PosePilot: Steering Camera Pose for Generative World Models with Self-supervised Depth

Bu Jin[1,3*], Baihan Yang[1,5*], Weize Li[1*], Zhenxin Zhu[1,4*], Junpeng Jiang[2], Huan-ang Gao[1]
Haiyang Sun[2], Kun Zhan[2], Hengtong Hu[2], Xueyang Zhang[2], Peng Jia[2], Hao Zhao[1†]

***Abstract*—Recent advancements in autonomous driving (AD) systems have highlighted the potential of world models in achieving robust and generalizable performance across both ordinary and challenging driving conditions. However, a key challenge remains: precise and flexible camera pose control, which is crucial for accurate viewpoint transformation and realistic simulation of scene dynamics. In this paper, we introduce PosePilot, a lightweight yet powerful framework that significantly enhances camera pose controllability in generative world models. Drawing inspiration from self-supervised depth estimation, PosePilot leverages structure-from-motion principles to establish a tight coupling between camera pose and video generation. Specifically, we incorporate self-supervised depth and pose readouts, allowing the model to infer depth and relative camera motion directly from video sequences. These outputs drive pose-aware frame warping, guided by a photometric warping loss that enforces geometric consistency across synthesized frames. To further refine camera pose estimation, we introduce a reverse warping step and a pose regression loss, improving viewpoint precision and adaptability. Extensive experiments on autonomous driving and general-domain video datasets demonstrate that PosePilot significantly enhances structural understanding and motion reasoning in both diffusion-based and auto-regressive world models. By steering camera pose with self-supervised depth, PosePilot sets a new benchmark for pose controllability, enabling physically consistent, reliable viewpoint synthesis in generative world models.**

***Index Terms*—world model, video generation, autonomous driving**

I. Introduction

In recent years, great advancements have been made in the development of autonomous driving systems. While existing AD methods [1, 2] demonstrate promising results in various scenarios, they still face significant challenges when confronted with long-tail distribution or out-of-distribution situations. These edge cases, which occur infrequently in typical training datasets but are critical in real-world driving environments, expose limitations in the robustness and generalization ability of current models. One promising solution is world models [3–10], which capture the structure and dynamics of an environment, enabling an autonomous driving system to simulate and predict future states by "imagining" the external world. Through building such models, these generative models can help AD systems anticipate future states, reason about complex dynamics and better generalize to novel or unexpected situations.

A critical factor in harnessing the full potential of world models lies in **pose controllability**, the ability to manipulate and reason about an agent's location, orientation, and viewpoint within the learned representation. Pose controllability ensures that the model can adapt its predictive capabilities to varying viewpoints, accurately simulate sensor feedback, and better handle environmental uncertainties. By granting precise and flexible control over the pose, the world model gains a robust mechanism for exploring and understanding the state space, thereby facilitating improved performance in both typical and hard-to-generalize driving conditions.

Some existing works [10–13] have tried to introduce the pose controllability to the video generation model. For example, AnimateDiff [14] introduces a transformer-based motion module along with a MotionLoRA, supporting certain types of camera movement. MotionCtrl [11] encodes the camera pose values RT with several fully connected layers and inserts the embeddings to the transformer layers to interact with the visual outputs, enabling a flexible camera pose control. CameraCtrl [13] further utilizes a plücker embeddings as the primary form of camera parameters. Vista [10] proposes a versatile action controllability with diverse control formats, like motion values, camera trajectory, driving command or goal point. However, these works typically learn the controllability through learnable attention layers, lacking a deeper understanding of the relation between camera pose and the structural evolution of the scene.

Inspired from self-supervised depth estimation [15–19], the camera pose is intrinsically tied to adjacent frames in a video. Specifically, the transformation between two neighboring frames, parameterized by camera intrinsics and their relative poses, enables warping pixels from one image onto another through the corresponding depth map. This structure-from-motion property ensures geometric consistency across different camera motions and provides an effective prior for pose controllability in generative world models.

We thus propose `PosePilot`, a light-weight camera pose control method that enhances the pose controllability in world models. Specifically, we introduce a depth readout and an ego-motion readout to generate the depth maps and camera poses of the input video. These outputs are then used to warp a source frame onto its neighboring frame with a specified camera pose. The warped frames guide the synthesized video through a photometric control loss [18], ensuring that the

* indicates equal contribution and † indicates the corresponding author.
[1]Institute for AI Industry Research (AIR), Tsinghua University
[2]Li Auto
[3]Institute of Automation, Chinese Academy of Sciences
[4]School of Automation Science and Electrical Engineering, Beihang University
[5]School of Computer Science and Technology, Beijing Jiaotong University

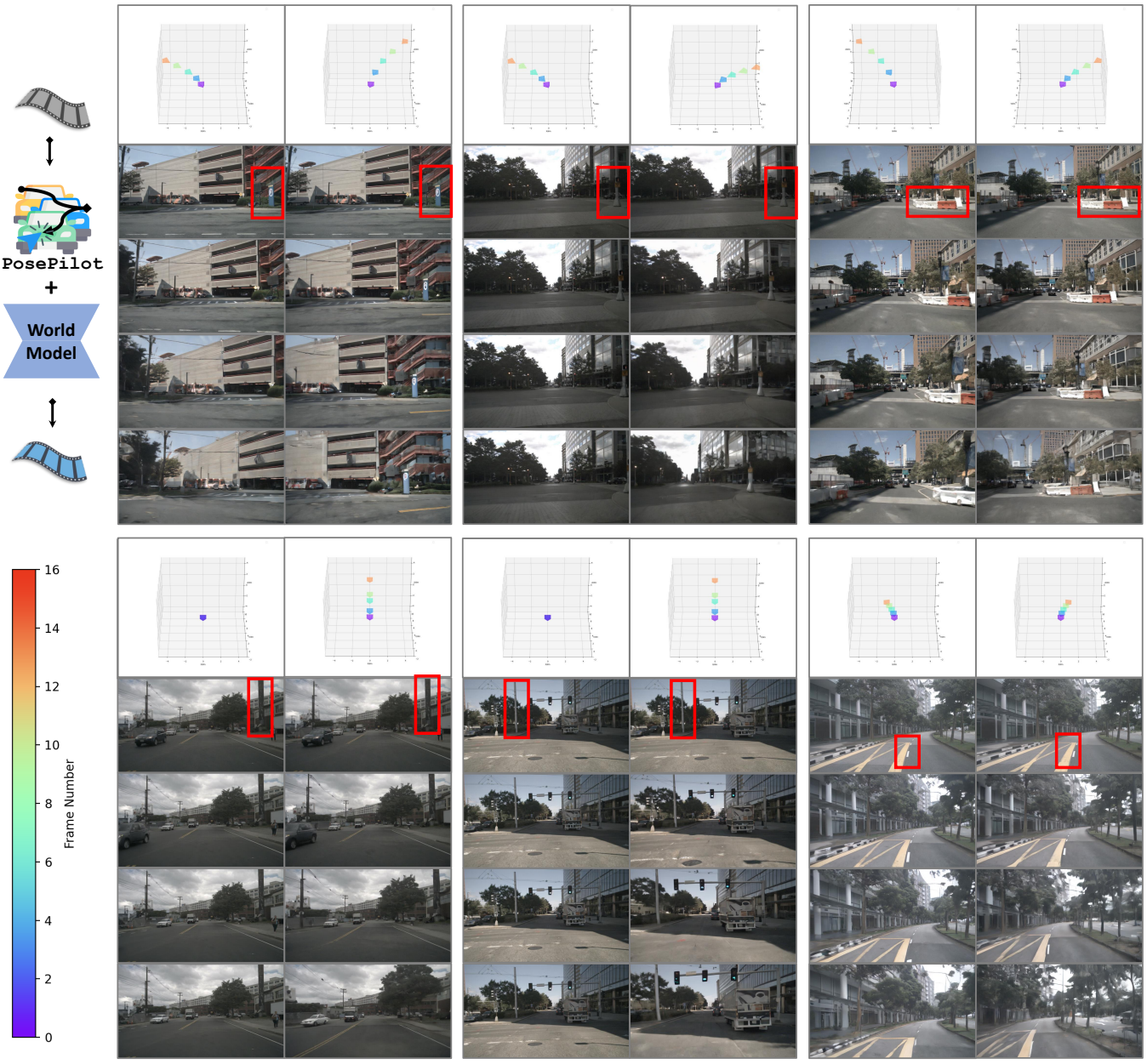

Fig. 1. **Illustration of `PosePilot`**. We propose `PosePilot`, a lightweight pose controllability enhancer for world models in autonomous driving. `PosePilot`can generate videos collaboratively aligned with the input camera pose. The red box indicates the reference objects for easy identification.

generated video remains consistent with the specified camera pose. By enforcing this warping consistency, `PosePilot` enhances the world model's ability to reason about its location, orientation, and viewpoint, thereby improving pose controllability. Furthermore, we introduce an inverse photometric control loss, which applies an inverse warping step to the synthesized video, aligning it with the original video via an additional pair of depth and pose readouts to further enhance the controllability. We also introduce a regression loss to the pose readouts, explicitly refining the pose estimation to achieve more precise pose control. This multi-stage procedure bolsters the overall pose controllability of the world models, ensuring more accurate and reliable camera viewpoints. Our contributions can be summarized as follows:

- We introduce `PosePilot`, a light-weight camera pose control method that enhances the pose controllability in world models, which can be easily applied to other world models to generate videos with flexible camera viewpoints.
- We propose a photometric loss, which integrates self-supervised depth and ego-motion estimation to enhance the structural and motion understanding of world models.
- We conduct comprehensive experiments and ablation study to analyze `PosePilot`, which boasts the control-

lability for both diffusion-based and autoregressive-based generative world models on autonomous driving and real world datasets.

II. Related Works

A. World Models for Driving Scenes

World Models [20, 21] are generative models that predict future frames to simulate dynamics. Recent advances enabling video generation to augment autonomous driving data with diverse road conditions and rare corner cases [22, 23]. Existing methods use various conditions including images, text, 3D layouts, and actions to generate specific driving scenes or predict future video sequences. GAIA-1 [24] generates realistic and diverse driving videos by integrating video, text, and action inputs, demonstrating a strong understanding of contextual information and physical principles. DriveDreamer [25] improves driving scenario generation by adding HD maps and 3D boxes for better video quality and enabling the generation of future driving actions, aiding in decision-making. WorldDreamer [26] frames world modeling as an unsupervised challenge within visual sequence modeling, inspired by large language models. MUVO [27] enhances world modeling by integrating LIDAR point clouds, improving the prediction of driving environments and generating 3D occupancy grids that integrate well with downstream tasks. OccWorld [28] leverages 3D occupancy data to predict environmental evolution and guide autonomous vehicle actions, marking a shift toward multimodal approaches in autonomous driving research. In contrast, our work does not design a brand new framework for world model but serves as a plug-and-play module that builds upon existing world models to provide effective camera pose control during video generation for driving scenes.

B. Controllable Video Generation

Controllability plays a crucial role in video generation applications [29–33], especially in enabling customized camera pose generation to better meet user needs. MotionCtrl [11] enables independent control of camera and object motion in video generation by explicitly modeling camera poses and trajectories. CameraCtrl [13] enhances text-to-video generation with precise camera pose control using a plug-and-play module trained on parameterized trajectories. CamTrol [30] achieves camera pose control in video diffusion models by leveraging latent layout priors to adjust noisy latents without fine-tuning. CamCtrl3D[31] improves fly-through video generation by conditioning an image-to-video diffusion model on camera trajectories using multiple techniques. In particular, controllable video generation in autonomous driving places special emphasis on camera pose control: MagicDrive [7] enables camera pose-controllable street view generation by conditioning on 3D geometry inputs and ensuring cross-view consistency. Vista [10] enhances autonomous driving video generation with explicit motion dynamics and structural learning for high-fidelity, generalizable viewpoint control. DiVE[34] achieves multi-view consistent video generation by enforcing BEV layout control and integrating spatial attention mechanisms for precise camera pose alignment.

What sets `PosePilot` apart is that we explicitly read out depth and camera pose between adjacent frames for direct warping of future frames, rather than relying on other modalities as conditions, which can lead to suboptimal optimization gaps.

III. Methodology

We propose `PosePilot`, a lightweight camera control enhancer in generative world models, focusing on enhancing the camera controllability in off-the-shelf world models for autonomous driving. As illustrated in Fig. 2, `PosePilot` is comprised of three losses: two photometric control losses and a pose regression loss. We adapt `PosePilot` on top of the powerful generative world models [10, 34, 35].

A. Photometric Control Loss

In contrast to traditional video generation models, a key factor in the world models for autonomous driving is the camera parameters. Once the extrinsic matrices (or pose) of each camera are available, we can use the structure-from-motion methods to generate the structural and dynamic context to guide the video generation. Inspired by monocular self-supervised depth estimation [17], we incorporate a photometric control loss into our generative model. The objective of the photometric control loss is to derive real-world scale information from the camera's extrinsic matrices, which provides prior information for video generation.

Firstly, we introduce a learnable depth readout and a learnable pose readout, which generates the depth maps and relative 6D camera pose for two consecutive frames. Given two frames (f^i, f^j), the outputs of the depth readout are their depth maps (D^i, D^j), and the output of the pose readout is the relative 6D camera pose $T^{i\to j}$. These two values can subsequently be employed to warp the initial frame f^i to f^j. Let x^i be the coordinates of a pixel in the frame f^i, the projected coordinates $x^{i\to j}$ in warped frame $f^{i\to j}$ can be calculated by:

$$x^{i\to j} = KT^{i\to j}D^iK^{-1}x^i, \tag{1}$$

where K are intrinsic matrices of the camera. The whole process is differentiable, following common practice in self-supervised depth estimation [15].

With the synthesized f'^j and the reference frame $\hat{f}^j$, the photometric control loss can be formulated as:

$$L_p = \frac{1}{|N|}\sum_{x\in N}\left\|\hat{f}^j(x) - f^{i\to j}(x)\right\|_1, \tag{2}$$

where $\hat{f}$ represents the predicted frame generated by the generation model and N denotes the collection of valid points mapped from f^i onto the plane of f^j. $|N|$ represents the point number of N. Note that to better integrate the camera pose prior to guide the models' generation, our reference frames are not the input frames f^j, but the frames generated by the world models $\hat{f}^j$. In this way, the generation model can get a

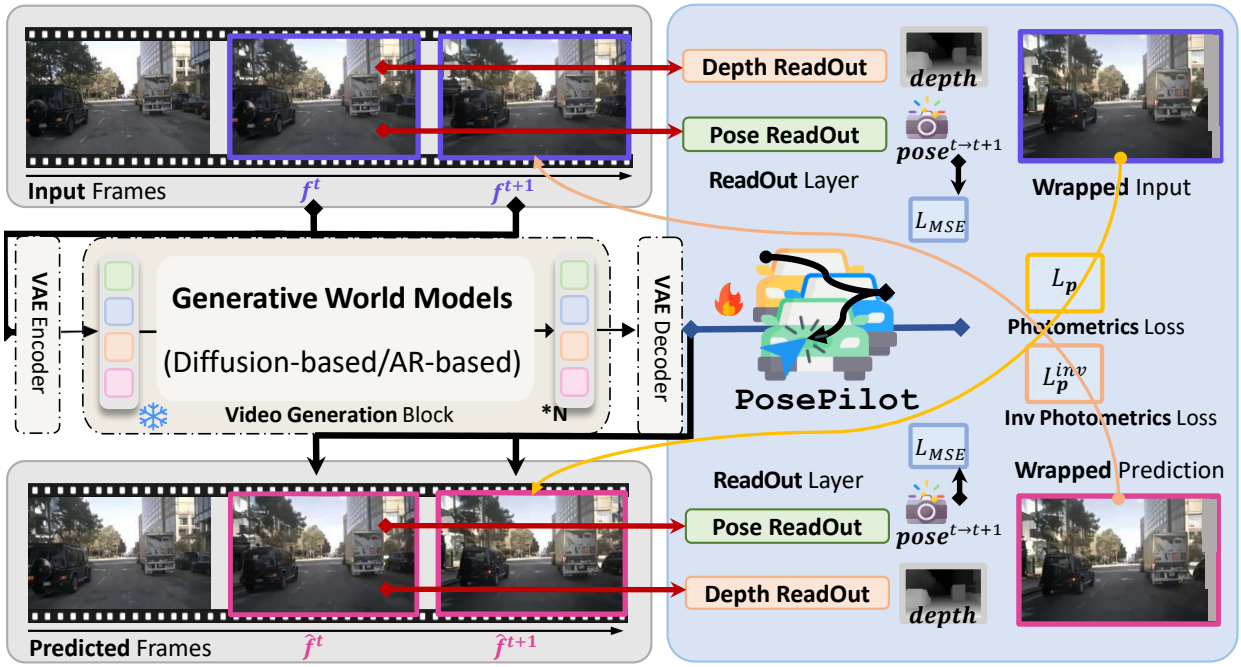

Fig. 2. **Overview of `PosePilot`.** We utilize off-the-shelf generative world model as our base model and build our `PosePilot` on top of them. We introduce two auxiliary tasks: depth estimation and ego-motion estimation. The depth estimation predicts per-pixel depth maps, while ego-motion prediction estimates the relative camera pose between consecutive frames. We introduce a photometric control loss and an inverse photometric control loss, which aligns the predicted frames with real frames by warping pixels based on the estimated depth and ego-motion.

feedback from the photometric control loss, which enhances the controllability of camera pose.

To better handle the complex illumination changes, we also add auxiliary SSIM [36] loss to the photometric control loss:

$$L_p = \frac{1}{|N|}\sum_{x\in N}\Big(\Big\|\hat{f}^j(x) - f^{i\to j}(x)\Big\|_1 + \frac{1-\mathrm{SSIM}_{\hat{j}j'}(x)}{2}\Big), \tag{3}$$

where $\mathrm{SSIM}_{jj'}$ represents the per-element structural similarity between $\hat{f}^j$ and f'^j.

Through this photometric loss, our model is capable of capturing structural information from the vehicle's motion, which is essential for pose controllability for world models.

B. Inverse Photometric Control Loss

Similar to the photometric control loss in the previous subsection, we introduce an inverse photometric control loss to further improve the video-pose consistency of the generated frames. Instead of solely warping the original frame to the viewpoint of the generated frame, we also inverse warp the generated frame back to the original viewpoint. This bidirectional alignment helps reduce the artifacts due to occlusions and dis-occlusions.

Let us denote the generated frames as $(\hat{f}^i, \hat{f}^j)$, which are obtained from the world model. The learnable depth readout provides their respective depth maps (D^i, D^j), and the pose readout outputs the relative 6D camera transformations $T^{i\to j}$ and $T^{j\to i}$. To map $\hat{f}^j$ back to $\hat{f}^i$, we warp using the same bilinear sampling mechanism as above. Specifically, the pixel projection from $\hat{f}^j$ to the warped inverse view $\hat{f}^{j\to i}$ follows the same procedure as in the forward warping. Denoting x^j as a pixel in the generated frame $\hat{f}^j$, its inverse-warped coordinate $x^{j\to i}$ is found by:

$$x^{j\to i} = KT^{j\to i}D^jK^{-1}p^j, \tag{4}$$

where K is the intrinsic matrix, and D^j is the depth at pixel x^j in the generated frame $\hat{f}^j$. Just as in the forward direction, this process is fully differentiable. Once we obtain the warped image $\hat{f}^{j\to i}$, we compute the inverse photometric control loss between $\hat{f}^i$ and $\hat{f}^{j\to i}$. Let $\tilde{N}$ denote the set of valid pixels that can be projected from $\hat{f}^j$ to $\hat{f}^i$. The inverse photometric control loss L_p^{inv} has the same form as the photometric control loss, which includes the per-pixel L1 term and the auxiliary SSIM loss. By enforcing consistency in this reverse mapping, the model refines the pose and depth estimation for generated frames and reduces warping artifacts in the generated views. This inverse photometric control loss, combined with the forward photometric control loss, ensures tighter motion consistency for video generation.

To ensure the pose readouts generate precise camera pose, we also utilize a regression loss L_{MSE} to refine the pose readouts.

C. Total Loss

We combine the forward and inverse photometric control losses with our generative objective into a comprehensive total loss. Let L_g represent the generative objective (e.g.,

TABLE I
COMPARISON WITH PREVIOUS WORKS

Method	TransErr↓	RotErr↓
Diffusion-based Generation:		
DiVE [34]	13.07	4.52
DiVE + PosePilot **(Ours)**	**6.37**	**1.40**
Vista [10]	6.83	1.74
Vista + PosePilot **(Ours)**	**6.52**	**1.53**
Autoregressive-based Generation:		
DrivingWorld [35]	3.17	1.64
DrivingWorld + PosePilot **(Ours)**	**2.95**	**1.48**

the diffusion-based reconstruction or adversarial loss), L_p be the forward photometric control loss, and L_p^{inv} be the inverse photometric control loss. We employ weighting factors α_p and $\alpha_{p^{\text{inv}}}$ to balance the relative importance of these terms, which yields the total loss:

$$L_{\text{total}} = L_g + \alpha_p L_p + \alpha_{p^{\text{inv}}} L_p^{\text{inv}} + \alpha_{mse} L_{MSE}, \quad (5)$$

where the L_{MSE} represents the regression loss to the pose readouts and α_{mse} is its factor, explicitly refining the pose estimation to achieve more precise pose control. In practice, α_p, $\alpha_{p^{\text{inv}}}$ and α_{mse} can be tuned to find an optimal trade-off between the generative fidelity of the generative model and the geometric consistency enforced by both the forward and inverse photometric control terms. By minimizing L_{total}, the framework simultaneously learns to produce high-quality, realistic frames while ensuring that the predicted depth and pose remain consistent in both directions of warping.

IV. EXPERIMENT

A. Experiment Setup

Dataset. We adopt the nuScenes [37] dataset for main experiments to evaluate the camera controllability of video generation in driving scenarios. We inherit the official split for evaluation, which includes 700 training videos and 150 validation videos, where each video sequence is recorded at 2 Hz and lasts about 20 seconds. Additionally, we follow [13] to use the RealEstate10K [38] dataset to conduct cross-domain application experiment in general video generation.

Evaluation Metrics. We adopt TranErr and RotErr [13] as camera alignment metrics to quantify the discrepancy between the translation and rotation vectors of the specified camera pose conditions and those of the generated video camera trajectories, assessing the quality of camera pose control during video generation. We also employ the Fréchet Inception Distance (FID) [39], along with parameters size, iteration time, and inference time, in ablation experiments to evaluate video quality, model size, and computational efficiency.

B. Main Results

We adapt our methods to the existing world model solutions on the nuScenes validation set. We divide the nuScenes validation set into four subsets. The TranErr and RotErr scores are measured on each subset and then averaged. We conduct experiments on most of the publicly available world models, including diffusion-based methods [10, 34] and auto-regression-based approach [35]. The results are shown in Tab. I. We can see that with our PosePilot, all of the models can get relatively lower TranErr & RotErr, which demonstrates the effectiveness of the proposed architecture. For example, in diffusion-based methods, with our PosePilot, Vista gets a 0.31 decrease on translation error and 0.21 decrease on rotation error. In auto-regression-based approach, Drivingworld also gets a 0.22 decrease on translation error and 0.16 decrease on rotation error, indicating the superior generalization of our method. We also show the qualitative results of PosePilot at Fig. 3, we can see that our PosePilot contributes significantly to the model's camera pose control.

C. Ablation Study

To evaluate the impact of each proposed loss terms in PosePilot, we conduct a comprehensive ablation study by analyzing their effects the key aspects of our architecture.

Camera Pose Control Accuracy. We examine how each loss term contributes to precise camera pose control. By ablating specific losses, we assess potential deviations in predicted viewpoints and the overall stability of camera motion across frames. This is quantified using Translation Error (TransErr) and Rotation Error (RotErr) following common practice [13]. This analysis reveals that certain loss terms are crucial for maintaining low translation errors, ensuring accurate camera positioning along the intended path, while other losses are essential for preserving correct angular orientation and minimizing rotation errors. By identifying the specific contributions of each loss term, we can optimize the camera pose control system according to the unique demands of different scenarios, ultimately enhancing the performance and reliability of systems dependent on precise camera pose estimation.

Video Generation Quality. We evaluate how the removal of each loss affects the realism and consistency of the generated video. This includes assessing structural coherence, temporal smoothness, and fidelity to the intended viewpoint. The quality of video generation is measured using the Fréchet Inception Distance (FID) [39].

Model Parameter Size. We analyze the influence of each loss term on the total parameter count of the model, determining whether certain losses introduce additional complexity or redundancy in the architecture. This is assessed through the parameter count of the model. Our findings indicate that, although the introduction of our method does increase parameters—specifically, an addition of 180 million parameters—this increase is accompanied by a substantial improvement in pose controllability. Importantly, even with the parameter increase, the size of our model remains considerably smaller than the original world models (more than 1 billion). This indicates a successful balance between maintaining a compact and efficient architecture while achieving greater capabilities in pose control. These insights suggest that carefully crafted loss terms can add value by refining model performance with min-

Fig. 3. **Controllability of `PosePilot`.** Our model can enhance the model's pose controllability, which generates videos that collaboratively align with the conditional inputs. More examples can be found in the supplementary video.

TABLE II
ELEMENT-WISE ABLATION STUDY

PosePilot			TransErr↓	RotErr↓	FID↓	Parameters (M)↓	Iter Time (s)↓	Inference Time (s)↓
L_{mse}	L_p	L_p^{inv}						
✔	✔	✔	**6.37**	**1.40**	**13.7**	+181.1	8.83	47.8
✔	✔		7.56	1.67	14.5	+90.5	8.71	47.7
✔		✔	7.09	1.51	13.8	+90.5	8.70	47.7
	✔	✔	6.97	1.43	14.7	+181.1	8.80	47.8

imal additional complexity, providing a strategic advantage in developing sophisticated yet efficient camera control systems.

Iteration and Inference Efficiency. We also measure how different loss configurations impact training iteration speed and inference time, identifying potential trade-offs between computational efficiency and model performance. In this experiment, we test the inference time with Tesla A100 GPU. The training time is tested on 8 GPUs and the inference time is tested on a single GPU. The resolution of the video is 1280×720 and the frame number is 8. We can see that the iteration time and inference are generally the same, indicating the lightweight and plug-and-play ability of our `PosePilot`.

D. Cross-domain Application

In previous experiments, we demonstrate that the introduced `PosePilot` significantly improves pose controllability in world models for autonomous driving. Surprisingly, when we apply this module to the general video generation task, it further enhances the precision of camera pose control, achieving a new state-of-the-art performance on RealEstate10K [38], as shown in Table III. This finding indicates that our approach possesses strong cross-domain adaptability and scalability, paving the way for more advanced video generation tasks in diverse scenarios.

TABLE III
CAMERA CONTROL ABILITY IN NATURAL SCENES.

Method	TransErr↓	RotErr↓
AnimateDiff [14]	9.81	1.03
MotionCtrl [11]	9.02	0.87
CameraCtrl [13]	8.83	0.95
CameraCtrl + `PosePilot` **(Ours)**	**6.52**	**0.70**

V. CONCLUSION

In this paper, we introduce `PosePilot`, a lightweight yet powerful framework that significantly enhances camera pose controllability in generative world models. Drawing inspiration from self-supervised depth estimation, PosePilot leverages structure-from-motion principles to establish a tight coupling between camera pose and video generation. We incorporate self-supervised depth and pose readouts, allowing the model to infer depth and relative camera motion directly from video sequences. These outputs drive pose-aware frame warping, guided by a photometric warping loss that enforces geometric consistency across synthesized frames. Extensive experiments on autonomous driving and general-domain video datasets demonstrate that PosePilot significantly enhances structural understanding and motion reasoning in both diffusion-based and auto-regressive world models. By steering camera pose with self-supervised depth, PosePilot sets a new benchmark for pose controllability, enabling physically consistent, reliable viewpoint synthesis in world models.

References

[1] Y. Hu, J. Yang, L. Chen, K. Li, C. Sima, X. Zhu, S. Chai, S. Du, T. Lin, W. Wang *et al.*, "Planning-oriented autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 17 853–17 862.

[2] H. Yang, S. Zhang, D. Huang, X. Wu, H. Zhu, T. He, S. Tang, H. Zhao, Q. Qiu, B. Lin *et al.*, "Unipad: A universal pre-training paradigm for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 15 238–15 250.

[3] A. Hu, G. Corrado, N. Griffiths, Z. Murez, C. Gurau, H. Yeo, A. Kendall, R. Cipolla, and J. Shotton, "Model-based imitation learning for urban driving," *Advances in Neural Information Processing Systems*, vol. 35, pp. 20 703–20 716, 2022.

[4] Y. Wang, J. He, L. Fan, H. Li, Y. Chen, and Z. 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*, 2024, pp. 14 749–14 759.

[5] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-driven world models for autonomous driving," *arXiv preprint arXiv:2309.09777*, 2023.

[6] K. Yang, E. Ma, J. Peng, Q. Guo, D. Lin, and K. Yu, "Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout," *arXiv preprint arXiv:2308.01661*, 2023.

[7] R. Gao, K. Chen, E. Xie, L. Hong, Z. Li, D.-Y. Yeung, and Q. Xu, "Magicdrive: Street view generation with diverse 3d geometry control," *arXiv preprint arXiv:2310.02601*, 2023.

[8] Y. Wen, Y. Zhao, Y. Liu, F. Jia, Y. Wang, C. Luo, C. Zhang, T. Wang, X. Sun, and X. Zhang, "Panacea: Panoramic and controllable video generation for autonomous driving," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 6902–6912.

[9] B. Huang, Y. Wen, Y. Zhao, Y. Hu, Y. Liu, F. Jia, W. Mao, T. Wang, C. Zhang, C. W. Chen *et al.*, "Subjectdrive: Scaling generative data in autonomous driving via subject control," *arXiv preprint arXiv:2403.19438*, 2024.

[10] S. Gao, J. Yang, L. Chen, K. Chitta, Y. Qiu, A. Geiger, J. Zhang, and H. Li, "Vista: A generalizable driving world model with high fidelity and versatile controllability," *arXiv preprint arXiv:2405.17398*, 2024.

[11] Z. Wang, Z. Yuan, X. Wang, Y. Li, T. Chen, M. Xia, P. Luo, and Y. Shan, "Motionctrl: A unified and flexible motion controller for video generation," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–11.

[12] S. Yang, L. Hou, H. Huang, C. Ma, P. Wan, D. Zhang, X. Chen, and J. Liao, "Direct-a-video: Customized video generation with user-directed camera movement and object motion," in *ACM SIGGRAPH 2024 Conference Papers*, 2024, pp. 1–12.

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

[14] Y. Guo, C. Yang, A. Rao, Z. Liang, Y. Wang, Y. Qiao, M. Agrawala, D. Lin, and B. Dai, "Animatediff: Animate your personalized text-to-image diffusion models without specific tuning," *arXiv preprint arXiv:2307.04725*, 2023.

[15] T. Zhou, M. Brown, N. Snavely, and D. G. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1851–1858.

[16] R. Mahjourian, M. Wicke, and A. Angelova, "Unsupervised learning of depth and ego-motion from monocular video using 3d geometric constraints," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5667–5675.

[17] J. Bian, Z. Li, N. Wang, H. Zhan, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth and ego-motion learning from monocular video," *Advances in neural information processing systems*, vol. 32, 2019.

[18] R. Garg, V. K. Bg, G. Carneiro, and I. Reid, "Unsupervised cnn for single view depth estimation: Geometry to the rescue," in *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VIII 14*. Springer, 2016, pp. 740–756.

[19] J.-W. Bian, H. Zhan, N. Wang, Z. Li, L. Zhang, C. Shen, M.-M. Cheng, and I. Reid, "Unsupervised scale-consistent depth learning from video," *International Journal of Computer Vision*, vol. 129, no. 9, pp. 2548–2564, 2021.

[20] T. Feng, W. Wang, and Y. Yang, "A survey of world models for autonomous driving," *arXiv preprint arXiv:2501.11260*, 2025.

[21] L. Chen, P. Wu, K. Chitta, B. Jaeger, A. Geiger, and H. Li, "End-to-end autonomous driving: Challenges and frontiers," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[22] X. Bai, Z. Hu, X. Zhu, Q. Huang, Y. Chen, H. Fu, and C.-L. Tai, "Transfusion: Robust lidar-camera fusion for 3d object detection with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1090–1099.

[23] W. Ding, L. Qiao, X. Qiu, and C. Zhang, "Pivotnet: Vectorized pivot learning for end-to-end hd map construction," 2023.

[24] A. Hu, L. Russell, H. Yeo, Z. Murez, G. Fedoseev, A. Kendall, J. Shotton, and G. Corrado, "Gaia-1: A generative world model for autonomous driving," *arXiv preprint arXiv:2309.17080*, 2023.

[25] X. Wang, Z. Zhu, G. Huang, X. Chen, J. Zhu, and J. Lu, "Drivedreamer: Towards real-world-drive world models for autonomous driving," in *European Conference on Computer Vision*. Springer, 2024, pp. 55–72.

[26] X. Wang, Z. Zhu, G. Huang, B. Wang, X. Chen, and J. Lu, "Worlddreamer: Towards general world models for video generation via predicting masked tokens," *arXiv preprint arXiv:2401.09985*, 2024.
[27] D. Bogdoll, Y. Yang, T. Joseph, and J. M. Zöllner, "Muvo: A multimodal world model with spatial representations for autonomous driving," *arXiv preprint arXiv:2311.11762*, 2023.
[28] W. Zheng, W. Chen, Y. Huang, B. Zhang, Y. Duan, and J. Lu, "Occworld: Learning a 3d occupancy world model for autonomous driving," in *European conference on computer vision*. Springer, 2024, pp. 55–72.
[29] J. Wu, X. Li, Y. Zeng, J. Zhang, Q. Zhou, Y. Li, Y. Tong, and K. Chen, "Motionbooth: Motion-aware customized text-to-video generation," *Advances in Neural Information Processing Systems*, vol. 37, pp. 34 322–34 348, 2025.
[30] C. Hou, G. Wei, Y. Zeng, and Z. Chen, "Training-free camera control for video generation," *arXiv preprint arXiv:2406.10126*, 2024.
[31] S. Popov, A. Raj, M. Krainin, Y. Li, W. T. Freeman, and M. Rubinstein, "Camctrl3d: Single-image scene exploration with precise 3d camera control," *arXiv preprint arXiv:2501.06006*, 2025.
[32] G. Zheng, T. Li, R. Jiang, Y. Lu, T. Wu, and X. Li, "Cami2v: Camera-controlled image-to-video diffusion model," *arXiv preprint arXiv:2410.15957*, 2024.
[33] W. Feng, J. Liu, P. Tu, T. Qi, M. Sun, T. Ma, S. Zhao, S. Zhou, and Q. He, "I2vcontrol-camera: Precise video camera control with adjustable motion strength," *arXiv preprint arXiv:2411.06525*, 2024.
[34] J. Jiang, G. Hong, L. Zhou, E. Ma, H. Hu, X. Zhou, J. Xiang, F. Liu, K. Yu, H. Sun *et al.*, "Dive: Dit-based video generation with enhanced control," *arXiv preprint arXiv:2409.01595*, 2024.
[35] X. Hu, W. Yin, M. Jia, J. Deng, X. Guo, Q. Zhang, X. Long, and P. Tan, "Drivingworld: Constructingworld model for autonomous driving via video gpt," *arXiv preprint arXiv:2412.19505*, 2024.
[36] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004.
[37] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.
[38] T. Zhou, R. Tucker, J. Flynn, G. Fyffe, and N. Snavely, "Stereo magnification: Learning view synthesis using multiplane images," in *SIGGRAPH*, 2018.
[39] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.