🌍 *Pandora*: Towards General World Model
with Natural Language Actions and Video States

Jiannan Xiang[*], **Guangyi Liu**[*], **Yi Gu**[*], **Qiyue Gao**, **Yuting Ning**, **Yuheng Zha**,
Zeyu Feng, **Tianhua Tao**, **Shibo Hao**, **Yemin Shi**, **Zhengzhong Liu**,
Eric P. Xing, **Zhiting Hu**
Maitrix.org, UC San Diego, MBZUAI
https://world-model.ai   *equal contribution

Figure 1: *Pandora* simulates future world states (videos) under action control (natural language).
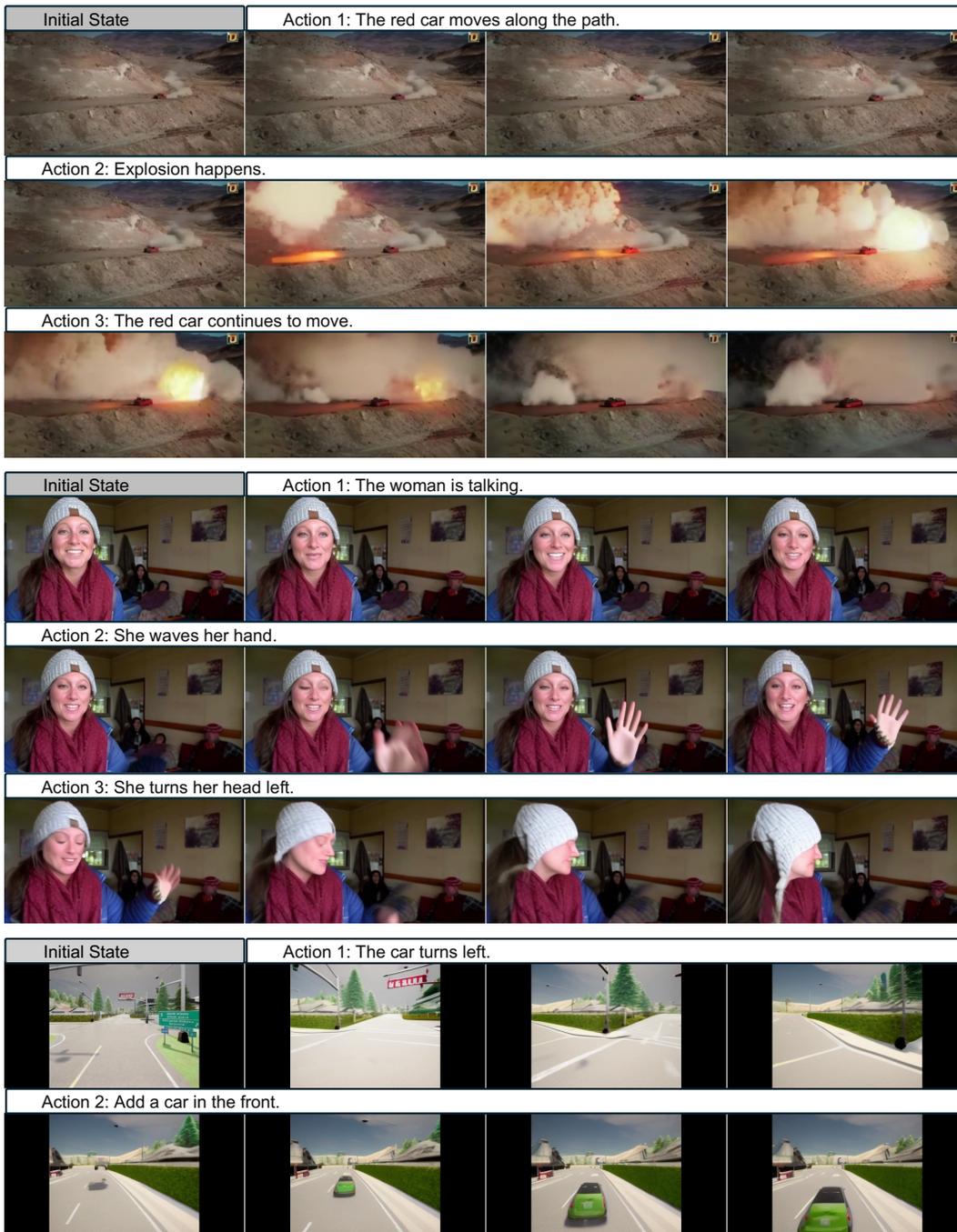
Abstract

World models simulate future states of the world in response to different actions. They facilitate interactive content creation and provides a foundation for grounded, long-horizon reasoning. Current foundation models do not fully meet the capabilities of general world models: large language models (LLMs) are constrained by their reliance on language modality and their limited understanding of the physical world, while video models lack interactive action control over the world simulations. This paper makes a step towards building a general world model by introducing *Pandora*, a hybrid autoregressive-diffusion model that simulates world states by generating videos and allows real-time control with free-text actions. *Pandora* achieves domain *generality*, video *consistency*, and *controllability* through large-scale pretraining and instruction tuning. Crucially, *Pandora* bypasses the cost of training-from-scratch by integrating a pretrained LLM (7B) and a pretrained video model, requiring only additional lightweight finetuning. We illustrate extensive outputs by *Pandora* across diverse domains (indoor/outdoor, natural/urban, human/robot, 2D/3D, etc.). The results indicate great potential of building stronger genearl world models with larger-scale training.

1 Introduction

A world model (WM) is an abstract representation that an intelligent system uses to understand and simulate the real world. The model encompasses various aspects of the environment, including physical laws, spatiotemporal knowledge, objects, scenes, agents, and their dynamic interactions. In particular, it allows to predict the future states of the world in response to different actions. Building a general world model, therefore, can serve for *interactive content creation*, such as generating realistic virtual scenes for video games and movies, developing immersive experiences in virtual and augmented reality, and creating dynamic simulations for training and educational purposes. Perhaps of even more significance is that a general WM provides a foundation for *robust, grounded reasoning* in AI systems, enabling them to anticipate complex environments and plan actions, such as robots navigating disaster scenes safely. WMs also hold the potential to power *long-horizon reasoning* that improves decision making in fields like logistics and healthcare, by simulating various scenarios and outcomes and identifying the most effective solutions.

Current *large language models (LLMs)* [1, 4, 37, 52, 68, 70, 71] are adept at generating human language and are used as surrogates for world models in certain reasoning tasks [30, 78]. However, language alone is a fundamentally insufficient and inefficient modality for describing various aspects of the world, such as intuitive physics (e.g., predicting fluid flow based on its viscosity) [36]. Moreover, LLMs lack a robust understanding of physical and temporal dynamics in the real world, relying on patterns in textual data without comprehending the underlying realities they describe [79, 43, 51]. On the other hand, contemporary *video generation models* can produce high-quality video content from given initial frames or text prompts [7, 9, 11, 25, 80, 83]. While these models can animate consistent sequences to visualize diverse scenes, they miss the complex interactive nature of the real world, lacking the ability for causal control and intervention with arbitrary actions during simulations. Recent work has also developed interactive world models at scale, such as GAIA-1 [35] for auto-driving, UniSim [81] for robotic manipulation, and Genie [12] for 2D games. These models are typically specific to certain domains, permitting limited sets of actions and/or states.

This work presents *Pandora*, a step towards a general world model that simulates world states across various domains by generating videos and allows real-time control through arbitrary actions expressed in natural language. *Pandora* is an autoregressive model that sequentially processes actions (free text) and previous states (videos) as inputs and generates next states (videos) as outputs (Figure 2). *Pandora* introduces a staged training strategy akin to the successful recipe of training LLMs [52, 46, 77], including: *(1) large-scale pretraining* with massive video and text data, respectively, to learn *domain-general* understanding of the world and production of *consistent* video simulations; and *(2) instruction tuning* with high-quality text-video sequential data to learn any-time text *controllability* during video generation.

Crucially, the pretraining stage allows for the separate training of text and video modeling. We thus can simply reuse existing pretrained LLMs and (text-to-)video generation models that have
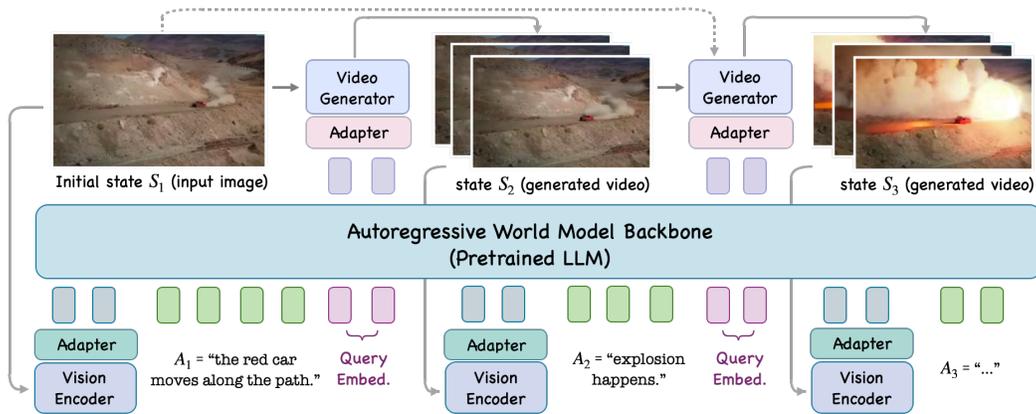
Figure 2: Model architecture of *Pandora*.

already achieved domain generality and video consistency in their own pretraining. We only need to stitch and align the language and video models together with necessary additional modules and lightweight tuning as described in §2. More specifically, in this work, we use the Vicuna-7B-v1.5 language model [17] and the DynamiCrafter text-to-video model [80] as the backbone. Using larger, more sophisticated pretrained models (such as GPT-4 and Sora) is expected to yield stronger performance. For the instruction tuning stage, we craft a large diverse set of action-state sequential data, by re-captioning general-domain videos and synthesizing with various simulators for robots, in-/out-door activities, driving, 2D games, and more. Similar to instruction tuning of LLMs that boosts their instructability in general unseen domains, tuning on the curated data boosts the world model's real-time controllability that generalizes to broad unseen states and actions.

We illustrate extensive outputs generated by *Pandora* across various domains in §3. The model demonstrates a range of desirable properties not exhibited by previous models. The results also indicate great potential for further enhancement with larger-scale training in the future.

- **The model simulates video states across broad domains:** *Pandora* is capable of generating videos across a wide range of general domains, such as indoor/outdoor, natural/urban, human/robot, 2D/3D, and other scenarios. This domain generality is primarily due to the large-scale video pretraining (inherited from the pretrained video model).

- **The model permits on-the-fly control with free-text actions:** *Pandora* accepts natural language actions as inputs during video generation to direct future world states. This differs crucially from previous text-to-video models which allow text prompts only at the beginning of the video. The on-the-fly control fulfills the promise of the world model to support interactive content generation and enhance robust reasoning and planning. The capability is enabled by the autoregressive architecture of the model (which permits text inputs at any time), the pretrained LLM backbone (which understands any text expressions), and the instruction tuning stage (which substantially enhances the effectiveness of control).

- **Action controllability transfers across domains:** As above, instruction tuning with high-quality data allows the model to learn effective action control and transfer to different unseen domains. We demonstrate that actions learned from a specific domain apply seamlessly to states in diverse new domains.

- **Autoregressive model backbone enables longer videos:** Existing video generation models based on diffusion architectures typically produce videos of a fixed length (e.g., 2 seconds). By integrating the pretrained video model with the LLM autoregressive backbone, *Pandora* is capable of extending the video duration indefinitely in an autoregressive manner. Together with the additional training (e.g., instruction tuning), we show *Pandora* can generate longer videos (e.g., 8 seconds) of higher quality.

2 Methods

2.1 Model Architecture

Pandora is an autoregressive world model. Given the previous states of the world, e.g., images or video clips, and a natural language action, it predicts the next state of the world, which is also a video

clip. Specifically, it formulates a state transition distribution:

$$P(S_t \mid S_{t-1}, A_{t-1}, ..., S_1, A_1), \qquad (1)$$

where S_i and A_i are the state and action at time step i, respectively. Each state $S_i = (s_1, ..., s_N)$ is a single or a sequence of video frames, and each action $A_i = (x_1, ..., x_M)$ is a sequence of text tokens. At the first time step, the state S_1 is one single image, and the states $S_2, ..., S_N$ at the following steps are video clips.

Figure 2 gives an overview of the model architecture. The two core components of of *Pandora* include the autoregressive backbone, which stems from a pretrained LLM, and the video generator, which is initialized with a pretrained video model. To stitch the two components together, other necessary components are added, including a vision encoder, and two adapters connecting the vision encoder to the LLM backbone, and the LLM backbone to the video generator, respectively.

At each time step t, the autoregressive backbone accepts three sets of embedding vectors as inputs: (1) the first is the sequence of visual embeddings, by the vision encoder followed by the adapter (a Q-Former [44]), that encodes the previous world state S_{t-1}; (2) the second is the token embeddings of the text words in action A_{t-1}; and (3) the third is a sequence of learnable embedding vectors (a.k.a. query embeddings). The length and positions of the query embeddings correspond exactly to those of the output embeddings by the autoregressive backbone to be fed to the video generator. Intuitively, the query embeddings stimulate the model to start generating videos [21]. The autoregressive backbone then generates a sequence of output embeddings. The adapter, which is a Q-Former, accepts the output embeddings and produces a new sequence of embeddings. Finally, the video generator takes the embeddings and generates the video clip outputs S_t. To improve the consistency of the new video clip with the preceding video clip S_{t-1}, the video generator additionally takes the last four frames of S_{t-1} as input (or the single image of S_{t-1} as input if S_{t-1} is the initial state S_1). In addition, the video generator will take an FPS number to control the motion level of the video. The number of frames generated in each video clip S_t depends on the specific pretrained video model used for initializing the video generator. As described below, we used the DynamiCrafter [80] which generates 16 frames.

2.2 Staged Training

A general world model needs to achieve *consistency*, *controllability*, and *generality*—it needs to generate consistent videos to describe the world state accurately, allow on-the-fly control by accepting natural language actions at any time during video generation, and perform the above well across all diverse domains (with different scenes and actions).

To this end, direct training of the world model requires massive high-quality (video S_1, text A_1, video $S_2, \dots$) sequences as training data, which is hard to obtain in practice. We instead devise a two-stage training strategy consisting of *pretraining* and *instruction tuning*.

The pretraining stage aims to acquire a few key capabilities, including (1) consistent general video generation of the video generator, (2) general text understanding of the autoregressive backbone to process actions, and (3) alignment of the representation spaces between the two components. The first two capabilities can be learned separately by training the video generator and the autoregressive backbone individually, or even by just plugging in existing pretrained video models and LLMs that already possess these capabilities during their own pretraining. The reuse of separately pretrained video and language models significantly reduces the training costs of the world model.

In the instruction tuning stage, we train the model on a curated video dataset with high-quality instructions (actions) that focus on the dynamics of the videos. This training is aimed at enhancing the model's ability to follow natural language instructions and accurately predict subsequent video states based on these directions.

We describe more details of the two training stages in the next sections, respectively.

2.2.1 Pretraining for Domain Generality and Generation Consistency

The pretraining stage aims to achieve the core capabilities of *consistency* and *generality* as described above. This is similar to the process of building an LLM where large-scale pretraining enables the LLM to generate consistent/fluent text in general domains.

General understanding of natural language can be achieved by massive training on text data, and generation of consistent general videos can be achieved by massive training on video data. Both of these have been done on existing pretrained LLMs and video generation models. We thus can directly reuse these models.

Specifically, we use Chat-Univi [38], which is a Vicuna-7B-v1.5 LLM equipped with a vision encoder, as our base LLM (autoregressive backbone), and DynamiCrafter [80] as our base video generation model. DynamiCrafter is a diffusion model pretrained to generate a video given an image and a text prompt.

In the pretraining stage, we additionally want to align the pretrained LLM backbone and video model, so that the output embeddings from the LLM can be passed along to the video model as input for video generation. We used a video caption dataset WebVid-10M [5] for the alignment training. For each (video, caption), we feed the first frame of the video and the caption into the LLM+vision encoder and get the output embeddings from the LLM. Meanwhile, we feed the caption into the text encoder of the video generation model and get the caption embeddings. We aim to match the two embeddings, so that the output embeddings from LLM can be understood by the vidoe generation model (just as how it understands the embeddings from its text encoder). Specifically, we minimize the L2 loss between the two sets of embeddings, and trains the parameters of the adapter between the LLM and video generator, as well as the query embeddings. Both the pretrained LLM and video generation model are fixed at this stage.

2.2.2 Instruction Tuning for Real-Time Controllability

This stage aims to gain real-time controllability by training the model on high-quality instruction tuning data. We construct such a dataset, which contains captions to precisely describe the dynamics of different clips in each video. With the data, we finetune the model by minimizing a diffusion loss on the videos given the instructions. In this stage, both the video generator and query embeddings are finetuned, while other components are fixed.

Below we describe the creation of the instruction tuning data in more details. An overview of the collected data is summarized in Table 1. The data come from both public corpus and simulators with careful data processing.

Public Video Datasets   To make the dataset general, we use a large-scale video dataset, Panda-70M [15]. We first filter the dataset by aesthetic score evaluation, optical flow magnitude assessment, cut detection, static video detection, and clip length filtering. Different from previous text-to-video models, our model emphasizes the controllability of natural language actions towards the next state. Therefore, we do re-captioning of the videos to get better captions that focus on the dynamics of the videos. we prompt GPT-4 Turbo [1] to generate captions describing the dynamics of four frames sampled from each video clip. This process yields a total of 500k video-text pairs. Besides Panda-70M, we also collect video-action pairs from existing action-annotated datasets, including Something-Something V2 [26], BridgeData V2 [73], and EPIC-KITCHENS [20]. This includes 260k examples.

Simulation Data   To provide our model with more diverse and accurate training experience, we use simulation environments to collect video-action pairs. CARLA [22] is a simulation platform for autonomous driving. It supports flexible modifications to the environment at runtime, making it suitable for simulating unexpected actions, such as *Change the weather to Sunset* or *Add a car to the front*. We sampled 75k video-action pairs from Carla. MP3D [13] and StreetLearn [50, 49] are indoor and urban panorama scans. We built simulation environments to render these 3D scans. Turning actions such as *turn right for 60 degrees* can be constructed by gradually changing camera poses and collecting corresponding image projections.

Dataset	Category	#Videos
Panda-70M	YouTube	500k
Something-Something	Human Activity	188k
BridgeData V2	Robot Arm	33k
HM3D	Indoor	152k
MP3D	Indoor	70k
StreetLearn	Street view	146k
Carla	Driving	75k
Coinrun	2D Game	30k
EPIC-KITCHENS	Kitchen	39k
Total	-	1.2M

Table 1: Instruction Tuning data statistics.

Besides, we prompt GPT-4 Turbo to generate scene descriptions, so that the instructions include both turning actions and the final scene descriptions. We got 70k data from MP3D and 146k data from StreetLearn, respectively. HM3D [57] is a 3D environment dataset of real-world indoor scenes. We used Habitat-Lab [56, 65, 60] to render these indoor scenes and collect data by sampling trajectories randomly. We created 152k data from it. Finally, we used Coinrun [19] for collecting 2D game simulation data, resulting in 30k data.

3 Qualitative Results

We show qualitative results that demonstrate the core capabilities of *Pandora* as a world simulator. Readers are encouraged to refer to `https://world-model.ai` for live video examples. We aim to report more quantitative results in the future.

### 3.1 On-The-Fly Control across Domains

Pandora is a general world model capable of generating videos across a broad range of domains. It permits on-the-fly control with free-text actions, *i.e.*, it can accept text action control anytime during the video generation and predict future world states accordingly. We show the generation results of indoor/outdoor videos in Figure 3, robot/human videos in Figure 4, and 2D/3D game videos in Figure 5. In Figure 6, we also show videos that correctly demonstrate basic physical phenomena, demonstrating the model's understanding of real-world physical concepts.

3.2 Action Controllability Transfer

Although some actions and their corresponding motion patterns only appear in some of the simulation data, we found that *Pandora* can transfer the action controllability to different unseen domains. As shown in Figure 7 and Figure 8, *Pandora* transfers 2D game ability from Coinrun and 3D simulator ability from HM3D to other unseen domains, respectively.

3.3 Autoregressively Generating Longer Videos

With the autoregressive backbone, *Pandora* is capable of generating longer videos of higher quality in an autoregressive manner. *Pandora* is trained on videos with up to 5 seconds (40 frames), but it is able to generate longer videos. We show the results of generating 8-second (64-frame) videos in Figure 9.
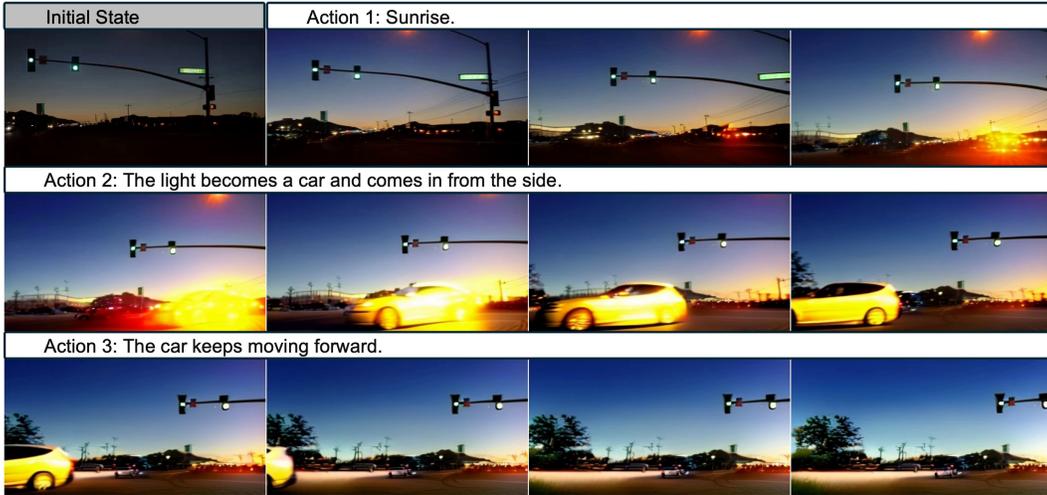
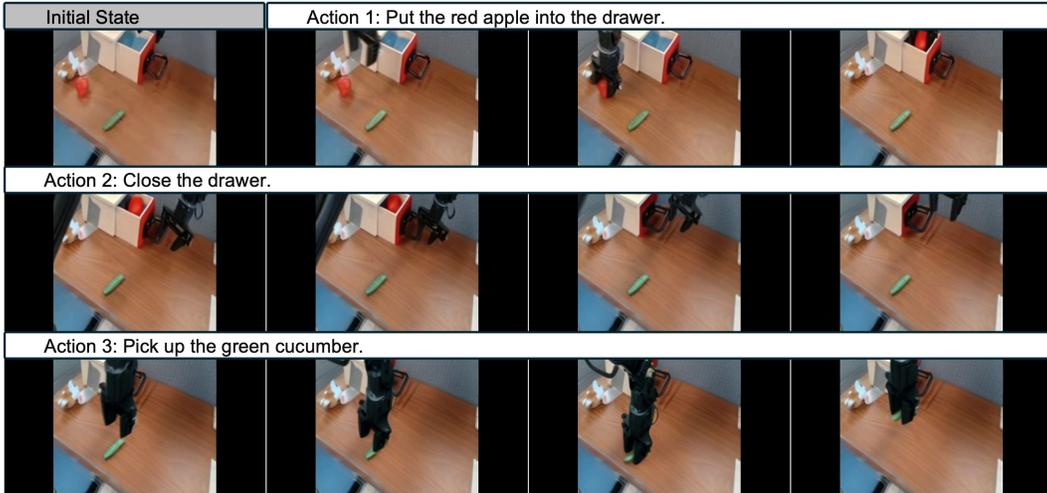(a) Sci-fi style movie scene.

(b) Everyday city scene.

(c) Household scene.

Figure 3: *Pandora* can generate sci-fi or real-life videos in both indoor and outdoor environments.

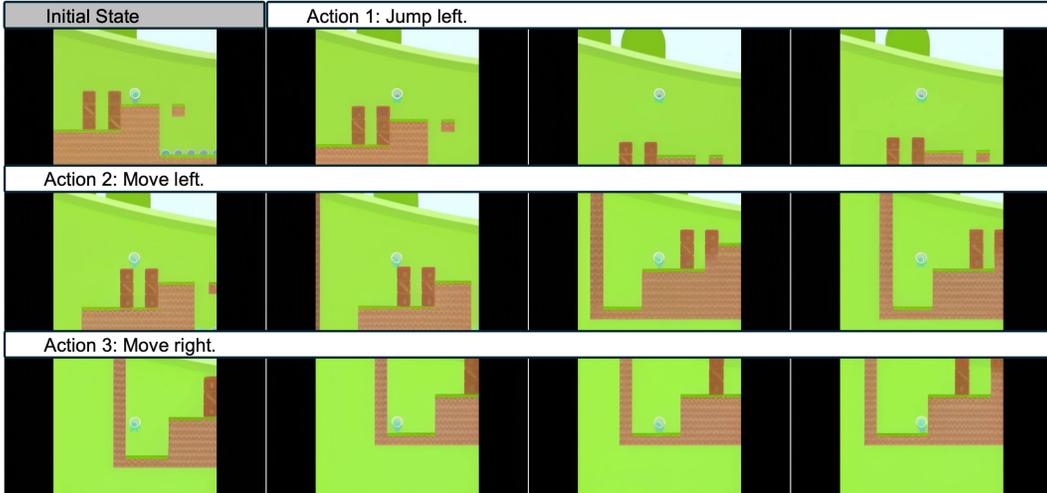(a) A robotic arm manipulating several stuff.

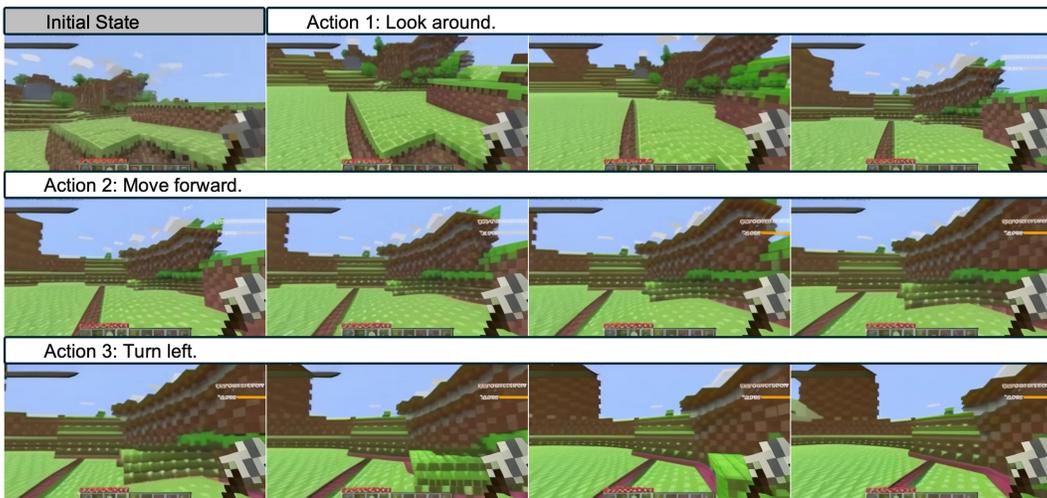(b) A real human doing some actions.

(c) An ego-view human activity

Figure 4: *Pandora* is capable of generating both robotics and human videos.
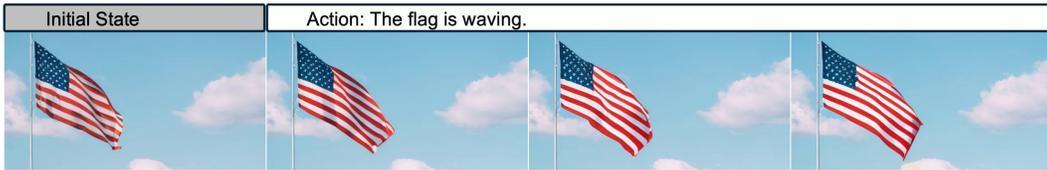
(a) 2D game Coinrun.

(b) 3D game Minecraft.

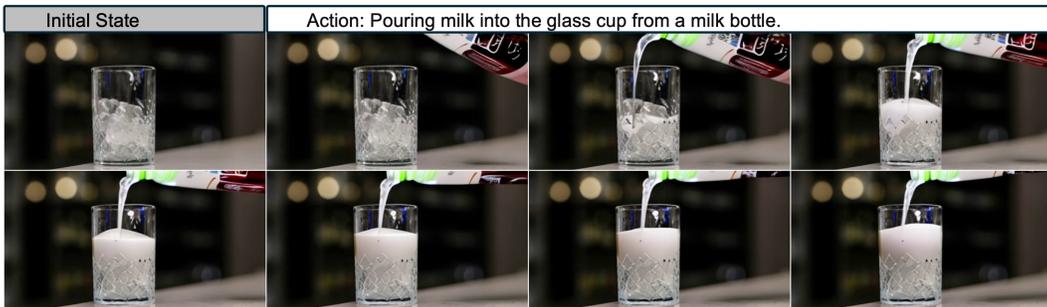(c) 3D game Grand Theft Auto V.

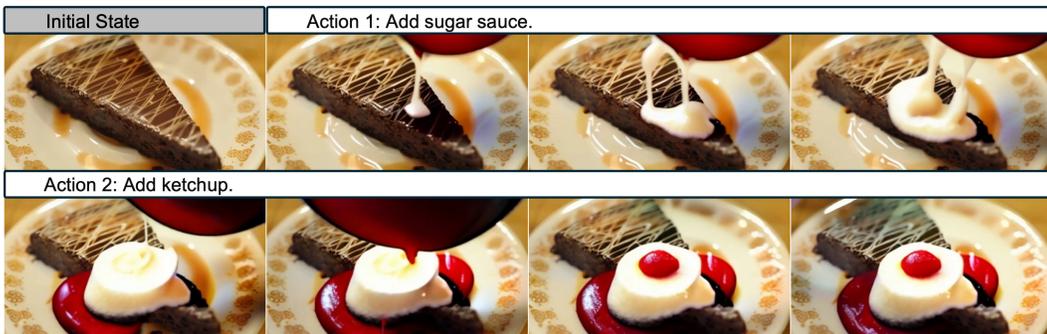Figure 5: *Pandora* can generate various 2D and 3D game videos.

(a) Object movements because of wind.

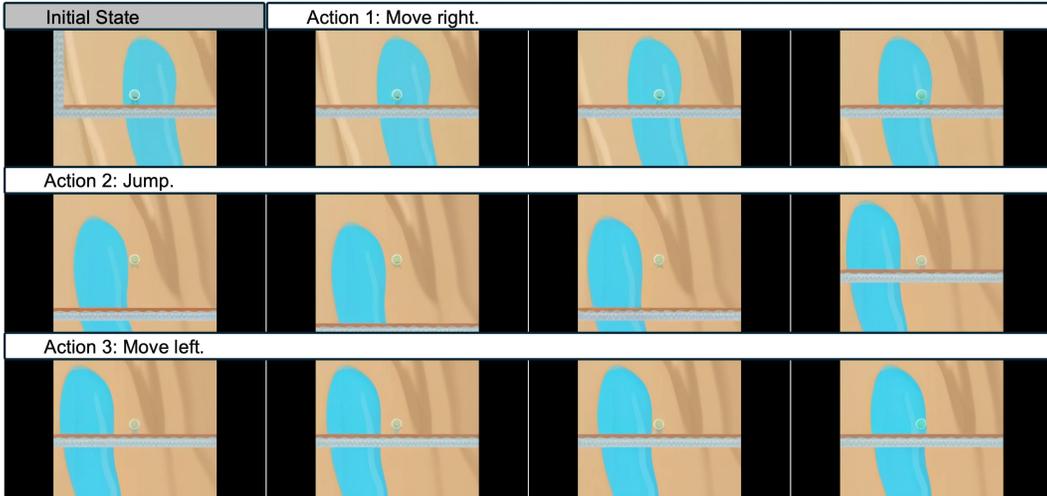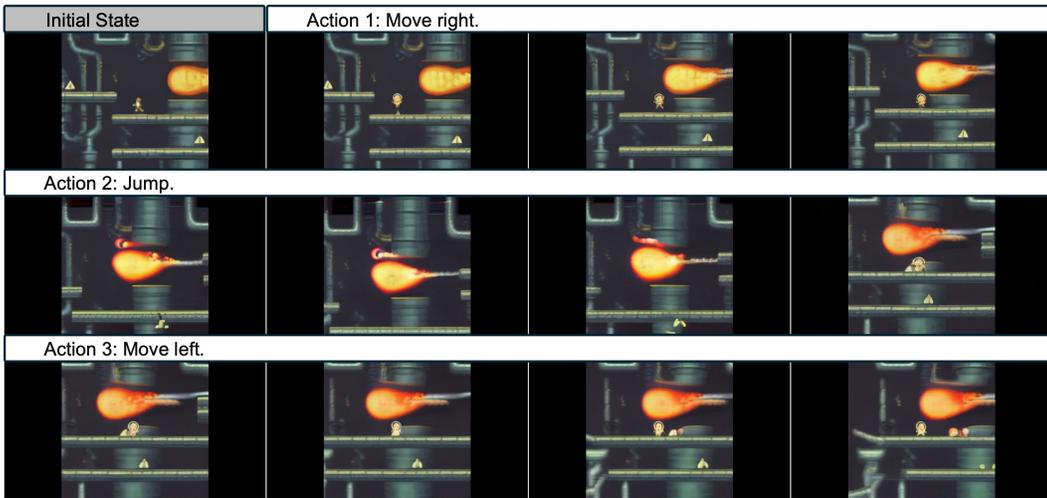(b) Object movements because of wind.

(c) The flow of liquid.

(d) The flow of viscous liquid.

Figure 6: *Pandora* is capable of generating videos that include common physical phenomena.

(a) Source domain (Coinrun).

(b) Target domain (Game 1).
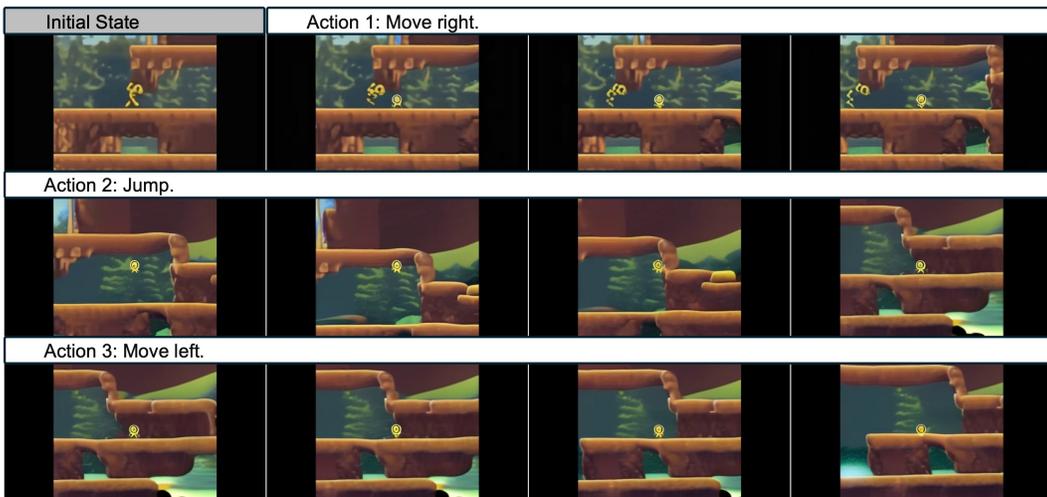
(c) Target domain (Game 2).

Figure 7: *Pandora* transfers the 2D game ability from the only 2D game in our training data, Coinrun, to other unseen 2D games.

(a) Source domain (HM3D).

(b) Target domain (Magic scene).

(c) Target domain (Colorful nature scene).

Figure 8: *Pandora* transfers the 3D indoor simulator ability to other unseen domains.

Figure 9: *Pandora* is capable of generating longer video autoregressively.

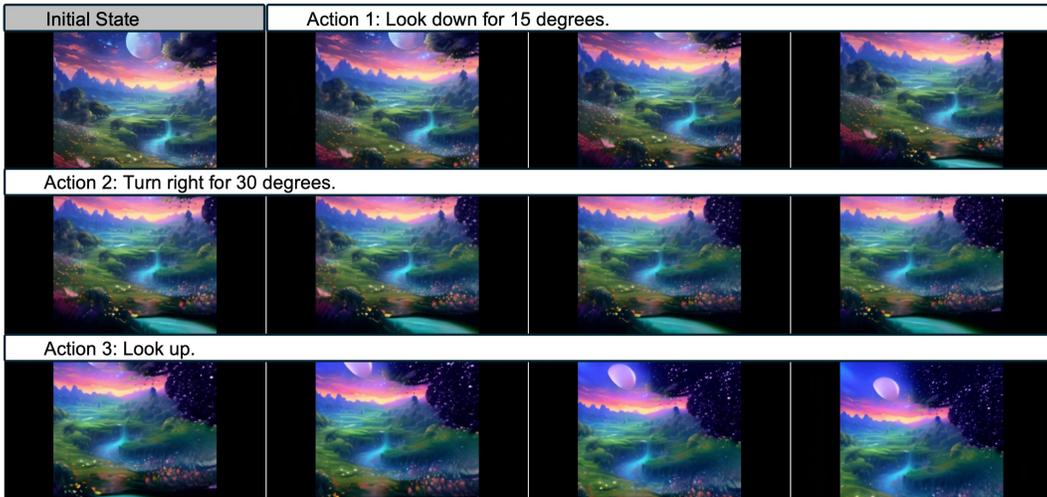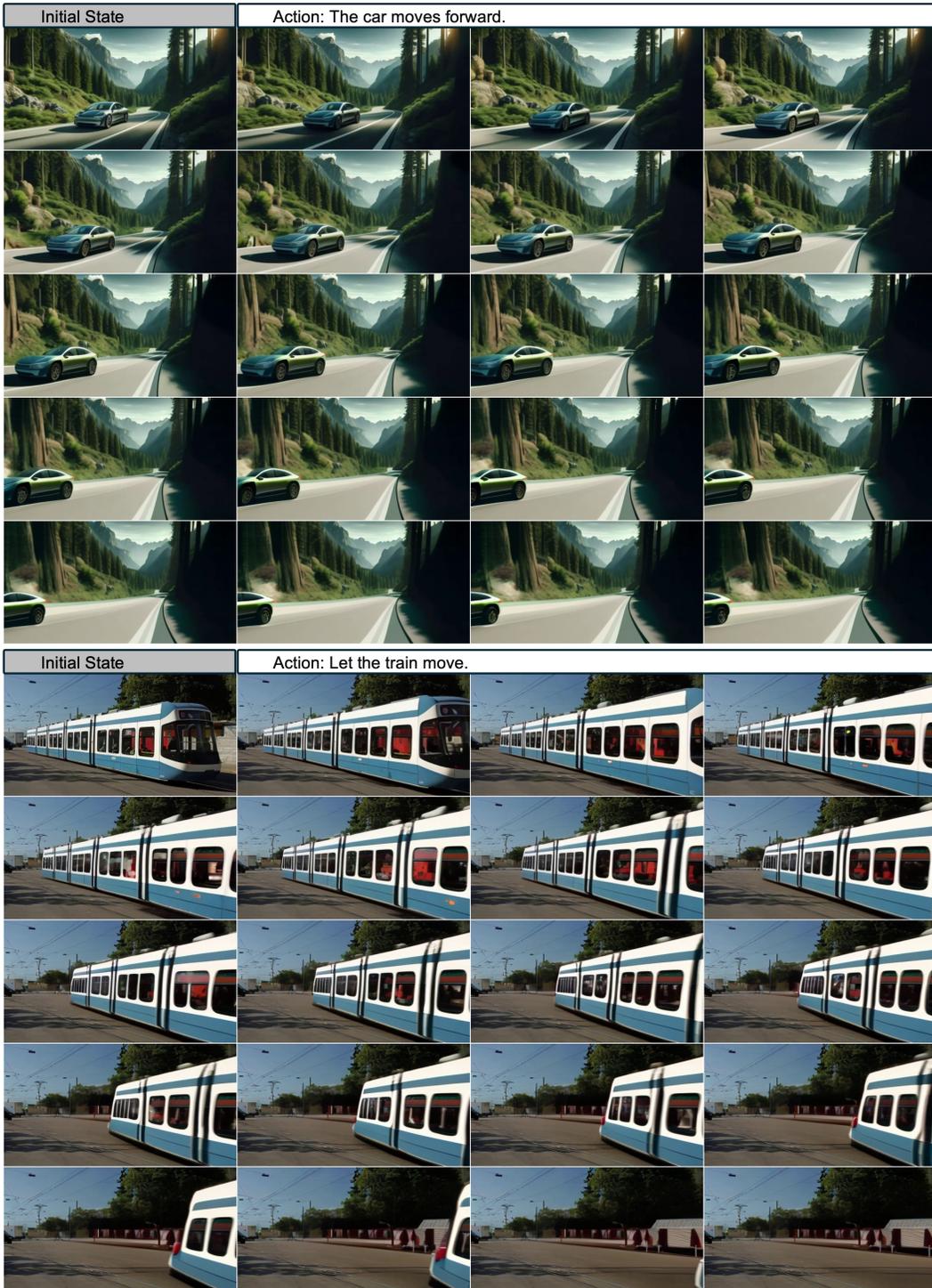## 3.4   Limitations

Pandora can struggle to generate videos with high quality and good controllability. Figure 10 shows failure cases about semantics understanding, motion control, and video consistency.
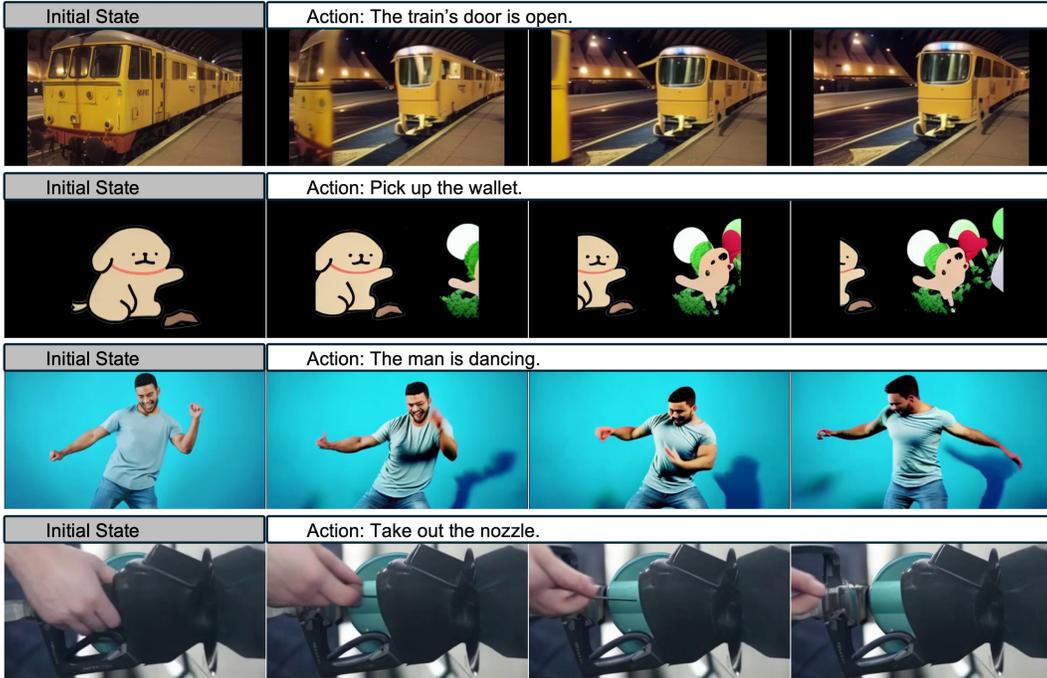

Figure 10: *Pandora* can fail in generating consistent videos, simulating complex scenarios, understanding commonsense and physical laws, and following instructions/actions.

(a) Results with small-scale training compute.

(b) Results with large-scale training compute.

Figure 11: Action: "the woman turns her head left". After scaling-up, the model shows better video quality and controllability.

When conducting small-scale exploratory experiments, we found that the data quality, *i.e.*, the precision of the dynamics descriptions, has great influence on the model performance. In the domains where high-quality simulation data exists, the model easily gains great controllability. But in the domains of public video datasets, where captions generated by GPT-4 Turbo are noisy, the model does not show good performance. However, when we increased the training compute, controllability across general domains emergents on the model. We show a result comparison between the models trained with small-scale and large-scale training compute in Figure 11. We hypothesize it is because

increasing data size can mitigate some of the noises in the data. The results indicate the great potential of building a stronger general world model by larger-scale training.

4 Related Works

World models World models simulate the future state of the world based on its previous states and given actions [69, 10, 8, 3, 55]. Previous world models in AI systems are usually designed for specific domains. For example, in robotics domain, world models are usually used for model-based reinforcement learning in specific simulators [27, 28, 47, 14, 39]. In robotics domain, world models [81, 88, 90, 2, 23] are capable of predicting future image or video states across diverse robotics environments. These predictive capabilities are important for robots to understand the environments, make informed decisions, and execute tasks accurately. Besides the robotics domain, world models are also widely used in autonomous driving [76, 75, 35, 45, 87, 84, 89], where they mainly focus on path planning and real-time decision-making, which is pivotal in enabling vehicles to navigate complex environments safely and efficiently. There are also world models for 2D games [6, 18, 24, 29, 41, 48, 58]. For example, Genie [12] is a generative model capable of simulating an interactive 2D game given an image. In this work, we make a step towards building a more general world model that simulates any-domain states given any-text actions at any time.

Video generation models Video generation models aim to synthesize realistic videos given text prompts or initial frames. Recent successes in diffusion models [33, 59, 62, 63] have paved the way for their application in the video generation domain [40, 54, 86, 16, 31, 61, 66, 72, 74]. For example, additional modules are introduced into the existing image diffusion models [34, 32, 9, 80, 85, 83] to facilitate video generation capabilities. However, the length of generated videos is limited due to the non-autoregressive nature. Consequently, the Diffusion Transformer (DiT) [53] has been proposed to allow for autoregressive generation, and Sora [11] has further scaled it up, achieving remarkable success in generating long, high-quality video. Furthermore, as the strong understanding and generation ability of LLMs, [67, 82] have explored the usage of LLMs in vision generation domain. Additionally, [42, 64] incorporate LLMs for video generation to enhance the semantic understanding. Previous models are designed to generate scenes from input descriptions, yet they frequently lack the ability to control actions or predict real-world states. On the contrary, *Pandora* is a hybrid autoregressive-diffusion model, thus it is capable of on-the-fly control over video generation.

5 Conclusion

We presented *Pandora* as a step towards building a general world model. The model is able to simulate world states by generating videos across different domains, and control the video on the fly with natural language actions. *Pandora* introduces a staged training recipe that allows to reuse and integrate existing pretrained language and video models. We believe larger-scale training with larger backbone models (e.g., GPT-4 and Sora) will lead to further improvement in terms of domain generality, video consistency, and action controllability. We are also excited about extending the model by incorporating other modalities, such as audio, to better measure and simulate the world.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[2] Anurag Ajay, Seungwook Han, Yilun Du, Shuang Li, Abhi Gupta, Tommi Jaakkola, Josh Tenenbaum, Leslie Kaelbling, Akash Srivastava, and Pulkit Agrawal. Compositional foundation models for hierarchical planning, 2023.

[3] Kelsey R Allen, Kevin A Smith, and Joshua B Tenenbaum. Rapid trial-and-error learning with simulation supports flexible tool use and physical reasoning. *PNAS*, 2020.

[4] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, Daniele Mazzotta, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. The falcon series of open language models, 2023.

[5] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021.

[6] Chris Bamford and Simon M Lucas. Neural game engine: Accurate learning of generalizable forward models from pixels. In *2020 IEEE Conference on Games (CoG)*, pages 81–88. IEEE, 2020.

[7] Fan Bao, Chendong Xiang, Gang Yue, Guande He, Hongzhou Zhu, Kaiwen Zheng, Min Zhao, Shilong Liu, Yaole Wang, and Jun Zhu. Vidu: a highly consistent, dynamic and skilled text-to-video generator with diffusion models. *arXiv preprint arXiv:2405.04233*, 2024.

[8] Peter W Battaglia, Jessica B Hamrick, and Joshua B Tenenbaum. Simulation as an engine of physical scene understanding. *PNAS*, 2013.

[9] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[10] Robert Eamon Briscoe. Mental imagery and the varieties of amodal perception. *Pacific Philosophical Quarterly*, 92(2):153–173, 2011.

[11] 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.

[12] Jake Bruce, Michael Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. *arXiv preprint arXiv:2402.15391*, 2024.

[13] Angel Chang, Angela Dai, Thomas Funkhouser, Maciej Halber, Matthias Niessner, Manolis Savva, Shuran Song, Andy Zeng, and Yinda Zhang. Matterport3d: Learning from rgb-d data in indoor environments. *International Conference on 3D Vision (3DV)*, 2017.

[14] Chang Chen, Jaesik Yoon, Yi-Fu Wu, and Sungjin Ahn. Transdreamer: Reinforcement learning with transformer world models. In *Deep RL Workshop NeurIPS 2021*, 2021.

[15] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers, 2024.

[16] Weifeng Chen, Yatai Ji, Jie Wu, Hefeng Wu, Pan Xie, Jiashi Li, Xin Xia, Xuefeng Xiao, and Liang Lin. Control-a-video: Controllable text-to-video generation with diffusion models. *arXiv preprint arXiv:2305.13840*, 2023.

[17] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, March 2023.

[18] Silvia Chiappa, Sébastien Racaniere, Daan Wierstra, and Shakir Mohamed. Recurrent environment simulators. In *International Conference on Learning Representations*, 2016.

[19] Karl Cobbe, Oleg Klimov, Chris Hesse, Taehoon Kim, and John Schulman. Quantifying generalization in reinforcement learning, 2019.

[20] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Scaling egocentric vision: The epic-kitchens dataset, 2018.

[21] Runpei Dong, Chunrui Han, Yuang Peng, Zekun Qi, Zheng Ge, Jinrong Yang, Liang Zhao, Jianjian Sun, Hongyu Zhou, Haoran Wei, et al. Dreamllm: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*, 2023.

[22] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017.

[23] Yilun Du, Mengjiao Yang, Bo Dai, Hanjun Dai, Ofir Nachum, Joshua B. Tenenbaum, Dale Schuurmans, and Pieter Abbeel. Learning universal policies via text-guided video generation, 2023.

[24] SM Ali Eslami, Danilo Jimenez Rezende, Frederic Besse, Fabio Viola, Ari S Morcos, Marta Garnelo, Avraham Ruderman, Andrei A Rusu, Ivo Danihelka, Karol Gregor, et al. Neural scene representation and rendering. *Science*, 360(6394):1204–1210, 2018.

[25] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7346–7356, 2023.

[26] Raghav Goyal, Samira Ebrahimi Kahou, Vincent Michalski, Joanna Materzyńska, Susanne Westphal, Heuna Kim, Valentin Haenel, Ingo Fruend, Peter Yianilos, Moritz Mueller-Freitag, Florian Hoppe, Christian Thurau, Ingo Bax, and Roland Memisevic. The "something something" video database for learning and evaluating visual common sense, 2017.

[27] David Ha and Jürgen Schmidhuber. Recurrent world models facilitate policy evolution. *Advances in neural information processing systems*, 31, 2018.

[28] David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[29] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[30] Shibo Hao, Yi Gu, Haodi Ma, Joshua Jiahua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with Language Model is Planning with World Model. *arXiv preprint arXiv:2305.14992*, 2023.

[31] William Harvey, Saeid Naderiparizi, Vaden Masrani, Christian Weilbach, and Frank Wood. Flexible diffusion modeling of long videos. *Advances in Neural Information Processing Systems*, 35:27953–27965, 2022.

[32] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022.

[33] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[34] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. Video diffusion models. *Advances in Neural Information Processing Systems*, 35:8633–8646, 2022.

[35] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving. *arXiv preprint arXiv:2309.17080*, 2023.

[36] Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023.

[37] Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. Mixtral of experts. *arXiv preprint arXiv:2401.04088*, 2024.

[38] Peng Jin, Ryuichi Takanobu, Caiwan Zhang, Xiaochun Cao, and Li Yuan. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *arXiv preprint arXiv:2311.08046*, 2023.

[39] Lukasz Kaiser, Mohammad Babaeizadeh, Piotr Milos, Blazej Osinski, Roy H Campbell, Konrad Czechowski, Dumitru Erhan, Chelsea Finn, Piotr Kozakowski, Sergey Levine, et al. Model-based reinforcement learning for atari. *arXiv preprint arXiv:1903.00374*, 2019.

[40] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023.

[41] Seung Wook Kim, Yuhao Zhou, Jonah Philion, Antonio Torralba, and Sanja Fidler. Learning to simulate dynamic environments with gamegan. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1231–1240, 2020.

[42] Dan Kondratyuk, Lijun Yu, Xiuye Gu, José Lezama, Jonathan Huang, Grant Schindler, Rachel Hornung, Vighnesh Birodkar, Jimmy Yan, Ming-Chang Chiu, Krishna Somandepalli, Hassan Akbari, Yair Alon, Yong Cheng, Josh Dillon, Agrim Gupta, Meera Hahn, Anja Hauth, David Hendon, Alonso Martinez, David Minnen, Mikhail Sirotenko, Kihyuk Sohn, Xuan Yang, Hartwig Adam, Ming-Hsuan Yang, Irfan Essa, Huisheng Wang, David A. Ross, Bryan Seybold, and Lu Jiang. Videopoet: A large language model for zero-shot video generation, 2024.

[43] Wenqiang Lai, Yuan Gao, and Tin Lun Lam. Vision-language model-based physical reasoning for robot liquid perception. *arXiv preprint arXiv:2404.06904*, 2024.

[44] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[45] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*, 2023.

[46] Zhengzhong Liu, Aurick Qiao, Willie Neiswanger, Hongyi Wang, Bowen Tan, Tianhua Tao, Junbo Li, Yuqi Wang, Suqi Sun, Omkar Pangarkar, et al. LLM360: Towards fully transparent open-source llms. *arXiv preprint arXiv:2312.06550*, 2023.

[47] Yutaka Matsuo, Yann LeCun, Maneesh Sahani, Doina Precup, David Silver, Masashi Sugiyama, Eiji Uchibe, and Jun Morimoto. Deep learning, reinforcement learning, and world models. *Neural Networks*, 2022.

[48] Vincent Micheli, Eloi Alonso, and François Fleuret. Transformers are sample-efficient world models. In *The Eleventh International Conference on Learning Representations*, 2022.

[49] Piotr Mirowski, Andras Banki-Horvath, Keith Anderson, Denis Teplyashin, Karl Moritz Hermann, Mateusz Malinowski, Matthew Koichi Grimes, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. The streetlearn environment and dataset, 2019.

[50] Piotr Wojciech Mirowski, Matthew Koichi Grimes, Mateusz Malinowski, Karl Moritz Hermann, Keith Anderson, Denis Teplyashin, Karen Simonyan, Koray Kavukcuoglu, Andrew Zisserman, and Raia Hadsell. Learning to navigate in cities without a map. *ArXiv*, abs/1804.00168, 2018.

[51] Melanie Mitchell and David C Krakauer. The debate over understanding in ai's large language models. *Proceedings of the National Academy of Sciences*, 120(13):e2215907120, 2023.

[52] OpenAI. Gpt-4 technical report, 2023.

[53] William Peebles and Saining Xie. Scalable diffusion models with transformers, 2023.

[54] Bo Peng, Xinyuan Chen, Yaohui Wang, Chaochao Lu, and Yu Qiao. Conditionvideo: Training-free condition-guided text-to-video generation. *arXiv preprint arXiv:2310.07697*, 2023.

[55] RT Pramod, Michael Cohen, Kirsten Lydic, Josh Tenenbaum, and Nancy Kanwisher. Evidence that the brain's physics engine runs forward simulations of what will happen next. *Journal of Vision*, 20(11):1521–1521, 2020.

[56] Xavier Puig, Eric Undersander, Andrew Szot, Mikael Dallaire Cote, Tsung-Yen Yang, Ruslan Partsey, Ruta Desai, Alexander William Clegg, Michal Hlavac, So Yeon Min, et al. Habitat 3.0: A co-habitat for humans, avatars and robots. *arXiv preprint arXiv:2310.13724*, 2023.

[57] Santhosh K. Ramakrishnan, Aaron Gokaslan, Erik Wijmans, Oleksandr Maksymets, Alex Clegg, John Turner, Eric Undersander, Wojciech Galuba, Andrew Westbury, Angel X. Chang, Manolis Savva, Yili Zhao, and Dhruv Batra. Habitat-matterport 3d dataset (hm3d): 1000 large-scale 3d environments for embodied ai, 2021.

[58] Jan Robine, Marc Höftmann, Tobias Uelwer, and Stefan Harmeling. Transformer-based world models are happy with 100k interactions. In *The Eleventh International Conference on Learning Representations*, 2022.

[59] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[60] Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A Platform for Embodied AI Research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019.

[61] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[62] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. PMLR, 2015.

[63] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

[64] Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Zhengxiong Luo, Yueze Wang, Yongming Rao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative multimodal models are in-context learners, 2024.

[65] Andrew Szot, Alex Clegg, Eric Undersander, Erik Wijmans, Yili Zhao, John Turner, Noah Maestre, Mustafa Mukadam, Devendra Chaplot, Oleksandr Maksymets, Aaron Gokaslan, Vladimir Vondrus, Sameer Dharur, Franziska Meier, Wojciech Galuba, Angel Chang, Zsolt Kira, Vladlen Koltun, Jitendra Malik, Manolis Savva, and Dhruv Batra. Habitat 2.0: Training home assistants to rearrange their habitat. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[66] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion. *Advances in Neural Information Processing Systems*, 36, 2024.

[67] Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models, 2024.

[68] Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

[69] Edward C Tolman. Cognitive maps in rats and men. *Psychological review*, 55(4):189, 1948.

[70] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[71] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and Yasmine et al. Babaei. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[72] Vikram Voleti, Alexia Jolicoeur-Martineau, and Chris Pal. Mcvd-masked conditional video diffusion for prediction, generation, and interpolation. *Advances in neural information processing systems*, 35:23371–23385, 2022.

[73] Homer Walke, Kevin Black, Abraham Lee, Moo Jin Kim, Max Du, Chongyi Zheng, Tony Zhao, Philippe Hansen-Estruch, Quan Vuong, Andre He, Vivek Myers, Kuan Fang, Chelsea Finn, and Sergey Levine. Bridgedata v2: A dataset for robot learning at scale, 2024.

[74] Weimin Wang, Jiawei Liu, Zhijie Lin, Jiangqiao Yan, Shuo Chen, Chetwin Low, Tuyen Hoang, Jie Wu, Jun Hao Liew, Hanshu Yan, et al. Magicvideo-v2: Multi-stage high-aesthetic video generation. *arXiv preprint arXiv:2401.04468*, 2024.

[75] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023.

[76] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. *arXiv preprint arXiv:2311.17918*, 2023.

[77] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

[78] Lionel Wong, Gabriel Grand, Alexander K Lew, Noah D Goodman, Vikash K Mansinghka, Jacob Andreas, and Joshua B Tenenbaum. From word models to world models: Translating from natural language to the probabilistic language of thought. *arXiv preprint arXiv:2306.12672*, 2023.

[79] Jiannan Xiang, Tianhua Tao, Yi Gu, Tianmin Shu, Zirui Wang, Zichao Yang, and Zhiting Hu. Language Models Meet World Models: Embodied Experiences Enhance Language Models. *arXiv preprint arXiv:2305.10626*, 2023.

[80] Jinbo Xing, Menghan Xia, Yong Zhang, Haoxin Chen, Xintao Wang, Tien-Tsin Wong, and Ying Shan. Dynamicrafter: Animating open-domain images with video diffusion priors. *arXiv preprint arXiv:2310.12190*, 2023.

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

[82] Lili Yu, Bowen Shi, Ramakanth Pasunuru, Benjamin Muller, Olga Golovneva, Tianlu Wang, Arun Babu, Binh Tang, Brian Karrer, Shelly Sheynin, Candace Ross, Adam Polyak, Russell Howes, Vasu Sharma, Puxin Xu, Hovhannes Tamoyan, Oron Ashual, Uriel Singer, Shang-Wen Li, Susan Zhang, Richard James, Gargi Ghosh, Yaniv Taigman, Maryam Fazel-Zarandi, Asli Celikyilmaz, Luke Zettlemoyer, and Armen Aghajanyan. Scaling autoregressive multi-modal models: Pretraining and instruction tuning, 2023.

[83] David Junhao Zhang, Dongxu Li, Hung Le, Mike Zheng Shou, Caiming Xiong, and Doyen Sahoo. Moonshot: Towards controllable video generation and editing with multimodal conditions. *arXiv preprint arXiv:2401.01827*, 2024.

[84] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Learning unsupervised world models for autonomous driving via discrete diffusion. *arXiv preprint arXiv:2311.01017*, 2023.

[85] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models. *arXiv preprint arXiv:2311.04145*, 2023.

[86] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023.

[87] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024.

[88] Haoyu Zhen, Xiaowen Qiu, Peihao Chen, Jincheng Yang, Xin Yan, Yilun Du, Yining Hong, and Chuang Gan. 3d-vla: A 3d vision-language-action generative world model. *arXiv preprint arXiv:2403.09631*, 2024.

[89] Wenzhao Zheng, Weiliang Chen, Yuanhui Huang, Borui Zhang, Yueqi Duan, and Jiwen Lu. Occworld: Learning a 3d occupancy world model for autonomous driving. *arXiv preprint arXiv:2311.16038*, 2023.

[90] Siyuan Zhou, Yilun Du, Jiaben Chen, Yandong Li, Dit-Yan Yeung, and Chuang Gan. Robodreamer: Learning compositional world models for robot imagination, 2024.