arXiv:2508.00299v1 [cs.CV] 1 Aug 2025

Controllable Pedestrian Video Editing for Multi-View Driving Scenarios via Motion Sequence

Danzhen Fu*, Jiagao Hu*, Daiguo Zhou*, Fei Wang, Zepeng Wang, Wenhua Liao
MiLM Plus, Xiaomi Inc.
{fudanzhen, hujiagao, zhoudaiguo, wangfei11, wangzepeng5, liaowenhua1}@xiaomi.com

Abstract

Pedestrian detection models in autonomous driving systems often lack robustness due to insufficient representation of dangerous pedestrian scenarios in training datasets. To address this limitation, we present a novel framework for controllable pedestrian video editing in multi-view driving scenarios by integrating video inpainting and human motion control techniques. Our approach begins by identifying pedestrian regions of interest across multiple camera views, expanding detection bounding boxes with a fixed ratio, and resizing and stitching these regions into a unified canvas while preserving cross-view spatial relationships. A binary mask is then applied to designate the editable area, within which pedestrian editing is guided by pose sequence control conditions. This enables flexible editing functionalities, including pedestrian insertion, replacement, and removal. Extensive experiments demonstrate that our framework achieves high-quality pedestrian editing with strong visual realism, spatiotemporal coherence, and cross-view consistency. These results establish the proposed method as a robust and versatile solution for multi-view pedestrian video generation, with broad potential for applications in data augmentation and scenario simulation in autonomous driving.

1. Introduction

The rapid advancement of autonomous driving technology has been driven by breakthroughs in artificial intelligence algorithms and the accumulation of large-scale driving datasets. Among the critical components of perception systems, pedestrian detection models require extensive training data to handle complex real-world scenarios. However, the long-tailed distribution characteristic of driving data presents significant challenges, particularly for rare but safety-critical situations such as jaywalking or close-proximity pedestrians. These edge cases, though constituting a minimal proportion in natural driving data, demand robust model performance due to their high-risk nature - creating a paradox where data scarcity hinders the very capabilities that safety requirements necessitate. This underscores the importance of synthetic pedestrian data generation for autonomous driving systems.

Recent developments in diffusion-based video generation have enabled significant progress in controllable human motion synthesis. General-domain human motion generation models [3, 9, 12, 28, 29, 33, 38, 42] demonstrate remarkable capabilities in entertainment applications, where reference images combined with motion sequences from source videos allow generating characters performing specific actions. However, these single-view approaches cannot meet the multi-view requirements of autonomous driving systems, which typically employ multi-camera setups to capture 360° environmental contexts. This domain-specific requirement necessitates specialized solutions distinct from generic video generation frameworks.

In the autonomous driving domain, various multi-view video generation methods have been proposed [5, 6, 19, 30, 31, 39, 40]. For example, MagicDrive3D [5] and DriveDreamer4D [39] utilize 3D Gaussian Splatting [13] for novel view synthesis, while Vista [6] integrates multi-modal control signals for future frame prediction. Panacea [31] adopts a two-stage pipeline that first generates multi-view images and subsequently uses them as initial frames for video generation. Although these methods achieve multi-view consistency in general driving scenes, they lack the ability to perform local pedestrian editing. To address this gap, we propose a dedicated framework for controllable pedestrian editing in multi-view driving scenarios.

In summary, We present a diffusion-based video editing framework that synergizes video inpainting with human motion sequence to achieve pedestrian manipulation in multi-view driving scenarios. Our principal contributions are threefold:

1) **Local Video Editing Architecture**: We introduce a local cropping-and-generation paradigm that enables spatially-

* Equal contribution.

aware video synthesis through a dedicated video inpainting scheme.

2) **Dynamic Pedestrian Region Cropping**: To tackle the scale-variation challenges in autonomous driving videos, we propose a two-stage adaptive processing pipeline: First, dynamically crop pedestrians according to their detection bounding box sizes, then resize the cropped regions to a uniform scale, enabling subsequent video generation to be performed consistently at this standardized scale.
3) **Multi-View Consistency Framework**: We further propose a cross-view temporal-consistent video generation approach that ensures geometric alignment and semantic consistency across multi-camera perspectives.

Experimental results validate our framework's superiority in generating photorealistic pedestrian sequences with precise motion control across camera perspectives. These results demonstrate our approach as a robust solution for safety-critical pedestrian scenario generation, with potential applications in perception model augmentation and autonomous vehicle simulation systems.

2. Related Works

2.1. Diffusion-Based Image Generation

Diffusion models [10, 20, 23, 24] have emerged as the dominant paradigm in image generation, achieving state-of-the-art performance across various benchmarks. To address computational efficiency, Latent Diffusion Model [23] introduces a two-stage framework that first encodes images into a compressed latent space, performs denoising in this lower-dimensional domain, and subsequently decodes to the pixel space - a strategy that balances generation quality and computational feasibility.

Particularly relevant to our work are conditional diffusion approaches that enhance controllability: ControlNet [37] and T2I-Adapter [18] introduce additional encoding pathways to preserve spatial constraints from edge maps, segmentation masks, and pose skeletons, enabling precise generation control through structural guidance. For identity-preserving generation, IP-Adapter [36] employs cross-attention mechanisms to inject image prompts as reference conditions, while ObjectStitch [26] utilizes CLIP [22] embeddings to maintain semantic consistency during image composition. These works demonstrate the critical role of condition injection strategies in achieving both visual fidelity and controllability in diffusion-based frameworks.

2.2. Diffusion-Based Video Generation

The diffusion model's remarkable success in image generation has driven its extension to video synthesis. Many approaches [4, 14, 21, 25, 32] focus on enhancing temporal coherence while maintaining spatial quality through two primary strategies: motion-aware modules integration and spatiotemporal architecture redesign. Pioneering works like AnimateDiff [8] demonstrate effective video generation by inserting motion modules into pre-trained text-to-image pipelines, leveraging inter-frame attention mechanisms to preserve temporal consistency without modifying base architectures. Meanwhile, recent advancements [15, 27, 35, 41] in native video diffusion models have generalized 2D denoising networks and variational autoencoders (VAEs) to 3D spatiotemporal domains. These 3D approaches achieve improved frame consistency but often require substantial computational resources.

2.3. Human Motion Video Generation

Recent advancements in pedestrian video generation have witnessed significant progress in controllable human motion synthesis. Early approaches primarily focused on text-driven generation, such as Text2Performer [11] which decoupled appearance descriptions from motion descriptions through dedicated language encoding. However, these purely language-based strategies often struggled to achieve precise control over both human appearance and motion dynamics. The subsequent Follow-Your-Pose [17] framework introduced a two-stage paradigm, first generating base character images via text-to-image synthesis, then guiding motion generation using pose sequences.

Current research trends have shifted towards more sophisticated control mechanisms through pose sequence conditioning and reference image guidance [3, 9, 12, 28, 29, 33, 38, 42]. These methods leverage motion sequences for precise posture control while employing human reference images (refs) to encode appearance features, enabling person-specific animation generation. Notable approaches include Disco [28], MagicPose [3], MagicAnimate [33], DreamPose [12], Animate Anyone [9], and Champ [42], which integrate action sequence encoding for motion control with CLIP-based reference image encoding for appearance conditioning. Among these, Disco [28] uniquely incorporates background encoding to facilitate scene replacement in generated videos. The Champ [42] framework further advances controllability by utilizing multi-modal motion sequences, including SMPL [16] (Skinned Multi-Person Linear) model representations that enhance 3D motion perception and produce more vivid, anatomically accurate animations.

The quality of motion guidance has also benefited from improved pose estimation techniques. Methods like Animate Anyone and Champ employ DWPose [34] - an enhanced version of OpenPose [2] - which demonstrates superior accuracy in skeletal keypoint detection. This technical advancement contributes significantly to the generation of high-fidelity human motion videos with natural articulation and spatial coherence. The work presented in Dense-

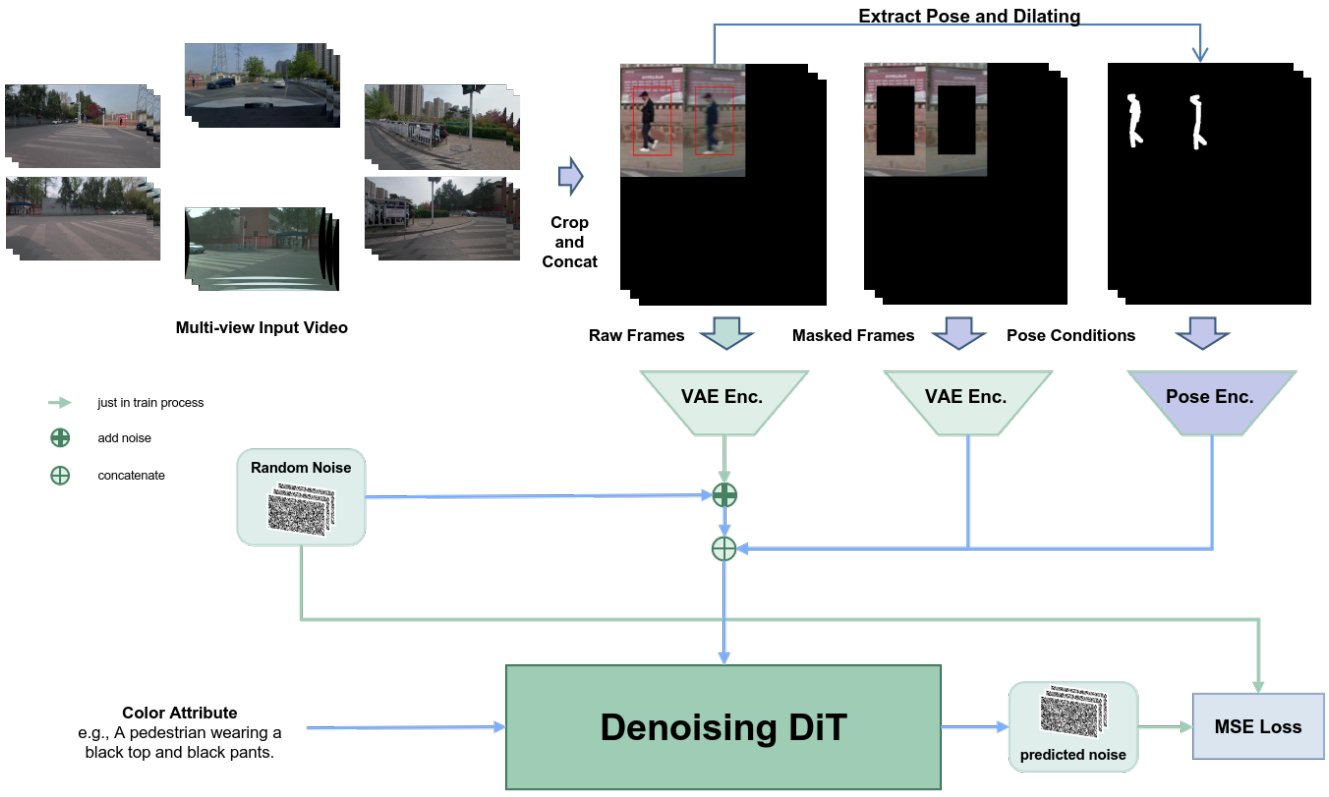


Figure 1. Architecture of the proposed pedestrian editing pipeline with motion sequence control.

Pose [7] aims to establish dense correspondences between an RGB image and a surface-based representation.

3. Method

Our proposed framework for controllable pedestrian generation in multi-view driving scenarios is illustrated in Figure 1. The approach adopts a diffusion-based denoising paradigm, leveraging a ControlNet-like architecture to construct a video inpainting pipeline. Specifically, we first crop the target regions from the multi-view footage where pedestrians are to be inserted or replaced. The model then synthesizes pedestrian video sequences conditioned on predefined motion templates. Finally, the generated pedestrian content is seamlessly integrated into the original scenes across all views, enabling temporally coherent and spatially consistent pedestrian insertion and replacement.

3.1. Dynamic Pedestrian Region Cropping

When fixed-size cropping was applied without dynamic adaptation, two critical visual artifacts emerged in multi-view driving scenarios. For close-range pedestrians, the rigid crop dimensions often failed to capture the entire body structure. This resulted in truncated limbs and incomplete poses, leading to visually implausible generations with distorted anatomical proportions or fragmented motion continuity. For distant pedestrians, the human body occupies an excessively small portion of the cropped frame, which significantly increases the difficulty of generation and hampers the recovery of fine-grained details. To overcome this limitation, we propose a dynamic pedestrian cropping strategy that standardizes pedestrian scales across views to simplify generation while enabling seamless multi-view integration. This dynamic crop-and-resize strategy is as illustrated in Figure 2.

3.2. Multi-View Consistency Framework

We utilize multi-view image sequences annotated with LiDAR-based object detection. Given precise camera intrinsic and extrinsic calibration parameters, we first project 3D pedestrian annotations onto all camera views to obtain view-specific bounding boxes (bbox). These bbox coordinates are then expanded by a fixed margin ratio to capture contextual information and cropped pedestrian regions are resized to a uniform resolution (e.g., 480×240 pixels). For occluded or out-of-view pedestrians in specific camera angles, we employ zero-filled placeholder images to maintain view consistency. Finally, the processed multi-view pedestrian sequences are spatially organized into a composite image (e.g., arranged in a 2-row by 3-column layout) for uni-

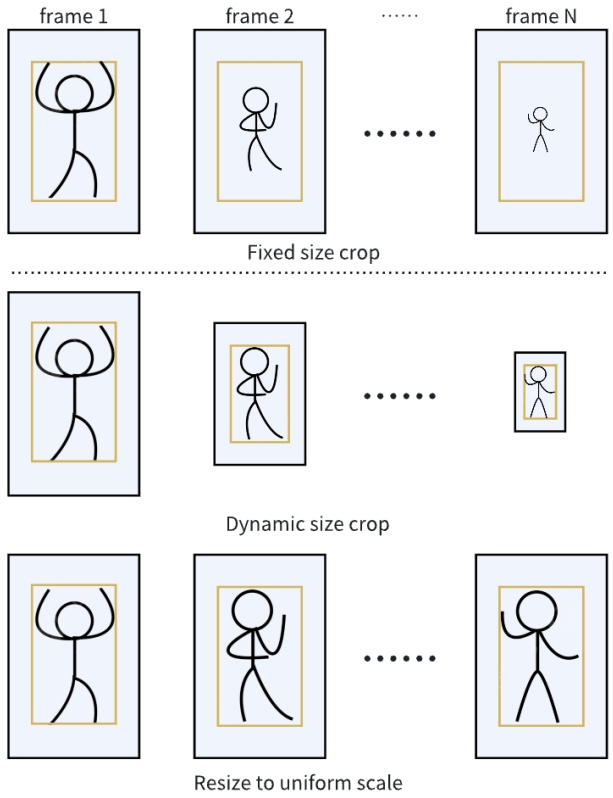


Figure 2. Dynamic-scale cropping with subsequent resizing to uniform scale

fied generation.

3.3. Video Inpainting

We implement a standard video inpainting framework through mask-based conditioning, where the target pedestrians to be replaced or generated are masked out during training. The unmasked regions remain fully visible to the model and are encoded via VAE to produce control latents that preserve background context. These latents serve as critical constraints for background reconstruction while allowing pedestrian content manipulation.

Specifically, our masking protocol employs pedestrian bounding boxes as base masks, followed by spatial expansion and cropping operations. Each bounding box undergoes independent isotropic expansion along width and height dimensions, followed by region cropping that results in 1.6× the original bounding box dimensions. The expanded buffer zone facilitates seamless blending between generated pedestrians and background by maintaining visible transition areas.

3.4. Pedestrian Motion Sequence Control

Recent advancements in controllable video generation have demonstrated the effectiveness of using human pose sequences as conditional inputs for synthesizing realistic human motions. Notable frameworks such as Disco [28], AnimateAnyone [9], and Champ [42] have explored different strategies in this domain. AnimateAnyone [9] employ 2D skeletal keypoint sequences as their primary control signals, while Champ innovatively combines four distinct pose representations - including depth images, normal maps, semantic maps obtained from SMPL sequences, and skeleton-based motion guidance - through collaborative control to achieve superior detail preservation in generated videos.

In our approach, we adopt the minimalist skeletal keypoint representation for pedestrian motion control, motivated by two critical observations: First, in autonomous driving scenarios, pedestrians typically occupy only a small portion of the field of view in individual camera feeds, making fine-grained pose details less perceptible. Second, maintaining cross-view consistency across multi-camera systems proves more crucial than local detail fidelity when synthesizing traffic scenarios. To implement this, we utilize the open-source DWPose [34] framework for keypoint extraction. For each camera view, we first perform cropping operations to isolate individual subjects before applying the pose estimation model to obtain temporally coherent keypoint sequences. These sequences are passed through a keypoint encoder to obtain motion latents, which are concatenated with inpainting latents from Section 3.2 to control pedestrian generation while maintaining background consistency.

This method follows standard denoising approaches: during training, VAE encodes videos into latents with added random Gaussian noise. The model predicts the noise through MSE loss minimization.

4. Experimental Results

4.1. Training Phase

We trained the pedestrian video generation model on the nuScenes [1] dataset using CogVideoX v1.5 [35] as the base architecture. For single-view processing, pedestrian regions were cropped and resized to 480×240 resolution, while six-view inputs were tiled in a 2-row×3-column layout to form a composite resolution of (480×2)×(240×3). The model was trained on 85-frame sequences extracted from continuous video clips.

Motion cues were obtained through the DWpose estimator, followed by morphological dilation operations to enhance pose conditioning robustness. For view selection, we annotated the dominant viewpoint (most frequent occurrence) for each candidate pedestrian instance and localized their 2D bounding boxes. A pretrained video understanding model was then employed to extract clothing attributes, generating descriptions like “a pedestrian wearing a white top and black pants” where only color terms (e.g.,

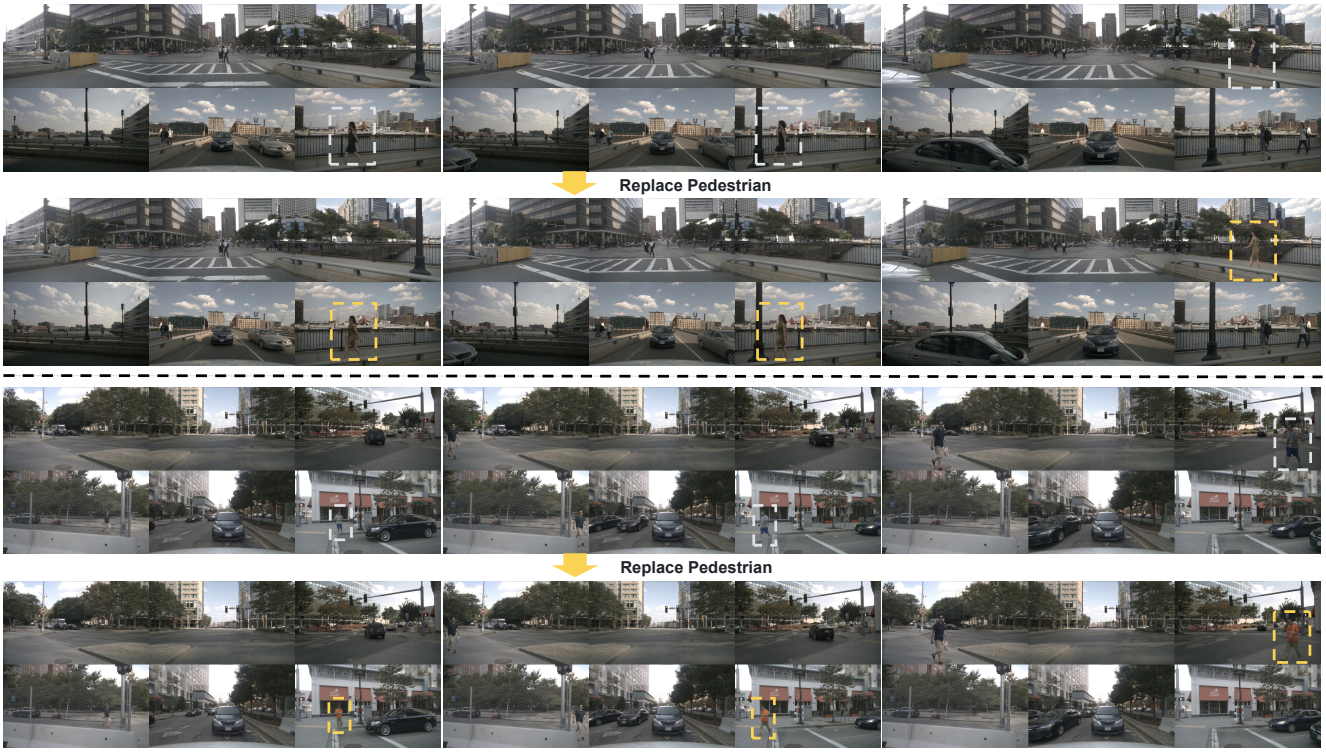


Figure 3. Qualitative results of pedestrian replacement in a multi-view scenario. Gray dashed boxes indicate the original pedestrian, while yellow dashed boxes highlight the generated pedestrian.

white/black) are replaceable within a fixed template structure.

During training, original videos were encoded into latent representations through the VAE module. The control conditions incorporated two components, Masked latent codes from the original video (with pedestrian regions masked out) and Encoded pose sequences. These were concatenated with noisy latent vectors following standard diffusion training protocols. The subsequent training procedures (noise prediction, optimization objectives) strictly followed the implementation of CogVideoX v1.5 [35].

4.2. Inference Phase

During inference, our framework first specifies a 3D motion sequence which is projected to multi-view perspectives using calibrated camera parameters to generate pose guidance maps. We then apply temporal-spatial masks to define editing regions and construct inpainting conditions. The pose maps and inpainting conditions are separately encoded via pose encoder and VAE encoder, then concatenated with noise latents in the DiT network architecture following training configurations. Text prompts (e.g., "A pedestrian wearing a white top and black pants.") are converted into cross-attention embeddings through CLIP-based text encoder and integrated into DiT's cross-attention layers.

BEVFormer	w/o Synthetic Data	w/ Synthetic Data
$AP_{dist_0.5}$	0.1012	0.1063
AP_{dist_1}	0.3552	0.3683
AP_{dist_2}	0.5971	0.6136
AP_{dist_4}	0.7173	0.7424
mAP	0.4427	**0.4577**

Table 1. 3D mAP metric for pedestrian detection.

4.3. Quantitative Assessment

We adopt the 3D mean Average Precision (mAP) as the primary evaluation metric, following the BEVFormer framework - a state-of-the-art multi-view perception model for autonomous driving. While mAP is a standard metric in 2D object detection tasks, its 3D counterpart employs distinct criteria for sample classification. Specifically, instead of using Intersection-over-Union (IoU) thresholds in image space, we utilize 2D center distance thresholds in Bird's Eye View (BEV) space, with standard thresholds set at 0.5m, 1m, 2m, 4m. The final mAP score is obtained by averaging precisions across all distance thresholds.

To validate the effectiveness of our pedestrian generation approach, we conduct controlled experiments on the nuScenes dataset. The baseline model is trained on orig-

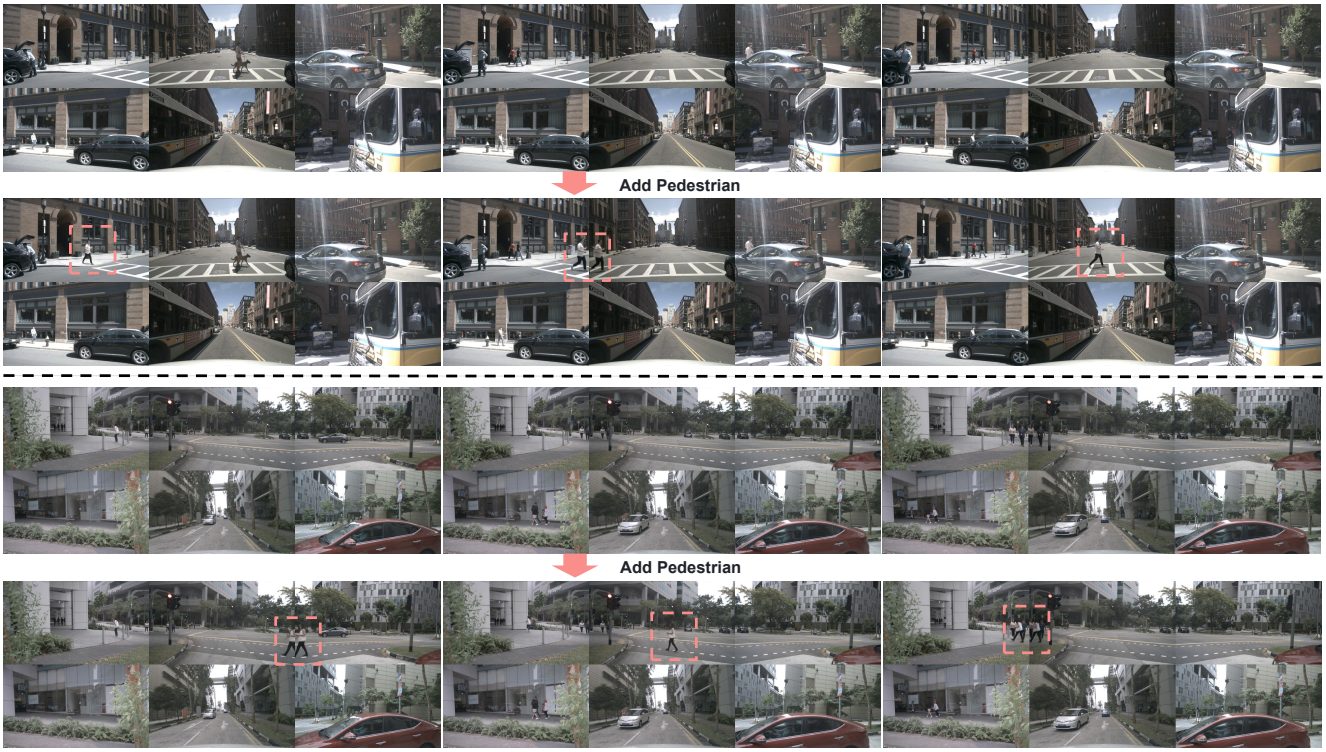


Figure 4. Qualitative results of pedestrian insertion in a multi-view scenario. Inserted pedestrians are highlighted by pink dashed boxes.

inal data, while our proposed method replaces pedestrian instances in the training set with synthesized counterparts. Both models are evaluated on the nuScenes val set using mAP specifically for pedestrian class. Performance improvement in this metric would indicate that our motion-controllable pedestrian generation approach positively contributes to perception model training.

Table 1 presents the mAP evaluation results for the BEVFormerV2-R50-T1-Base model, a representative variant in the BEVFormer series. Compared to the baseline model trained on real-world data, our approach demonstrates significant performance gains in pedestrian detection when incorporating synthesized pedestrian samples for training. Specifically, the model trained with our generated data achieves notable improvements across all BEV distance thresholds. These results validate our method's capability in generating high-fidelity, multi-view pedestrian sequences that effectively enhance perception models' ability to detect pedestrians in complex driving scenarios.

The quantitative improvements in mAP scores provide empirical evidence that our motion-controllable pedestrian generation framework addresses critical challenges in multi-view detection tasks. By preserving spatial-temporal consistency while enabling motion sequence control, our synthesized data contributes to better generalization of the perception model.

4.4. Qualitative Assessment

In this section, we qualitatively evaluate the capabilities of our method. Beyond modifying existing pedestrians, our framework also supports their insertion and removal, enabling a wide range of flexible editing operations within multi-view video sequences.

Pedestrian Replacement Figure 3 showcases our method's effectiveness in replacing pedestrians across synchronized multi-camera video streams. The generated pedestrians maintain consistent appearance across views and frames, and exhibit natural, temporally coherent motion. Furthermore, their behaviors are fully controllable via user-specified input conditions, enabling precise and customizable edits in complex, dynamic environments.

Pedestrian Insertion As illustrated in Figure 4, our approach enables the realistic insertion of synthetic pedestrians into vehicle-mounted multi-view footage. The inserted figures are visually indistinguishable from real pedestrians, and the synthesized content exhibits no observable artifacts or inconsistencies in either spatial or temporal dimensions. These results demonstrate strong generalization and highlight the model's ability to maintain coherence across multiple viewpoints.

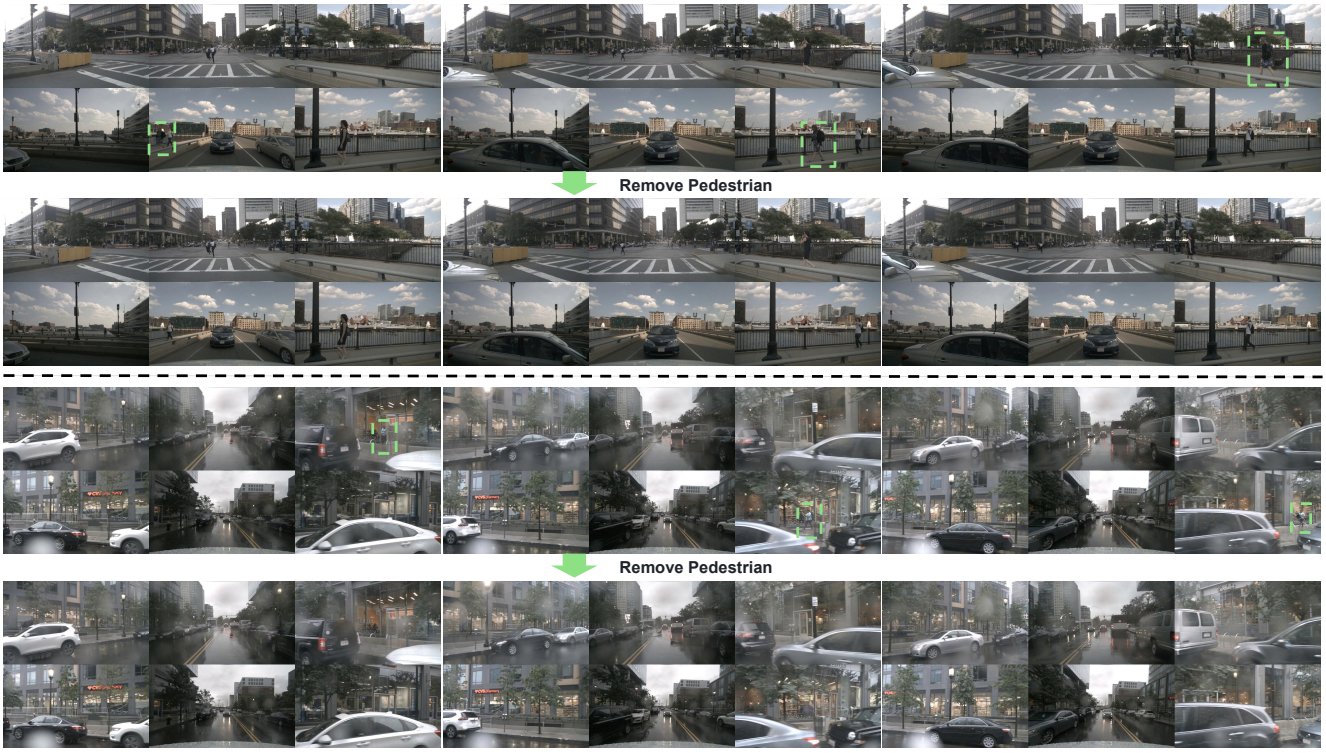


Figure 5. Qualitative results of pedestrian removal in a multi-view scenario. Pedestrians targeted for removal are indicated by green dashed boxes.

Pedestrian Removal Pedestrian removal is achieved by setting the corresponding pose masks to zero, thereby suppressing the pedestrian signal while preserving the surrounding scene context. As shown in Figure 5, the model successfully eliminates the specified pedestrians and reconstructs the occluded regions with high visual fidelity. The inpainted backgrounds blend seamlessly into the original scene, exhibiting no discernible discontinuities or visual artifacts.

Clothing Color Control Figure 6 demonstrates the ability of our model to precisely control pedestrian clothing color through simple modifications in textual prompts. The generated results exhibit photo-realistic appearance and maintain contextual consistency across frames and views. This attribute-level control showcases the model's versatility in enabling fine-grained, text-driven visual edits while preserving both spatial alignment and temporal coherence.

In summary, these qualitative results highlight the robustness and flexibility of our method in executing comprehensive pedestrian editing tasks—ranging from replacement and insertion to removal and appearance customization. The edited sequences exhibit high perceptual quality, strong spatio-temporal consistency, and seamless integration within complex, multi-view scenes. These capabilities position our approach as a powerful and generalizable tool for controllable video editing in real-world applications.

5. Conclusion

This paper presents a novel multi-view pedestrian editing framework tailored for autonomous driving scenarios. Our approach synergistically integrates video inpainting and ControlNet-like architectural components to enable dynamic pedestrian synthesis through pose-guided cropping and stitching mechanisms. By leveraging clothing color prompts and pose sequence controls, the framework achieves comprehensive editability over pedestrian attire color and motion patterns in multi-view video generation. Notably, the proposed method ensures spatio-temporal consistency across all views while maintaining strict geometric alignment with the driving context.

Through empirical evaluation, we demonstrate the framework's effectiveness in generating photorealistic, multi-view consistent pedestrian sequences that preserve both appearance attributes and motion dynamics. The successful application of our synthesized data in BEVFormer model training highlights its practical value in expanding limited real-world pedestrian samples for autonomous driving perception tasks. Experimental results show that the generated data improves detection performance. This validates our method's capability to

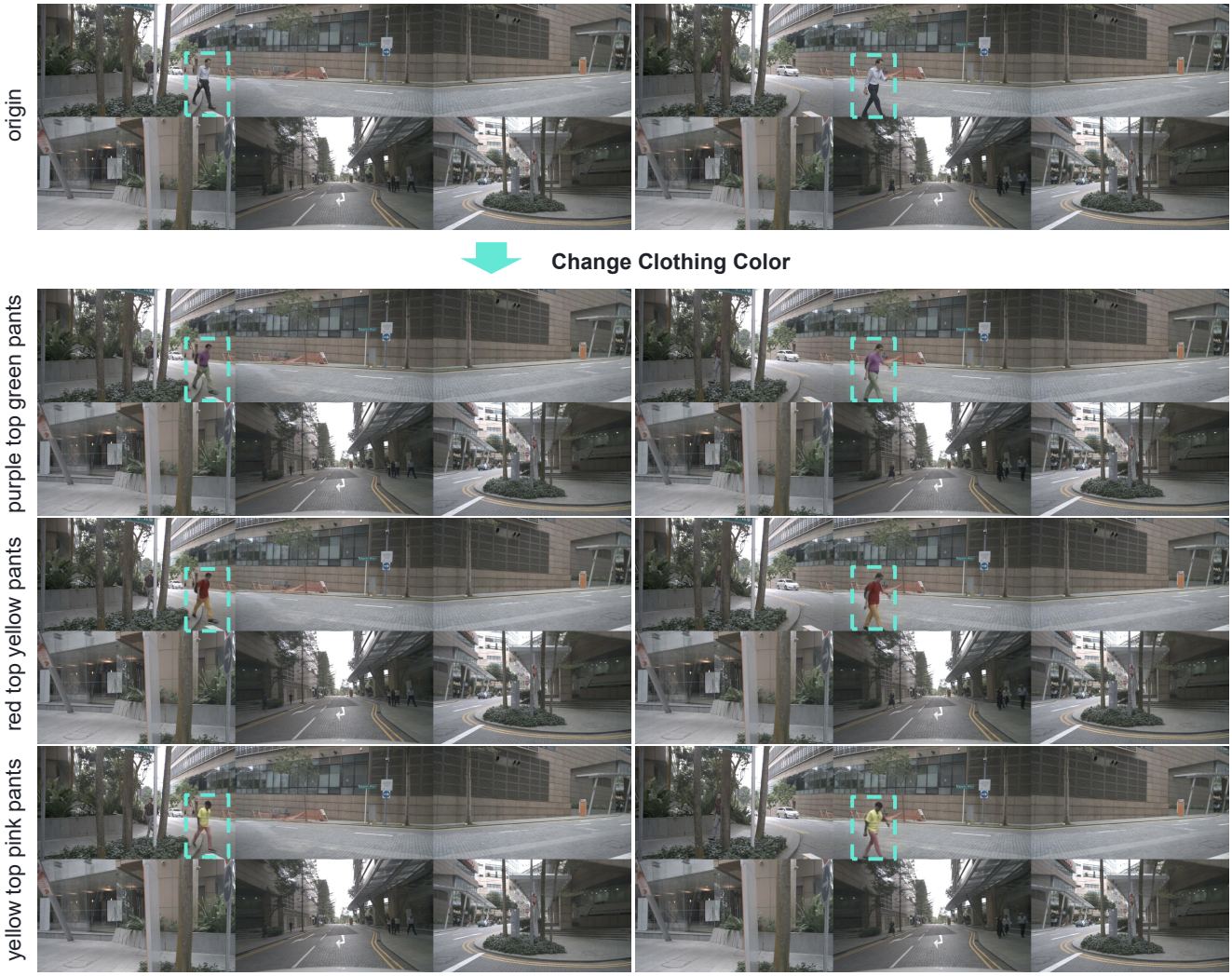


Figure 6. Qualitative results of clothing color manipulation in a multi-view scenario. Edited pedestrians are highlighted with cyan dashed boxes.

produce high-quality, spatial-temporal consistency pedestrian samples that faithfully reflect real-world driving contexts.

References

[1] 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. 4

[2] Zhe Cao, Tomas Simon, Shih-En Wei, and Yaser Sheikh. Realtime multi-person 2d pose estimation using part affinity fields. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7291–7299, 2017. 2

[3] Di Chang, Yichun Shi, Quankai Gao, Jessica Fu, Hongyi Xu, Guoxian Song, Qing Yan, Yizhe Zhu, Xiao Yang, and Mohammad Soleymani. Magicpose: Realistic human poses and facial expressions retargeting with identity-aware diffusion. *arXiv preprint arXiv:2311.12052*, 2023. 1, 2

[4] Patrick Esser, Johnathan Chiu, Parmida Atighehchian, Jonathan Granskog, and Anastasis Germanidis. Structure and content-guided video synthesis with diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 7346–7356, 2023. 2

[5] Ruiyuan Gao, Kai Chen, Zhihao Li, Lanqing Hong, Zhenguo Li, and Qiang Xu. Magicdrive3d: Controllable 3d generation for any-view rendering in street scenes. *arXiv preprint arXiv:2405.14475*, 2024. 1

[6] 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. *arXiv preprint arXiv:2405.17398*, 2024. 1

[7] Rıza Alp Güler, Natalia Neverova, and Iasonas Kokkinos.

Densepose: Dense human pose estimation in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7297–7306, 2018. 3

[8] 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. *arXiv preprint arXiv:2307.04725*, 2023. 2

[9] Li Hu. Animate anyone: Consistent and controllable image-to-video synthesis for character animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 1, 2, 4

[10] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: creative and controllable image synthesis with composable conditions. In *Proceedings of the 40th International Conference on Machine Learning*, pages 13753–13773, 2023. 2

[11] Yuming Jiang, Shuai Yang, Tong Liang Koh, Wayne Wu, Chen Change Loy, and Ziwei Liu. Text2performer: Text-driven human video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22747–22757, 2023. 2

[12] Johanna Karras, Aleksander Holynski, Ting-Chun Wang, and Ira Kemelmacher-Shlizerman. Dreampose: Fashion image-to-video synthesis via stable diffusion. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22623–22633. IEEE, 2023. 1, 2

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 42(4):139–1, 2023. 1

[14] Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15954–15964, 2023. 2

[15] Weijie Kong, Qi Tian, Zijian Zhang, Rox Min, Zuozhuo Dai, Jin Zhou, Jiangfeng Xiong, Xin Li, Bo Wu, Jianwei Zhang, et al. Hunyuanvideo: A systematic framework for large video generative models. *arXiv preprint arXiv:2412.03603*, 2024. 2

[16] Matthew Loper, Naureen Mahmood, Javier Romero, Gerard Pons-Moll, and Michael J Black. Smpl: a skinned multi-person linear model. *ACM Transactions on Graphics (TOG)*, 34(6):1–16, 2015. 2

[17] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024. 2

[18] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI conference on artificial intelligence*, pages 4296–4304, 2024. 2

[19] Chaojun Ni, Guosheng Zhao, Xiaofeng Wang, Zheng Zhu, Wenkang Qin, Guan Huang, Chen Liu, Yuyin Chen, Yida Wang, Xueyang Zhang, et al. Recondreamer: Crafting world models for driving scene reconstruction via online restoration. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 1559–1569, 2025. 1

[20] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*, 2021. 2

[21] Chenyang Qi, Xiaodong Cun, Yong Zhang, Chenyang Lei, Xintao Wang, Ying Shan, and Qifeng Chen. Fatezero: Fusing attentions for zero-shot text-based video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15932–15942, 2023. 2

[22] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 2

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2

[24] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 2

[25] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022. 2

[26] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. Objectstitch: Object compositing with diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 2

[27] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 2

[28] Tan Wang, Linjie Li, Kevin Lin, Yuanhao Zhai, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. Disco: Disentangled control for realistic human dance generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9326–9336, 2024. 1, 2, 4

[29] Xiang Wang, Shiwei Zhang, Changxin Gao, Jiayu Wang, Xiaoqiang Zhou, Yingya Zhang, Luxin Yan, and Nong Sang. Unianimate: Taming unified video diffusion models for consistent human image animation. *arXiv preprint arXiv:2406.01188*, 2024. 1, 2

[30] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, Jiagang Zhu, and Jiwen Lu. Drivedreamer: Towards real-world-drive world models for autonomous driving. In *European Conference on Computer Vision*, pages 55–72. Springer, 2024. 1

[31] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6902–6912, 2024. 1

[32] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7623–7633, 2023. 2

[33] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. Magicanimate: Temporally consistent human image animation using diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 1, 2

[34] 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. 2, 4

[35] Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang, Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024. 2, 4, 5

[36] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2

[37] 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

[38] 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. 1, 2

[39] Guosheng Zhao, Chaojun Ni, Xiaofeng Wang, Zheng Zhu, Xueyang Zhang, Yida Wang, Guan Huang, Xinze Chen, Boyuan Wang, Youyi Zhang, et al. Drivedreamer4d: World models are effective data machines for 4d driving scene representation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12015–12026, 2025. 1

[40] 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. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10412–10420, 2025. 1

[41] Zangwei Zheng, Xiangyu Peng, Tianji Yang, Chenhui Shen, Shenggui Li, Hongxin Liu, Yukun Zhou, Tianyi Li, and Yang You. Open-sora: Democratizing efficient video production for all. *arXiv preprint arXiv:2412.20404*, 2024. 2

[42] Shenhao Zhu, Junming Leo Chen, Zuozhuo Dai, Zilong Dong, Yinghui Xu, Xun Cao, Yao Yao, Hao Zhu, and Siyu Zhu. Champ: Controllable and consistent human image animation with 3d parametric guidance. In *European Conference on Computer Vision*, pages 145–162. Springer, 2024. 1, 2, 4