Preparation of Papers for IEEE Sponsored Conferences & Symposia*

Albert Author[1] and Bernard D. Researcher[2]

***Abstract*— Recent development in multi-view diffusion models have significantly enhanced the street image synthesis with 3D layout control. However, it is very tricky to generate similar street scenes images using only text prompt and object bounding boxes. Global features consistency can assistant training place recognition network. This work focuses on scene level to achieve nuanced scene prompt capability for the pretrained multi-view diffusion models, leveraging a visual place recognition network and Scene-Adapter. The key design of our Scene-Adapter is using linear projection and contrastive learning to map place-ID embedding from place recognition prior to CLIP image space. In addition, we introduce Light-Adapter, an approach offering light estimation prior to diffusion models, controllable creating a large dataset of images under different lighting targets with similar scene description distributions. With Place-ID-Adapter and Light-Adapter, we achieve high-fidelity street-view image synthesis that captures scene descriptions and various light condition, enhancing visual place recognition tasks. We comprehensively compare our approach with other existing methods, using both qualitative side-by-side comparisons and augmentation evaluations. The results show that our method achieves state-of-the-art performance in reconstruction and 3D object detection tasks. Remarkably, our method represents the first generative approach capable of enhancing the performance of place recognition models beyond what is possible with pre-trained models on the nuScenes dataset.**

I. Introduction

Generative models [**?**], [**?**], [**?**], [**?**] have made significant progress recently, excelling in producing high-quality, realistic visual contents. Diffusion models [**?**], [**?**], 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.

Recent development in multi-view diffusion models have significantly enhanced the street image synthesis with 3D geometry control. Early studies focused on creating street-view images to enhance image-based BEV perception methods, introducing solutions such as BEVGen [**?**], BEVControl [**?**], and MagicDrive [**?**]. More recent efforts have ventured into generating driving scene videos [**?**], [**?**], [**?**], aiming to bolster more. However, it is difficult to generate background-consistent (in other words, place-aware) street view images only using text, BEV maps and object bounding boxes, which are not informative to express background of a scene,

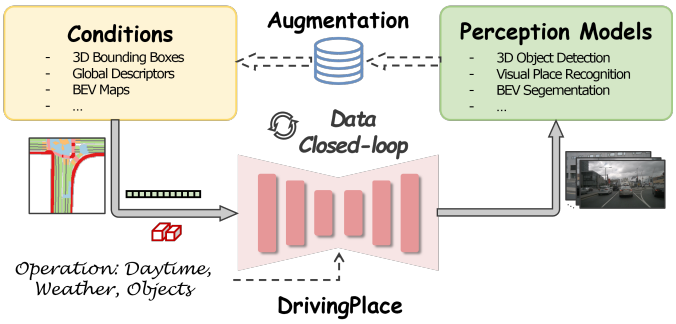

Fig. 1: Place encoder details.

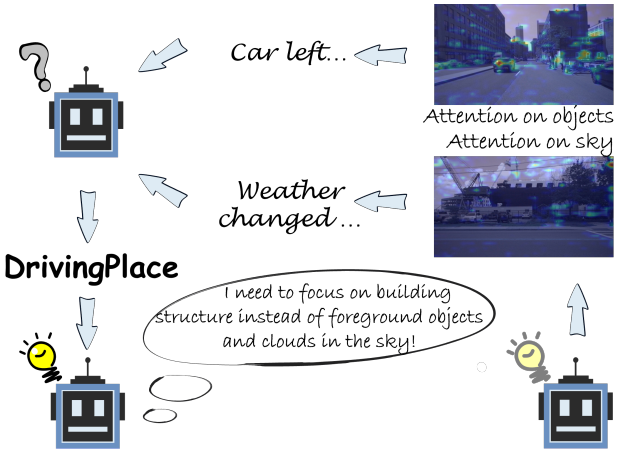

Fig. 2: Place encoder details.

which can be a hindrance to content creation. It means these methods are impossible to create synthetic samples nearly indistinguishable from real-exist places, impeding increase in using generative models for effective natural data synthesis in discriminative place recognition tasks [**?**], [**?**], [**?**]. Place recognition or loop closure in autonomous driving and robotics is one of the most crucial yet challenging discriminative tasks as object detection. Amidst the significant strides in generative models, a pressing question emerges: Could these innovations catalyze its advancement?

A naive idea to achieve that is fine-tuning the text-conditioned diffusion models directly on place-ID embedding to achieve place-aware prompt capabilities. However, the disadvantages of this approach are obvious. First, it eliminates the original ability to generate images using text, BEV maps and 3D boxes. Then, large computing resources are often required for such fine-tuning.

In order to solve the above problems, we have designed a novel task for existing multi-view diffusion. The goal of our

*This work was not supported by any organization

[1]Albert Author is with Faculty of Electrical Engineering, Mathematics and Computer Science, University of Twente, 7500 AE Enschede, The Netherlands albert.author@papercept.net

[2]Bernard D. Researcheris with the Department of Electrical Engineering, Wright State University, Dayton, OH 45435, USA b.d.researcher@ieee.org

mission is to generate background-controllable multi-view images in a place conditioning on an additional place-ID embeddings, easily integrating with original control remaining. Intuition of this is that, by training such a generative model, the synthesis images have a distribution close to the place-ID embedding, which can be used to assist in the training of place recognition algorithms. The learned place-ID controller can be treated as a distribution of learned "descriptions" of the reference place and should be able to model the commonalities and variations of visual attributes in the background.

In summary, we focuses on scene level to achieve nuanced scene-level prompt capability for the pre-trained multi-view diffusion models, leveraging a visual place recognition network and Place-ID controller. The key design of our Place-ID controller is using linear projection, perceiver transformer and contrastive learning to map place-ID embedding from place recognition prior to fixed CLIP text space. With Place-ID controller, we achieve high-fidelity street-view image synthesis that captures universal scene descriptions under various light and foreground objects simply achieved by changing the prompts, enhancing both visual place recognition and object detection tasks. We comprehensively compare our approach with other existing methods, using both qualitative side-by-side comparisons and augmentation evaluations. The results show that our method achieves better performance in generation quality and place recognition tasks.

II. Related Work

A. Latent Diffusion Models

Diffusion models represent a family of probabilistic generative models that progressively introduce noise to data and subsequently learn to reverse this process for the purpose of generating samples [**?**]. These models have recently garnered significant attention due to their exceptional performance in various applications, setting new benchmarks in image synthesis [**?**], [**?**], [**?**], [**?**], [**?**], video generation [**?**], [**?**], [**?**], [**?**], [**?**], [**?**], and 3D content generation [**?**], [**?**], [**?**], [**?**]. To enhance the controllable generation capability, ControlNet [**?**], GLIGEN [**?**], T2I-Adapter [**?**], and Composer [**?**] have been introduced to utilize various control inputs, including depth maps, segmentation maps, canny edges, and sketches.

The text-to-image (T2I) generation task aims at generating realistic images based on text inputs. Early works solve the problem by turning it into a sequence-to-sequence problem. For example, DALL-E [**?**] translates text tokens to discrete image embeddings obtained by VQ-VAE [**?**]. Subsequent works improve the quality of the generated images by using more advanced architectures such as image tokenizers [**?**], encoder-decoder architectures [**?**], or hierarchical transformers [**?**]. Recently, Denoising Diffusion Probabilistic Models (DDPM) [**?**] have begun to conquer the T2I task. The quality of the generated image is further increased by utilizing the excellent ability to generate realistic images of diffusion models [**?**], [**?**] or improving the text-image alignments [**?**] using powerful text encoders [**?**].

B. Visual Place Recognition

Visual Place Recognition (VPR) has evolved from traditional feature-based methods like SURF [7], and RootSIFT [25], to deep learning approaches. Early CNN-based models [4,26,38] achieved success, but recently Vision Transformers (ViTs) [22,27,33,34,46,49] have gained attention, offering improved performance through their ability to model long-range dependencies. Two-stage methods [21,34,46] combine global feature ranking with local feature re-ranking, but they struggle with viewpoint variations.

Visual Foundation Models (VFMs) like CLIP [39] and DINOv2 [37] have become popular, with models such as AnyLoc [27] using dense local features for zero-shot VPR. While these models excel in certain scenarios, they face challenges with large time gaps and environmental changes. Fine-tuned models like SALAD [22] improve performance, but at the cost of increased feature dimensionality and memory usage. Methods like SelaVPR [34] and CricaVPR [33] incorporate trainable adapters into ViT architectures to avoid catastrophic forgetting, though they still face computational challenges.

Overall, despite advances, existing VPR methods struggle with changing foregrounds and lighting conditions. This paper proposes generating richer place images with various foreground objects and weather conditions through diffusion model, enhancing the training of VPR algorithms to improve robustness in real-world scenarios.

C. Street View Image generation

This task can be seen as the reverse process of object detection or other visual perception, which generates images with the input of bounding boxes et. al. Previous works utilize GANs [**?**], [**?**], [**?**], [**?**] or diffusion models [**?**], [**?**] to generate synthetic images based on 2D layout. These methods encode the layout as a conditional image that is downsampled and upsampled jointly with the data. Recently, BEVGen [**?**] generates multi-view urban scene images from BEV layout based on VQ-VAE [**?**]. Concurrently, BEVControl [**?**], MagicDrive [**?**], and DrivingDiffusion [**?**] incorporate layout conditions to enhance image generation. The fundamental essence of diffusion-based generative models lies in their capacity to comprehend and understand the intricacies of the world. Harnessing the power of these diffusion models, DriveDreamer seeks to comprehend the complex realm of dynamic autonomous-driving scenarios. Compared with MagicDrive, we tackle a more complex task that generates place-aware multi-view images. In addition, the generated images were used as auxiliary datasets to significantly improve the place recognition effect.

III. Methods

While advanced generative methods can create high-quality images of driving scenes, their impact on downstream perception task such as place recognition (shown in Fig. 1), remains limited. We believe this is mainly due to the inadequate control over the generated background informations,

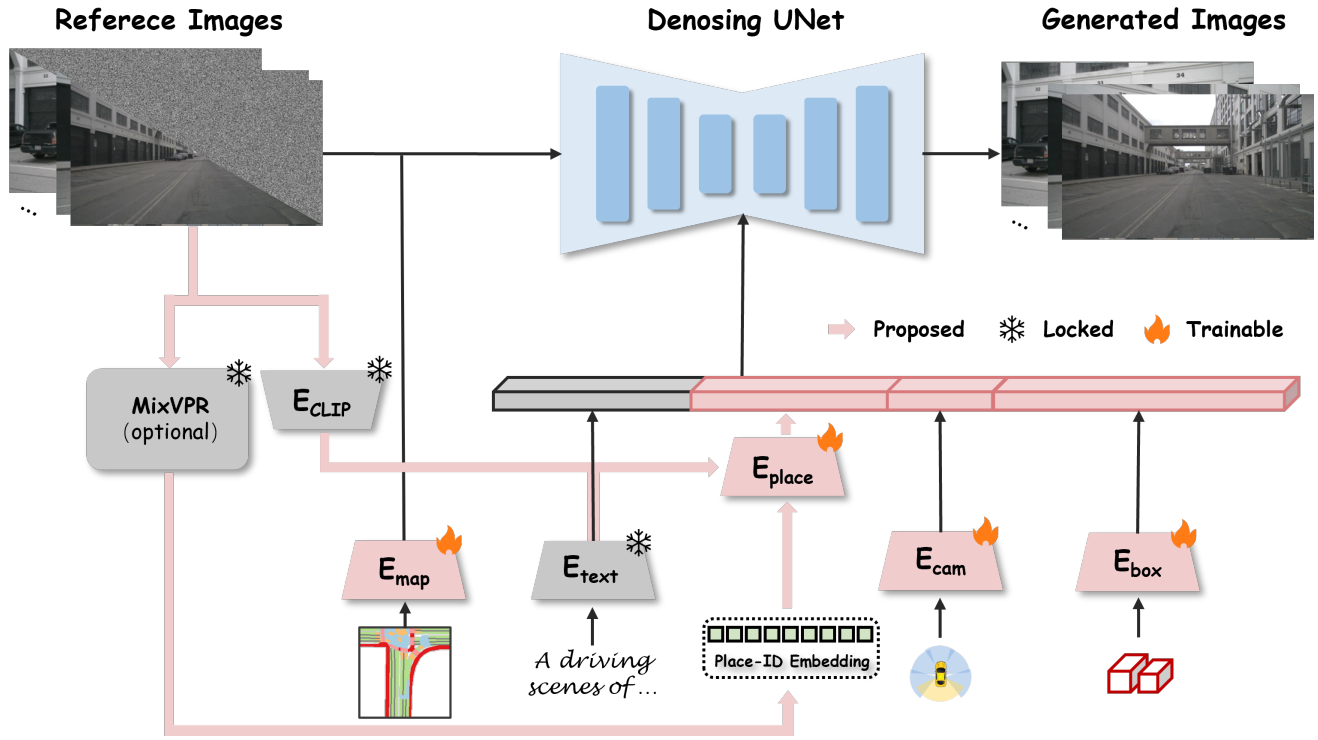

Fig. 3: The pipeline of our proposed RangePlace. Given the range images projected from point clouds, we first utilize Local Swin Transformer to capture the geometric information F_l of the range images at varying resolutions, emphasizing long-range dependencies to improve the distinction of local and global features. Following that, we build a Feature Pyramid Network to facilitate feature maps $F_l{}'$ at multiple scales. Then, we feed these feature maps to the Feature Mix module, where we generate multi-scale descriptors. In the end, we exploit a context gating mechanism to produce unified global descriptor for range image retrieval.

which are vital for effective scene understanding and place recognition.

As depicted in Figure 3, various strategies are implemented to inject information into the denoising UNet of multi-view diffusion models. We propose a Place-ID controller that maps place-ID embedding from the place recognition network to the aligned CLIP text space. Key components of this approach include place-ID encoding, a attribute perceiver transformer, and a contrastive learning strategy.

We begin by outlining the fundamental concepts of multi-view diffusion (Section 3.1). Next, we introduce our overall diffusion architecture (Section 3.2), which is generated through a place recognition network, incorporating specially designed tokens to enhance the diffusion models. Finally, we provide a detailed discussion on controlling the diffusion of place features.

A. Preliminary

Latent Diffusion models are designed to learn a probability distribution $p_\theta(x_0) = p_\theta(x_{0:T})\, dx_{1:T}$, where x_0 represents the data and $x_{1:T} := x_1, \ldots, x_T$ are latent variables. The joint distribution is characterized by a Markov chain (reverse process):

$$p_\theta(x_{0:T}) = p(x_T) \prod_{t=1}^{T} p_\theta(x_{t-1}|x_t),$$

with $p(x_T) = \mathcal{N}(x_T; 0, I)$ and $p_\theta(x_{t-1}|x_t) = \mathcal{N}(x_{t-1}; \mu_\theta(x_t, t), \sigma_t^2 I)$. Here, $\mu_\theta(x_t, t)$ is a trainable component, while the variance σ_t^2 consists of untrained time-dependent constants. The aim is to learn μ_θ for generation purposes.

To achieve this, a Markov chain known as the forward process is constructed:

$$q(x_{1:T}|x_0) = \prod_{t=1}^{T} q(x_t|x_{t-1}),$$

where

$$q(x_t|x_{t-1}) = \mathcal{N}(x_t; \sqrt{1-\beta_t}x_{t-1}, \beta_t I),$$

and β_t are constants. The DDPM approach demonstrates that by defining

$$\mu_\theta(x_t, t) = \frac{1}{\sqrt{\alpha_t}} x_t - \beta_t \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t),$$

with α_t and $\bar{\alpha}_t$ being constants derived from β_t and ϵ_θ functioning as a noise predictor, we can learn ϵ_θ by minimizing the following loss function:

$$\mathcal{L} = \mathbb{E}_{t,x_0,\epsilon} \left\| \epsilon - \epsilon_\theta \left(\sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, t \right) \right\|^2,$$

where ϵ is a random variable sampled from $\mathcal{N}(0, I)$.

To ensure viewpoint consistency, multi-view diffusion models [**?**], [**?**] usually adopt cross-view attention module. Considering the sparse camera arrangement in driving scenarios, each cross-view attention mechanism enables the target view to access information from its adjacent left and right views, as expressed in Equation. In this context, t, l, and r refer to the target, left, and right views, respectively. The target view then consolidates this information through a skip connection, as outlined in Equation , where h_v denotes the hidden state of the target view. The attention computation can be described as follows:

$$\text{Attention}_{cv}(Q_t, K_i, V_i) = \text{softmax}\left(\frac{Q_t K_i^T}{\sqrt{d}}\right) \cdot V_i, \quad i \in \{l, r\}$$

B. Overall Architecture

Our latent multi-view diffusion model comprises an VAE encoder E, a denoiser U-Net , and a VAE decoder D. Given the text, BEV maps and 3D geometric information akin to MagicDrive [**?**] doesn't ensure precise guidance for background generation. Thus, we introduce additional encoder for place-ID coined Eplace, given the original input-output pair, x (camera poses, BEV maps, scene captions, 3D bounding boxes). Following MagicDrive, we treat the coordinates of the LiDAR system as those of the ego vehicle and parameterize all geometric information accordingly. Let $S = \{MAP, BOX, TEXT, PLACE\_ID\}$ denote the representation of a driving scene surrounding the ego vehicle, where M is a binary map representing a $w \times h$ meter area of the road in Bird's Eye View (BEV) with c semantic classes. $BOX = \{(c_i, b_i)\}_{i=1}^{N}$ indicates the positions of 3D bounding boxes for each object within the scene. In scene level, $TEXT$ contains textual descriptions that provide basic context about the scene (e.g., weather conditions and time of day). Moreover, $PLACE\_ID$ provide detailed background information to recognize the place. Given a camera pose $P = [K, R, T]$ (which includes intrinsics, rotation, and translation), the objective of the generator $G(\cdot)$ is determined to synthesize multi-view consistent images with foreground object-level, mid map-level and background scene-level.

C. Place-ID Encoding

Place recognition networks are specifically designed for describing a place. We optionally utilize an existing visual place recognition method known as MixVPr, and adopt a implementation employs a ResNet50 backbone along with an all-MLP aggregation, and produce a discriminative place-ID embedding of dimension 4096 for retrieval purposes.

To ensure the alignment of the place-ID embedding generated by the place recognition network with other conditions, we implement two trainable linear projection layers, as inspired by IP-Adapter-Face-ID [**?**]. These layers operate on the place-ID embedding, producing a sequence of features with a length of N_S (set to 4 in our implementation), with the same dimensionality as other prompt embeddings. It is found that using two linear layers yields better performance than employing an MLP with successive blocks.

Unlike MagicDrive, we do not apply a masking mechanism to these place-ID embedding, as various scenes consistently yield a fixed-length place-ID embedding through aggregation, eliminating the need for random masking operations. Additionally, we opted not to incorporate a multi-view approach in place recognition, as we believe it could compromise place information. Instead, the camera parameters encoding and cross-view attention in U-Net present pre-trained diffusion model maintain the consistency across views when generating images.

D. Attribute Perceiver Transformer

We utilize a perceiver-based transformer to map the place-ID with assistance from reference images. Prior to feature extraction using the CLIP Image Encoder, we mask out the sky regions by [**?**], thereby forcing the network to focus on the street scene and avoiding interference from the sky and lighting conditions. This ensures that the attention mechanism concentrates solely on the relevant portions of the scene.

In our proposed Place-ID controller, the place-ID embeddings, after being processed through a linear projector, are enhanced by the CLIP image features from the BEV map via several cross-attention layers (we implement 3 layers in this study). Given the query place-ID embeddings Z and the CLIP features c_t, the output of the cross-attention mechanism can be represented by the following equation:

$$Z = \text{Attention}(Q, K, V) = \text{Softmax}\left(\frac{QK^T}{\sqrt{d}}\right) V$$

where $Q = ZW_q$ represents the query matrix of the place-ID embeddings, $K = c_t W_k$, and $V = c_t W_v$ denote the key and value matrices derived from the CLIP hidden states.

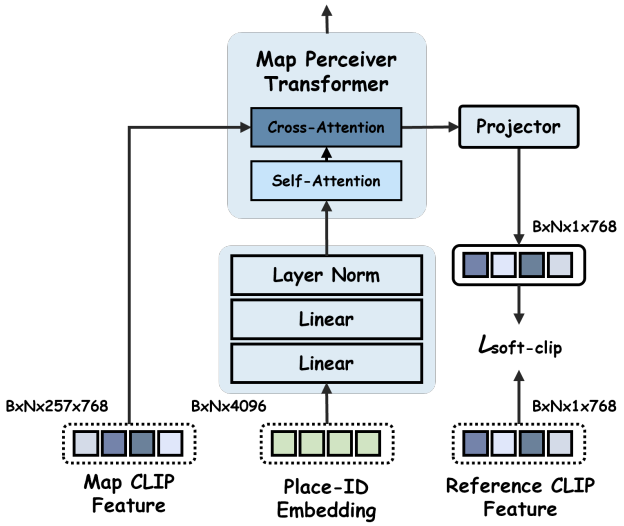

Fig. 4: Place encoder details.

E. Contrastive Learning

We also incorporated contrastive learning into the ControlNet training process, which helps mapp place-ID embeddings

into the CLIP text space, facilitating the alignment distribution of place-ID conditions with original text description.

Contrastive learning is a powerful approach for learning representations across different modalities by maximizing the cosine similarity of positive pairs while minimizing it for negative pairs. Previous studies have demonstrated the advantages of using contrastive learning in conjunction even with neural data [**?**]. An example of this is CLIP [**?**], a multimodal contrastive model that projects both images and text captions into a shared embedding space. In our approach, we apply contrastive learning to introduce additional place-ID embedding conditions while keeping the CLIP text space static.

Specifically, we use a projector to transform features with a shape of NxC to match the length of CLIP text tokens (after pooling layer in our implementation). We then normalize these tokens and compute the InfoNCE loss between them. It's important to note that this strategy is only applied during the training phase.

Our soft CLIP loss draws inspiration from knowledge distillation [], which posits that the softmax probability distribution generated by a strong teacher model serves as a more effective teaching signal for a student model compared to hard labels. To create the soft labels, we calculate the dot product of CLIP text embeddings within a batch. The soft CLIP loss is computed as our contrastive objective between the CLIP-CLIP and Place ID-CLIP matrices as follows:

$$\mathcal{L}_{\text{SoftCLIP}} = \sum_{i=1}^{N}\sum_{j=1}^{N} \left[\frac{\exp\left(\frac{t_i \cdot t_j}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{t_i \cdot t_m}{\tau}\right)} \cdot \log\left(\frac{\exp\left(\frac{p_i \cdot t_j}{\tau}\right)}{\sum_{m=1}^{N}\exp\left(\frac{p_i \cdot t_m}{\tau}\right)}\right)\right]$$

IV. Experiments

A. Experiments Setups

1) Dataset and Baselines: We present experiments on the nuScenes dataset to assess the efficiency of DrivingScene. It is a popular dataset in 3D object detection and place recognition for autonomous driving. We follow most of methods, utilizing 700 street-view scenes for training and 150 for validation. Our baseline is MagicDrive, which is recent SoTA propositions for street view generation.

We also present training support for place recognition experiment on Pitts30k-test [], which is collected from Google Street View, and comprises 8k queries and 8k references. Pittsburgh datasets show significant viewpoint and light changes. We use a MixVPR model trained on NuScenes train split to generate original Place-ID embedding.

2) Evaluation Metrics: We evaluate the realism of generated images using Fr´echet Inception Distance (FID). For object generation accuracy controlled by the input 3D bounding boxes, we evaluate through BEVFusion [**?**]. For place controllability evaluation, we follow the same evaluation metric of MixVPR, where the recall@1 and recall@5 are measured. The query image is determined to be successfully retrieved if at least one of the top-k retrieved reference images is located within d = 25 meters from the query one. We generate images aligned with the validation set annotations and use perception models pre-trained with real data to assess image quality and control accuracy.

To examine the support for training, synthesis images are generated based on the training set as data augmentation for 3D object detection and place recognition models.

3) Implementation Details: We use CLIP ViT-L/14 [**?**] as freeze image encoder to extract attribute feature in the place-ID controller. Our DrivingPlace utilizes pre-trained weights from MagicDrive to ease training costs, which is implemented with Stable Diffusion v1.5. During training, we only optimize the newly layers in ControlNet while keeping the original fixed. We gradient the model with a constant learning rate at 8e-5 and a linear warmup of 3000 iterations. The training process was carried out on 6 NVIDIA 3090 GPUs with delay gradient descent batch size set to 24. We adopt AdamW as optimizer and set decay at 0.01. We adopt the resolutions of 224×400 to reconcile discrepancies in down-stream tasks. For place recognition input, we resize the image to 320 x 320 resolution to meet the need of MixVPR. For image reconstructions, we use 20 denoising timesteps with UniPC [**?**] multi-step noise scheduler.

B. Qualitative Results

1) Realism and Controllability Validation. : To confirm the efficacy of DrivingScene in generating realistic images, we employ the nuScenes validation set to synthesize street-view images and show the metrics in Table 1. DrivingScene surpasses baseline method in image quality, delivering significantly lower FID scores. Despite the inherent domain gap between place-ID embedding and pre-trained models, our refinement process effectively mitigates this, yielding highly convincing results. Regarding controllability, while maintaining the same effect in the 3D object detection task, DrivingScene dramatic exceeds baseline results at 224×400 resolution in place recognition task. This is mainly attributed to the distinct encoding design which significantly boosts generation precision on background information.

Nonetheless, MixVPR's recognition results further corroborate the efficacy of our place-ID alignment method. As can be seen in table, our generated images in validation set can be better recognized by a pre-trained place recognition network, which indicates its scene level and background controllability.

2) Training Support: DrivingScene can produce augmented data with accurate annotation and place-ID controls, enhancing the training for perception and recognition tasks. For 3D objects detection and place recognition, we augment an equal number of images as in the original dataset, ensuring consistent training iterations and batch sizes for fair comparisons to the baseline. To optimize data augmentation, we randomly exclude 50% of bounding boxes in each generated scene. This demonstrates that our method maintains its

Fig. 5: Qualitative comparison with baselines. Both scenes are from nuScenes validation set. We highlight some areas with rectangles to ease comparison. Compared with MagicDrive, quality of background from DrivingPlace is much better.

TABLE I: **Comparison of generation fidelity with driving-view generation methods. Conditions for data synthesis are from nuScenes validation set. For each task, we test the corresponding models rained on the nuScenes training set.**

Method	FID↓	3D object detection		Place Recognition	
		mAP↑	NDS↑	AR@1↑	AR@5%↑
MagicDrive [**?**]	16.2	12.30	23.32	45.9	76.1
Ours[1]	13.4	13.5	22.87	67.6	85.4

TABLE II: **Training Support for 3D Object Detection.**

Data	mAP↑	NDS↑
w/o synthetic data	64.92	69.42
w/ MagicDrive	67.28	70.14
Ours	67.91	70.78

original control capabilities while supplementing place-aware control.

As shown in Table 3, DrivingScene provides a marginal improvement over BEVFusion in CAM+LiDAR (C+L) settings, similar to MagicDrive. It is important to note that in the CAM+LiDAR setting, BEVFusion utilizes both modalities for object detection, which requires the precise generation of long-distance information due to the incorporation of LiDAR data.

Moreover, as can be seen in Fig. 2, it generate a tree in the original area of clouds, and this visualizes the pre-trained place recognition network's unexpected attention on "clouds" when extracting place features. The similarity in other parts

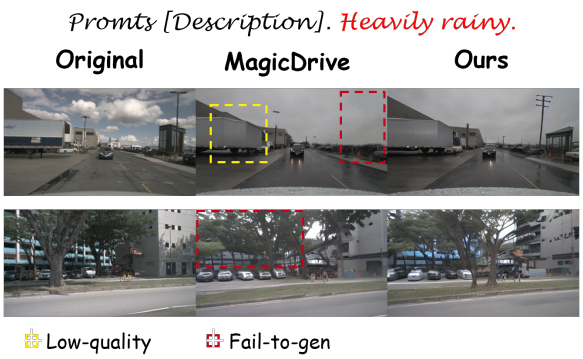

Fig. 6: Place controlablility under different weather prompts.

between the generated image and the reference image, along with the enhanced data diversity, ensures strong augmentation support for the training of place recognition network. Table 2 shows the advantageous impact of DrivingScene's data in place recognition tasks. It's crucial to note that in place recognition models requires more precise background information, highlighting the background controllability high fidelity.

TABLE III: **Training Support for place recognition.**

Data	AR@1↑	AR@5↑
w/o synthetic data	83.5	90.1
MagicDrive	84.2	91.4
Ours	**87.9**	**95.2**

C. Visualization Experiments

In addition, we utilize the Eud Distance visualization of the place-ID embedding generated by MixVPR on the synthetic validation set for visualization experiments. As shown in Fig. 5(a), (b) and (c), compared with MagicDrive, the Eud Distance results of our method shows a smaller gap, which indicates the images generated by our method can further generate images with aligned place feature with original ones.

D. Ablation Study

DrivingScene utilizes cross-attention to encode place-ID embedding. To demonstrate the efficacy, we train a model that directly takes the dimensional projection of place-ID embedding, denoted as "w/o cross-attention" in Table 4. Given the relative small size of place-ID embedding, it is necessary for a additional cross-attention to effectively help the model accurately represent background information, evidenced by the performance gap in recall rate. Applying BEV maps (denoted as "w/ BEV maps"), both the deviders and the xxx, improve controllability notably. This is due to that sementic BEV maps reduces the burden of attention localization.

Additionally, we evaluate the generation result of DrivingScene without contrastive loss. In this case, without external force to make place conditions close to the CLIP text space, the stability of training phase is reduced and it easily leads to bigger gap that reduce controllability.

TABLE IV: **Ablation study on different fusion approaches.**

Data	AR@1↑	AR@1%↑
w/o cross-attention	56.3	78.5
w/o contrastive learning	57.1	77.2
Ours	**67.6**	**85.4**

V. Conclusion

Recent development in multi-view diffusion models have significantly enhanced the street image synthesis with 3D layout control. However, it is very tricky to generate similar street scenes images using only text prompt and object bounding boxes. Global features consistency can assistant training place recognition network. This work focuses on scene level to achieve nuanced scene prompt capability for the pretrained multi-view diffusion models, leveraging a visual place recognition network and Scene-Adapter. The key design of our Scene-Adapter is using linear projection and contrastive learning to map place-ID embedding from place recognition prior to CLIP image space. In addition, we introduce Light-Adapter, an approach offering light estimation prior to diffusion models, controllable creating a large dataset of images under different lighting targets with similar scene description distributions. With Place-ID-Adapter and Light-Adapter, we achieve high-fidelity street-view image synthesis that captures scene descriptions and various light condition, enhancing visual place recognition tasks. We comprehensively compare our approach with other existing methods, using both qualitative side-by-side comparisons and augmentation evaluations. The results show that our method achieves state-of-the-art performance in reconstruction and 3D object detection tasks. Remarkably, our method represents the first generative approach capable of enhancing the performance of place recognition models beyond what is possible with pre-trained models on the nuScenes dataset.