

---

# Whole-Body Conditioned Egocentric Video Prediction

---

**Yutong Bai\***<sup>1</sup>

**Danny Tran\***<sup>1</sup>

**Amir Bar\***<sup>2</sup>

**Yann LeCun†<sup>2,3</sup>**

**Trevor Darrell†<sup>1</sup>**

**Jitendra Malik†<sup>1,2</sup>**

<sup>1</sup>UC Berkeley (BAIR)

<sup>2</sup>FAIR, Meta

<sup>3</sup>New York University

## Abstract

We train models to Predict Ego-centric Video from human Actions (**PEVA**), given the past video and an action represented by the relative 3D body pose. By conditioning on kinematic pose trajectories, structured by the joint hierarchy of the body, our model learns to simulate how physical human actions shape the environment from a first-person point of view. We train an auto-regressive conditional diffusion transformer on Nymeria, a large-scale dataset of real-world egocentric video and body pose capture. We further design a hierarchical evaluation protocol with increasingly challenging tasks, enabling a comprehensive analysis of the model’s embodied prediction and control abilities. Our work represents an initial attempt to tackle the challenges of modeling complex real-world environments and embodied agent behaviors with video prediction from the perspective of a human.<sup>1</sup>

## 1 Introduction

Human movement is rich, continuous, and physically grounded (Rosenhahn et al., 2008; Aggarwal and Cai, 1999). The way we walk, lean, turn, or reach—often subtle and coordinated—directly shapes what we see from a first-person perspective. For embodied agents to simulate and plan like humans, they must not only predict future observations (Von Helmholtz, 1925), but also understand how visual input arises from whole-body action (Craik, 1943). This understanding is essential because many aspects of the environment are not immediately visible—we need to move our bodies to reveal new information and achieve our goals.

Vision serves as a natural signal for long-term planning (LeCun, 2022; Hafner et al., 2023; Ebert et al., 2018; Ma et al., 2022). We look at our environment to plan and act, using our egocentric view as a predictive goal (Sridhar et al., 2024; Bar et al., 2025). When we consider our body movements, we should consider both actions of the feet (locomotion and navigation) and the actions of the hand (manipulation), or more generally, whole-body control (Nvidia et al., 2025; Cheng et al., 2024; He et al., 2024b; Radosavovic et al., 2024; He et al., 2024a; Hansen et al., 2024). For example, when reaching for an object, we must anticipate how our arm movement will affect what we see, even before the object comes into view. This ability to plan based on partial visual information is crucial for embodied agents to operate effectively in real-world environments.

Building a model that can effectively learn from and predict based on whole-body motion presents several fundamental challenges. First, representing human actions requires capturing both global body dynamics and fine-grained joint articulations, which involves high-dimensional, structured data with complex temporal dependencies. Second, the relationship between body movements and visual perception is highly nonlinear and context-dependent—the same arm movement can result in

---

\* Equal contribution; † Equal advising.

<sup>1</sup>Project page: <https://dannytran123.github.io/PEVA>.

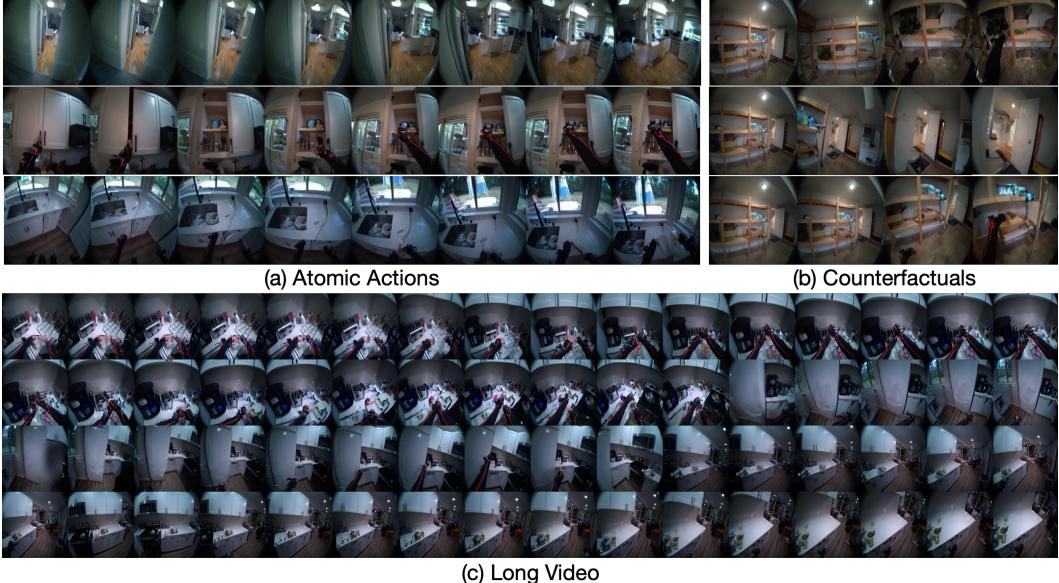


Figure 1: **Predicting Ego-centric Video from human Actions (PEVA)**. Given past video frames and an action specifying a desired change in 3D pose, PEVA predicts the next video frame. Our results show that, given the first frame and a sequence of actions, our model can generate videos of atomic actions (a), simulate counterfactuals (b), and support long video generation (c).

different visual outcomes depending on the environment and the agent’s current state. Third, learning these relationships from real-world data is particularly challenging due to the inherent variability in human motion and the subtle, often delayed visual consequences of actions.

To address these challenges, we develop a novel approach **PEVA** that combines several key innovations. First, we design a structured action representation that preserves both global body dynamics and local joint movements, using a hierarchical encoding that captures the kinematic tree structure of human motion. This representation enables the model to understand both the overall body movement and the fine-grained control of individual joints. Second, we develop a novel architecture based on conditional diffusion transformers that can effectively model the complex, nonlinear relationship between body movements and visual outcomes. The architecture incorporates temporal attention mechanisms to capture long-range dependencies and a specialized action embedding component that maintains the structured nature of human motion. Third, we leverage a large-scale dataset of synchronized egocentric video and motion capture data (Ma et al., 2024), which provides the necessary training signal to learn these complex relationships. Our training strategy includes random timeskips to handle the delayed visual consequences of actions and sequence-level training to maintain temporal coherence.

For evaluation, we first assess the model’s ability to predict immediate visual consequences by evaluating its performance on single-step predictions at 2-second intervals, measuring both perceptual quality (LPIPS (Zhang et al., 2018)) and semantic consistency (DreamSim (Fu et al., 2023)). Second, we decompose complex human movements into atomic actions—such as hand movements (up, down, left, right) and whole-body movements (forward, rotation)—to test the model’s understanding of how specific joint-level movements affect the egocentric view. This fine-grained analysis reveals whether the model can capture the nuanced relationship between individual joint movements and their visual effects. Third, we examine the model’s capability to predict long-term visual consequences by evaluating its performance across extended time horizons (up to 16 seconds), where the effects of actions may be delayed or not immediately visible. Finally, we explore the model’s ability to serve as a world model for planning by using it to simulate actions and choose the ones that lead to a predefined goal. This layered approach allows us to systematically analyze the strengths and limitations of our model, revealing both its capacity to simulate embodied perception and the open challenges that remain in bridging the gap between physical action and visual experience.

To conclude, we introduce PEVA, a model that predicts future egocentric video conditioned on whole-body human motion. By leveraging structured action representations derived from 3D pose trajectories, our model captures the intricate relationship between physical movement and visual

perception. We develop a diffusion-based architecture that can be trained auto-regressively over a sequence of images in a parallelized fashion. Our model utilizes random time-skips that enable covering long-term videos efficiently. Trained on Nymeria (Ma et al., 2024), a large-scale real-world dataset of egocentric video and synchronized motion, PEVA advances embodied simulation with physically grounded, visually realistic predictions. Our comprehensive evaluation framework demonstrates that whole-body control significantly improves video quality, semantic consistency, and simulating counterfactual.

## 2 Related Works

**World Models.** The concept of a “world model”, an internal representation of the world used for prediction and planning, has a rich history across multiple disciplines. The idea was first proposed in psychology by Craik (1943), who hypothesized that the brain uses “small-scale models” of reality to anticipate events. This principle found parallel development in control theory, where methods like the Kalman Filter and Linear Quadratic Regulator (LQR) rely on an explicit model of the system to be controlled (Kalman, 1960). The idea of internal models became central to computational neuroscience for explaining motor control, with researchers proposing that the brain plans and executes movements by simulating them first (Jordan, 1996; Kawato et al., 1987; Kawato, 1999).

With the rise of deep learning, the focus shifted to learning these predictive models directly from data. Early work in computer vision demonstrated that models could learn intuitive physics from visual data to solve simple control tasks like playing billiards or poking objects (Fragkiadaki et al., 2015; Agrawal et al., 2016). This paved the way for modern, large-scale world models that predict future video frames conditioned on actions, enabling planning by “imagining” future outcomes (Ha and Schmidhuber, 2018; Hafner et al.; Liu et al., 2024; Li et al., 2022; Zhou et al., 2024; Yang et al., 2023, 2024; Assran et al., 2025). In reinforcement learning, models like Dreamer have shown that learning a world model improves sample efficiency (Hafner et al., 2023). Recent approaches have used diffusion models for more expressive generation; for example, DIAMOND generates multi-step rollouts via autoregressive diffusion (Alonso et al., 2024). In the egocentric domain, Navigation World Models (NWM) used conditional diffusion transformers (CDiT) to predict future frames from a planned trajectory (Bar et al., 2025). However, these models use low-dimensional controls and neglect the agent’s own body dynamics. We build on this extensive line of work by conditioning video prediction on whole-body pose, enabling a more physically-grounded simulation.

**Human Motion Generation and Controllable Prediction.** Human motion modeling has advanced from recurrent and VAE-based methods (Rempe et al., 2021; Petrovich et al., 2021; Ye et al., 2023) to powerful diffusion-based generators (Tevet et al., 2022; Zhang et al., 2024). These models generate diverse, realistic 3D pose sequences conditioned on text (Hong et al., 2024; Guo et al., 2022; Dabral et al., 2023), audio (Ng et al., 2024; Dabral et al., 2023; Ao et al., 2023), and head pose (Li et al., 2023; Castillo et al., 2023; Yi et al., 2025). Recent works like Animate Anyone (Hu et al., 2023) and MagicAnimate (Xu et al., 2023) generate high-fidelity human animations from a reference image and pose sequence. Physically-aware extensions like PhysDiff (Yuan et al., 2023) incorporate contact into the denoising loop. While prior works treat pose as the target, our model uses it as input for egocentric video prediction, reversing the typical motion generation setup. This enables fine-grained visual control, bridging pose-conditioned video generation (Wu et al., 2023; Zhang et al., 2023) with embodied simulation. Unlike Make-a-Video (Singer et al., 2022) or Tune-A-Video (Wu et al., 2023), which focus on text/image prompts, we condition directly on physically realizable body motion.

**Egocentric Perception and Embodied Forecasting.** Egocentric video datasets such as Ego4D (Grauman et al., 2022), Ego-Exo4D (Grauman et al., 2024) and EPIC-KITCHENS (Damen et al., 2018) were used to study human action recognition, object anticipation (Furnari and Farinella, 2020), future video prediction (Girdhar and Grauman, 2021), and even animal behavior (Bar et al., 2024). To study pose estimation, EgoBody (Zhang et al., 2022) and Nymeria (Ma et al., 2024) provide synchronized egocentric video and 3D pose. Unlike these works, we treat future body motion as a control signal, enabling visually grounded rollout. Prior works in egocentric pose forecasting (Yuan and Kitani, 2019) and visual foresight (Finn and Levine, 2017) show that predicting future perception supports downstream planning. Our model unifies these lines by predicting future egocentric video from detailed whole-body control, enabling first-person planning with physical and visual realism.

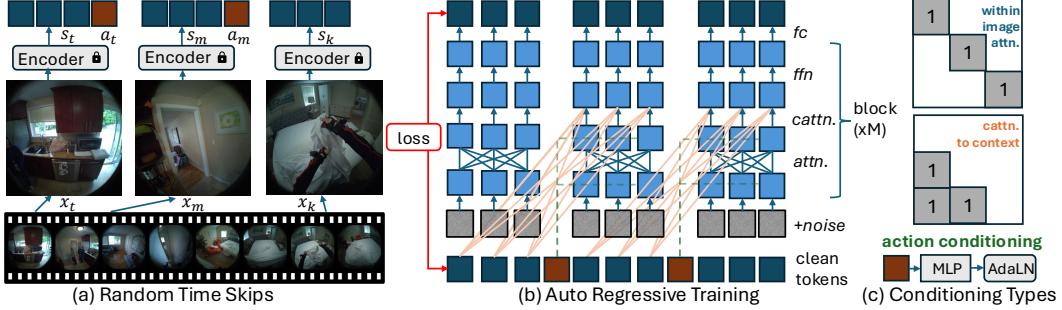


Figure 2: **Design of PEVA.** To train on an input video, we choose a random subset of frames and encode them via a fixed encoder (a). They are then fed to a CDiT that is trained autoregressively with teacher forcing (b). During the denoising process, each token attends to same-image tokens and cross-attends to clean tokens from past image(s). Action conditioning is done via AdaLN layers.

### 3 PEVA

In this section we describe our whole-body-conditioned ego-centric video prediction model. We start by describing how to represent human actions (Section 3.1), then move on to describe the model and the training objective (Section 3.2). Finally, we describe the model architecture in Section 3.3.

#### 3.1 Structured Action Representation from Motion Data

To effectively capture the relationship between human motion and egocentric visual perception, we define each action as a high-dimensional vector encoding both global body dynamics and detailed joint articulations. Rather than relying on simplified or discrete controls, our framework uses full-body motion information, including global translation (via the root joint) and relative joint rotations structured by the kinematic tree. This design ensures that the action space richly represents human movement at both coarse and fine levels.

To construct this representation, we synchronize motion capture data with video frames based on timestamps, then transform global coordinates into a local frame centered at the pelvis. This transformation makes the data invariant to initial position and orientation. Global positions are converted to local coordinates, quaternions to relative Euler angles, and joint relationships are preserved using the kinematic hierarchy. We normalize all motion parameters for stable learning: positions are scaled to  $[-1, 1]$  and rotations bounded within  $[-\pi, \pi]$ . Each action reflects the change between consecutive frames, capturing motion over time and enabling the model to learn how physical actions translate into visual outcomes.

#### 3.2 Ego-Centric Video Prediction for Whole-Body Control

Next, we describe our formulation of **PEVA** from the perspective of an embodied agent. Intuitively, the model is an autoregressive diffusion model that receives an input video and a corresponding sequence of actions describing how the agent moves and acts. Given any prefix of frames and actions, the model predicts the resulting state of the world after applying the last action and considering other environment dynamics.

More formally, we are given a dataset  $D = \{(x_0, a_0, \dots, x_T, a_T)\}_{i=1}^n$  of agents videos from egocentric view and their associated body controls, such that every  $x_j \in \mathbb{R}^{H \times W \times 3}$  is a video frame and  $a_j \in \mathbb{R}^{d_{act}}$  an action in the Xsens skeleton ordering (Movella, 2021) for the upper body (everything above the pelvis), representing the change in translation, together with the delta rotation of all joints relative to the previous joint rotation. We represent motion in 3D space, thus we have 3 degrees of freedom for root translation, 15 joints for the upper body and represent relative joint rotations as Euler angles in 3D space leaving  $d_{act} = 3 + 15 \times 3 = 48$ .

We start by encoding each individual frame  $s_i = \text{enc}(x_i)$  into a corresponding state representation, through a pre-trained VAE encoder (Rombach et al., 2022). Given a sequence of controls  $a_0, \dots, a_T$ ,

our goal is to build a generative model that captures the dynamics of the environment:

$$P(s_T, \dots, s_0 | a_T, \dots, a_0) = P(s_0) \prod_{t=0}^{T-1} P(s_{t+1} | s_t, \dots, s_0, a_T, \dots, a_0) \quad (1)$$

To simplify the model, we factorize the distribution and make a Markov assumption that the next state is dependent on the last  $k$  states and a single past action:

$$P(s_{t+1} | s_t, \dots, s_0, a_T, \dots, a_0) = P(s_{t+1} | s_t, \dots, s_{t-k+1}, a_t) \quad (2)$$

We aim to train a model parametrized by  $\theta$  that minimizes the negative log-likelihood:

$$\hat{\theta} = \arg \min_{\theta} \left[ -\log P_{\theta}(s_0) - \sum_{t=0}^{T-1} \log P_{\theta}(s_{t+1} | s_t, \dots, s_{t-k+1}, a_t) \right]$$

We model each transition  $P_{\theta}(s_{t+1} | s_t, \dots, s_{t-k+1}, a_t)$  using a Denoising Diffusion Probabilistic Model (DDPM) (Ho et al., 2020), which maximizes the (reweighted) evidence lower bound (ELBO) of the log-likelihood. For each transition, we define the forward diffusion process  $q(z_{\tau} | s_{t+1}) = \mathcal{N}(z_{\tau}; \sqrt{\bar{\alpha}_{\tau}} s_{t+1}, (1 - \bar{\alpha}_{\tau}) \mathbf{I})$ , where  $z_{\tau}$  is the noisy version of  $s_{t+1}$  at noise timestep  $\tau$ , and  $\bar{\alpha}_{\tau}$  is the cumulative product of noise scales. The reverse process is learned by training a neural network  $\epsilon_{\theta}$  to predict the noise given  $z_{\tau}$  and the conditioning context  $c_t = (s_t, \dots, s_{t-k+1}, a_t)$ .

Then denoising loss term for a transition is given by:

$$\mathcal{L}_{\text{simple}, t} = \mathbb{E}_{\tau, \epsilon \sim \mathcal{N}(0, I)} \left[ \left\| \epsilon - \epsilon_{\theta} \left( \sqrt{\bar{\alpha}_{\tau}} s_{t+1} + \sqrt{1 - \bar{\alpha}_{\tau}} \epsilon, c_t, \tau \right) \right\|^2 \right] \quad (3)$$

Where  $\mathcal{L}_{\text{simple}, 0}$  is the loss term corresponding to the unconditional generation of  $s_0$ . Additionally, we also predict the covariances of the noise, and supervise them using the full variational lower bound loss  $\mathcal{L}_{\text{vib}, t}$  as proposed by (Nichol and Dhariwal, 2021).

Hence the final objective yields a (weighted) version of the ELBO for each term in the sequence:

$$\mathcal{L} = \sum_{t=0}^{T-1} \mathcal{L}_{\text{simple}, t} + \lambda \mathcal{L}_{\text{vib}, t} \quad (4)$$

Despite not being a lower bound of the log-likelihood, the reweighted ELBO works well in practice for image generation with transformers (Nichol and Dhariwal, 2021; Peebles and Xie, 2023).

The advantage of our formulation is that it allows training in parallelized fashion using causal masking. Given a sequence of frames and actions, we can train on every prefix of the sequence in a single forward-backward pass. Next, we elaborate on the architecture of our model.

### 3.3 Autoregressive Conditional Diffusion Transformer

While prior work in navigation world models (Bar et al., 2025) focuses on simple control signals like velocity and heading, modeling whole-body human motion presents significantly greater challenges. Human activities involve complex, coordinated movements across multiple degrees of freedom, with actions that are both temporally extended and physically constrained. This complexity necessitates architectural innovations beyond standard CDiT approaches.

To address these challenges, we extend the Conditional Diffusion Transformer (CDiT) architecture with several key modifications that enable effective modeling of whole-body motion:

**Random Timeskips.** Human activities often span long time horizons with actions that can take several seconds to complete. At the same time, videos are a raw signal which requires vast amounts of compute to process. To handle video more efficiently, we introduce random timeskips during training (see Figure 2a), and include the timeskip as an action to inform the model’s prediction. This allows the model to learn both short-term motion dynamics and longer-term activity patterns. Learning long-term dynamics is particularly important for modeling activities like reaching, bending, or walking, where the full motion unfolds over multiple seconds. In practice, we sample 16 video frames from a 32 second window.

**Sequence-Level Training.** Unlike NWM which predicts single frames, we model the entire sequence of motion by applying the loss over each prefix of frames following Eq. 4. We include an example of

this in Figure 2b. This is crucial because human activities exhibit strong temporal dependencies - the way someone moves their arm depends on their previous posture and motion. We enable efficient training by parallelizing across sequence prefixes through spatial-only attention in the current frame and past-frame-only attention for historical context (Figure 2c). In practice we train models with sequences of 16 frames.

**Action Embeddings.** The high-dimensional nature of whole-body motion (joint positions, rotations, velocities) requires careful handling of the action space. We take the most simple strategy: we concatenate all actions in time  $t$  into a  $1D$  tensor which is fed to each AdaLN layer for conditioning (see Figure 2c).

These architectural innovations are essential for modeling the rich dynamics of human motion. By training on sequence prefixes and incorporating timeskips, our model learns to generate temporally coherent motion sequences that respect both short-term dynamics and longer-term activity patterns. The specialized action embeddings further enable precise control over the full range of human movement, from subtle adjustments to complex coordinated actions.

### 3.4 Inference and Planning with PEVA

**Sampling procedure at test time.** Given a set of context frames  $(x_t, \dots, x_{t-k+1})$ , we encode these frames to get  $(s_t, \dots, s_{t-k+1})$  and pass the encoded context as the clean tokens in Figure 2b and pass in randomly sampled noise as the last frame. We then follow the DPPM sampling process to denoise the last frame conditioning on our action. For faster inference time, we employ special attention masks where we change the mask in Figure 2c for within image attention to only be applied on the tokens of the last frame and change the mask for cross attention to context so that cross attention is only applied for the last frame.

**Autoregressive rollout strategy.** To follow a set of actions we use an autoregressive rollout strategy. Given an initial set of context frames we  $(x_t, \dots, x_{t-k+1})$  we start by encoding each individual frame to get  $(s_t, \dots, s_{t-k+1})$  and add the current action to create the conditioning context  $c_t = (s_t, \dots, s_{t-k+1}, a_t)$ . We then sample from our model parameterized by  $\theta$  to generate the next state:  $s_{t+1} = P_\theta(s_{t+1}|c_t)$ . We then discard the first encoding and append the generated  $s_{t+1}$  and add the next action to produce the next context  $c_{t+1} = (s_{t+1}, s_t, \dots, s_{t-k+1}, a_{t+1})$ . We then repeat the process for our entire set of actions. Finally, to visualize the predictions, we decode the latent states to pixels using the VAE decoder Rombach et al. (2022).

## 4 Experiments and Results

### 4.1 Experiment Setting

**Dataset.** We use the Nymeria dataset (Ma et al., 2024), which contains synchronized egocentric video and full-body motion capture, recorded in diverse real-world settings using an XSens system (Movella, 2021). Each sequence includes RGB frames and 3D body poses in the XSens skeleton format, covering global translation and rotations of body joints. We sample body motions at 4 FPS. Videos are center-cropped and resized to  $224 \times 224$ . We split the dataset 80/20 for training and evaluation, and report all metrics on the validation set.

**Training Details.** We train variants of Conditional Diffusion Transformer (CDiT-S to CDiT-XXL, up to 32 layers) using a context window of 3–15 frames and predicting 64-frame trajectories. Models operate on  $2 \times 2$  patches and are conditioned on both pose and temporal embeddings. We use AdamW (lr=8e $-5$ , betas=(0.9, 0.95), grad clip=10.0) and batch size 512. Action inputs are normalized to  $[-1, 1]$  for translation and  $[-\pi, \pi]$  for rotation. All experiments use Stable Diffusion VAE tokenizer and follow NWM’s hardware and evaluation setup. Metrics are averaged over 5 samples per sequence.

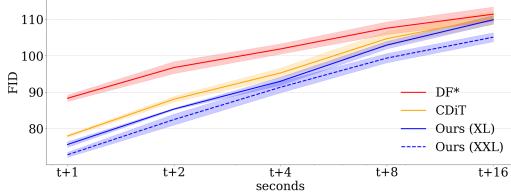
### 4.2 Comparison with Baselines

To comprehensively evaluate our model, we compare PEVA with (CDiT Bar et al. (2025) and Diffusion Forcing (Chen et al., 2024)) along two key dimensions. First, to assess whether the model faithfully simulates future observations conditioned on actions, we use LPIPS (Zhang et al., 2018) and DreamSim (Fu et al., 2023), which measure perceptual and semantic similarity to ground truth. Second, to evaluate the overall quality and realism of the generated samples, we report FID (Heusel

et al., 2017). As shown in Table 1, our model achieves better results on both action consistency and generative quality. Furthermore, Figure 3 shows that our models tend to maintain lower FID scores than the baselines as the prediction horizon increases, suggesting improved visual quality and temporal consistency over longer rollouts. Qualitative results for 16 second rollouts can be seen in Figure 1c and Figure 5. We implement Diffusion Forcing (DF\*) on top of PEVA by applying the diffusion forward process to the entire sequence of encoded latents, then predicting the next state given the previous (noisy) latents. At test time, we autoregressively predict the next state as in PEVA, without injecting noise into previously predicted frames, like Chen et al. (2024).

**Table 1: Baseline Perceptual Metrics.** Comparison of baselines on single-step prediction 2 seconds ahead.

Model	LPIPS ↓	DreamSim ↓	FID ↓
DF*	0.352 <sup>0.003</sup>	0.244 <sup>0.003</sup>	73.052 <sup>1.101</sup>
CDiT	0.313 <sup>0.001</sup>	0.202 <sup>0.002</sup>	63.714 <sup>0.491</sup>
PEVA	0.303 <sup>0.001</sup>	0.193 <sup>0.002</sup>	62.293 <sup>0.671</sup>



**Figure 3: Video Quality Across Time (FID).** Comparison of generation accuracy and quality as a function of time for up to 16 seconds. Qualitative results for 16 second rollouts can be seen in Figure 1c and Figure 5.

### 4.3 Atom Actions Control

To better evaluate PEVA’s ability to follow structured physical control, we decompose complex motions into atomic actions. By analyzing joint trajectories over short windows, we extract video segments exhibiting fundamental movements—such as hand motions (up, down, left, right) and whole-body actions (forward, rotate)—based on thresholded positional deltas. We sample 100 examples per action type to ensure balanced coverage, and evaluate single-step prediction 2 seconds ahead. Qualitative results are shown in Figure 1a and Figure 4, and quantitative results in Table 2.

**Table 2: Atomic Action Performance.** Comparison of models in generating videos of atomic actions.

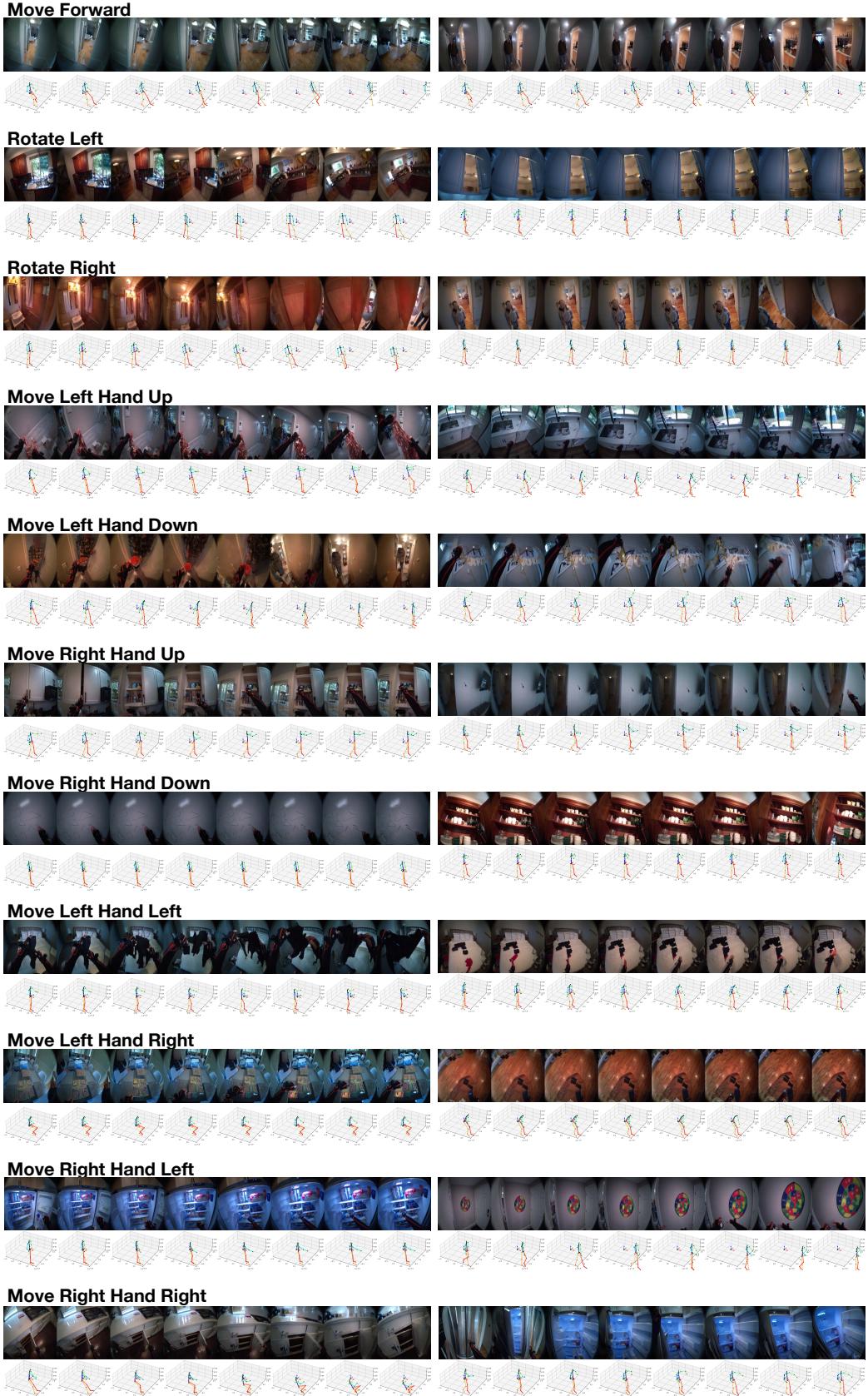
Model	Navigation			Left Hand				Right Hand			
	Forward	Rot.L	Rot.R	Left	Right	Up	Down	Left	Right	Up	Down
DF*	0.393 <sup>0.011</sup>	0.314 <sup>0.006</sup>	0.279 <sup>0.005</sup>	0.292 <sup>0.009</sup>	0.306 <sup>0.005</sup>	0.332 <sup>0.008</sup>	0.323 <sup>0.006</sup>	0.304 <sup>0.006</sup>	0.315 <sup>0.007</sup>	0.305 <sup>0.005</sup>	0.296 <sup>0.008</sup>
CDiT	0.348 <sup>0.004</sup>	0.284 <sup>0.003</sup>	0.249 <sup>0.004</sup>	0.258 <sup>0.005</sup>	0.265 <sup>0.009</sup>	0.279 <sup>0.008</sup>	0.267 <sup>0.004</sup>	0.286 <sup>0.007</sup>	0.273 <sup>0.004</sup>	0.277 <sup>0.004</sup>	0.268 <sup>0.002</sup>
Ours (XL)	0.337 <sup>0.006</sup>	0.277 <sup>0.006</sup>	0.242 <sup>0.007</sup>	0.244 <sup>0.005</sup>	0.257 <sup>0.004</sup>	0.272 <sup>0.008</sup>	0.263 <sup>0.003</sup>	0.271 <sup>0.005</sup>	0.267 <sup>0.003</sup>	0.268 <sup>0.004</sup>	0.256 <sup>0.009</sup>
Ours (XXL)	0.325 <sup>0.006</sup>	0.269 <sup>0.005</sup>	0.234 <sup>0.004</sup>	0.236 <sup>0.003</sup>	0.241 <sup>0.003</sup>	0.251 <sup>0.004</sup>	0.247 <sup>0.005</sup>	0.256 <sup>0.007</sup>	0.254 <sup>0.005</sup>	0.252 <sup>0.004</sup>	0.245 <sup>0.005</sup>

### 4.4 Ablation Studies

We conduct ablation studies to assess the impact of key design choices in PEVA, summarized in Table 3. First, increasing the context window from 3 to 15 frames consistently improves performance across all metrics, highlighting the importance of temporal context for egocentric prediction. Second, model scale plays a significant role: larger variants from PEVA-S to PEVA-XXL show steady gains in perceptual and semantic fidelity. Lastly, we compare two action embedding strategies—MLP-based encoding versus simple concatenation—and find that the latter performs competitively despite its simplicity, suggesting that our structured action representation already captures sufficient motion information. The gray-highlighted row denotes the default configuration in main experiments.

### 4.5 Long-Term Prediction Quality

We evaluate the model’s ability to maintain visual and semantic consistency over extended prediction horizons. As shown in Figure 5, PEVA generates coherent 16-second rollouts conditioned on full-body motion. Table 3 reports DreamSim scores at increasing time steps, showing a gradual degradation from 0.178 (1s) to 0.390 (16s), indicating that predictions remain semantically plausible even far into the future.



**Figure 4: Atom Actions Generation.** We include video generation examples of different atomic actions specified by 3D-body poses.

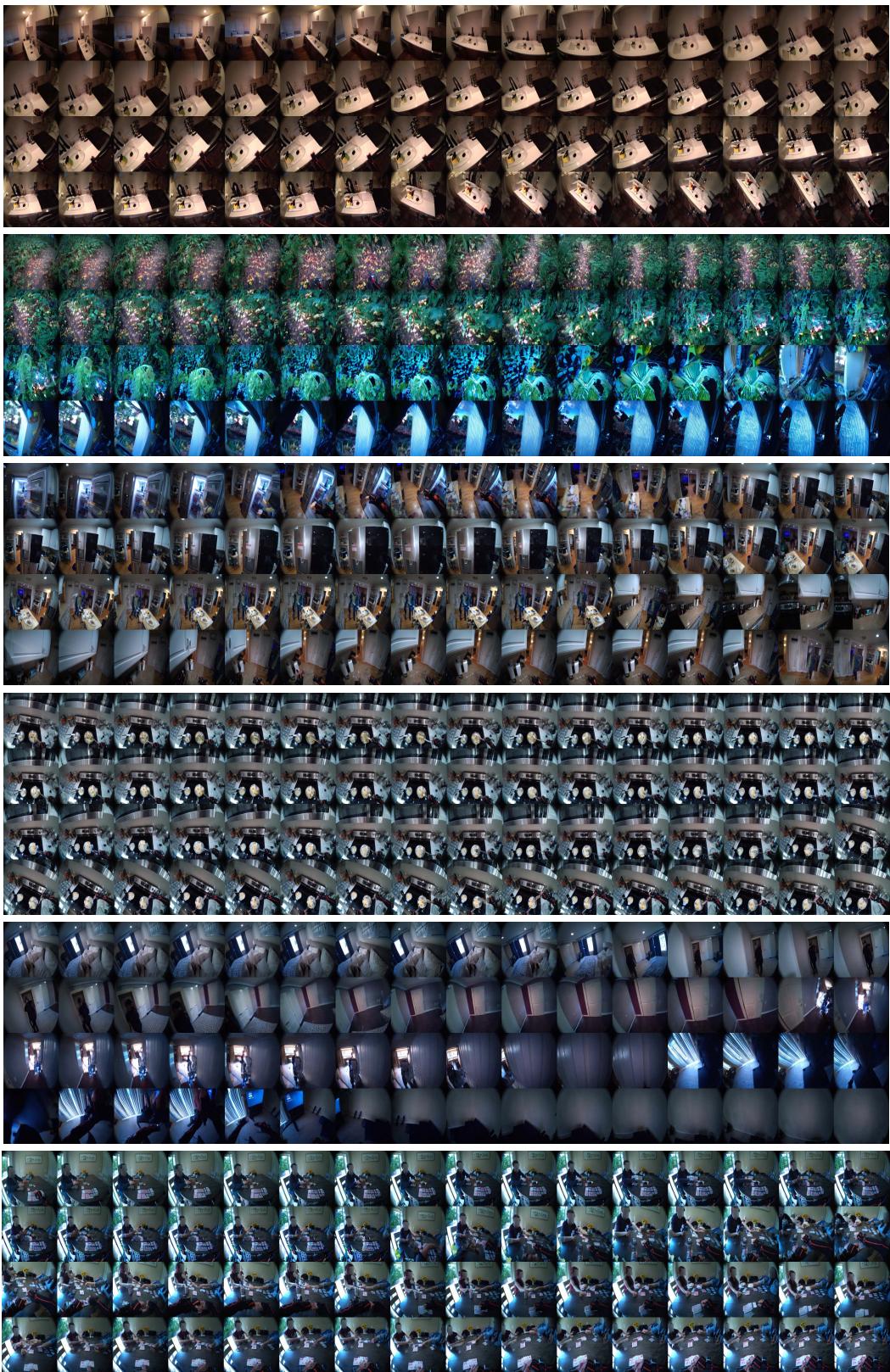


Figure 5: **Generation Over Long-Horizons.** We include 16-second video generation examples.



**Figure 6: Planning with Counterfactuals.** We demonstrate a planning example by simulating multiple action candidates using PEVA and scoring them based on their perceptual similarity to the goal, as measured by LPIPS (Zhang et al., 2018). In the first case, we show that PEVA enables us to rule out action sequences that leads us to the sink in the top row, and outdoors in the second row. In the second case we show PEVA allows us to find a reasonable sequence of actions to open the refrigerator in the third row. PEVA enables us to rule out action sequences that grab the nearby plants and go to the kitchen and mix ingredients. PEVA allows us to choose the most correct action sequences that grab the box from the shelf.

Table 3: **Model Ablations.** We evaluate the impact of different context lengths, action embedding methods, and model sizes on single-step prediction performance (2 seconds into the future).

Configuration	Metrics			
	LPIPS ↓	DreamSim ↓	PSNR ↑	FID ↓
<i>Context Length</i>				
3 frames	0.304 <sup>0.002</sup>	0.199 <sup>0.003</sup>	16.469 <sup>0.044</sup>	63.966 <sup>0.421</sup>
7 frames	0.304 <sup>0.001</sup>	0.195 <sup>0.002</sup>	16.443 <sup>0.068</sup>	62.540 <sup>0.314</sup>
15 frames	0.303 <sup>0.001</sup>	0.193 <sup>0.002</sup>	16.511 <sup>0.061</sup>	62.293 <sup>0.671</sup>
<i>Action Representation</i>				
Action Embedding ( $d = 512$ )	0.317 <sup>0.003</sup>	0.202 <sup>0.002</sup>	16.195 <sup>0.081</sup>	63.101 <sup>0.341</sup>
Action Concatenation	0.303 <sup>0.001</sup>	0.193 <sup>0.002</sup>	16.511 <sup>0.061</sup>	62.293 <sup>0.671</sup>
<i>Model Size</i>				
PEVA-S	0.370 <sup>0.002</sup>	0.327 <sup>0.002</sup>	15.743 <sup>0.060</sup>	101.38 <sup>0.450</sup>
PEVA-B	0.337 <sup>0.001</sup>	0.246 <sup>0.002</sup>	16.013 <sup>0.091</sup>	74.338 <sup>1.057</sup>
PEVA-L	0.308 <sup>0.002</sup>	0.202 <sup>0.001</sup>	16.417 <sup>0.037</sup>	64.402 <sup>0.496</sup>
PEVA-XL	0.303 <sup>0.001</sup>	0.193 <sup>0.002</sup>	16.511 <sup>0.061</sup>	62.293 <sup>0.671</sup>
PEVA-XXL	0.298 <sup>0.002</sup>	0.186 <sup>0.003</sup>	16.556 <sup>0.060</sup>	61.100 <sup>0.517</sup>

#### 4.6 Planning with Multiple Action Candidates.

We demonstrate a sample in which PEVA enables planning with multiple action candidates in Figure 1b and Figure 6. We start by sampling multiple action candidates and simulate each action candidate using PEVA via autoregressive rollout. Finally, we rank each action candidate’s final prediction by measuring LPIPS similarity with the goal image. We find that PEVA is effective in enabling planning through simulating action candidates.

### 5 Failure Cases, Limitations and Future Directions

While our model demonstrates promising results in predicting egocentric video from whole-body motion, several limitations remain that suggest directions for future work. First, our planning evaluation is preliminary—we only explore a simulation-based selection over candidate actions for only the left or right arm. While this provides an early indication that the model can anticipate visual consequences of body movement, it does not yet support long-horizon planning or full trajectory optimization. Extending PEVA to closed-loop control or interactive environments is a key next step. Second, the model currently lacks explicit conditioning on task intent or semantic goals. Our evaluation uses image similarity as a proxy objective. Future work could explore combining PEVA with high-level goal conditioning or integrating object-centric representations.

#### 5.1 Some planning attempts with PEVA

Here we describe how to use a trained PEVA to plan action sequences to achieve a visual target. We formulate planning as an energy minimization problem and perform standalone planning in the same way as NWM (Bar et al., 2025) using the Cross-Entropy Method (CEM) (Rubinstein, 1997) besides minor modifications in the representation and initialization of the action.

For simplicity, we conduct two experiments where we only predict moving either the left or right arm controlled by predicting the relative joint rotations represented as euler angles. For each respective arm we control only the shoulder, upper arm, forearm, and hand leaving our actions space as  $4 \times 3 = 12$  where we have  $(\Delta\phi_{\text{shoulder}}, \Delta\theta_{\text{shoulder}}, \Delta\psi_{\text{shoulder}}, \dots, \Delta\phi_{\text{forearm}}, \Delta\theta_{\text{forearm}}, \Delta\psi_{\text{forearm}})$ . We initialize mean  $(\mu_{\Delta\phi_{\text{shoulder}}}, \mu_{\Delta\theta_{\text{shoulder}}}, \mu_{\Delta\psi_{\text{shoulder}}}, \dots, \mu_{\Delta\phi_{\text{forearm}}}, \mu_{\Delta\theta_{\text{forearm}}}, \mu_{\Delta\psi_{\text{forearm}}})$  and variance  $(\sigma^2_{\Delta\phi_{\text{shoulder}}}, \sigma^2_{\Delta\theta_{\text{shoulder}}}, \sigma^2_{\Delta\psi_{\text{shoulder}}}, \dots, \sigma^2_{\Delta\phi_{\text{forearm}}}, \sigma^2_{\Delta\theta_{\text{forearm}}}, \sigma^2_{\Delta\psi_{\text{forearm}}})$  as the mean and variance of the next action across the training dataset for these segments.

We assume the action is a straight continuous motion. Thus we repeat this action for our sequence length, in our case  $T = 8$  and optimize the delta actions. The time interval between steps is fixed at  $k = 0.25$  seconds. All other hyperparameters remain the same as in NWM (Bar et al., 2025).

Table 4: Mean and Variance of relative rotation as euler angles ( $\phi, \theta, \psi$ ) for arm segments computed across the training dataset.

Segment	Statistic	Right Arm	Left Arm
Shoulder	Mean	(0.0027, -0.0012, -0.0015)	(0.0624, 0.0687, 0.1494)
	Variance	(0.0010, -0.0006, 0.0003)	(0.0625, 0.0697, 0.1496)
Upper Arm	Mean	(0.0107, -0.0011, -0.0020)	(0.1119, 0.1647, 0.1791)
	Variance	(-0.0062, -0.0004, -0.0013)	(0.0991, 0.1593, 0.1611)
Forearm	Mean	(0.0068, -0.0035, 0.0077)	(0.1937, 0.2107, 0.2261)
	Variance	(-0.0036, -0.0063, 0.0002)	(0.1791, 0.2012, 0.2186)
Hand	Mean	(0.0065, 0.0001, 0.004, )	(0.2417, 0.229, 0.2631)
	Variance	(-0.0024, -0.0032, -0.0001)	(0.2126, 0.2237, 0.2475)

## 5.2 Qualitative Results

Due to time constraints, we focus our investigation on arm movements—arguably the most complex among body actions. While this remains an open problem, we present preliminary results using PEVA with CEM for standalone planning. This setting simplifies the high-dimensional control space while still capturing key challenges of full-body coordination.

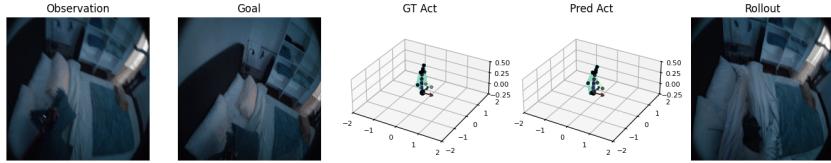


Figure 7: In this case, we are able to predict a sequence of actions that pulls our left arm in, similar to the goal.

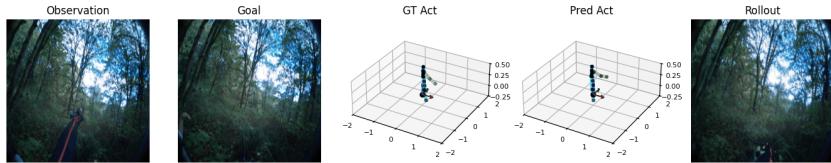


Figure 8: In this case, we are able to predict a sequence of actions that lowers our left arm, but not the same amount as the groundtruth sequence as we can see in the pose and hand at the bottom of our rollout.

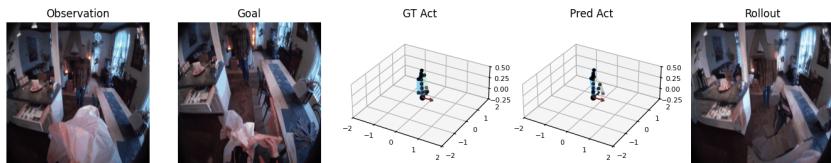


Figure 9: In this case, we are able to predict a sequence of actions that lowers our left arm that lowers the tissue. However, the goal image still has the tissue visible.

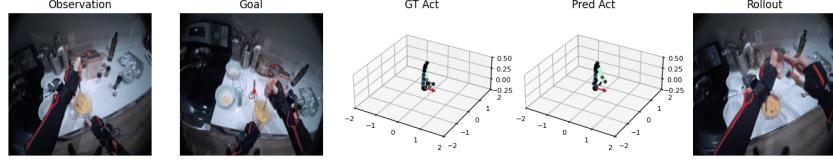


Figure 10: In this case, we are able to predict a sequence of actions that raises our right arm to the mixing stick. We see a limitation with our method as we only predict the right arm so we do not predict to move the left arm down accordingly.

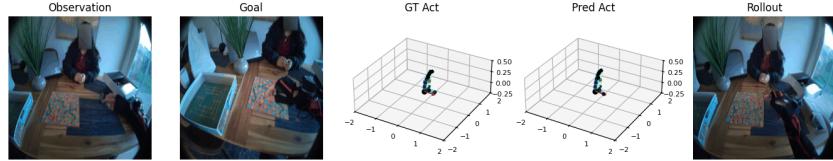


Figure 11: In this case, we are able to predict a sequence of actions that moves our right arm toward the left but not quite enough. We see a limitation with our method as we only predict the right arm so we do not predict any necessary additional body rotations.

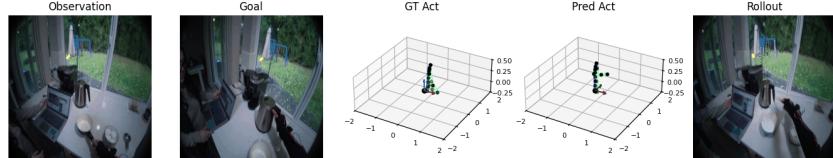


Figure 12: In this case, we are able to predict a sequence of actions that reaches toward the kettle but does not quite grab it as in the goal.

## 6 Conclusion

We introduced PEVA, a model that predicts egocentric video conditioned on detailed 3D human motion. Unlike prior work that uses low-dimensional or abstract control, PEVA leverages full-body pose sequences to simulate realistic and controllable visual outcomes. Built on a conditional diffusion transformer and trained on Nymeria, it captures the link between physical movement and first-person perception. Experiments show that PEVA improves prediction quality, semantic consistency, and fine-grained control over strong baselines. Our hierarchical evaluation highlights the value of whole-body conditioning across short-term, long-horizon, and atomic action tasks. While our planning results are preliminary, they demonstrate the potential for simulating action consequences in embodied settings. We hope this work moves toward more grounded models of perception and action for physically embodied intelligence.

## Acknowledgment

The authors thank Rithwik Nukala for his help in annotating atomic actions. We thank Katerina Fragiadaki, Philipp Krähenbühl, Bharath Hariharan, Guanya Shi, Shubham Tsunami and Deva Ramanan for the useful suggestions and feedbacks for improving the paper; Jianbo Shi for the discussion regarding control theory; Yilun Du for the support on Diffusion Forcing; Brent Yi for his help in human motion related works and Alexei Efros for the discussion and debates regarding world models. This work is partially supported by the ONR MURI N00014-21-1-2801.

## References

- Jake K Aggarwal and Quin Cai. Human motion analysis: A review. *Computer vision and image understanding*, 73(3):428–440, 1999.
- Pulkit Agrawal, Ashvin V Nair, Pieter Abbeel, Jitendra Malik, and Sergey Levine. Learning to poke by poking: Experiential learning of intuitive physics. *Advances in neural information processing systems*, 29, 2016.
- Eloi Alonso et al. Diamond: Diffusion as a model of environment dreams. *arXiv preprint arXiv:2401.02644*, 2024.
- Tenglong Ao, Zeyi Zhang, and Libin Liu. Gesturediffuclip: Gesture diffusion model with clip latents. *ACM Transactions on Graphics (TOG)*, 42(4):1–18, 2023.
- Mido Assran, Adrien Bardes, David Fan, Quentin Garrido, Russell Howes, Matthew Muckley, Ammar Rizvi, Claire Roberts, Koustuv Sinha, Artem Zholus, et al. V-jepa 2: Self-supervised video models enable understanding, prediction and planning. *arXiv preprint arXiv:2506.09985*, 2025.
- Amir Bar, Arya Bakhtiar, Danny Tran, Antonio Loquercio, Jathushan Rajasegaran, Yann LeCun, Amir Globerson, and Trevor Darrell. Egopet: Egomotion and interaction data from an animal’s perspective. In *European Conference on Computer Vision*, pages 377–394. Springer, 2024.
- Amir Bar, Gaoyue Zhou, Danny Tran, Trevor Darrell, and Yann LeCun. Navigation world models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 15791–15801, 2025.
- Angela Castillo, Maria Escobar, Guillaume Jeanneret, Albert Pumarola, Pablo Arbeláez, Ali Thabet, and Artsiom Sanakoyeu. Bodiffusion: Diffusing sparse observations for full-body human motion synthesis. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4221–4231, 2023.
- Boyuan Chen, Diego Martí Monsó, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 37:24081–24125, 2024.
- Xuxin Cheng, Yandong Ji, Junming Chen, Ruihan Yang, Ge Yang, and Xiaolong Wang. Expressive whole-body control for humanoid robots. *arXiv preprint arXiv:2402.16796*, 2024.
- Kenneth JW Craik. *The Nature of Explanation*. Cambridge University Press, 1943.
- Rishabh Dabral, Muhammad Hamza Mughal, Vladislav Golyanik, and Christian Theobalt. Mofusion: A framework for denoising-diffusion-based motion synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9760–9770, 2023.
- Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pages 720–736, 2018.
- Frederik Ebert, Chelsea Finn, Sudeep Dasari, Annie Xie, Alex Lee, and Sergey Levine. Visual foresight: Model-based deep reinforcement learning for vision-based robotic control. *arXiv preprint arXiv:1812.00568*, 2018.
- Chelsea Finn and Sergey Levine. Deep visual foresight for planning robot motion. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*, pages 2786–2793. IEEE, 2017.
- Katerina Fragkiadaki, Pulkit Agrawal, Sergey Levine, and Jitendra Malik. Learning visual predictive models of physics for playing billiards. *arXiv preprint arXiv:1511.07404*, 2015.
- Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.
- Antonino Furnari and Giovanni Maria Farinella. Rolling-unrolling lstms for action anticipation from first-person video. *IEEE transactions on pattern analysis and machine intelligence*, 43(11):4021–4036, 2020.
- Rohit Girdhar and Kristen Grauman. Anticipative video transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 13505–13515, 2021.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18995–19012, 2022.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. 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*, pages 19383–19400, 2024.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5152–5161, 2022.

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

Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by latent imagination. In *International Conference on Learning Representations*.

Danijar Hafner, Jurgis Pasukonis, Jimmy Ba, and Timothy Lillicrap. Mastering diverse domains through world models. *arXiv preprint arXiv:2301.04104*, 2023.

Nicklas Hansen, Jyothir SV, Vlad Sobal, Yann LeCun, Xiaolong Wang, and Hao Su. Hierarchical world models as visual whole-body humanoid controllers. *arXiv preprint arXiv:2405.18418*, 2024.

Tairan He, Zhengyi Luo, Xialin He, Wenli Xiao, Chong Zhang, Weinan Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Omnih2o: Universal and dexterous human-to-humanoid whole-body teleoperation and learning. *arXiv preprint arXiv:2406.08858*, 2024a.

Tairan He, Zhengyi Luo, Wenli Xiao, Chong Zhang, Kris Kitani, Changliu Liu, and Guanya Shi. Learning human-to-humanoid real-time whole-body teleoperation. In *2024 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 8944–8951. IEEE, 2024b.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 6840–6851, 2020.

Fangzhou Hong, Vladimir Guzov, Hyo Jin Kim, Yuting Ye, Richard Newcombe, Ziwei Liu, and Lingni Ma. Egolm: Multi-modal language model of egocentric motions. *arXiv preprint arXiv:2409.18127*, 2024.

Li Hu, Xin Gao, Peng Zhang, Ke Sun, Bang Zhang, and Liefeng Bo. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. *arXiv preprint arXiv:2311.17117*, 2023.

Michael I Jordan. Computational aspects of motor control and motor learning. *Handbook of perception and action*, 2:71–120, 1996.

Rudolph Emil Kalman. A new approach to linear filtering and prediction problems. 1960.

Mitsuo Kawato. Internal models for motor control and trajectory planning. *Current opinion in neurobiology*, 9(6):718–727, 1999.

Mitsuo Kawato, Kazunori Furukawa, and Ryoji Suzuki. A hierarchical neural-network model for control and learning of voluntary movement. *Biological cybernetics*, 57:169–185, 1987.

Yann LeCun. A path towards autonomous machine intelligence version 0.9. 2, 2022-06-27. *Open Review*, 62(1):1–62, 2022.

Jiaman Li, Karen Liu, and Jiajun Wu. Ego-body pose estimation via ego-head pose estimation. In *CVPR*, pages 17142–17151, 2023.

Yunzhu Li, Shuang Li, Vincent Sitzmann, Pulkit Agrawal, and Antonio Torralba. 3d neural scene representations for visuomotor control. In *Conference on Robot Learning*, pages 112–123. PMLR, 2022.

Huihan Liu, Yu Zhang, Vaarij Betala, Evan Zhang, James Liu, Crystal Ding, and Yuke Zhu. Multi-task interactive robot fleet learning with visual world models. *arXiv preprint arXiv:2410.22689*, 2024.

- Qianli Ma et al. Nymeria: A massive collection of multimodal egocentric daily motion in the wild. *arXiv preprint arXiv:2406.09905*, 2024.
- Yecheng Jason Ma, Shagun Sodhani, Dinesh Jayaraman, Osbert Bastani, Vikash Kumar, and Amy Zhang. Vip: Towards universal visual reward and representation via value-implicit pre-training. *arXiv preprint arXiv:2210.00030*, 2022.
- Movella. *MVN User Manual*. Movella, 2021. [https://www.movella.com/hubfs/MVN\\_User\\_Manual.pdf](https://www.movella.com/hubfs/MVN_User_Manual.pdf).
- Evronne Ng, Javier Romero, Timur Bagautdinov, Shaojie Bai, Trevor Darrell, Angjoo Kanazawa, and Alexander Richard. From audio to photoreal embodiment: Synthesizing humans in conversations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1001–1010, 2024.
- Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.
- J Björck Nvidia, F Castaneda, N Cherniadev, X Da, R Ding, L Fan, Y Fang, D Fox, F Hu, S Huang, et al. Gr00t n1: An open foundation model for generalist humanoid robots. *arXiv preprint arXiv:2503.14734*, 2025.
- William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4195–4205, 2023.
- Mathis Petrovich, Michael J Black, and Gü̈l Varol. Action-conditioned 3d human motion synthesis with transformer vae. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10985–10995, 2021.
- Ilija Radosavovic, Sarthak Kamat, Trevor Darrell, and Jitendra Malik. Learning humanoid locomotion over challenging terrain. *arXiv preprint arXiv:2410.03654*, 2024.
- Davis Rempe, Tolga Birdal, Aaron Hertzmann, Jimei Yang, Srinath Sridhar, and Leonidas J Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11488–11499, 2021.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022.
- Bodo Rosenhahn, Reinhard Klette, and Dimitris Metaxas. Human motion. *Understanding, Modeling, Capture*, 2008.
- Reuven Y Rubinstein. Optimization of computer simulation models with rare events. *European Journal of Operational Research*, 99(1):89–112, 1997.
- 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.
- Ajay Sridhar, Dhruv Shah, Catherine Glossop, and Sergey Levine. Nomad: Goal masked diffusion policies for navigation and exploration. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pages 63–70. IEEE, 2024.
- Guy Tevet, Sigal Raab, Brian Gordon, Yonatan Shafir, Daniel Cohen-Or, and Amit H Bermano. Human motion diffusion model. *arXiv preprint arXiv:2209.14916*, 2022.
- Hermann Von Helmholtz. *Helmholtz's treatise on physiological optics*. Optical Society of America, 1925.
- Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023.
- Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. *arXiv preprint arXiv:2311.16498*, 2023.
- Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 1(2):6, 2023.

- Sherry Yang, Jacob Walker, Jack Parker-Holder, Yilun Du, Jake Bruce, Andre Barreto, Pieter Abbeel, and Dale Schuurmans. Video as the new language for real-world decision making. *arXiv preprint arXiv:2402.17139*, 2024.
- Vickie Ye, Georgios Pavlakos, Jitendra Malik, and Angjoo Kanazawa. Decoupling human and camera motion from videos in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21222–21232, 2023.
- Brent Yi, Vickie Ye, Maya Zheng, Yunqi Li, Lea Müller, Georgios Pavlakos, Yi Ma, Jitendra Malik, and Angjoo Kanazawa. Estimating body and hand motion in an ego-sensed world. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 7072–7084, 2025.
- Ye Yuan and Kris Kitani. Ego-pose estimation and forecasting as real-time pd control. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10082–10092, 2019.
- Ye Yuan, Jiaming Song, Umar Iqbal, Arash Vahdat, and Jan Kautz. Physdiff: Physics-guided human motion diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 10082–10092, 2023.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023.
- Mingyuan Zhang, Zhongang Cai, Liang Pan, Fangzhou Hong, Xinying Guo, Lei Yang, and Ziwei Liu. Motion-diffuse: Text-driven human motion generation with diffusion model. *IEEE transactions on pattern analysis and machine intelligence*, 46(6):4115–4128, 2024.
- Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.
- Siwei Zhang, Qianli Ma, Yan Zhang, Zhiyin Qian, Taein Kwon, Marc Pollefeys, Federica Bogo, and Siyu Tang. Egobody: Human body shape and motion of interacting people from head-mounted devices. *arXiv preprint arXiv:2204.06953*, 2022.
- Gaoyue Zhou, Hengkai Pan, Yann LeCun, and Lerrel Pinto. Dino-wm: World models on pre-trained visual features enable zero-shot planning. *arXiv preprint arXiv:2411.04983*, 2024.

## More Qualitative Results

In the main paper, we provide three types of visualization: PEVA can simulate counterfactuals, generate videos of atomic actions, and long video generation.

Here, we show more qualitative results following the settings in main paper:



Figure 13: **Generation Over Long-Horizons**. We include 16-second video generation examples.

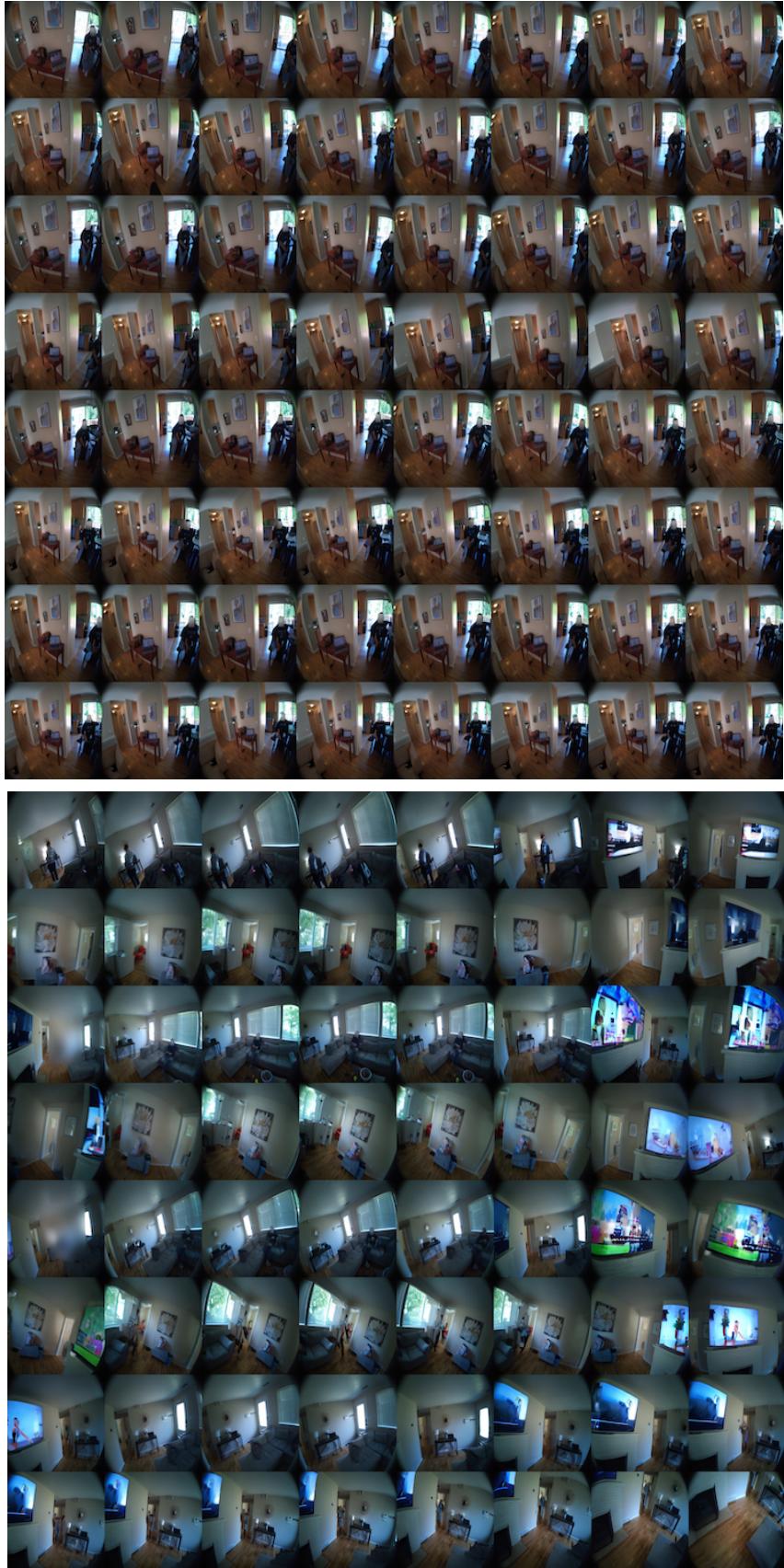


Figure 14: **Generation Over Long-Horizons**. We include 16-second video generation examples.

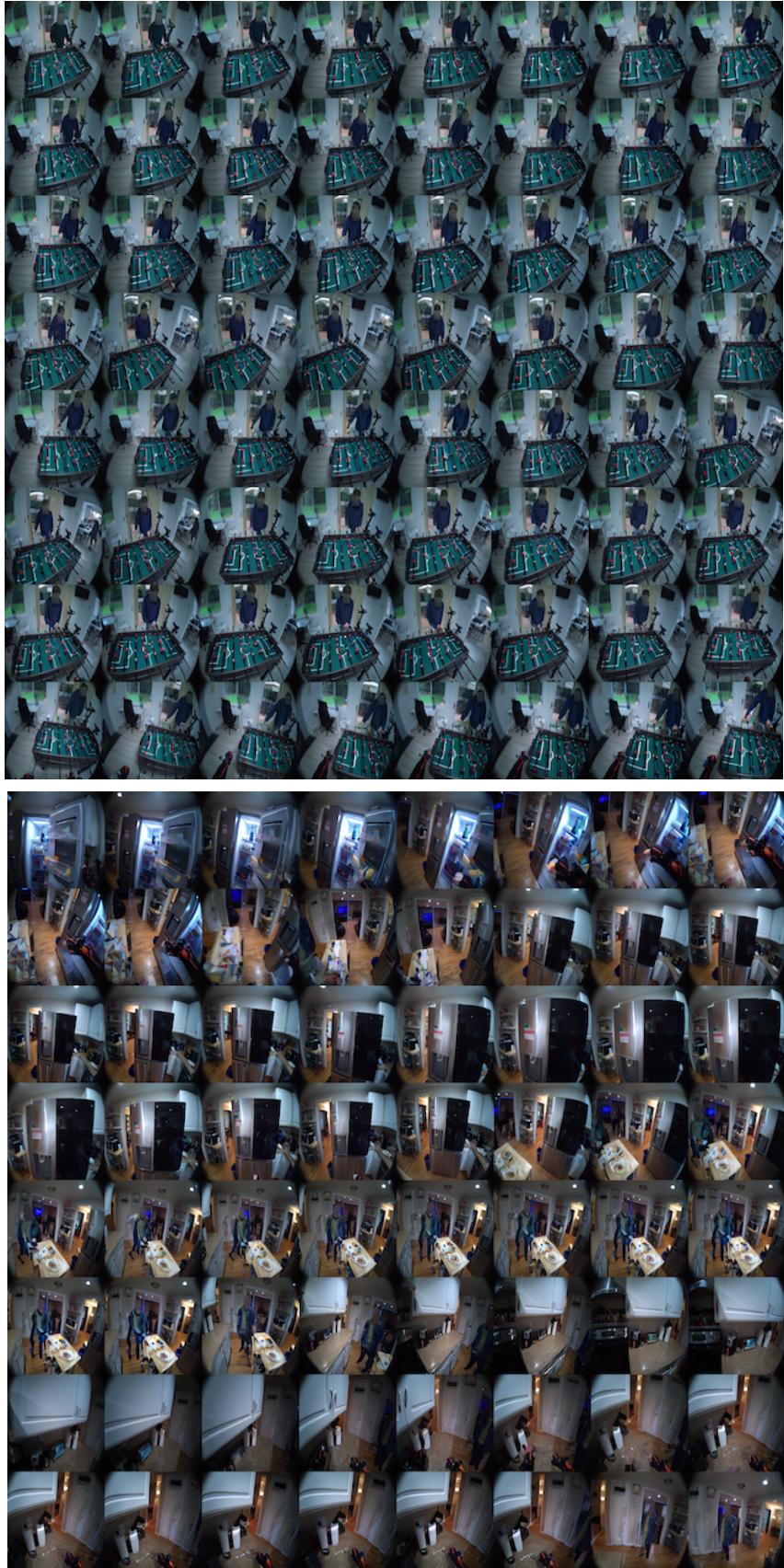


Figure 15: **Generation Over Long-Horizons**. We include 16-second video generation examples.

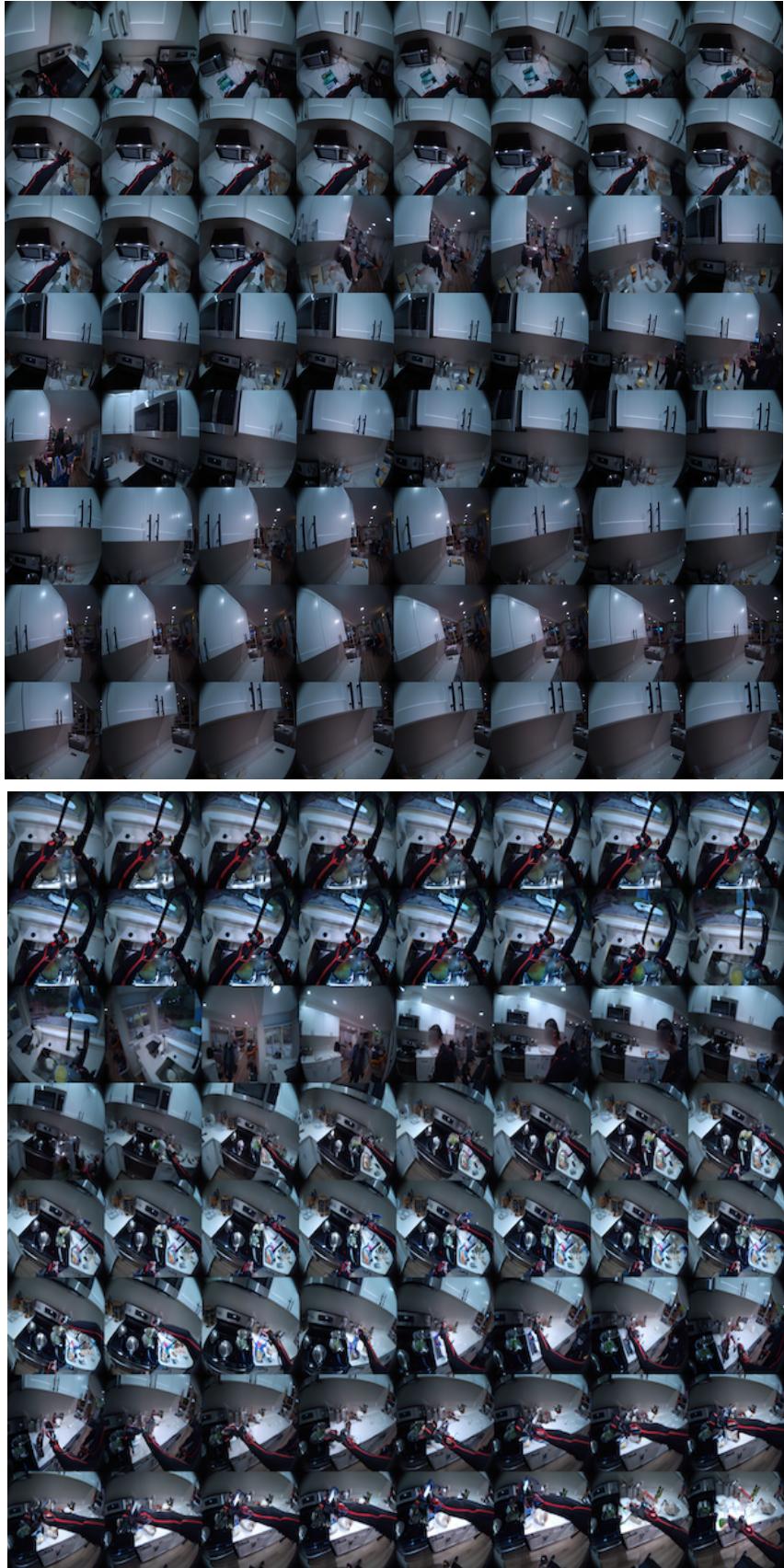


Figure 16: **Generation Over Long-Horizons.** We include 16-second video generation examples.

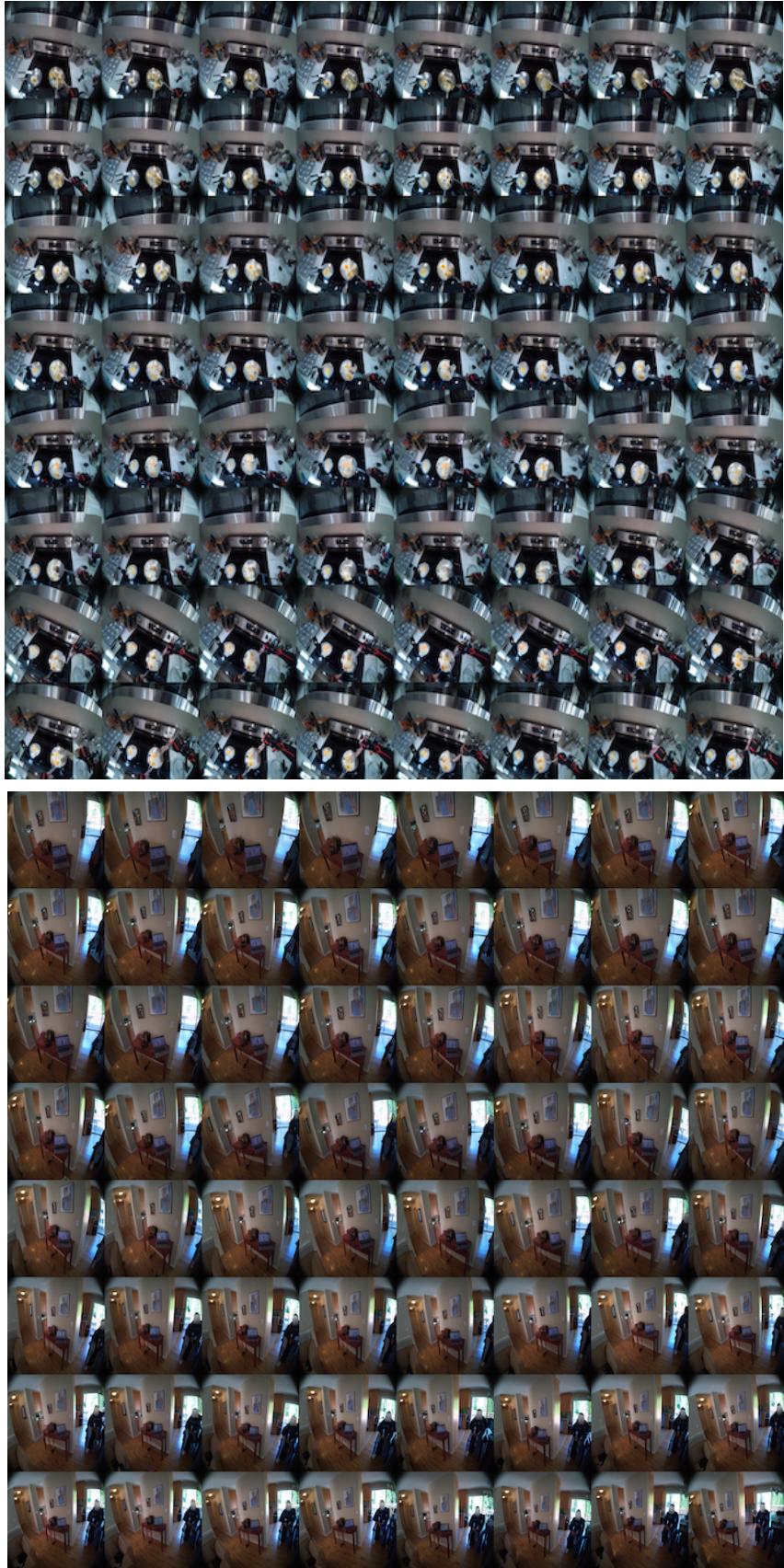


Figure 17: **Generation Over Long-Horizons.** We include 16-second video generation examples.

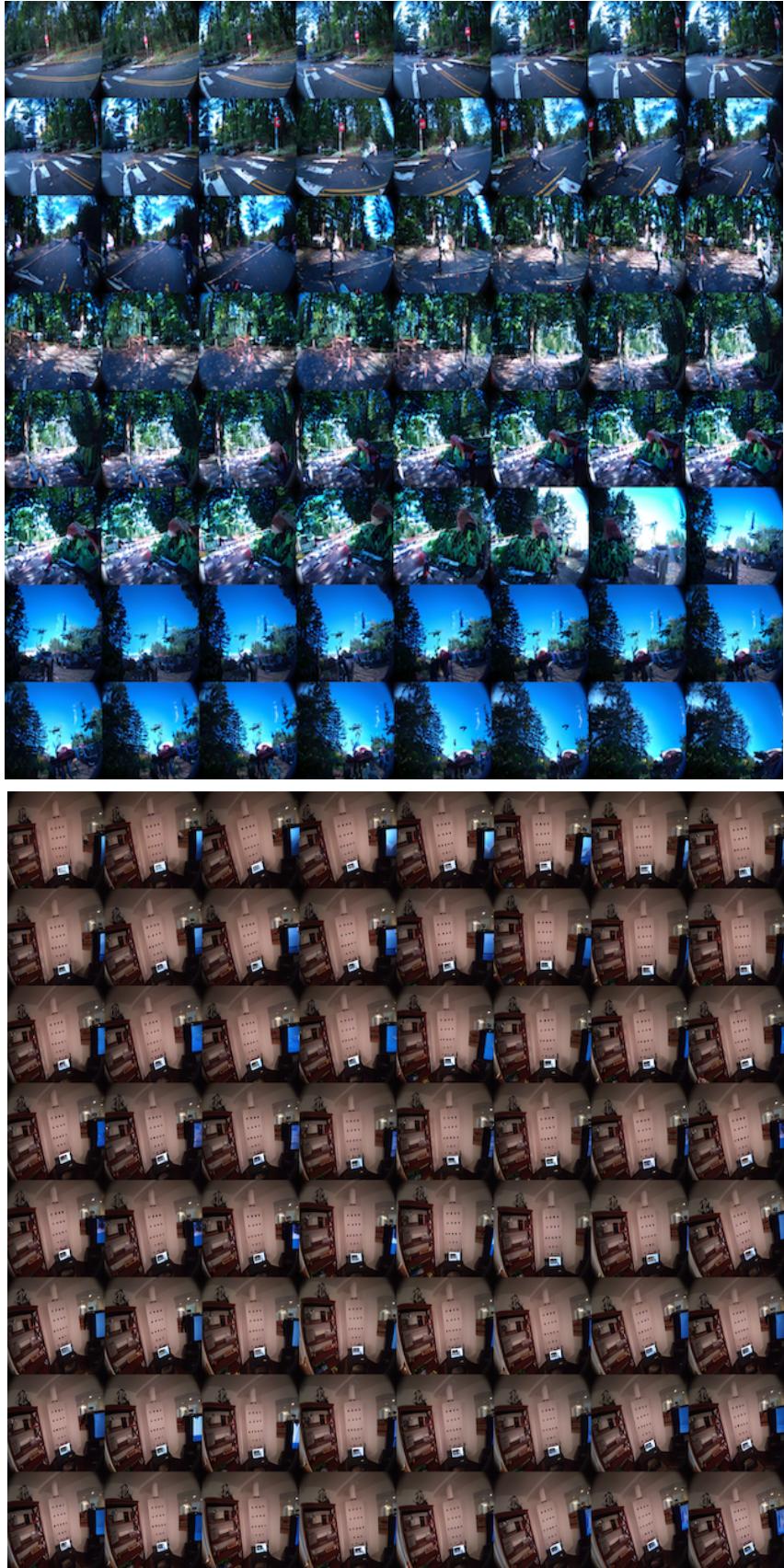


Figure 18: **Generation Over Long-Horizons**. We include 16-second video generation examples.

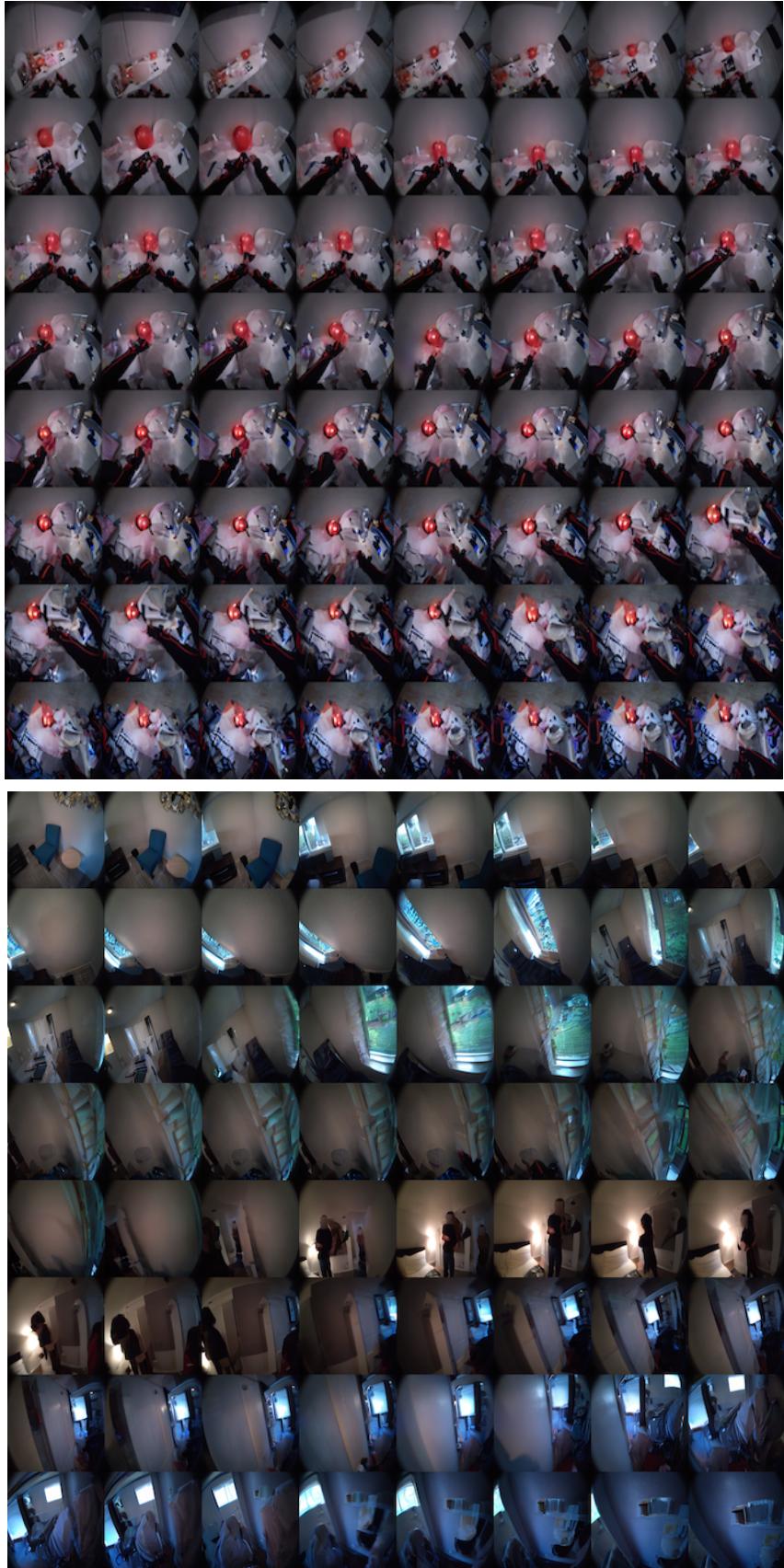


Figure 19: **Generation Over Long-Horizons**. We include 16-second video generation examples.

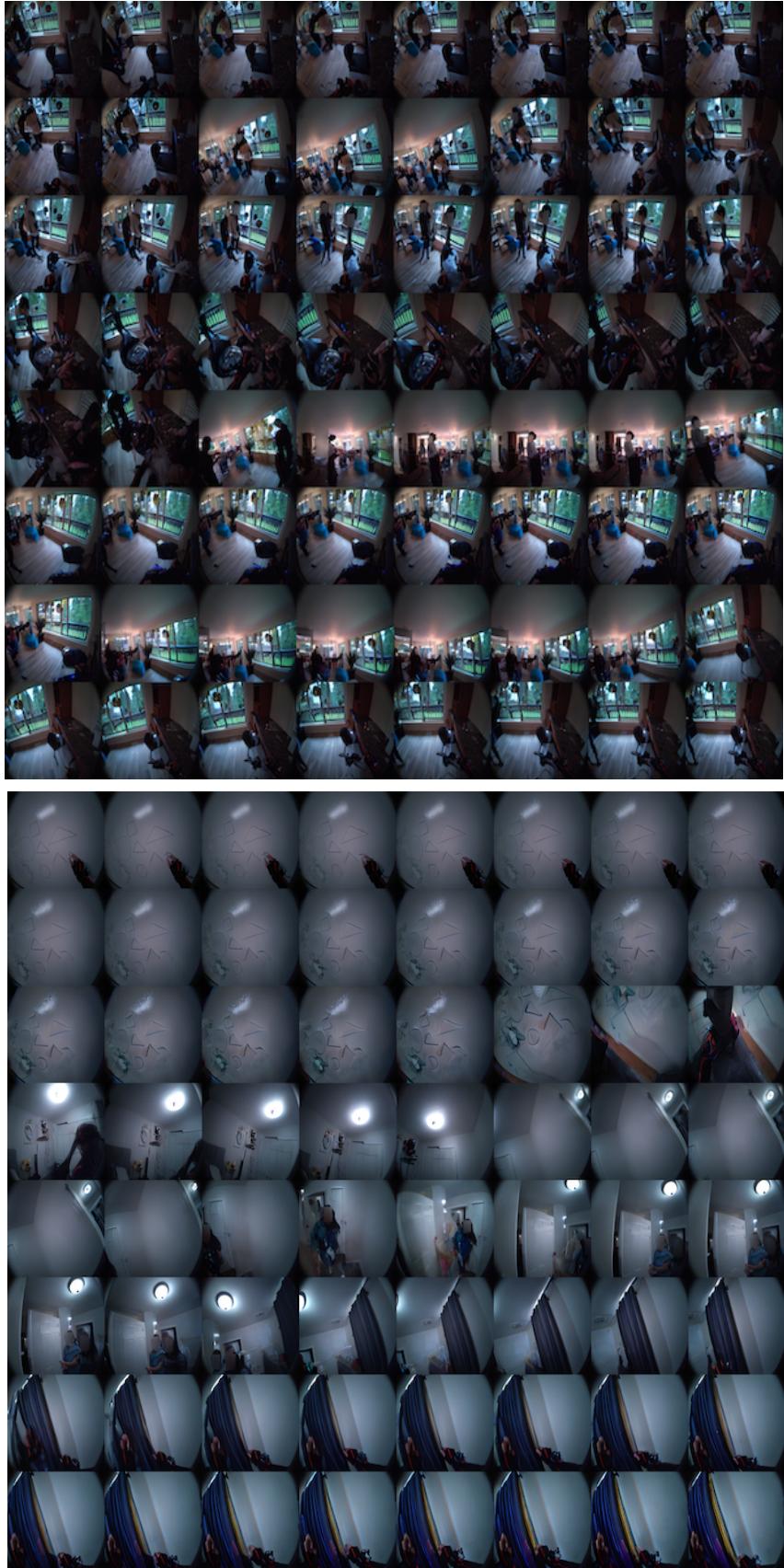


Figure 20: **Generation Over Long-Horizons**. We include 16-second video generation examples.

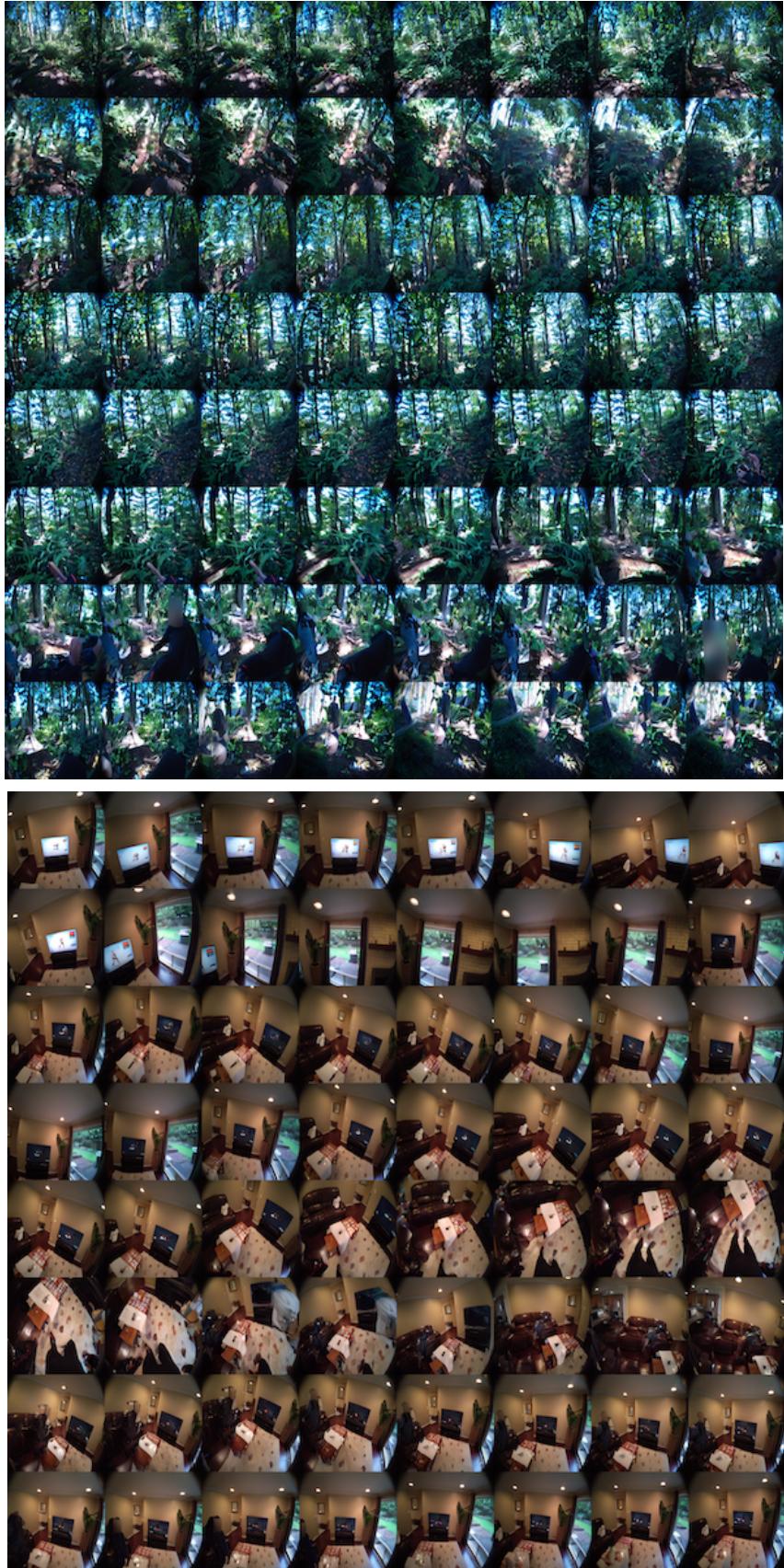


Figure 21: **Generation Over Long-Horizons**. We include 16-second video generation examples.

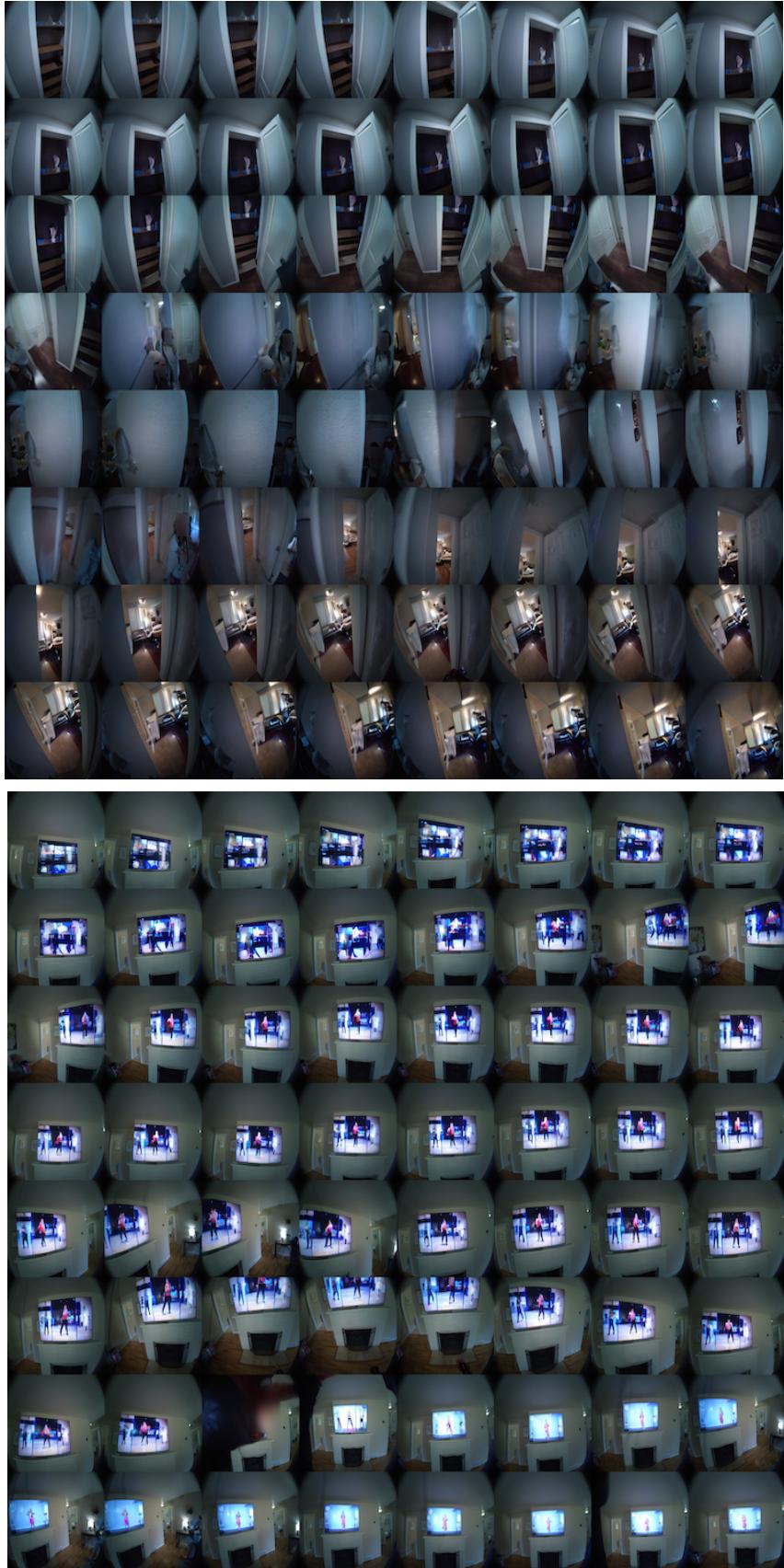


Figure 22: **Generation Over Long-Horizons**. We include 16-second video generation examples.

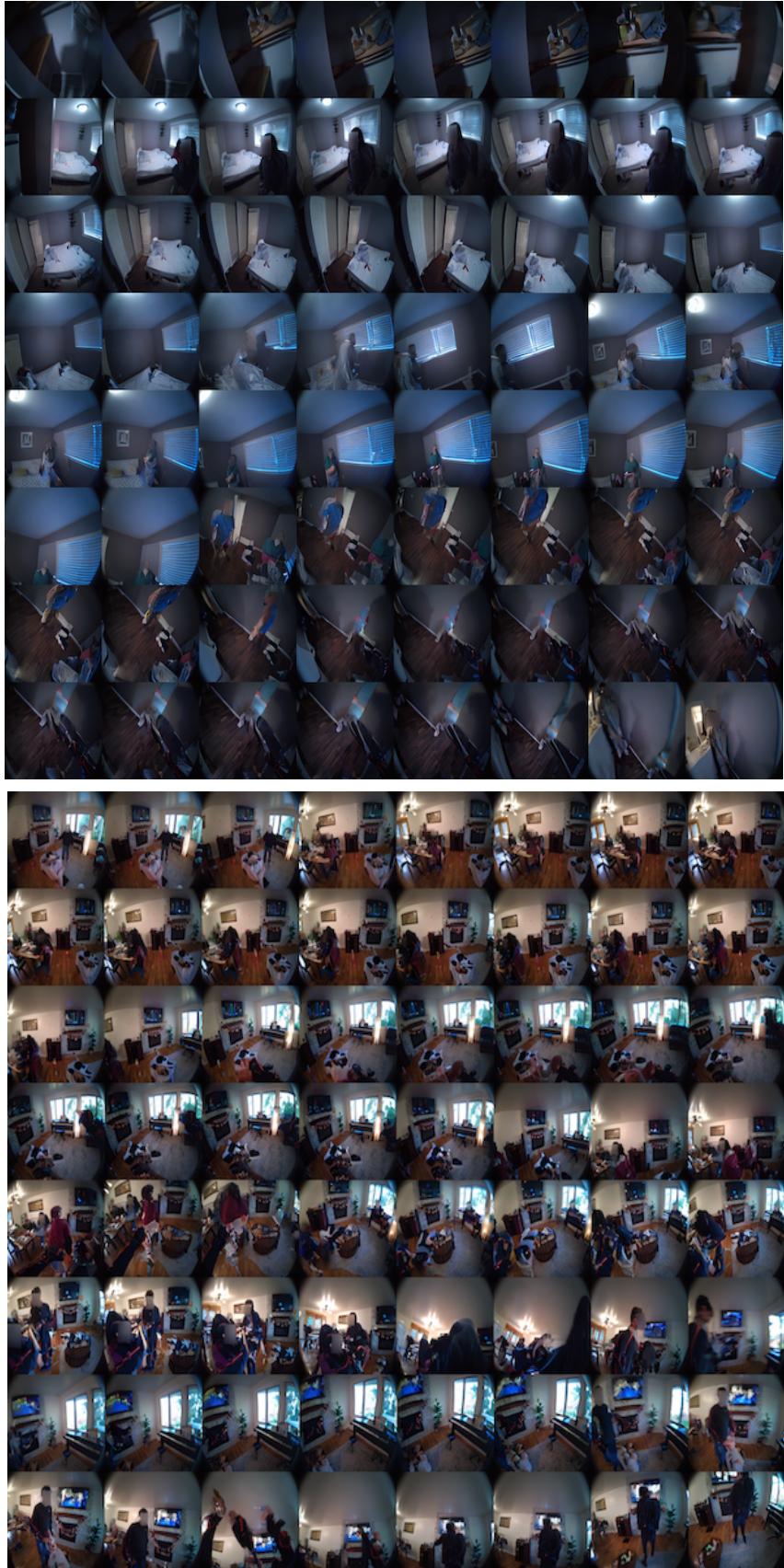


Figure 23: **Generation Over Long-Horizons.** We include 16-second video generation examples.



Figure 24: **Generation Over Long-Horizons**. We include 16-second video generation examples.

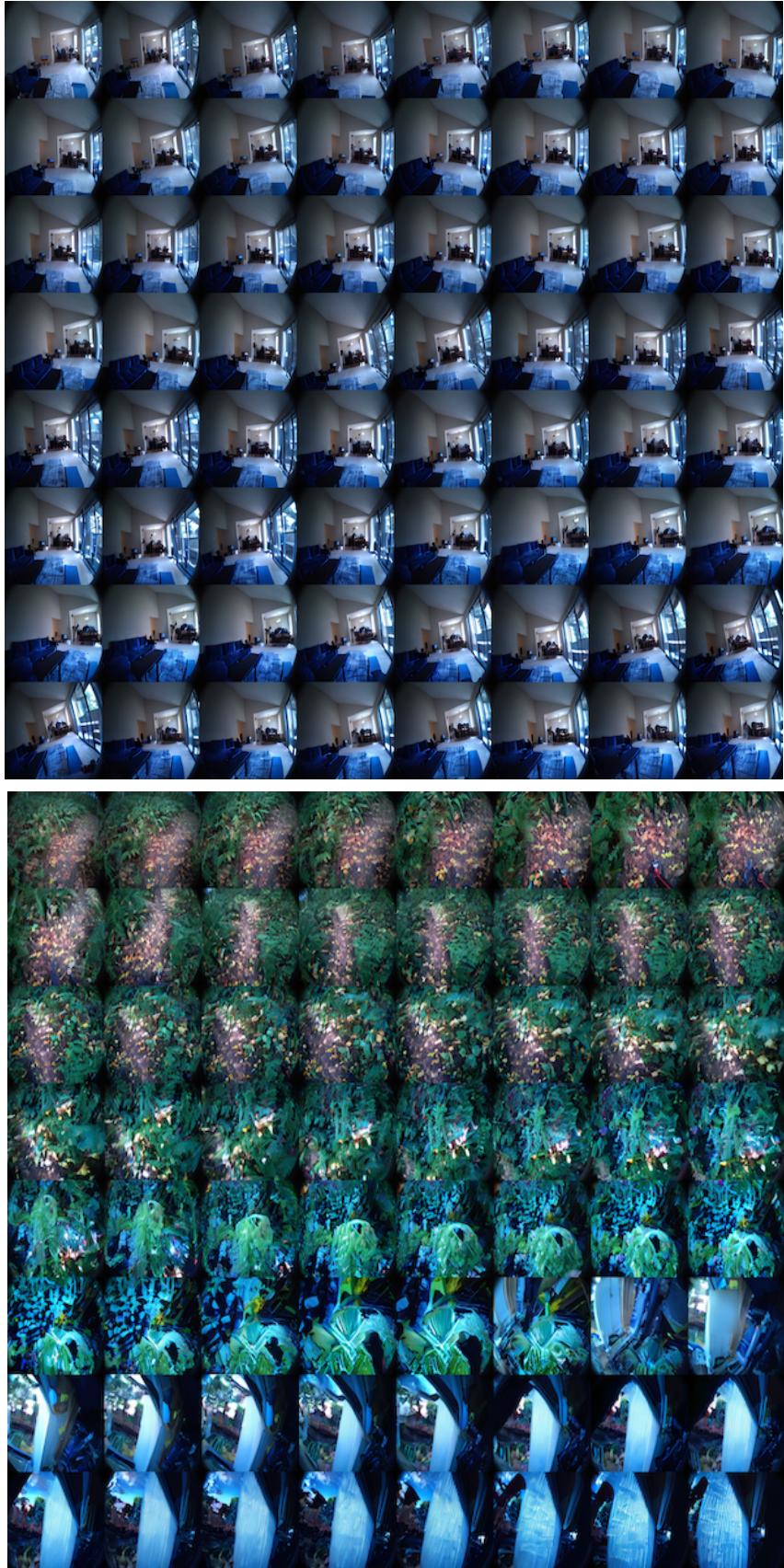


Figure 25: **Generation Over Long-Horizons.** We include 16-second video generation examples.