GEM: A Generalizable Ego-Vision Multimodal World Model for Fine-Grained Ego-Motion, Object Dynamics, and Scene Composition Control

Mariam Hassan[⋆1], Sebastian Stapf[⋆2], Ahmad Rahimi[⋆1], Pedro M B Rezende[⋆2], Yasaman Haghighi[◊1], David Brüggemann[◊3], Isinsu Katircioglu[◊3], Lin Zhang[◊3], Xiaoran Chen[◊3], Suman Saha[◊3], Marco Cannici[◊4], Elie Aljalbout[◊4], Botao Ye[◊5], Xi Wang[◊5], Aram Davtyan[2], Mathieu Salzmann[1,3], Davide Scaramuzza[4], Marc Pollefeys[5], Paolo Favaro[2], Alexandre Alahi[1]

[1] École Polytechnique Fédérale de Lausanne (EPFL), [2] University of Bern,
[3] Swiss Data Science Center, [4] University of Zurich, [5] ETH Zurich

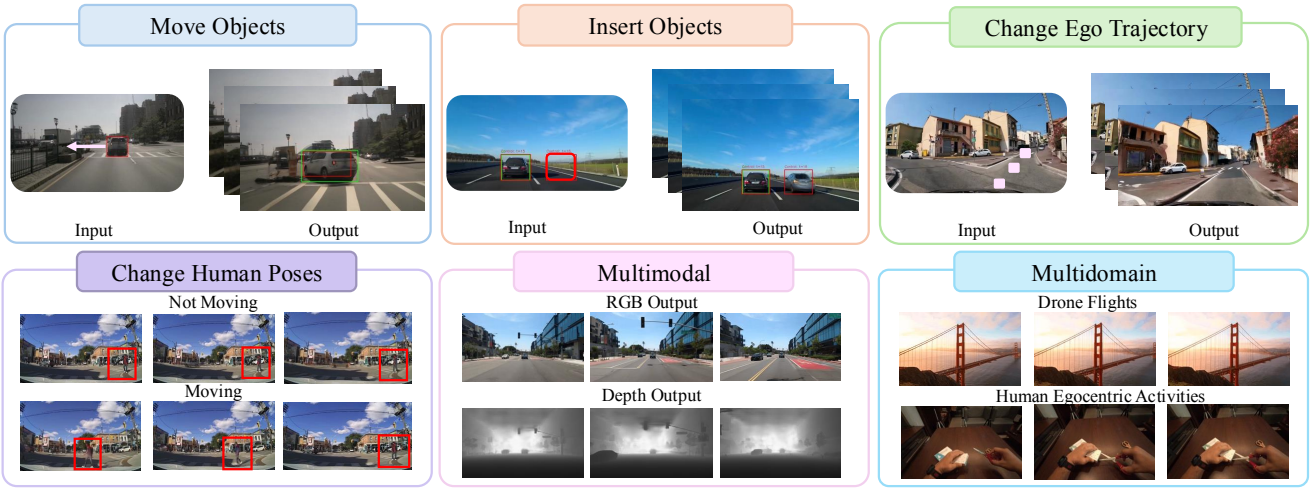

Figure 1. Overview of the capabilities of our proposed world model. GEM enables a range of features, including object manipulation (move, and insert objects), dynamic ego-trajectory adjustments, human poses changes and adaptability to multimodal outputs (i.e., images and depth maps) and multiple domains (i.e., drones and human egocentric activities). All images are generated by GEM.

Abstract

*We present **GEM**, a **G**eneralizable **E**go-vision **M**ultimodal world model that predicts future frames using a reference frame, sparse features, human poses, and ego-trajectories. Hence, our model has precise control over object dynamics, ego-agent motion and human poses. GEM generates paired RGB and depth outputs for richer spatial understanding. We introduce autoregressive noise schedules to enable stable long-horizon generations. Our dataset is comprised of 4000+ hours of multimodal data across domains like autonomous driving, egocentric human activities, and drone flights. Pseudo-labels are used to get depth maps, ego-trajectories, and human poses. We use a comprehensive evaluation framework, including a new Control of Object Manipulation (COM) metric, to assess controllability. Experiments show GEM excels at generating diverse, controllable scenarios and temporal consistency over long generations. Code, models, and datasets are fully open-sourced*[1].

1. Introduction

Different ego-vision tasks, such as autonomous driving, egocentric human activities, and drone navigation, share a common set of challenges centered around understanding and interacting with the environment from a first-person perspective. Whether it is a car moving, a drone flying, or

⋆Main Contributors, ◊ Data Contributors

[1] https://vita-epfl.github.io/GEM.github.io/

a human preparing a meal, ego-agents are inherently highly dynamic and interactive within their environment. Consequently, planning for ego-vision tasks requires understanding the dynamics and interactions occurring in respective environments, as well as understanding the effects that ego-agent's actions have on their surroundings.

World models predict plausible futures given past observations and control signals [33, 40]. By replicating the distribution of appearances and dynamics in observed visual data, they capture patterns and principles that drive interactions. This imaginative capability makes them excellent tools for decision-making in ego-vision tasks [33]. Existing egocentric world models [16, 35, 39, 41, 61, 65, 84] perform well with different controls but mainly focus on a single ego-vision task, such as autonomous driving, with domain-dependent control technique. The control in such models is primarily egocentric, *i.e.*, only capturing the motion and actions of the ego-agent. This limits the diversity of the generated scenes, making it hard to model complex interactions such as changing agents' locations in the scene. Overcoming such limitations comes with key challenges in scaling datasets, generalizing controls, and developing tailored evaluation frameworks for controllability.

To address the gaps highlighted above, we propose GEM, a **G**eneralizable **E**go-vision **M**ultimodal world model with high-fidelity controls. As summarized in Fig. 1, GEM is a multimodal and multidomain model designed to adapt to different ego-vision tasks while enabling fine-grained control over the scene in an unsupervised manner. GEM's controlling technique is threefold: 1) ego-motion control through ego trajectories, 2) scene composition control by inpainting the future content and the dynamics from a sparse set of visual tokens, and 3) a more fine-grained control over human motions through human poses. For scene composition control, we use sparse visual tokens extracted by the DINOv2 [44] encoder augmented with unique object identification codes to enable precise control over the motion and appearance of all objects in the scene. We support the insertion of entirely new objects, and allow for highly flexible and controllable future prediction. Alongside the frames, GEM is capable of generating depth providing rich spatial context. To achieve that, we pseudo-label our data with depth maps, ego-trajectories and human poses.

GEM is trained on a large corpus of open-source datasets, with contributions in both methods and datasets. The methods are summarized as follows:

- We present GEM, a generalizable world model that predicts future frames given a reference, sparse DINOv2 features, human pose, and ego-trajectories. It enables control over ego motion, object dynamics, and human poses.
- We introduce autoregressive noise schedules to our framework, enabling stable long-horizon generations.
- We train on autonomous driving domain and explore multimodal and multidomain generation by (1) integrating depth as an extra generation modality; and (2) fine-tuning our model on different ego-vision domains, *i.e.*, human ego activities, and drone navigations.
- We present a comprehensive evaluation of GEM's controllability and introduce a metric, Control of Object Manipulation (COM) to evaluate control of object motion.

To address limitations in scale and diversity of existing open-source datasets, we propose the following dataset-related contributions:

- We utilize a large-scale open-source corpus with over 3200 hours of driving videos, 1000 hours of egocentric human activity datasets, and 27.4 hours of self-collected drone footage from YouTube. The driving datasets are further curated for diverse interactions and dynamics.
- Given scarcity of labels, we implement pseudo-labeling approaches to generate depth maps, ego-trajectories, and human pose annotations. We show the effectiveness of our control strategy given pseudo-labels.

Our work is fully open-source, sharing the curated datasets, codebase, and models[2].

2. Related Work

We briefly review the previous works on controllable video generation models and world models including autonomous driving and egocentric human activity world models.

Controllable Video Generation. Recent advancements in video generation models have enabled realistic, high-quality video rendering. Several pioneering models leverage Large Language Models (LLMs) for text-to-video generation [42, 70]. Since the success of diffusion models [15, 51], diffusion-based video generation has become prominent. Methods can be categorized as: text-to-video [12, 20, 28, 29, 53, 60, 62] or image-to-video, [4, 11, 80]. Diffusion models adapt to various control inputs like text, edge maps, and depth maps [78]; they also offer superior realism [4]. However, generic video generation models are not trained to encode the intricate dynamics of egocentric environments [71], and many do not offer detailed motion controls over the generations.

World Models. World models are large-scale generative models that infer dynamics and predict plausible futures based on past observations [33, 40, 74]. They are valuable in many tasks such as real-world simulations [74, 87], reinforcement learning [2, 21, 24, 45, 66], model-predictive control [22, 23], and representation learning [25, 43].

Autonomous Driving World Models. World models for autonomous driving represent the world using sensor obser-

[2] https://vita-epfl.github.io/GEM.github.io/

vations, such as lidar-generated point clouds [6, 76, 79, 85], with limited datasets often constraining their scale, or images [16, 30, 39, 41, 61, 65, 71, 84]. Recent visual world models use LLMs as backbones [35, 67, 84], but these models rely heavily on LLMs' spatial reasoning, which remains limited [34, 47, 83]. This makes them better suited for high-level scene control, like weather or lighting adjustments, rather than precise motion control [16]. Diffusion-based models, in contrast, use low-level controls like ego-trajectories and maps [16, 41, 61, 65, 71, 84], but focus primarily on ego-centric control, limiting their ability to generate complex scenarios such as controlling over any other motions in the scene. Additionally, efforts to improve multimodal world models for spatial metric understanding [6] rely on limited simulation-based point cloud datasets, which are difficult to generalize to real-world data.

Egocentric Human Activities World Models. Recent large-scale egocentric video datasets (e.g., Ego4D [18] and Ego-Exo4D [19]) have advanced human egocentric vision. However, research on comprehensive world models for this domain remains limited. To the best of our knowledge, UniSim [74] is the first approach in this direction, using a video diffusion model conditioned on action labels. In contrast, our approach provides higher-fidelity control allowing for a greater diversity in the generated content.

3. Uncovering the Real GEM

In this section, we present GEM's key components and capabilities. As shown in Fig. 2, GEM has two output modalities—images and depth—and three control signals: ego-trajectories, DINOv2 features, and human poses. We begin with background (Sec. 3.1), detail our control methodology (Sec. 3.2), long-horizon generation (Sec. 3.3), multimodal generation (Sec. 3.4), and training strategy (Sec. 3.5).

3.1. Preliminaries

We cast the training of our world model as video generation. Thus, we employ the current SotA open-source image-to-video model, Stable Video Diffusion (SVD) [4], as backbone for GEM and fine-tune it on ego-centric data. In SVD, videos are represented as sequences of N RGB frames of size $H \times W$. The frames are independently encoded into the latent space of a pre-trained autoencoder, resulting in a sequence of N feature maps, each having 4 channels, height $\tilde{H} = \frac{H}{8}$, and width $\tilde{W} = \frac{W}{8}$. We denote the distribution of the encoded videos from the dataset as $p_{\text{data}}(x)$, where $x \in \mathbb{R}^{N \times 4 \times \tilde{H} \times \tilde{W}}$. SVD operates within the Elucidated Diffusion Model (EDM) framework [36]. In this framework, a network $D_\theta(x; \sigma, \mathcal{C})$ is trained to denoise a noisy sample x given the noise level σ and conditioning variables $\mathcal{C}$. These conditioning variables may include embeddings from text, images, or videos. In the case of SVD, $\mathcal{C} = \{x_0\}$ only includes the embedding of the first frame in the sequence, enabling image-to-video synthesis.

3.2. Controlling Ego-Vision Generation

We decompose the control space of our model into three main components: 1) the ego-motion, 2) the object-level control, and 3) human pose control. The first component, Sec. 3.2.1, allows us to specify the motion of the ego-agent through ego-trajectories. The second component, Sec. 3.2.2, facilitates object-specific control, enabling editing of scene composition and dynamics across space and time by adjusting the location of object features. This also enables insertion of new objects. The last component, Sec. 3.2.3, enables control of pedestrian poses.

3.2.1. Ego-Motion Control

To control the ego-motion, we expand the set of conditioning variables in $D_\theta(x; \sigma, \mathcal{C})$ with the ego-trajectories c_{traj}, *i.e.*, $\mathcal{C} = \{x_0, c_{\text{traj}}\}$. Our ego-trajectories are metric sequences of 2D positions that quantify the motion of the ego-agent when projected to the birds-eye-view plane. Inspired by Vista [16], we integrate c_{traj} into the network by first flattening and then applying Fourier embeddings [54]. Since the ego-motion control provides solely a global context and does not encode direct spatial information in the image space, we condition the network on c_{traj} by fusing them through additional LoRA modules [31] in the cross-attention layers of the UNet backbone (see Fig. 2).

3.2.2. Object-Level Control

For object-level control, we leverage DINOv2 tokens [44], following previous work [14]. DINOv2 tokens are data-agnostic, abstract, and inexpensive to obtain, making them well-suited as conditioning signals. They encode high-level semantic information about objects in the scene (*e.g.*, object categories or basic style features) and are invariant to small appearance perturbations, ensuring their semantic consistency across frames. We condition the model D_θ on a sparse set of DINOv2 tokens that specify where and when a certain object should appear in the generated video, as illustrated in Fig. 2. The model generates the output by inpainting the missing information both spatially and temporally. More precisely, the object-level control tokens (DINOv2 tokens) extracted from either unrelated images or the reference frame x_0, are inserted into zero-initialized feature maps $\{z_0, \ldots, z_N\}, z_i \in \mathbb{R}^{N \times d \times h \times w}$ at the desired locations and times. Here, N is the number of frames, d is the dimension of each DINOv2 feature, and $h \times w$ is its resolution. These feature maps form the object-level control c_{dino} that is fed to the model, conditioning it to place the objects at specific coordinates and time steps. Thus, the set of conditioning variables in D_θ becomes $\mathcal{C} = \{x_0, c_{\text{traj}}, c_{\text{dino}}\}$.

The training of the object-level control is done in an unsupervised manner. During training, we randomly sample

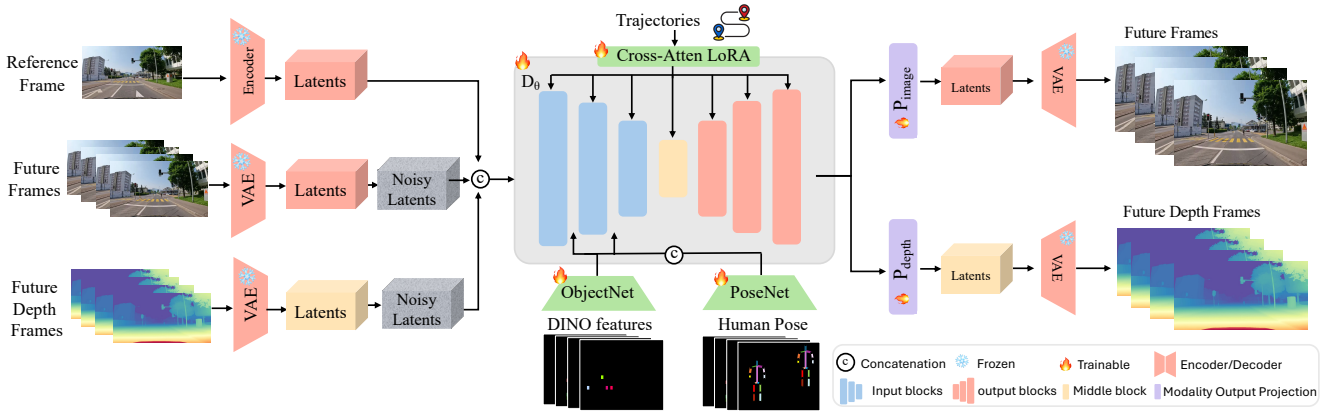

Figure 2. GEM generates two modalities by taking as inputs a reference frame and noisy latents of images and depth modalities. The denoiser network, D_θ is conditioned on ego trajectories, DINOv2 features and human poses. Ego-trajectories are added using a cross attention LoRA at every block of the network. DINOv2 features and human poses are added to the output of each block in the input layers of the denoiser. To handle multimodal outputs, we use different output convolution-based projection layers P.

k frames $\{x_{t_1}, \ldots, x_{t_k}\}$ from a given video $x \sim p_{\text{data}}(x)$. We then process the original frames with DINOv2, and extract the corresponding dense feature maps $\{z_{t_1}, \ldots, z_{t_k}\}$, where $z_{t_i} \in \mathbb{R}^{d \times h \times w}$. From each of these feature maps, we randomly mask all but $m \sim U[0, M]$ tokens, where M is a hyperparameter that we set to 32 in our experiments. The masked feature maps are then padded with zeros to match the original frame count. We thus obtain $c_{\text{dino}} = \{z_{t_1}^{\text{masked}}, \ldots, z_{t_k}^{\text{masked}}\}^{\text{pad}}$. By employing this randomized approach, we learn both the spatial composition and the temporal dynamics of the scene.

Identity Embeddings. One challenge in the object-level control using DINOv2 features arises when the inserted tokens are both spatially and feature-wise similar to the visual features of objects already present in the reference frame. This creates an ambiguity between moving an existing object or inserting a new one. To address this, we propose using a learned identity embedding to associate individual tokens over time. This approach, visualized in Fig. 3, involves adding the same identity embedding to the control tokens representing the same moving entity across different time steps. More specifically, we start with $\{z_{t_1}^{\text{masked}}, \ldots, z_{t_k}^{\text{masked}}\}$, as before, and add individual learned identity embeddings $\text{ID}_\phi : \{1, \ldots L\} \rightarrow \mathbb{R}^d$ to the nonzero tokens in each map. Here L is chosen to be large enough to ensure that different tokens from the same feature map do not receive the same identity embedding. For each feature map, we then sample a target time $\tau_i > t_i$ and translate the tokens from $z_{t_i}^{\text{masked+ID}}$ to z_{τ_i} using the optical flow between frames x_{t_i} and x_{τ_i}, as shown in Fig. 3. During inference, we disambiguate the generation by using the same identity embeddings in the reference and in the target frames to guide the model towards moving the underlying object instead of introducing a new object at the desired location.

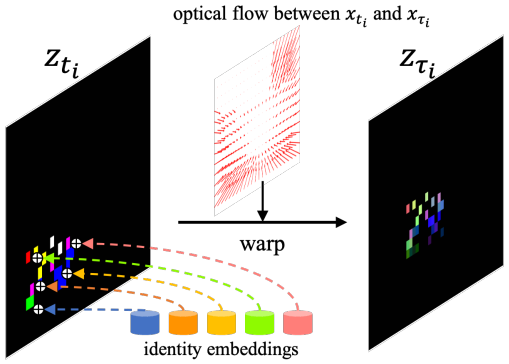

Figure 3. During training the sparse DINOv2 features from frame t_i are translated to frame τ_i using the corresponding optical flow.

Conditioning Technique. In contrast to the ego-motion, the object-level control encodes fine-grained details about the scene composition. This influences the design of the conditioning technique as it now requires incorporating spatial information. We start by processing the sequence of sparse DINOv2 feature maps using a network with a similar architecture to the input blocks of the denoising UNet. We call this network *ObjectNet*. ObjectNet is meant to capture and inpaint both the spatial and the temporal information in the sparse DINOv2 feature maps. Following prior work [81], the encoded tokens are directly added to the outputs of the UNet's input blocks, as depicted in Fig. 2. Empirically, we observe that this technique for object-level control outperforms feeding the DINOv2 tokens via cross-attention layers, as done for the ego-motion control. Moreover, ObjectNet serves as a transition layer, helping to align the representations from DINOv2 with the UNet's internal feature space, and performs better than directly fusing raw

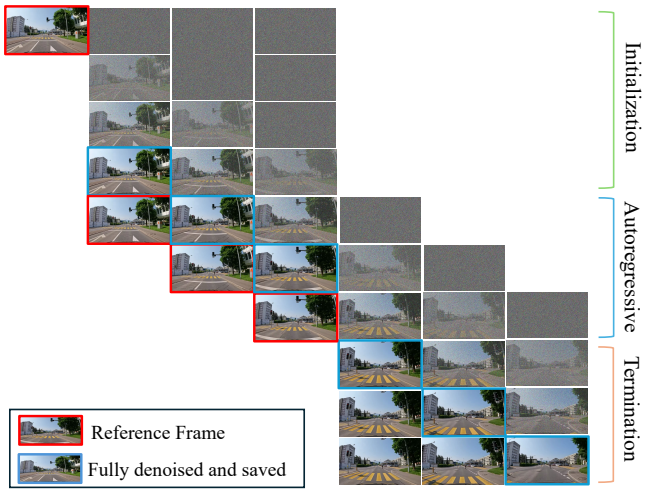

Figure 4. Visualization of our dynamic autoregressive sampling noise schedule for denoising 6 frames in total with a window size of 3 frames and 3 sampling steps.

DINOv2 feature maps into the denoiser.

3.2.3. Human Pose Control

We find that the aforementioned object-level control performs well for objects with few moving parts. However, accurately modeling humans remains challenging, yet crucial for safe navigation and effective human-robot interaction. To address this, we extend the object-level control with a human pose component, *i.e.*, $\mathcal{C} = \{x_0, c_{\text{traj}}, c_{\text{dino}}, c_{\text{pose}}\}$. To condition the model D_θ on the extracted human poses, we follow prior techniques for generating human motion [81]. We render the skeletons onto an empty image plane and pass the image through PoseNet, a small CNN trained from scratch. The resulting human pose feature maps are then added to the network features of D_θ, in a manner similar to the object-level controls, as shown in Fig. 2.

3.3. Stable Long Video Generation

Generating long videos beyond the training horizon is a challenging task for diffusion models [10, 13, 68, 82]. A simple approach is to generate sequential short clips with overlapping frames. However, this causes temporal discontinuities and abrupt scene changes [68]. Inspired by recent works [10, 68], we introduce progressive denoising and autoregressive sampling, using a per-frame noise schedule to reinforce causal relationships between consecutive frames.

Autoregressive Sampling. The goal of sampling is to autoregressively denoise all frames over a long horizon. To this end, we adopt a dynamic per-frame noise schedule, as illustrated in Fig. 4. The schedule has three phases: initialization, autoregressive, and termination. Initially, the schedule controls the noise levels of each frame so that the denoising of frame i only starts after the denoising of frame $i-1$ has been initiated. This enables the denoising of each frame to benefit from some cleaner information in its preceding frames. Once a frame is fully denoised, it is saved and replaces the current reference frame. At this point, the autoregressive phase starts where at each step, a fully denoised frame is removed and a new noisy frame is appended. This process continues until only N frames still need to be denoised, signaling the start of the termination stage. At this stage, no new frames are added; the fully denoised frames are saved. Refer to supplementary material in Section 8.2 for the full algorithm in detail.

Training Noise Schedule. To support the inference with the proposed custom noise schedule, we modify the training noise distribution in the following way. We first sample a random noise level $\log(\sigma) \sim \mathcal{N}(p_{\text{mean}}, p_{\text{std}})$. Using SVD's noise-to-time step mapping, we compute the corresponding denoising time step, $t_{\text{intercept}}$. Next, we sample a random shift $t_{\text{shift}} \sim \text{Beta}(\alpha, \beta)$, where α and β are selected to favor lower shift values. The per-frame time steps are then calculated as $t_{\text{intercept}} - (\frac{i}{N-1} - t_{\text{shift}})$ for $i \in \{0, \dots, N-1\}$, ensuring a consistent noise increase over the frame axis. To add variability, we add small random noise to the time steps, which are subsequently converted back to σ values. This approach keeps the essential information within the attention window for the autoregressive component.

3.4. Multimodal Generation

We incorporate depth as an additional generated modality to leverage its rich spatial information, proven to enhance tasks such as scene perception, planning, object localization, and more [1, 63]. By generating depth alongside RGB images, GEM can generate spatial information alongside the structural context of the scene. To encode and decode depth, we use the same VAE used for images, following [32], which shows that the pretrained VAE of SVD has negligible reconstruction error on depth images. We concatenate both modalities at the input and introduce an output convolution projection layer (P_{depth}) to the denoising network to predict the noise for the depth. D_θ simultaneously denoises both inputs ensuring consistency between both modalities. The final denoiser becomes $D_\theta(x, x_{\text{depth}}; \sigma, \{x_0, c_{\text{traj}}, c_{\text{dino}}, c_{\text{pose}}\})$.

3.5. Training Strategy

For efficiency, we divide our training into two stages. The first stage focuses on learning new control signals, while the second one emphasizes high-resolution generation. We begin with the pre-trained SVD [4] and initially fine-tune it on low-resolution videos (320×576) using all added control signals and modalities. In the second stage, the training continues with the same configuration, but at a higher res-

Domain	Dataset	Hours	Front-View Frames	Diversity Cities	GT Traj
Driving	OpenDV [71]	1747	60.2M	>244	
	BDD [69]	1000	100M	1	
	Nuscenes [8]	5.5	241k	2	✓
	Driving Dojo [64]	150	-	9	✓
	Honda HDD [49]	104	1.1M	1	
	Honda HAD [38]	30	-	1	
	Drive360 [26]	55	-	-	
	D^2City [9]	100	700k	5	
	DoTA✳ [77]	13.9	504k	-	
	CarCrashDataset✳ [3]	6.25	-	-	
	Total	3211	>162.4M	>244	
Human	EgoExo4D [19]	1000	108M	-	✓
Drone	self-collected	27.4	961k	-	

Table 1. Overview of ego video datasets used for training, totaling 4238 hours of training data. ✳ denotes accident-focused datasets.

olution (576×1024). As detailed in Sec. 4, we apply data filtering to improve diversity and quality during both stages. Please refer to the supplementary material Sec. 9 for more details about our training strategy.

4. Dataset Preparation

We combine various open source datasets across different domains presented in Tab. 1. In total, we use 3211 hours of driving, 1000 hours of human egocentric videos, and 27.4 hours of drone footage collected from YouTube.

Data Curation. To achieve precise control over object movements, the training data must include (1) diverse interactions and dynamics, (2) fine-grained object details. We curate the dataset by removing low-quality and low-motion sequences, segmenting videos into 2.5-second clips, and applying two types of filters: *quality* and *diversity*. Quality filtering excludes clips with poor camera quality or high blur using aesthetic scores from the LAION dataset [52] and PIQE metrics [59], similar to [48]. Diversity filtering assesses motion diversity via optical flow, similar to [17, 48], and semantic variation using DINO feature encodings [44]. Clips with low intra-clip diversity or high cross-clip similarity are excluded to balance motion and content. Additional details can be found in the in the supplementary Sec. 8.3.

Pseudo-labeling. Given scarcity of labeled datasets, we pseudo-label all the data with depth, ego-trajectories, and human skeletons. We generate metric depth using Depth Anything V2 [73] for trajectory labeling and geometric understanding. We estimate ego-trajectories with GeoCalib [58] for intrinsics followed by DroidSLAM [56] for RGB-D SLAM, with pseudo-depth resolving scale ambiguity. Finally, we extract human skeletons using DWPose [75] for efficient and high-quality pose estimation. Refer to the supplementary material Sec. 8.1 for more details.

5. Experiments

In this section, we conduct experiments to evaluate our model based on quality and controllability. We start by introducing our metrics in Sec. 5.1. Then, we outline our experiments for quality evaluation followed by controllability evaluation. We conduct the experiments on Nuscenes' validation set and a randomly sampled subset of equal number of videos from the OpenDV's validation set. We use Vista [16] as a baseline, as its architecture aligns closely with ours, making it the most comparable model to GEM. We additionally show qualitative results in Fig. 6. and more ablation studies in supplementary material Sec. 11.

5.1. Evaluation Metrics

In this section, we outline the metrics utilized to assess various aspects of our world model including the visual quality and controllability of the generations.

Video Quality. We evaluate the generation quality of our world model using standard metrics: Frechet Inception Distance (FID) [27] and Frechet Video Distance (FVD) [57].

Ego Motion. To evaluate the ego-motion control, we estimate the ego-motion, $\hat{p}$, from the generated videos using our pseudo-labeling technique (Sec. 4). We use the Average Displacement Error (ADE) to compare the generated trajectory against the ground truth trajectories. If a dataset has no ground truth labels, we estimate the pseudo-ground truth using the same pipeline (Sec. 4).

Control of Object Manipulation (COM). To evaluate the object-level control, we use YOLOv11 [37] to detect and track the objects through frames. We extract the bounding boxes of the largest vehicle in the scene and compare its bounding boxes across the frames in the generated and ground-truth videos. We then calculate the absolute difference in pixels between the centers of the bounding boxes as follows: $\text{COM} = |\text{BBox}_{\text{gen}} - \text{BBox}_{\text{GT}}|$.

Human Pose. For human poses, we select video clips that contain at least five pedestrians and extract their joints using DWPose [75]. We use the COCO toolkit to evaluate based on 17 keypoints; we calculate Average Precision (AP) of poses extracted from the generated videos against those from the ground truth videos. We compare AP of poses in unconditional versus conditional generation to evaluate the effectiveness of the control technique.

Depth Evaluation. We compare our generations to the pseudo-labels. The evaluation is based on two commonly used metrics [32, 72, 73]: Absolute Relative Error (AbsRel) defined as $\frac{|\hat{d}-d|}{d}$, and δ defined as percentage of

	Nuscenes		OpenDV	
	FID ↓	FVD ↓	FID ↓	FVD ↓
DriveGAN [39]	73.4	502.3	-	-
DriveDreamer [61]	14.9	340.8	-	-
DriveDreamer-2 [84]	25.0	105.1	-	-
WoVoGen [41]	27.6	417.7	-	-
Drive-WM [65]	15.8	122.7	-	-
GenAD [71]	15.4	184.0	-	-
Vista [16]	6.6⋆	167.7⋆	5.5⋆	163⋆
GEM (Ours)	10.5	158.5	6.3	131

⋆ our reproduced results using the official code from Vista [16].

Table 2. Quality comparison of generations using FID and FVD on Nuscenes and OpenDV datasets. While GEM shows slightly worse FID results than Vista [16], GEM outperforms it on FVD.

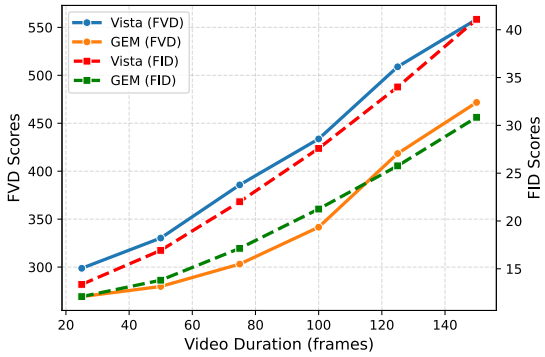

Figure 5. FVD and FID comparisons of generations across different video duration for GEM and Vista [16]. While the generations of both models degrade with longer videos, GEM shows better visual quality in long generations.

$\max\left(\frac{d}{\hat{d}}, \frac{\hat{d}}{d}\right) < 1.25$. Refer to Sec. 10 in the supplementary material for the results.

5.2. Comparisons of Generation Quality

Training-Horizon Generation Quality. Tab. 2 compares GEM's generation quality with existing autonomous driving world models based on the standard quality metrics, FID and FVD. As shown in Tab. 2, GEM outperforms Vista on FVD results for both datasets and achieves competitive FID results for OpenDV. While GEM's FID results on Nuscenes are marginally lower than Vista's, it is likely due to Vista's fine-tuning on Nuscenes during the final training stage. Notably, GEM achieves these comparable results despite being trained on a dataset curated specifically to enhance controllability performance rather than visual quality.

Long-Horizon Generation Quality. We quantitatively evaluate long-generation quality by randomly generating 500 videos of 150 frames using GEM's autoregressive sampler and Vista's triangular sampler (window size: 25 frames, 3-frame overlap). Due to the computational cost, we limited the experiment to 500 randomly generated videos. We calculate FID and FVD on video lengths of 25, 50, 75, 100, 125, and 150 frames. Results (Fig. 5) show GEM outperforming Vista, with consistently lower FVD and FID scores across all durations, reflecting better temporal consistency and quality. This demonstrates GEM's superior sampling approach for long-horizon videos.

5.3. Human Evaluation

To address the limitations of FVD in evaluating video perceptual quality, particularly for dynamic scenes [5, 7], we conduct a human evaluation. Participants performed pairwise comparisons between GEM and Vista's unconditional video generations. Using a random selector, we sampled 50 2.5s and 50 15s videos from each model, along with 50 videos of each length from the OpenDV validation set. Each participant evaluated 20 randomly selected videos, equally split between short and long generations, focusing on realistic dynamics, visual quality, and temporal consistency. Results, based on 116 responses, are shown in Fig. 7.

For short videos, most participants found minimal differences between GEM and Vista, with 76, 48, and 60 noting no distinction in realism, visual quality, and temporal consistency, respectively. This suggests that short generations are generally perceived as highly similar.

For long videos, GEM was strongly preferred: 75 vs. 23 votes for realism, 79 vs. 29 for visual quality, and 81 vs. 27 for temporal consistency. These results indicate that GEM's long generations are perceived as superior in realism, visual quality, and temporal consistency as per human judgment.

5.4. Comparisons of Controllability

In this section, we evaluate controllability of ego motion, object motion and human pose. We compare our conditional generation against the unconditional one to evaluate the effectiveness of our control strategy.

Ego-Motion Tab. 3 reports the ADE results of generated trajectories for the Nuscenes dataset and OpenDV. Results on Nuscenes show slight improvement in our controlled generation. This is due to the fact that in most cases, the ego-motion is obvious in the 2.5s window (*e.g.*, moving straight). Therefore, we manually choose a subset of 50 videos, Nuscenes^{sub}, where ambiguity is present (*e.g.*, multiple paths ahead). Results on Nuscenes^{sub} in Tab. 3 show 18% improvement of the conditional generation, demonstrating effective ego-motion controllability. We additionally provide qualitative examples on our website.

Object Motion Tab. 3 reports the results of object manipulation controllability based COM introduced in Sec. 5.1.

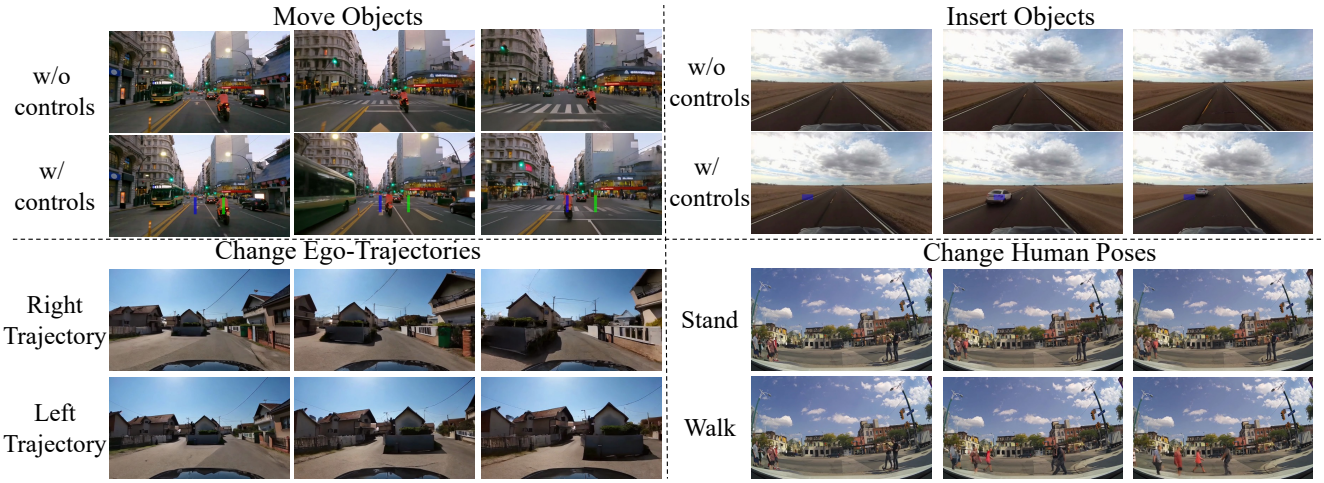

Figure 6. Qualitative results for GEM's controllability. GEM can flexibly move objects (top-left), insert new objects (top-right), change ego trajectories (bottom-left) and change human poses (bottom-right). Refer to our website for more videos.

	Ego-Motion Controllability			Object Motion Controllability		Human Pose Controllability	
	Nuscenes	Nuscenes^{sub}	OpendDV	Nuscenes	OpenDV	OpenDV	
	ADE ↓	ADE ↓	ADE ↓	COM ↓	COM ↓	AP@IoU=.5 ↑ area=all	AP@IoU=.5:.95 ↑ area=large
GEM *w/o controls*	3.24	3.59	5.39	38.8	55.2	0.00	0.00
GEM *w/ controls*	**3.07**	**2.85**	**3.47**	**12.2**	**11.5**	**0.12**	**0.12**

Table 3. Controllability evaluation of ego-motion, object motion and human pose. We compare the metrics with unconditional generations of GEM and consistently observe better performance with controls, indicating that GEM accurately follows the given control signals. Nuscenes^{sub} denotes the subset of Nuscenes with ego-motion ambiguity, where we observe a higher ADE gap.

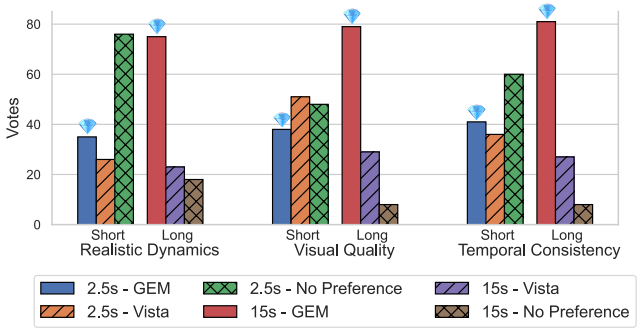

Figure 7. Human evaluation on short (2.5s) and long (15s) videos. The gem denotes GEM. While results show no preference on short videos, GEM significantly outperforms Vista [16] on long videos.

The results show that our conditional generation consistently outperforms the unconditional one by 68.8% on Nuscenes and 79% on OpenDV. This highlights the effectiveness of our control strategy for moving objects.

Human Pose We report two types of AP metrics: AP based on (1) loose Intersection over Union (IoU) of at least 50% on all sizes of skeletons, (2) strict IoU (50% to 95%) on large skeletons. Tab. 3 presents the results, highlighting the control effectiveness particularly for large human poses.

6. Conclusion

We introduced GEM, a multimodal world model for ego-vision tasks, capable of generating videos in environments with complex dynamics. By leveraging egocentric trajectories, DINO features, and human skeletons, GEM enables precise control over ego-motion, object movements, as well as humans. Its multimodal features, including image and depth frame generation, provide both rich semantic and spatial context. Our evaluation results have highlighted GEM's effectiveness, with conditional generation significantly surpassing unconditional generation for ego trajectories, object motion, and modeling humans. While GEM advances the state of the art in controllable ego-vision world models, it is not without limitations. Notably, while GEM demonstrates strong performance in long generations, further improvements are needed to enhance their quality and consistency over extended sequences. Despite these limitations, we hope GEM will serve in the future as a foundation for adaptable and controllable world models.

Acknowledgments. This work was supported as part of the Swiss AI Initiative by a grant from the Swiss National Supercomputing Centre (CSCS) under project ID a03 on Alps. Sebastian Stapf, Aram Davtyan, and Yasaman Haghighi have been supported by SNSF Projects 200020-188690, 200021-228098, and 10003100. Ahmad Rahimi has been supported by Hasler Foundation under the Responsible AI program. This work was also supported by the Swiss National Science Foundation Advanced Grant 216260: Beyond Frozen Worlds: Capturing Functional 3D Digital Twins from the Real World

References

[1] Markus Achtelik, Abraham Bachrach, Ruijie He, Samuel Prentice, and Nicholas Roy. Stereo vision and laser odometry for autonomous helicopters in gps-denied indoor environments. In *Unmanned Systems Technology XI*, pages 336–345. SPIE, 2009. 5

[2] Elie Aljalbout, Nikolaos Sotirakis, Patrick van der Smagt, Maximilian Karl, and Nutan Chen. Limt: Language-informed multi-task visual world models. *arXiv preprint arXiv:2407.13466*, 2024. 2

[3] Wentao Bao, Qi Yu, and Yu Kong. Uncertainty-based traffic accident anticipation with spatio-temporal relational learning. In *Proceedings of the 28th ACM International Conference on Multimedia*, pages 2682–2690, 2020. 6

[4] A. Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, and Dominik Lorenz. Stable video diffusion: Scaling latent video diffusion models to large datasets. *ArXiv*, abs/2311.15127, 2023. 2, 3, 5, 4

[5] Andreas Blattmann, Robin Rombach, Huan Ling, Tim Dockhorn, Seung Wook Kim, Sanja Fidler, and Karsten Kreis. Align your latents: High-resolution video synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22563–22575, 2023. 7

[6] Daniel Bogdoll, Yitian Yang, Tim Joseph, and J. Marius Zöllner. Muvo: A multimodal world model with spatial representations for autonomous driving, 2024. 3

[7] Tim Brooks, Janne Hellsten, Miika Aittala, Ting-Chun Wang, Timo Aila, Jaakko Lehtinen, Ming-Yu Liu, Alexei Efros, and Tero Karras. Generating long videos of dynamic scenes. *Advances in Neural Information Processing Systems*, 35:31769–31781, 2022. 7

[8] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 6

[9] Zhengping Che, Guangyu Li, Tracy Li, Bo Jiang, Xuefeng Shi, Xinsheng Zhang, Ying Lu, Guobin Wu, Yan Liu, and Jieping Ye. D2-city: a large-scale dashcam video dataset of diverse traffic scenarios. *arXiv preprint arXiv:1904.01975*, 2019. 6

[10] Boyuan Chen, Diego Marti Monso, Yilun Du, Max Simchowitz, Russ Tedrake, and Vincent Sitzmann. Diffusion forcing: Next-token prediction meets full-sequence diffusion. *Advances in Neural Information Processing Systems*, 2024. 5

[11] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter1: Open diffusion models for high-quality video generation, 2023. 2

[12] Haoxin Chen, Yong Zhang, Xiaodong Cun, Menghan Xia, Xintao Wang, Chao Weng, and Ying Shan. Videocrafter2: Overcoming data limitations for high-quality video diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7310–7320, 2024. 2

[13] Xinyuan Chen, Yaohui Wang, Lingjun Zhang, Shaobin Zhuang, Xin Ma, Jiashuo Yu, Yali Wang, Dahua Lin, Yu Qiao, and Ziwei Liu. Seine: Short-to-long video diffusion model for generative transition and prediction. In *The Twelfth International Conference on Learning Representations*, 2024. 5

[14] Aram Davtyan, Sepehr Sameni, Bjorn Ommer, and Paolo Favaro. Enabling visual composition and animation in unsupervised video generation. *ArXiv*, abs/2403.14368, 2024. 3, 4

[15] Prafulla Dhariwal and Alex Nichol. Diffusion models beat gans on image synthesis, 2021. 2

[16] Shenyuan Gao, Jiazhi Yang, Li Chen, Kashyap Chitta, Yihang Qiu, Andreas Geiger, Jun Zhang, and Hongyang Li. Vista: A generalizable driving world model with high fidelity and versatile controllability. *Advances in Neural Information Processing Systems*, 2024. 2, 3, 6, 7, 8

[17] Rohit Girdhar, Mannat Singh, Andrew Brown, Quentin Duval, Samaneh Azadi, Sai Saketh Rambhatla, Akbar Shah, Xi Yin, Devi Parikh, and Ishan Misra. Emu video: Factorizing text-to-video generation by explicit image conditioning. *ArXiv*, abs/2311.10709, 2023. 6

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

[19] Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, et al. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400, 2024. 3, 6

[20] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning, 2024. 2

[21] Danijar Hafner, Timothy Lillicrap, Jimmy Ba, and Mohammad Norouzi. Dream to control: Learning behaviors by la-

tent imagination. In *International Conference on Learning Representations*, 2020. 2

[22] Nicklas Hansen, Jyothir SV, Vlad Sobal, Yann LeCun, Xiaolong Wang, and Hao Su. Hierarchical world models as visual whole-body humanoid controllers. *arXiv preprint arXiv:2405.18418*, 2024. 2

[23] Nicklas A Hansen, Hao Su, and Xiaolong Wang. Temporal difference learning for model predictive control. In *International Conference on Machine Learning*, pages 8387–8406. PMLR, 2022. 2

[24] Shibo Hao, Yi Gu, Haodi Ma, Joshua Hong, Zhen Wang, Daisy Zhe Wang, and Zhiting Hu. Reasoning with language model is planning with world model. In *NeurIPS 2023 Workshop on Generalization in Planning*, 2023. 2

[25] Haoran He, Chenjia Bai, Ling Pan, Weinan Zhang, Bin Zhao, and Xuelong Li. Large-scale actionless video pre-training via discrete diffusion for efficient policy learning, 2024. 2

[26] Simon Hecker, Dengxin Dai, and Luc Van Gool. End-to-end learning of driving models with surround-view cameras and route planners. In *Proceedings of the european conference on computer vision (eccv)*, pages 435–453, 2018. 6

[27] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. In *Neural Information Processing Systems*, 2017. 6

[28] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. Imagen video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303*, 2022. 2

[29] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J. Fleet. Video diffusion models, 2022. 2

[30] Anthony Hu, Lloyd Russell, Hudson Yeo, Zak Murez, George Fedoseev, Alex Kendall, Jamie Shotton, and Gianluca Corrado. Gaia-1: A generative world model for autonomous driving, 2023. 3

[31] J. Edward Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *ArXiv*, abs/2106.09685, 2021. 3

[32] Wenbo Hu, Xiangjun Gao, Xiaoyu Li, Sijie Zhao, Xiaodong Cun, Yong Zhang, Long Quan, and Ying Shan. Depthcrafter: Generating consistent long depth sequences for open-world videos. *arXiv preprint arXiv:2409.02095*, 2024. 5, 6

[33] Yafei Hu, Quanting Xie, Vidhi Jain, Jonathan Francis, Jay Patrikar, Nikhil Keetha, Seungchan Kim, Yaqi Xie, Tianyi Zhang, Shibo Zhao, Yu Quan Chong, Chen Wang, Katia Sycara, Matthew Johnson-Roberson, Dhruv Batra, Xiaolong Wang, Sebastian Scherer, Zsolt Kira, Fei Xia, and Yonatan Bisk. Toward general-purpose robots via foundation models: A survey and meta-analysis, 2023. 2

[34] Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023. 3

[35] Fan Jia, Weixin Mao, Yingfei Liu, Yucheng Zhao, Yuqing Wen, Chi Zhang, Xiangyu Zhang, and Tiancai Wang. Adriver-i: A general world model for autonomous driving. *arXiv preprint arXiv:2311.13549*, 2023. 2, 3

[36] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *ArXiv*, abs/2206.00364, 2022. 3, 4

[37] Rahima Khanam and Muhammad Hussain. Yolov11: An overview of the key architectural enhancements. *arXiv preprint arXiv:2410.17725*, 2024. 6

[38] Jinkyu Kim, Teruhisa Misu, Yi-Ting Chen, Ashish Tawari, and John Canny. Grounding human-to-vehicle advice for self-driving vehicles. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. 6

[39] Seung Wook Kim, Jonah Philion, Antonio Torralba, and Sanja Fidler. Drivegan: Towards a controllable high-quality neural simulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5820–5829, 2021. 2, 3, 7

[40] Yann LeCun. A path towards autonomous machine intelligence. 2

[41] Jiachen Lu, Ze Huang, Zeyu Yang, Jiahui Zhang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. *arXiv preprint arXiv:2312.02934*, 2023. 2, 3, 7

[42] Zeyu Lu, Zidong Wang, Di Huang, Chengyue Wu, Xihui Liu, Wanli Ouyang, and Lei Bai. Fit: Flexible vision transformer for diffusion model, 2024. 2

[43] Russell Mendonca, Shikhar Bahl, and Deepak Pathak. Structured world models from human videos, 2023. 2

[44] Maxime Oquab, Timoth'ee Darcet, Théo Moutakanni, Huy Q. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russ Howes, Po-Yao (Bernie) Huang, Shang-Wen Li, Ishan Misra, Michael G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Huijiao Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *ArXiv*, abs/2304.07193, 2023. 2, 3, 6, 4

[45] Aj Piergiovanni, Alan Wu, and Michael S. Ryoo. Learning real-world robot policies by dreaming. In *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 7680–7687, 2019. 2

[46] Ed Pizzi, Sreya . Dutta Roy, Sugosh Nagavara Ravindra, Priya Goyal, and Matthijs Douze. A self-supervised descriptor for image copy detection. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14512–14522, 2022. 4

[47] Aske Plaat, Annie Wong, Suzan Verberne, Joost Broekens, Niki van Stein, and Thomas Back. Reasoning with large language models, a survey. *arXiv preprint arXiv:2407.11511*, 2024. 3

[48] Adam Polyak, Amit Zohar, Andrew Brown, Andros Tjandra, Animesh Sinha, Ann Lee, Apoorv Vyas, Bowen Shi, Chih-Yao Ma, Ching-Yao Chuang, David Yan, Dhruv Choudhary, Dingkang Wang, Geet Sethi, Guan Pang, Haoyu Ma, Ishan Misra, Ji Hou, Jialiang Wang, Kiran Jagadeesh, Kunpeng Li, Luxin Zhang, Mannat Singh, Mary Williamson, Matt

Le, Matthew Yu, Mitesh Kumar Singh, Peizhao Zhang, Peter Vajda, Quentin Duval, Rohit Girdhar, Roshan Sumbaly, Sai Saketh Rambhatla, Sam Tsai, Samaneh Azadi, Samyak Datta, Sanyuan Chen, Sean Bell, Sharadh Ramaswamy, Shelly Sheynin, Siddharth Bhattacharya, Simran Motwani, Tao Xu, Tianhe Li, Tingbo Hou, Wei-Ning Hsu, Xi Yin, Xiaoliang Dai, Yaniv Taigman, Yaqiao Luo, Yen-Cheng Liu, Yi-Chiao Wu, Yue Zhao, Yuval Kirstain, Zecheng He, Zijian He, Albert Pumarola, Ali Thabet, Artsiom Sanakoyeu, Arun Mallya, Baishan Guo, Boris Araya, Breena Kerr, Carleigh Wood, Ce Liu, Cen Peng, Dimitry Vengertsev, Edgar Schonfeld, Elliot Blanchard, Felix Juefei-Xu, Fraylie Nord, Jeff Liang, John Hoffman, Jonas Kohler, Kaolin Fire, Karthik Sivakumar, Lawrence Chen, Licheng Yu, Luya Gao, Markos Georgopoulos, Rashel Moritz, Sara K. Sampson, Shikai Li, Simone Parmeggiani, Steve Fine, Tara Fowler, Vladan Petrovic, and Yuming Du. Movie gen: A cast of media foundation models, 2024. 6

[49] Vasili Ramanishka, Yi-Ting Chen, Teruhisa Misu, and Kate Saenko. Toward driving scene understanding: A dataset for learning driver behavior and causal reasoning. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 6

[50] Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020. 4

[51] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. 2

[52] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models. *ArXiv*, abs/2210.08402, 2022. 6

[53] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, Devi Parikh, Sonal Gupta, and Yaniv Taigman. Make-a-video: Text-to-video generation without text-video data, 2022. 2

[54] Matthew Tancik, Pratul Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. *Advances in neural information processing systems*, 33:7537–7547, 2020. 3

[55] Zachary Teed and Jia Deng. Raft: Recurrent all-pairs field transforms for optical flow. In *European Conference on Computer Vision*, 2020. 4

[56] 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. 6, 2

[57] Thomas Unterthiner, Sjoerd van Steenkiste, Karol Kurach, Raphaël Marinier, Marcin Michalski, and Sylvain Gelly. Towards accurate generative models of video: A new metric & challenges. *ArXiv*, abs/1812.01717, 2018. 6

[58] Alexander Veicht, Paul-Edouard Sarlin, Philipp Lindenberger, and Marc Pollefeys. GeoCalib: Single-image Calibration with Geometric Optimization. In *ECCV*, 2024. 6, 2

[59] Features Venkatanath, Praneeth, Maruthi Chandrasekhar Bh., Sumohana S. Channappayya, and Swarup S. Medasani. Blind image quality evaluation using perception based features. *2015 Twenty First National Conference on Communications (NCC)*, pages 1–6, 2015. 6, 3

[60] Vikram Voleti, Alexia Jolicoeur-Martineau, and Christopher Pal. Mcvd: Masked conditional video diffusion for prediction, generation, and interpolation, 2022. 2

[61] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. *arXiv preprint arXiv:2309.09777*, 2023. 2, 3, 7

[62] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. Videocomposer: Compositional video synthesis with motion controllability. *Advances in Neural Information Processing Systems*, 36, 2024. 2

[63] Yan Wang, Wei-Lun Chao, Divyansh Garg, Bharath Hariharan, Mark Campbell, and Kilian Q Weinberger. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8445–8453, 2019. 5

[64] Yuqi Wang, Ke Cheng, Jiawei He, Qitai Wang, Hengchen Dai, Yuntao Chen, Fei Xia, and Zhaoxiang Zhang. Drivingdojo dataset: Advancing interactive and knowledge-enriched driving world model. *arXiv preprint arXiv:2410.10738*, 2024. 6

[65] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14749–14759, 2024. 2, 3, 7

[66] Jialong Wu, Haoyu Ma, Chaoyi Deng, and Mingsheng Long. Pre-training contextualized world models with in-the-wild videos for reinforcement learning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, Red Hook, NY, USA, 2024. Curran Associates Inc. 2

[67] Jiannan Xiang, Guangyi Liu, Yi Gu, Qiyue Gao, Yuting Ning, Yuheng Zha, Zeyu Feng, Tianhua Tao, Shibo Hao, Yemin Shi, et al. Pandora: Towards general world model with natural language actions and video states. *arXiv preprint arXiv:2406.09455*, 2024. 3

[68] Desai Xie, Zhan Xu, Yicong Hong, Hao Tan, Difan Liu, Feng Liu, Arie Kaufman, and Yang Zhou. Progressive autoregressive video diffusion models. *arXiv preprint arXiv:2410.08151*, 2024. 5

[69] Huazhe Xu, Yang Gao, Fisher Yu, and Trevor Darrell. End-to-end learning of driving models from large-scale video datasets. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2174–2182, 2017. 6

[70] Wilson Yan, Yunzhi Zhang, Pieter Abbeel, and Aravind Srinivas. Videogpt: Video generation using vq-vae and transformers, 2021. 2

[71] Jiazhi Yang, Shenyuan Gao, Yihang Qiu, Li Chen, Tianyu Li, Bo Dai, Kashyap Chitta, Penghao Wu, Jia Zeng, Ping Luo, Jun Zhang, Andreas Geiger, Yu Qiao, and Hongyang Li. Generalized Predictive Model for Autonomous Driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024. 2, 3, 6, 7

[72] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything: Unleashing the power of large-scale unlabeled data. In *CVPR*, 2024. 6, 5

[73] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. *arXiv preprint arXiv:2406.09414*, 2024. 6, 2

[74] Sherry Yang, Yilun Du, Seyed Kamyar Seyed Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. In *NeurIPS 2023 Workshop on Generalization in Planning*, 2023. 2, 3

[75] Zhendong Yang, Ailing Zeng, Chun Yuan, and Yu Li. Effective whole-body pose estimation with two-stages distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4210–4220, 2023. 6, 2

[76] Zetong Yang, Li Chen, Yanan Sun, and Hongyang Li. Visual point cloud forecasting enables scalable autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14673–14684, 2024. 3

[77] Yu Yao, Xizi Wang, Mingze Xu, Zelin Pu, Ella Atkins, and David Crandall. When, where, and what? a new dataset for anomaly detection in driving videos. *arXiv preprint arXiv:2004.03044*, 2020. 6

[78] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2

[79] Lunjun Zhang, Yuwen Xiong, Ze Yang, Sergio Casas, Rui Hu, and Raquel Urtasun. Copilot4d: Learning unsupervised world models for autonomous driving via discrete diffusion, 2024. 3

[80] Shiwei Zhang, Jiayu Wang, Yingya Zhang, Kang Zhao, Hangjie Yuan, Zhiwu Qin, Xiang Wang, Deli Zhao, and Jingren Zhou. I2vgen-xl: High-quality image-to-video synthesis via cascaded diffusion models, 2023. 2

[81] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *ArXiv*, abs/2406.19680, 2024. 4, 5

[82] Yuang Zhang, Jiaxi Gu, Li-Wen Wang, Han Wang, Junqi Cheng, Yuefeng Zhu, and Fangyuan Zou. Mimicmotion: High-quality human motion video generation with confidence-aware pose guidance. *arXiv preprint arXiv:2406.19680*, 2024. 5

[83] Yizhuo Zhang, Heng Wang, Shangbin Feng, Zhaoxuan Tan, Xiaochuang Han, Tianxing He, and Yulia Tsvetkov. Can llm graph reasoning generalize beyond pattern memorization? *arXiv preprint arXiv:2406.15992*, 2024. 3

[84] Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Xinze Chen, Guan Huang, Xiaoyi Bao, and Xingang Wang. Drivedreamer-2: Llm-enhanced world models for diverse driving video generation. *arXiv preprint arXiv:2403.06845*, 2024. 2, 3, 7

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

[86] Yi Zhou, Connelly Barnes, Jingwan Lu, Jimei Yang, and Hao Li. On the continuity of rotation representations in neural networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5745–5753, 2019. 2

[87] Zheng Zhu, Xiaofeng Wang, Wangbo Zhao, Chen Min, Nianchen Deng, Min Dou, Yuqi Wang, Botian Shi, Kai Wang, Chi Zhang, Yang You, Zhaoxiang Zhang, Dawei Zhao, Liang Xiao, Jian Zhao, Jiwen Lu, and Guan Huang. Is sora a world simulator? a comprehensive survey on general world models and beyond, 2024. 2