Xiaodan Liang, Jiaya Jia, and Xiu Li. Physeditworld: A large-scale dataset toward physics-editable world models. *arXiv preprint arXiv:2606.26694*, 2026.

Hongrui Huang, Junke Wang, Quanhao Li, Yu-Gang Jiang, and Zuxuan Wu. Disco: World models with discrete camera motion control. *arXiv preprint arXiv:2606.07967*, 2026a.

Jiahui Huang, Qunjie Zhou, Hesam Rabeti, Aleksandr Korovko, Huan Ling, Xuanchi Ren, Tianchang Shen, Jun Gao, Dmitry Slepichev, Chen-Hsuan Lin, et al. Vipe: Video pose engine for 3d geometric perception. *arXiv preprint arXiv:2508.10934*, 2025a.

Junchao Huang, Guian Fang, Shengju Qian, Xianghao Kong, Zhuoran Zhao, Wei Huang, Yihua Du, Zixin Zhang, Justin Cui, Yuchao Gu, Yukang Chen, Xinting Hu, Tianyu He, Shaoshuai Shi, Zhuotao Tian, Xin Wang, Mike Zheng Shou, and Li Jiang. Solarwm: Open data and scalable training for long-horizon video world models. *arXiv preprint arXiv:2609.02886*, 2026b.

Siqiao Huang, Jialong Wu, Qixing Zhou, Shangchen Miao, and Mingsheng Long. Vid2world: Crafting video diffusion models to interactive world models. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 10763–10806, 2026c. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/1214d67bb7fa38d40e0b3c2be677e39b-Paper-Conference.pdf.

Xun Huang, Zhengqi Li, Guande He, Mingyuan Zhou, and Eli Shechtman. Self forcing: Bridging the train-test gap in autoregressive video diffusion. In *Advances in Neural Information Processing Systems*, volume 38, pp. 185411–185436, 2025b. doi: 10.52202/085713-5576. URL https://doi.org/10.52202/085713-5576.

Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. Deepspeed ulysses: System optimizations for enabling training of extreme long sequence transformer models. *arXiv preprint arXiv:2309.14509*, 2023.

Fan Jiang, Zhaoxu Sun, Mengchao Wang, Ziyu Zhu, Chiyu Wang, Yunpeng Zhang, Wenlin Liu, Yun Wang, Xue Zheng, Rui Sun, Junfeng Ni, Hongyu Pan, Zhongxu Sun, Fei Yu, Zengye Ge, Mengmeng Du, Nianfei Fan, Mingchao Sun, Yu Liu, Yongchang, Yanqing Zhu, Jiahang Wang, Ning Ying, Yuze Xuan, Di Yang, Zhicheng Liu, Zhe Gao, Tingbing Xu, Jiacheng Sui, Wenjin Yang, Junnan Lai, Shufeng Liu, Yuan Liu, Zheng Zhou, Yingliang Peng, Dawei Cao, Kaifeng Sheng, Yuxiang Cai, Fei Lu, Mu Xu, and Ning Guo. ABot-World-0: Infinite interactive world rollout on a single desktop gpu, 2026.

Xinye Li, Lingshuai Lin, Lei Wang, Liuzhou Zhang, Jialin Cui, Qingshan Li, Guanchu Wang, Qingbin Liu, Xi Chen, Jiang Bian, and Wai Lam. Forgewm: Progressive causal training for few-step action-conditioned video world models. *arXiv preprint arXiv:2608.14022*, 2026a.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Jiangmiao Pang, Yu Qiao, Yunde Jia, and Kaipeng Zhang. Sekai: A video dataset towards world exploration. *arXiv preprint arXiv:2506.15675*, 2025.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, et al. Sekai: A video dataset towards world exploration. *Advances in Neural Information Processing Systems*, 38, 2026b.

Zhen Li, Zian Meng, Shuwei Shi, Wenshuo Peng, Yuwei Wu, Bo Zheng, Chuanhao Li, and Kaipeng Zhang. Wildworld: A large-scale dataset for dynamic world modeling with actions and explicit state toward generative arpg. *arXiv preprint arXiv:2603.23497*, 2026c.

Markov AI. Gaming-500-hours. https://huggingface.co/datasets/markov-ai/gaming-500-hours, 2026.

MiniMax. MiniMax H3. https://github.com/MiniMax-AI/MiniMax-H3, 2026.

Runjia Qian, Zile Wang, Jihai Zhang, Kai Zou, Wei Yu, Jiaxing Li, Zexiang Liu, Yaokun Li, Fei Kang, Kaichen Huang, Mengyin An, Haobo Zhang, Biao Jiang, Jiahua Wang, Haofeng Sun, Yang Liu, and Yangguang Li. Matrix-game 3.5: Enhancing real-time streaming interactive world models with patch memory. *arXiv preprint arXiv:2608.29910*, 2026.

Wenxuan Shen, Dongna Jin, and Dongping Chen. Game2world engine: Unlocking in-the-wild gameplay videos for world model training. *arXiv preprint arXiv:2608.24680*, 2026.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2026.

Pardis Taghavi, Tingyu Guo, Jonas Lossner, Gaurav Pandey, and Reza Langari. Actionsplice: In-flight action editing for interactive world models. *arXiv preprint arXiv:2609.08230*, 2026.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Ruicong Liu, et al. Alayaworld: Long-horizon and playable video world generation. *arXiv preprint arXiv:2607.06291*, 2026a.

DreamX Team, Yancheng Bai, Rui Chen, Xiangxiang Chu, Rujing Dang, Hao Dou, Bingjie Gao, Qiwen Gu, Siyu Hong, Jiachen Lei, et al. Dreamx-world 1.0: A general-purpose interactive world model. *arXiv preprint arXiv:2606.16993*, 2026b.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/b71ecea210f7159f31e46631fe5c838f-Paper-Conference.pdf.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Haoyu Wang, Songchun Zhang, Haoran Li, Haoyang Huang, Zeyue Xue, and Nan Duan. Building pretraining data for world models: An unreal engine-based pipeline for action-conditioned video generation. *arXiv preprint arXiv:2609.03557*, 2026a.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. Spatialvid: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025.

Yunnan Wang, Kecheng Zheng, Jianyuan Wang, Minghao Chen, David Novotny, Christian Rupprecht, Yinghao Xu, Xing Zhu, Wenjun Zeng, Xin Jin, and Yujun Shen. Scenescribe-1m: A large-scale video dataset with comprehensive geometric and semantic annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12628–12639, June 2026b.

Zile Wang, Zexiang Liu, Jiaxing Li, Kaichen Huang, Baixin Xu, Fei Kang, Mengyin An, Peiyu Wang, Biao Jiang, Yichen Wei, et al. Matrix-game 3.0: Real-time and streaming interactive world model with long-horizon memory. *arXiv preprint arXiv:2604.08995*, 2026c.

Yuhui Wu, Liyi Chen, Ruibin Li, Shihao Wang, Chenxi Xie, and Lei Zhang. Insvie-1m: Effective instruction-based video editing with elaborate dataset construction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 16692–16701, October 2025.

Xiaokunfeng2022. EmbodiedWorld-200K: Toward embodied world agents via embodied-planning dataset and interactive world models. https://huggingface.co/datasets/Xiaokunfeng2022/EmbodiedWorld-200K, 2026.

Zhexiao Xiong, Yizhi Song, Hao Kang, Qing Yan, Liming Jiang, Jenson Yang, Zhoujie Fu, Stathi Fotiadis, Angtian Wang, Zichuan Liu, et al. Actworld: From explorable to interactive world model via action-aware memory. *arXiv preprint arXiv:2606.17730*, 2026.

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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22963–22974, June 2025.

Kaining Ying, Hengrui Hu, Siyu Ren, Jiamu Li, Fengjiao Chen, Ziwen Wang, Xuezhi Cao, Xunliang Cai, and Henghui Ding. Wbench: A comprehensive multi-turn benchmark for interactive video world model evaluation. *arXiv preprint arXiv:2605.25874*, 2026.

Jiwen Yu, Yiran Qin, Xintao Wang, Pengfei Wan, Di Zhang, and Xihui Liu. Gamefactory: Creating new games with generative interactive videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 11590–11599, October 2025.

Yuguang Yue, Irakli Salia, Samuel Hunt, Chris Green, Wenzhe Shi, and Jonathan J. Hunt. Scaling behavior cloning improves causal reasoning: An open model for real-time video game playing, 2026.

Cheng Zhang, Boying Li, Meng Wei, Yan-Pei Cao, Camilo Gambardella, Dinh Phung, and Jianfei Cai. Unified camera positional encoding for controlled video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 38027–38037, 2026.

Peiyuan Zhang, Yongqi Chen, Haofeng Huang, Will Lin, Zhengzhong Liu, Ion Stoica, Eric Xing, and Hao Zhang. Vsa: Faster video diffusion with trainable sparse attention. *arXiv preprint arXiv:2505.13389*, 2025a.

Peiyuan Zhang, Yongqi Chen, Runlong Su, Hangliang Ding, Ion Stoica, Zhengzhong Liu, and Hao Zhang. Fast video generation with sliding tile attention. *arXiv preprint arXiv:2502.04507*, 2025b.

Min Zhao, Hongzhou Zhu, Bokai Yan, Zihan Zhou, Yimin Chen, Wenqiang Sun, Kaiwen Zheng, Guande He, Xiao Yang, Chongxuan Li, et al. minwm: A full-stack open-source framework for real-time interactive video world models. *arXiv preprint arXiv:2605.30263*, 2026a.

Min Zhao, Hongzhou Zhu, Kaiwen Zheng, Zihan Zhou, Bokai Yan, Xinyuan Li, Xiao Yang, Chongxuan Li, and Jun Zhu. Causal forcing++: Scalable few-step autoregressive diffusion distillation for real-time interactive video generation. *arXiv preprint arXiv:2605.15141*, 2026b.

Yizhou Zhao, Yifan Wang, Xiaoyuan Wang, Yushu Wu, Hao Zhang, Moayed Haji-Ali, Rameen Abdal, Ashkan Mirzaei, Yanyu Li, Willi Menapace, Laszlo Jeni, Sergey Tulyakov, Peter Wonka, and Chaoyang Wang. Geostream: Toward precise camera controlled streaming video generation. *arXiv preprint arXiv:2606.15162*, 2026c.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, Mingyu Liu, Dingning Liu, Jiange Yang, Zhoujie Fu, Junyi Chen, Chunhua Shen, Jiangmiao Pang, Kaipeng Zhang, and Tong He. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025.

Haoyi Zhu, Haozhe Liu, Yuyang Zhao, Tian Ye, Junsong Chen, Jincheng Yu, Tong He, Song Han, and Enze Xie. Sana-wm: Efficient minute-scale world modeling with hybrid linear diffusion transformer. *arXiv preprint arXiv:2605.15178*, 2026a.

Hongzhou Zhu, Min Zhao, Guande He, Hang Su, Chongxuan Li, and Jun Zhu. Causal forcing: Autoregressive diffusion distillation done right for high-quality real-time interactive video generation. *arXiv preprint arXiv:2602.02214*, 2026b.

Shangwen Zhu, Qianyu Peng, Zhao Pu, Zhilei Shu, Xiangrui Ke, Zhaohu Xing, Zizhao Tong, Zeqing Wang, Xinyu Cui, Zian Zheng, et al. Incantation: Natural language as the action interface for multi-entity video world models. *arXiv preprint arXiv:2605.18601*, 2026c.

A Appendix

A.1 Implementation Details

The training process is built upon FastVideo (Zhang et al., 2025a;b) and consists of three main stages: Bidirectional Pre-training/SFT (BI Stage), Multi-step Causal Autoregressive SFT (AR Teacher Stage) and Four-step Causal ODE Endpoint Distillation (ODE Student Stage). Across all training stages, a consistent global batch size of 96 is maintained. In the BI Stage, training begins with an initial warmup phase of 2,000 steps using approximately 70,000 pure camera data samples, followed by 10,000 steps of full-data training incorporating all dataset samples, and concludes with 2,000 steps of fine-tuning on a dataset blending curated synthetic data with previous data at a 1:2 ratio, all optimized using AdamW with a learning rate of 1×10^{-5} and a weight decay of 0.001. In the AR Teacher Stage, the model transitions to a causal teacher setup and is trained for 8,000 steps using AdamW with a lower learning rate of 1×10^{-6} and a weight decay of 0.001. In the ODE Student Stage, the student model is trained for 8,000 steps using the Adam optimizer with a learning rate of 2×10^{-6} and no weight decay (0) to directly predict trajectory endpoints sampled from a 4-step discrete grid $\{0, 12, 24, 36\}$.

A.2 Inference Protocol and Streaming Decoding

During inference, the model receives the initial latent L_0 and a sequence of chunk-aligned control inputs. To maximize throughput, we apply Ulysses sequence parallelism (Gao et al., 2026b; Jacobs et al., 2023), using All-to-All communication to transpose visual tokens between sequence-sharded and head-sharded layouts for parallel multi-head attention. Furthermore, DiT denoising is decoupled from visual decoding through an asynchronous pipeline queue. A dedicated causal VAE worker decodes completed latents over separate CUDA streams, overlapping DiT forward passes of chunk $i + 1$ with video playback of chunk i.

A.3 More Results

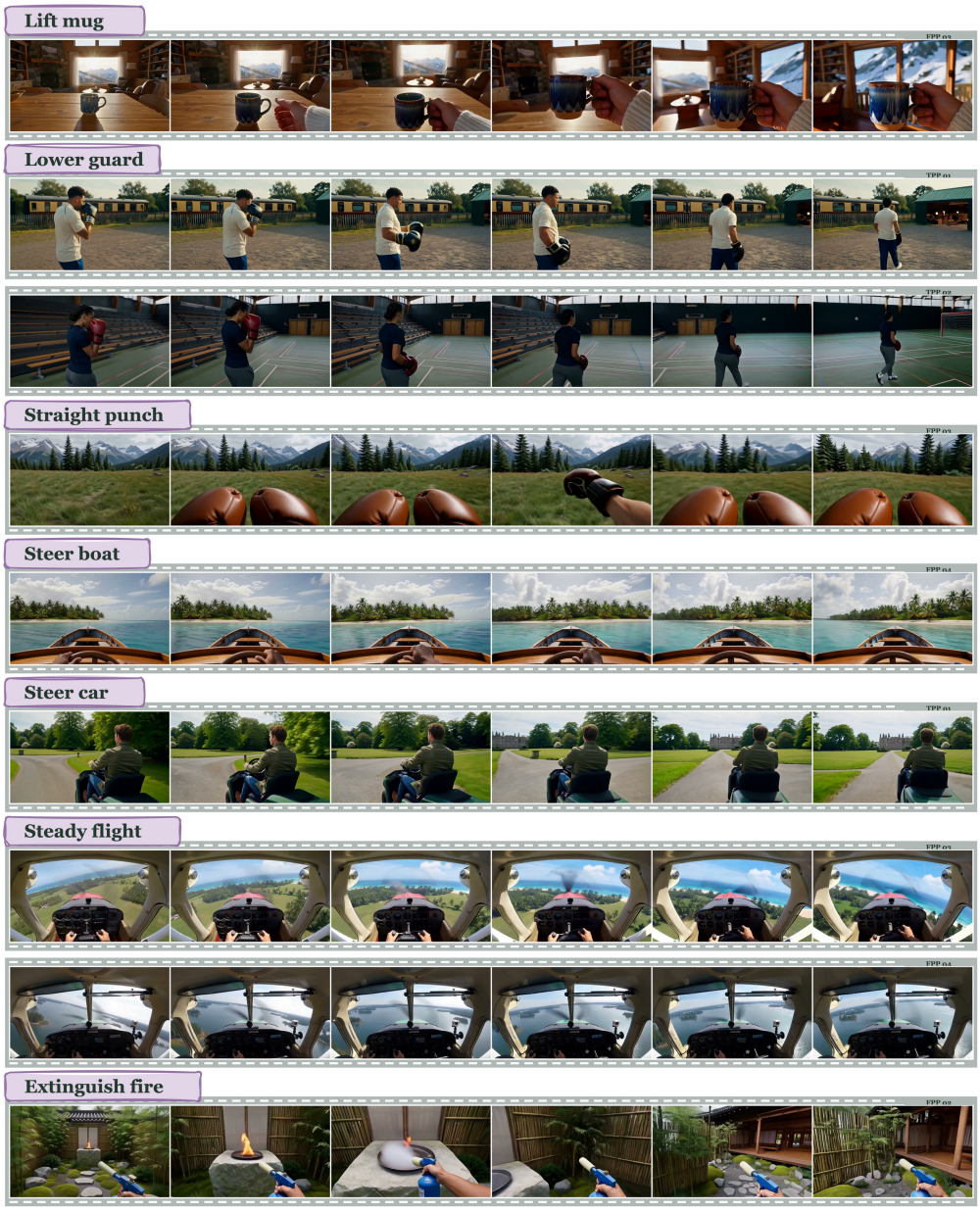

Figure 8: **More results.**

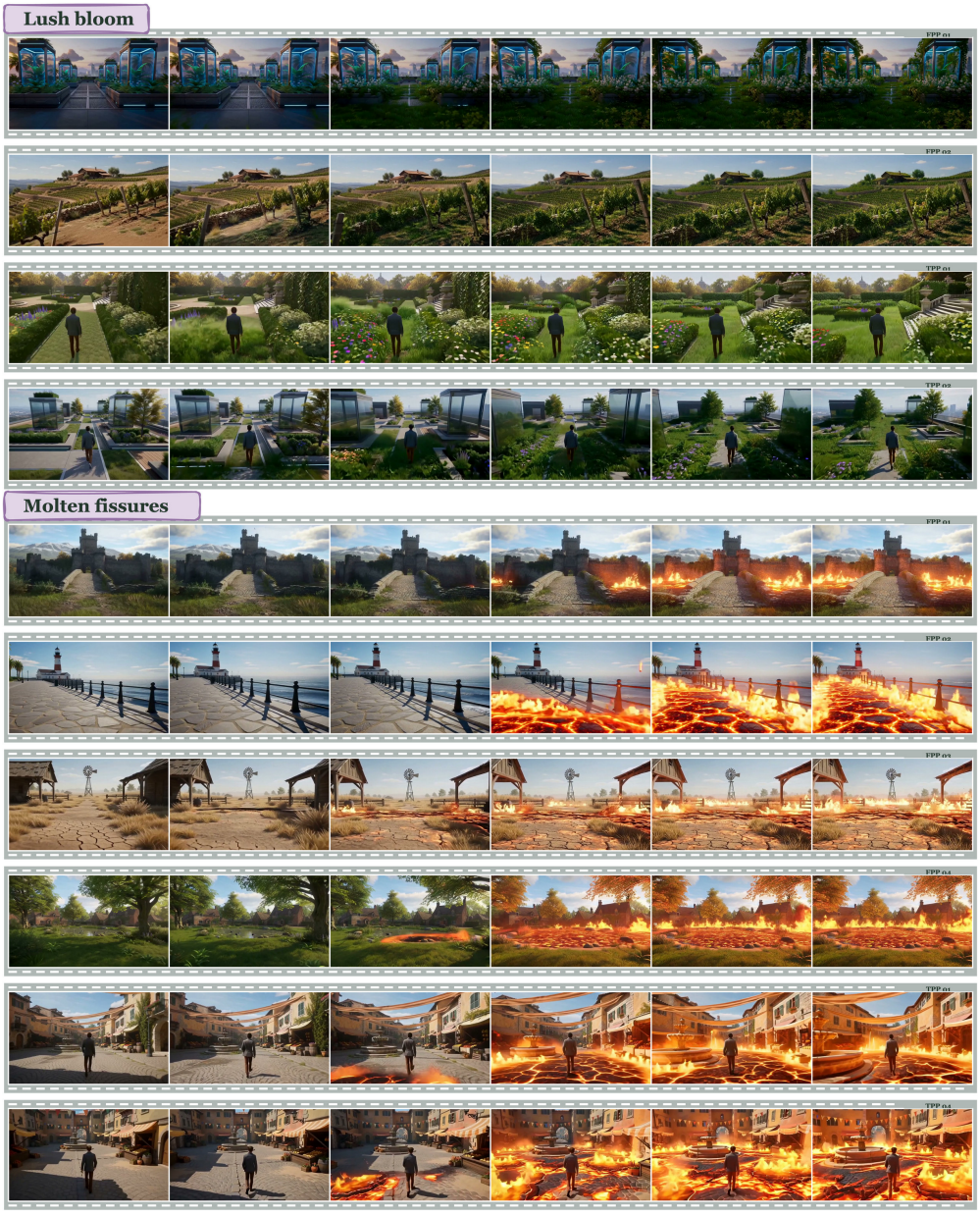

Figure 9: **More results.**

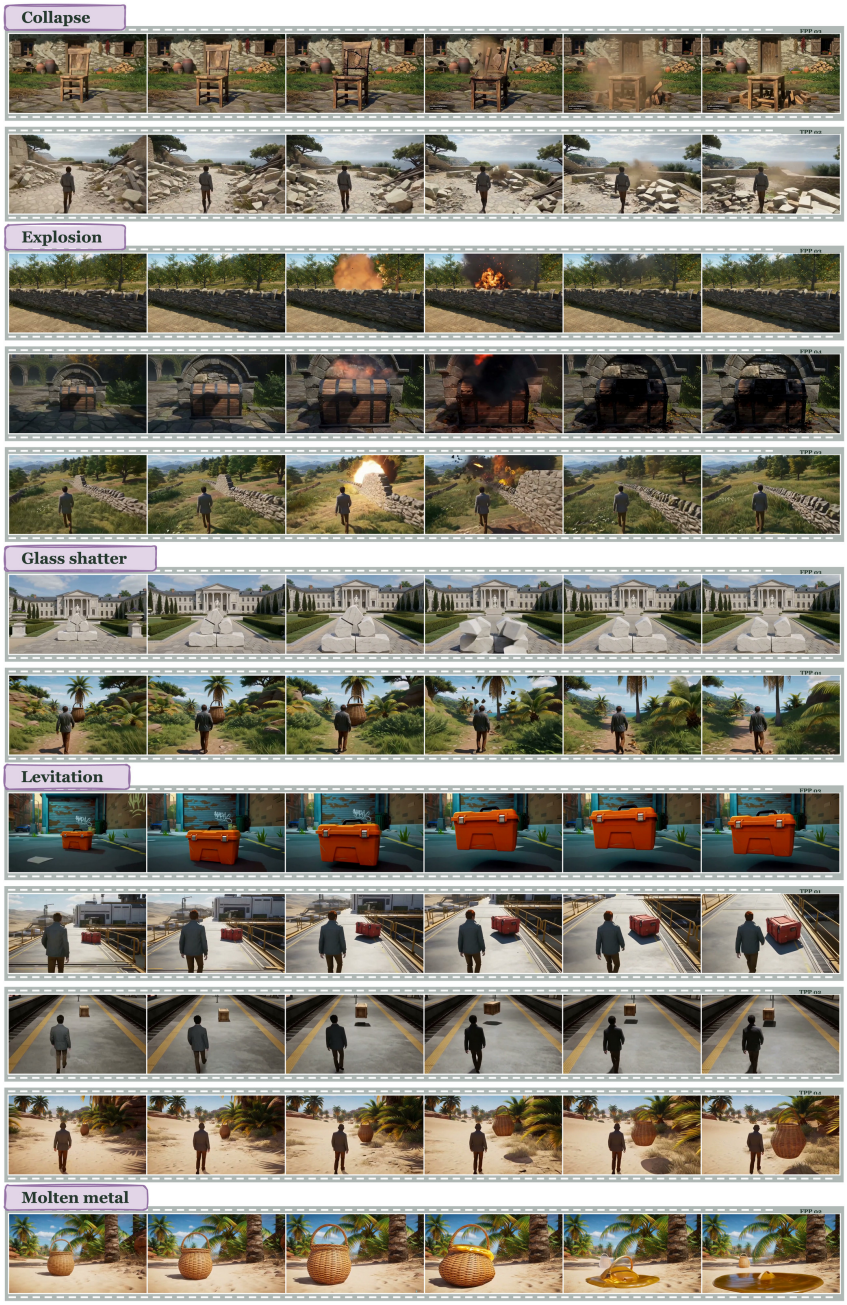

Figure 10: **More results.**