SyntheOcc: Synthesize Geometric-Controlled Street View Images through 3D Semantic MPIs

Anonymous Author(s)
Affiliation
Address
email

Abstract

The advancement of autonomous driving is increasingly reliant on high-quality annotated datasets, especially in the task of 3D occupancy prediction, where the occupancy labels require dense 3D annotation with significant human effort. In this paper, we propose **SytheOcc**, which denotes a diffusion model that Synthesize photorealistic and geometric-controlled images by conditioning Occupancy labels in driving scenarios. This yields an unlimited amount of diverse, annotated, and controllable datasets for applications like training perception models and simulation. SyntheOcc addresses the critical challenge of how to efficiently encode 3D geometric information as conditional input to a 2D diffusion model. Our approach innovatively incorporates 3D semantic multi-plane images (MPIs) to provide comprehensive and spatially aligned 3D scene descriptions for conditioning. As a result, SyntheOcc can generate photorealistic multi-view images and videos that faithfully align with the given geometric labels (semantics in 3D voxel space). Extensive qualitative and quantitative evaluations of SyntheOcc on the nuScenes dataset prove its effectiveness in generating controllable occupancy datasets that serve as an effective data augmentation to perception models.

1 Introduction

With the rapid development of generative models, they have shown realistic image synthesis and diverse controllability. This progress has opened up new avenues for dataset generation in autonomous driving [5, 12, 23, 30]. The task of dataset generation is usually modeled as controllable image generation, where the ground truth (*e.g.* 3D Box) is employed to control the generation of new datasets in downstream tasks (*e.g.* 3D detection). This approach helps to mitigate the data collection and annotation effort as it can generate labeled data for free. However, a novel task of vital importance, occupancy prediction [24, 27], poses new challenges for dataset generation compared with 3D detection. It requires finer and more nuanced geometry controllability, which refers to use the occupancy state and semantics of voxels in the whole 3D space to control the image generation. We argue that solving this problem not only allows us to synthesize occupancy datasets, but also empowers valuable applications such as editing geometry to generate rare data for corner case evaluation, as shown in Fig. 1. In the following, we first illustrate why prior work struggles to achieve the above objective, and then demonstrate how we address these challenges.

In the area of diffusion models, several representative works have displayed high-quality image synthesis; however, they are constrained by limited 3D controllability: they are incapable of editing 3D voxels for precise control. For example, BEVGen [23] generates street view images by conditioning BEV layouts using diffusion models. MagicDrive [5] extend BEVGen and additionally converts the 3D box parameters into text embedding through Fourier mapping that is similar to NeRF [19], and uses cross-attention to learn conditional generation. Although these methods achieve satisfactory results in image generation, their 3D controllability is inherently limited. These approaches are

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.

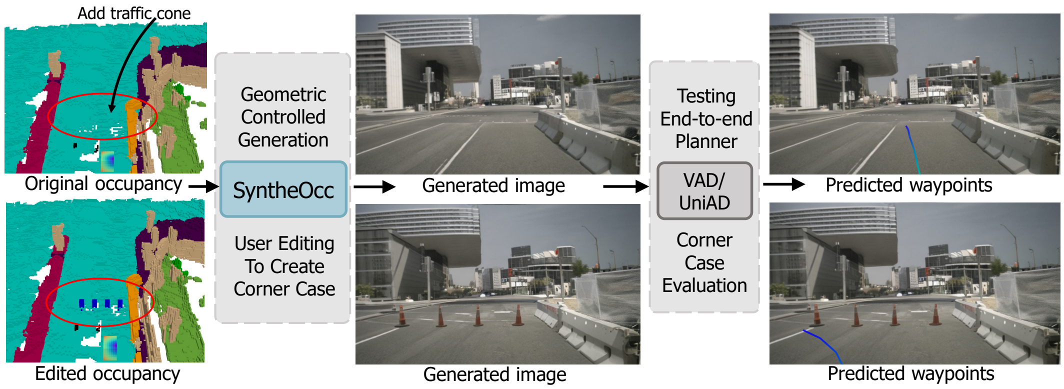


Figure 1: A showcase of application of **SytheOcc**. We enable geometric-controlled generation that conveys the user editing in 3D voxel space to generate realistic street view images. In this case, we create a rare scene that traffic cones block the way. This advancement facilitates the evaluation of autonomous systems, such as the end-to-end planner VAD [9], in simulated corner case scenes.

restricted to manipulating the scene in types of 3D boxes and BEV layouts, and hardly adapt to finer geometry control such as editing the shape of objects and scenes. Meanwhile, they usually convert conditional input into 1D embedding that aligns with prompt embedding, which is less effective in 3D-aware generation due to lack of spatial alignment with the generated images. This limitation hinders their utility in downstream applications, such as occupancy prediction and editing scene geometry to create long-tailed scenes, where granular volumetric control is paramount in both tasks.

ControlNet [41] and GLIGEN [14] is another type of prominent method in the field of controllable image generation. These approaches exhibit several desirable attributes in terms of controllability. They leverage conditional images such as semantic masks for control, thereby offering a unified framework to manipulate both foreground and background. However, despite its precise spatial control, ControlNet does not align with our specific requirements. Their conditions of pixel-level images differ fundamentally from what we require in 3D contexts. Our experimental results also find that ControlNet struggles to handle overlapping objects with varying depths (see Fig. 6 (a)), as it only utilizes an ambiguous 2D semantic map as conditional input. As a result, it is non-trivial to extend the ControlNet framework and convey their desirable attributes for 3D conditioning.

To address the above challenges, we propose an innovative representation, 3D semantic multi-plane images (MPIs), which contribute to image generation with finer geometric control. In detail, we employ multi-plane images [43] to represent the occupancy, where each plane represents a slice of semantic label at a specific depth. Our 3D semantic MPIs not only preserve accurate and authentic 3D information, but also keep pixel-wise alignment with the generated images. We additionally introduce the MPI encoder to encode features, and the reweighing methods to ease the training with long-tailed cases. As a collection, our framework enables 3D geometry and semantic control for image generation and further facilitates corner case evaluation as depicted in Fig. 1. Finally, experimental results demonstrate that our synthetic data achieve better recognizability, and are effective in improving the perception model on occupancy prediction. In summary, our contributions include:

- We present **SytheOcc**, a novel image generation framework to attain finer and precise 3D geometric control, thereby unlocking a spectrum of applications such as 3D editing, dataset generation, and long-tailed scene generation.
- Incorporating the proposed 3D semantic MPI, MPI encoder, and reweighing strategy, we deliver a substantial advancement in image quality and recognizability over prior works.
- Our extensive experimental results demonstrate that our synthetic data yields an effective data augmentation in the realm of 3D occupancy prediction.

2 Related Work

2.1 3D Occupancy Prediction

The task of 3D occupancy prediction aims to predict the occupancy status of each voxel in 3D space, as well as its semantic label if occupied. Compared with previous perception methods like 3D object detection, occupancy prediction offers a more detailed and nuanced understanding of the environment, as it provides finer geometric details, is capable of handling general, out-of-vocabulary objects, and finally, enriches the planning stack with comprehensive 3D information. Early methods exploited LiDAR as inputs to complete the 3D occupancy of the entire 3D scene [18, 33]. Recent methods began to explore the more challenging vision-based 3D occupancy prediction [24, 25, 27, 29]. By predicting the geometric and semantic properties of both dynamic and static elements, 3D occupancy prediction offers a more comprehensive understanding of the surrounding environment.

2.2 Diffusion-based Image Generation

Recent advancements in diffusion models (DMs) have achieved remarkable progress in image generation. In particular, Stable Diffusion (SD) [21] employs DMs within the latent space of autoencoders, striking a balance between computational efficiency and high image quality. Beyond text control, there is also the introduction of additional control signals. A noteworthy work is ControlNet [41], which incorporates a trainable copy of the SD encoder to extract the feature of conditional images and adds it to the UNet feature. It significantly enhances the controllability and unlocking pathways for advanced applications. We refer readers to recent survey [35] for more details.

2.3 Image Generation in Autonomous Driving

As training neural networks relies heavily on labeled data, numerous studies are delving into dataset generation to boost training. Lift3D [12] designs generative NeRF to synthesize labeled datasets for 3D detection for the first time. Several other works employ BEV layouts to synthesize image data, proving beneficial for perception models. For example, BEVGen [23] conditions BEV layouts to generate multi-view street images, while BEVControl [34] separately generates foregrounds and backgrounds from BEV layouts. MagicDrive [5] generates images with 3D geometry controls by independently encoding objects and maps through a text encoder or map encoder. Compared with MagicDrive, our geometry control is characterized by a more detailed and lossless representation of 3D scenes for control, which poses significant challenges than projected layout or box embedding.

Recently, DriveDreamer [26], DrivingDiffusion [13], Drive-WM [28] and Panacea [30] use a ControlNet framework, which involves projecting bounding boxes and road maps onto 2D FoV images as a conditioning input. This approach has proven to be effective for geometric control. However, it is limited in that it only achieves alignment at the 2D-pixel level. Consequently, this method falls short in capturing the depth hierarchy and fails to account for the occlusion relationships present in the 3D real world. Besides, adding a depth channel like Panacea [30] may address the limitations of depth order, but it discards the occluded part and only contains partial observation. UrbanGiraffe [37] train a generative NeRF to perform image generation. WoVoGen [17] creates a 4D world volume feature using occupancy to guide the generation, but seems to rely on object mask guidance.

As described above, most of the prior work is restricted by only modeling a projected primitive of 3D boxes and road maps as conditions. They suffer from ill-posed un-projection ambiguity. In contrast, we model 3D occupancy labels as conditions, as they provide finer geometric details and semantic information. However, designing an input representation of 3D occupancy labels into a 2D diffusion model is challenging. In this paper, we propose a novel representation: 3D semantic Multi-Plane Images (MPIs) as conditional inputs, which not only provide spatial alignment that improves visual consistency, but also encode comprehensive 3D geometric information including occluded parts.

3 Method

Overview The overview of our method is depicted in Fig. 2. Built upon the SD pipeline, we aim to perform geometry-controlled image generation by conditioning on 3D geometry labels with semantics (occupancy labels). One requirement is that the images should faithfully align with the given label. This task is more challenging than conditioned on 3D box due to the sparse and irregular nature of occupancy. We first discuss how to efficiently represent occupancy in Sec. 3.2, followed by our designed MPI encoder to enhance generation quality in Sec. 3.3, and reweighing strategy to handle the long-tailed depth and category in Sec. 3.5.

3.1 Representation of Condition: Local Control Aligns Better than Global Control

One of the key challenges is how to represent our conditional occupancy input. A straightforward method [3, 5] is to convert the 3D occupancy voxel to 1D global embedding that is similar to text embedding, and then use cross-attention to learn controllable generation. However, these global methods can be less effective when dealing with dense or irregular data due to the following reasons:

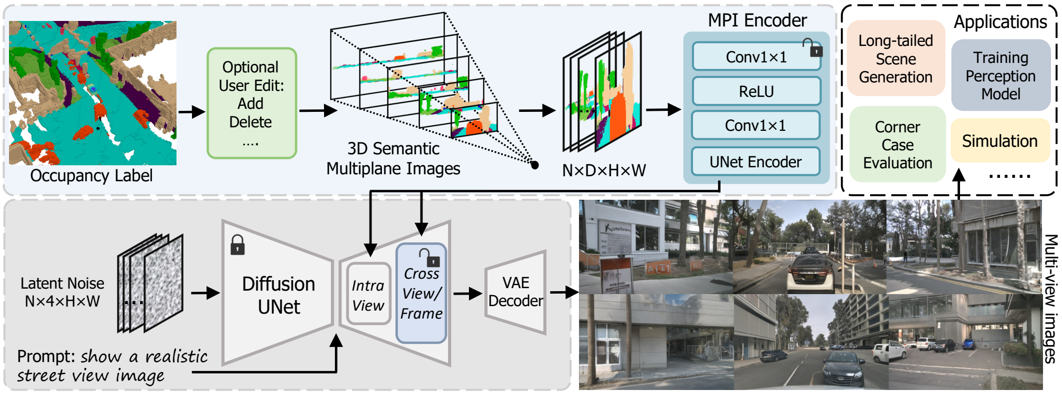


Figure 2: The overall architecture of **SytheOcc**. We achieve 3D geometric control in image generation by utilizing our proposed 3D semantic multiplane images to encode scene occupancy. In our framework, we can edit the occupied state and semantics of every voxel in 3D space to control the image generation, thereby opening up a wide spectrum of applications as shown in the top right.

(i) They perform controllable generation through hard encoding the spatial relationship between 1D global embedding and 2D UNet features. **(ii)** Ignore the underlying geometry alignment between the conditional input and the generated image. In contrast, local methods like ControlNet, directly add spatial features to the UNet features, providing 2D local control with pixel-level spatial alignment. They are better than the global method (see Tab. 1), but suffer from 3D ambiguity (see Fig. 6 (a)). Consequently, this comparison motivates us to seek a more compact and efficient manner to encode and condition our 3D occupancy labels.

3.2 Represent Occupancy as 3D Semantic Multiplane Images

It is non-trivial to design a 3D representation for conditioning. To efficiently store both the semantic and geometric information of the irregular occupancy input, we propose to use multiplane images (MPIs) [43] as representation. An MPI is composed of a series of fronto-parallel RGBA layers within the frustum of the source camera with a specific viewpoint. These planes are arranged at varying depths, from d_{min} to d_{max}, starting from the nearest to the farthest. Each layer of these images contains both an RGB image and an alpha map, which collectively capture the visual and geometric details of the scene at the respective depth. In our work, instead of storing RGB value and alpha map in the original MPI, we store our 3D semantic labels. Each layer of MPI represents the semantic index at the corresponding depth. We display the colored MPI in the top row of Fig. 2 for visual clarity, but we actually use the integer index for learning. We obtain our 3D semantic MPI by:

$$P_l = (u \times d_l,\ v \times d_l,\ d_l)^T,\ d_l = d_{min} + (d_{max} - d_{min}) \times l/D, \tag{1}$$

$$\texttt{MPI}_{n,l} = \texttt{Interpolate}(\texttt{Occupancy},\ \mathbf{T_n} \cdot \mathbf{K_n^{-1}} \cdot P_l), \tag{2}$$

$$\texttt{MPI} = \texttt{Concatenate}(\texttt{MPI}_{i,j}),\ i \in (0, N),\ j \in (0, D), \tag{3}$$

where (u, v) is a pixel coordinate in image space, d_l is depth value of the l^{th} layer, n denotes the n^{th} camera view. This equation implies we first back project points P in camera frustum space (u, v, d) to Euclid space (x, y, z) by multiplying inverse intrinsic $\mathbf{K^{-1}}$. Then we use transformation matrix $\mathbf{T}$ to map points from camera coordinates to occupancy coordinates. We then use the point coordinates to interpolate the nearest semantic index from the dense occupancy voxel to form a slice of MPI. Finally, we concatenate all slices to form $\texttt{MPI} \in \mathbb{R}^{N \times D \times H \times W}$, where D is the number of layers that is set at 256, N is the number of camera views in the case of batch size = 1.

By representing occupancy as 3D semantic MPI, every pixel in MPI contains geometry and semantic information with implicit depth, seamlessly integrating occluded elements, and ensuring a precise spatial alignment with the generated images.

3.3 3D Semantic MPI Encoder

To enable local control with spatially aligned conditions, we develop a simple but effective MPI encoder that aligns the 3D multi-plane feature to the latent space of the diffusion model. The purpose of the MPI encoder is to obtain features from multi-plane images to perform 3D-aware

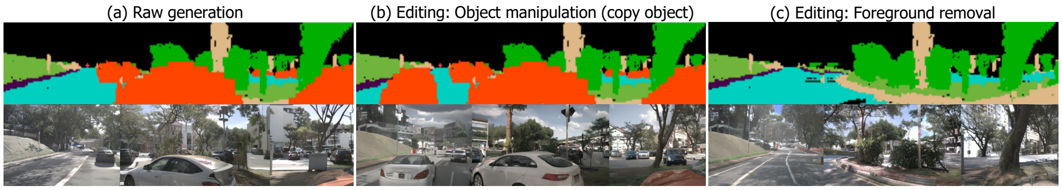


Figure 3: Visualizations of geometric controlled generation. **Top row**: Fusion of 3D semantic MPI. **Bottom row**: our generation concatenated from neighboring views.

image synthesis. Unlike the original ControlNet which downsampling conditional input through 3×3 convolutions with padding, we design a 1×1 convolutional encoder without downsampling to encode features. In detail, the 3D multiplane features which have the sample resolution with latent features, are transformed by a 1×1 convolution layer and ReLU activation [1] in the MPI encoder.

After obtaining the multi-scale feature after the MPI encoder, we add the feature to the decoder of diffusion UNet to provide spatial features. Experimental results in Tab. 3 will show that our 1×1 conv in MPI encoder is more effective than 3×3 conv, as the 1×1 conv with receptive field = 1 provides a spatial align feature to the latent feature in the diffusion UNet. In contrast, 3×3 conv is conducted in a camera frustum space rather than Euclid space, making an imprecise correspondence between 3D multiplane features and 2D image features. Moreover, using 3×3 conv to process 3D semantic MPI will introduce a large computational burden as the channel number increases from 3 channels of RGB to 256 planes. We display our 3D geometry and semantic control property in Fig. 3.

In summary, we chose MPIs as the representation because they **(i)** Incorporate lossless 3D information, including scene geometry rather than 2.5D depth. **(ii)** Provide spatially aligned conditional features that naturally extend the ControlNet framework from image level to 3D level. **(iii)** Capable of representing geometry and semantics including occluded elements.

3.4 Cross-View and Cross-Frame Attention

The sensor arrangement in a self-driving car usually requires a full surround view of cameras to capture the entire 360-degree environment. To effectively simulate the multi-view and subsequent multi-frame generation, zero-initialized [41] cross-view and cross-frame attention are integrated into the diffusion model to maintain consistency between views and frames. Following prior work [5, 28, 30, 31], each cross-view attention allows the target view to access information from its neighboring left and right views, thus training cross-view attention using multi-view consistent images will enforce it to generate the same instance in the overlapping region of multi-view cameras.

$$\texttt{Attention}(Q, K, V) = \texttt{softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V, \tag{4}$$

$$h_{out} = h_{in} + \textstyle\sum_{i \in \{l,r\}} \texttt{Attention}(Q_{in}, K_i, V_i), \tag{5}$$

where l, and r is the camera view of left and right. Q_{in} and h_{in} denotes the query and the hidden state of input view. Similarly, we add cross-frame attention that attend previous frame and future frame to enable video generation. In this case, we use the same formulation while $i \in \{f, h\}$, where f and h is the camera view of future and history frames.

3.5 Importance Reweighing

To deal with the extreme imbalance problem between foreground, background, and object categories, and also to ease the training, we propose three types of reweighting methods to improve the generation quality of foreground objects.

Progressive Foreground Enhancement To mitigate the complexity of the learning task, we propose a progressive reweighting method that incrementally enhances the loss associated with the foreground regions (based on semantic class) as the training progresses. The detailed formulation is:

$$w(x, m, n) = \frac{(m-1)}{2} \cdot (1 + \cos(\frac{x}{n} \cdot \pi + \pi)) + 1, \tag{6}$$

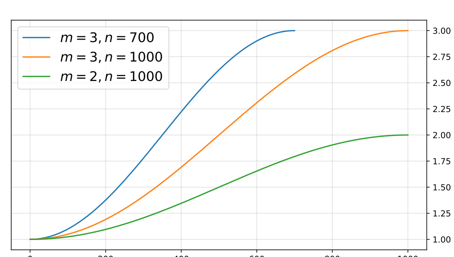
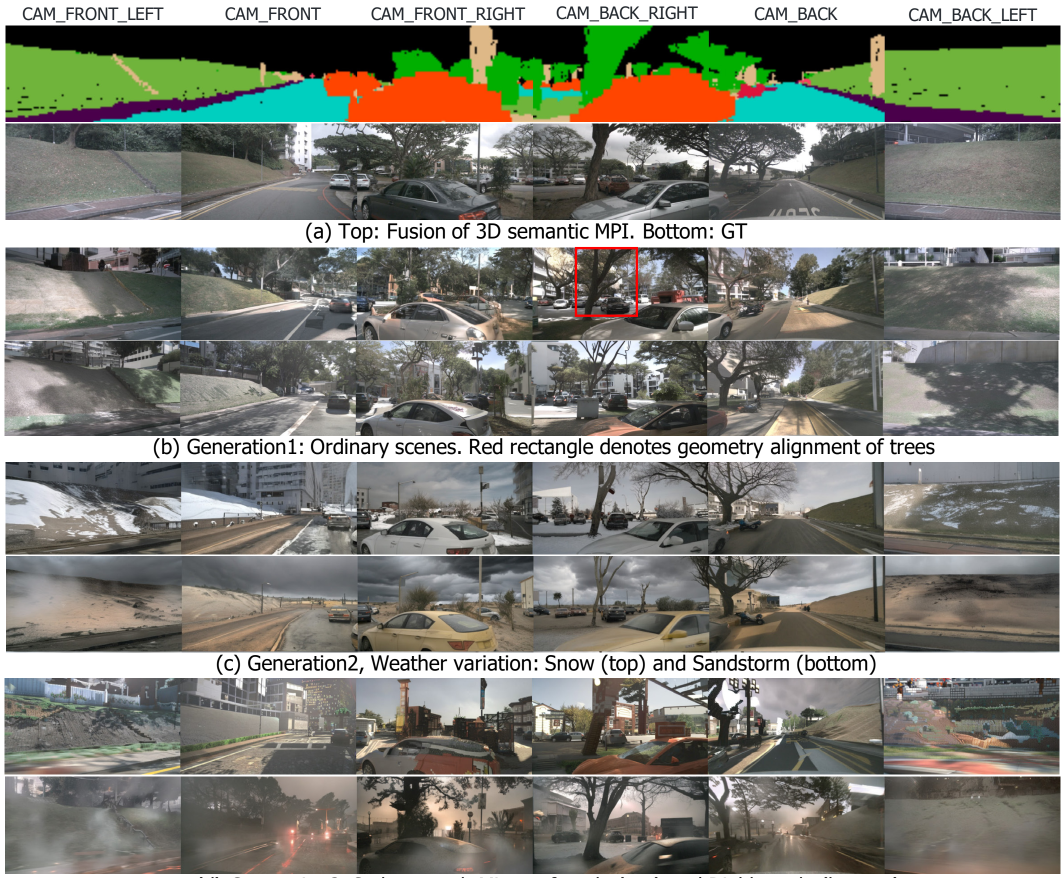


Figure 4: Visualizations of the reweighing function in Eq. 6.



(d) Generation3, Style control: Minecraft style (top) and Diablo style (bottom)

Figure 5: Visualizations of generated multi-view images. The generation conditions (occupancy labels) are from nuScenes validation set. We highlight that **(i)** Geometry alignment of trees in red rectangle in (b). **(ii)** Use text prompt to control high-level appearance in (c,d).

where x is the current training step, m is the maximum value of weights that set at 2, and n is the total training steps. This approach is engineered to facilitate a learning trajectory that progresses from simplicity to complexity, thereby aiding in the convergence of the model. This curve can be interpreted as a cosine annealing but inverted to amplify the importance of the foreground region.

Depth-aware Foreground Reweighing In the meantime, we acknowledge the learning difficulty in different depth places in 3D scenes. Following GeoDiffusion [3], we perform depth reweighing to foreground objects by adaptively assigning higher weights to farther foreground areas. This enables the model to focus more thoroughly on hard examples with depth-aware importance reweighting. Instead of using their exponential function to increase weights, we use our designed cosine function Eq. 6 for stability. Here x is the input depth value, and n is the maximum depth that set at 50.

CBGS Sampling To deal with the class imbalance problem in driving scenarios, where certain object categories appear infrequently, we employ the Class-Balanced Grouping and Sampling (CBGS) [44] to better handle the long-tailed classes. CBGS addresses the challenge of class imbalance by grouping and re-sampling training data to ensure each group has a balanced distribution of sample frequency across different object categories. This method reduces the bias towards more frequent classes and enables better generalization to rare scenarios.

3.6 Model Training

To ease the training of the MPI encoder and added attention module, we use a two stage training pipeline. We first train MPI encoder and cross-view attention in a multi-view image generation setting. Then we train cross-frame attention and freeze other components in a video generation setting.

Method	Train	Val	mIoU	barrier	bicycle	bus	car	cons. veh.	moto.	pedes.	traf. cone	trailer	truck	drive. suf.	other flat	sidewalk	terrain	manmade	vegetation
Oracle (FB-Occ [15])	Real	Real	39.3	45.4	28.2	44.1	49.4	25.9	28.8	28.0	27.7	32.4	37.3	80.4	42.2	49.9	55.2	42.0	37.7
SytheOcc-Aug	Real+Gen	Real	40.3	45.4	27.2	46.6	49.5	26.4	27.8	28.4	29.4	34.0	37.2	81.3	46.0	52.4	56.5	43.3	38.9
MagicDrive	Real	Gen	13.4	0.7	0.0	11.8	32.4	0.0	6.6	2.8	0.3	2.6	19.6	60.1	12.1	26.2	23.4	15.5	12.8
ControlNet	Real	Gen	17.3	17.7	0.2	13.6	21.0	0.6	0.8	8.6	10.4	6.9	11.9	67.4	18.8	36.4	36.9	20.8	22.4
ControlNet+depth	Real	Gen	17.5	19.3	0.3	14.0	23.7	1.0	0.6	9.2	9.2	5.7	12.1	68.8	19.2	36.0	35.3	19.8	22.8
SytheOcc-Gen	Real	Gen	25.5	32.6	13.8	27.7	33.4	7.5	6.5	15.7	16.5	16.5	25.6	74.3	24.5	39.4	40.5	28.6	28.8

Table 1: Downstream evaluation on the **nuScenes-Occupancy** validation set. Based on the used train and val data, two types of settings are reported. The first is to use generated training set to augment the real training set, and evaluate on the real validation set, denoted as Aug. The second is to use pretrained models trained on the real training datasets to test on the generated validation set, denoted as Gen.

Objective Function Our final objective function can be formulated as a standard denoising objective with reweighing:

$$\mathcal{L} = \mathbb{E}_{\mathcal{E}(x),\epsilon,t}\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2 \odot w, \tag{7}$$

where w is the multiplication of progressive reweighing and depth-aware reweighing.

4 Experiments

4.1 Dataset and Setups

We conduct our experiments on the nuScenes dataset [2], which is collected using 6 surrounded-view cameras that cover the full 360° field of view around the ego-vehicle. It contains 700 scenes for training and 150 scenes for validation. We resize the original image from 1600 × 900 to 800 × 448 for training. In our work, we use the occupancy label with a resolution of $0.2m$ from OpenOccupancy [27] as condition input, while the benchmark of occupancy prediction uses a resolution of $0.4m$ from Occ3D [24] dataset for its popularity.

Networks We use Stable Diffusion [21] v2.1 checkpoint as initialization and only train occupancy encoder, cross-view attention. We additionally add cross-frame attention if in video experiments. We adopt FB-Occ [15] as the target model for occupancy prediction for its SOTA performance in this task. The pretrained checkpoint of the network is obtained from their official repository. Since FB-Occ predicts occupancy using only single frame images, we thus train SyntheOcc without cross-frame attention in related experiments. For video generation, we provide experimental results in appendix.

Metrics We use Frechet Inception Distance (FID) [6] to measure the perceptual quality of generated images, and use mIoU to measure the precision of occupancy prediction.

Hyperparameters We set $D = 256$, $d_{min} = 0$ and $d_{max} = 50$. The depth resolution of MPI is thus higher than occupancy voxel. We train our model in 6 epochs with batch size $= 8$. The learning rate is set at $2e^{-5}$. The training phase takes around 1 day using 8 NVIDIA A100 80G GPUs. We use UniPC scheduler [42] with the classifier-free guidance (CFG) [7] that is set as 7.0. During inference, we use 20 denoising steps for dataset generation.

Baselines We compare our method with prior methods in Tab. 1. ControlNet denotes we train a ControlNet using an RGB semantic mask as the condition. ControlNet+depth denotes we add a depth channel after the semantic mask to provide 2.5D depth information. The depth map rendered by occupancy is normalized to [0-255] to accommodate the RGB value. The ControlNet+depth can be regarded as a degradation of SytheOcc which is reduced to a single plane. Then we evaluate MagicDrive since it is the only open-sourced method in this area. MagicDrive separately encodes foreground and background using prompt and BEV layout. Furthermore, we evaluate the image quality (FID [6]) of our method in Tab. 2. Compared with prior methods, we use a unified 3D representation that seamlessly handles foreground and background, surpassing them by a large margin.

4.2 Qualitative Results

High-level Control using Prompt In Fig. 5 (c,d) and Fig. 6 (c), we demonstrate the capability to employ user-defined prompts to generate images with specific weather conditions and high-level

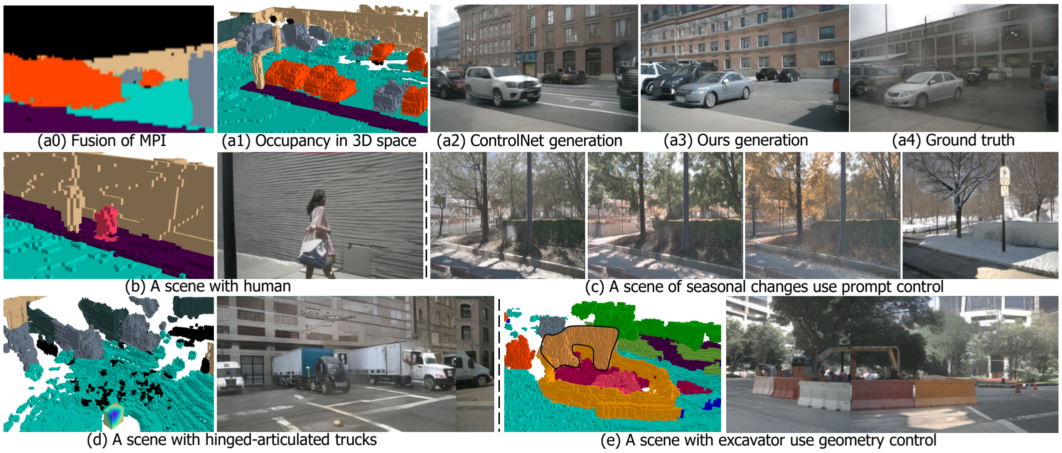


Figure 6: **Top row**: Comparison with ControlNet. We achieve a precise alignment between conditional labels and synthesized images, while ControlNet generates objects with incorrect pose due to ambiguous 2D condition. **Mid and Bottom row**: Visualizations of geometry-controlled image generation. We can faithfully generate objects with the desired topology in a specific 3D position.

style. Although the nuScenes dataset doesn't contain rare weather images like snow and sandstorms, our method successfully conveys prior knowledge pretrained from stable diffusion to our scenes. Compared with visualization results in prior work like Fig. 8 of MagicDrive, our method shows better alignment with the text prompt, demonstrating the cross-domain generalization ability of our method.

3D Geometric Control Our flexible framework enables us to create novel scenes by manipulating voxels as displayed in Fig. 1 and Fig. 3. Basically, we can edit the occupied state and semantics of every voxel in our scenes for generation. We highlight that we can create a hinged-articulated truck and an excavator as shown in Fig. 6 (d,e). The generated excavator image exhibits a remarkable alignment with the input occupancy that is delineated by a black outline.

Long-tailed Scene Generation The flexibility of 3D semantic MPI has conferred significant advantages upon our approach. In the following, we create long-tail scenes that rarely occur in our real world for evaluation. In Fig. 1, we show that we manually add parallel traffic cones in front of the ego vehicle. This scene has never happened in the training dataset, but our geometric controllability provides us the capability to create such data. We then use the created scene to test autonomous driving systems such as end-to-end planner VAD [9] to validate its effectiveness. In this case, VAD successfully predicts correct waypoints with the high-level command 'turn left'. Moreover, in appendix Sec. B, we generate long-tailed scenes with extreme weather such as snow and sandstorms, and evaluate perception model on it to examine its generalizability of rare weather.

Comparison with Baselines In Fig. 6 (a), we visualize a comparison with ControlNet. We find that ControlNet struggles to distinguish the overlapping instances in 2D-pixel space. This leads to the two parked cars being merged into a single car with incorrect pose. In contrast, our 3D semantic MPIs contain more than 2D semantic mask, but also account for complete scene geometry with occluded

Method	Condition Type	FID
BEVGen [23]	BEV map	25.54
BEVControl [34]	BEV map	24.85
DriveDreamer [26]	Box + FoV map	52.60
MagicDrive [5]	Box + BEV map	16.20
Panacea [30]	Box + FoV map	16.96
Ours	3D Semantic MPI	**14.75**

Table 2: Comparison of FID with previous methods on the nuScenes dataset.

MPI Encoder	**Reweighing Method**			**Metric**
design	Progressive	Depth	CBGS	mIoU
3×3	-	-	-	21.96
1×1	-	-	-	23.05
1×1	✓	-	-	23.63
1×1	✓	✓	-	24.40
1×1	✓	✓	✓	25.50

Table 3: Ablation of different designs of the MPI encoder and reweighing methods.

parts. Together with our proposed MPI encoder and reweighing strategy, our framework yields a realistic image generation with high-quality label alignment. More comparison is provided in Sec. D.

4.3 Quantitative Results

Recognizability, Realism and Controllability Evaluation To evaluate whether our generated images aligned with given annotations, we provide Gen experiment in Tab. 1. Using the annotation of val set, we synthesize a copy of val set's images, then use perception model trained on real training set to perform evaluation. The performance will be more effective as it is close to the oracle performance. We find that local method (ControlNet) perform better than global method (MagicDrive). Furthermore, SytheOcc generalizes the locality for 3D conditioning to yield better performance.

Data Augmentation for 3D Occupancy Prediction Notably, we conduct experiments using our synthesized dataset to enhance the real training set in Tab. 1. We first use the occupancy labels from training set to create a synthetic training set. Then we modify the loading pipeline in perception model to randomly sample images from real dataset or synthetic dataset and train network from scratch. Therefore, our approach preserves the inherent training dynamics of the neural network by solely modifying the training images, without any alteration to the number of training iterations or epochs. As MagicDrive-Aug exhibits numerical overflow when training FB-Occ, which may attributed to unsatisfactory recognizability, we have to omit it and only provide MagicDrive-Gen experiments.

As shown in Tab. 1, where SytheOcc-Aug denotes the augmentation experiments using our generated dataset, shows a satisfactory improvement over the prior state of the art. We emphasize that surpassing the performance of the original dataset is not the primary objective of our work; rather, it is an ancillary benefit that emerges from our framework for geometry-controlled generation.

Ablations In Tab. 3, we present ablation studies across several design spaces of our model, analogous to the Gen experiment in Tab. 1. We find that our designed MPI encoder of 1×1 conv have significant improvement when compared to the conventional 3×3 conv approach. Besides, our proposed three types of reweighing methods demonstrate a consistent improvement over the baseline. As a result, the improved image quality and label alignment enable higher precision in downstream tasks.

5 Limitation and Broader Impacts

Layout Genereation Our method is restricted in a conditional generation framework that should have a conditional input at first. Our condition signal is from the original dataset annotation. Thus most of the augmented data is generated using the same occupancy layout, or with minimal human editing. Future research can incorporate the recent research [10, 17, 32, 40] that generates occupancy descriptions of the scenes to synthesize images with novel occupancy layouts.

Closed-loop Simulation Given the underlying diverse and controllable image generation of our method, it would be advantageous and valuable to extend our work to a broader domain such as closed-loop simulation [16, 38], to enable high-fidelity autonomous systems testing. This line of work can be conducted by utilizing motion conditions to generate future frames as in world model [17, 28, 36], or by explicitly modeling scene graph as in the case of UniSim [20, 38] and NeuroNCAP [16].

Long-tailed Scene Generation In this paper, we only investigate a limited number of long-tailed scene generation and corner case evaluations such as rare layout in Fig. 1 and extreme weather in Sec. B. Future work can extend our framework to **(i)** Synthesize more samples for tail classes to boost performance. **(ii)** Generate or replicate large-scale databases of corner cases [11] for robust perception.

6 Conclusion

In this paper, we propose **SytheOcc**, an innovative image generation framework that is empowered with geometry-controlled capabilities using occupancy. We introduce a novel 3D representation, 3D semantic MPIs, to address the critical challenge of how to efficiently encode occupancy. This representation not only preserves the authentic and complete 3D geometry details with semantics, but also provides a spatial-align feature representation for 2D diffusion models. With this property, our method enjoys photorealistic appearances and fine-grained 3D controllability, serves as a generative data engine to enable a broad range of applications. Extensive experiments demonstrate that our synthetic data facilitate the training for perception models on occupancy prediction, and provide valuable corner case evaluation in a simulated world.

References

[1] Abien Fred Agarap. Deep learning using rectified linear units (relu). arXiv preprint arXiv:1803.08375, 2018. 5

[2] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In CVPR, 2020. 7

[3] Kai Chen, Enze Xie, Zhe Chen, Lanqing Hong, Zhenguo Li, and Dit-Yan Yeung. Integrating geometric control into text-to-image diffusion models for high-quality detection data generation via text prompt. arXiv preprint arXiv:2306.04607, 2023. 3, 6

[4] Jaeyoung Chung, Suyoung Lee, Hyeongjin Nam, Jaerin Lee, and Kyoung Mu Lee. Luciddreamer: Domain-free generation of 3d gaussian splatting scenes. arXiv preprint arXiv:2311.13384, 2023. 12

[5] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. Magicdrive: Street view generation with diverse 3d geometry control. In ICLR, 2024. 1, 3, 5, 8, 14

[6] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. NeurIPS, 2017. 7

[7] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. arXiv preprint:2207.12598, 2022. 7

[8] Lukas Höllein, Ang Cao, Andrew Owens, Justin Johnson, and Matthias Nießner. Text2room: Extracting textured 3d meshes from 2d text-to-image models. In ICCV, 2023. 12

[9] Bo Jiang, Shaoyu Chen, Qing Xu, Bencheng Liao, Jiajie Chen, Helong Zhou, Qian Zhang, Wenyu Liu, Chang Huang, and Xinggang Wang. Vad: Vectorized scene representation for efficient autonomous driving. In ICCV, 2023. 2, 8, 12, 13

[10] Jumin Lee, Sebin Lee, Changho Jo, Woobin Im, Juhyeong Seon, and Sung-Eui Yoon. Semcity: Semantic scene generation with triplane diffusion. arXiv preprint arXiv:2403.07773, 2024. 9

[11] Kaican Li, Kai Chen, Haoyu Wang, Lanqing Hong, Chaoqiang Ye, Jianhua Han, Yukuai Chen, Wei Zhang, Chunjing Xu, Dit-Yan Yeung, et al. Coda: A real-world road corner case dataset for object detection in autonomous driving. In ECCV, 2022. 9

[12] Leheng Li, Qing Lian, Luozhou Wang, Ningning Ma, and Ying-Cong Chen. Lift3d: Synthesize 3d training data by lifting 2d gan to 3d generative radiance field. In CVPR, 2023. 1, 3

[13] Xiaofan Li, Yifu Zhang, and Xiaoqing Ye. Drivingdiffusion: Layout-guided multi-view driving scene video generation with latent diffusion model. arXiv preprint arXiv:2310.07771, 2023. 3

[14] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. Gligen: Open-set grounded text-to-image generation. In CVPR, 2023. 2

[15] Zhiqi Li, Zhiding Yu, David Austin, Mingsheng Fang, Shiyi Lan, Jan Kautz, and Jose M Alvarez. Fb-occ: 3d occupancy prediction based on forward-backward view transformation. arXiv preprint arXiv:2307.01492, 2023. 7

[16] William Ljungbergh, Adam Tonderski, Joakim Johnander, Holger Caesar, Kalle Åström, Michael Felsberg, and Christoffer Petersson. Neuroncap: Photorealistic closed-loop safety testing for autonomous driving. arXiv preprint arXiv:2404.07762, 2024. 9

[17] Jiachen Lu, Ze Huang, Jiahui Zhang, Zeyu Yang, and Li Zhang. Wovogen: World volume-aware diffusion for controllable multi-camera driving scene generation. arXiv preprint arXiv:2312.02934, 2023. 3, 9

[18] Jianbiao Mei, Yu Yang, Mengmeng Wang, Tianxin Huang, Xuemeng Yang, and Yong Liu. Ssc-rs: Elevate lidar semantic scene completion with representation separation and bev fusion. In IROS, 2023. 3

[19] Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. In ECCV, 2020. 1

[20] Julian Ost, Fahim Mannan, Nils Thuerey, Julian Knodt, and Felix Heide. Neural scene graphs for dynamic scenes. In CVPR, 2021. 9

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In CVPR, 2022. 3, 7

[22] Liangchen Song, Liangliang Cao, Hongyu Xu, Kai Kang, Feng Tang, Junsong Yuan, and Yang Zhao. Roomdreamer: Text-driven 3d indoor scene synthesis with coherent geometry and texture. arXiv preprint arXiv:2305.11337, 2023. 12

[23] Alexander Swerdlow, Runsheng Xu, and Bolei Zhou. Street-view image generation from a bird's-eye view layout. IEEE RAL, 2024. 1, 3, 8

[24] Xiaoyu Tian, Tao Jiang, Longfei Yun, Yucheng Mao, Huitong Yang, Yue Wang, Yilun Wang, and Hang Zhao. Occ3d: A large-scale 3d occupancy prediction benchmark for autonomous driving. NeurIPS, 2024. 1, 3, 7

[25] Wenwen Tong, Chonghao Sima, Tai Wang, Li Chen, Silei Wu, Hanming Deng, Yi Gu, Lewei Lu, Ping Luo, Dahua Lin, et al. Scene as occupancy. In ICCV, 2023. 3

[26] Xiaofeng Wang, Zheng Zhu, Guan Huang, Xinze Chen, and Jiwen Lu. Drivedreamer: Towards real-world-driven world models for autonomous driving. arXiv preprint arXiv:2309.09777, 2023. 3, 8

[27] Xiaofeng Wang, Zheng Zhu, Wenbo Xu, Yunpeng Zhang, Yi Wei, Xu Chi, Yun Ye, Dalong Du, Jiwen Lu, and Xingang Wang. Openoccupancy: A large scale benchmark for surrounding semantic occupancy perception. In ICCV, 2023. 1, 3, 7

[28] Yuqi Wang, Jiawei He, Lue Fan, Hongxin Li, Yuntao Chen, and Zhaoxiang Zhang. Driving into the future: Multiview visual forecasting and planning with world model for autonomous driving. arXiv preprint arXiv:2311.17918, 2023. 3, 5, 9

[29] Yi Wei, Linqing Zhao, Wenzhao Zheng, Zheng Zhu, Jie Zhou, and Jiwen Lu. Surroundocc: Multi-camera 3d occupancy prediction for autonomous driving. In ICCV, 2023. 3

[30] Yuqing Wen, Yucheng Zhao, Yingfei Liu, Fan Jia, Yanhui Wang, Chong Luo, Chi Zhang, Tiancai Wang, Xiaoyan Sun, and Xiangyu Zhang. Panacea: Panoramic and controllable video generation for autonomous driving. arXiv preprint arXiv:2311.16813, 2023. 1, 3, 5, 8

[31] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. Tune-a-video: One-shot tuning of image diffusion models for text-to-video generation. In ICCV, 2023. 5, 14

[32] Zhennan Wu, Yang Li, Han Yan, Taizhang Shang, Weixuan Sun, Senbo Wang, Ruikai Cui, Weizhe Liu, Hiroyuki Sato, Hongdong Li, et al. Blockfusion: Expandable 3d scene generation using latent tri-plane extrapolation. arXiv preprint arXiv:2401.17053, 2024. 9

[33] Xu Yan, Jiantao Gao, Jie Li, Ruimao Zhang, Zhen Li, Rui Huang, and Shuguang Cui. Sparse single sweep lidar point cloud segmentation via learning contextual shape priors from scene completion. In AAAI, 2021. 3

[34] 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. 3, 8

[35] Ling Yang, Zhilong Zhang, Yang Song, Shenda Hong, Runsheng Xu, Yue Zhao, Wentao Zhang, Bin Cui, and Ming-Hsuan Yang. Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys, 2023. 3

[36] Mengjiao Yang, Yilun Du, Kamyar Ghasemipour, Jonathan Tompson, Dale Schuurmans, and Pieter Abbeel. Learning interactive real-world simulators. arXiv preprint arXiv:2310.06114, 2023. 9

[37] Yuanbo Yang, Yifei Yang, Hanlei Guo, Rong Xiong, Yue Wang, and Yiyi Liao. Urbangiraffe: Representing urban scenes as compositional generative neural feature fields. In ICCV, 2023. 3

[38] Ze Yang, Yun Chen, Jingkang Wang, Sivabalan Manivasagam, Wei-Chiu Ma, Anqi Joyce Yang, and Raquel Urtasun. Unisim: A neural closed-loop sensor simulator. In CVPR, 2023. 9

[39] Hong-Xing Yu, Haoyi Duan, Junhwa Hur, Kyle Sargent, Michael Rubinstein, William T Freeman, Forrester Cole, Deqing Sun, Noah Snavely, Jiajun Wu, et al. Wonderjourney: Going from anywhere to everywhere. arXiv preprint arXiv:2312.03884, 2023. 12

[40] Junge Zhang, Qihang Zhang, Li Zhang, Ramana Rao Kompella, Gaowen Liu, and Bolei Zhou. Urban scene diffusion through semantic occupancy map. arXiv preprint arXiv:2403.11697, 2024. 9

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In ICCV, 2023. 2, 3, 5

[42] Wenliang Zhao, Lujia Bai, Yongming Rao, Jie Zhou, and Jiwen Lu. Unipc: A unified predictor-corrector framework for fast sampling of diffusion models. NeurIPS, 2023. 7

[43] Tinghui Zhou, Richard Tucker, John Flynn, Graham Fyffe, and Noah Snavely. Stereo magnification: Learning view synthesis using multiplane images. arXiv preprint arXiv:1805.09817, 2018. 2, 4

[44] Benjin Zhu, Zhengkai Jiang, Xiangxin Zhou, Zeming Li, and Gang Yu. Class-balanced grouping and sampling for point cloud 3d object detection. arXiv preprint arXiv:1908.09492, 2019. 6

Appendix

In the appendix, we provide the following content:

Sec. A: Statement of Geometric Control.	Sec. E: Results of Video Generation.
Sec. B: Long-Tailed Scene Evaluation.	Sec. F: Generalize to New Cameras.
Sec. C: Ablation of plane number in MPIs.	Sec. G: Impact of Amount of Augment Data.
Sec. D: Additional Qualitative Comparison.	Sec. H: Visualization of Failure Cases.

A Statement of Geometric Control

In our paper, we refer the geometric controllable generation as using a voxel grid in 3D space to control the image generation. Although the voxel is a quantized representation of the 3D world, when the resolution goes larger, it can already faithfully represent the geometry detail of scenes. Currently, we are limited by the precision of ground truth labels. The $0.2m$ occupancy grid is a tensor of 500×500×40 that cover a space in x-axis spanning $[-50m, 50m]$, y-axis spanning $[-50m, 50m]$, z-axis spanning $[-5m, 3m]$. In the future, we plan to explore a higher resolution of geometric control to refine our generation.

Except for occupancy, several other 3D representations can be expressed by 3D semantic MPI, such as mesh, dense point clouds, and even 3D boxes or HD maps. The underlying mechanism is to cast several slices of multi-plane images at different depths to retrieve geometric information. Thus, our 3D semantic MPI can be regarded as a general 3D conditioning representation to benefit a wide spectrum of practical systems. These encompass but are not limited to 3D generation such as text2room [8], RoomDreamer [22], WonderJourney [39], and LucidDreamer [4], each of which stands to benefit from the rich geometric context provided by our approach.

B Long-Tailed Scene Evaluation

In this section, we explore to use SytheOcc to create long-tailed scenes for downstream evaluation. This also stands for evaluating our model using several corner cases. Similar to the SytheOcc-Gen experiment in Tab. 1, we generate a synthetic validation set but use prompts control to manipulate weather patterns or the intensity of illumination.

As depicted in Fig. 7. We create a variety of weather conditions including sandstorms, snow, foggy, rainy, day night, and day time. The motivation behind the creation of these scenes lies in their extreme rarity compared to the ordinary scenes we have captured. The generation of such data is of significant value, as it aids in addressing the long-tailed distribution of scenes, thereby enriching the diversity of our dataset. More visualization is provided in Fig. 13 to Fig. 14.

In Tab. 4, we observe that all kinds of extreme weather lead to a degradation in performance. This observation underscores the limitations of the perception model in terms of its generalizability to infrequent weather scenarios. Among them, we find that foggy, rainy, and day night exert the most severe impact, as they contribute to a large reduction in visibility as shown in Fig. 7. To improve the generalizability to handle various weather conditions, future work can leverage our generated data to cover the long-tailed scenes, or use adversarial search to find severe scenes based on our framework.

Scenes	**Sandstorm**	**Snow**	**Foggy**	**Rainy**	**Day night**	**Day time** (raw data)
FB-Occ mIOU	22.88	18.25	10.29	9.71	9.95	25.50

Table 4: Experiments of downstream evaluation on long-tailed scenes with extreme weather.

Furthermore, we perform long-tailed scene evaluation in Fig. 8. We display the failure of the downstream model VAD [9] in our synthetic long-tailed scene. In this case, we simulate a foggy environment that the dense fog obscures the majority of the ego view. Our experiment reveals that due to the lack of training images of foggy scenes, VAD erroneously predicts waypoints that would result in a collision with the bus. This experiment elucidates the boundaries and failure cases of the VAD model [9]. It exposes the limitations of the system under certain conditions, thereby providing insights into scenarios where the model's performance may be compromised.

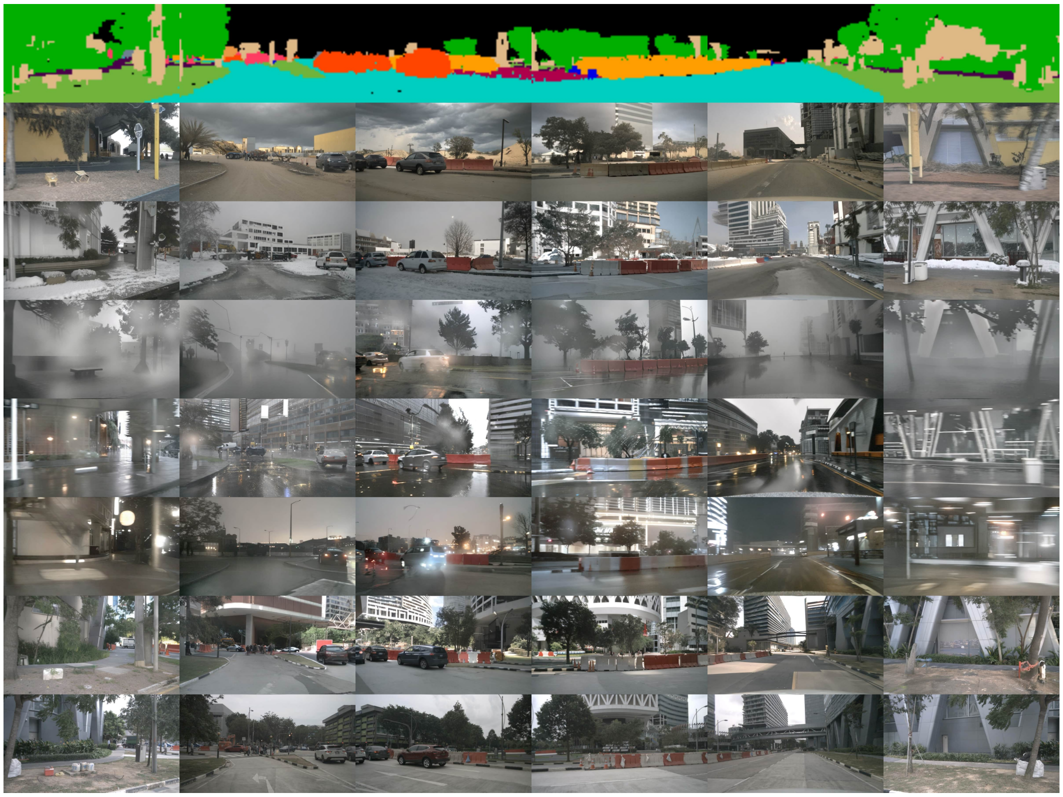


Figure 7: From top to bottom, we display images of fusion of 3D semantic MPI, synthesized images of sandstorm, snow, foggy, rainy, day night, day time, and ground truth.

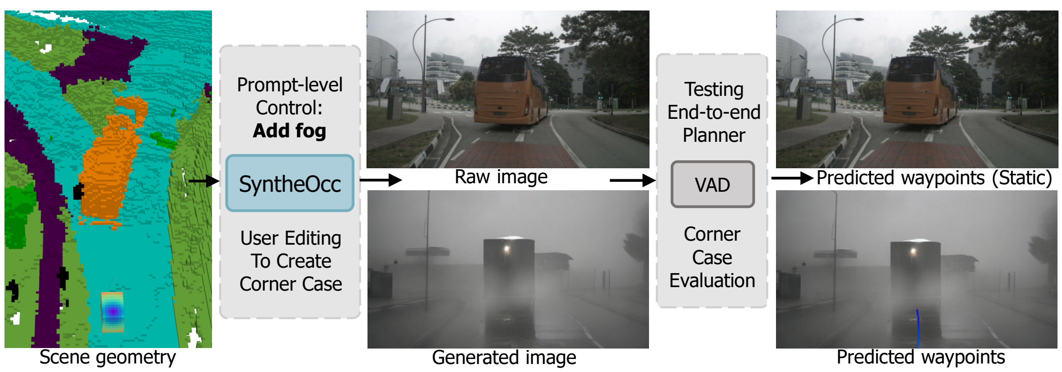


Figure 8: Use **SytheOcc** to create long-tailed scenes for testing. **Top**: In the ordinary scene of a bus placed in front of the ego vehicle, the end-to-end planner VAD [9] predicts future waypoints without movement, thus not plotted in the image. **Bottom**: By harnessing the prompt-level control in our framework, we simulate a scene with the same layout but filled with fog. VAD predicts wrong waypoints that will collide with the bus.

C Ablation of plane number of MPIs

In our proposed 3D semantic MPIs, the number of planes is a hyperparameter that affects the precision of 3D representation. The plane number can be regarded as the 3D resolution in depth axis. The larger the plane number, the MPI will contain more details. We find that an increase in the number of planes is associated with improved accuracy in downstream tasks. This finding denotes that more condition information leads to better downstream task performance.

Figure 9: Comparison with baselines.

Number of Planes	**96**	**128**	**256**
FB-Occ mIOU	23.36	24.28	25.50

Table 5: Ablation of the number of multi-plane images.

D Qualitative Comparison with Baselines and SOTA

In Fig. 9, we conduct a qualitative comparison of our method against MagicDrive, ControlNet, and ControlNet+depth. We find that all the methods display a satisfactory image quality, as they build upon the foundation of the stable diffusion model. The generation of MagicDrive fails to synthesize barriers as shown in the bottom row. ControlNet struggles to generate objects with the correct pose solely from only 2D conditions as shown in the second row. ControlNet+depth, a degradation of our method, an enhancement over ControlNet in terms of alignment, nevertheless suffers from a loss of finer detail in scenes with heavy occlusion, as shown in the human of the third row. Our method, in contrast, aims to address these challenges and provide a more nuanced and accurate generation of complex scenes.

E Extend to Video Generation

As described in the main paper Sec. 3.4, we further extend the cross-view attention to cross-frame attention to perform video generation. Our generation results are Fig. 11, Fig. 12 and Fig. 16. Our implementation is adopted from MagicDrive [5] which is similar to Tune-a-video [31]. The formulation of cross-frame attention is:

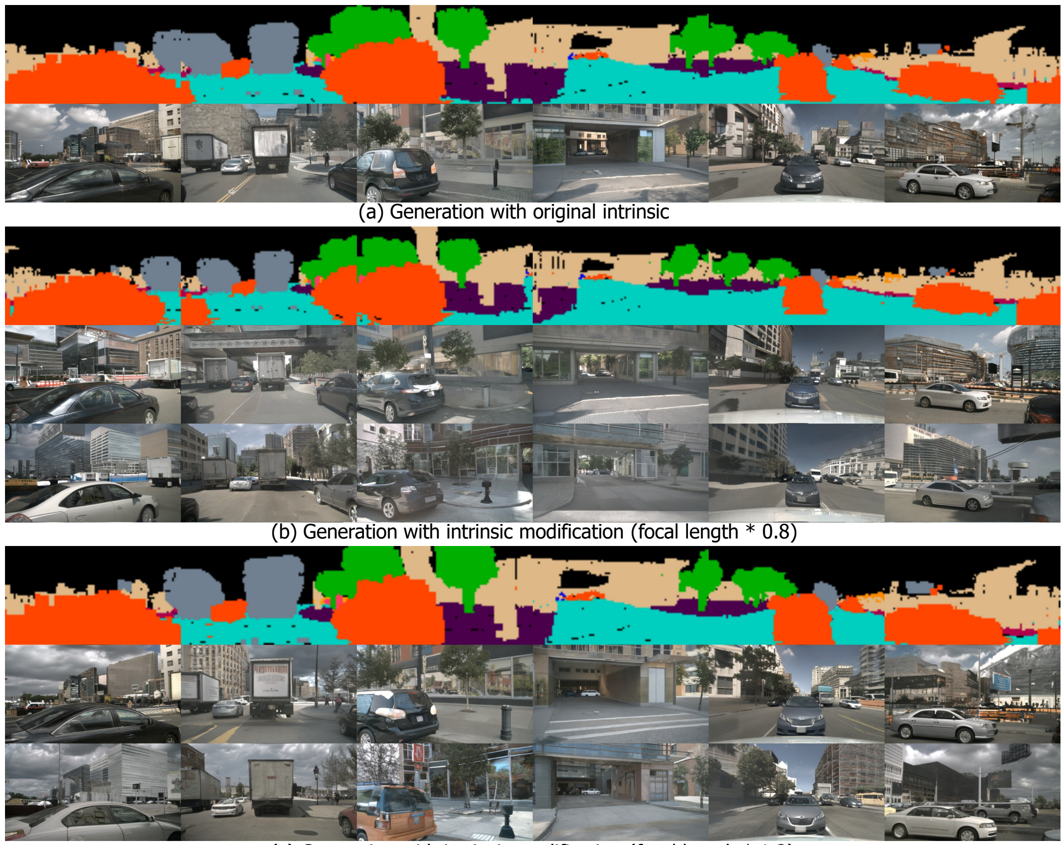
$$\texttt{Attention}(Q, K, V) = \texttt{softmax}(\frac{QK^T}{\sqrt{d}}) \cdot V, \tag{8}$$

$$h_{out} = h_{in} + \textstyle\sum_{i \in \{f,h\}} \texttt{Attention}(Q_{in}, K_i, V_i), \tag{9}$$

where f, and h are the camera view of future and history frames. Q_{in} and h_{in} denotes the query and the hidden state of input view. We train our model in a two-stage pipeline. We first train the MPI encoder and cross-view attention in a multi-view image generation setting. Then we train cross-frame attention and freeze other components in a video generation setting.

In practice, we use the keyframe annotation of the nuScenes dataset to train our video model. We start with our pretrained MPI encoder and cross-view attention and only train our cross-frame attention while keeping others frozen. We employ a sequence of 7 frames as a batch, resulting in a batch size of 42 images for the training process.

Given that our primary contribution does not lie in video generation, this experiment serves as a proof of concept, demonstrating the potential of our framework. Future research may extend our methodology to facilitate the generation of longer video sequences, thereby expanding the scope and applicability of our framework.



(a) Generation with original intrinsic

(b) Generation with intrinsic modification (focal length * 0.8)

(c) Generation with intrinsic modification (focal length * 1.2)

Figure 10: We demonstrate the generalizability of SytheOcc to new camera intrinsic. We multiply factors to the focal length while keeping the resolution the same. In (b,c), focal length $\times 0.8$ denotes a camera with a larger field of view similar to zoom out, focal length $\times 1.2$ denotes a camera with a smaller field of view similar to zoom in.

F Generalize to New Cameras

In this section, we investigate the adaptability of our method to a new set of cameras with different intrinsic. Given that our training set has a fixed camera intrinsic and extrinsic, generalizing to novel cameras indicates that our approach possesses robust generalization capabilities. As shown in Fig. 10, benefiting from our local type of condition, SytheOcc generates images that faithfully align with the new intrinsic, proving that SytheOcc do not over-fit certain parameters. Regarding extrinsic parameters, we can cast our MPI at the desirable locations to retrieve geometric information, thus inherently ensuring generalizability without doubt.

G The Influence of the Amount of Augmented Data

As SytheOcc is capable of generating an infinite number of synthetic data, we investigate the influence of the amount of augmented data on downstream tasks in Tab. 6. We find that when our augmented data is expanded from one-fold to two-fold of the training dataset, the performance of perception model slightly decreases. This may indicate the generated data has an optimal ratio for downstream tasks. Due to limited computational resources, we only experiment with a limited amount of ratio. Future work can conduct more thorough experiments to find a universal theorem.

Amount of Augmented Data	**0** (no augmentation)	**1**	**2**
FB-Occ mIOU	39.3	40.3	40.1

Table 6: Ablation of the amount of augmented data.

Figure 11: Video generation results. In the temporal progression, the distant buildings maintain a high degree of consistency, and objects retain their identical shapes and textures across different views and frames.

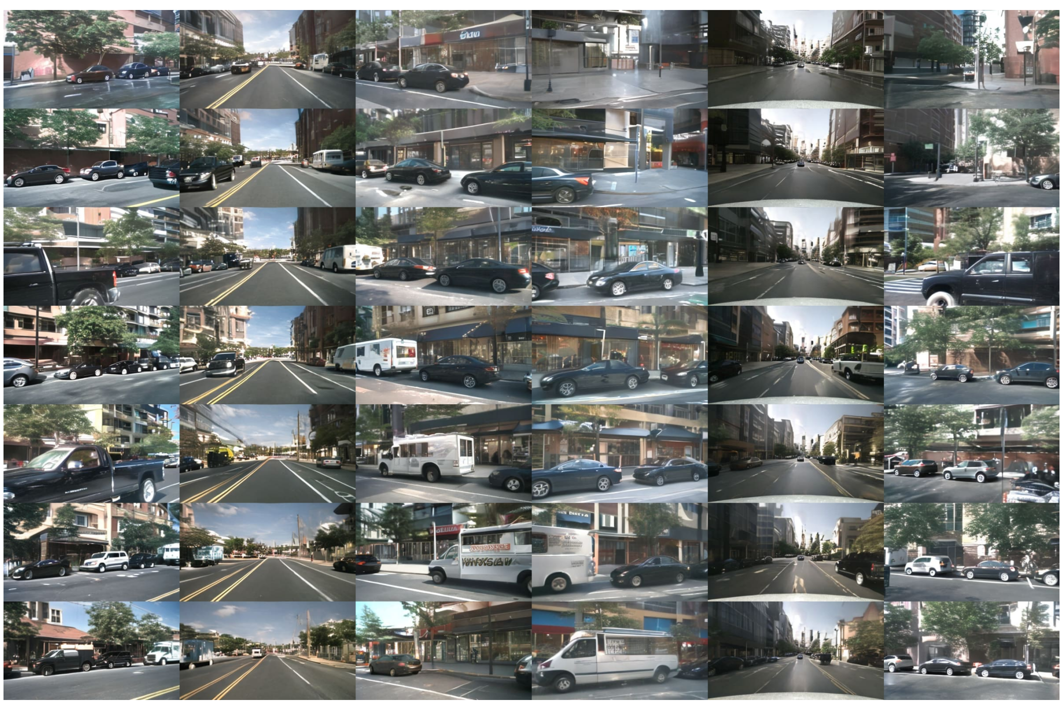


Figure 12: Video generation results of large dynamics scenes. The white car comes across different views and frames depicting consistent shapes with only a slight appearance change.

Figure 13: From top to bottom, we display images of fusion of 3D semantic MPI, synthesized images of sandstorm, snow, foggy, rainy, day night, day time, and ground truth.

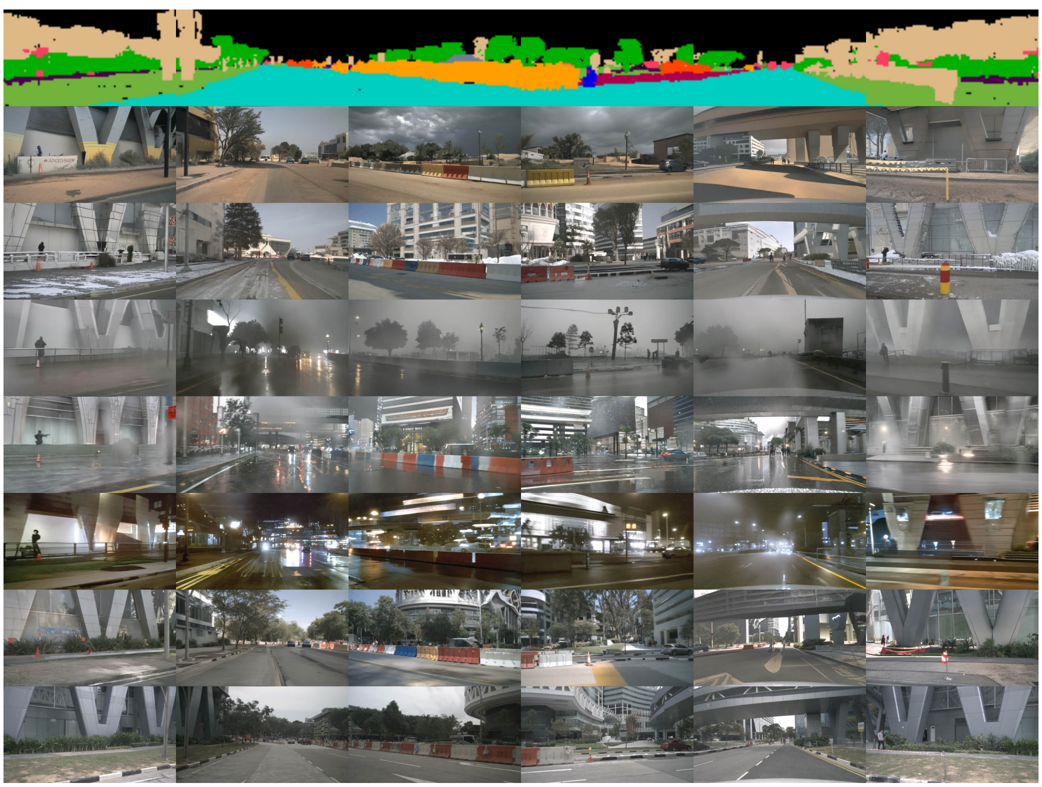


Figure 14: Weather variation. Same structure with Fig. 13.

Figure 15: Out of distribution generation. We use prompts to control the high-level appearance of images with specific styles. From top to bottom, we display (1) fusion of 3D semantic MPI. (2) Sunny day. (3) Science fiction style. (4) 8-bit pixel art style. (5) Snowfall. (6) Minecraft style. (7) Pokémon style. (8) Diablo style. (9) Ghibli style. (10) Metropolis style. (11) Gotham style. (12) Ground truth.

H Failure Cases

We display several failure cases of our method. In Fig. 16, we show a crowd scenes. In this scenario, the excessive number of pedestrians presents a challenge to the cross-view attention and cross-frame attention modules. We find our method incapable of discerning individual entities with clarity. Future research can improve the model capacity or enrich high-quality data to mitigate this problem.

Figure 16: Failure case of video generation results. Our cross-frame attention module is challenging to distinguish a crowd of people across different views and frames.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS paper checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: Please find this part in Sec. 3.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Please find this part in Sec. 5.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines: Do not have theoretical results.

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Please find this part in Sec. 4.

Guidelines:

- The answer NA means that the paper does not include experiments.

- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Please find this part in Sec. 4.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: Please find this part in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Please find this part in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
   - If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please find this part in Sec. 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.

- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: It should be fine.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please find this part in Sec. 5.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.

- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Please find this part in Sec. 4.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: Our paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.