InstaDrive: Instance-Aware Driving World Models for Realistic and Consistent Video Generation

Zhuoran Yang[1] Xi Guo[2] Chenjing Ding[2] Chiyu Wang[2] Wei Wu[2,3] Yanyong Zhang[1*]

[1]University of Science and Technology of China [2]SenseAuto [3]Tsinghua University

shanpoyang@mail.ustc.edu.cn, {guoxi,dingchenjing}@sensetime.com

{wangchiyu, wuwei}@senseauto.com

{yanyongz}@ustc.edu.cn

[*] Corresponding Author

InstaDrive: Instance-Aware Driving World Models for Realistic and Consistent Video Generation

Supplementary Material

1. More Experimental Details

We provide a project page for additional video results: https://github.com/InstaDrive.

1.1. Training Details

Our method is implemented based on OpenSora [7]. Initially, we train for 20k iterations on the front-view videos from the NuScenes training set. Next, to adapt to multi-view positional encoding, we froze the backbone and fine-tuned the patch embedder for 2k iterations. Finally, we added all the control modules and trained the entire model for 100k iterations with a mini-batch size of 1. All training inputs were set to 16x256x448 and performed on 8 A100 GPUs. Additionally, during training, we set a 0.2 probability of not adding noise to the first frame and assigned a timestep of 0 to the first frame, enabling the model to have image-to-video generation capability. As a result, during testing, the model can autoregressively iterate. Experimental results show that our method can stably generate over 200 frames.

1.2. Perception Evaluation Details

In Tab. **??**, StreamPETR was re-implemented from official config due to resolution differences with Panacea. Stronger baselines are harder to surpass, further highlighting our data's effectiveness.

2. More Quantitative Results

2.1. Evaluation on Planning Task.

We also assess our model in the planning task in autonomous driving. A high-quality planning system not only perceives the current environment but also maintains stable temporal understanding of dynamic objects (e.g., pedestrians, vehicles) to generate smooth and safe trajectories. Therefore, planning task performance serves as a comprehensive measure of both instance-level temporal consistency and spatial geometric fidelity.

To validate our approach, we evaluate the generated nuScenes validation data using pretrained planning model UniAD [4] in Tab. 1. The L2 loss and collision rates closely match the performance of the original data, demonstrating clear benefits from improved temporal consistency and spatial fidelity.

Eval Data	L2(m)↓				Col. Rate↓			
	1s	2s	3s	Avg.	1s	2s	3s	Avg.
Oracle	0.48	0.96	1.65	1.03	0.0005	0.0017	0.0071	0.0031
InstaDrive	0.67	1.54	2.69	1.63	0.0033	0.0095	0.0189	0.0105

Table 1. Comparison on planning task using the pre-trained planning model UniAD [4]. We resize all generated images to 900×1600 to ensure a unified evaluation standard. The L2 loss and collision rates closely match the performance of the original data, highlighting the benefits of enhanced temporal consistency and spatial fidelity.

2.2. Ablation Study for SGA

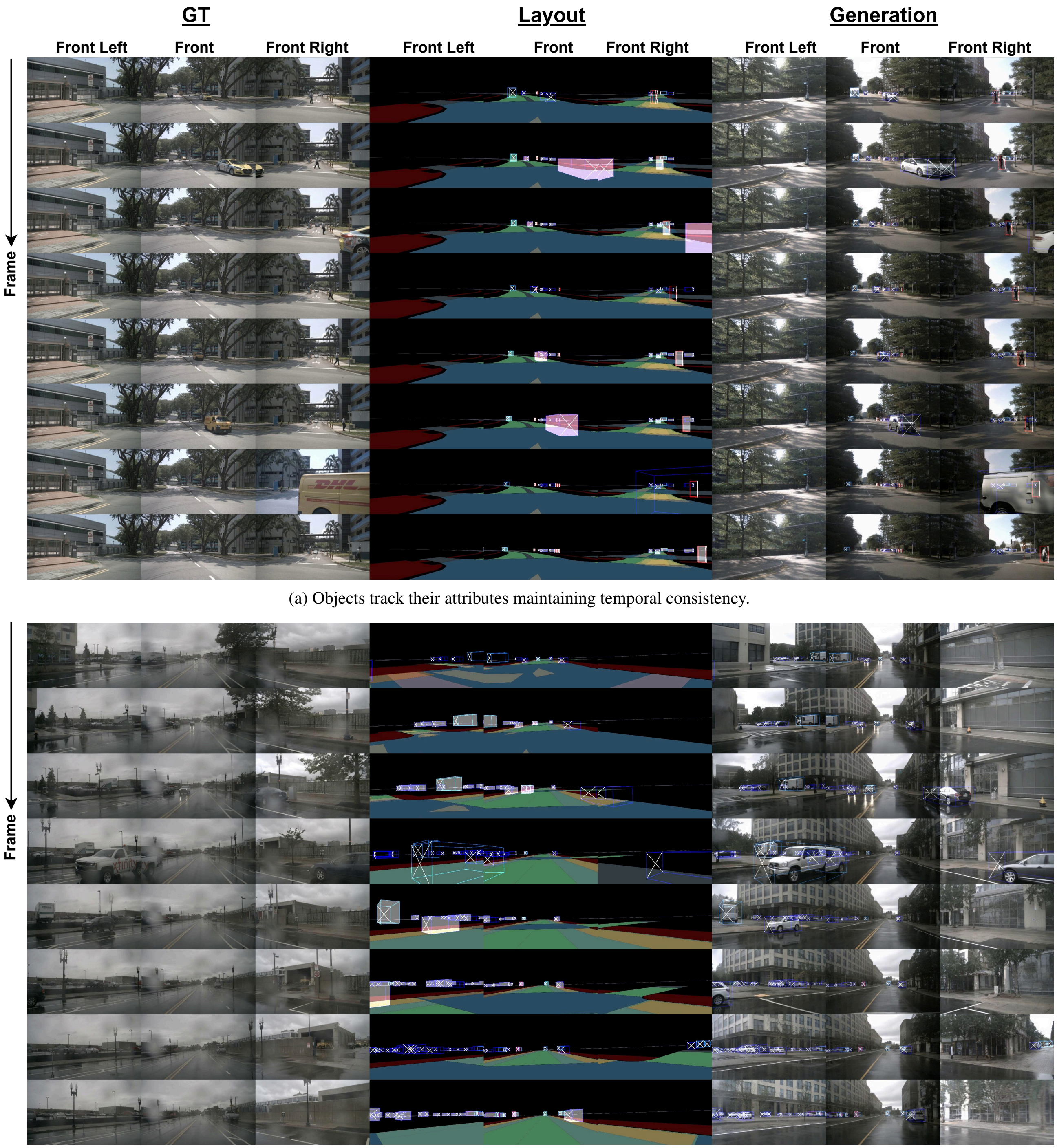
We conduct ablation on the SGA module by gradually removing its two components: depth injection and box projection. As shown in Tab. **??**, removing both causes a large drop of 3.20 in NDS and a big increase of 187 in IDS, confirming the importance of SGA module. Individually, removing box projection leads to a larger NDS drop, while removing depth injection results in more ID switches, indicating their respective contributions to spatial localization and temporal consistency.

2.3. Ablation Study in T2V Scenario

We also conduct an ablation study in the T2V scenario to evaluate the effectiveness of two key modules—the Instance Flow Guider (IFG) module and the Spatial Geometric Aligner (SGA) module. The results are presented in Tab.2.

Instance Flow Guider. To assess the impact of the IFG module, we eliminate the instance flow injection. In the T2V mode, where no guidance is provided from the first frame, the IFG module plays a more critical role compared to its role in the (T+I)2V mode. As shown in Tab.2, removing the instance flow injection leads to a significant degradation, with the FVD score increasing by 46.81. This highlights the module's importance in ensuring smooth and coherent video generation.

Spatial Geometric Aligner. To evaluate the influence of the SGA module, we remove the injection of depth order and replace the box projection with box coordinates as the control signal. This ablation significantly impacts the model's ability to accurately localize objects and understand spatial relationships. As demonstrated in Tab.2, the absence of the SGA module results in a notable performance drop, further emphasizing its essential role in achieving high spatial fidelity.



(a) Objects track their attributes maintaining temporal consistency.

(b) Small and densely packed objects rendered at correct locations.

Figure 1. Precise control mechanisms. We overlay the 3D bounding box projections onto the generated videos. The precision of control is reflected in: **(1)** Objects in the scene are accurately placed and sized to align with their *projected bounding boxes*, as shown in 1a and 1b. **(2)** Drivable areas, sidewalks, and zebra crossings are faithfully generated following the *road map projections*, as shown in 1a and 1b. **(3)** Objects track their previous attributes as guided by the *instance flow*, ensuring temporal consistency across frames. As shown in Figure 1a, the pink-rendered instance flow directs the model to generate the white sedan, maintaining its consistent attributes over time. **(4)** Small and densely packed objects are precisely rendered at their correct locations, following *3D bounding box coordinates*, as shown in 1b.

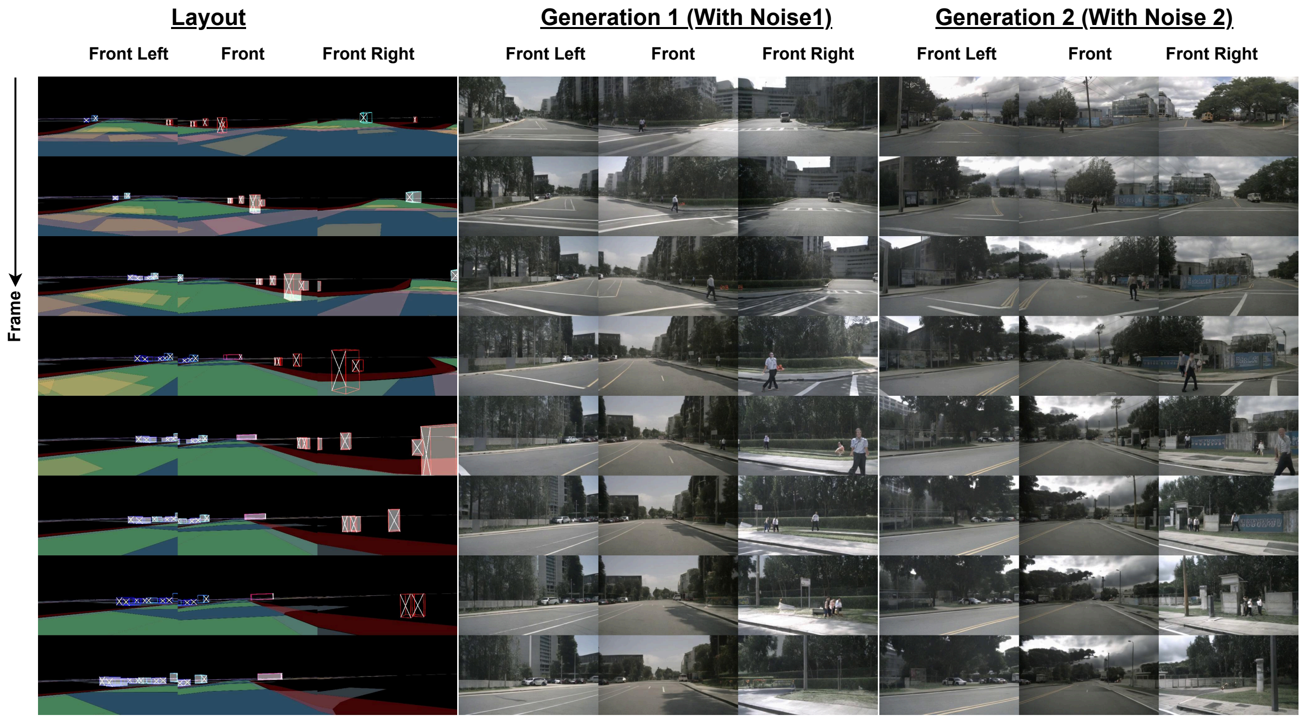


Figure 2. Diverse videos using **varying noise inputs** and **the same control conditions**. By introducing stochastic noise while maintaining consistent control signals—such as 3D bounding box coordinates, lane line projections, and instance flow—our model can produce a variety of videos that adhere to the defined constraints.

Settings	FVD↓	FID↓
InstaDrive	107.50	12.91
w/o Instance Flow Guider	154.31 (+46.81)	16.73 (+3.82)
w/o Spatial Geometric Aligner	117.25 (+9.75)	14.52 (+1.61)

Table 2. Ablation study results in T2V scenarios on the generated nuScenes validation set.

3. More Visualization Results

Here, we provide additional visualization results to showcase our model's strong ability to generate high-fidelity, realistic, and diverse multi-view driving videos. We sample 8 frames from each generated video as a demo to save space in the paper. Our model is capable of generating high-quality, long-duration driving videos through iterative processing. We provide a web page in the supplementary materials for additional results. Please refer to the webpage in the supplementary materials for visualization results.

3.1. Prompt Edit

InstaDrive enables video editing by modifying only the text prompt condition while keeping all other conditions fixed. In Fig.4 in main text, we demonstrate the model's editing capability by altering the weather and time of day in the text prompt. Specifically, we add ”Sunny,” ”Rainy,” and ”Night” to the original text prompt, while maintaining other conditions such as camera pose, 3D bounding box coordinates, 3D bounding box projections, road map projections, and instance flow unchanged. The generated videos showcase high quality and effective editing:

- **Sunny**: Displays clear skies with sunlight shining on the scene, reflecting bright and vivid environmental details.
- **Rainy**: Captures wet road surfaces and blurred camera views caused by raindrops, adding realistic weather dynamics.
- **Night**: Depicts dimly lit scenes with streetlights and reduced visibility, accurately simulating nighttime driving conditions.

These results emphasize the strong editing capability of ***InstaDrive*** , producing diverse and realistic driving videos with minimal changes to the input conditions.

3.2. Control Precision

InstaDrive excels at generating videos that adhere closely to various control conditions, including 3D bounding box projections, road map projections, depth order, and instance flow. In Fig. 1, we overlay the 3D bounding box projections onto the generated videos to illustrate the precision of our control mechanisms. The precision of control is reflected in:

- **Object alignment with bounding box projections**: Objects in the scene are accurately placed and sized to align with their projected bounding boxes, as shown in Fig. 1 (a)(b).
- **Road and pedestrian area fidelity**: Drivable areas, sidewalks, and zebra crossings are faithfully generated following the road map projections, as shown in Fig. 1 (a)(b).
- **Precision in dense and small objects**: Small and densely packed objects are precisely rendered at their correct locations, as shown in Fig. 1 (b).
- **Temporal consistency through instance flow**: Objects track their previous attributes as dictated by the *instance flow*, enabling consistent temporal consistency across frames, as shown in Fig. 1 (a).

These results highlight the superior control and fidelity of ***InstaDrive*** in generating realistic and controllable driving videos.

3.3. Carla-Generated Layout Control

InstaDrive demonstrates the ability to generate high-quality driving videos based on layout conditions provided by the Carla simulator [1], which include 3D bounding box projections, lane line projections, and scene description text prompts. The use of Carla-generated layouts addresses a critical limitation in real-world driving video datasets: the lack of diversity in scene types, especially for rare but critical events like lane cutting and sudden braking. By leveraging Carla's highly configurable simulation environment, we can create synthetic layouts that represent complex and diverse driving scenarios, such as multi-vehicle intersections, narrow streets, or sudden obstacles, which are difficult to capture in real-world data.

In Fig.5 in main text, we showcase our model's ability to generate rare videos corresponding to these layouts. The generated videos highlight ***InstaDrive***'s capacity to faithfully adhere to the control signals while producing realistic outputs. Moreover, by effectively handling corner cases, our approach bridges the gap in scene diversity, making it a valuable tool for training and validating driving models under challenging scenarios. The results demonstrate that ***InstaDrive*** can not only replicate realistic conditions but also adapt seamlessly to a wide range of complex layouts generated by Carla, further enhancing its applicability in autonomous driving research.

3.4. Diversity with Varying Noise

InstaDrive demonstrates the ability to generate diverse driving videos from identical control conditions with varying noise inputs, as illustrated in Fig. 2. By introducing stochastic noise while maintaining consistent control signals—such as 3D bounding box coordinates, lane line projections, and instance flow—our model produces a variety of plausible video outputs that adhere to the defined constraints.

3.5. Results on our private dataset

In addition to public datasets, we trained on a 200 hour private dataset. The results, as shown in the Fig. 3, demonstrate that we achieved similar generation quality and control performance as on nuScenes, highlighting the generalization capability of our method.

4. Limitations

Our work establishes a robust, physically informed framework for generating high-quality, multi-view driving videos, achieving state-of-the-art performance in both video generation quality and downstream perception task validation. However, certain limitations remain, mainly due to time and resource constraints.

Currently, our model's design has not been exhaustively optimized, leaving room for improvement in the quality of the generated videos. For example, the training process is conducted at a relatively low spatial resolution of 256×448, which constrains visual fidelity. Scaling to higher resolutions would require fine-tuning the position embeddings to ensure compatibility, an aspect not yet addressed in this work.

Future research could explore the integration of more advanced generative models, such as SD-XL [3], and develop more efficient methods to produce high-fidelity videos at larger spatial resolutions. Additionally, the computational cost of inference for InstaDrive is relatively high, which presents another avenue for improvement. Enhancing the efficiency of InstaDrive will be a key focus in future developments to make the model more practical for real-world applications.

5. Backbone

Spatial-Temporal DiT (ST-DiT). We use Spatial-Temporal DiT (ST-DiT) [2] as our backbone, which introduces a novel architecture that merges the strengths of diffusion models with transformer architectures [5]. This integration aims to address the limitations of traditional U-Net-based latent diffusion models (LDMs), improving their performance, versatility, and scalability. While keeping the overall framework consistent with existing LDMs, the key shift lies in replacing the U-Net with a transformer architecture for learning the denoising function $\epsilon_\theta(\cdot)$, thereby marking a pivotal advance in the realm of generative modelling.

The ST-DiT architecture incorporates two distinct block types: the Spatial DiT Block (S-DiT-B) and the Temporal DiT Block (T-DiT-B), arranged in an alternating sequence. The S-DiT-B comprises two attention layers, each performing Spatial Self-Attention (SSA) and Cross-Attention se-

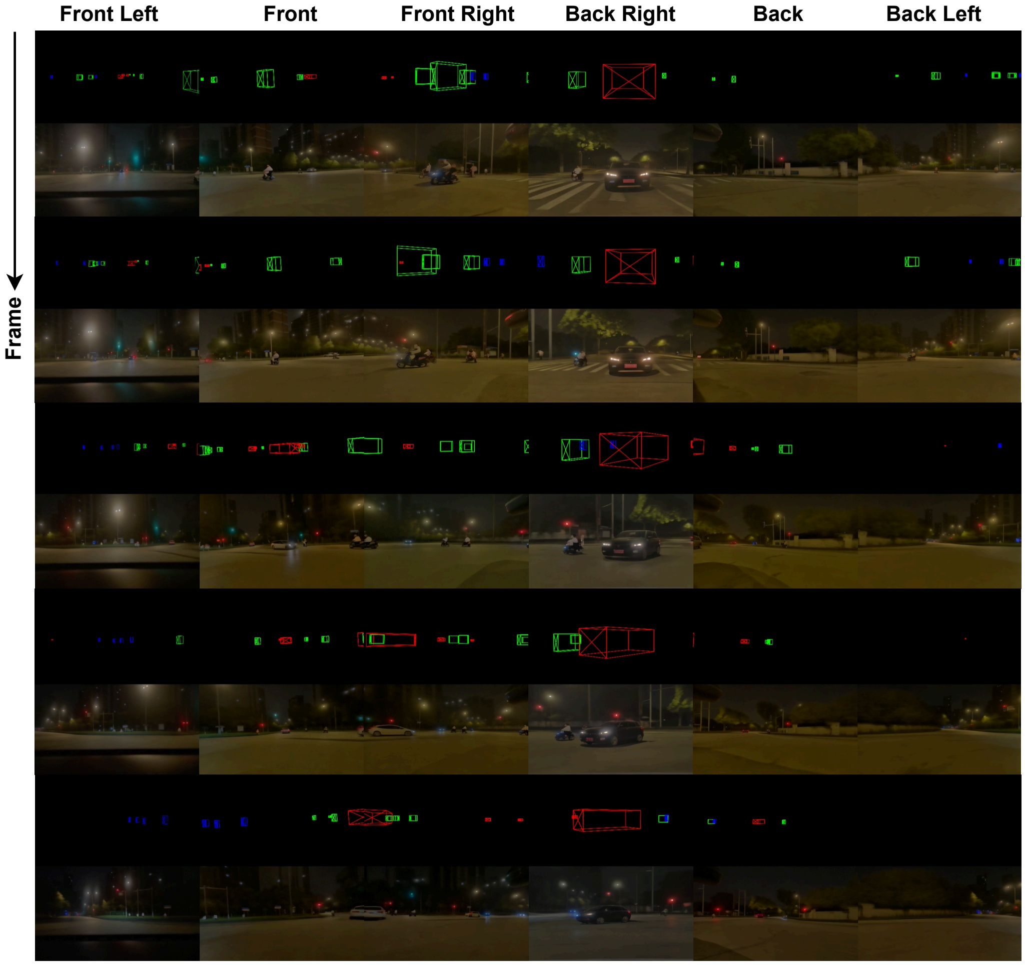


Figure 3. Results on private dataset.

quentially, succeeded by a point-wise feed-forward layer that serves to connect adjacent T-DiT-B block. Notably, the T-DiT-B modifies this schema solely by substituting SSA with Temporal Self-Attention (TSA), preserving architectural coherence. Within each block, the input, upon undergoing normalization, is concatenated back to the block's output via skip-connections. Leveraging the ability to process variable-length sequences, the denoising ST-DiT can handle videos of variable durations.

During processing, a video autoencoder [6] is first employed to diminish both spatial and temporal dimensions of videos. To elaborate, it encodes the input video $X \in \mathbb{R}^{T\times H\times W\times 3}$ into video latent $z_0 \in \mathbb{R}^{t\times h\times w\times 4}$, where L denotes the video length and $t = T, h = H/8, w = W/8$. z_0 is next "patchified", resulting in a sequence of input tokens $I \in \mathbb{R}^{t\times s\times d}$. Here, $s = hw/p^2$ and p denote the patch size. I is then forwarded to the ST-DiT, which models these compressed representations. In both SSA and TSA, standard Attention is performed using Query (Q), Key (K), and Value (V) matrices:

$$Q = W_Q \cdot I_{norm}; K = W_K \cdot I_{norm}; V = W_V \cdot I_{norm},$$

I_{norm} is the normalized I, W_Q, W_K, W_V are learnable matrices. The textual prompt is embedded with a T5 encoder and integrated using a cross-attention mechanism.

References

[1] Alexey Dosovitskiy, German Ros, Felipe Codevilla, Antonio Lopez, and Vladlen Koltun. CARLA: An open urban driving simulator. In *Proceedings of the 1st Annual Conference on Robot Learning*, pages 1–16, 2017. 4

[2] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 4

[3] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: improving latent diffusion models for high-resolution image synthesis. *CoRR*, abs/2307.01952, 2023. 4

[4] Xiaogang Shi, Bin Cui, Gillian Dobbie, and Beng Chin Ooi. Uniad: A unified ad hoc data processing system. *ACM Transactions on Database Systems (TODS)*, 42(1):1–42, 2016. 1

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 4

[6] Lijun Yu, Yong Cheng, Kihyuk Sohn, José Lezama, Han Zhang, Huiwen Chang, Alexander G Hauptmann, Ming-Hsuan Yang, Yuan Hao, Irfan Essa, et al. Magvit: Masked generative video transformer. In *CVPR*, 2023. 5

[7] 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, 2024. 1