# HELLOWORLD: ENABLING SOCIALLY INTERACTIVE CHARACTERS IN VIDEO WORLD MODELS

**Anonymous authors**
Paper under double-blind review

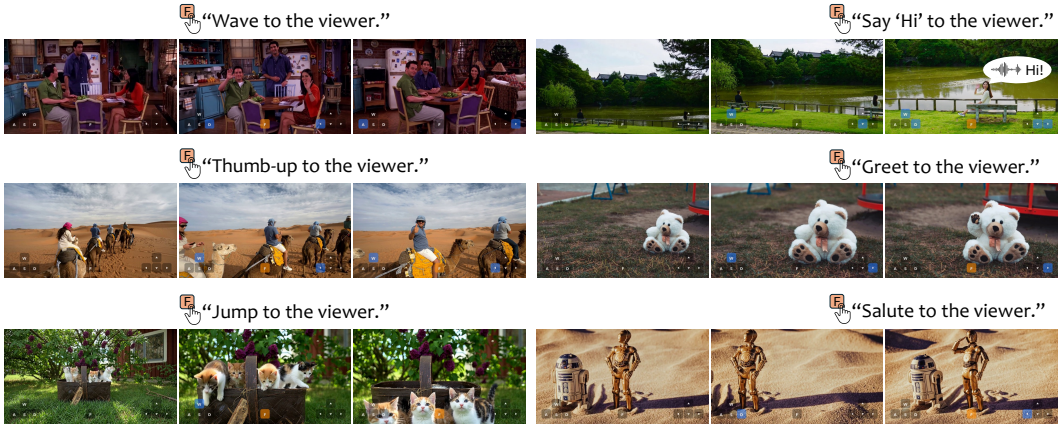



Figure 1: We propose HelloWorld, a video world model with socially interactive characters. With the interaction button F, users can prompt the on-screen character to interact with the viewer. HelloWorld supports diverse interactions across diverse characters, including humans, animals, and toys, while maintaining high-quality scene and camera-trajectory reconstruction.

## ABSTRACT

Despite the remarkable recent progress of video world models, social interaction between users and the characters within these worlds remains unsupported. To fill this gap, we present **HelloWorld**, a video world model that enables social interaction with in-world characters. With a single button press, users can prompt the on-screen character to respond toward the camera, *e.g.*, turning to the viewer, waving, nodding, or speaking a short greeting. To make these interactions natural, we propose a self-distillation pipeline that finetunes the video generation model on data synthesized by itself. Each synthesized clip contains both social interactions and camera motion, allowing the model to learn camera-pose conditioning without degrading interaction quality. At inference, we further introduce a training-free module that determines when the interaction occurs. Upon a button press, it modulates the cross-attention masks of the DiT so that the interaction-related text prompt attends only to the frames within the press window, temporally localizing the character's response. We further build **HelloWorldBench**, a 400-sample benchmark with three social interaction metrics alongside three conventional metrics, for evaluation. Experiments demonstrate that HelloWorld surpasses a variety of baselines in interaction quality, while maintaining state-of-the-art picture aesthetics and camera-pose following.

## 1 INTRODUCTION

Video world models learn to dream in pixels, simulating visual worlds whose futures unfold in response to the evolving environment. Recent works make this simulation interactive, conditioning on user inputs that steer the camera through the world (Sun et al., 2025) or trigger events within it (Ball



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

et al., 2025; Team et al., 2026a). These capabilities open up applications in game production (Che et al., 2025), world simulation (Brooks et al., 2024), and film making.

However, existing world models offer no support for social interaction between the user and the characters within these worlds. In some generated worlds, characters remain static pixels (Wang et al., 2026c). Some models are able to animate characters (Team et al., 2026b), yet their motions are ambient behaviors rather than social interactions with the user.

To fill this gap, we propose HelloWorld, an interactive video world model that enables users to actively create social interactions with in-world characters. Besides the camera trajectory and the text prompt, HelloWorld accepts an additional input, interaction button `F`. As illustrated in Fig. 1, when `F` is pressed, the in-world character interacts with the user, following the description in the text prompt.

HelloWorld consists of two components: *self-distillation* for finetuning (training) and a *temporal cross-attention mask* for inference. The self-distillation converts a pretrained video generation model into a controllable world model using its own generations. We first curate a set of prompts to generate videos containing both social interactions and camera motions. For each generated video, we annotate its camera trajectory and lift the first frame into a point cloud via off-the-shelf tools. Given the annotated trajectory, we re-render the point cloud along the camera path to synthesize a warped video, which serves as an explicit encoding of the target camera motion. During finetuning, the warped video is injected into the model as a history condition, from which the model learns to follow the specified camera trajectory while preserving the interaction quality. The temporal cross-attention mask is training-free and applied at inference time to control *when* an interaction occurs. Specifically, upon the button `F` press, the mask modulates the cross-attention layers of the DiT such that the interaction-related text tokens attend only to the frames within the press window, thereby temporally localizing the character's response.

To evaluate interactions, we introduce HelloWorldBench, a benchmark built upon 120 high-quality images spanning diverse humans, animals, and stylized characters. Pairing these images with designed text prompts and camera trajectories yields 400 evaluation samples. Beyond standard metrics for camera accuracy and video quality, we propose three interaction-specific metrics that respectively assess *what* interaction is performed, *when* it occurs, and *whether* it is directed toward the viewer.

Experiments on HelloWorldBench demonstrate that HelloWorld substantially outperforms existing methods on all three interaction metrics. Meanwhile, it remains competitive with or surpasses recent world models in video quality and camera following, achieving state-of-the-art performance overall. Our main contributions are summarized as:

- We propose HelloWorld, an interactive video world model that enables users to actively create social interactions with in-world characters.
- We build HelloWorldBench, the first benchmark for social interactions in world models. It provides newly collected images, designed social prompts, camera trajectories, and a suite of interaction metrics.
- We conduct comprehensive evaluations, demonstrating that HelloWorld substantially outperforms existing world models in social interaction, while maintaining state-of-the-art video quality and camera-pose following.

## 2 Related Work

### 2.1 Interactive Video World Models

Video world models learn to simulate visual worlds (Bai et al., 2025a) from large-scale videos of real and virtual environments (Zhou et al., 2018; Ling et al., 2024; Li et al., 2025b; Zhou et al., 2025b; Wang et al., 2026a; Che et al., 2025). Built upon video generation models (Wan et al., 2025; Yuan et al., 2026), recent interactive world models (Bruce et al., 2024; Agarwal et al., 2026; Feng et al., 2026; Zhu et al., 2025; Gao et al., 2025) further introduce controls of camera (Sun et al., 2025; Wang et al., 2026c; Huang et al., 2025; Sun et al., 2026), keyboard (Li et al., 2025a; Valevski et al., 2025; Decart et al., 2024; Guo et al., 2025; Yu et al., 2025; Savva et al., 2026), skill (Team et al., 2026a; Li et al., 2026d), and event (Ball et al., 2025; Mao et al., 2026), enabling interactive exploration of

the generated world. However, while the reconstruction and exploration of the environment have been extensively studied (Duan et al., 2025; Li et al., 2026a; Xu et al., 2026), interaction with the subjects within it remains underexplored. Recent works support object interactions such as picking up and placing items (Xiong et al., 2026) and controlling object dynamics (Yin et al., 2026), and the concurrent ReactiveGWM (Wang et al., 2026b) controls interactions between NPCs, yet it is limited to a single game. Compared with these works, HelloWorld is the first video world model focusing on social interaction between in-world characters and the user, supporting diverse stylized characters across a vocabulary of social behaviors.

### 2.2 Multimodal Social Interactions

Multimodal social interaction refers to human communication across multiple modalities, including spoken language, facial expressions, gaze (Liu et al., 2021), gestures (Cao et al., 2025; Beg et al., 2026), and body movements (Müller et al., 2021). Earlier works focus on understanding social interactions between humans, including social video question answering (Zadeh et al., 2019; Kang et al., 2025; Peng et al., 2026a), social behavior classification (Kim et al., 2026; Ouyang et al., 2025), and social conversation modeling (Lee et al., 2024; Li et al., 2026b; Ouyang et al., 2026a; Li et al., 2026c). With the development of generative models, recent works turn to the generation of multi-person social interactions, including talking video generation (Kong et al., 2026; Zhong et al., 2025; Huang et al., 2026; Wang et al., 2025c;a; Ma et al., 2025; Chu et al., 2026; Peng et al., 2026b; Lin et al., 2026; Zhou et al., 2025c), social interaction video generation (Ouyang et al., 2026b), and 3D motion generation (Liang et al., 2024; Xu et al., 2024; Ruiz-Ponce et al., 2026; Yu et al., 2026b). More recently, a series of social world models (Zhou et al., 2025a; Yu et al., 2026a; Zhang et al., 2025) employ LLM-based world models to simulate social dynamics, but they are limited to the text modality and lack multimodal interactions. Compared with these works, HelloWorld is the first to study multimodal social interaction in video world models, covering actions, gestures, facial expressions, and speech. Beyond that, HelloWorld is the first to introduce social interaction with non-human characters such as animals, robots, and toys.

## 3 Proposed Method

### 3.1 Preliminary

**Task formulation.** Our video world model $\mathcal{G}$ receives four inputs: a first-frame image $\mathbf{x}^0$ containing the scene and its characters; a text prompt $\mathbf{y}$ describing the world and the interaction event; a camera trajectory $\mathcal{C} = \{\mathbf{c}^i\}_{i=1}^N$ specifying the camera movement across the $N$ frames of the video, where $\mathbf{c}^i \in \mathrm{SE}(3)$ denotes the camera pose of frame $i$; and an interaction window $\mathcal{W} = [\tau_s, \tau_e]$ specifying when the social interaction occurs. The model generates a video $\mathcal{V} = \{\mathbf{x}^i\}_{i=1}^N$ in which the camera follows $\mathcal{C}$ throughout the clip, and the character interacts with the viewer within $\mathcal{W}$.

**Warp video condition.** Recent work (Wang & He, 2026) shows that a video generation model can be finetuned into a camera-controllable world model with only lightweight training, by conditioning on a warp video. The warp video is a pseudo-video that re-renders the first frame along the target camera trajectory. The warp video is fed to the DiT $\mathcal{G}$ as a history condition, providing an explicit, frame-aligned specification of the desired camera motion:

$$\mathcal{V}_{\mathrm{warp}} = \mathrm{warp}(\mathbf{x}^0, \mathcal{C}), \tag{1}$$

where $\mathrm{warp}(\cdot)$ lifts $\mathbf{x}^0$ into a point cloud with an off-the-shelf 3D reconstructor and reprojects it onto each target camera $\mathbf{c}^i$. Note that the warp video is inherently incomplete: regions occluded or outside the field of view in $\mathbf{x}^0$ appear as holes after reprojection. It thus serves as a geometric guidance for camera motion rather than a complete target, leaving the model to fill the missing regions. Following this insight, HelloWorld adopts the warp video as the camera condition:

$$\mathcal{V} = \mathcal{G}\big(\boldsymbol{\epsilon};\, \mathbf{x}^0,\, \mathbf{y},\, \mathcal{V}_{\mathrm{warp}}\big), \quad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}). \tag{2}$$

Inside $\mathcal{G}$, the warp video is tokenized and injected as history tokens, each sharing the temporal position embedding of its corresponding target frame.

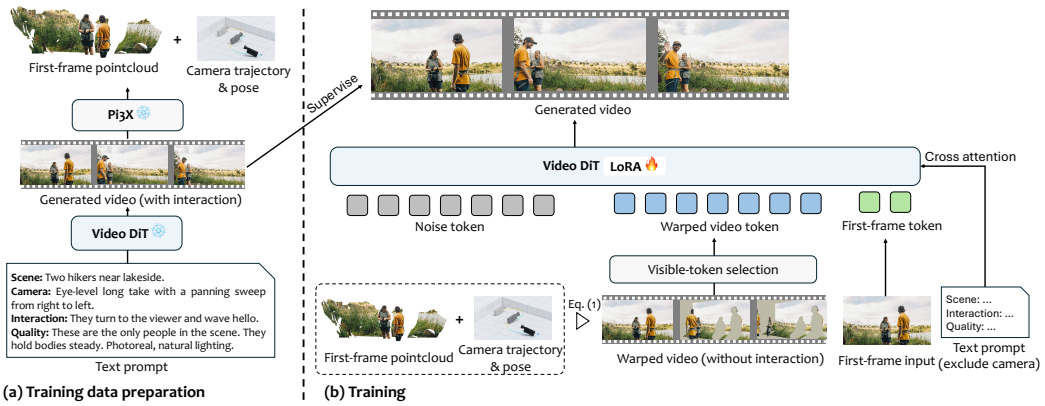



Figure 2: Training pipeline of HelloWorld. (a) Training data preparation: The frozen base model generates interaction-rich clips, from which Pi3X recovers the first-frame point cloud and camera poses. (b) Training process: The point cloud is re-rendered along the trajectory into the warp video, whose visible tokens condition the DiT. A lightweight LoRA is finetuned to reconstruct the original clip, with the camera motion removed from the text prompt.

### 3.2 Training

The base video generation model itself is capable of producing rich social interactions (Ouyang et al., 2026b). To preserve this ability while converting the model into a world model, we propose a self-distillation training pipeline: the base model generates interaction-rich videos with camera motions, and is then finetuned on these self-generated videos under the warp-video conditioning scheme in Sec. 3.1.

First, as shown in Fig. 2 (a), we prompt the base model to synthesize training data. The text prompt $\mathbf{y}$ contains four parts, $\mathbf{y}_{\text{scene}}$, $\mathbf{y}_{\text{inter}}$, $\mathbf{y}_{\text{camera}}$, and $\mathbf{y}_{\text{quality}}$, describing the scene, characters' interactions, camera trajectory, and video quality, respectively. After generation, following Wang & He (2026), we apply Pi3X (Wang et al., 2025b) to recover the first-frame point cloud and the per-frame camera trajectory. Detailed prompt design are provided in appendix.

Fig. 2 (b) shows the training process. For each clip, we render its warp video via Eq. (1). A visible-token selection module then discards the warp tokens without valid source observations, *i.e.*, the holes left by reprojection Wang & He (2026). The remaining warp tokens are concatenated with the noise tokens and first-frame tokens, and fed into the video DiT. The DiT is finetuned with a lightweight LoRA under flow matching loss:

$$\mathcal{L} = \mathbb{E}_{\mathbf{z}_0,\, t,\, \boldsymbol{\epsilon}} \Big\| \mathbf{v}_\theta\big(\mathbf{z}_t,\, t,\, \mathbf{x}^0,\, \mathbf{y}_{\text{scene}},\, \mathbf{y}_{\text{inter}},\, \mathbf{y}_{\text{quality}},\, \mathcal{V}_{\text{warp}}\big) - (\boldsymbol{\epsilon} - \mathbf{z}_0) \Big\|_2^2, \tag{3}$$

where $\mathbf{z}_0$ is the clean latent of the training video $\mathcal{V}$, $\mathbf{z}_t = (1-t)\,\mathbf{z}_0 + t\,\boldsymbol{\epsilon}$ is its noisy interpolation at timestep $t \in [0, 1]$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and $\mathbf{v}_\theta$ predicts the velocity field. Notably, the camera prompt $\mathbf{y}_{\text{camera}}$ is excluded to make sure that camera following is controlled solely by the warp video. The entire self-distillation pipeline requires no external data collection and no human annotation.

### 3.3 Inference

Inference is illustrated in Fig. 3. Given a first frame, a text prompt, and keyboard inputs, HelloWorld generates a video in which the character interacts with the user. The keyboard inputs are translated into a camera trajectory $\mathcal{C}$, from which the warp video is constructed via Eq. (1). The forward pass then follows the same procedure as training.

To further control when an interaction occurs, we propose a training-free temporal cross-attention mask. During inference, the interaction button `F` specifies the interaction window $\mathcal{W}$: a press at time $\tau_s$ opens $\mathcal{W} = [\tau_s, \tau_e]$, spanning the subsequent seconds. We then apply a temporal mask $M$ to the

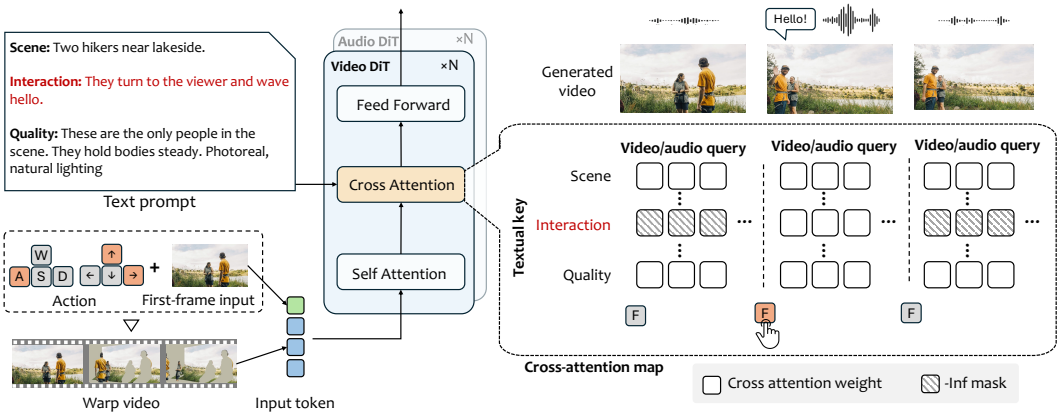



Figure 3: Inference pipeline of HelloWorld. Keyboard actions are translated into a camera trajectory and rendered as the warp video to condition the DiT (left). The temporal cross-attention mask blocks queries outside the F press window (hatched) from attending to the interaction prompt, temporally localizing the character's response (right).

cross-attention from the audio and video tokens to the text tokens:

$$M_{ij} = \begin{cases} -\infty, & i \notin \mathcal{W} \text{ and } j \in \mathbf{y}_{\text{inter}}, \\ 0, & \text{otherwise}, \end{cases}$$
$$\text{Attention} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d}} + M\right) V, \tag{4}$$

where $i$ indexes the temporal position of the query (video and audio) tokens and $j$ indexes the text tokens, with $j \in \mathbf{y}_{\text{inter}}$ denoting tokens of the interaction prompt. The mask prevents frames outside $\mathcal{W}$ from attending to $\mathbf{y}_{\text{inter}}$, so that the character responds to the viewer precisely within the press window and behaves ambiently elsewhere. This temporal control is training-free and incurs negligible overhead.

## 4 HelloWorldBench

Existing video world model benchmarks (Duan et al., 2025; Li et al., 2026a; Xu et al., 2026; Ying et al., 2026) center on scene fidelity and camera controllability, overlooking the characters within the world. We therefore propose **HelloWorldBench**, the first benchmark for evaluating viewer-directed social interactions in world models.

### 4.1 Dataset Construction

We collect 120 high-quality images from Unsplash (Unsplash, 2026), covering humans, animals, toys, and robots across diverse scenes and visual styles. For each image, an LLM agent designs two to four subject-appropriate interactions, *e.g.*, a human waving to the camera or a dog wagging its tail. We further design four common camera trajectories: static, scan, dolly-in, and orbit. The interaction timing is randomly assigned to early, middle, or late. Combining the images, interactions, camera trajectories, and timings yields 400 samples, spanning 264 character instances and 101 interaction types. More details are provided in the appendix.

### 4.2 Evaluation Metrics

As shown in Fig. 4, we design three metrics decouple social interaction into *what* is performed, *when* it occurs, and *whether* it is directed toward the viewer.

**ActAcc↑.** ActAcc measures whether the generated character performs the prompted interaction. We feed the generated video to a VLM judge (Bai et al., 2025b) with an eight-way multiple-choice

Table 1: Comparison of HelloWorld with baseline methods on HelloWorldBench.

| Method | Camera Input | Social Interaction | | | Video Quality | | Camera |
|---|---|---|---|---|---|---|---|
| | | ActAcc ↑ | TimeAcc ↑ | GazeDev° ↓ | BgCons ↑ | Aesthetic ↑ | CamCtrl ↑ |
| LTX-2.3 (HaCohen et al., 2026) | None | 42.5 | 52.6 | 38.1 | 94.7 | 5.19 | 31.4 |
| WorldPlay (Sun et al., 2025) | Keyboard | 10.5 | 41.2 | 63.5 | 94.7 | 5.14 | 48.1 |
| Matrix-Game 3.0 (Wang et al., 2026c) | Keyboard | 8.0 | 37.5 | 77.2 | 88.3 | 4.95 | 51.1 |
| LingBot-World (Team et al., 2026b) | Keyboard | **50.5** | 39.5 | 59.0 | 93.1 | 5.21 | 62.6 |
| SANA-WM (Zhu et al., 2026) | Trajectory | 38.5 | 30.9 | 56.8 | 92.5 | 5.01 | 70.0 |
| Warp-as-History (Wang & He, 2026) | Trajectory | 33.8 | 35.2 | 52.8 | 95.3 | 5.14 | 65.1 |
| **HelloWorld (ours)** | Trajectory | 41.4 | **81.7** | **40.2** | **96.9** | **5.27** | **82.9** |

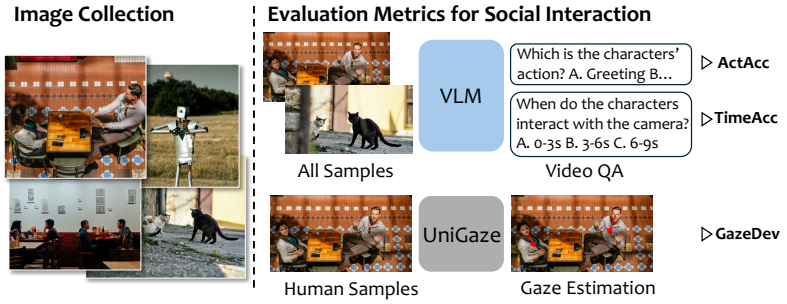



Figure 4: HelloWorldBench overview. Left: we collect 120 high-quality and diverse first-frame images. Right: we use VLMs and a gaze estimator to evaluate social interactions.

question, *e.g.*, "*Which is the characters' action? A. Greeting. B. Dancing. . . .*". The ground truth is the prompted action while the distractors are sampled from other interaction types in HelloWorldBench. ActAcc is the fraction of videos for which the VLM selects the correct answer.

**TimeAcc↑.** TimeAcc measures whether the interaction occurs at the user-specified moment. The VLM judge is asked "*When do the characters interact with the camera?*" and selects among uniformly divided three temporal segments (*e.g.*, 0–3 s, 3–6 s, and 6–9 s for a 9-second clip) and an additional *no interaction* option. TimeAcc is the fraction of videos whose selected segment matches the assigned interaction window. Note that videos where the judge selects *no interaction* are excluded from the calculation.

**GazeDev↓.** GazeDev measures whether the interaction is directed toward the viewer. Over the 217 human samples, we estimate the characters' gaze (Qin et al., 2026) and compute the mean angular deviation between the gaze direction and the camera's optical axis within the interaction window. A deviation of 90° is assigned if no face detected.

**Other metrics.** Following previous works, we also report background consistency (**BgCons**), aesthetic score (**Aesthetic**), and camera controllability (**CamCtrl**). More details can be found in the appendix.

## 5 Experiments

### 5.1 Baseline Methods

We compare HelloWorld with five recent competitive video world models: WorldPlay (Sun et al., 2025), Matrix-Game 3.0 (Wang et al., 2026c), LingBot-World (Team et al., 2026b), SANA-WM (Zhu et al., 2026), and Warp-as-History (Wang & He, 2026). We also include LTX-2.3 (HaCohen et al., 2026), the base model itself, as a pure video generation model without a trajectory interface. More details about the baseline methods are provided in appendix.




Figure 5: Qualitative comparison with baseline methods.

### 5.2 Implementation Details

Our self-distillation finetunes the base model LTX-2.3 on 156 synthesized videos for 2,000 steps, with a learning rate of $1 \times 10^{-4}$. The LoRA is rank-32, applied to all projection matrices of the self-attention layers in the video branch. The attention strength of the warp reference tokens is set to 0.3. At inference, HelloWorld and all baseline methods receive the same text prompt and camera motion input for each sample to generate 10-second videos. The resolution and frame rate of HelloWorld are set to $1280 \times 704$ and 24 fps. Each sample is run with three different random seeds, reporting the averaged results. The VLM judge used in all experiments is Qwen3.6-35B-A3B (Qwen Team, 2026). All training and testing are conducted on a single NVIDIA H200 GPU.

### 5.3 Main Results

**Comparison with baselines.** As shown in Table 1, HelloWorld outperforms existing world models by a clear margin. WorldPlay and Matrix-Game 3.0 tend to generate static scenes and thus score low across all three social interaction metrics. LingBot-World, SANA-WM, and Warp-as-History follow the text prompt and produce character actions, LingBot-World even attains the highest ActAcc. However, none of them controls *when* the interaction occurs: their TimeAcc stays around 30%, close to the 33.3% random-guess level of the three-way timing question, whereas HelloWorld reaches 81.7%. Their larger gaze deviations further indicate that the generated characters fail to look toward the camera. On video quality, HelloWorld performs on par with or slightly better than the best baseline, showing that self-distillation preserves the quality of the base model. Finally, HelloWorld attains the best camera-following score, confirming that warp-video conditioning is an effective approach to converting a video generation model into a camera-controllable world model.

**Visualizations.** Fig. 5 presents the qualitative comparison with baseline methods. As highlighted by the red boxes, the baseline models often fail to generate the correct interactions: the characters neither look toward the camera nor perform the prompted actions, such as crossing arms or approaching. LingBot-World and SANA-WM do generate some actions, but are prone to hallucination. For "make a heart", an extra character abruptly appears in the scene. In contrast, HelloWorld produces faithful viewer-directed interactions: the characters act toward the camera with vivid and natural motions, consistent with the quantitative results in Table 1.

### 5.4 Ablation Studies

**Effect of training data.** We study the effect of training data by comparing three settings: *Real-video* trains the LoRA on real data Wang & He (2026), which contains no social interaction; *Human-*

Table 2: Ablation on the training data.

| Training data | Social Interaction | | | Video Quality | | Cam. |
|---|---|---|---|---|---|---|
| | ActAcc ↑ | TimeAcc ↑ | GazeDev° ↓ | BgCons ↑ | Aesthetic ↑ | CamCtrl ↑ |
| Real-video | 36.4 | 81.3 | 51.3 | **97.2** | 5.23 | **83.1** |
| Human-only | 40.4 | **82.4** | 42.3 | 96.8 | **5.27** | 82.7 |
| Full | **41.4** | 81.7 | **40.2** | 96.9 | **5.27** | 82.9 |




Figure 6: Qualitative ablation on the training data. Without interaction data, the model fails to face the viewer and hallucinates (red boxes).

Table 3: Ablation on the temporal cross-attention masks.

| $M_v$ | $M_a$ | ActAcc ↑ | TimeAcc ↑ | SpeechInWin ↑ |
|---|---|---|---|---|
| | | **42.5** | 36.7 | 52.5 |
| ✓ | | 41.5 | 80.9 | 62.8 |
| ✓ | ✓ | 41.4 | **81.7** | **69.1** |

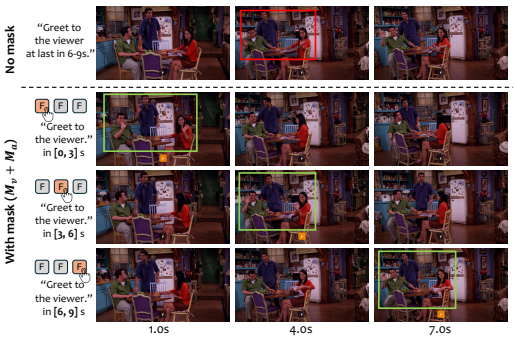



Figure 7: Qualitative ablation on the temporal cross-attention masks. Without the temporal mask, the greeting mistimes (red box). With the mask, it precisely follows the F-press window (green boxes).

*only* trains on self-generated videos with interactions, but restricted to human characters; and *Full* further extends the coverage to non-human subjects such as animals and cartoon characters. As shown in Table 2, the self-generated data (Human-only and Full) substantially improves ActAcc and GazeDev, while other metrics are barely affected. TimeAcc remains unchanged as expected, since the interaction timing is controlled by the training-free mask rather than training data. Video quality and camera following are likewise stable, indicating that our self-distillation improves the interaction ability without side effects.

Fig. 6 further illustrates the difference: model trained with real videos performs the prompted action, but not toward the viewer. We attribute this to the training data. Real videos without interactions merely teach the model to inpaint the missing regions of the warp video, providing no signal for engaging with the camera. In contrast, the self-generated interaction data natively contains characters gazing and acting toward the camera, and the model trained on it thus learns to produce viewer-directed actions.

**Effect of temporal cross-attention masks.** We compare three settings of the temporal cross-attention mask: no mask, masking the video stream only ($M_v$), and masking both the video and audio streams ($M_v + M_a$). We additionally report SpeechInWin, which transcribes the generated audio with Whisper (Radford et al., 2023) and checks whether the speech starts within the interaction window. Its accuracy is computed in the same way as TimeAcc. As shown in Table 3, the mask on each stream improves the temporal control of the corresponding modality, and applying both yields the best overall performance. Interestingly, the no-mask setting attains the highest ActAcc. This is expected, as without temporal constraints, the character may perform the prompted action throughout the video, which ActAcc rewards regardless of timing. The slight drop is thus the cost of temporal localization, rather than a loss of interaction ability.

As shown in Fig. 7, without the temporal mask, the interaction still occurs but at an arbitrary time, even when the timing is explicitly specified in the text prompt. With the mask applied, the interaction precisely falls within the designated window, confirming that our training-free mask equips the world model with temporal control over interactions.

Table 4: Computational cost of HelloWorld and baselines.

| **Method** | Time (s) | Time/frame (s) | Res. / frames | FLOPs ($\times 10^{15}$) |
|---|---|---|---|---|
| LTX-2.3 | 50.3 | 0.21 | $1280 \times 704$ / 241 | 6.9 |
| WorldPlay | 131.8 | 0.56 | $832 \times 480$ / 237 | 29.3 |
| Matrix-Game 3.0 | 62.5 | **0.15** | $1280 \times 704$ / 417 | 11.2 |
| LingBot-World | 63.3 | 0.39 | $832 \times 464$ / 153 | 18.2 |
| SANA-WM | **44.9** | 0.27 | $1280 \times 704$ / 168 | **3.2** |
| **HelloWorld** | <u>60.2</u> | <u>0.26</u> | $1280 \times 704$ / 241 | <u>9.4</u> |

Table 5: User study. Each cell reports the percentage of judgments preferring HelloWorld over the competing method, with the bootstrap 95% confidence interval. $N$ is the number of judgments.

| **HelloWorld *vs.*** | $N$ | Action Naturalness | Interaction | Scene Quality |
|---|---|---|---|---|
| Real-video LoRA | 270 | 83.7 [79.6, 88.1] | 88.5 [84.8, 92.2] | 82.6 [78.1, 87.0] |
| SANA-WM | 330 | 90.9 [87.6, 93.9] | 85.5 [81.8, 89.1] | 91.2 [87.9, 94.2] |
| Warp-as-History | 330 | 81.2 [77.0, 85.5] | 80.6 [76.4, 85.2] | 83.6 [79.4, 87.6] |
| LingBot-World | 300 | 77.3 [72.7, 82.0] | 71.7 [66.7, 77.0] | 88.0 [84.0, 91.7] |

**Computational cost.** Table 4 reports the average computational cost of generating a 10-second clip. Compared with the base model LTX-2.3, the additional warp-video tokens increase the inference time by roughly 20% (50.3s $\rightarrow$ 60.2s) and the FLOPs by 36% (6.9 $\rightarrow$ 9.4$\times 10^{15}$), which is the modest price of camera controllability. Despite operating at high resolution (1280$\times$704, 241 frames), HelloWorld remains competitive with the baselines: its per-frame time (0.26 s) is on par with SANA-WM (0.27 s) and well below WorldPlay (0.56 s) and LingBot-World (0.39 s).

### 5.5 User Study

To evaluate the perceptual quality of the interactions, we conduct a user study on 41 samples of HelloWorldBench. We invite 30 human raters, and present them with pairs of videos generated by HelloWorld and by a competing method from the same text prompt and camera motion, in a randomized order. For every pair, the raters answer three questions: (1) *In which video is the intended action performed more naturally?* (2) *In which video does the character feel more like it is interacting with you, the viewer?* and (3) *Which video has higher scene quality?* The results are reported in Table 5. HelloWorld is preferred over all four competing methods on all three questions, with every confidence interval lying above 66%. This confirms that it achieves state-of-the-art performance in the social interaction dimension as perceived by human viewers. In particular, the comparison against the LoRA trained on real videos yields the largest margin on the interaction question. This further verifies that our self-distillation strategy preserves the social prior and substantially improves the quality of the interaction.

## 6 Conclusion

This paper presents HelloWorld, a video world model that enables social interaction with in-world characters. We propose a self-distillation pipeline that turns a base video generation model into an interactive world model without any manually collected or annotated data. We further introduce a temporal cross-attention mask that achieves training-free control over the interaction timing. In addition, we contribute HelloWorldBench, the first benchmark for socially interactive world models. Experiments demonstrate that HelloWorld surpasses existing world models on the interaction metrics while maintaining state-of-the-art video quality.

**Limitation and future work.** Constrained by the design of the base model and the computation cost, HelloWorld does not yet support real-time interaction with users. World generation is driven by pre-specified camera trajectories and interaction scripts. Our future work will primarily explore autoregressive architectures to enable real-time interaction. Other directions include supporting long video generation and consistent modeling of individual characters, *e.g.*, persistent identities and sustained multi-round interactions.

### AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure. (i) *Generating synthetic data*: the text prompts used to construct HelloWorldBench and the self-distillation training set were drafted by a large language model. The interaction vocabulary and the per-character interaction designs were specified by the authors; the LLM expanded them into the scene, interaction, camera, and quality prompt components (Sec. 3.2 and Appendix A), and every generated prompt was manually reviewed and corrected by the authors before use. (ii) *Implementing methods*: parts of the training, inference, and evaluation code were written with the assistance of AI coding tools; all AI-generated code was verified and tested for correctness by the authors. We have not used generative AI tools for developing the theoretical models or conceptual frameworks, formulating mathematical claims, or interpreting results; assisting with proofs is not applicable to this work.

Additionally, we used generative AI tools for tasks with recommended disclosure, namely editing and polishing the writing of this manuscript. The research questions, method design, experimental protocol, and analysis of results were carried out by the authors, and all AI-assisted text was checked by the authors for accuracy and consistency with our results. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

## REFERENCES

Niket Agarwal, Arslan Ali, Jon Allen, Martin Antolini, Adeline Aubame, Alisson Azzolini, Junjie Bai, Maciej Bala, Yogesh Balaji, Josh Bapst, et al. Cosmos 3: Omnimodal world models for physical ai. *arXiv preprint arXiv:2606.02800*, 2026.

Jinbin Bai, Yu Lei, Hecong Wu, Yuchen Zhu, Shufan Li, Yi Xin, Xiangtai Li, Molei Tao, Aditya Grover, and Ming-Hsuan Yang. From masks to worlds: A hitchhiker's guide to world models. *arXiv preprint arXiv:2510.20668*, 2025a.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025b.

Philip J. Ball, Jakob Bauer, Frank Belletti, Bethanie Brownfield, Ariel Ephrat, Shlomi Fruchter, Agrim Gupta, Kristian Holsheimer, Aleksander Holynski, Jiri Hron, Christos Kaplanis, Marjorie Limont, Matt McGill, Yanko Oliveira, Jack Parker-Holder, Frank Perbet, Guy Scully, Jeremy Shar, Stephen Spencer, Omer Tov, Ruben Villegas, Emma Wang, Jessica Yung, Cip Baetu, Jordi Berbel, David Bridson, Jake Bruce, Gavin Buttimore, Sarah Chakera, Bilva Chandra, Paul Collins, Alex Cullum, Bogdan Damoc, Vibha Dasagi, Maxime Gazeau, Charles Gbadamosi, Woohyun Han, Ed Hirst, Ashyana Kachra, Lucie Kerley, Kristian Kjems, Eva Knoepfel, Vika Koriakin, Jessica Lo, Cong Lu, Zeb Mehring, Alex Moufarek, Henna Nandwani, Valeria Oliveira, Fabio Pardo, Jane Park, Andrew Pierson, Ben Poole, Helen Ran, Tim Salimans, Manuel Sanchez, Igor Saprykin, Amy Shen, Sailesh Sidhwani, Duncan Smith, Joe Stanton, Hamish Tomlinson, Dimple Vijaykumar, Luyu Wang, Piers Wingfield, Nat Wong, Keyang Xu, Christopher Yew, Nick Young, Vadim Zubov, Douglas Eck, Dumitru Erhan, Koray Kavukcuoglu, Demis Hassabis, Zoubin Gharamani, Raia Hadsell, Aäron van den Oord, Inbar Mosseri, Adrian Bolton, Satinder Singh, and Tim Rocktäschel. Genie 3: A new frontier for world models. 2025.

Hajra Anwar Beg, Mohamed Daoudi, and Angela Bartolo. Ambigest: A dataset of social gestures with inter-class similarity and intra-class variability. In *2026 IEEE 20th International Conference on Automatic Face and Gesture Recognition (FG)*, pp. 1–5. IEEE, 2026.

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. URL `https://openai.com/research/video-generation-models-as-world-simulators`.

Jake Bruce, Michael D Dennis, Ashley Edwards, Jack Parker-Holder, Yuge Shi, Edward Hughes, Matthew Lai, Aditi Mavalankar, Richie Steigerwald, Chris Apps, et al. Genie: Generative interactive environments. In *Forty-first International Conference on Machine Learning*, 2024.

Xu Cao, Pranav Virupaksha, Wenqi Jia, Bolin Lai, Fiona Ryan, Sangmin Lee, and James M Rehg. Socialgesture: Delving into multi-person gesture understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19509–19519, 2025.

Haoxuan Che, Xuanhua He, Quande Liu, Cheng Jin, and Hao Chen. Gamegen-x: Interactive open-world game video generation. In *International Conference on Learning Representations*, volume 2025, pp. 37546–37593, 2025.

Xuangeng Chu, Ruicong Liu, Yifei Huang, Yun Liu, Yichen Peng, and Bo Zheng. Unils: End-to-end audio-driven avatars for unified listening and speaking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25142–25152, 2026.

Etched Decart, Quinn McIntyre, Spruce Campbell, Xinlei Chen, and Robert Wachen. Oasis: A universe in a transformer. *URL: https://oasis-model. github. io*, 2(3):6, 2024.

Haoyi Duan, Hong-Xing Yu, Sirui Chen, Li Fei-Fei, and Jiajun Wu. Worldscore: A unified evaluation benchmark for world generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 27713–27724, 2025.

Ruili Feng, Han Zhang, Zhilei Shu, Zhantao Yang, Longxiang Tang, Zhicai Wang, Andy Zheng, Jie Xiao, Zhiheng Liu, Ruihang Chu, et al. The matrix: Infinite-horizon world generation with real-time moving control. *Advances in Neural Information Processing Systems*, 38:87318–87344, 2026.

Jianxiong Gao, Zhaoxi Chen, Xian Liu, Junhao Zhuang, Chengming Xu, Jianfeng Feng, Yu Qiao, Yanwei Fu, Chenyang Si, and Ziwei Liu. Longvie 2: Multimodal controllable ultra-long video world model. *arXiv preprint arXiv:2512.13604*, 2025.

Junliang Guo, Yang Ye, Tianyu He, Haoyu Wu, Yushu Jiang, Tim Pearce, and Jiang Bian. Mineworld: a real-time and open-source interactive world model on minecraft. *arXiv preprint arXiv:2504.08388*, 2025.

Yoav HaCohen, Benny Brazowski, Nisan Chiprut, Yaki Bitterman, Andrew Kvochko, Avishai Berkowitz, Daniel Shalem, Daphna Lifschitz, Dudu Moshe, Eitan Porat, et al. Ltx-2: Efficient joint audio-visual foundation model. *arXiv preprint arXiv:2601.03233*, 2026.

Tianyu Huang, Wangguandong Zheng, Tengfei Wang, Yuhao Liu, Zhenwei Wang, Junta Wu, Jie Jiang, Hui Li, Rynson Lau, Wangmeng Zuo, et al. Voyager: Long-range and world-consistent video diffusion for explorable 3d scene generation. *ACM Transactions on Graphics (TOG)*, 44 (6):1–15, 2025.

Yubo Huang, Weiqiang Wang, Sirui Zhao, Tong Xu, Lin Liu, and Enhong Chen. Bind-your-avatar: Multi-character-talking video generation with dynamic 3d-mask-based embedding router. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4440–4449, 2026.

Caixin Kang, Yifei Huang, Liangyang Ouyang, Mingfang Zhang, Ruicong Liu, and Yoichi Sato. Can mllms read the room? a multimodal benchmark for assessing deception in multi-party social interactions. *arXiv preprint arXiv:2511.16221*, 2025.

Junho Kim, Xu Cao, Houze Yang, Bikram Boote, Ana Jojic, Fiona Ryan, Bolin Lai, Sangmin Lee, and James M Rehg. Grasp: Learning to ground social reasoning in multi-person non-verbal interactions. *arXiv preprint arXiv:2605.15764*, 2026.

Zhe Kong, Feng Gao, Yong Zhang, Zhuoliang Kang, Xiaoming Wei, Xunliang Cai, Guanying Chen, and Wenhan Luo. Let them talk: Audio-driven multi-person conversational video generation. *Advances in Neural Information Processing Systems*, 38:70990–71013, 2026.

Sangmin Lee, Bolin Lai, Fiona Ryan, Bikram Boote, and James M Rehg. Modeling multimodal social interactions: new challenges and baselines with densely aligned representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14585–14595, 2024.

Dacheng Li, Yunhao Fang, Yukang Chen, Shuo Yang, Shiyi Cao, Justin Wong, Michael Luo, Xiaolong Wang, Hongxu Yin, Joseph Gonzalez, et al. Worldmodelbench: Judging video generation models as world models. *Advances in Neural Information Processing Systems*, 38, 2026a.

Jiaqi Li, Junshu Tang, Zhiyong Xu, Longhuang Wu, Yuan Zhou, Shuai Shao, Tianbao Yu, Zhiguo Cao, and Qinglin Lu. Hunyuan-gamecraft: High-dynamic interactive game video generation with hybrid history condition. *arXiv preprint arXiv:2506.17201*, 2(3):6, 2025a.

Xinpeng Li, Shijian Deng, Bolin Lai, Weiguo Pian, James M Rehg, and Yapeng Tian. Towards online multi-modal social interaction understanding. *Transactions on Machine Learning Research (TMLR)*, 2026b.

Xinpeng Li, Bolin Lai, Hardy Chen, Shijian Deng, Cihang Xie, Yuyin Zhou, James M Rehg, and Yapeng Tian. Omni-mmsi: Toward identity-attributed social interaction understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8684–8696, 2026c.

Zhen Li, Chuanhao Li, Xiaofeng Mao, Shaoheng Lin, Ming Li, Shitian Zhao, Zhaopan Xu, Xinyue Li, Yukang Feng, Jianwen Sun, Zizhen Li, Fanrui Zhang, Jiaxin Ai, Zhixiang Wang, Yuwei Wu, Tong He, Yunde Jia, and Kaipeng Zhang. Sekai: A video dataset towards world exploration. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025b. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/2368d4a3f24122b9f3779aaffa261719-Paper-Datasets_and_Benchmarks_Track.pdf.

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

Han Liang, Wenqian Zhang, Wenxuan Li, Jingyi Yu, and Lan Xu. Intergen: Diffusion-based multi-human motion generation under complex interactions. *International Journal of Computer Vision*, 132(9):3463–3483, 2024.

Zhi-Yi Lin, Thomas Markhorst, Jouh Yeong Chew, and Xucong Zhang. Polyslgen: Online multimodal speaking-listening reaction generation in polyadic interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 29379–29390, 2026.

Lu Ling, Yichen Sheng, Zhi Tu, Wentian Zhao, Cheng Xin, Kun Wan, Lantao Yu, Qianyu Guo, Zixun Yu, Yawen Lu, et al. Dl3dv-10k: A large-scale scene dataset for deep learning-based 3d vision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22160–22169, 2024.

Yunfei Liu, Ruicong Liu, Haofei Wang, and Feng Lu. Generalizing gaze estimation with outlier-guided collaborative adaptation. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 3835–3844, 2021.

Xingpei Ma, Shenneng Huang, Jiaran Cai, Yuansheng Guan, Shen Zheng, Hanfeng Zhao, Qiang Zhang, and Shunsi Zhang. Playmate2: Training-free multi-character audio-driven animation via diffusion transformer with reward feedback. *arXiv preprint arXiv:2510.12089*, 2025.

Xiaofeng Mao, Zhen Li, Chuanhao Li, Xiaojie Xu, Kaining Ying, and Kaipeng Zhang. Yume1. 5: A text-controlled interactive world generation model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7752–7761, 2026.

Philipp Müller, Michael Dietz, Dominik Schiller, Dominike Thomas, Guanhua Zhang, Patrick Gebhard, Elisabeth André, and Andreas Bulling. Multimediate: Multi-modal group behaviour analysis for artificial mediation. In *Proceedings of the 29th ACM International Conference on Multimedia*, pp. 4878–4882, 2021.

Liangyang Ouyang, Yuki Sakai, Ryosuke Furuta, Hisataka Nozawa, Hikoro Matsui, and Yoichi Sato. Leadership assessment in pediatric intensive care unit team training. *arXiv preprint arXiv:2505.24389*, 2025.

Liangyang Ouyang, Yifei Huang, Mingfang Zhang, Caixin Kang, Ryosuke Furuta, and Yoichi Sato. Multi-speaker attention alignment for multimodal social interaction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24608–24619, 2026a.

Liangyang Ouyang, Ruicong Liu, Caixin Kang, Yifei Huang, and Yoichi Sato. Socialdirector: Training-free social interaction control for multi-person video generation. *arXiv preprint arXiv:2605.10079*, 2026b.

Wenshuo Peng, Gongxuan Wang, Tianmeng Yang, Chuanhao Li, Xiaojie Xu, Hui He, and Kaipeng Zhang. Svbench: Evaluation of video generation models on social reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32872–32881, 2026a.

Ziqiao Peng, Yi Chen, Yifeng Ma, Guozhen Zhang, Zhiyao Sun, Zixiang Zhou, Youliang Zhang, Zhengguang Zhou, Zhaoxin Fan, Hongyan Liu, et al. Actavatar: Temporally-aware precise action control for talking avatars. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 39983–39993, 2026b.

Jiawei Qin, Xucong Zhang, and Yusuke Sugano. Unigaze: Towards universal gaze estimation via large-scale pre-training. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 5809–5820, 2026.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015.

Pablo Ruiz-Ponce, Sergio Escalera, José García-Rodríguez, Jiankang Deng, and Rolandos Alexandros Potamias. Interact2ar: Full-body human-human interaction generation via autoregressive diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 23559–23569, 2026.

Georgy Savva, Oscar Michel, Daohan Lu, Suppakit Waiwitlikhit, Timothy Meehan, Dhairya Mishra, Srivats Poddar, Jack Lu, and Saining Xie. Solaris: Building a multiplayer video world model in minecraft. *arXiv preprint arXiv:2602.22208*, 2026.

Huiqiang Sun, Zhan Peng, Size Wu, Kun Wang, Kang Liao, Dianyi Wang, Xingyu Zeng, Sheng Jin, Yangguang Li, Zhiguo Cao, et al. Prisma-world: Camera-controllable multi-agent video world model. *arXiv preprint arXiv:2606.09507*, 2026.

Wenqiang Sun, Haiyu Zhang, Haoyuan Wang, Junta Wu, Zehan Wang, Zhenwei Wang, Yunhong Wang, Jun Zhang, Tengfei Wang, and Chunchao Guo. Worldplay: Towards long-term geometric consistency for real-time interactive world modeling. *arXiv preprint arXiv:2512.14614*, 2025.

AlayaWorld Team, Kaipeng Zhang, Chuanhao Li, Yifan Zhan, Yongtao Ge, Yuanyang Yin, Jiaming Tan, Kang He, Liaoyuan Fan, Ruicong Liu, Xiaojie Xu, Xuangeng Chu, Zhen Li, Zhengyuan Lin, Zhixiang Wang, Zian Meng, and Zihui Gao. Alayaworld: Long-horizon and playable video world generation, 2026a. URL https://arxiv.org/abs/2607.06291.

Robbyant Team, Zelin Gao, Qiuyu Wang, Yanhong Zeng, Jiapeng Zhu, Ka Leong Cheng, Yixuan Li, Hanlin Wang, Yinghao Xu, Shuailei Ma, et al. Advancing open-source world models. *arXiv preprint arXiv:2601.20540*, 2026b.

Zachary Teed and Jia Deng. Droid-slam: Deep visual slam for monocular, stereo, and rgb-d cameras. *Advances in neural information processing systems*, 34:16558–16569, 2021.

Unsplash. Unsplash: Beautiful, free images and photos. https://unsplash.com, 2026. Accessed: 2026-07-17.

Dani Valevski, Yaniv Leviathan, Moab Arar, and Shlomi Fruchter. Diffusion models are real-time game engines. In *International Conference on Learning Representations*, volume 2025, pp. 73754–73776, 2025.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, et al. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42592–42603, 2026a.

Qiang Wang, Mengchao Wang, Fan Jiang, Yaqi Fan, Yonggang Qi, and Mu Xu. Fantasyportrait: Enhancing multi-character portrait animation with expression-augmented diffusion transformers. *arXiv preprint arXiv:2507.12956*, 2025a.

Yifan Wang and Tong He. Warp-as-history: Generalizable camera-controlled video generation from one training video. *arXiv preprint arXiv:2605.15182*, 2026.

Yifan Wang, Jianjun Zhou, Haoyi Zhu, Wenzheng Chang, Yang Zhou, Zizun Li, Junyi Chen, Jiangmiao Pang, Chunhua Shen, and Tong He. $\pi^3$: Permutation-equivariant visual geometry learning. *arXiv preprint arXiv:2507.13347*, 2025b.

Zeqing Wang, Danze Chen, Zhaohu Xing, Zizhao Tong, Yinhan Zhang, Xingyi Yang, and Yeying Jin. Reactivegwm: Steering npc in reactive game world models. *arXiv preprint arXiv:2605.15256*, 2026b.

Zhenzhi Wang, Jiaqi Yang, Jianwen Jiang, Chao Liang, Gaojie Lin, Zerong Zheng, Ceyuan Yang, Yuan Zhang, Mingyuan Gao, and Dahua Lin. Interacthuman: Multi-concept human animation with layout-aligned audio conditions. *arXiv preprint arXiv:2506.09984*, 2025c.

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

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

Liang Xu, Xintao Lv, Yichao Yan, Xin Jin, Shuwen Wu, Congsheng Xu, Yifan Liu, Yizhou Zhou, Fengyun Rao, Xingdong Sheng, et al. Inter-x: Towards versatile human-human interaction analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 22260–22271, 2024.

Xiaojie Xu, Zhengyuan Lin, Kang He, Yukang Feng, Xiaofeng Mao, Yuanyang Yin, Kaipeng Zhang, and Yongtao Ge. Worldmark: A unified benchmark suite for interactive video world models. *arXiv preprint arXiv:2604.21686*, 2026.

Xiangchen Yin, Wenzhang Sun, Jiahui Yuan, Zijie Liu, Yinda Chen, Wei Li, Dachun Kai, Chunfeng Wang, and Xiaoyan Sun. Holo-world: Unified camera, object and weather control for video world model. *arXiv preprint arXiv:2606.20083*, 2026.

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

Haofei Yu, Yining Zhao, Guanyu Lin, and Jiaxuan You. Building social world models with large language models. *arXiv preprint arXiv:2606.11482*, 2026a.

Heng Yu, Juze Zhang, Changan Chen, Tiange Xiang, Yusu Fang, Juan Carlos Niebles, and Ehsan Adeli. Socialgen: Modeling multi-human social interaction with language models. In *2026 International Conference on 3D Vision (3DV)*, pp. 1–17. IEEE, 2026b.

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

Shenghai Yuan, Yuanyang Yin, Zongjian Li, Xinwei Huang, Xiao Yang, and Li Yuan. Helios: Real real-time long video generation model. *arXiv preprint arXiv:2603.04379*, 2026.

Amir Zadeh, Michael Chan, Paul Pu Liang, Edmund Tong, and Louis-Philippe Morency. Social-iq: A question answering benchmark for artificial social intelligence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8807–8817, 2019.

Xinnong Zhang, Jiayu Lin, Xinyi Mou, Shiyue Yang, Xiawei Liu, Libo Sun, Hanjia Lyu, Yihang Yang, Weihong Qi, Yue Chen, et al. Socioverse: A world model for social simulation powered by llm agents and a pool of 10 million real-world users. *arXiv preprint arXiv:2504.10157*, 2025.

Zhizhou Zhong, Yicheng Ji, Zhe Kong, Yiying Liu, Jiarui Wang, Jiasun Feng, Lupeng Liu, Xiangyi Wang, Yanjia Li, Yuqing She, et al. Anytalker: Scaling multi-person talking video generation with interactivity refinement. *arXiv preprint arXiv:2511.23475*, 2025.

Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: learning view synthesis using multiplane images. *ACM Transactions on Graphics (TOG)*, 37(4): 1–12, 2018.

Xuhui Zhou, Jiarui Liu, Akhila Yerukola, Hyunwoo Kim, and Maarten Sap. Social world models. *arXiv preprint arXiv:2509.00559*, 2025a.

Yang Zhou, Yifan Wang, Jianjun Zhou, Wenzheng Chang, Haoyu Guo, Zizun Li, Kaijing Ma, Xinyue Li, Yating Wang, Haoyi Zhu, et al. Omniworld: A multi-domain and multi-modal dataset for 4d world modeling. *arXiv preprint arXiv:2509.12201*, 2025b.

Yingjie Zhou, Xilei Zhu, Siyu Ren, Ziyi Zhao, Ziwen Wang, Farong Wen, Yu Zhou, Jiezhang Cao, Xiongkuo Min, Fengjiao Chen, et al. Evaltalker: Learning to evaluate real-portrait-driven multi-subject talking humans. *arXiv preprint arXiv:2512.01340*, 2025c.

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

Yixuan Zhu, Jiaqi Feng, Wenzhao Zheng, Yuan Gao, Xin Tao, Pengfei Wan, Jie Zhou, and Jiwen Lu. Astra: General interactive world model with autoregressive denoising. *arXiv preprint arXiv:2512.08931*, 2025.