EgoSim: Egocentric Exploration in Virtual Worlds with Multi-modal Conditioning

Wei Yu [1 2] **Songheng Yin** [3] **Steve Easterbrook** [1] **Animesh Garg** [4 5]

Abstract

The recent advancements in video diffusion models have created a strong basis for developing world models with practical value. The upcoming challenge is to investigate how an agent can leverage this foundation model for understanding, interacting with, and planning within observed environments. This requires incorporating additional controllability into the model, transforming it into a versatile game engine that can be dynamically manipulated and controlled. To this end, we investigated the three key conditioning factors: camera, context frame, and text, and identified the current model design's shortcomings. More specifically, the fusion of camera embedding and features results in camera control being influenced by video features. On the other hand, while the injection of textual information compensates for unobserved spatiotemporal structures, it also intrudes into the already observed parts. To address these two issues, we propose the Spacetime Epipolar Attention Layer, which ensures that the egomotion generated by the model strictly adheres to the camera's movement. Additionally, we integrate the injection of text and context frame in a mutually exclusive manner to alleviate the intrusion problem. Through extensive experiments, we demonstrate that our new model **EgoSim** achieves excellent results on both the RealEstate and EpicKitchen datasets. For more results, please refer to https://egosim.github.io/EgoSim/.

1. Introduction

The success of diffusion models (Ho et al., 2020) has revolutionized the field of generative models, enabling advancements from realistic image generation (Rombach et al., 2022) to consistent video generation (Blattmann et al., 2023b). Recent works on long-term video generative pre-training (OpenAI, 2024) have further shown that diffusion models can effectively capture the complex dynamics of the physical world, laying a strong foundation for developing world models (Ha & Schmidhuber, 2018) with practical value.

The next challenge is to investigate how an agent can leverage this foundational model for understanding, interacting with, and planning within observed environments. To this end, we need to first identify what conditioning factors are necessary for constructing an effective world model. Our ultimate aim is to empower an agent to visualize potential scenarios based on observed environmental data, much like playing a game, thereby enhancing its ability to predict and respond to different situations (Yang et al., 2023).

Conditioning with Camera, Frame and Text To create a playable simulation engine, the primary desired feature is to enable agents to freely explore the simulated world. This necessitates using the agents' egomotion data, e.g. camera poses, as conditions to guide the generation process of the video diffusion model. Fortunately, as the majority of pixel changes in the video stem from the observer's egomotions rather than dynamic alterations in the surrounding environment, the pre-trained video diffusion models have already successfully internalized prior knowledge about typical patterns and transformations in the 3D world (Blattmann et al., 2023b; Voleti et al., 2024). Therefore, our task is to determine how to efficiently extract this prior knowledge and ensure that the videos generated by the model precisely align with the specified camera motion instructions.

Existing methods (Wang et al., 2023; He et al., 2024) typically combine the transformed camera information directly with intermediate features before feeding them into the temporal transformer layer. This approach can indeed roughly learn the camera motion. However, it is quite intuitive to notice that the intermediate features, along with the camera poses, jointly determine how the video is generated. As shown in Figure 1a, influenced by the distribution of its training data, the model may determine that even if the input motion is valid in certain situations, the camera cannot move forward because it has never seen an example of forward

[1]University of Toronto [2]Vector Institute [3]Columbia University [4]Georgia Tech [5]Nvidia. Correspondence to: Wei Yu <gnosis@cs.toronto.edu>.

1st Workshop on Controllable Video Generation at ICML, Vienna, Austria. PMLR 235, 2024.

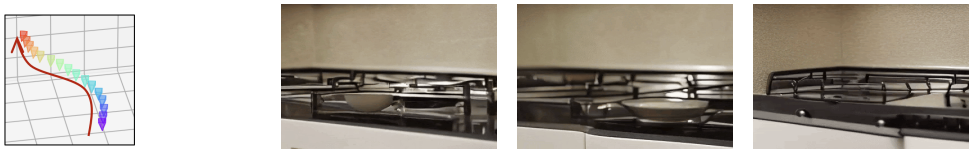


Figure 1. (a) Prompt: a kitchen with a stove top oven and a sink.

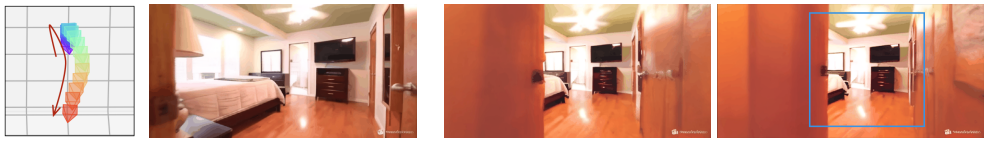


Figure 1. (b) Image-to-video generation: No prompt.

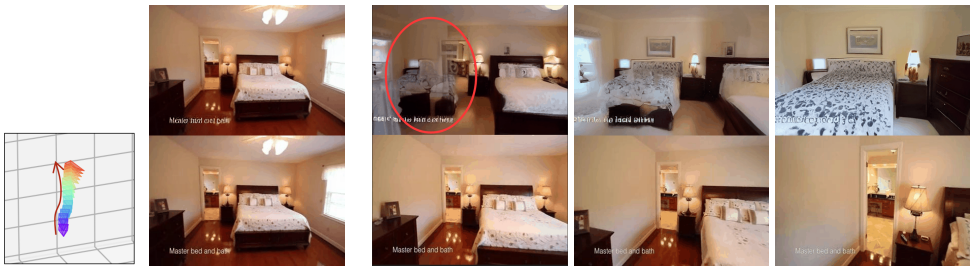


Figure 1. (c) Prompt: a bedroom with a bed and dresser in it.

Figure 1. To create a playable world simulator that supports free exploration and imagination based on observation, we have identified several primary bottlenecks: (a) We attempt to generate a video of moving forward in front of a stove, but the model fails to fulfill it because it has never seen such training data of moving above a stove. (b) Image-to-video model equipped with camera control, as emphasized inside the blue box, can only depict details already observed in the first frame. (c) Enable text-to-video models with image conditioning. The first row is a generated video while the second row is the groundtruth. As illustrated inside the red ellipse, introducing textual information can erode the already observed part, leading to a disruptive transition from realistic scenes to complete fantasy.

movement in such scenarios. Therefore, the generated video in this case will severely deviate from the user-specified motion.

Additionally, we aim for the model to seamlessly integrate observed environmental information, usually presented as context frames, into the video generation process while simultaneously inferring unobserved structures and predicting future interactions with the environment. Our preliminary experiments indicate that although the pretrained image-to-video model (Blattmann et al., 2023a) can produce authentic videos with minor movements, increasing the amplitude of the movement results in outputs that only depict the details visible in the context frame, as illustrated in Figure 1b. This suggests that the model needs supplementary information to imagine the unobserved parts.

Moreover, through further analysis on text-to-video models (Guo et al., 2023c), we found that adding textual information can effectively compensate for previous shortcomings by injecting extra spatiotemporal structure, leading to more coherent and realistic video sequences. However, as shown in Figure 1c, it became evident that this added spatiotemporal structure intrudes upon already observed parts, causing a disruptive transition from realistic scenes to complete fantasy.

In summary, our detailed examination reveals that while each condition in multi-condition inputs has an effective controlling method, *integrating these conditions together often leads to their embeddings interfering with or negating each other, significantly undermining the reliability of the generated video*. This indicates that using a naively unified embedding control interface may inevitably harm the accuracy of video generation. Consequently, this paper focuses on **resolving the interactions between different conditioning modalities to construct a coordinated and compositional world model**. The contributions of our paper can be categorized as follows:

1. We introduce a novel plug-in DiT (Peebles & Xie, 2023) module called the **S**pacetime **E**pipolar **A**ttention layer (SEAL) to override the interference of other features in video generation, making it more accurately aligned with camera motion.

2. We propose to arrange text information and visual conditions in a mutually exclusive relationship and leverage camera movement information to further assist in clearly defining the boundaries between the text and visual elements, ensuring they do not interfere with each other during the generation process.

3. We evaluate the proposed method **EgoSim** on two com-

petitive benchmarks, RealEstate and Epic Kitchen datasets, in a multi-condition-input setting. Extensive experimental results show that our model achieves precise control that previous methods could not accomplish. It not only generates video that accurately follows camera movements but also fills in unobserved new information into existing observations and interacts with the environment.

2. Orchestrating Diverse Conditions

Given a reference image, a text prompt and a sequence of camera poses, our goal is to generate a video sequence which starts from the context frame, faithfully obeys the user-specified motion and interact with environment in accordance with the textual description. As discussed earlier, the current control methods for these conditioning factors often interfere with each other, leading to the generation of videos that are inconsistent and unexpected.

Our initial analysis has pinpointed two main sources of this conflict: the interaction between camera control and intermediate features, and the interplay between text and context frames. This chapter will begin by providing an overview of video generation models and then address these two critical issues separately to resolve the problems.

2.1. Preliminary

Diffusion Models (Ho et al., 2020) are a class of generative models designed to produce high-quality samples through a multi-step process. Starting with Gaussian noise, these models iteratively refine and denoise the initial random input. In the case of video generation (Blattmann et al., 2023b), a sequence of N images (or their latent features) $z_0^{1:N}$ are progressively subjected to noise ϵ, transforming them into a normal distribution over T steps. A neural network ϵ^σ is then trained to predict the added noise from these noised inputs. During training, the network aims to minimize the mean squared error (MSE) between its predictions and the actual noise. The training objective function is defined as follows:

$$\mathcal{L}(\theta) = \mathbb{E}_{z_{0:N},\epsilon,c_t,t}\left[\left\|\epsilon - \hat{\epsilon}_\theta(z_t^{1:N}, c_t, t)\right\|_2^2\right]$$

where c_t represents conditional embeddings.

2.1.1. Controllable Video Generation

Controllability plays an important role in video generation as it enables users to craft content precisely as they envision. In this work, we aim to integrate three distinct control conditions— **autonomous camera motion, reference images, and text** —into a single pipeline.

First, regarding camera motion, it's important to note that the field of video diffusion models is still relatively new. Control over camera motion has only been tentatively explored so far. Recent methods (Wang et al., 2023; He et al., 2024) primarily achieve control by combining camera embeddings with intermediate features. On the other hand, there is a wealth of research on both text-to-video (Guo et al., 2023c) and image-to-video generation (Blattmann et al., 2023a), but combining these two approaches remains rare. Typically, text-to-video generation incorporates CLIP embeddings (Radford et al., 2021) of texts through multiple cross-attention layers while image-to-video models are mainly trained from scratch by concatenating repeated first-frame image features to noised input.

2.2. Camera Control

In this paper, we attempt to develop a more robust and reliable method for controlling camera trajectories in video generation. Existing approaches, which concatenate or add camera embeddings with intermediate features (Wang et al., 2023; He et al., 2024), often result in camera movements being influenced by these features. Therefore, we need a control method that operates relatively independently of these features.

2.2.1. Spacetime Epipolar Attention Layer

We propose using geometric constraints derived from epipolar lines to precisely guide camera movements. Epipolar attention (Kant et al., 2024; Du et al., 2023) has been extensively studied in the context of 3D generation, but its potential for video generation remains largely unexplored. Current publicly available pre-trained models all utilize factored space-time attention, which means that a patch in one frame cannot directly attend to a patch in another frame. Inspired by SORA (OpenAI, 2024), we addressed this limitation by first rearranging the latent representation into spacetime patches and introducing the DiT (Peebles & Xie, 2023) structure to enable interaction between patches.

Next, we can utilize the newly established channel to enforce camera control by leveraging the epipolar geometry. More specifically, consider a patch coordinate (u, v) in the target frame I_t, where the intrinsic parameters $\mathcal{K}_t$ and extrinsic parameters, including rotation $\mathcal{R}_t$ and translation $\mathcal{T}_t$ relative to canonical frame, are known. The epipolar line $\mathbf{l}_i$ corresponding to this patch can be calculated using the fundamental matrix:

$$\mathbf{l}_i = \mathcal{F}_i \left[u, v, 1\right]^T = \mathcal{K}_i^{-T}\left(\left[\mathcal{T}_t\right]_\times \mathcal{R}_t\right)\mathcal{K}_t^{-1}\left[u, v, 1\right]^T \quad (1)$$

As illustrated in Figure 2, with the help of epipolar lines, model can identify which regions in other frames to focus

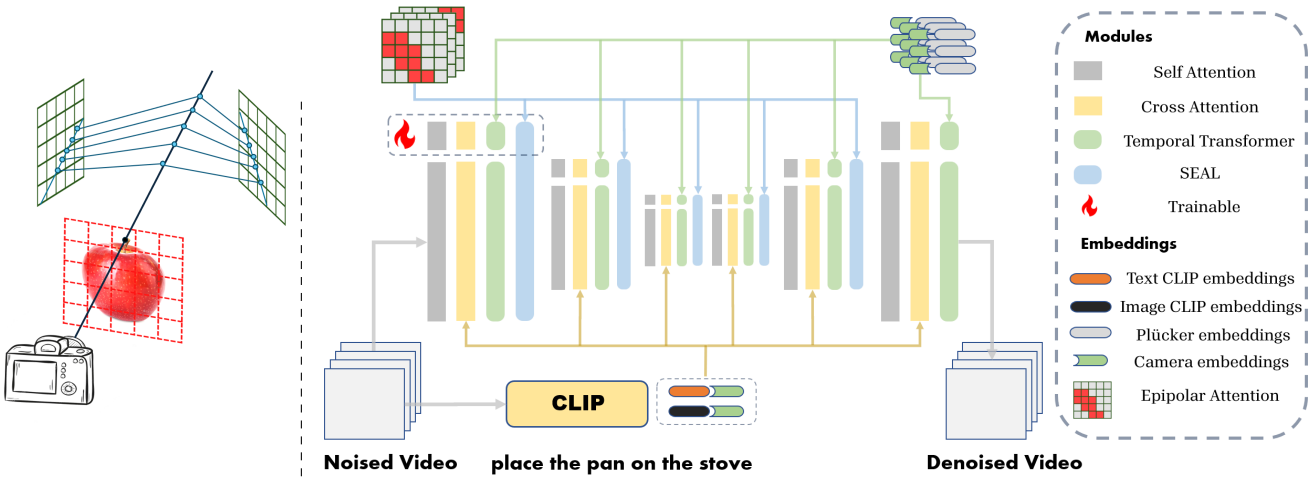


Figure 2. **EgoSim Overview**. Left: A simple illustration of how epipolar attention is calculated. Each patch only attends to patches from other frames which directly intersect or are near the re-projection of the unprojected epipolar line. Right: This diagram shows how various embeddings are injected into the video diffusion model. All attention modules along with the entire SEAL inside the dashed box with a flammable sign are set trainable. For simplicity, ResNets are omitted and trainable parts are not repeatedly emphasized with flammable sign for other blocks.

on for any given patch as follows:

$$\text{EpipolarAttention}(Q, K, V) = \text{softmax}\left(\frac{QK_\star^T}{\sqrt{d_k}}\right)V_\star \quad (2)$$

where $K_\star$ and $V_\star$ are keys and values calculated from patches directly crossed by or near the projection of epipolar lines. After constructing such links between all patches, the model can learn to effectively extract knowledge about viewpoint changes in the 3D world from video generative pre-training.

2.2.2. Camera Embeddings

Since the temporal transformer mainly focuses on learning motion information, integrating the camera embedding (Wang et al., 2023; He et al., 2024) into the temporal transformer can have the most direct impact on egomotion generation. More importantly, it is compatible with epipolar attention. Therefore, this paper retains and improves the method of injecting camera embedding. Methods like MotionCtrl (Wang et al., 2023), which directly flatten the camera pose, can also capture camera motion but overlook intrinsic parameters. Similar to a concurrent work, CameraCtrl (He et al., 2024), we concatenate the camera positional encoding (Vaswani et al., 2017) with Plücker embedding (Sitzmann et al., 2021) to get a more precise geometric interpretation for each patch.

2.3. Frame v.s. Text

As previously mentioned, after we equipped the video model with camera movement control capabilities, relying solely on the context frame was insufficient for the model to envision the details of unobserved scenes. This underscores the necessity of text input, which supplements the unobserved spatiotemporal structure. However, we also realized that textual information is intrusive, and thus it should be mutually exclusive with the observed visual information. Specifically, a particular patch in a specific frame should be explained either by text or by the context frame, but not both. To determine which source of information to use, we find that integrating camera motion data can help provide a clearer distinction between the two.

2.3.1. Better Embedding Coordination

To preserve the text-to-video capabilities that the cross-attention layer has learned, we decided to begin with text-to-video generation model and then modify it to support image-to-video generation. Therefore, we decided to make improvements based on the I2V-adapter (Guo et al., 2023a) pipeline. In I2V-adapter, an additional trainable cross-attention operation is performed in the self-attention layer for the first frame where ground truth is provided. In some cases, we find it would be helpful to use a precomputed epipolar attention mask into the cross-attention to achieve more precise feature fusion.

Additionally, we modify the cross-attention component of the UNet (Ronneberger et al., 2015). The I2V-adapter follows the IP-adapter (Ye et al., 2023) by adding the output of extra cross-attention computations with image embeddings. In contrast, we directly concatenate the text and image embeddings to mitigate mutual interference between the two types of information. Furthermore, we incorporate camera embedding, processed through a fully connected layer, into

both embeddings to inform the model which of these two embeddings should be more adopted under the current camera trajectory. Note that cross-attention layers of the spatial blocks are set trainable as well.

3. Evaluation

We will now evaluate the proposed method, EgoSim, and demonstrate its applications across various scenarios. In particular, we will highlight the following key capabilities of the model: (a). precise control of camera movement, and (b). generating meaningful interactions with the environment based on both observation and imagination.

3.1. Experimental Setup

Datasets: We select RealEstate (Zhou et al., 2018) and EpicKitchen (Damen et al., 2018) as evaluation benchmarks because they both come with camera poses (Tschernezki et al., 2024). Additionally, we used BLIP (Li et al., 2022) to label every frame for both datasets.

Pretrained weights: We leverage two pretrained video diffusion model as base to implement our method. More specifically, for the image-to-video case, we use Stable Video Diffusion (SVD) (Blattmann et al., 2023a). For experiments that require text input, we use AnimateDiff V3 (Guo et al., 2023c).

Baselines: To the best of our knowledge, there is no video model that can simultaneously accept camera information, video frames, and text as conditional inputs. Consequently, we can only attempt some simplified experimental scenarios, or we can modify existing methods for comparative purposes. CameraCtrl (He et al., 2024) and MotionCtrl (Wang et al., 2023) are included as baselines for video diffusion model equipped with camera motion control. Methods which can modify AnimateDiff for image conditioning involve I2V-adapter (Guo et al., 2023a) and SparseCtrl (Guo et al., 2023b). We reproduced the above models following the respective research papers and publicly available code.

Metrics: To estimate the fidelity of egocentric video prediction, SSIM ((Wang et al., 2004)) and the Fréchet Video Distance (FVD) (Unterthiner et al., 2019) are calculated between the predictions and groundtruths. Besides, we recruit Colmap (Sayab et al.) to assess the camera poses of generated videos to see if they authentically follow camera movements. **TransErr** and **RotErr** (He et al., 2024) are computed by comparing the ground truth camera poses $[R—T]$ and estimation $[R_*— T_*]$, as follows.

$$\text{RotErr} = \sum_{j=1}^{n} \arccos\left(\frac{\text{tr}(R_*^j R^{jT}) - 1}{2}\right)$$

$$\text{TransErr} = \sum_{j=1}^{n} \|T^j - T_*^j\|_2$$

We use a mixture of groundtruth trajectories and more difficult random trajectories to calculate the above errors.

3.2. RealEstate

RealEstate (Zhou et al., 2018) consists of a large number of open house video tours. In the case of AnimateDiff, we resize the video resolution to 256×384. We train and test the model with a length of 16 frames. For SVD, we resize to 256×512 and use $T = 14$.

3.2.1. Text-to-Video Generation with Camera Control

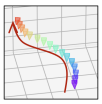
In this section, we will first evaluate the effectiveness of SEAL for text-to-video model equipped with camera control on the RealEstate dataset. As presented in left part of Table 1, the proposed method outperforms all previous methods by a wide margin on all metrics.

Qualitative analysis in Fig 3a further demonstrates the limitations of previous methods and the superiority of our approach. We can see that the baseline model is unable to generate out-of-distribution movements because its control is influenced by intermediate features. In contrast, our method perfectly adheres to the specified camera movements. In addition, SEAL also improves the consistency and realism of the videos thanks to its spacetime structure.

3.2.2. Image-to-Video Generation with Camera Control

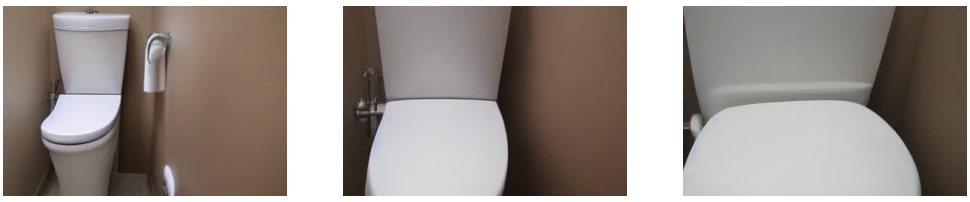
Next, we move to the pure image-to-video setting and compare our method with SVD + MotionCtrl. it should be noted that estimating SSIM becomes practically meaningful in the case of frame-conditioned generation because SSIM can more effectively tell us if the model strictly follows the given camera motion for a given groundtruth-generation pair. Besides, as SVD uses more parameters, calculations and data compared to AnimateDiff, the overall generated video quality is better.

Results: As shown in the first two rows of Table 1, EgoSim significantly improves on all metrics compared to the baseline methods. In Figure 3b, we can see that because the context frame we provide is a long corridor, MotionCtrl would assume that movement to the lower right corner should not occur even if it's a valid movement. On the contrary, our model can faithfully generate the corresponding movement to the lower right corner as expected.

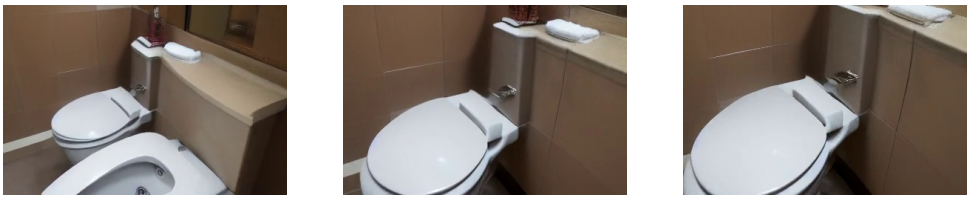


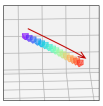
a. Text-to-video generation. Prompt: A bathroom with a toilet.
Input camera moving forward, left and then forward are shown on the left.
CameraCtrl depicts it wrong by not moving left and by stopping before toilet.

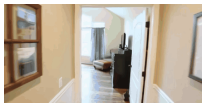
EgoSim



CameraCtrl

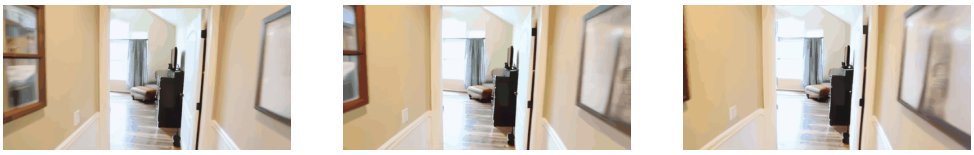




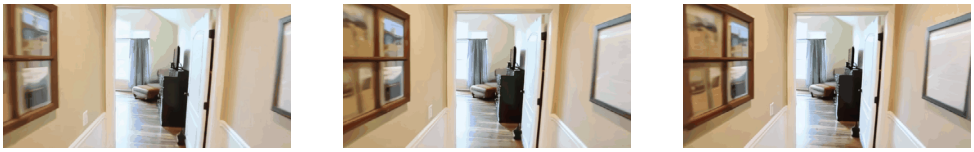


b. Image-to-video generation: No prompt. SVD base.
Input camera moving toward lower right are shown on the left.
MotionCtrl depicts it wrong by moving backward.

EgoSim



MotionCtrl



c. Prompt: An aerial view of a suburban neighborhood
Camera focuses on the center while moving in a circular motion.
Baseline depicts it wrong by adding irrelvant details.

EgoSim

CameraCtrl
+
I2V-Adapter

Figure 3. Qualitative Comparison on RealEstate Datasets

Method	RealEstate-T2V			RealEstate-I2V				EpicKitchen			
	TransErr	RotErr	FVD	TransErr	RotErr	FVD	SSIM	TransErr	RotErr	FVD	SSIM
MotionCtrl	-	-	-	7.74	0.88	330.1	0.822	-	-	-	-
EgoSim (SVD)	-	-	-	**5.21**	**0.53**	**223.8**	**0.934**	-	-	-	-
MotionCtrl	9.88	1.23	793.5	-	-	-	-	-	-	-	-
CameraCtrl	9.67	1.23	782.4	-	-	-	-	-	-	-	-
+ I2V	-	-	-	13.95	1.68	472.9	0.808	15.95	1.77	1172.1	0.733
+ SparseCtrl	-	-	-	-	-	1360.8	0.45	-	-	1566.2	0.42
EgoSim	**8.01**	**0.80**	**722.0**	**6.75**	**0.77**	**293.7**	**0.903**	**12.41**	**1.27**	**663.2**	**0.839**

Table 1. Quantitative comparison among baselines. Note that for TransErr, RotErr and FVD, lower number indicates better performance while higher SSIM means better.

3.2.3. Camera, Frame and Text

As we mentioned earlier, due to the lack of existing baselines in this setting, we can only create baseline models for comparison by combining existing methods. More specifically, we tried the following two combinations: CameraCtrl + I2V-Adapter and CameraCtrl + SparseCtrl (Guo et al., 2023b). It is worth noting that CameraCtrl + I2V-Adapter in general performed well, but we were unable to successfully reproduce the results with CameraCtrl + SparseCtrl and cannot calculate TransErr and RotErr even though we directly used the publicly available code and the pre-trained weights provided by the authors.

Results: Quantitative results are summarized in the middle part of Table 1 and EgoSim achieves the best performance. From the generated videos in Figure 3c, we can see that CameraCtrl + I2V-Adapter suffers from the intrusion issue while the proposed method can add reasonable imagination on top of observed details.

3.3. EpicKitchen

Finally, we move to the most challenging setting, EpicKitchen (Damen et al., 2018). Compared with RealEstate, EpicKitchen involves not only camera movements but also a significant amount of interaction with the environment. Thus, the model needs to learn additionally how to generate videos of these actions. We trained an additional LORA for EpicKitchen, and as in the previous section, we conducted experiments using CameraCtrl + I2V-Adapter and CameraCtrl + SparseCtrl as baselines. We resized the video frames to 256×448 and use $T = 14$.

Results: The quantitative comparisons are provided in the right part of Table 1 and EgoSim achieves the best scores on all metrics. The qualitative analysis in Fig 4 further reveals the advantage of our method. We encourage readers to view more impressive visual results in the project page. EgoSim not only learned to generate precise autonomous movements but also can perform environmental interactions such as washing dishes and opening drawers. This significantly broadens the range of potential applications for EgoSim. Past inverse dynamics approaches (Du et al., 2024) were limited in their ability to control autonomous movements, restricting test scenarios to fixed camera positions. However, with the advent of EgoSim, we can now attempt to learn more advanced inverse dynamics models. This enables dynamic camera positioning and broader testing capabilities, paving the way for more robust and versatile autonomous systems.

4. Related Work

4.1. Video generation

Since the advent of the deep learning era, video generation has been an active area of research. Initially, ConvLSTM model (Shi et al., 2015; Wang et al., 2017; Yu et al., 2020) was popular, followed by GAN models (Skorokhodov et al., 2022), and more recently, video diffusion models (Blattmann et al., 2023b) have become the mainstream. Video diffusion model (Blattmann et al., 2023a) usually extends a 2D image diffusion architecture to handle video data, enabling joint training on both images and videos from the ground up. To leverage powerful pre-trained image generators like Stable Diffusion (Rombach et al., 2022), subsequent approaches expand the 2D architecture by integrating temporal layers between the pre-trained 2D layers. This new model is then fine-tuned on a large video dataset. The latest advancement is SORA (OpenAI, 2024), which abandons the aforementioned strategy and directly uses diffusion transformers (Peebles & Xie, 2023) to train on large-scale videos of different sizes. Its generation quality is exceptionally impressive.

4.2. 3D generation

Our work is also closely related to 3D generation. With the introduction of the Scene Representation Transformer (Sajjadi et al., 2022), researchers have begun attaching ray embeddings directly to image patches. This allows neural networks to automatically learn the associations of different patches in 3D space. Many methods also directly employ epipolar geometry (Kant et al., 2024; Du et al., 2023). These methods typically aggregate information from different viewpoints through ray projection and unprojection to

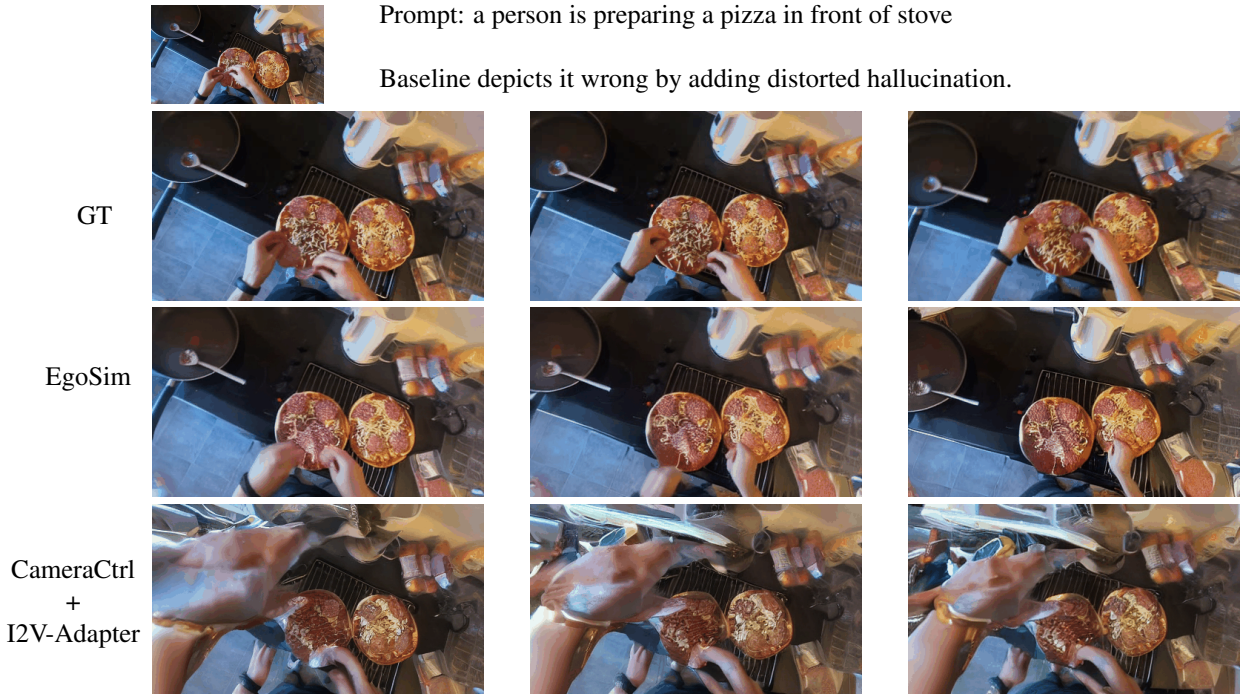


Figure 4. Qualitative Comparison on EpicKitchen Dataset

achieve consistent 3D generation. Our work adopts a similar strategy.

5. Conclusion and Limitation

This paper introduces EgoSim, a compositional world simulator that can egocentrically explore and interact with the observed environment. To achieve this, we identify two major obstacles that prevent the creation of meaningful videos that accurately adhere to user-specified instructions and tackle them with spacetime epipolar attention and better embedding coordination. Our designed model has demonstrated unprecedented controllability and the potential uses of such an egocentric world simulator are also diverse and impactful.

Despite the impressive results demonstrated by our model, its potential has not yet been fully tapped.

More powerful pretrain weights: Firstly, although the base we used for our model is among the strongest publicly available, it is not the strongest existing base. If our method could be linked to a more powerful base model, we expect to produce more impressive and consistent videos.

Hallucination: Secondly, like all diffusion models, the hallucination problem still plagues EgoSim. We need to research further methods to address this issue.

Inverse dynamics: Finally, the most suitable application scenario for this method is inverse dynamics. We hope to test EgoSim soon for training robots, giving embodied agents the ability to plan and simulate the future through imagination.

References

Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023a.

Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S. W., Fidler, S., and Kreis, K. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 22563–22575, 2023b.

Damen, D., Doughty, H., Farinella, G. M., Fidler, S., Furnari, A., Kazakos, E., Moltisanti, D., Munro, J., Perrett, T., Price, W., et al. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 720–736, 2018.

Du, Y., Smith, C., Tewari, A., and Sitzmann, V. Learning to render novel views from wide-baseline stereo pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4970–4980, 2023.

Du, Y., Yang, S., Dai, B., Dai, H., Nachum, O., Tenenbaum, J., Schuurmans, D., and Abbeel, P. Learning universal policies via text-guided video generation. *Advances in Neural Information Processing Systems*, 36, 2024.

Guo, X., Zheng, M., Hou, L., Gao, Y., Deng, Y., Ma, C., Hu, W., Zha, Z., Huang, H., Wan, P., et al. I2v-adapter: A general image-to-video adapter for video diffusion models. *arXiv preprint arXiv:2312.16693*, 2023a.

Guo, Y., Yang, C., Rao, A., Agrawala, M., Lin, D., and Dai, B. Sparsectrl: Adding sparse controls to text-to-video diffusion models. *arXiv preprint arXiv:2311.16933*, 2023b.

Guo, Y., Yang, C., Rao, A., Wang, Y., Qiao, Y., Lin, D., and Dai, B. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. *arXiv preprint arXiv:2307.04725*, 2023c.

Ha, D. and Schmidhuber, J. World models. *arXiv preprint arXiv:1803.10122*, 2018.

He, H., Xu, Y., Guo, Y., Wetzstein, G., Dai, B., Li, H., and Yang, C. Cameractrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101*, 2024.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

Kant, Y., Wu, Z., Vasilkovsky, M., Qian, G., Ren, J., Guler, R. A., Ghanem, B., Tulyakov, S., Gilitschenski, I., and Siarohin, A. Spad: Spatially aware multiview diffusers. *arXiv preprint arXiv:2402.05235*, 2024.

Li, J., Li, D., Xiong, C., and Hoi, S. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

OpenAI. Sora. https://www.openai.com/sora, 2024. URL https://www.openai.com/sora. Sora is an AI model that can create realistic and imaginative scenes from text instructions.

Peebles, W. and Xie, S. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 4195–4205, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Ronneberger, O., Fischer, P., and Brox, T. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pp. 234–241. Springer, 2015.

Sajjadi, M. S., Meyer, H., Pot, E., Bergmann, U., Greff, K., Radwan, N., Vora, S., Lučić, M., Duckworth, D., Dosovitskiy, A., et al. Scene representation transformer: Geometry-free novel view synthesis through set-latent scene representations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6229–6238, 2022.

Sayab, M., Paananen, M., Aerden, D., Saarela, P., and Mining, D. 'structure-from-motion'and 'multi-view stereo'drone-based photogrammetry: An efficient virtual toolkit for measuring structural elements in 3-d on inaccessible outcrops. In *3rd Finnish National Colloquium of Geosciences Espoo, 15–16 March 2017*, pp. 97.

Shi, X., Chen, Z., Wang, H., Yeung, D.-Y., Wong, W.-K., and Woo, W.-c. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pp. 802–810, 2015.

Sitzmann, V., Rezchikov, S., Freeman, B., Tenenbaum, J., and Durand, F. Light field networks: Neural scene representations with single-evaluation rendering. *Advances in Neural Information Processing Systems*, 34:19313–19325, 2021.

Skorokhodov, I., Tulyakov, S., and Elhoseiny, M. Stylegan-v: A continuous video generator with the price, image quality and perks of stylegan2. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3626–3636, 2022.

Tschernezki, V., Darkhalil, A., Zhu, Z., Fouhey, D., Laina, I., Larlus, D., Damen, D., and Vedaldi, A. Epic fields: Marrying 3d geometry and video understanding. *Advances in Neural Information Processing Systems*, 36, 2024.

Unterthiner, T., van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., and Gelly, S. Fvd: A new metric for video generation. 2019.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Voleti, V., Yao, C.-H., Boss, M., Letts, A., Pankratz, D., Tochilkin, D., Laforte, C., Rombach, R., and Jampani, V. Sv3d: Novel multi-view synthesis and 3d generation from a single image using latent video diffusion. *arXiv preprint arXiv:2403.12008*, 2024.

Wang, Y., Long, M., Wang, J., Gao, Z., and Philip, S. Y. Predrnn: Recurrent neural networks for predictive learning using spatiotemporal lstms. In *Advances in Neural Information Processing Systems*, pp. 879–888, 2017.

Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

Wang, Z., Yuan, Z., Wang, X., Chen, T., Xia, M., Luo, P., and Shan, Y. Motionctrl: A unified and flexible motion controller for video generation. *arXiv preprint arXiv:2312.03641*, 2023.

Yang, M., Du, Y., Ghasemipour, K., Tompson, J., Schuurmans, D., and Abbeel, P. Learning interactive real-world simulators. *arXiv preprint arXiv:2310.06114*, 2023.

Ye, H., Zhang, J., Liu, S., Han, X., and Yang, W. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

Yu, W., Lu, Y., Easterbrook, S., and Fidler, S. Efficient and information-preserving future frame prediction and beyond. 2020.

Zhou, T., Tucker, R., Flynn, J., Fyffe, G., and Snavely, N. Stereo magnification: Learning view synthesis using multiplane images. *arXiv preprint arXiv:1805.09817*, 2018.