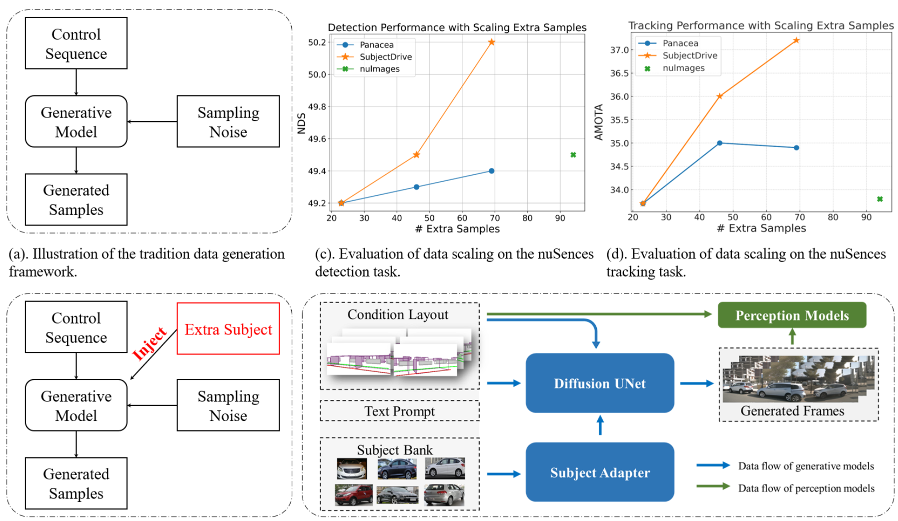
SubjectDrive: Scaling Generative Data in Autonomous Driving via Subject Control

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], and 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

Project Page: https://subjectdrive.github.io/



(a). Illustration of the tradition data generation framework.

(c). Evaluation of data scaling on the nuSences detection task.

(d). Evaluation of data scaling on the nuSences tracking task.

(b). Illustration of the SubjectDrive data generation framework.

(e). Illustration of using SubjectDrive to produce perception training data in autonomous driving.

Fig. 1: An overview of the proposed SubjectDrive framework and its effectiveness in enhancing BEV perception tasks. (a) Traditional data generation framework that uses the control sequence and sampling noise to generate synthetic data. (b) Compared with the traditional framework, our SubjectDrive introduces additional synthesis diversity by incorporating extra subject control. (c)-(d) Evaluation of detection and tracking performance with data scaling. (e) Illustration of using the SubjectDrive framework to produce perception training data in autonomous driving.

---

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

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.

Keywords: Video Generation · Autonomous Driving · Data Scaling · Subject Control

1 Introduction

Deep generative models [4, 15, 34, 37, 47, 48] have made significant progress recently, excelling in producing high-quality, realistic visual contents. Diffusion models [15,37], a key contributor to this advancement, are noted for their stable and superior-quality sample generation. Leveraging modern diffusion models, recent breakthroughs in controllable technology [48] now enable precise, flexible content customization. These advancements make it possible to create synthetic samples nearly indistinguishable from human-annotated data, sparking increased interest in using generative models for effective natural data synthesis in complex discriminative tasks [2, 30, 36].

BEV (Bird's Eye View) perception in autonomous driving is one of the most crucial yet challenging discriminative tasks in computer vision. Training robust BEV perception models, such as those for 3D object detection [23–25, 32, 40] and tracking [9, 29, 32], demands large-scale annotated samples. However, acquiring such datasets is not only costly and time-intensive but also encumbered by significant concerns regarding data privacy and usage rights. These barriers significantly hinder the processes of data collection, labeling, release, and exchange, ultimately slowing progress in the field. Amidst the significant strides in deep generative models, a pressing question emerges: Could these innovations transform the data production pipelines for autonomous driving, thereby catalyzing its advancement?

There have been a few initial attempts at exploring generative data in autonomous driving. Early studies focused on creating street-view images to enhance image-based BEV perception methods, introducing solutions such as BEVGen [38], BEVControl [44], and MagicDrive [11]. More recent efforts have ventured into generating driving scene videos [20, 27, 43], aiming to bolster more sophisticated temporal BEV perception models [23, 40]. A notable example, Panacea [43], excels in producing multi-view videos with BEV controlling capabilities, significantly aiding in the training of the advanced BEV perception model [40].

Albeit promising, most existing works focus on small-scale training [20,27,43], which fails to fully exploit the potential of generative data in producing an inexhaustible supply of samples. Notably, early methods have yet to demonstrate superior performance in supporting downstream perception tasks, especially when compared to models pre-trained on large-scale, realistic datasets. For instance, the advanced generative method Panacea achieves a nuScenes Detection Score (NDS) of 49.2 in the 3D detection tasks. This score is lower than that of the model pre-trained on the nuImages [5] dataset, as illustrated in Fig. 1(c).

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. For instance, as shown in Fig. 1(c-d), tripling the generative data via Panacea yields only a minor improvement. 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, as illustrated in Fig. 1(b). 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.

Specifically, SubjectDrive is designed around a trio of innovative modules that collectively enable robust subject control capabilities. Initially, the model leverages a subject prompt adapter, integrating subject control seamlessly 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, as demonstrated by the visualizations in Fig. 1(e).

We have conducted extensive experiments to validate the effectiveness of our proposed method on the widely used nuScenes [5] dataset. Compared to existing methods, SubjectDrive not only achieves superior performance but also offers improved scalability. Remarkably, our method represents the first generative approach capable of enhancing the performance of perception models beyond what is possible with pre-trained models on the nuImages dataset. These outstanding results underscore the potential of generative data to revolutionize autonomous driving technologies, marking a promising path forward in this field.

2 Related Work

2.1 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 [2, 36], semantic segmentation [30], and visual object tracking [19]. Recently, there has been a burgeoning interest in employing generative data for the challenging BEV perception tasks [23, 40, 49], which demand precise geometric and appearance alignment from generative models. Initial works aim at synthesizing street-view images [11, 38, 44], with subsequent studies extending into the generation of driving scene videos [27, 41, 43]. 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 [2] 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 [1], SVD [3], and Sora [31]. 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.

2.2 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 [8]. These approaches [7, 12, 18, 21, 28] initially emerged in image generation research to facilitate identity-preserving applications. Early methods [10, 12, 18, 21, 35] adopted a fine-tuning framework to adjust a specific model capable of generating images of the desired subject. A notable example is DreamBooth [35], 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 [28, 46] were developed. Subject Diffusion [28], 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 [26] and others [7], have expanded tuning-free subject control to scenarios involving multiple subjects.

In the video domain, there is also growing interest in subject-controllable generation [6, 16, 42]. For instance, VideoBooth [16] proposed a framework capable of generating consistent videos containing the subjects specified in image prompts. VideoDreamer [6] introduced the Disen-Mix Fine-tuning and Human-in-the-Loop Refine-tuning strategy. Additionally, CustomVideo [42] 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. However, unlike previous efforts, we explore the significance of such subject-controlled generation for data scaling, especially in the context of autonomous driving.

3 Method

3.1 Preliminary: Latent Diffusion Models

Diffusion Models (DMs) [15,37] 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 $\varSigma_\theta$ 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.

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

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

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

Given the significant computational burden of diffusion models in generating high-resolution images or videos, Latent Diffusion Models (LDMs) [34] have become popular in recent studies. With a pre-trained auto-encoder [17] 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.

3.2 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 [43] to produce high-quality driving-scene videos, they narrowly uplifts the performance on downstream perception tasks as illustrated in Fig. 1(c)(d). 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 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 [43], which is a strong baseline for multi-view video generation. The overall training framework of SubjectDrive is illustrated in Fig. 2. 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 [48] 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 [33] 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 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

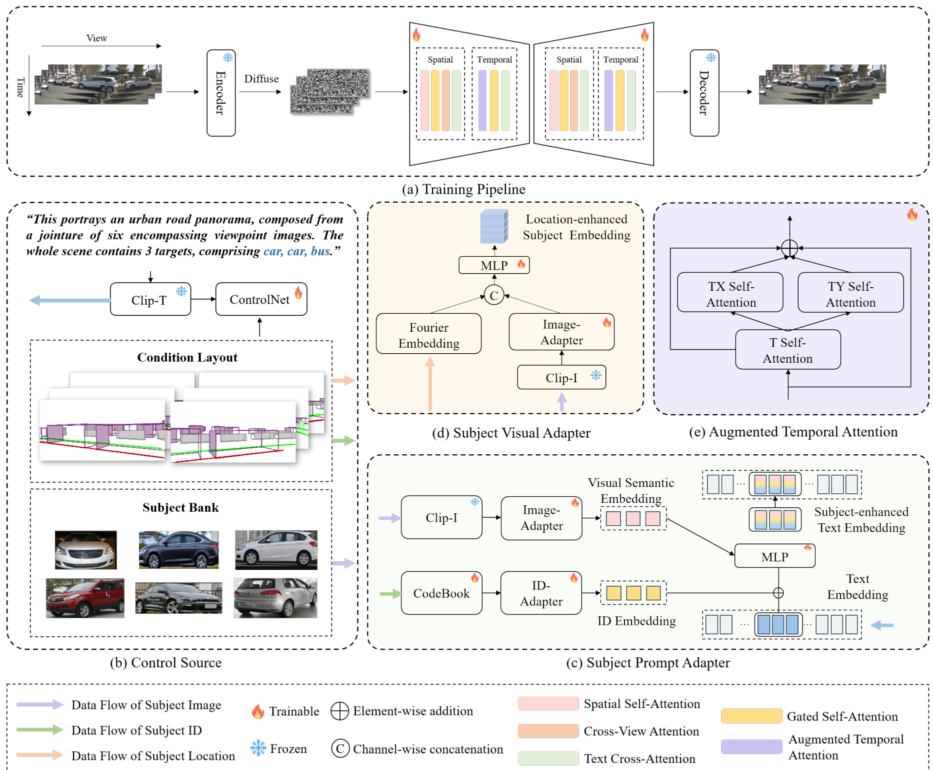


Fig. 2: Overview of SubjectDrive. (a) The pipeline of SubjectDrive involves a frozen auto-encoder and a trainable UNet-based diffusion model. (b) Different control signal sources including extended text prompt, condition layout, and subject bank. (c) The Subject Prompt Adaptor which augments the original text embedding of extended prompt with corresponding ID identifier and visual semantic information to enhance the expressity of subjects. (d) The Subject Visual Adapter which injects location-enhanced subject information into visual features cooperated with gated self-attention. (e) The Augmented Temporal Attention integrates conventional temporal 1D attention with decomposed attention on temporal-horizontal (TX) plane and temporal-vertical (TY) plane to effectively capture long-range movements of subjects.

semantic embeddings z_i^v by

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

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.

Inspired by [22], the 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

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

To inject the location-enhanced subject embedding into generated frame, we employ the gated self-attention layer [22] 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

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

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 effectively 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 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

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

4 Experiment

4.1 Evaluation Datasets and Metrics

Datasets. We employ the nuScenes dataset to train SubjectDrive and utilize it to evaluate the visual fidelity and controllability of the generated data. The nuScenes dataset comprises 1,000 scenes, each lasting 20 seconds, with annotations provided at a frequency of 2Hz and featuring a comprehensive 360° camera field of view. It includes 1.4 million 3D bounding boxes, spanning 10 categories: car, truck, bus, trailer, construction vehicle, pedestrian, motorcycle, bicycle, barrier, and traffic cone.

Evaluation Metrics. Following Panacea, we employ perceptual model evaluation to validate two key objectives. First, we assess the controllability of the generative model, specifically how closely the data generated aligns with given BEV annotations. We utilize a pre-trained StreamPETR model to compare the validation performance of our generated data with that of real data. Second, we explore the extent to which the generative model's data improves the performance of the perceptual model. We conduct evaluations on both detection and tracking tasks. For the detection task, we use the nuScenes Detection Score (NDS), mean Average Precision (mAP), mean Average Orientation Error (mAOE), and mean Average Velocity Error (mAVE) as metrics. For the tracking task, our metrics include Average Multi-Object Tracking Accuracy (AMOTA), Average Multi-Object Tracking Precision (AMOTP), Recall (RECALL), and Multi-Object Tracking Accuracy (MOTA). Additionally, we verify the visual fidelity of our synthesized samples using both the Fréchet Inception Distance (FID) [14] and the Fréchet Video Distance (FVD) [39].

4.2 Implementation Details

SubjectDrive employs a two-stage video generation strategy: the first stage generates images, and the second stage generates videos. The first stage is optimized for 56k steps, while the second stage is optimized for 84k steps. During sampling, we utilize the Denoising Diffusion Implicit Models (DDIM) sampler, configured to 25 steps, to yield video clips at a resolution of 256×512 across 8 frames. Furthermore, we have chosen StreamPETR with a ResNet50 [13] backbone as our evaluating perception model, which is trained at a resolution of 256×512.

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 [45] 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.

4.3 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 perception 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.

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. Furthermore, when combining

Table 1: Evaluation of data scaling on detection and tracking tasks.

#Extra Samples	Source	Tracking		Detection	
		AMOTA ↑	AMOTP ↓	NDS ↑	MAP ↑
0	nuScenes [5]	30.1	1.379	46.9	34.5
93k	nuImages [5]	33.8 (+3.7%)	1.324	49.5 (+2.6%)	37.8
23k	Panacea [43]	33.7 (+3.6%)	1.353	49.2 (+2.3%)	37.1
46k	Panacea [43]	35.0 (+4.9%)	1.346	49.3 (+2.4%)	37.3
69k	Panacea [43]	34.9 (+4.8%)	1.348	49.4 (+2.5%)	37.1
23k	SubjectDrive	33.7 (+3.6%)	1.353	49.2 (+2.3%)	37.1
46k	SubjectDrive	36.0 (+5.9%)	1.322	49.5 (+2.6%)	37.7
69k	SubjectDrive	37.2 (+**7.1%**)	1.317	50.2 (+**3.3%**)	38.1

Table 2: Comparison of performance 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.

DataType	Method	MAP↑	NDS↑	MAOE↓	MAVE↓
Real	Real Only	34.5	46.9	59.4	29.1
Generated	Panacea [43]	22.5	36.1	72.7	46.9
	SubjectDrive	**28.0 (+5.5%)**	**41.1 (+5.0%)**	**69.5**	**37.0**
Real+Generated	DriveDreamer [41]	35.8	39.5	-	-
	WoVoGen∗ [27]	36.2	18.1	15.7	123.4
	MagicDrive [11]	35.4	39.8	-	-
	Panacea [43]	37.1	49.2	54.2	27.3
	SubjectDrive	**38.1**	**50.2**	**52.2**	**26.4**

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

DataType	Method	AMOTA ↑	MOTA ↑	RECALL ↑	AMOTP ↓
Real	Real Only	30.1	27.1	41.8	1.379
Generated	SubjectDrive	23.4	23.5	32.3	1.544
Real+Generated	Panacea [43]	33.7	30.6	44.8	1.353
	SubjectDrive	**37.2 (+3.5%)**	**33.3 (+2.7%)**	**47.3 (+2.5%)**	**1.317**

synthetic data with real data, SubjectDrive surpasses all other methods by a significant margin.

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.

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.

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

Method	Multi-View	Multi-Frame	FVD↓	FID↓
BEVGen [38]	✓		-	25.54
BEVControl [44]	✓		-	24.85
MagicDrive [11]	✓		-	16.20
DriveDreamer [41]		✓	452	52.6
WoVoGen [27]	✓	✓	418	27.6
Panacea [43]	✓	✓	139	16.96
SubjectDrive	✓	✓	**124**	**15.98**

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

SVA	SPA	ATA	mAP↑	NDS↑	AMOTA↑	RECALL ↑
–	–	–	18.8	32.1	11.4	21.3
✓	–	–	20.0	34.3	15.6	24.5
✓	✓	–	22.1	36.1	17.4	27.1
✓	✓	✓	22.8 (+**4.0%**)	36.3 (+**4.2%**)	18.1 (+**6.7%**)	28.1 (+**6.8%**)
Real Validation Data			34.5	46.9	30.1	41.8

4.4 Ablation Studies

The alignment between generated samples and the provided BEV conditional labels is essential for evaluating generative models' controllability. It also acts as a key indicator of the synthetic data's applicability. In this section, we will conduct ablation studies to justify the design choices in SubjectDrive using this metric.

Table 5 presents our detailed evaluation results for the Subject Visual Adapter (SVA), Subject Prompt Adapter (SPA), and Augmented Temporal Attention (ATA). First, compared to the baseline, introducing the SVA resulted in a significant improvement of 2.2 in NDS and 4.2 in AMOTA, demonstrating its effectiveness in enhancing label alignment. This improvement is achieved by injecting extra subject control through visual features and their corresponding positional embeddings. Second, by integrating the SPA with the SVA, we observed an additional increase of 2.1 in mAP and 1.8 in AMOTA. These results indicate that these two subject control modules are complementary, with each playing a crucial role in the SubjectDrive framework. Third, the final model includes ATA, complementing both SVA and SPA, and yields further improvements of 0.7 in AMOTA and 1.0 in RECALL. These results underscore the efficacy of ATA in enhancing alignment accuracy and temporal consistency across tasks.

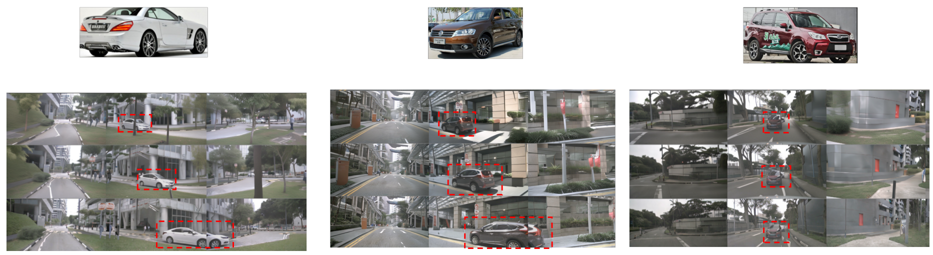


Fig. 3: Generate subject-controlled videos by SubjectDrive.

4.5 Visualisation results

Consistent Multi-View Video Generation. As illustrated in Fig. 4, for the six-view, eight-frame generated video, SubjectDrive produces temporally and view-consistent videos on the nuScenes validation set.

Fig. 4: Multi-view videos generated by SubjectDrive.

Subject-controlled Video Generation. Fig. 3 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. 5 illustrates the driving scene video generated by our method, which adheres to the BEV layout conditions. From this visualization, it is evident that our generated synthetic data closely aligns with the specified BEV conditions, showcasing superior layout control and alignment capabilities.

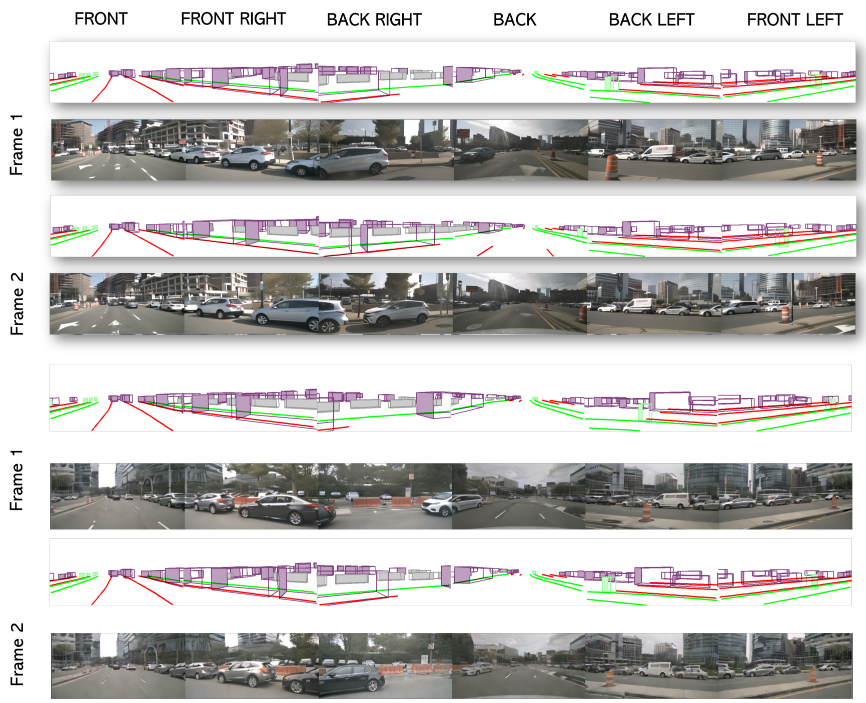


Fig. 5: Controllable multi-view videos generated by SubjectDrive.

5 Conclusion

In this work, we introduce SubjectDrive, a novel video generation framework designed to enhance the scalability and sampling diversity of generative models. The architecture of SubjectDrive features three key innovations: a subject prompt adapter, a subject visual adapter, and augmented temporal attention. Collectively, these innovations enable a robust subject control capability. Empirically, we find that this subject control feature significantly enhances our model's ability to produce diverse samples. Through extensive experimentation, SubjectDrive not only demonstrates superior performance but also exhibits strong scaling capabilities. Impressively, our model is the first generative approach to improve the performance of perception models beyond the capabilities of pretrained models on the nuImages dataset. These results underscore the transformative potential of generative data in advancing autonomous driving technologies, pointing to a promising direction for future developments in the field.

References

1. Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 4
2. Azizi, S., Kornblith, S., Saharia, C., Norouzi, M., Fleet, D.J.: Synthetic data from diffusion models improves imagenet classification. arXiv preprint arXiv:2304.08466 (2023) 2, 4
3. Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., et al.: Stable video diffusion: Scaling latent video diffusion models to large datasets. arXiv preprint arXiv:2311.15127 (2023) 4
4. Blattmann, A., Rombach, R., Ling, H., Dockhorn, T., Kim, S.W., Fidler, S., Kreis, K.: Align your latents: High-resolution video synthesis with latent diffusion models. In: CVPR (2023) 2
5. Caesar, H., Bankiti, V., Lang, A.H., Vora, S., Liong, V.E., Xu, Q., Krishnan, A., Pan, Y., Baldan, G., Beijbom, O.: nuscenes: A multimodal dataset for autonomous driving. In: CVPR (2020) 3, 10
6. Chen, H., Wang, X., Zeng, G., Zhang, Y., Zhou, Y., Han, F., Zhu, W.: Videodreamer: Customized multi-subject text-to-video generation with disen-mix finetuning. arXiv preprint arXiv:2311.00990 (2023) 4, 5
7. Chen, X., Huang, L., Liu, Y., Shen, Y., Zhao, D., Zhao, H.: Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481 (2023) 4
8. Choi, J., Choi, Y., Kim, Y., Kim, J., Yoon, S.: Custom-edit: Text-guided image editing with customized diffusion models. arXiv preprint arXiv:2305.15779 (2023) 4
9. Fischer, T., Yang, Y.H., Kumar, S., Sun, M., Yu, F.: Cc-3dt: Panoramic 3d object tracking via cross-camera fusion. arXiv preprint arXiv:2212.01247 (2022) 2
10. Gal, R., Alaluf, Y., Atzmon, Y., Patashnik, O., Bermano, A.H., Chechik, G., Cohen-Or, D.: An image is worth one word: Personalizing text-to-image generation using textual inversion. arXiv preprint arXiv:2208.01618 (2022) 4
11. Gao, R., Chen, K., Xie, E., Hong, L., Li, Z., Yeung, D.Y., Xu, Q.: Magicdrive: Street view generation with diverse 3d geometry control. arXiv preprint arXiv:2310.02601 (2023) 2, 4, 11, 12
12. Han, L., Li, Y., Zhang, H., Milanfar, P., Metaxas, D.N., Yang, F.: Svdiff: Compact parameter space for diffusion fine-tuning. In: ICCV (2023) 4
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: CVPR (2016) 9
14. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS **30** (2017) 9

15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. NeurIPS **33** (2020) 2, 5
16. Jiang, Y., Wu, T., Yang, S., Si, C., Lin, D., Qiao, Y., Loy, C.C., Liu, Z.: Videobooth: Diffusion-based video generation with image prompts. arXiv preprint arXiv:2312.00777 (2023) 4
17. Kingma, D.P., Welling, M.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013) 5
18. Kumari, N., Zhang, B., Zhang, R., Shechtman, E., Zhu, J.: Multi-concept customization of text-to-image diffusion. In: CVPR (2023) 4
19. Li, P., Liu, Z., Chen, K., Hong, L., Zhuge, Y., Yeung, D.Y., Lu, H., Jia, X.: Trackdiffusion: Multi-object tracking data generation via diffusion models. arXiv preprint arXiv:2312.00651 (2023) 4
20. Li, X., Zhang, Y., Ye, X.: Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. arXiv preprint arXiv:2310.07771 (2023) 2, 3
21. Li, Y., Liu, H., Wen, Y., Lee, Y.J.: Generate anything anywhere in any scene. arXiv preprint arXiv:2306.17154 (2023) 4
22. Li, Y., Liu, H., Wu, Q., Mu, F., Yang, J., Gao, J., Li, C., Lee, Y.J.: Gligen: Open-set grounded text-to-image generation. In: CVPR (2023) 8
23. Li, Z., Wang, W., Li, H., Xie, E., Sima, C., Lu, T., Qiao, Y., Dai, J.: Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In: ECCV (2022) 2, 3, 4
24. Liu, Y., Wang, T., Zhang, X., Sun, J.: Petr: Position embedding transformation for multi-view 3d object detection. In: ECCV (2022) 2
25. Liu, Y., Yan, J., Jia, F., Li, S., Gao, A., Wang, T., Zhang, X.: Petrv2: A unified framework for 3d perception from multi-camera images. In: ICCV (2023) 2
26. Liu, Z., Zhang, Y., Shen, Y., Zheng, K., Zhu, K., Feng, R., Liu, Y., Zhao, D., Zhou, J., Cao, Y.: Cones 2: Customizable image synthesis with multiple subjects. arXiv preprint arXiv:2305.19327 (2023) 4
27. Lu, J., Huang, Z., Zhang, J., Yang, Z., Zhang, L.: Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. arXiv preprint arXiv:2312.02934 (2023) 2, 3, 4, 11, 12
28. Ma, J., Liang, J., Chen, C., Lu, H.: Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410 (2023) 4
29. Marinello, N., Proesmans, M., Van Gool, L.: Triplettrack: 3d object tracking using triplet embeddings and lstm. In: CVPR (2022) 2
30. Nguyen, Q., Vu, T., Tran, A., Nguyen, K.: Dataset diffusion: Diffusion-based synthetic data generation for pixel-level semantic segmentation. NeurIPS **36** (2024) 2, 4
31. OpenAI: Sora: Creating video from text. https://openai.com/sora 4
32. Pang, Z., Li, J., Tokmakov, P., Chen, D., Zagoruyko, S., Wang, Y.X.: Standing between past and future: Spatio-temporal modeling for multi-camera 3d multi-object tracking. In: CVPR (2023) 2
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021) 6
34. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022) 2, 5

35. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In: CVPR (2023) 4
36. Sarıyıldız, M.B., Alahari, K., Larlus, D., Kalantidis, Y.: Fake it till you make it: Learning transferable representations from synthetic imagenet clones. In: CVPR (2023) 2, 4
37. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: ICLR (2020) 2, 5
38. Swerdlow, A., Xu, R., Zhou, B.: Street-view image generation from a bird's-eye view layout. arXiv preprint arXiv:2301.04634 (2023) 2, 4, 12
39. Unterthiner, T., Van Steenkiste, S., Kurach, K., Marinier, R., Michalski, M., Gelly, S.: Towards accurate generative models of video: A new metric & challenges. arXiv preprint arXiv:1812.01717 (2018) 9
40. Wang, S., Liu, Y., Wang, T., Li, Y., Zhang, X.: Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In: ICCV (2023) 2, 3, 4
41. Wang, X., Zhu, Z., Huang, G., Chen, X., Lu, J.: Drivedreamer: Towards real-world-driven world models for autonomous driving. arXiv preprint arXiv:2309.09777 (2023) 4, 11, 12
42. Wang, Z., Li, A., Xie, E., Zhu, L., Guo, Y., Dou, Q., Li, Z.: Customvideo: Customizing text-to-video generation with multiple subjects. arXiv preprint arXiv:2401.09962 (2024) 4, 5
43. Wen, Y., Zhao, Y., Liu, Y., Jia, F., Wang, Y., Luo, C., Zhang, C., Wang, T., Sun, X., Zhang, X.: Panacea: Panoramic and controllable video generation for autonomous driving. arXiv preprint arXiv:2311.16813 (2023) 2, 3, 4, 5, 6, 10, 11, 12
44. Yang, K., Ma, E., Peng, J., Guo, Q., Lin, D., Yu, K.: Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. arXiv preprint arXiv:2308.01661 (2023) 2, 4, 12
45. Yang, L., Luo, P., Change Loy, C., Tang, X.: A large-scale car dataset for fine-grained categorization and verification. In: CVPR (2015) 10
46. Ye, H., Zhang, J., Liu, S., Han, X., Yang, W.: Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. arXiv preprint arXiv:2308.06721 (2023) 4
47. Yu, L., Cheng, Y., Sohn, K., Lezama, J., Zhang, H., Chang, H., Hauptmann, A.G., Yang, M.H., Hao, Y., Essa, I., et al.: Magvit: Masked generative video transformer. In: CVPR (2023) 2
48. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: ICCV (2023) 2, 6
49. Zhou, B., Krähenbühl, P.: Cross-view transformers for real-time map-view semantic segmentation. In: CVPR (2022) 4