This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

DriveGEN: Generalized and Robust 3D Detection in Driving via Controllable Text-to-Image Diffusion Generation

Hongbin Lin[1,2] Zilu Guo[1,2] Yifan Zhang[3] Shuaicheng Niu[4]
Yafeng Li[5] Ruimao Zhang[6] Shuguang Cui[2,1] Zhen Li[2,1*]
[1]FNii-Shenzhen [2] SSE, CUHK-Shenzhen [3] National University of Singapore
[4] Nanyang Technological University [5]Baoji University of Arts and Sciences [6]Sun Yat-sen University



Figure 1. An illustration of DriveGEN for enhancing vision-centric 3D detectors. DriveGEN extends original training images to diverse Out-of-Distribution (OOD) scenarios without additional diffusion model training, preserving all annotated objects. Even with a single augmentation, i.e., 'Snow', the augmented detector [63] is improved not only on the original KITTI [10] val set (Ideal) and already known 'Snow' scenario, but also in **unseen** 'Fog' and 'Defocus' scenarios, demonstrating a comprehensive model generalizability improvement.

Abstract

In autonomous driving, vision-centric 3D detection aims to identify 3D objects from images. However, high data collection costs and diverse real-world scenarios limit the scale of training data. Once distribution shifts occur between training and test data, existing methods often suffer from performance degradation, known as Out-of-Distribution (OOD) problems. To address this, controllable Text-to-Image (T2I) diffusion offers a potential solution for training data enhancement, which is required to generate diverse OOD scenarios with precise 3D object geometry. Nevertheless, existing controllable T2I approaches are restricted by the limited scale of training data or struggle to preserve all annotated 3D objects. In this paper, we present ***DriveGEN****, a method designed to improve the robustness of 3D detectors in Driving via Training-Free Controllable Text-to-Image Diffusion Generation. Without extra diffusion model training, DriveGEN consistently preserves objects with precise 3D geometry across diverse OOD generations, consisting of 2 stages: 1) Self-Prototype Extraction: We empirically find that self-attention features are semantic-aware but require accurate region selection for 3D objects. Thus, we extract precise object features via layouts to capture 3D object geometry, termed self-prototypes. 2) Prototype-Guided Diffusion: To preserve objects across various OOD scenarios, we per-*

*Corresponding author.

form semantic-aware feature alignment and shallow feature alignment during denoising. Extensive experiments demonstrate our effectiveness in improving 3D detection. The code is available at github.com/Hongbin98/DriveGEN.

1. Introduction

Three-dimensional (3D) object detection intends to identify objects and assess their spatial and dimensional attributes via various sensor inputs, which is widely explored in Autonomous Driving (AD) and robotics [3–5, 19, 49, 50, 58]. To save sensor costs, there is a growing trend towards implementing perception systems by vision-centric 3D detection which relies solely on single or multi-view RGB images along with camera calibration information. Despite challenges in this field, existing methods [53, 56, 60] have demonstrated promising results across various benchmarks.

Behind the success, one prerequisite is that test images share the same distribution as training images. However, this assumption is often not held in diverse real-world scenarios [21, 32, 62, 64], especially for perception systems that are required for long durations, like autonomous vehicles navigating through different weather conditions. Once the environment changes, well-trained detectors may suffer performance degradation due to the existence of *data distribution shifts* between training images and test images. As shown in Figure 1, the original detector performs well in the ideal (sunny) scenario while failing to maintain stable performance in the Snow and Fog scenario due to the out-of-distribution (OOD) issue. Considering the widespread application of 3D detection, significant performance degradation on OOD test data may result in traffic accidents and serious safety risks. Thus, it is essential to address the OOD generalization problem of vision-centric 3D detection.

Recently, two attempts sought to solve the distribution shifts in AD by test-time model adaptation [22] or restoring OOD scenarios to ideal conditions with a weather-adaptive diffusion model [28]. However, they require additional consumption during the test phase, leading to additional energy costs while the driving range is crucial in AD. To save test-time computation costs, we explore controllable Text-to-Image (T2I) diffusion generation for training data augmentation, thereby improving the robustness of 3D detectors. Prior training-based controllable T2I approaches such as ControlNet [59] and T2i-adapter [27] offer users fine-grained spatial control but rely on substantial training data to train auxiliary modules, making them impractical for handling OOD scenarios in AD. As for training-free methods, they [25, 46] leverage low-resolution self-attention features to capture image structures for diffusion model guidance. However, these features are relatively coarse for preserving object geometric characteristics of different classes in 3D object detection. As shown in Figure 2, the object geometry may be lost after multiple rounds of downsampling (32x), leading to severe orientation error, position misalignment and omissions, particularly for tiny objects. Therefore, an eligible training-free controllable T2I diffusion method for training data enhancement in 3D detection should preserve all objects with precise 3D geometry.

To address these challenges, we propose a method namely **DriveGEN** for Generalized and Robust 3D Detection in **Driv**ing via Training-Free Controllable Text-to-Image Diffusion **GEN**eration. DriveGEN enhances the model robustness within various OOD scenarios generation, consisting of two stages: 1) Self-Prototype Extraction. To capture accurate object geometric characteristics, we first extract principal components of self-attention features [25, 46] to obtain coarse semantic-aware features. Then, we adopt layouts (*i.e.,* bounding boxes) to re-weight features in object regions via a peak function, termed self-prototypes. 2) Prototype-Guided Diffusion. To preserve objects with precise 3D geometry, we perform semantic-aware feature alignment with self-prototypes during the denoising phase. However, these semantic-aware features are relatively coarse to represent tiny objects as shown in Figure 2. Therefore, we devise a shallow feature alignment strategy to constrain the diffusion process and retrain tiny objects.

Contributions: 1) To the best of our knowledge, we are the first to explore training-free controllable T2I diffusion generation to enhance the model robustness of vision-centric 3D detection. Even with a single augmentation, DriveGEN improves the 3D detector [63] by an average of 7.6 mAP across 13 OOD scenarios. 2) With self-prototype extraction and prototype-guided diffusion stages, DriveGEN enables OOD scenario augmentation while preserving precise 3D geometry for all objects, enhancing the applicability of diffusion models in autonomous driving. 3) Extensive experiments show DriveGEN significantly improves existing monocular 3D detectors across 13 OOD corruption on KITTI and multi-view 3D detectors on the real scenarios (Night, Rainy) of nuScenes, demonstrating the effectiveness of DriveGEN in boosting model generalizability.

2. Related Work

We first review the literature on explorations of model robustness in autonomous driving, and then discuss controllable T2I diffusion generation methods. More discussions on vision-centric 3D object detection are in Appendix A.

Model Robustness in Autonomous Driving. The reliability of vision-centric autonomous driving systems depends on the robustness of perception models, especially under challenging scenarios like adverse weather, varied lighting, and complex urban environments [52]. To improve robustness, existing methods employ diffusion models to generate synthetic driving scenes for training data augmentation. Specifically, BEVGen [44] and BEVControl [57] employ

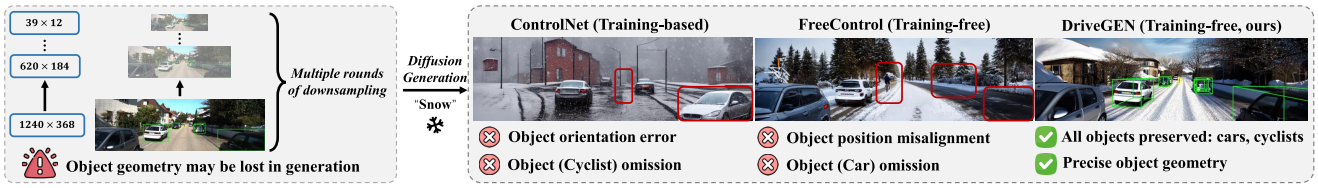


Figure 2. Illustration of the importance of object preservation in 3d detection based on KITTI. ControlNet [59] suffers object orientation errors and omissions even if accurate segmentation masks [34] and rich prompts [6] are provided, showing that training-based methods may struggle with spatial control with limited training data. Additionally, Freecontrol [25] relies on coarse low-resolution features to capture semantic structures that may encounter potential object geometry loss, resulting in object position misalignment and omission issues.

layouts to create flexible scene configurations. Innovative frameworks like Panacea [43] and MagicDrive [14] extend this controllability to multi-view scenes, while models such as DriveDreamer [66] and DrivingDiffusion [16] focus on generating realistic videos for model validation. Recently, RoboFusion [42] has sought to tackle OOD noise scenarios by leveraging visual foundation models like SAM.

Controllable T2I Diffusion. Controllable text-to-image (T2I) diffusion models [35, 61, 69] are increasingly applied in real-world scenarios, where text prompts or scene descriptions guide the creation of synthetic data. By leveraging pre-trained models like Stable Diffusion [35] or other large-scale T2I models [33, 37], existing methods can generate high-fidelity and diverse images based on textual input, allowing for controlled variations [2, 65]. Techniques such as Uni-ControlNet [67], UniControl [30] and Layoutdiffusion [68] further enhance control by integrating multimodal inputs, creating contextually accurate and condition-specific scenes. Recent works also explore the utility of cross-attention mechanisms and latent embeddings to align generated images closely with text descriptions, supporting the generation for perception tasks like semantic segmentation and depth estimation [15, 17]. However, these methods require a large amount of training data to train auxiliary modules, making them impractical for handling OOD scenarios in AD for the limited data scale.

Another alternative choice is training-free controllable T2I diffusion, PnP [46] directly injects the self-attention features of condition images to guide the diffusion process while Freecontrol [25] utilizes PCA to extract principle components of self-attention features for semantic structures extraction. However, existing approaches solely rely on the relatively coarse self-attention features, which often results in object region omissions and misalignment as we mentioned before. To this end, we propose our DriveGEN to preserve all objects with precise 3D geometry.

3. Preliminary

Pre-trained diffusion models [7, 13, 35] are probabilistic generative models, which consists of two complementary stochastic processes: 1) In the *forward* process, Gaussian noise is incrementally added to a clean image $\mathbf{x}_0$. 2) As for the *backward* process, it conducts the iterative process of denoising the initial Gaussian noise image $\mathbf{z}_{N_t}$, where a cleaner image is given at each step. To be specific, this process leverages a denoising network $\epsilon_\theta(\mathbf{z}_t, t)$ to estimate the added noise in the forward process. Subsequently, for $\mathbf{z}_t$ at each step, the network removes the estimated noise perturbation to achieve a cleaner $\mathbf{z}_{t-1}$.

For T2I diffusion models, a previous method [41] has shown that ϵ_θ can approximate the score function for the marginal distributions p_t, *i.e.,* $-\sigma_t \nabla_{\mathbf{z}_t} \log p_t(\mathbf{z}_t|\mathbf{c})$ where σ_t denotes a noise schedule and $\mathbf{c}$ is a text prompt. To incorporate auxiliary information y into the diffusion sampling process, existing methods [7, 8, 25] guide the diffusion process by adding a time-dependent energy function termed $g(\mathbf{z}_t; t, y)$ with a strength coefficient s:

$$\hat{\epsilon}_\theta(\mathbf{z}_t; t, c) = \epsilon_\theta(\mathbf{z}_t; t, c) - s \cdot g(\mathbf{z}_t; t, y),$$

where g can be defined via bounding boxes [51]. In this paper, we adopt a standard choice ϵ_θ following existing methods [25, 46], *i.e.,* U-Net architecture [36] with multi-level self-attention and cross-attention layers [47].

4. Training-Free Controllable T2I Diffusion

4.1. Problem Statement

Without loss of generality, we denote the well-trained 3D detector as $f_{\Theta_d}(\cdot)$ which is obtained by training on labeled training images $\mathcal{D}=\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$. The training images follow the original training distribution $P(\mathbf{x})$ (*i.e.,* $\mathbf{x} \sim P(\mathbf{x})$). Here, Θ_d represents the parameters of the detector and N is the total number of training images. After training, the model is applied to unlabeled test images $\mathcal{D}_t=\{\mathbf{x}^t\}$. Unfortunately, the well-trained model often encounters Out-of-distribution (OOD) test scenarios in autonomous driving (AD) due to prevalent natural corruptions, namely distribution shifts, *i.e.,* $\mathbf{x}^t \sim Q(\mathbf{x})$ and $P(\mathbf{x}) \neq Q(\mathbf{x})$.

As previously discussed, existing methods [22, 28] are relatively time-consuming due to their additional computation costs at test time. Hence, we investigate controllable

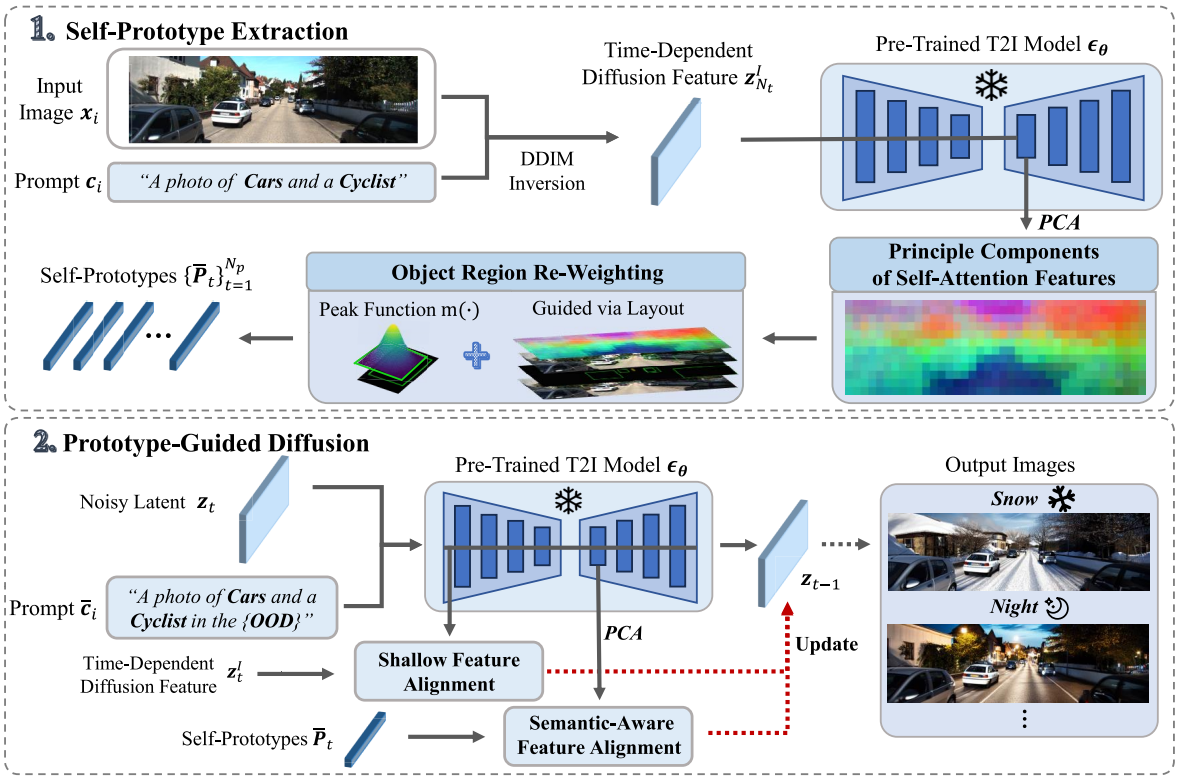


Figure 3. An overview of our DriveGEN, consists of two stages: 1) The Self-Prototype Extraction stage is devised to extract accurate semantic structures of multiple objects. To capture precise locations, we achieve fine-grained self-prototypes via leveraging layouts and the peak function to re-weight object regions rather than directly using coarse self-attention features. 2) The Prototype-Guided Diffusion stage conducts semantic-aware feature alignment for semantic structure matching and shallow feature alignment for tiny object preservation.

Text-to-Image (T2I) diffusion generation for training data augmentation, thereby improving the robustness of 3D detectors. Particularly, it is crucial to preserve object information because the annotations would be reused during detector training. However, training-based approaches [27, 59] require substantial training data to train auxiliary modules for spatial control, which is often unsatisfying in AD. Besides, training-free methods [25, 46] mainly consider the relatively coarse self-attention features, making it challenging to preserve relatively tiny objects.

4.2. Overall Scheme

After examining the characteristics and challenges of robust 3D object detection, we introduce **DriveGEN**, a Training-Free Controllable T2I Diffusion Generation method devised toimprove the robustness of vision-centric 3D detectors. As illustrated in Figure 3, DriveGEN consists of two stages: 1) Self-Prototype Extraction and 2) Prototype-Guided Diffusion. We briefly introduce each stage below.

First, we develop the self-prototype extraction stage (c.f. Section 4.3) to capture the geometric characteristics of objects accurately. Existing methods apply principal component analysis (PCA) on the self-attention features of the first decoder layer of the diffusion model ϵ_θ to capture semantic structures of condition images [25, 46]. However, we empirically find that solely relying on cross-attention relevance between self-attention features and object prompts for region selection may result in object preservation errors (c.f. Figure 2). To address this, DriveGEN leverages layouts to re-weight object regions in the principal components of self-attention features via a peak function and then achieves the self-prototypes which help the diffusion model focus on the accurate object regions during generation.

Second, we introduce the prototype-guided diffusion stage (c.f. Section 4.4) to preserve all objects during the generation process. On the one hand, we conduct semantic-aware feature alignment between self-attention features and self-prototypes to capture accurate object semantic structures. On the other hand, DriveGEN performs shallow feature alignment to further preserve object information by constraining the noisy latent at the shallow level, compensating for the gaps in semantic-aware feature alignment regarding tiny object preservation. The Pseudo-code of DriveGEN is summarized in Algorithm 1.

4.3. Self-Prototype Extraction

To accurately capture the geometric characteristics of multiple objects, we propose a self-prototype extraction strategy that consists of two components: 1) Image structure extraction and 2) Object region re-weighting.

Image Structure Extraction. To achieve zero-shot spatial control, capturing the semantic structures of input images is essential. Prior works [25, 46] demonstrate that self-attention features (*e.g.,* keys) of the T2I diffusion model ϵ_θ are semantic-aware [35]. Thus, given the input image $\mathbf{x}_i$, we first generate the text prompt $\mathbf{c}_i$ via a simple template *'A photo of* $\{object\}$*, ..., and* $\{object\}$*'*. Then, we leverage DDIM Inversion [39] to achieve a series of time-dependent diffusion features $\{\mathbf{z}_t^I \in \mathbb{R}^{N_c \times H \times W}\}_{t=1}^{N_t}$ where N_t denotes the number of DDIM inversion and sampling steps, N_c represents the number of feature channels, and H, W indicate the height and width, respectively. During sampling, we apply PCA on self-attention features to extract principal components $\{\mathbf{P}_t\}_{t=1}^{N_p}$, each $\mathbf{P}_t \in \mathbb{R}^{N_b \times H \times W}$ and $\mathbf{P}_t = \{\mathbf{p}_t^1, \ldots, \mathbf{p}_t^{N_b}\}$. Here N_b denotes the number of principal components and we store $\mathbf{P}_t$ of the initial N_p steps.

Object Region Re-weighting. Existing methods [9, 25] select foreground regions by leveraging cross-attention maps between specific words (*i.e.,* object names) and image features to derive a concept region mask $\mathbf{M}$. However, relying solely on such attention maps may result in omissions and misalignment of object regions, as illustrated in Figure 2. To address this, we introduce a peak function $m(\cdot)$ to re-weight $\mathbf{M}$ for each object region based on layouts $\mathbf{L}$ after the coordinate transformation through downsampling. To be specific, given a point (p, q) from $\mathbf{L}_i$ of $\mathbf{x}_i$, the corresponding object center is defined as $(\hat{p}, \hat{q})$ and then we can calculate $m(p, q)$ by:

$$m(p,q) = \exp\left(-\frac{(p-\hat{p})^2 + (q-\hat{q})^2}{2\sigma^2}\right) \tag{1}$$

where $\sigma \in (0, 1)$ controls the range of the influence area. Then, we update the concept token mask $\mathbf{M}$ by:

$$\hat{\mathbf{M}}(p,q) = \mathbf{M}(p,q) + m(p,q) \cdot \mathbf{L}_i(p,q). \tag{2}$$

The object regions in $\mathbf{L}_i$ are marked as 1 while all other regions as 0. Eventually, we obtain the self-prototypes by $\{\hat{\mathbf{P}}_t\}_{t=1}^{N_p} = \{\mathbf{P}_t\}_{t=1}^{N_p}\big|_{\hat{\mathbf{M}}}$ which guide the diffusion model to localize precise locations of objects and combine $\hat{\mathbf{M}}$ for accurate object region selection.

4.4. Prototype-Guided Diffusion

To preserve all objects with precise 3D geometry during diffusion, we leverage the self-prototypes $\{\hat{\mathbf{P}}_t\}_{t=1}^{N_p}$ and time-dependent diffusion features $\{\mathbf{z}_t^I\}_{t=1}^{N_t}$ for feature alignment, including: 1) semantic-aware feature alignment and 2) shallow feature alignment.

Algorithm 1 The pipeline of the proposed DriveGEN

Require: Training data $\{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^N$; Pretrained T2I diffusion model ϵ_θ; Hyper-parameters s, σ, τ; The OOD scenario.

Stage 1: Self-Prototype Extraction

for each training image x_i **do**
  Get the text prompt c_i and layout $\mathbf{L}_i$ based on $\mathbf{y}_i$;
  Undergo DDIM Inversion to get $\mathbf{P}_t$;
  Get updated $\hat{\mathbf{M}}$ via Eqn. (1), Eqn. (2) for self-prototypes $\hat{\mathbf{P}}_t$.
end for

Stage 2: Prototype-Guided Diffusion

for each training image $\mathbf{x}_i$ **do**
  Update $\mathbf{c}_i$ with specific OOD type;
  for diffusion step $t = 1 \rightarrow N_t$ **do**
    if $t \leq N_p$ **then**
      Calculate g_{sa}, g_{sl} based on Eqn. (3), Eqn. (4);
      Get the gradient based on Eqn.(5);
    end if
    Update the noisy latent $\mathbf{z}_t \rightarrow \mathbf{z}_{t-1}$;
  end for
end for

return Output images for all $\mathbf{x}_i$ in the OOD scenario.

Semantic-aware Feature Alignment. As mentioned before, the self-attention features of the first decoder layer can represent the semantic structures of input images. Thus, we first initialize the noisy latent $\mathbf{z}_{N_t}$ via the $\mathbf{z}_{N_t}^I$. For the step $t \leq N_p$, we perform PCA to obtain the principal components $\mathbf{P}_t$ of the self-attention features. Then, we adopt a pre-defined threshold τ following [9] to select informative regions by $\hat{\mathbf{M}} = \mathbb{I}(\hat{\mathbf{M}} > \tau)$. Specifically, we calculate the semantic-aware feature alignment loss with $\mathbf{P}_t$ and $\hat{\mathbf{P}}_t$:

$$g_{sa} = \|\hat{\mathbf{M}} \odot (\mathbf{P}_t - \hat{\mathbf{P}}_t)\|_2^2, \tag{3}$$

where $\odot$ represents element-wise multiplication, applying the mask $\hat{\mathbf{M}}$ to the difference between $\mathbf{P}_t$ and $\hat{\mathbf{P}}_t$.

Shallow Feature Alignment. As shown in Figure 2, we empirically find that principal components lack sufficient details to represent objects, especially tiny objects. For instance, given a bounding box of a cyclist with a height of 20 pixels and a width of 5 pixels, this object is unable to occupy even one element within principal components after multiple rounds of downsampling (e.g., 32x).

Therefore, DriveGEN introduces shallow feature alignment before the forward process of diffusion, which aims to preserve fine-grained object details by constraining the noisy latent. Specifically, based on the layout $\mathbf{L}_i$ with N_i objects, we calculate the shallow feature alignment loss by:

$$g_{sl} = \frac{1}{N_i} \sum_{p,q} \|\mathbf{L}_i(p,q) \odot (\mathbf{z}_t^I(p,q) - \mathbf{z}_t(p,q))\|_2^2, \tag{4}$$

Overall, the total scheme of DriveGEN is as follows:

$$\hat{\epsilon}_t = (1+s)\epsilon_\theta(\mathbf{z}_t; t, c) - s\epsilon_\theta(\mathbf{z}_t; t, \emptyset) + g_{sa} + g_{sl}. \tag{5}$$

Note that $\emptyset$ denotes null and DriveGEN is a classifier-free guidance [12] process with hyper-parameters s, σ and τ.

Table 1. Comparison on the KITTI-C dataset, severity **level 1** regarding Mean $AP_{3D|R_{40}}$. The **bold** number indicates the best result.

Method	Training-free diffusion	Noise			Blur			Weather				Digital			Avg.
		Gauss.	Shot	Impul.	Defoc.	Glass	Motion	Snow	Frost	Fog	Brit.	Contr.	Pixel	Sat.	
Car, IoU @ 0.7, 0.5, 0.5															
Monoflex [63]		13.06	20.91	14.09	20.17	28.59	30.34	33.64	30.31	19.58	45.22	20.01	29.07	38.85	26.45
• Color Jitter (Traditional aug.)		9.55	15.81	11.90	22.67	25.38	30.12	34.08	30.00	19.29	42.10	19.93	17.17	36.48	24.19
• Brightness (Traditional aug.)		11.44	19.42	12.73	11.18	18.95	22.03	26.64	21.70	13.04	39.61	13.08	21.11	29.73	20.05
• ControlNet (Only Snow aug.)	✗	0.32	1.18	1.28	4.65	11.32	17.04	22.84	19.72	9.53	34.79	8.73	1.25	16.88	11.50
• ControlNet (3 scenarios aug.)	✗	0.70	1.14	0.29	0.35	0.37	0.83	4.07	3.05	1.20	9.43	0.94	0.48	3.20	2.00
• ControlNet (6 scenarios aug.)	✗	0.00	0.00	0.00	0.00	0.00	1.94	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.15
• Freecontrol (Only Snow aug.)	✓	20.39	28.05	22.27	12.52	21.67	21.09	27.97	17.91	10.07	35.99	9.90	28.56	35.27	22.44
• Freecontrol (3 scenarios aug.)	✓	15.43	21.25	14.04	15.33	19.22	16.53	22.01	17.41	15.18	24.66	16.66	25.39	31.44	19.58
• Freecontrol (6 scenarios aug.)	✓	12.85	18.71	16.31	12.06	16.31	12.86	17.64	16.18	14.23	24.52	15.64	22.56	24.28	17.24
• DriveGEN (Only Snow aug.)	✓	16.48	26.72	24.98	31.17	35.55	38.13	**41.39**	38.61	27.85	**49.76**	29.28	38.88	**44.12**	34.07
• DriveGEN (3 scenarios aug.)	✓	**25.63**	**37.04**	**29.13**	34.13	39.15	36.81	38.58	37.98	33.93	45.39	34.66	39.36	43.81	36.58
• DriveGEN (6 scenarios aug.)	✓	24.77	33.79	28.27	**36.92**	**40.33**	**40.45**	40.60	**40.56**	**38.10**	44.83	**39.28**	**41.81**	44.05	**37.98**
MonoGround [31]		13.05	21.77	18.87	20.79	30.74	32.02	34.43	27.02	14.15	46.21	14.63	33.41	35.60	26.36
• Color Jitter (Traditional aug.)		12.88	24.31	18.95	23.07	30.44	31.42	35.94	30.43	19.89	44.66	20.61	29.75	36.65	26.36
• Brightness (Traditional aug.)		14.02	23.52	20.14	23.95	31.78	28.79	35.08	31.87	18.87	42.94	17.75	25.55	37.18	27.03
• ControlNet (Only Snow aug.)	✗	1.76	3.23	4.63	5.20	12.95	14.11	17.70	11.58	3.04	35.21	2.98	7.29	13.98	10.28
• ControlNet (3 scenarios aug.)	✗	0.00	0.00	0.24	1.42	1.68	1.61	4.90	5.40	0.57	17.90	1.12	7.79	6.82	3.80
• ControlNet (6 scenarios aug.)	✗	0.00	0.00	0.00	1.68	1.26	0.35	1.13	0.52	0.44	4.08	0.38	2.22	1.77	1.06
• Freecontrol (Only Snow aug.)	✓	11.75	21.89	15.76	17.70	21.45	21.69	32.08	20.60	13.57	36.05	14.03	26.75	38.35	22.43
• Freecontrol (3 scenarios aug.)	✓	16.31	20.75	17.61	13.10	16.84	14.82	17.72	15.89	11.87	25.20	13.55	22.57	24.49	17.75
• Freecontrol (6 scenarios aug.)	✓	15.20	22.59	15.35	22.00	21.18	18.95	17.69	14.85	14.82	24.02	16.97	22.99	26.12	19.44
• DriveGEN (Only Snow aug.)	✓	17.07	26.78	23.78	32.89	37.52	39.06	**40.61**	34.91	25.29	**46.21**	27.12	38.25	44.45	33.38
• DriveGEN (3 scenarios aug.)	✓	19.79	31.44	27.63	36.84	40.10	39.35	39.17	36.42	28.79	45.42	29.29	42.60	44.99	35.53
• DriveGEN (6 scenarios aug.)	✓	**23.84**	**32.59**	**30.34**	**38.57**	**41.20**	**40.19**	38.16	**38.40**	**32.53**	43.95	**34.80**	**44.10**	**45.13**	**37.21**
Pedestrian, IoU @ 0.5, 0.25, 0.25															
Monoflex [63]		1.16	3.54	0.78	8.06	17.70	15.26	12.72	9.25	5.61	19.87	5.35	1.49	8.65	8.42
• Color Jitter (Traditional aug.)		0.99	3.53	1.41	10.05	14.63	12.00	14.73	12.13	7.72	19.02	9.23	1.11	11.29	9.07
• Brightness (Traditional aug.)		0.63	1.85	1.20	5.01	13.72	12.15	7.93	5.84	1.87	16.89	2.30	0.48	3.77	5.67
• ControlNet (Only Snow aug.)	✗	0.00	0.00	0.00	1.78	8.32	5.84	4.03	3.75	1.31	11.00	1.23	0.00	0.93	2.94
• ControlNet (3 scenarios aug.)	✗	0.00	0.00	0.00	1.15	3.23	1.14	1.64	1.35	0.83	4.06	2.01	0.00	0.00	1.19
• ControlNet (6 scenarios aug.)	✗	0.00	0.00	0.00	0.00	0.36	0.00	0.00	0.00	0.00	2.50	0.00	0.00	0.00	0.22
• Freecontrol (Only Snow aug.)	✓	3.62	5.17	3.21	3.23	5.16	5.94	4.68	3.52	4.28	7.70	4.22	8.51	7.70	5.15
• Freecontrol (3 scenarios aug.)	✓	2.26	3.54	3.04	3.27	4.97	4.25	7.42	4.41	7.27	10.91	6.18	6.83	7.43	5.52
• Freecontrol (6 scenarios aug.)	✓	5.99	9.10	9.08	4.80	7.36	5.84	7.22	7.58	7.10	10.81	9.40	9.90	10.24	8.03
• DriveGEN (Only Snow aug.)	✓	1.25	3.22	4.56	16.21	**19.88**	**20.61**	**19.80**	14.46	8.80	**24.62**	9.11	9.77	18.04	13.10
• DriveGEN (3 scenarios aug.)	✓	6.09	**9.85**	**9.60**	14.96	18.59	16.47	15.45	**16.39**	**15.88**	21.98	**16.85**	**14.55**	**18.13**	14.98
• DriveGEN (6 scenarios aug.)	✓	**6.34**	9.22	7.71	**17.60**	19.21	20.20	16.88	16.19	15.72	23.57	16.75	14.37	17.31	**15.47**
MonoGround [31]		2.67	3.25	5.76	17.57	18.91	17.71	12.96	9.35	4.37	24.15	5.89	3.27	7.16	10.23
• Color Jitter (Traditional aug.)		2.44	3.24	4.11	15.37	18.46	16.50	15.45	12.38	9.14	24.71	9.82	2.07	7.81	10.89
• Brightness (Traditional aug.)		2.79	4.10	7.61	14.51	14.13	14.52	12.12	12.66	5.55	20.80	5.36	2.36	10.31	9.75
• ControlNet (Only Snow aug.)	✗	1.85	1.03	0.81	7.18	9.97	8.32	1.32	2.92	1.37	12.99	1.40	0.28	0.90	3.87
• ControlNet (3 scenarios aug.)	✗	0.00	0.00	0.00	3.60	3.89	1.96	1.14	1.03	1.50	4.21	1.25	0.26	0.77	1.51
• ControlNet (6 scenarios aug.)	✗	0.00	0.00	0.00	0.00	0.00	1.67	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.13
• Freecontrol (Only Snow aug.)	✓	**10.04**	13.09	11.70	14.91	12.61	13.03	15.14	12.35	8.13	15.69	12.34	16.32	14.57	13.07
• Freecontrol (3 scenarios aug.)	✓	0.77	1.81	1.98	2.19	3.08	3.96	3.29	0.52	2.55	3.76	3.10	4.10	5.37	2.81
• Freecontrol (6 scenarios aug.)	✓	6.81	7.54	7.67	5.54	5.45	5.10	2.19	1.31	3.63	8.59	5.35	8.86	6.26	5.72
• DriveGEN (Only Snow aug.)	✓	6.22	6.89	9.27	15.03	17.24	18.81	16.40	13.38	9.81	23.28	11.50	12.04	13.92	13.37
• DriveGEN (3 scenarios aug.)	✓	6.53	9.26	11.17	16.15	**19.60**	**21.02**	**18.67**	16.17	13.72	**24.80**	14.52	19.98	**19.36**	16.23
• DriveGEN (6 scenarios aug.)	✓	9.70	**13.68**	**13.52**	**17.00**	17.74	20.39	17.23	**18.79**	**15.50**	23.09	**15.80**	**20.99**	18.34	**17.06**

5. Experiments

We conduct experiments to validate DriveGEN on monocular and multi-view 3D object detection, including four different training settings with various augmented OOD scenarios: 1) Augmented with traditional techniques (Traditional aug.); 2) Only augmented with Snow (Only Snow aug.); 3) Augmented with Snow, Rain and Fog (3 scenarios aug.); 4) Augmented with Snow, Rain, Fog, Night, Defocus and Sandstorm (6 scenarios aug.).

Datasets. For monocular 3d object detection, we split the images of KITTI [10] into a training set (3712 images) and a validation set (3769 images) following the protocol from Monoflex [63], encompassing three classes: Car, Pedestrian, and Cyclist. We adopt the KITTI-C dataset [22] for validation, which includes 13 types of data corruptions on the validation set across four categories: Noise, Blur, Weather, and Digital [11].

As for multi-view 3D object detection, we first split real-world Night and Rainy validation scenarios based on descriptions of nuScenes [1] following [24]. Next, we augment 500 daytime training scenes under the Snow condition. The augmented detectors are then evaluated on real-world Night and Rainy scenarios, as well as on the widely used Robo3D benchmark [52]. More details of datasets are provided in Appendix B.

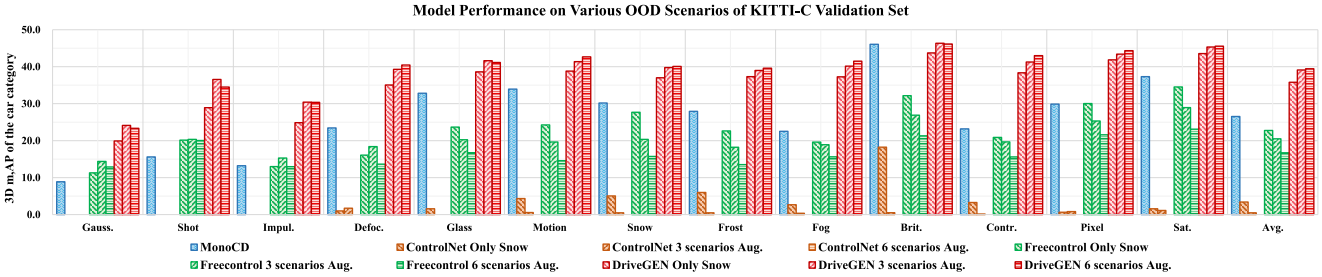


Figure 4. Based on MonoCD [54], we provide more comparisons with baselines on KITTI-C, regarding Mean $AP_{3D|R_{40}}$.

Table 2. Detection results on nuScenes-C and real-world Night and Rainy scenarios of nuScenes, regarding mAP and NDS.

Metric	Method	nuScenes-C									Real-world Scenarios		
		Brightness	CameraCrash	ColorQuant	Fog	FrameLost	LowLight	MotionBlur	Snow	Avg.	nuScenes-Night	nuScenes-Rainy	Avg.
mAP	BEVFormer-tiny	24.26	22.01	23.91	21.98	20.43	16.11	20.78	10.83	20.04	9.31	26.76	18.04
	• DriveGEN (3k Snow)	**26.04**	**24.13**	**25.78**	**24.13**	**21.13**	**17.20**	**22.39**	**11.72**	**21.57**(↑1.53)	**12.19**	**29.11**	**20.65** (↑2.61)
NDS	BEVFormer-tiny	34.93	33.06	34.56	33.04	31.42	28.42	31.81	23.27	31.31	17.92	37.35	27.64
	• DriveGEN (3k Snow)	**37.14**	**35.34**	**36.93**	**35.35**	**33.32**	**30.42**	**34.23**	**24.09**	**33.35**(↑2.04)	**19.37**	**40.95**	**30.16** (↑2.52)

Implementation Details. We implement our method and other baselines in PyTorch [29]. Following Freecontrol [25], we adopt keys from the self-attention of the first U-Net decoder layer as the features. In the self-prototype extraction stage, we run DDIM inversion with 200 steps. Subsequently, we also run 200 steps of DDIM sampling in the prototype-guided diffusion stage. We set s and τ to 7.5 and 0.3 following FreeControl [25], while σ is set to 0.1. All 3D detectors are trained on a mix of original and augmented data. All results in the manuscript are based on Stable Diffusion (SD) 1.5 [35]. More details and the results of SD 2.1 and XL 1.0 are put in Appendix C.

Compared Methods. Based on typical or state-of-the-art (SOTA) 3D detectors [20, 31, 54, 63], we fully compare DriveGEN with following methods: 1) source-only, *i.e.,* directly apply the well-trained model to corrupted test data; 2) Traditional data augmentation, *i.e.,* Color Jitter and Brightness; 3) training-based controllable T2I diffusion: ControlNet [59] trains auxiliary modules for spatial control; 4) training-free controllable T2I diffusion: Freecontrol [25] enables zero-shot control of pretrained diffusion models.

Evaluation Protocols. We report the experimental results in the Average Precision (AP) for 3D bounding boxes, denoted as $AP_{3D|R_{40}}$. On the KITTI-C dataset, the results present the mean values of three difficulty levels and Intersection over Union (IoU) thresholds are set to 0.7, 0.5, 0.5 for Cars and 0.5, 0.25, 0.25 for Pedestrians and Cyclists.

As for nuScenes, the mean average precision (mAP) and nuScenes detection score (NDS) are calculated following BEVFormer [20]. The metrics also include five true positive metrics, including ATE, ASE, AOE, AVE, and AAE for measuring errors in translation, scale, orientation, velocity, and attributes.

5.1. Comparisons with Previous Methods

For monocular 3d object detection, Table 1 gives the following observations: 1) Applying the well-trained detector to corrupted scenarios suffers significant performance degradation due to the data distribution shifts and traditional data augmentation offers limited gains. 2) ControlNet (pretrained on ade20k [70]) fails to enhance 3d detectors with limited training data even if it is provided with accurate segmentation masks [34] and rich prompts [6] for fine-tuning. 3) Freecontrol achieves minor gains in a few cases but generally degrades, indicating that training-free T2I diffusion is a feasible solution but requires careful object preservation. 4) DriveGEN consistently outperforms all compared methods over all categories within various base models for both known and unseen scenarios, like improving MonoFlex by an average of 7.6 mAP across 13 various OOD scenarios with the single Snow scenario augmentation.

As for multi-view 3D object detection, we randomly select 3k daytime training images and apply Snow augmentation (3k Snow) to enhance the detector [20]. Table 2 shows DriveGEN consistently enhances BEVFormer-tiny across 8 OOD scenarios with an average of 1.53 mAP and 2.04 NDS, further confirming our effectiveness. Moreover, we validate the detector on real Night and Rainy scenarios of nuScenes. With the help of DriveGEN, it also achieves notable gains in real-world tasks with an average improvement of 2.61 mAP and 2.52 NDS, further demonstrating our superiority. More detailed results are put in the Appendix D.

5.2. More Augmented Scenarios for Training

An intuitive concern is whether the robustness of detectors progressively improves with the increasing of additional augmented scenarios (*i.e.,* Only Snow aug.→3 scenarios

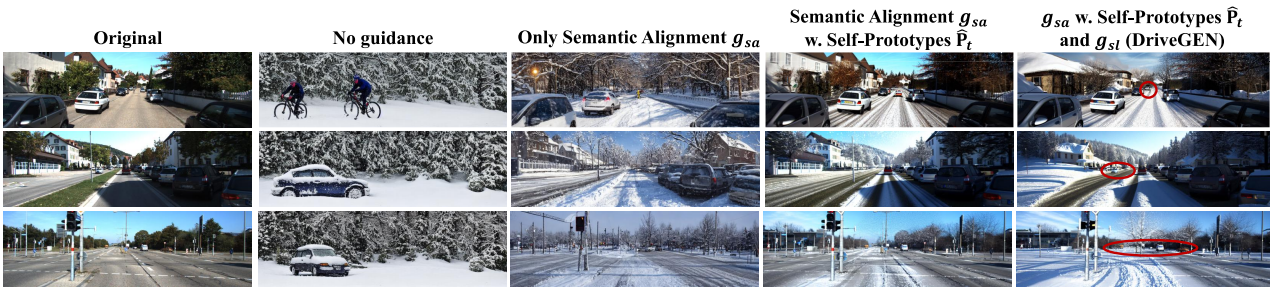


Figure 5. Ablation studies on semantic feature alignment loss g_{sa}, self-prototypes $\hat{\mathbf{P}}_t$ and shallow feature alignment loss g_{sl}.

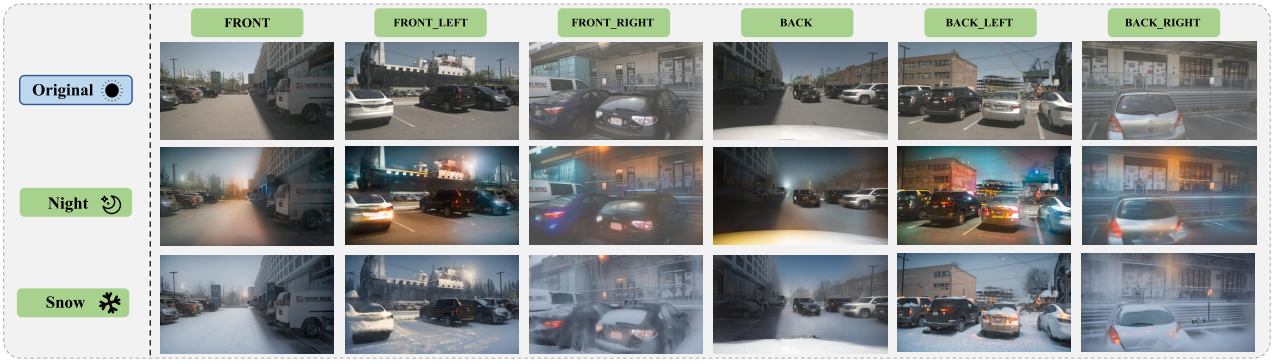


Figure 6. Qualitative results of DriveGEN for multi-view 3D object detection on nuScenes. DriveGEN supports vision-centric 3D detection tasks since our method only requires input images and corresponding object annotations without any additional diffusion model training.

aug.→6 scenarios aug.). Based on Table 1 and Figure 4, we are given additional observations: 1) For ControlNet, the model performance drop for OOD scenarios grows as the number of augmentation scenarios increases, demonstrating that severe object omission leads to noisy training data. 2) For Freecontrol, the model performance across various OOD scenarios remains stable or tends to decline, indicating its object preservation may be unstable. 3) Owing to the precise object preservation, DriveGEN obtains continuous overall performance (*i.e.,* Avg.) improvement across all categories with the incorporation of OOD scenarios.

5.3. Ablation Studies

To examine DriveGEN, we present qualitative results guided by different loss settings. As shown in Figure 5, compared with only generated via text prompts, introducing semantic-aware feature alignment g_{sa} helps preserve more image structures. Next, by guiding the diffusion process with our self-prototypes $\hat{\mathbf{P}}_t$, we achieve more accurate object geometry, alleviating potential object geometry loss. Eventually, introducing our shallow feature alignment g_{sl} further improves object preservation with precise 3D geometry, especially for small objects (highlighted by red circles). More ablation studies are provided in Appendix D.

5.4. Quantitative Results

We provide visualizations for multi-view 3D object detection on nuScenes in Snow scenarios as shown in Figure 6. It is evident that DriveGEN supports existing vision-centric 3D detection tasks since it requires only input images and corresponding object annotations without any additional diffusion model training, demonstrating that DriveGEN achieves superior results even in challenging multi-view tasks. More visualizations are put in Appendix E.

6. Conclusion

In this paper, we propose a method for robust 3D detection in driving via training-free controllable text-to-image diffusion generation, namely DriveGEN. Specifically, DriveGEN consists of: 1) Self-Prototype Extraction: To improve the guidance of self-attention features, we extract the self-prototypes by layouts to capture accurate object geometry, leveraging more precise features to guide diffusion. 2) Prototype-Guided Diffusion: To further preserve 3D object geometry, we conduct semantic-aware feature and shallow feature alignment during diffusion, alleviating the object misalignment and omission issues compared with previous methods. Experiments on KITTI-C and nuScenes demonstrate the effectiveness of DriveGEN in improving model robustness for vision-centric 3D detection.

Acknowledgements

This work was supported by NSFC with Grant No. 62293482, by the Basic Research Project No. HZQB-KCZYZ-2021067 of Hetao Shenzhen HK S&T Cooperation Zone, by Shenzhen General Program No. JCYJ20220530143600001, by Shenzhen-Hong Kong Joint Funding No. SGDX20211123112401002, by the Shenzhen Outstanding Talents Training Fund 202002, by Guangdong Research Project No. 2017ZT07X152 and No. 2019CX01X104, by the Guangdong Provincial Key Laboratory of Future Networks of Intelligence (Grant No. 2022B1212010001), by the Guangdong Provincial Key Laboratory of Big Data Computing, CHUK-Shenzhen, by the NSFC 61931024&12326610, by the Key Area R&D Program of Guangdong Province with grant No. 2018B030338001, by the Shenzhen Key Laboratory of Big Data and Artificial Intelligence (Grant No. ZDSYS201707251409055), by Shaanxi Mathematical Basic Science Research Project(No.23JSY047), and by Tencent & Huawei Open Fund, by China Association for Science and Technology Youth Care Program.

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, and et.al. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11621–11631, 2020. 6

[2] Changjian Chen, Fei Lv, Yalong Guan, Pengcheng Wang, Shengjie Yu, Yifan Zhang, and Zhuo Tang. Human-guided image generation for expanding small-scale training image datasets. *arXiv preprint arXiv:2412.16839*, 2024. 3

[3] Xiaozhi Chen, Kaustav Kundu, Yukun Zhu, and et. al. 3d object proposals for accurate object class detection. *Advances in neural information processing systems*, 28, 2015. 2

[4] Xiaozhi Chen, Huimin Ma, Ji Wan, Bo Li, and Tian Xia. Multi-view 3d object detection network for autonomous driving. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1907–1915, 2017.

[5] Yukang Chen, Jianhui Liu, Xiangyu Zhang, Xiaojuan Qi, and Jiaya Jia. Voxelnext: Fully sparse voxelnet for 3d object detection and tracking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21674–21683, 2023. 2

[6] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 3, 7

[7] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021. 3, 1

[8] Dave Epstein, Allan Jabri, Ben Poole, Alexei Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. *Advances in Neural Information Processing Systems*, 36:16222–16239, 2023. 3

[9] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023. 5

[10] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3354–3361. IEEE, 2012. 1, 6

[11] Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. In *International Conference on Learning Representations*, 2018. 6

[12] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 5

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 3, 1

[14] Yuming Hong, Weiyao Xu, Yi Yang, Hao Zhou, and Xiaoxue Wang. Magicdrive: 3d controllable driving video synthesis using layout and 3d representations. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021. 3

[15] Xinyu Huang, Mingyu Chen, Yunqi Li, Hao Wang, and Yuxing Zhao. Diffumask: Synthetic image and semantic mask annotation using cross-attention in diffusion models. *arXiv preprint arXiv:2211.09123*, 2022. 3

[16] Junsoo Kim, Yeongjae Ye, Dongheon Cho, Jaechan Seo, and Jiseob Jeong. Drivingdiffusion: Realistic multi-view video generation with diffusion models. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. 3

[17] Feng Li, Yuan Zhang, Xinyang Wang, and Meng Liu. Vpd: Vision-based perception diffusion models for human pose estimation and dense visual annotations. *arXiv preprint arXiv:2109.05270*, 2021. 3

[18] Hongyu Li, Pengfei Zhao, Wei Jiang, and Zihan Chen. Sensor fusion with diffusion models for enhanced 3d object detection. *arXiv preprint arXiv:2205.02574*, 2022. 1

[19] Peiliang Li, Xiaozhi Chen, and Shaojie Shen. Stereo r-cnn based 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7644–7652, 2019. 2

[20] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *European conference on computer vision*, pages 1–18. Springer, 2022. 7, 2, 4

[21] Hongbin Lin, Yifan Zhang, Zhen Qiu, Shuaicheng Niu, Chuang Gan, Yanxia Liu, and Mingkui Tan. Prototype-guided continual adaptation for class-incremental unsupervised domain adaptation. In *European Conference on Computer Vision*, pages 351–368. Springer, 2022. 2

[22] Hongbin Lin, Yifan Zhang, Shuaicheng Niu, Shuguang Cui, and Zhen Li. Monotta: Fully test-time adaptation for monocular 3d object detection. In *European Conference on Computer Vision*, pages 96–114. Springer, 2025. 2, 3, 6, 1, 4

[23] Xiaoyang Liu, Yikang Wang, Xinyu Zhang, Wei Ma, Chao Wang, and Yun Chen. Multi-modal fusion for 3d object detection in autonomous driving. *arXiv preprint arXiv:2204.01234*, 2022. 1

[24] Zhijian Liu, Haotian Tang, Alexander Amini, Xinyu Yang, Huizi Mao, Daniela L Rus, and Song Han. Bevfusion: Multi-task multi-sensor fusion with unified bird's-eye view representation. In *2023 IEEE international conference on robotics and automation (ICRA)*, pages 2774–2781. IEEE, 2023. 6, 1

[25] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7465–7475, 2024. 2, 3, 4, 5, 7

[26] Seokhwa Moon, Sanghyuk Hwang, Hyunsu Lee, and Minje Kim. Temporal fusion with attention mechanisms for driving scenarios. *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2022. 1

[27] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4296–4304, 2024. 2, 4

[28] Youngmin Oh, Hyung-Il Kim, Seong Tae Kim, and Jung Uk Kim. Monowad: Weather-adaptive diffusion model for robust monocular 3d object detection. In *European Conference on Computer Vision*, 2025. 2, 3

[29] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems*, 2019. 7, 2

[30] Yuheng Qin, Zhaoyang Zhang, Han Zhang, Dong Li, Baining Chen, Lu Zhang, Jiuxiang Gu, and Dongdong Zhang. Unicontrol: A unified diffusion model for controllable visual generation in the wild. *arXiv preprint arXiv:2305.11147*, 2023. 3

[31] Zequn Qin and Xi Li. Monoground: Detecting monocular 3d objects from the ground. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3793–3802, 2022. 6, 7, 4

[32] Zhen Qiu, Yifan Zhang, Hongbin Lin, Shuaicheng Niu, Yanxia Liu, Qing Du, and Mingkui Tan. Source-free domain adaptation via avatar prototype generation and adaptation. *arXiv preprint arXiv:2106.15326*, 2021. 2

[33] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[34] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024. 3, 7

[35] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3, 5, 7, 1

[36] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical image computing and computer-assisted intervention–MICCAI 2015: 18th international conference, Munich, Germany, October 5-9, 2015, proceedings, part III 18*, pages 234–241. Springer, 2015. 3, 2

[37] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, et al. Imagen: Photorealistic text-to-image diffusion models. *arXiv preprint arXiv:2205.11487*, 2022. 3

[38] Jascha Sohl-Dickstein, Eric A Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. *Proceedings of the 32nd International Conference on Machine Learning*, 2015. 1

[39] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 5

[40] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2021. 1

[41] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochast ic differential equations. In *International Conference on Learning Representations*, 2021. 3

[42] Ziying Song, Guoxing Zhang, Lin Liu, Lei Yang, Shaoqing Xu, Caiyan Jia, Feiyang Jia, and Li Wang. Robofusion: towards robust multi-modal 3d object detection via sam. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, pages 1272–1280, 2024. 3

[43] Lin Sun, Liangliang Cao, Yang Zheng, Bo Wang, Zhiming Li, Wenzhe Xie, and Xu Sun. Panacea: A framework for diverse, controllable generation of urban scenes. *arXiv preprint arXiv:2207.10701*, 2022. 3

[44] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. *IEEE Robotics and Automation Letters*, 2024. 2

[45] Sergey Tulyakov, Ming-Yu Liu, Xiaodong Yang, and Jan Kautz. Mocogan: Decomposing motion and content for video generation. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018. 1

[46] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 2, 3, 4, 5

[47] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017. 3

[48] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. *Proceedings of the 25th International Conference on Machine Learning*, 2008. 1

[49] Yan Wang, Wei-Lun Chao, Divyansh Garg, and et. al. Pseudo-lidar from visual depth estimation: Bridging the gap in 3d object detection for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8445–8453, 2019. 2

[50] Hai Wu, Chenglu Wen, Shaoshuai Shi, Xin Li, and Cheng Wang. Virtual sparse convolution for multimodal 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21653–21662, 2023. 2

[51] Jinheng Xie, Yuexiang Li, Yawen Huang, Haozhe Liu, Wentian Zhang, Yefeng Zheng, and Mike Zheng Shou. Boxdiff: Text-to-image synthesis with training-free box-constrained diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7452–7461, 2023. 3

[52] Shaoyuan Xie, Lingdong Kong, Wenwei Zhang, Jiawei Ren, Liang Pan, Kai Chen, and Ziwei Liu. Benchmarking and improving bird's eye view perception robustness in autonomous driving. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 2, 6

[53] Junkai Xu, Liang Peng, Haoran Cheng, Hao Li, Wei Qian, Ke Li, Wenxiao Wang, and Deng Cai. Mononerd: Nerf-like representations for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 6814–6824, 2023. 2

[54] Longfei Yan, Pei Yan, Shengzhou Xiong, Xuanyu Xiang, and Yihua Tan. Monocd: Monocular 3d object detection with complementary depths. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10248–10257, 2024. 7

[55] Xu Yan, Weiming Zeng, Jiaming Zhu, Hailin Chen, Yaowei Li, and Deng Cai. Videogpt: Video generation using vq-vae and transformers. *arXiv preprint arXiv:2104.10157*, 2021. 1

[56] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17830–17839, 2023. 2

[57] Kairui Yang, Enhui Ma, Jibin Peng, Qing Guo, Di Lin, and Kaicheng Yu. Bevcontrol: Accurately controlling street-view elements with multi-perspective consistency via bev sketch layout. *arXiv preprint arXiv:2308.01661*, 2023. 2

[58] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11784–11793, 2021. 2

[59] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3, 4, 7

[60] Renrui Zhang, Han Qiu, Tai Wang, Ziyu Guo, Ziteng Cui, Yu Qiao, Hongsheng Li, and Peng Gao. Monodetr: Depth-guided transformer for monocular 3d object detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9155–9166, 2023. 2

[61] Yifan Zhang and Bryan Hooi. Hipa: enabling one-step text-to-image diffusion models via high-frequency-promoting adaptation. *arXiv preprint arXiv:2311.18158*, 2023. 3

[62] Yifan Zhang, Ying Wei, Qingyao Wu, Peilin Zhao, Shuaicheng Niu, Junzhou Huang, and Mingkui Tan. Collaborative unsupervised domain adaptation for medical image diagnosis. *IEEE Transactions on Image Processing*, 29: 7834–7844, 2020. 2

[63] Yunpeng Zhang, Jiwen Lu, and Jie Zhou. Objects are different: Flexible monocular 3d object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3289–3298, 2021. 1, 2, 6, 7, 3, 4

[64] Yifan Zhang, Bingyi Kang, Bryan Hooi, Shuicheng Yan, and Jiashi Feng. Deep long-tailed learning: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 45(9):10795–10816, 2023. 2

[65] Yifan Zhang, Daquan Zhou, Bryan Hooi, Kai Wang, and Jiashi Feng. Expanding small-scale datasets with guided imagination. In *Advances in neural information processing systems*, pages 76558–76618, 2023. 3

[66] Bowen Zhao, Feng Lu, Cheng Zhang, Jiaqi Hu, and Lin Song. Drivedreamer: Realistic multi-view video generation of driving scenes with future prediction. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 3

[67] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Dongdong Zhang, Lu Yuan, Han Zhang, Dong Li, Baining Chen, and Lu Zhang. Uni-controlnet: All-in-one control to text-to-image diffusion models. *arXiv preprint arXiv:2305.16322*, 2023. 3

[68] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22490–22499, 2023. 3

[69] Longtao Zheng, Yifan Zhang, Hanzhong Guo, Jiachun Pan, Zhenxiong Tan, Jiahao Lu, Chuanxin Tang, Bo An, and Shuicheng Yan. Memo: Memory-guided diffusion for expressive talking video generation. *arXiv preprint arXiv:2412.04448*, 2024. 3

[70] Bolei Zhou, Hang Zhao, Xavier Puig, Sanja Fidler, Adela Barriuso, and Antonio Torralba. Scene parsing through ade20k dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 633–641, 2017. 7