SubjectDrive: Scaling Generative Data in Autonomous Driving via Subject Control

Binyuan Huang[1*†], Yuqing Wen[2*†], Yucheng Zhao[3*], Yaosi Hu[4*], Yingfei Liu[3], Fan Jia[3], Weixin Mao[3], Tiancai Wang[3‡], Chi Zhang[5], Chang Wen Chen[4], Zhenzhong Chen[1], Xiangyu Zhang[3]

[1]Wuhan University
[2]University of Science and Technology of China
[3]MEGVII Technology
[4]The Hong Kong Polytechnic University
[5]Mach Drive

Abstract

Autonomous driving progress relies on large-scale annotated datasets. In this work, we explore the potential of generative models to produce vast quantities of freely-labeled data for autonomous driving applications and present SubjectDrive, the first model proven to scale generative data production in a way that could continuously improve autonomous driving applications. We investigate the impact of scaling up the quantity of generative data on the performance of downstream perception models and find that enhancing data diversity plays a crucial role in effectively scaling generative data production. Therefore, we have developed a novel model equipped with a subject control mechanism, which allows the generative model to leverage diverse external data sources for producing varied and useful data. Extensive evaluations confirm SubjectDrive's efficacy in generating scalable autonomous driving training data, marking a significant step toward revolutionizing data production methods in this field.

Introduction

Deep generative models (Rombach et al. 2022; Blattmann et al. 2023b; Yu et al. 2023; Zhang, Rao, and Agrawala 2023; Song, Meng, and Ermon 2020; Ho, Jain, and Abbeel 2020) have achieved remarkable progress recently, demonstrating excellence in generating high-quality and realistic visual content. Diffusion models(Song, Meng, and Ermon 2020; Ho, Jain, and Abbeel 2020), a key contributor to this advancement, are renowned for their stable and top-quality sample generation. Through the utilization of modern diffusion models, the recent breakthroughs in controllable technology (Zhang, Rao, and Agrawala 2023) now facilitate precise and flexible content customization. These developments enable the creation of synthetic samples that are almost indistinguishable from real, human-annotated data. By leveraging the generative data, methods dependent on supervised training can be substantially enhanced for tasks such as image classification(Fan et al. 2024), object detection(Wang et al. 2024a), and object tracking(Li et al. 2023a). Motivated by these successes, there is a growing interest in applying generative models to the synthesis of natural data for more discriminative tasks (Sarıyıldız et al. 2023; Azizi et al. 2023; Nguyen et al. 2024).

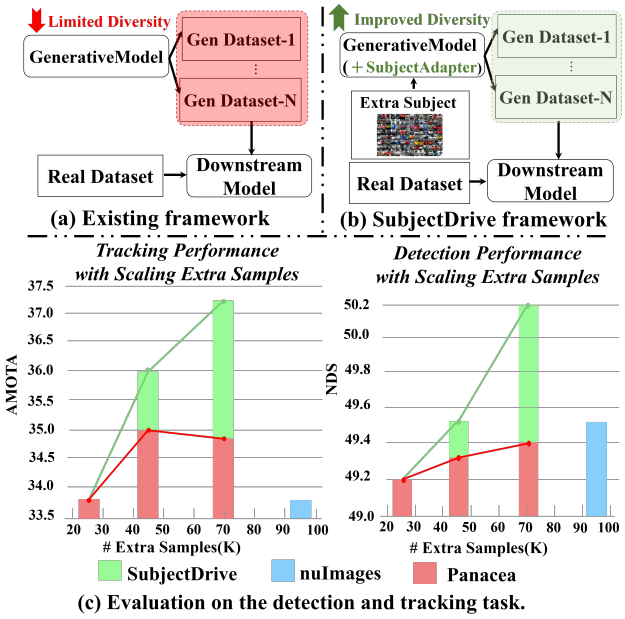

Figure 1: The comparison between existing method and SubjectDrive framework.(a) Existing data generation framework that uses the control sequence and sampling noise to generate synthetic data with limited sample diversity and scalability. (b) Compared with the traditional framework, SubjectDrive introduces additional synthesis diversity by incorporating extra subject control to enhance the scalability of generative model. (c) Evaluation of data scaling on the nuScenes detection and tracking task.

Autonomous driving, as a key application domain, heavily relies on large-scale annotated data and has increasingly paid attention to generative data collection methods. One of the most challenging tasks in autonomous driving is BEV (Bird's Eye View) perception, which includes 3D object de-

*Equal Contribution.
†This work was done during the internship at MEGVII.
‡Corresponding author.

tection(Wang et al. 2023a; Pang et al. 2023; Li et al. 2022; Liu et al. 2022, 2023a), 3D object tracking(Pang et al. 2023; Fischer et al. 2022; Marinello, Proesmans, and Van Gool 2022), and 3D lane detection. Acquiring annotated training data for these tasks is not only costly but also burdened by significant concerns regarding data privacy and usage rights, creating barriers that hinder data collection, labeling, release, and exchange processes, ultimately slowing progress in the field. To mitigate this issue, previous work has developed generative solutions such as BEVGen (Swerdlow, Xu, and Zhou 2023) and BEVControl (Yang et al. 2023), which could produce annotated training samples by adapting an image synthesis model. More recent efforts (Gao et al. 2023; Wen et al. 2023; Li, Zhang, and Ye 2023; Lu et al. 2023) have ventured into crafting driving scene videos, which successfully bolster more sophisticated temporal BEV perception models (Wang et al. 2023a).

Albeit promising, most existing work focuses on exploring the potential of generative data on a small scale, which fails to fully exploit the capacity of generative models to produce an inexhaustible supply of training samples. Notably, although early methods have demonstrated superior performance when incorporating generated samples alongside real data (Wen et al. 2023), there is still a gap compared to using existing large-scale real image data (Caesar et al. 2020) for pre-training, as illustrated in Fig. 1(a).

To further ignite the potential of synthetic data, this paper will first investigate the scalability of generative data produced by existing methods. As shown in Fig. 1(a), we observe that when the number of synthetic samples generated by the existing method increased from 46K to 69K, the performance in 3D object tracking notably decreased. We found that this lack of scalability is intrinsically related to **the limited diversity of the generated samples**. As shown in Fig. 2, different generative samples with the same annotation exhibit remarkably similar appearances in their foreground objects, i.e., the car, despite differences in the background. We also found that when the diversity issue is addressed by the method we introduce later, the scalability of generative data can be significantly improved.

In this paper, we tackle the challenge of scaling generative data for autonomous driving. We present SubjectDrive, an advanced video generation framework designed to enhance the scalability of generative models. Our initial findings reveal that conventional video generation pipelines struggle to scale effectively with increased data volumes due to its limited diversity. To overcome this limitation, we propose a novel generation framework centered on augmenting sampling diversity. Specifically, we integrate a feature termed subject control into existing generation pipelines. This feature empowers generative models to manipulate the diversity of the synthesis process by providing a mechanism to dictate the visual appearance of foreground elements in generated samples. This innovative feature enables the blending of the inherent stochastic nature of the generative process with the diversity drawn from external data sources, thereby crafting a more powerful model capable of producing scalable and diverse samples.

More concretely, SubjectDrive is designed around a trio

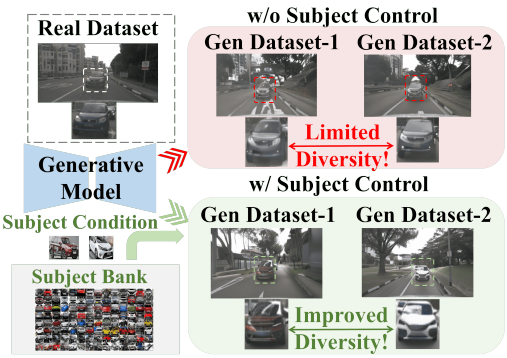

Figure 2: Above: Existing autonomous driving generative models struggle to produce diverse foreground samples. Below: By enhancing the sampling diversity capabilities with our subject control methods, the diversity of the generated foreground sample has significantly improved.

of innovative modules that collectively enable robust subject control capabilities. Initially, the model leverages a **Subject Prompt Adapter**, integrating subject control with the existing text-conditioned branch. Subsequently, to enhance the model's ability to inject spatial information, we introduce a **Subject Visual Adapter** that directly utilizes visual features, incorporating them into the existing diffusion U-Net architecture. Lastly, to ensure consistent injection of these features over time, we deploy **Augmented Temporal Attention** that expands the model's temporal-spatial context. Together, these modules empower SubjectDrive to perform subject-conditioned video generation with compelling efficacy. Overall, our research includes three main contributions:

- Unlike existing synthetic methods for autonomous driving, which are limited to small-scale validations, we propose a novel and critical research direction: investigating the scalability of generative data in autonomous driving. Our exploration provides a deeper understanding of the challenges and solutions associated with leveraging large-scale synthetic data in this domain.

- To address the challenges faced by existing methods during data scaling, particularly the issue of limited diversity, we introduce a novel generative framework called *SubjectDrive*. This framework enhances sampling diversity and improves the scalability of generative data through an innovative mechanism known as subject control.

- Extensive experiments on the nuScenes dataset (Caesar et al. 2020) validate the effectiveness of our proposed method. Remarkably, our approach is the first generative technique to surpass the performance of perception models pre-trained on a large-scale real dataset, specifically nuImages(Caesar et al. 2020). These outstanding results underscore the transformative potential of generative data in advancing autonomous driving technologies, offering a promising direction for future research in this field.

Related Work

Scalable Data Synthesis for Autonomous Driving

The use of synthetic data has been widely explored for visual perception tasks that require great of efforts in labeled data collection. Examples include applications in image classification (Azizi et al. 2023; Sarıyıldız et al. 2023), semantic segmentation (Nguyen et al. 2024), and object tracking (Li et al. 2023b). Recently, there has been a burgeoning interest in employing generative data for the challenging BEV perception tasks (Wang et al. 2023a; Li et al. 2022), which demand precise geometric and appearance alignment from generative models. Initial works aim at synthesizing street-view images (Swerdlow, Xu, and Zhou 2023; Yang et al. 2023), with subsequent studies extending into the generation of driving scene videos (Wen et al. 2023; Lu et al. 2023; Wang et al. 2023b). Despite these successful efforts, the majority of existing research has been limited to producing a small quantity of training samples, not fully exploiting the generative models' capacity to offer an unlimited reservoir of samples. Data scaling efforts (Azizi et al. 2023) are sparse and, thus far, not particularly successful. It's imperative to recognize that recent breakthroughs highlight the paramount importance of data scaling in training sophisticated deep learning models, such as GPT (Achiam et al. 2023), SVD (Blattmann et al. 2023a), and Sora (Brooks et al. 2024). Therefore, addressing the challenge of amplifying generative data volumes is of essence. This work delves into scalable data synthesis for autonomous driving, providing analysis results and proposing an enhanced framework.

Diffusion-based Generative Models with Subject Controls

Diffusion models with subject control are designed to embed target subjects into generated visual content, guided by reference images (Choi et al. 2023). These approaches (Han et al. 2023; Kumari et al. 2023; Li et al. 2023c; Chen et al. 2023b; Ma et al. 2023; Xiao et al. 2024) initially emerged in image generation research to facilitate identity-preserving applications. Early methods (Han et al. 2023; Kumari et al. 2023; Li et al. 2023c; Gal et al. 2022; Ruiz et al. 2023) adopted a fine-tuning framework to adjust a specific model capable of generating images of the desired subject. A notable example is DreamBooth (Ruiz et al. 2023), which fine-tunes a diffusion model using a small set of subject images. While these fine-tuning methods achieved high-quality results, they were limited by the significant resources required for model tuning, making them impractical for scenarios with numerous target entities. To address these limitations, tuning-free methods (Ma et al. 2023; Ye et al. 2023) were developed. Subject Diffusion (Ma et al. 2023), for instance, achieved impressive customization by introducing a subject conditioning module and training on a bespoke dataset of subject pairs. Subsequent works, such as Cones2 (Liu et al. 2023b) and others (Chen et al. 2023b), have expanded tuning-free subject control to scenarios involving multiple subjects.

In the video domain, there is also growing interest in subject-controllable generation (Jiang et al. 2023; Chen et al. 2023a; Wang et al. 2024b). For instance, VideoBooth (Jiang et al. 2023) proposed a framework capable of generating consistent videos containing the subjects specified in image prompts. Additionally, CustomVideo (Wang et al. 2024b) introduced a multi-subject-driven text-to-video model powered by a simple yet effective co-occurrence and attention control mechanism. Our work falls into the category of subject-driven video generation. Unlike previous efforts, we focus on the importance of subject-controlled generation for data scaling, particularly in autonomous driving.

Method

Preliminary: Latent Diffusion Models

Diffusion Models (DMs) (Song, Meng, and Ermon 2020; Ho, Jain, and Abbeel 2020) iteratively denoise a random noise with Gaussion distribution $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ over T steps to generate target data according to Eq. 1, where the functions μ_θ and Σ_θ are derived from the denoising model ϵ_θ. The training of DMs includes both a diffusion process and a denoising process. The diffusion process begins by adding Gaussion noise to the original data sample x_0 over t steps, controlled by the scheduled noise strength β_t, which can be simplified as Eq. 2 where $\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_t)$. To learn the denoising process, the denoising model ϵ_θ, which aims to estimate the original noise ϵ from the noisy data x_t, is optimized by minimizing the loss function as shown in Eq. 3.

$$p_\theta(x_{t-1}|x_t) = \mathcal{N}(\mathbf{x}_{t-1}; \mu_\theta(\mathbf{x}_t, t), \mathbf{\Sigma}_\theta(\mathbf{x}_t, t)) \quad (1)$$

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \mathbf{x}_0 \sim p(x) \quad (2)$$

$$\min_\theta \mathbb{E}_{t,x,\epsilon}\|\epsilon - \epsilon_\theta(x_t, t)\|^2 \quad (3)$$

Given the significant computational burden of diffusion models in generating high-resolution images or videos, Latent Diffusion Models (LDMs) (Rombach et al. 2022) have become popular in recent studies. With a pre-trained autoencoder (Kingma and Welling 2013) to handle the compression and reconstruction between high-dimensional visual data and low-redundancy latent representations, the diffusion model can concentrate exclusively on generation in latent space, significantly reducing computational costs. Consequently, our method employs LDMs for video generation.

SubjectDrive

SubjectDrive is designed to enhance the scalability of generative data, thereby promoting perception models for autonomous driving applications. Despite the ability of advanced generative methods (Wen et al. 2023) to produce high-quality driving-scene videos, they narrowly uplifts the performance on downstream perception tasks as illustrated in Fig. 1(c). We believe this is primarily due to the limited diversity of generated foreground elements, which are crucial for autonomous driving. Thus, different from tradition generation pipeline using control sequence to guide global scene generation, we innovatively integrate subject control mechanism into generation process, allowing for the injection of external subjects from extensive open-source data. The integration of controlled subjects not only boosts controllability but also effectively enhances the diversity of generated

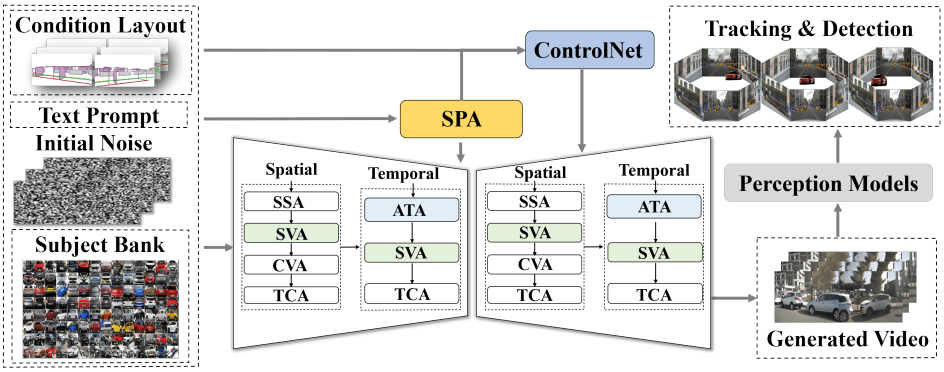

Figure 3: Overview of SubjectDrive. The pipeline of SubjectDrive involves a frozen auto-encoder and a trainable UNet-based diffusion model. Different control signal sources including extended text prompt, condition layout, and subject bank. SSA represents Spatial Self-Attention, SVA represents Subject Visual Adapter, TCA represents Text Cross-Attention, ATA represents Augmented Temporal Attention, and CVA represents Cross-View Attention.

foreground elements, which shows strong augmentation for autonomous driving applications.

To inject subjects into generation process, we propose the novel Subject Prompt Adapter (SPA) and Subject Visual Adapter (SVA) to augment expressity of text embedding with regard to subjects and integrate subjects' spatial information into frames, respectively. To further improve the appearance consistency of injected subjects across frames, the Augmented Temporal Attention (ATA) is introduced to effectively capture long-range movements in driving videos.

Overview Our framework is built on a text-to-video diffusion model, i.e. Panacea (Wen et al. 2023), which is a strong baseline for multi-view video generation. The overall training framework of SubjectDrive is illustrated in Fig. 3. The diffused noisy input is fed into the trainable diffusion model to generate latent video under the guidance of text, condition layout, and injected subjects. We inherit the guidance of condition layout with ControlNet (Zhang, Rao, and Agrawala 2023) from Panacea. For subjects guidance, we first extend text prompt and augment the text embedding of subject part with Subject Prompt Adapter. On the other hand, we insert the gated self-attention layer into diffusion model to enhance the location guidance of subjects captured by Subject Visual Adapter. Furthermore, we replace the conventional temporal 1D attention with proposed Augmented Temporal Attention to improve the temporal consistency of injected subjects.

Subject Prompt Adapter Subject Prompt Adapter introduces an enhancement to textual cues through the integration of subject attributes including the category semantic, ID identifier, and visual semantic information.

To achieve this, we first extend the scene prompt with injected subject category description. For example, the prompt for the N_{th} frame that contains M subjects is organised as "*[Scene Prompt], including [Subject* X_1*], [Subject* X_2*].... [Subject* X_M*].*" *[Subject* X_i*]* here refers to the category word of subject X_i, e.g., car, bus. The extended prompt is input into the pre-trained CLIP (Radford et al. 2021) text encoder to obtain the original text embedding $z^t \in \mathbb{R}^{L\times d}$, where L represents the length of the text embedding and

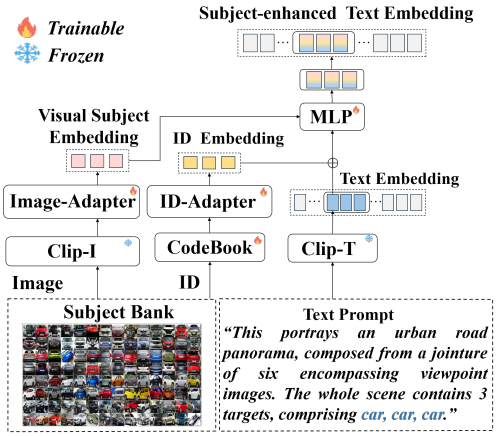

Figure 4: The Subject Prompt Adaptor which augments the original text embedding of extended prompt with corresponding ID identifier and visual semantic information to enhance the expressity of subjects.

d is the dimension of the text embedding. To further integrate subject attributes for better expressity, we extract the ID identifier of each subject from condition layout to ensure coherent trajectories across frames, and encode it with a learnable codebook and an ID-adapter comprised of MLP to get the ID embedding $z_i^{id} \in \mathbb{R}^d$ for subject X_i. Similarly, we encode the image of X_i by the CLIP image encoder and an image-adapter to obtain the corresponding visual semantic embeddings $z_i^v \in \mathbb{R}^d$.

For the original text embedding z_i^t at the index of subject X_i in z^t, we successively enhance it with the corresponding ID embeddings z_i^{id} and visual semantic embeddings z_i^v by

$$\hat{z}_i^t = \text{MLP}\left([z_i^t + z_i^{id}, z_i^v]\right), \quad i \in \{\text{Index}_{1,2,\cdots,M}\} \tag{4}$$

where $[,]$ represents the concatenation operation. The resulted subject-enhanced text embedding $\hat{z}^t$ contains not only the semantic of whole scene, but also specific identifying information for each subject. The subject-enhanced text embedding

is further injected into the UNet through the cross-attention layer to guide the generation of frames.

Subject Visual Adapter The Subject Visual Adapter is proposed to inject subject spatial information into video feature to further enhance content alignment of generated video with provided visual clue and location of subjects.

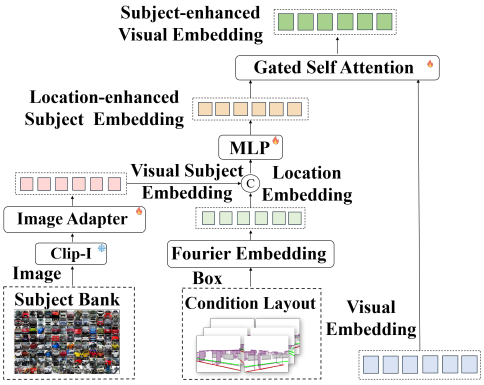

Figure 5: The Subject Visual Adapter which injects location-enhanced subject information into visual features cooperated with gated self-attention.

Inspired by (Li et al. 2023d), SVA first combines subject images and corresponding locations to form location-enhanced subject embeddings, and then injects them into frames with a control signal to guide the generation of subjects at specified location. To obtain the location-enhanced subject embedding f^{vl}, the location of each subject, represented by its coordinates $l = [x_{min}, y_{min}, x_{max}, y_{max}]$ in the current frame, is encoded with fourier embedding, further integrated into visual embedding f^v of corresponding subject as

$$f^{vl} = \text{MLP}[f^v, \text{Fourier}(l)]). \quad (5)$$

To inject the location-enhanced subject embedding into generated frame, we employ the gated self-attention layer (Li et al. 2023d) that locates between each paired self-attention layer and cross-attention layer in the UNet. In the gated self-attention layer, the location-enhanced subject embedding f^{vl} interacts with frame visual tokens z by an attention operation to capture the dependency between subject embedding and visual tokens, followed by a token selection operation TS to preserve only visual tokens. To adaptively adjust the guidance scale of location information over frames, a gating factor is learned which is operated as

$$z = z + \tanh(\gamma) \cdot \text{TS}(\text{SelfAttn}([z, f^{vl}])), \quad (6)$$

where γ represents the gating factor, a learnable parameter initialized to 0 for stable training.

Augmented Temporal Attention The Augmented Temporal Attention is designed to capture long-range movement of subjects with feasible computation cost. The conventional temporal attention layer utilized in video diffusion models solely incorporates self-attention operation within the temporal dimension. However, due to the substantial movements typically involved with subjects in autonomous driving

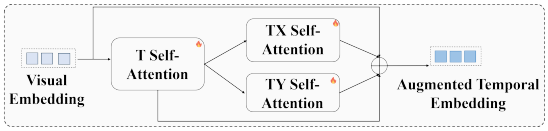

Figure 6: Augmented Temporal Attention combines 1D temporal attention with decomposed attention on the temporal-horizontal (TX) and temporal-vertical (TY) planes to capture long-range subject movements.

videos, it becomes challenging for the temporal-dimension attention to effectively capture the long-range dependency of inter-frame subjects. Therefore, we propose the Augmented Temporal Attention which incorporates interaction within temporal-horizontal (TX) plane and temporal-vertical (TY) plane to improve subject consistency across sequences.

Specifically, given input video feature $z^i \in R^{T\times H\times W\times C}$ where T, H, W, and C denote the video length, spatial height, width, and number of channels, respectively, it is processed by 1D attention along temporal dimension (T) to obtain z^T. z^T is reshaped to respective $z^{TX} \in R^{H\times(T\times W)\times C}$ and $z^{TY} \in R^{W\times(T\times H)\times C}$ to aggregate information from two decomposed planes by the parallel TX self-attention and TY self-attention. In this way, the fused video feature not only integrates small-range variation from temporal 1D attention, but also captures large-range movements by decomposed TX and TY attention, which can effectively enhance the temporal coherence of the generated video. The overall operation of this layer can be written as

$$z^o = \text{SelfAttn}(z^{TX}) + \text{SelfAttn}(z^{TY}) + z^T + z^i. \quad (7)$$

Experiment

Evaluation Datasets and Metrics

Datasets. We use the nuScenes dataset to train SubjectDrive and assess the visual fidelity and controllability of the generated data. nuScenes contains 1,000 scenes, each 20 seconds long, with 2Hz annotations and a 360° camera view. It includes 1.4 million 3D bounding boxes across 10 categories: car, truck, bus, trailer, construction vehicle, pedestrian, motorcycle, bicycle, barrier, and traffic cone.

Evaluation Metrics. Following Panacea, we assess two key objectives using perceptual models. First, we evaluate the generative model's controllability by comparing the alignment of generated data with BEV annotations using a pre-trained StreamPETR model. Second, we measure how the generated data enhances perceptual model performance in detection and tracking. Detection metrics include NDS, mAP, mAOE, and mAVE, while tracking metrics cover AMOTA, AMOTP, Recall, and MOTA. Visual fidelity is validated with FID(Heusel et al. 2017) and FVD)(Unterthiner et al. 2018).

Implementation Details

SubjectDrive adopts a two-stage video generation approach: image generation in the first stage (optimized for 56k steps) and video generation in the second (84k steps). Experiments

are conducted on 8A100 GPUs using the DDIM sampler with 25 steps to produce 256×512 resolution video clips spanning 8 frames. The evaluation uses StreamPETR with a ResNet50 backbone (He et al. 2016), trained at 256×512 resolution. To construct the subject bank, we draw from two distinct sources, categorizing them into internal and external subject banks. The internal subject bank is curated by collecting subjects from the training set of the nuScenes dataset. The external subject bank is established by integrating external vehicle datasets from the open-source CompCars (Yang et al. 2015) dataset. The internal subject bank is used during the training phase, while the external subject bank is mixed with the internal one during the sampling phase to promote the generation of diverse data.

Main Results

Analysis of Synthetic Data Scale-up. We investigate the impact of scaling up the quantity of generative data on the performance of downstream tasks and present the results in Table 1. Two interesting findings emerge. First, **increasing the quantity of synthetic data has a positive influence on the performance of the perception model.** Although using a small amount of synthetic data is initially less effective than using real nuImages data, increasing the data volume can boost the models to achieve superior performance in both NDS and AMOTA. This demonstrates that scaling up generative data is essential to unlock the potential of generative data production. Second, **the integration of external subjects can substantially improve scaling performance.** For instance, while tripling the generative data from Panacea yields a mere 0.2 improvement in NDS, incorporating SubjectDrive at the same data scale elevates the NDS by 1.0. Similar trends are also observed in AMOTA. These results demonstrate that utilizing external subjects is an effective way to enhance the generative model's scaling capability.

#Extras	Source	Tracking		Detection	
		AMOTA ↑	AMOTP ↓	NDS ↑	MAP ↑
0	nuScenes	30.1	1.379	46.9	34.5
93K	nuImages	33.8 (+3.7%)	1.324	49.5 (+2.6%)	37.8
23K	Panacea	33.7 (+3.6%)	1.353	49.2 (+2.3%)	37.1
46K		35.0 (+4.9%)	1.346	49.3 (+2.4%)	37.3
69K		34.9 (+4.8%)	1.348	49.4 (+2.5%)	37.1
23K	Ours	33.7 (+3.6%)	1.353	49.2 (+2.3%)	37.1
46K		36.0 (+5.9%)	1.322	49.5 (+2.6%)	37.7
69K		**37.2(+7.1%)**	**1.317**	**50.2(+3.3%)**	**38.1**

Table 1: Evaluation of data scaling on detection and tracking tasks. Extras denote extra samples beyond nuScenes dataset.

Analysis of Performance in 3D Object Detection Task. Next, we present a quantitative analysis that compares SubjectDrive with other data generation methods in the 3D detection task. We utilize SubjectDrive to generate novel synthetic data as additional training source for the StreamPETR model, and then evaluate its performance on the real nuScenes validation set. The results are showcased in Table 2. When trained solely on the synthetic data, SubjectDrive manages to outperform Panacea by 5.5 mAP and 5.0 NDS. Specifically, SubjectDrive achieves an NDS of 41.1, reaching 88% of the performance compared to that trained with real nuScenes data. When combining synthetic data with real data, SubjectDrive surpasses all other methods by a significant margin.

DataType	Method	MAP↑	NDS↑	MAVE↓
Real	Real Only	34.5	46.9	29.1
Gen	Panacea	22.5	36.1	46.9
	Ours	**28.0 (+5.5%)**	**41.1 (+5.0%)**	**37.0**
Real+Gen	DriveDreamer	35.8	39.5	-
	WoVoGen∗	36.2	18.1	123.4
	MagicDrive	35.4	39.8	-
	Panacea	37.1	49.2	27.3
	Ours	**38.1**	**50.2**	**26.4**

Table 2: Comparison on the 3D object detection task with other generation methods. ∗ indicates the evaluation of WoVoGen is only on the vehicle classes of cars, trucks, and buses.

Analysis of Performance in 3D Object Tracking Task. In addition to the 3D object detection task, the 3D object tracking task is an important and more challenging fundamental task in autonomous driving. As shown in Table 3, using solely our synthetic data can produce a model with a 23.5 MOTA, attaining 86% of the performance compared to that trained with real nuScenes data. When combining synthetic data with real data, SubjectDrive achieves a 37.2 AMOTA, which is 3.5 points higher than the model trained with Panacea. It is worth noting that the performance gain in tracking is notably higher than that in detection. This is because our subject control mechanism not only improves data diversity but also has a beneficial side effect of enhancing temporal coherency by ensuring all generated objects across frames align with the given reference subjects.

DataType	Method	AMOTA ↑	AMOTP ↓	MOTA ↑
Real	Real Only	30.1	1.379	27.1
Gen	Ours	23.4	1.544	23.5
Real+Gen	Panacea	33.7	1.353	30.6
	Ours	**37.2 (+3.5%)**	**1.317**	**33.3 (+2.7%)**

Table 3: Comparison on the 3D object tracking task.

Comparison of Generation Quality. In order to validate the visual quality of our generated samples, we compared them with state-of-the-art methods for driving scene generation. We generated the validation set of nuScenes without applying any pre-processing or post-processing to the selected samples. As shown in Table 4, our method achieves the best performance, with an FVD of 124 and an FID of 15.98, compared to both video-level generation methods—WoVoGen, Panacea, and DriveDreamer—and image-level generation methods, BEVGen , BEVControl and MagicDrive.

Ablation Studies

Aligning generated samples with BEV conditional labels is crucial for evaluating generative model controllability and assessing synthetic data applicability. Here, we conduct ablation studies to validate SubjectDrive's design choices using

Method	Multi-View	Multi-Frame	FVD↓	FID↓
BEVGen	✓		-	25.54
BEVControl	✓		-	24.85
MagicDrive	✓		-	16.20
DriveDreamer		✓	353	26.8
WoVoGen	✓	✓	418	27.6
Panacea	✓	✓	139	16.96
Ours	✓	✓	**124**	**15.98**

Table 4: Comparison of FID and FVD metrics with state-of-the-art methods on the validation set of the nuScenes dataset.

this metric. Table 5 presents evaluation results for the Subject Visual Adapter (SVA), Subject Prompt Adapter (SPA), and Augmented Temporal Attention (ATA). First, adding SVA to the baseline improves NDS by 2.2 and AMOTA by 4.2, demonstrating its effectiveness in enhancing label alignment through visual features and positional embeddings. Second, integrating SPA with SVA further boosts NDS and AMOTA by 1.8, highlighting their complementary roles in SubjectDrive. Finally, incorporating ATA improves AMOTA by 0.7 and RECALL by 1.0, reinforcing its contribution to alignment accuracy and temporal consistency.

SVA	SPA	ATA	NDS↑	AMOTA↑	RECALL ↑
–	–	–	32.1	11.4	21.3
✓	–	–	34.3	15.6	24.5
✓	✓	–	36.1	17.4	27.1
✓	✓	✓	36.3 (+**4.2%**)	18.1 (+**6.7%**)	28.1 (+**6.8%**)
Real Validation			46.9	30.1	41.8

Table 5: Ablation studies of different modules in SubjectDrive, with the last row showing the alignment performance on the real validation data.

Visualisation Results

Consistent Multi-View Video Generation. As illustrated in Fig. 7, for the six-view, eight-frame generated video, SubjectDrive produces temporally and view-consistent videos on the nuScenes validation set.

Subject-controlled Video Generation. Fig. 8 shows the visualization of subject-controlled driving video generation. Given the image of a reference subject, SubjectDrive can generate layout-aligned driving videos featuring the desired subject. By using reference subjects as control signals, SubjectDrive offers a mechanism for incorporating external diversity into the generated data.

Controllable Video Generation. Fig. 9 shows the driving scene video generated by our method, which adheres closely to the specified BEV layout, demonstrating strong layout control and alignment.

Conclusion

In this work, we present SubjectDrive, a novel video generation framework that enhances the scalability and sampling diversity of generative models. The architecture incorporates

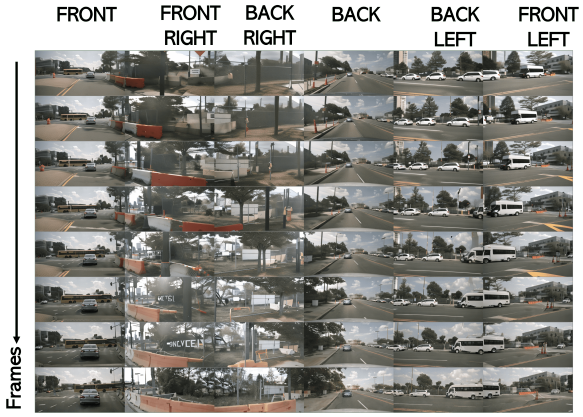

Figure 7: Multi-view videos generated by Ours.

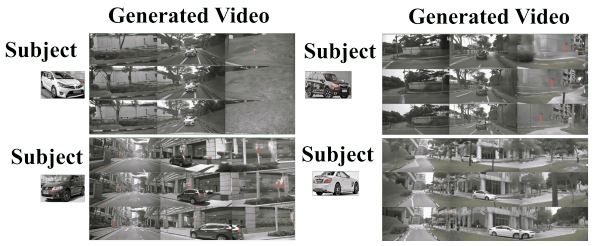

Figure 8: Subject-controlled videos generated by Ours.

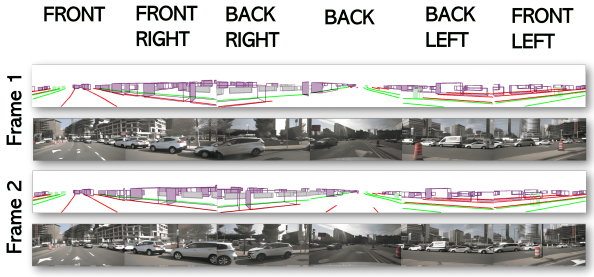

Figure 9: Controllable multi-view videos generated by Ours.

three key innovations: a subject prompt adapter, a subject visual adapter, and augmented temporal attention, collectively enabling robust subject control. This feature significantly improves the model's ability to generate diverse samples. Extensive experiments demonstrate that SubjectDrive not only outperforms existing methods but also scales effectively. Notably, it is the first generative model to enhance perception performance beyond pre-trained capabilities on the nuImages dataset, showcasing the transformative potential of generative data in advancing autonomous driving technologies and pointing to promising future directions in the field.

Acknowledgments

The work was supported by National Science and Technology Major Project of China (2023ZD0121300).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Azizi, S.; Kornblith, S.; Saharia, C.; Norouzi, M.; and Fleet, D. J. 2023. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*.

Blattmann, A.; Dockhorn, T.; Kulal, S.; Mendelevitch, D.; Kilian, M.; Lorenz, D.; Levi, Y.; English, Z.; Voleti, V.; Letts, A.; et al. 2023a. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*.

Blattmann, A.; Rombach, R.; Ling, H.; Dockhorn, T.; Kim, S. W.; Fidler, S.; and Kreis, K. 2023b. Align your latents: High-resolution video synthesis with latent diffusion models. In *CVPR*.

Brooks, T.; Peebles, B.; Holmes, C.; DePue, W.; Guo, Y.; Jing, L.; Schnurr, D.; Taylor, J.; Luhman, T.; Luhman, E.; Ng, C.; Wang, R.; and Ramesh, A. 2024. Video generation models as world simulators.

Caesar, H.; Bankiti, V.; Lang, A. H.; Vora, S.; Liong, V. E.; Xu, Q.; Krishnan, A.; Pan, Y.; Baldan, G.; and Beijbom, O. 2020. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*.

Chen, H.; Wang, X.; Zeng, G.; Zhang, Y.; Zhou, Y.; Han, F.; and Zhu, W. 2023a. Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. *arXiv preprint arXiv:2311.00990*.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2023b. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*.

Choi, J.; Choi, Y.; Kim, Y.; Kim, J.; and Yoon, S. 2023. Custom-edit: Text-guided image editing with customized diffusion models. *arXiv preprint arXiv:2305.15779*.

Fan, L.; Chen, K.; Krishnan, D.; Katabi, D.; Isola, P.; and Tian, Y. 2024. Scaling laws of synthetic images for model training... for now. In *CVPR*, 7382–7392.

Fischer, T.; Yang, Y.-H.; Kumar, S.; Sun, M.; and Yu, F. 2022. CC-3DT: Panoramic 3D object tracking via cross-camera fusion. *arXiv preprint arXiv:2212.01247*.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-Or, D. 2022. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*.

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

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D. N.; and Yang, F. 2023. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. In *ICCV*.

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *CVPR*.

Heusel, M.; Ramsauer, H.; Unterthiner, T.; Nessler, B.; and Hochreiter, S. 2017. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *NeurIPS*, 30.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *NeurIPS*, 33.

Jiang, Y.; Wu, T.; Yang, S.; Si, C.; Lin, D.; Qiao, Y.; Loy, C. C.; and Liu, Z. 2023. VideoBooth: Diffusion-based Video Generation with Image Prompts. *arXiv preprint arXiv:2312.00777*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J. 2023. Multi-Concept Customization of Text-to-Image Diffusion. In *CVPR*.

Li, P.; Liu, Z.; Chen, K.; Hong, L.; Zhuge, Y.; Yeung, D.-Y.; Lu, H.; and Jia, X. 2023a. Trackdiffusion: Multi-object tracking data generation via diffusion models. *arXiv preprint arXiv:2312.00651*.

Li, P.; Liu, Z.; Chen, K.; Hong, L.; Zhuge, Y.; Yeung, D.-Y.; Lu, H.; and Jia, X. 2023b. Trackdiffusion: Multi-object tracking data generation via diffusion models. *arXiv preprint arXiv:2312.00651*.

Li, X.; Zhang, Y.; and Ye, X. 2023. DrivingDiffusion: Layout-Guided multi-view driving scene video generation with latent diffusion model. *arXiv preprint arXiv:2310.07771*.

Li, Y.; Liu, H.; Wen, Y.; and Lee, Y. J. 2023c. Generate anything anywhere in any scene. *arXiv preprint arXiv:2306.17154*.

Li, Y.; Liu, H.; Wu, Q.; Mu, F.; Yang, J.; Gao, J.; Li, C.; and Lee, Y. J. 2023d. Gligen: Open-set grounded text-to-image generation. In *CVPR*.

Li, Z.; Wang, W.; Li, H.; Xie, E.; Sima, C.; Lu, T.; Qiao, Y.; and Dai, J. 2022. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*.

Liu, Y.; Wang, T.; Zhang, X.; and Sun, J. 2022. Petr: Position embedding transformation for multi-view 3d object detection. In *ECCV*.

Liu, Y.; Yan, J.; Jia, F.; Li, S.; Gao, A.; Wang, T.; and Zhang, X. 2023a. Petrv2: A unified framework for 3d perception from multi-camera images. In *ICCV*.

Liu, Z.; Zhang, Y.; Shen, Y.; Zheng, K.; Zhu, K.; Feng, R.; Liu, Y.; Zhao, D.; Zhou, J.; and Cao, Y. 2023b. Cones 2: Customizable image synthesis with multiple subjects. *arXiv preprint arXiv:2305.19327*.

Lu, J.; Huang, Z.; Zhang, J.; Yang, Z.; and Zhang, L. 2023. WoVoGen: World volume-aware diffusion for controllable multi-camera driving scene generation. *arXiv preprint arXiv:2312.02934*.

Ma, J.; Liang, J.; Chen, C.; and Lu, H. 2023. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. *arXiv preprint arXiv:2307.11410*.

Marinello, N.; Proesmans, M.; and Van Gool, L. 2022. TripletTrack: 3D object tracking using triplet embeddings and LSTM. In *CVPR*.

Nguyen, Q.; Vu, T.; Tran, A.; and Nguyen, K. 2024. Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. *NeurIPS*.

Pang, Z.; Li, J.; Tokmakov, P.; Chen, D.; Zagoruyko, S.; and Wang, Y.-X. 2023. Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking. In *CVPR*.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *CVPR*.

Sarıyıldız, M. B.; Alahari, K.; Larlus, D.; and Kalantidis, Y. 2023. Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In *CVPR*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *ICLR*.

Swerdlow, A.; Xu, R.; and Zhou, B. 2023. Street-View Image Generation from a Bird's-Eye View Layout. *arXiv preprint arXiv:2301.04634*.

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

Wang, S.; Liu, Y.; Wang, T.; Li, Y.; and Zhang, X. 2023a. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*.

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

Wang, Y.; Gao, R.; Chen, K.; Zhou, K.; Cai, Y.; Hong, L.; Li, Z.; Jiang, L.; Yeung, D.-Y.; Xu, Q.; et al. 2024a. Detdiffusion: Synergizing generative and perceptive models for enhanced data generation and perception. In *CVPR*.

Wang, Z.; Li, A.; Xie, E.; Zhu, L.; Guo, Y.; Dou, Q.; and Li, Z. 2024b. CustomVideo: Customizing Text-to-Video Generation with Multiple Subjects. *arXiv preprint arXiv:2401.09962*.

Wen, Y.; Zhao, Y.; Liu, Y.; Jia, F.; Wang, Y.; Luo, C.; Zhang, C.; Wang, T.; Sun, X.; and Zhang, X. 2023. Panacea: Panoramic and Controllable Video Generation for Autonomous Driving. *arXiv preprint arXiv:2311.16813*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2024. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *IJCV*.

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

Yang, L.; Luo, P.; Change Loy, C.; and Tang, X. 2015. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Yu, L.; Cheng, Y.; Sohn, K.; Lezama, J.; Zhang, H.; Chang, H.; Hauptmann, A. G.; Yang, M.-H.; Hao, Y.; Essa, I.; et al. 2023. Magvit: Masked generative video transformer. In *CVPR*.

Zhang, L.; Rao, A.; and Agrawala, M. 2023. Adding conditional control to text-to-image diffusion models. In *ICCV*.