QuaDreamer: Controllable Panoramic Video Generation for Quadruped Robots

Sheng Wu[1,*], **Fei Teng**[1,*], **Hao Shi**[2,3,*], **Qi Jiang**[2], **Kai Luo**[1], **Kaiwei Wang**[2], **Kailun Yang**[1,†]
[1]Hunan University [2]Zhejiang University [3]Nanyang Technological University

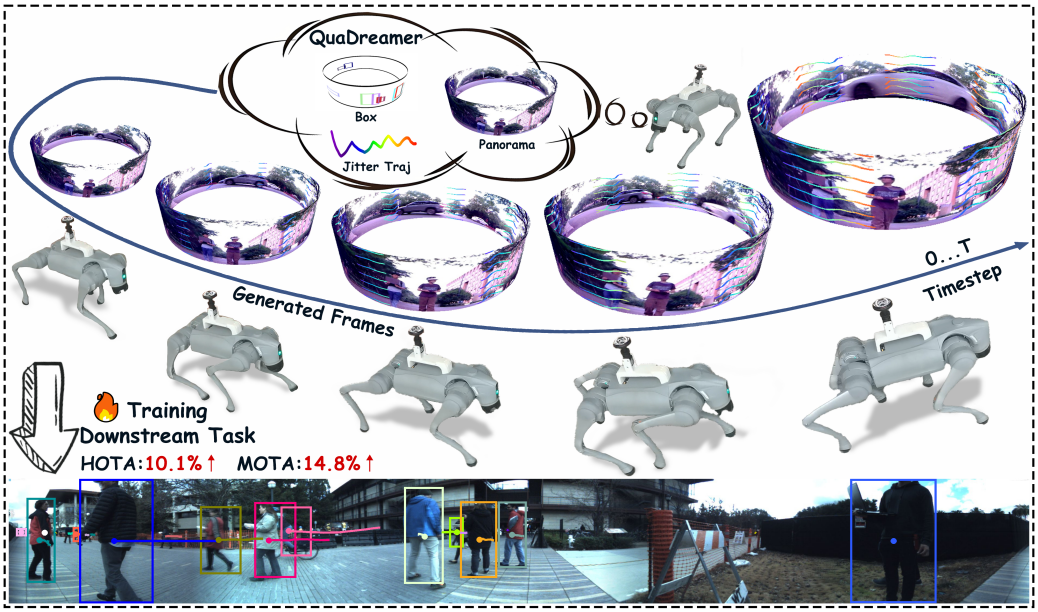

Figure 1: Illustration of the proposed QuaDreamer, the first panoramic video generation framework tailored for quadruped robots. QuaDreamer enables generation with control via box and jitter prompts, serving as a data source to enhance the performance of downstream tasks.

Abstract: Panoramic cameras, capturing comprehensive 360-degree environmental data, are suitable for quadruped robots in surrounding perception and interaction with complex environments. However, the scarcity of high-quality panoramic training data — caused by inherent kinematic constraints and complex sensor calibration challenges — fundamentally limits the development of robust perception systems tailored to these embodied platforms. To address this issue, we propose QuaDreamer—the first panoramic data generation engine specifically designed for quadruped robots. QuaDreamer focuses on mimicking the motion paradigm of quadruped robots to generate highly controllable, realistic panoramic videos, providing a data source for downstream tasks. Specifically, to effectively capture the unique vertical vibration characteristics exhibited during quadruped locomotion, we introduce Vertical Jitter Encoding (VJE). VJE extracts controllable vertical signals through frequency-domain feature filtering and provides high-quality prompts. To facilitate high-quality panoramic video generation under jitter signal control, we propose a Scene-Object Controller (SOC) that effectively manages object motion and boosts background jitter control through the attention mechanism. To address panoramic distortions in wide-FoV video generation, we propose the Panoramic Enhancer (PE) – a dual-stream architecture that synergizes frequency-texture refinement for local detail enhancement with spatial-structure correction for global geometric consistency. We further demonstrate that the generated video

[*]Equal Contribution. [†]Correspondence to kailun.yang@hnu.edu.cn.

9th Conference on Robot Learning (CoRL 2025), Seoul, Korea.

sequences can serve as training data for the quadruped robot's panoramic visual perception model, enhancing the performance of multi-object tracking in 360-degree scenes. The source code and model weights will be publicly available at https://github.com/losehu/QuaDreamer.

Keywords: Panoramic Video Generation, World Model, Quadruped Robots

1 Introduction

Quadruped robots have become a crucial component of advanced embodied intelligence technologies, demonstrating significant potential across a variety of application scenarios such as inspection[1, 2], search and rescue [3], and security [4]. An increasing number of researchers are focusing on enhancing robotic agents' understanding of their environments [5, 6, 7, 8]. With the advantages of a higher Field-of-View (FoV) relative to device size ratio [9, 10, 11, 12, 13] and better alignment with LiDARs, panoramic cameras with a 360-degree FoV offer a convenient solution for quadruped robots to achieve a comprehensive understanding of the surrounding scenes. However, an embodied agent's ability to understand its environment is fundamentally a data-driven process, heavily reliant on training with accurately annotated datasets [14]. Large-scale, diverse, and photorealistic data is essential for achieving robust perception in dynamic environments. Currently, panoramic datasets captured from the perspective of quadruped robots remain scarce due to multiple challenges, including the limited endurance of quadruped platforms [15] and the instability of the ego-agent [4]. These challenges introduce the issue of homogeneous scene distribution in existing datasets [16], along with the technical difficulties of image stitching and time synchronization [17], making the data collection process highly labor-intensive.

To enable mass production of high-quality panoramic data for scene understanding of embodied agents, we propose **QuaDreamer**—the first panoramic motion generation engine (world model) specifically designed for quadruped robots, as shown in Fig. 1. By providing a single panoramic image, object motion trajectories, and the vibration signals generated during the robot's movement, high-quality panoramic videos are generated to recreate the motion and vibration characteristics of the quadruped robot. QuaDreamer consists of three main components: the Vertical Jitter Encoding (VJE), Scene-Object Controller (SOC), and Panoramic Enhancer (PE).

Specifically, VJE employs a high-pass filter to decouple the low-frequency object-relative trajectories from the high-frequency vertical jitter. The extracted jitter signal represents spatial-domain positional offsets of the quadruped robot caused by locomotion dynamics. Subsequently, these spatial offsets are projected into the feature domain via a camera encoder, enabling generation models to explicitly learn and leverage vibration patterns in both geometric and latent spaces. In latent spaces, control signals are integrated using an attention mechanism, while in the geometric space, a position-aware diffusion process is constructed: by using the SOC to merge the object bounding box with the features encoded by the camera, and seamlessly integrating this into the pre-trained video diffusion model, jitter control is enhanced, enabling precise object position control. Meanwhile, panoramic imagery substantially expands the FoV but simultaneously introduces considerable challenges to maintaining global consistency during the diffusion process. To address this challenge, we propose the Panoramic Enhancer, a dual-stream module that simultaneously processes frequency-domain features and spatial-domain structures. This module includes a Fourier CNN Texture Restorer (FTR) [18], which employs frequency-domain convolutions with global receptive fields to mitigate resolution-sensitive artifacts while preserving high-frequency details, and a State Space Model (SSM) [19], which enforces structural continuity across distorted regions through selective state transitions. The bidirectional interaction between these components allows for effective resolution of both local distortions and global inconsistencies.

To verify the effectiveness of our proposed method, we evaluate the model in terms of controllability and video quality, ensuring a comprehensive assessment of the framework. Specifically, the LPIPS [20] score decreases by 3.68% and the SSIM [21] score increases by 3.46% compared to our baseline. On the evaluation metric PTrack designed for assessing controllability, QuaDreamer

improves by 43.86% over the baseline, demonstrating a significant enhancement in synthesized video control. Furthermore, our architecture without control modules outperforms SVD in terms of video quality. We also evaluate the synthesized videos on downstream tasks, demonstrating that the generated videos serve effectively as data augmentation, benefiting multi-object tracking tasks in complex, unconstrained surroundings.

In summary, the contributions of this work are as follows:

- We introduce QuaDreamer, the first panoramic data generation engine specifically designed for quadruped robots, capable of generating controllable panoramic videos with natural vibrations from the perspective of a quadruped robot.
- An in-depth analysis of control signals is presented, with Vertical Jitter Encoding (VJE) put forward to extract jitter signals. The Scene-Object Controller (SOC) is developed to effectively regulate video vibrations and object movements. Additionally, we propose a dual-stream design that integrates frequency-domain features and spatial-domain structures — Panoramic Enhancer (PE), aimed at enhancing image quality.
- Through extensive experiments, we demonstrate the effectiveness of QuaDreamer in achieving precise controllability. Additionally, our results indicate that the synthetic data significantly enhances performance in downstream multi-object tracking tasks for quadruped robots, with HOTA improving by 10.14% and MOTA improving by 14.75%.

2 Related Work

Perception for Quadruped Robots. Quadruped robots are increasingly vital in embodied AI for tasks like disaster response [22, 3, 23], assistive navigation for the visually impaired [24], and industrial inspection [1, 25, 2]. Enhancing their perception [6, 26, 27] is critical, with advances in visual navigation [19], vision-language models [28], and object tracking [16]. Panoramic perception [29, 30, 31] is particularly beneficial for quadrupeds, offering comprehensive situational awareness essential for navigating complex terrains and managing dynamic body movements inherent to their locomotion. Complementing perception, accurate state estimation is fundamental for stable control and autonomous navigation [32, 33]. Methods often achieve this by fusing data from various sensors, such as cameras (including panoramic [34, 35, 36]), LiDAR, IMUs, and leg odometry, using techniques like factor graph optimization [37, 38] or specialized algorithms [39]. Despite progress, the complex motion patterns of quadrupeds present significant hurdles for data collection and annotation, especially for multi-sensor or panoramic systems. Furthermore, inherent vibrations can degrade camera data quality, particularly impacting wide-fov panoramic sensors. To overcome these data bottlenecks, our work focuses on efficiently and affordably generating controllable, high-quality panoramic videos from existing data. This approach preserves quadruped-specific motion consistency while enabling diverse styles, providing a valuable resource for advancing perception research for these agile robots.

Diffusion Models for Conditional Generation. The diffusion model [40, 41, 42] was originally designed for image generation by learning the gradual denoising process from a Gaussian noise distribution to an image distribution. This paradigm achieves great success in high-quality video generation [43, 44, 43], such as SVD [44], AnimateDiff [45], and VideoCrafter [46]. Based on that, several works have introduced control signals and diverse control mechanisms to generate diverse controllable video, such as camera pose-driven video generation [47, 48, 49, 50], trajectory-based video generation [51, 52, 53, 54], and layout-to-image generation [55, 56, 14, 57]. Furthermore, works such as GeoDiffusion [58], TrackDiffusion [57], and MagicDrive [14] have shown that generated data can serve as datasets to boost downstream tasks. However, these works focus on global style and viewpoint transformations but fail to capture the subtle spatial jitters typical of quadruped robots. Therefore, this paper explores the potential of Image-to-Video (I2V) diffusion models in generating panoramic videos from the perspective of a quadruped robot, and, based on this, to advance the development of downstream perception models for quadruped robots.

3 Methodology

To generate controllable panoramic videos from the quadruped robot's perspective under intense motion jitter, we propose the QuaDreamer architecture. QuaDreamer integrates control signals into the Latent Diffusion Model [59]. VJE efficiently extracts jitter signals and projects them into the feature domain. The SOC seamlessly integrates jitter and geometric features into U-Net [60] through an attention mechanism. To enhance image quality and adapt to distortion, a dual-stream structure, PE, is designed to refine image quality in the frequency domain and accommodate panoramic distortion in the spatial domain. The architecture of QuaDreamer is shown in Figure 2.

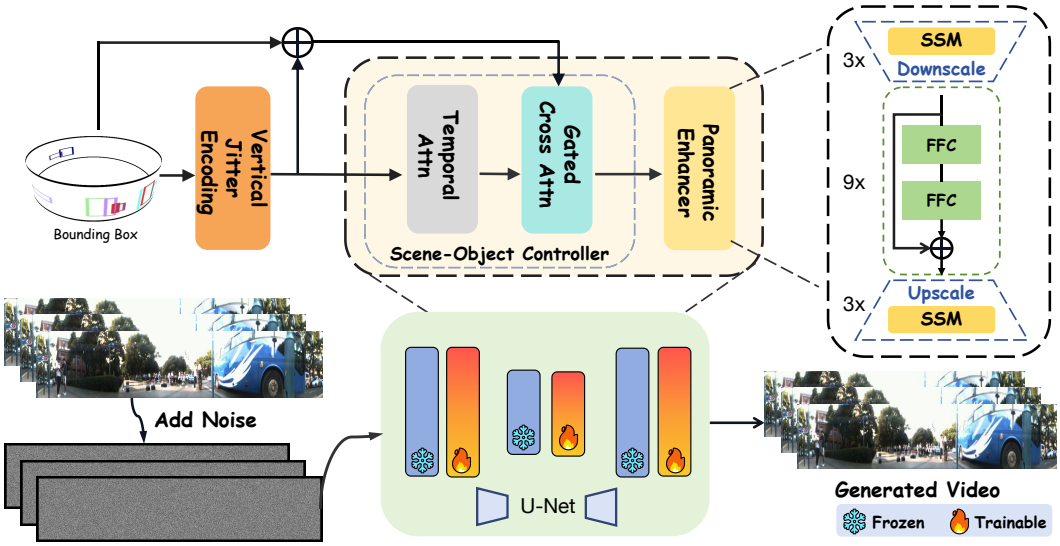

Figure 2: **The overall framework of our QuaDreamer.** Vertical Jitter Encoding extracts jitter signals from bounding boxes and combines them with box information to accurately model motion patterns. To further enhance realism, we incorporate a Scene-Object Controller and a Panoramic Enhancer, which jointly manage object dynamics and refine the representation of panoramic motion.

Preliminary. Latent Diffusion Models [59] combine the advantages of autoencoders and diffusion models by performing the diffusion process in the latent space, significantly reducing computational cost while maintaining high-quality image generation. During the diffusion process, noise is progressively added to the latent representation z until it transforms into standard Gaussian noise at step T. Specifically, starting from the initial latent representation $z_0 \sim q(z_0)$, the latent representation z_t is iteratively updated as follows:

$$q(z_t|z_{t-1}) = \mathcal{N}(z_t; \sqrt{1-\beta_t} z_{t-1}, \beta_t I), \quad t = 1, \ldots, T, \tag{1}$$

where β_t regulates the strength of the noise added at each step during the diffusion process. A U-Net is trained to predict the denoising process, and the MSE loss is used to optimize the network:

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim u(1,T),\epsilon_t\sim\mathcal{N}(0,I)} \left[\|\epsilon_t - \epsilon_\theta(z_t; t, y)\|^2 \right], \tag{2}$$

where y represents the conditional information. By minimizing the loss value, the network learns to reconstruct latent representations that closely match the original data.

Vertical Jitter Encoding. Since the data collector is always present in the panoramic images, we decompose the vertical coordinates of the collector's bounding box $y_{\text{human}}(t)$ into low-frequency and high-frequency components through spectral analysis. The low-frequency component represents the relative displacement between the collector and the quadruped robot, exhibiting stable, slow-changing characteristics. The high-frequency component reflects the vertical jitter induced by the gait characteristics of the quadruped robot, manifested as oscillatory fluctuations in the vertical direction, as shown in Figure 3. To extract the high-frequency part, we used a first-order Butterworth high-pass filter [61]. The frequency response $H(f)$ of the filter is given by the following formula:

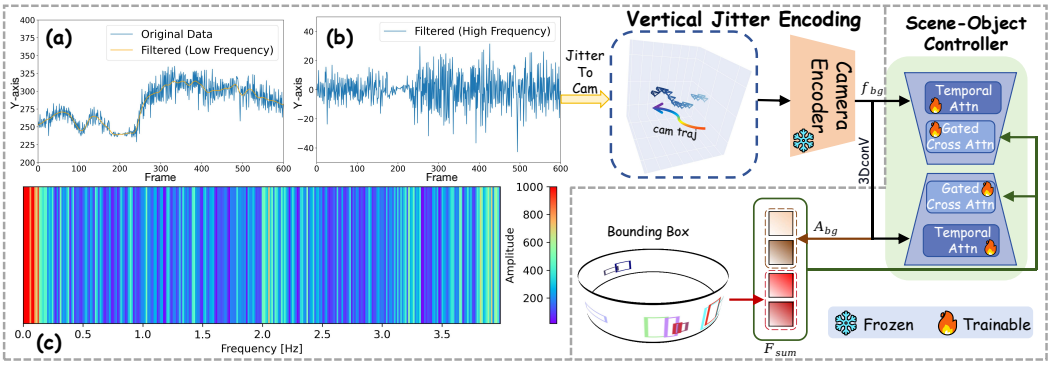

Figure 3: **Illustration of the control components, including VJE and SOC.** (a) shows the original y-axis pixel coordinate data and its low-frequency component; (b) displays the filtered high-frequency jitter data; (c) illustrates the frequency spectrum of the original data.

$$H(f) = \frac{(f/f_c)^{2n}}{1 + (f/f_c)^{2n}}, \tag{3}$$

where f is the frequency, f_c is the cutoff frequency, and n is the order of the filter. The cutoff frequency f_c was set to 0.3 Hz, and the filter order n was set to 1.0. By applying this high-pass filter, we obtained the filtered signal $y_{\text{w}}(t)$, which represents the extracted vertical jitter data of the quadruped robot, expressed as:

$$y_w(t) = \mathcal{F}^{-1}\left[\mathcal{F}(y_{\text{human}}(t)) \cdot H(f)\right], \tag{4}$$

where $\mathcal{F}$ and $\mathcal{F}^{-1}$ represent the Fourier transform and its inverse, respectively.

We first convert the vertical jitter data into the world coordinate system, and the world coordinates for each frame are represented as $P_w^i = (x_w^i, y_w^i, z_w^i)$, where x_w^i is set to half the image width, z_w^i is fixed at 1 to represent the global depth information of the image, and $y_w^i = y_{\text{w}}(i)$ represents the vertical jitter data in the world coordinate system. Next, using the camera intrinsic matrix $K = [[f_x, 0, c_x], [0, f_y, c_y], [0, 0, 1]]$, the world coordinates P_w^i can be mapped into the camera coordinate system $C^i = (x_c^i, y_c^i, z_c^i)$. Following previous works [62, 63], the matrix K can be roughly estimated based on the spatial dimensions of the generated video. Using the formula $C^i = K \cdot P_w^i$, we convert the world coordinates P_w^i into the camera coordinates C^i. Compared to directly using camera coordinates, Plücker embeddings [64] enable more precise control over visual details by effectively representing camera pose. This avoids the imprecision of raw parameters while allowing efficient per-pixel manipulation in image space, thereby enhancing visual consistency and accuracy. To leverage this, we first convert the camera pose C^i into its Plücker embedding P_u^i. Then, the camera encoder processes this representation to compute per-frame background features, yielding $f_{bg}^i = \text{PoseEncoder}(P_u^i)$.

Scene-Object Controller. The core idea of SOC is to decompose the scene into two interrelated components: the background motion field and the object motion field. Specifically, the background motion field is modeled by embedding the background features f_{bg}^i into a temporal attention module to capture the temporal dependencies of the camera trajectory. These features are combined with the image latent features through pixel-level addition, enhancing the control over background jitter. On the other hand, the background motion field is also processed using 3D convolutions to extract spatiotemporal features $A_{\text{bg}}^i = \text{Conv3D}(f_{\text{bg}}^i)$, modeling camera jitter and global motion patterns.

The operation aggregates features across both time and space, enabling the network to learn the time-varying spatial jitter characteristics. Meanwhile, the object motion field is encoded using Fourier Embedding, which maps the bounding box coordinates into a high-dimensional frequency space. This representation enables it to seamlessly integrate with spatiotemporal features in subsequent operations, as expressed by:

$$\Gamma(b_t) = \bigoplus_{k=1}^{K} \left[\sin(\omega_k b_t), \cos(\omega_k b_t)\right], \tag{5}$$

where $\omega_k = \tau^{\frac{k}{K}}$ forms geometric frequency bands ($\tau = 100$), allowing us to capture the multi-scale nature of object motion. Additionally, by incorporating the binary switching mechanism of the visibility mask $m_t \in \{0, 1\}$, the model is able to learn differentiated motion representations of the object in both visible and occluded states:

$$B_t = m_t \cdot \Gamma(b_t) + (1 - m_t) \cdot \varphi_{\text{null}}, \tag{6}$$

Here, φ_{null} represents a learnable zero embedding, used to fill in the missing detection frame information. To integrate object motion with background dynamics, the object motion field $\mathbf{B} = \{B_0, B_1, \ldots, B_t\}$ is fused with the background features $\mathbf{A}_{\text{bg}} = \{A^0_{\text{bg}}, A^1_{\text{bg}}, \ldots, A^i_{\text{bg}}\}$, generating the final feature representation:

$$F_{\text{sum}} = [\text{MLP}(\text{Fourier}(\mathbf{B})), \text{MLP}(\mathbf{A}_{\text{bg}})]. \tag{7}$$

The fused representation F_{sum} is subsequently passed through a gated self-attention mechanism [55] to integrate it with the input visual features.

Panoramic Enhancer. The proposed Panoramic Enhancer (PE) is based on an encoder-decoder architecture, which injects the State Space Model (SSM) [19] symmetrically at specific layers: before the first downsampling layer and after the final downsampling layer. This design performs frequency domain optimization at the intermediate layers, collaboratively addressing the issues of geometric distortion and detail degradation in wide-FoV image generation [16]. As illustrated in Figure 2, the input image first passes through an SSM module at the initial encoder layer, where multi-directional scanning mechanisms model long-range geometric dependencies to suppress edge distortions caused by panoramic unwrapping. For the feature tensor $D \in \mathbb{R}^{B \times C \times H \times W}$ from the encoder downsampling layer, the SSM operation is defined as follows:

$$D^*[b, c, x, y] = \frac{1}{N} \sum_{d \in \text{scan}} F_{S6}(S_d(D[b, c, x, y])), \tag{8}$$

where N represents the number of scans, S_d represents the scanning function, and F_{S_6} is the transformation function for the S6 block [19]. Following three successive downsampling stages, the intermediate layers further process the output features. To enhance frequency-domain representations, we employ residual blocks with Fast Fourier Convolution (FFC) [65]:

$$X_{out} = X + FFC(FFC(X_l, X_g)), \tag{9}$$

where $X = (X_l, X_g)$ is the input divided into local and global components at a 25% and 75% ratio [65]. This design enables effective recovery of periodic structures while avoiding grid artifacts [18]. The decoder progressively reconstructs high-resolution images through a three-stage upsampling process, with a secondary SSM injection applied after the final upsampling layer to enhance geometric continuity. This hierarchical architecture enables phased collaboration between spatial-domain correction and frequency-domain refinement: geometric constraints propagate bidirectionally through SSM-enhanced encoder-decoder layers, while FFC-driven intermediate optimization focuses on detail enhancement.

4 Experimental Results

Dataset and Baselines. The experiments are conducted on the QuadTrack dataset [16], which contains $19,200$ panoramic frames captured from the quadruped robot's perspective, along with $189,876$ high-quality manually annotated object bounding boxes. The training set comprises 702 video sequences, each with 8 frames, and an additional $3,600$ frames are reserved separately for validating downstream tasks. The test set includes 378 sequences (each with 14 frames), specifically used for evaluating generation quality. We adopt TrackDiffusion [57] as the baseline method, where the model inputs consist of panoramic images and object bounding boxes. Additional implementation details are provided in the supplementary material.

Evaluation Metrics. We employed FVD [67], LPIPS [20], PSNR, and SSIM [21] to assess the video quality. To evaluate the controllability, *i.e.*, the correspondence between the input bounding

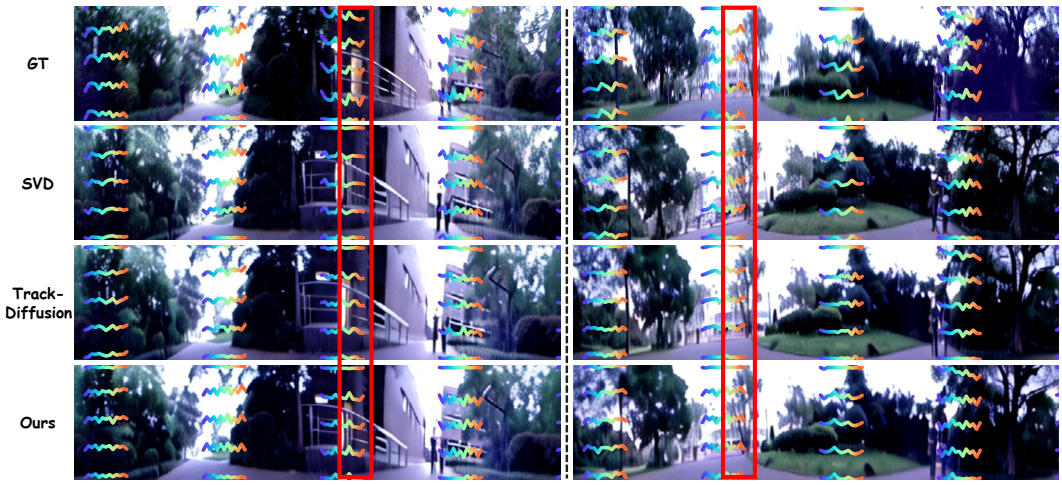

Figure 4: **Visualization results generated on the QuadTrack dataset [16].** The rainbow-colored trajectory is the CoTracker [66] jittering trajectory. The trajectory in the *red boxes* clearly demonstrates that our jitter control is more similar to the ground truth.

boxes and the generated instances, we used tracking metrics such as HOTA [68] and MOTA [69]. These metrics were applied using the OmniTrack model [16] to track objects in the generated video, and then compared with the ground truth tracklets. Additionally, we utilized CoTracker [66] to measure point-tracking consistency for assessing video jitter control performance, denoted as PTrack. Following the conventions of TrackDiffusion [57] and MagicDrive [14], we augmented the training set for downstream tasks with the generated images, validating their contribution to improving task performance.

Method	**Control**	FVD ↓	LPIPS ↓	PSNR↑	SSIM↑	HOTA ↑	MOTA ↑	PTrack↓
SVD [44]	×	831.31	0.2711	14.50	0.3874	9.6113	-22.927	**24.1699**
QuaDreamer*	×	**830.92**	**0.2669**	**14.51**	**0.3893**	**10.332**	**-20.395**	25.5571
TrackDiffusion [57]	box	**887.31**	0.2714	14.27	0.3815	9.6528	-21.451	25.2496
QuaDreamer (Ours)	box&jitter	895.70	**0.2614**	**14.51**	**0.3947**	**9.6542**	**-21.285**	**14.1744**

Table 1: **Comparison of generation fidelity on the QuadTrack dataset** [16]. TrackDiffusion [57] is our baseline. **QuaDreamer*** is a model that does not integrate any control modules.

4.1 Main Results

The Tab. 1 presents the video quality and controllability test results of our proposed QuaDreamer and other baseline models on the QuadTrack dataset. The visualization results are shown in Figure 4. Through comparative experiments with the uncontrolled SVD [44] method of the same type, the effectiveness of the Panoramic Enhancer in improving video quality is validated. Compared to baseline methods, QuaDreamer demonstrates superior overall performance. Specifically, in terms of image quality, LPIPS decreased by 3.68%, SSIM increased by 3.46%, and PSNR improved by 1.68%. Meanwhile, our method is the first to achieve controllable generation based on jitter. Although FVD increased slightly, this is primarily due to the rise in control dimensions. The significant reduction in the PTrack metric (computed by tracking frame-wise jitter with CoTracker [66]) effectively validates the controllability of high-frequency jitter in the generated quadruped robot videos.

4.2 Ablation Study

Ablation of QuaDreamer Framework. In Tab. 2, to verify the effectiveness of our designed modules, we conducted an ablation study on two key modules. Setting (a) represents the baseline method, illustrating the model's initial performance. In setting (b), SOC was used, and improvements in

MOTA (increased by 6.74%) and PTrack (decreased by 36.76%) performance were observed, indicating an enhancement in control effectiveness. In setting (c), the introduction of PE led to a significant improvement in FVD, with a decrease of 6.36%. Finally, setting (d) represents our final model, which integrates both modules, thereby enhancing both controllability and video quality.

Setting	SOC	PE	FVD ↓	LPIPS ↓	PSNR↑	SSIM↑	HOTA ↑	MOTA ↑	PTrack↓
(a)			887.31	0.2714	14.27	0.3815	9.6528	-21.451	25.2496
(b)	✓		896.07	0.2620	14.51	0.3950	9.3924	**-20.006**	15.9676
(c)		✓	**830.92**	0.2669	14.51	0.3893	**10.332**	-20.395	25.5571
(d)	✓	✓	895.70	**0.2614**	**14.51**	**0.3947**	9.6542	-21.285	**14.1744**

Table 2: Ablation of Scene-Object Controller and Panoramic Enhancer. (a) indicates the baseline.

Analysis of Benefits for Downstream Task. In this section, we analyze the advantages of using frames generated by QuaDreamer for data augmentation to train a multi-object tracker. Experiments are conducted on OmniTrack [16], an omnidirectional multi-object tracking task specifically for quadruped robots navigating in unconstrained environments. In OmniTrack, the model is designed to handle the distortions of panoramic images while maintaining multi-object tracking performance and robustness, even when the robot dog undergoes uneven motion. We incorporate these generated frames into the training process to explore how generated data contributes to this challenging downstream task. Using single-frame images and object bounding box annotations from the QuadTrack test set as input, and referring to MagicDrive [14] and TrackDiffusion [57], we generate additional frames for augmented samples and combine them with real images to enrich the training dataset. Subsequently, the model is trained using the default settings of OmniTrack and evaluated on the validation set. The results in Tab. 3 demonstrate the effectiveness of QuadTrack in enhancing the quadruped robot multi-object tracking model. When comparing the model trained with only real data to the model augmented with frames generated by QuaDreamer, a noticeable gain in performance is observed. Specifically, QuaDreamer improved HOTA by **1.437** and MOTA by **10.095**, indicating a significant increase in tracking accuracy. In contrast, the MOTA performance of other models showed a slight decline. This indicates that the data generated by QuaDreamer can effectively support downstream tasks for the quadruped robot. Its strengthened performance stems from its ability to more realistically simulate jitter and provide higher-quality images, thereby enhancing the data's authenticity and diversity, which helps downstream task models better adapt to complex real-world environments.

Method	**HOTA ↑**	**MOTA ↑**	**DetA ↑**	**AssA ↑**
Real Only	14.168	-68.427	12.871	16.052
SVD [44]	14.539(+2.62%)	-72.153(-5.45%)	12.361(-3.96%)	17.666(+10.1%)
TrackDiffusion [57]	14.793(+4.41%)	-72.102(-5.37%)	12.758(-0.88%)	17.596(+9.62%)
QuaDreamer (Ours)	**15.605(+10.1%)**	**-58.332(+14.8%)**	**13.595(+5.63%)**	**18.384(+14.5%)**

Table 3: **The improvement of downstream task performance by different models.** "Real Only" means that OmniTrack is trained without using any data augmentation.

5 Conclusion

We propose QuaDreamer, the first controllable panoramic data generation engine designed explicitly for quadruped robots. QuaDreamer explores a new data extraction method and is capable of producing highly controllable and realistic panoramic videos for perception tasks in complex environments.

6 Limitation

This paper focuses on the generation conditions under the small high-frequency jitter of quadruped robots, but lacks control generation for the movement of other degrees of freedom. In the future, we will explore other control methods for generation models. In addition, leveraging large language models to enable generation under coarsely annotated conditions also presents a valuable and underexplored avenue.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (Grant No. 62473139), in part by the Hunan Provincial Research and Development Project (Grant No. 2025QK3019), and in part by the Open Research Project of the State Key Laboratory of Industrial Control Technology, China (Grant No. ICT2025B20).

References

[1] B. Parkinson, A. Wolf, P. Galambos, and K. Széll. Assessment of the utilization of quadruped robots in pharmaceutical research and development laboratories. In *INES*, pages 000221–000228, 2023.

[2] M. A. V. Torres and F. Pfitzner. Investigating robot dogs for construction monitoring: A comparative analysis of specifications and on-site requirements. *arXiv preprint arXiv:2409.15253*, 2024.

[3] M. Patel, G. Waibel, S. Khattak, and M. Hutter. LiDAR-guided object search and detection in subterranean environments. In *SSRR*, pages 41–46, 2022.

[4] P. Saraf, A. Sarkar, and A. Javed. Terrain adaptive gait transitioning for a quadruped robot using model predictive control. In *ICAC*, pages 1–6, 2021.

[5] X. Xu, E. L. Li, M. Khamis, G. Zhao, and R. Bretin. Understanding dynamic human-robot proxemics in the case of four-legged canine-inspired robots. *arXiv preprint arXiv:2302.10729*, 2023.

[6] N. D. W. Mudalige, E. Nazarova, I. Babataev, P. Kopanev, A. Fedoseev, M. A. Cabrera, and D. Tsetserukou. DogTouch: CNN-based recognition of surface textures by quadruped robot with high density tactile sensors. In *VTC*, pages 1–5, 2022.

[7] A. Li, C. Yang, J. Frey, J. Lee, C. Cadena, and M. Hutter. Seeing through the grass: Semantic pointcloud filter for support surface learning. *IEEE Robotics and Automation Letters*, 8(11): 7687–7694, 2023.

[8] Y. Ouyang, J. Li, Y. Li, Z. Li, C. Yu, K. Sreenath, and Y. Wu. Long-horizon locomotion and manipulation on a quadrupedal robot with large language models. *arXiv preprint arXiv:2404.05291*, 2024.

[9] K. Yang, X. Hu, and R. Stiefelhagen. Is context-aware CNN ready for the surroundings? Panoramic semantic segmentation in the wild. *IEEE Transactions on Image Processing*, 30: 1866–1881, 2021.

[10] K. Yang, X. Hu, H. Chen, K. Xiang, K. Wang, and R. Stiefelhagen. DS-PASS: Detail-sensitive panoramic annular semantic segmentation through SwaftNet for surrounding sensing. In *IV*, pages 457–464, 2020.

[11] K. Tateno, N. Navab, and F. Tombari. Distortion-aware convolutional filters for dense prediction in panoramic images. In *ECCV*, pages 707–722, 2018.

[12] Z. Shen, C. Lin, K. Liao, L. Nie, Z. Zheng, and Y. Zhao. PanoFormer: Panorama transformer for indoor 360° depth estimation. In *ECCV*, pages 195–211, 2022.

[13] H. Ai, Z. Cao, and L. Wang. A survey of representation learning, optimization strategies, and applications for omnidirectional vision. *International Journal of Computer Vision*, pages 1–40, 2025.

[14] 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. In *ICLR*, 2024.

[15] I. D. Miller, A. Cohen, A. Kulkarni, J. Laney, C. J. Taylor, V. Kumar, F. C. Ojeda, A. Cowley, S. S. Shivakumar, E. S. Lee, L. Jarin-Lipschitz, A. Bhat, N. Rodrigues, and A. Zhou. Mine tunnel exploration using multiple quadrupedal robots. *IEEE Robotics and Automation Letters*, 5(2):2840–2847, 2020.

[16] K. Luo, H. Shi, S. Wu, F. Teng, M. Duan, C. Huang, Y. Wang, K. Wang, and K. Yang. Omni-directional multi-object tracking. In *CVPR*, pages 21959–21969, 2025.

[17] J. Wei, J. Zheng, R. Liu, J. Hu, J. Zhang, and R. Stiefelhagen. OneBEV: Using one panoramic image for bird's-eye-view semantic mapping. In *ACCV*, 2024.

[18] Q. Dong, C. Cao, and Y. Fu. Incremental transformer structure enhanced image inpainting with masking positional encoding. In *CVPR*, pages 11348–11358, 2022.

[19] T. Dao and A. Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *ICML*, 2024.

[20] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[21] 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*, 13(4):600–612, 2004.

[22] M. Oh, B. Yu, I. M. A. Nahrendra, S. Jang, H. Lee, D. Lee, S. Lee, Y. Kim, K. C. Marsim, H. Lim, and H. Myung. TRIP: Terrain traversability mapping with risk-aware prediction for enhanced online quadrupedal robot navigation. *arXiv preprint arXiv:2411.17134*, 2024.

[23] D. Lee, I. M. A. Nahrendra, M. Oh, B. Yu, and H. Myung. TRG-Planner: Traversal risk graph-based path planning in unstructured environments for safe and efficient navigation. *IEEE Robotics and Automation Letters*, 10(2):1736–1743, 2025.

[24] H. Hwang, K. Suzuki, N. A. Giudice, J. Biswas, S. I. Lee, and D. Kim. Lessons learned from developing a human-centered guide dog robot for mobility assistance. *arXiv preprint arXiv:2409.19778*, 2024.

[25] J. Lee, J. Kim, and A. D. Ames. Safety-critical autonomous inspection of distillation columns using quadrupedal robots equipped with roller arms. In *IROS*, pages 9094–9101, 2024.

[26] G. Bellegarda, M. Shafiee, and A. Ijspeert. Visual CPG-RL: Learning central pattern generators for visually-guided quadruped locomotion. In *ICRA*, pages 1420–1427, 2024.

[27] M. Elnoor, K. Weerakoon, A. J. Sathyamoorthy, T. Guan, V. Rajagopal, and D. Manocha. AMCO: Adaptive multimodal coupling of vision and proprioception for quadruped robot navigation in outdoor environments. In *IROS*, pages 7687–7694, 2024.

[28] P. Ding, H. Zhao, W. Zhang, W. Song, M. Zhang, S. Huang, N. Yang, and D. Wang. QUAR-VLA: Vision-language-action model for quadruped robots. In *ECCV*, pages 352–367, 2024.

[29] S. Gao, K. Yang, H. Shi, K. Wang, and J. Bai. Review on panoramic imaging and its applications in scene understanding. *IEEE Transactions on Instrumentation and Measurement*, 71: 1–34, 2022.

[30] H. Shi, Y. Zhou, K. Yang, X. Yin, Z. Wang, Y. Ye, Z. Yin, S. Meng, P. Li, and K. Wang. PanoFlow: Learning 360° optical flow for surrounding temporal understanding. *IEEE Transactions on Intelligent Transportation Systems*, 24(5):5570–5585, 2023.

[31] J. Zhang, K. Yang, H. Shi, S. Reiß, K. Peng, C. Ma, H. Fu, P. H. Torr, K. Wang, and R. Stiefelhagen. Behind every domain there is a shift: Adapting distortion-aware vision transformers for panoramic semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(12):8549–8567, 2024.

[32] Y. Nisticò, J. C. V. Soares, L. Amatucci, G. Fink, and C. Semini. MUSE: A real-time multi-sensor state estimator for quadruped robots. *IEEE Robotics and Automation Letters*, 10(5): 4620–4627, 2025.

[33] M. Menner and K. Berntorp. Simultaneous state estimation and contact detection for legged robots by multiple-model kalman filtering. In *ECC*, pages 2768–2773, 2024.

[34] Z. Wang, K. Yang, H. Shi, P. Li, F. Gao, and K. Wang. LF-VIO: A visual-inertial-odometry framework for large field-of-view cameras with negative plane. In *IROS*, pages 4423–4430, 2022.

[35] Z. Wang, K. Yang, H. Shi, P. Li, F. Gao, J. Bai, and K. Wang. LF-VISLAM: A SLAM framework for large field-of-view cameras with negative imaging plane on mobile agents. *IEEE Transactions on Automation Science and Engineering*, 21(4):6321–6335, 2024.

[36] Z. Wang, K. Yang, H. Shi, Y. Zhang, Z. Xu, F. Gao, and K. Wang. LF-PGVIO: A visual-inertial-odometry framework for large field-of-view cameras using points and geodesic segments. *IEEE Transactions on Intelligent Vehicles*, 2024.

[37] D. Wisth, M. Camurri, and M. Fallon. Robust legged robot state estimation using factor graph optimization. *IEEE Robotics and Automation Letters*, 4(4):4507–4514, 2019.

[38] D. Wisth, M. Camurri, and M. Fallon. VILENS: Visual, inertial, lidar, and leg odometry for all-terrain legged robots. *IEEE Transactions on Robotics*, 39(1):309–326, 2023.

[39] J.-H. Kim, S. Hong, G. Ji, S. Jeon, J. Hwangbo, J.-H. Oh, and H.-W. Park. Legged robot state estimation with dynamic contact event information. *IEEE Robotics and Automation Letters*, 6 (4):6733–6740, 2021.

[40] J. Ho, A. Jain, and P. Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pages 6840–6851, 2020.

[41] Y. Song, J. Sohl-Dickstein, D. P. Kingma, A. Kumar, S. Ermon, and B. Poole. Score-based generative modeling through stochastic differential equations. In *ICLR*, 2020.

[42] Z. Zheng, R. Gao, and Q. Xu. Non-cross diffusion for semantic consistency. *arXiv preprint arXiv:2312.00820*, 2023.

[43] J. Ho, T. Salimans, A. A. Gritsenko, W. Chan, M. Norouzi, and D. J. Fleet. Video diffusion models. In *NeurIPS*, pages 8633–8646, 2022.

[44] A. Blattmann, T. Dockhorn, S. Kulal, D. Mendelevitch, M. Kilian, D. Lorenz, Y. Levi, Z. English, V. Voleti, A. Letts, V. Jampani, and R. Rombach. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[45] 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. In *ICLR*, 2024.

[46] H. Chen, Y. Zhang, X. Cun, M. Xia, X. Wang, C. Weng, and Y. Shan. VideoCrafter2: Overcoming data limitations for high-quality video diffusion models. In *CVPR*, pages 7310–7320, 2024.

[47] 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.

[48] 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 *SIGGRAPH*, pages 1–11, 2024.

[49] 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.

[50] D. Xu, W. Nie, C. Liu, S. Liu, J. Kautz, Z. Wang, and A. Vahdat. CamCo: Camera-controllable 3D-consistent image-to-video generation. *arXiv preprint arXiv:2406.02509*, 2024.

[51] W. Wang, Q. Wang, K. Zheng, H. Ouyang, Z. Chen, B. Gong, H. Chen, Y. Shen, and C. Shen. Framer: Interactive frame interpolation. *arXiv preprint arXiv:2410.18978*, 2024.

[52] W. Wu, Z. Li, Y. Gu, R. Zhao, Y. He, D. J. Zhang, M. Z. Shou, Y. Li, T. Gao, and D. Zhang. DragAnything: Motion control for anything using entity representation. In *ECCV*, pages 331–348, 2024.

[53] Z. Wang, Y. Lan, S. Zhou, and C. C. Loy. ObjCtrl-2.5D: Training-free object control with camera poses. *arXiv preprint arXiv:2412.07721*, 2024.

[54] S. Yin, C. Wu, J. Liang, J. Shi, H. Li, G. Ming, and N. Duan. DragNUWA: Fine-grained control in video generation by integrating text, image, and trajectory. *arXiv preprint arXiv:2308.08089*, 2023.

[55] Y. Li, H. Liu, Q. Wu, F. Mu, J. Yang, J. Gao, C. Li, and Y. J. Lee. GLIGEN: Open-set grounded text-to-image generation. In *CVPR*, pages 22511–22521, 2023.

[56] J. Cheng, X. Liang, X. Shi, T. He, T. Xiao, and M. Li. LayoutDiffuse: Adapting foundational diffusion models for layout-to-image generation. *arXiv preprint arXiv:2302.08908*, 2023.

[57] P. Li, K. Chen, Z. Liu, R. Gao, L. Hong, G. Zhou, H. Yao, D.-Y. Yeung, H. Lu, and X. Jia. TrackDiffusion: Tracklet-conditioned video generation via diffusion models. In *WACV*, pages 3539–3548, 2025.

[58] K. Chen, E. Xie, Z. Chen, Y. Wang, L. Hong, Z. Li, and D.-Y. Yeung. GeoDiffusion: Text-prompted geometric control for object detection data generation. In *ICLR*, 2024.

[59] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10674–10685, 2022.

[60] O. Ronneberger, P. Fischer, and T. Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, pages 234–241, 2015.

[61] S. Butterworth. On the theory of filter amplifiers. *Wireless Engineer*, 7(6):536–541, 1930.

[62] Q. Wang, V. Ye, H. Gao, J. Austin, Z. Li, and A. Kanazawa. Shape of motion: 4D reconstruction from a single video. *arXiv preprint arXiv:2407.13764*, 2024.

[63] X. Wang, H. Yuan, S. Zhang, D. Chen, J. Wang, Y. Zhang, Y. Shen, D. Zhao, and J. Zhou. VideoComposer: Compositional video synthesis with motion controllability. In *NeurIPS*, pages 7594–7611, 2023.

[64] V. Sitzmann, S. Rezchikov, B. Freeman, J. Tenenbaum, and F. Durand. Light field networks: Neural scene representations with single-evaluation rendering. In *NeurIPS*, pages 19313–19325, 2021.

[65] R. Suvorov, E. Logacheva, A. Mashikhin, A. Remizova, A. Ashukha, A. Silvestrov, N. Kong, H. Goka, K. Park, and V. Lempitsky. Resolution-robust large mask inpainting with fourier convolutions. In *WACV*, pages 3172–3182, 2022.

[66] N. Karaev, I. Makarov, J. Wang, N. Neverova, A. Vedaldi, and C. Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. *arXiv preprint arXiv:2410.11831*, 2024.

[67] T. Unterthiner, S. Van Steenkiste, K. Kurach, R. Marinier, M. Michalski, and S. Gelly. Towards accurate generative models of video: A new metric & challenges. *arXiv preprint arXiv:1812.01717*, 2018.

[68] J. Luiten, A. Osep, P. Dendorfer, P. Torr, A. Geiger, L. Leal-Taixé, and B. Leibe. HOTA: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 129(2):548–578, 2021.

[69] K. Bernardin and R. Stiefelhagen. Evaluating multiple object tracking performance: The CLEAR MOT metrics. *EURASIP Journal on Image and Video Processing*, 2008:1–10, 2008.