

Supplementary Materials:

MVPbev: Multi-view Perspective Image Generation from BEV with Test-time Controllability and Generalizability

Anonymous Authors

This supplementary material consists of five sections. We provide more implementation details in Sec. 1. We then highlight our training-free instance-level control in Sec. 2, followed by human analysis and extension to objects can be found in Sec.3 and Sec. 4 respectively. Finally, we kindly ask the readers to check more qualitative results in Sec. 5.

1 IMPLEMENTATION DETAILS

1.1 Data preparation

Instead of using all frames in NuScenes [1], which can be highly redundant due to temporal consistency between consecutive frames, we propose to perform sampling on frames according to the geographic locations of ego car. Specifically, starting from the very first frame of a scene, we keep only the frames as long as their pairwise distances are greater than d meters. We then build a subset of NuScenes based on these remaining frames. In practice, d is set to 10 and we will have frames from 7288 and 1634 time stamps for training and validation. To further boost the efficiency, we then randomly sample 6000 and 1200 frames from them to train our model and report our overall performance respectively.

1.2 Baselines

In this section, we will provide more details about the baselines. Please note that we make some revisions to them so that they are fitted to our task. For instance, since neither of our baselines is capable of handling large changes in viewpoints, we assume that they are utilized at the second stage of our method, meaning that both of them take the perspective semantic and text prompts as input and aim to output multi-view perspective images. Otherwise notified, we launch all our experiments with one NVIDIA A40 GPU with PyTorch [5]. T is set to 50.

SD+Controlnet¹ We utilize publicly available code and pre-trained model in Diffusers [8] to re-implement Stable Diffusion(SD) [6] and Controlnet [9] model. Specifically, version 1.5 architecture and weights are used for the former, and version 1.1 architecture and semantic-conditioned pre-trained weights are used for the latter. To adapt these models to our task, we first fine-tune the SD on NuScenes for 8 epochs and then incorporate the Controlnet such that the control signal can be effectively leveraged. Given the control signal coming from $\{S_m\}_{m=1}^M$, we introduce a binary mask in each layer of Controlnet so that only the regions where signals are provided will be updated during training. Afterward, we fine-tune the SD+Controlnet for 2 more epochs, with parameters in Controlnet fixed. In practice, We find that our design gives better performance compared to jointly fine-tuning SD and Controlnet.

¹In our main paper, we use SD+Controlnet and Controlnet interchangeably.

During the fine-tuning process of SD, we set its batch size and learning rate to 6 and 1e-6 respectively. And Adam[4] is used as our optimizer. During inference, we set the guidance scale to 5.0.

MVDiffusion We re-implement MVDiffusion [7] based on its official code². To allow semantic conditions, we include a pre-trained Controlnet to its pipeline, followed by fine-tuning its original SD on NuScenes. Finally, we re-train the entire model of MVDiffusion with parameters of SD and Controlnet frozen. All hyper-parameters and training configurations, such as the number of epochs and learning rate, are chosen according to the official code of MVDiffusion.

1.3 More details about MVPbev

In this section, we provide more details about the second stage of our MVPbev. In practice, we follow the SD+Controlnet as our initial step. Then we implement our multi-view attention module and include it in the SD+Controlnet baseline. The multi-view module is further trained on NuScenes for 4 epochs. In practice, We set the learning rate and batch size to 1e-5 and 6. Again, Adam is used as the optimizer. As described in our main paper, we introduce novel initialization and denoising processes to explicitly enforce local consistency at overlapping FOVs. We observe that our design would improve the visual results if applied to up to $\frac{3*T}{5} = 30$ denoising steps.

2 TRAINING-FREE OBJECTS CONTROL

As described in our paper, MVPbev can be extended with instance-level controllability at test-time without extra training cost. To achieve this, we propose a special mechanism that manipulates the responses of cross-attention layers in multi-view LDM to accurately guide instance-level synthesis. In practice, the users will click on the target instance, or its 3D bounding box, and then choose the target color $\langle COLOR \rangle$, e.g., "red" or "deep blue". Then we generate one text description from the target color with format "A car colored $\langle COLOR \rangle$ ". Rather than working on its 3D bounding box, we turn to the 2D mask of this instance in perspective view. This instance-level mask can be obtained with either existing methods [2] or simple retrieval. For instance, one can retrieve 3D bounding boxes in training data and use the 2D mask of the one with the closest 6D distance. Let's denote the binary mask for the n -th instance as Y_n and $n \in \{1, \dots, N\}$. We further refer $A \in \mathbb{R}^{h \times w \times c}$ as to the response (i.e. output) from the cross-attention layer. Denoting the original text-prompt as \mathbf{o}_0 and the descriptions of other instances as $\{\mathbf{o}_n\}_n$, we can obtain their corresponding attention map as A_0 and $\{A_n\}_n$ by parsing them to pre-trained MVPbev model, together with BEV semantic $\{S_m\}_m$. We then effectively combine them with

²Please find their official release here <https://github.com/Tangshitao/MVDiffusion>

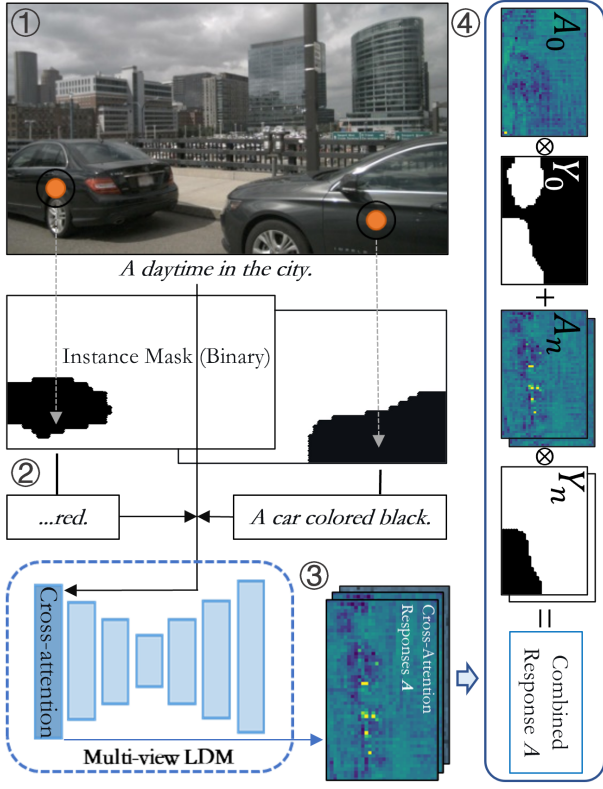


Figure 1: Our method to achieve test-time controllability by combining multiple attention responses from paired instance masks and text description, in a training-free manner.

the following equation:

$$A = A_0 \otimes (1 - \sum_n Y_n) + \sum_n (A_n \otimes Y_n) \quad (1)$$

where \otimes is the layer-wise multiplication. Our design ensures that each text prompts \mathbf{o}_n acts on instance region only, leading to more spatial-consistent performance. By manipulating cross-attention layers only, we are able to achieve the training-free goal, without introducing post-processings. Please see Fig.1 for its detailed structure.

3 HUMAN ANALYSIS

Human analysis provides a more reliable and intuitive tool for image quality measurement. Therefore, we conduct comprehensive human analyses of our tasks, which encompass human perception of multi-view consistency, view-point generalizability, and instance color controllability.

3.1 Cross-view consistency

We first focus on cross-view consistency where humans are asked to make decisions that which set of generated images reflects cross-view consistency in a better manner. Specifically, we provide two sets of generated images, which are generated from two different methods with the same input signal, to humans. Then we ask humans to decide which set of images is perceptually more realistic,

considering the image quality and visual consistency. We also allow humans to label them as 'undecided' but this option is not encouraged. In addition, GT and perspective semantics are also visualized for annotators' reference. We would like to note that all methods are compared anonymously. In practice, we invite 20 people with different backgrounds to perform quality comparisons. Each person is in charge of results on 30 exclusive frames. We report the percentage of *win*, *loss*, and *undecided* cases in a pair-wise form. For instance, .71 in the top-right of Tab. 1.(b) in our main paper means that 71% of MVPbev outperforms baseline MVD from the perspective of humans.

3.2 View-point generalizability

We conduct another human analysis to showcase whether pre-trained models can be adapted to unseen camera mountings. Rather than applying random camera mountings, we start from the camera setup from the original NuScenes, and rotate cameras w.r.t. pre-defined angle. We argue that this assumption is valid as compared to following the absolute angle from NuScenes, amounting to cameras w.r.t. their relative poses is much easier. Moreover, rotating all cameras by a fixed angle ensures that correspondences can be found across different views. Meanwhile, ground truth images can be used as good references for consistency in overlapping regions. Specifically, we revise the camera rotation w.r.t the direction of car head (i.e. yaw rotation) by $\{-25^\circ, -15^\circ, -5^\circ, 5^\circ, 15^\circ, 25^\circ\}$, respectively, which is equivalent to changing the $\{R_m\}_{m=1}^M$. For each rotation angle, we randomly select 200 sets of images and obtain results from MagicDrive [3] and ours with the same input signal. Subsequently, results from both methods, GT images, and projected road semantics are presented to humans. Humans will judge which method reflects the changes in camera pose better. For situations that are difficult to judge, we also have an "undecided" option. This option is discouraged. Finally, we report our results in the form of *win*, *lose*, and *undecided* rate. Please see Tab. 1.(b) in our main paper for the final results.

3.3 Instance-level controllability

The last human analysis we conduct is about instance-level controllability. In particular, we choose the photo-metric appearance as our control for the following two reasons. First of all, compared to other control signals such as shape, appearance, especially colors, are easiest to provide from an interaction perspective. Secondly, appearance is more user-friendly from an evaluation point of view. In practice, we first select 151 sets of images from NuScenes validation set, including 195 objects. Then we generate text descriptions for these with format "A car colored <COLOR>.". In particular, <COLOR> is a color randomly chosen from our palette, and each text description is associated with an instance-level mask. They are regarded as our new signals for instance-level control. Humans are asked to give their judgment on whether the generated instance color can be regarded as <COLOR>. As long as more than one person votes for "Yes", we believe the instance-level color control is fulfilled. In our experiment, 93.5% of objects are correctly generated.

4 EXTENSION TO OBJECTS



Figure 2: We provide three sets of experiments in this figure. Specifically, the first two sets showcase the multi-view consistency ability of our MVPbev, especially on objects that have been highlighted with orange bounding boxes. The last set of examples demonstrates that MVPbev can be easily applied to multi-object setting.

Though not shown in the main paper, our MVPbev can be extended to foreground road participants as well. To this end, we assume the 3D object information as well as their visibility are available, together with the original B . At the projection stage, objects are mapped to 2D perspective view if they are visible. Rather than utilizing the projected 2D bounding boxes in perspective view, we propose to introduce their masks. Specifically, for each object on the validation set of NuScenes, we can find its nearest neighbor in the training set such that their overall distance, which is measured by the averaged relative distances and poses w.r.t. ego car in L_2 space, is minimized. We then perform instance-level segmentation on this nearest-neighbor object and obtain its binary mask on M views. These masks are further pasted to $\{S_m\}_{m=1}^M$, leading to an additional semantic class in c_b . To effectively leverage the updated $\{S_m\}_m$ at the second stage, we further fine-tune our multi-view LDM with additional control signals.

We provide examples of our extended MVPbev model in Fig. 2. As can be found in this figure, our MVPbev is able to handle various objects in a cross-view consistent manner.

5 MORE QUALITATIVE RESULTS

Qualitative results of MVPbev We provide more regular results generated from NuScenes validation set in Fig. 3. As can be found in this figure, MVPbev can generate photo-realistic, multi-view consistent, and diverse images from complex road layouts in BEV

and text prompts. In addition, we further provide visual examples to demonstrate our controllability over diverse prompts in Fig. 4. Again, we observe that with the same input BEV semantics while various text prompts as the control signal, our method can generalize to unseen text prompts beyond our training settings in both prompt format and textual semantics.

View-point generalizability Similar to Fig. 9 in our main paper, more visual examples are provided in Fig. 5 to show how our method generalizes well to different camera setups, which is beyond SOTA MagicDrive [3] that requires far more training samples.

Instance-level controllability Please see Fig. 6 for more generated instances with its paired text prompts.

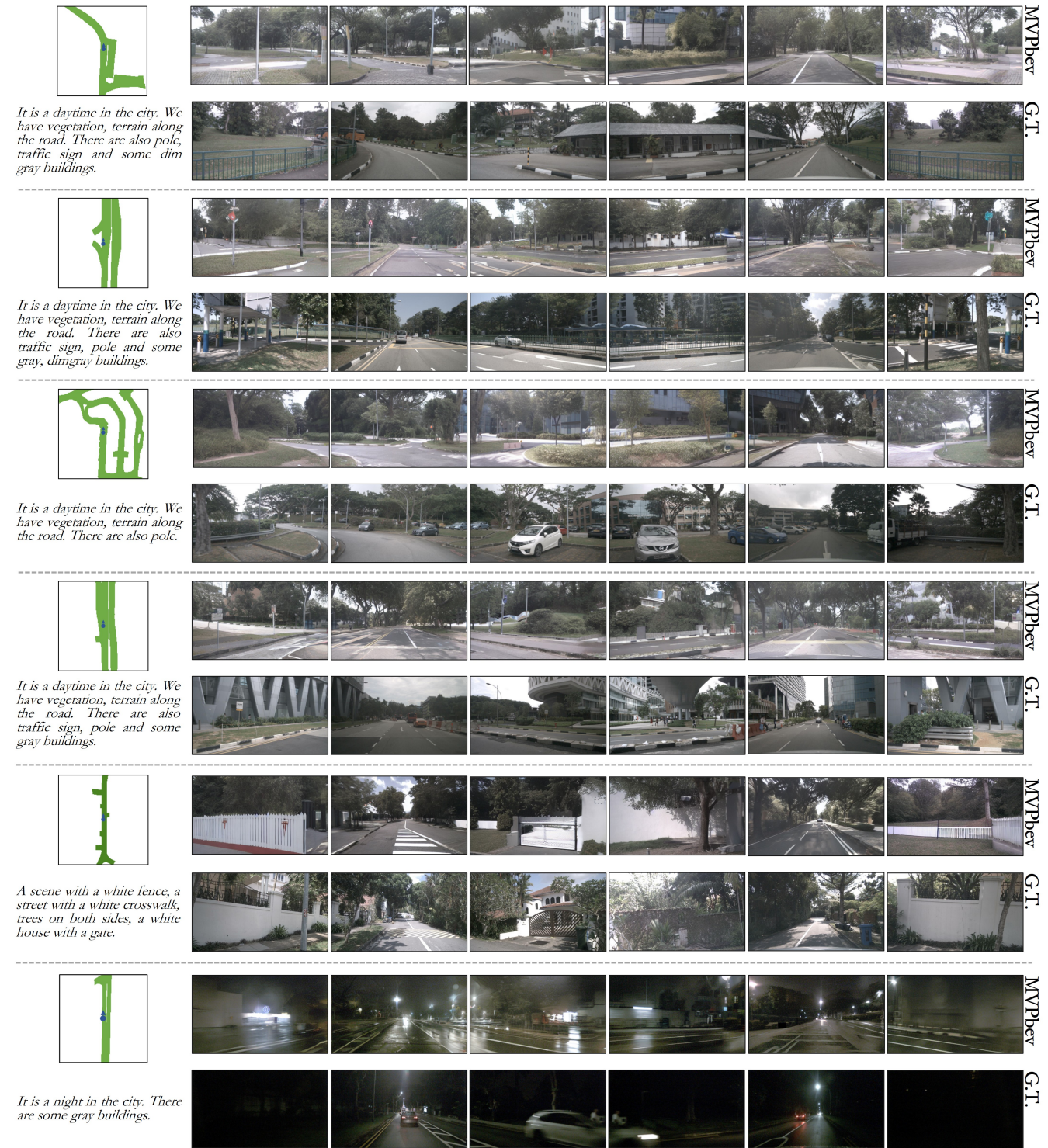


Figure 3: We provide five examples from NuScenes validation set. Our MVPbev is able to generate photo-realistic, multi-view consistent, and diverse images from complex road layouts in BEV and text prompts.

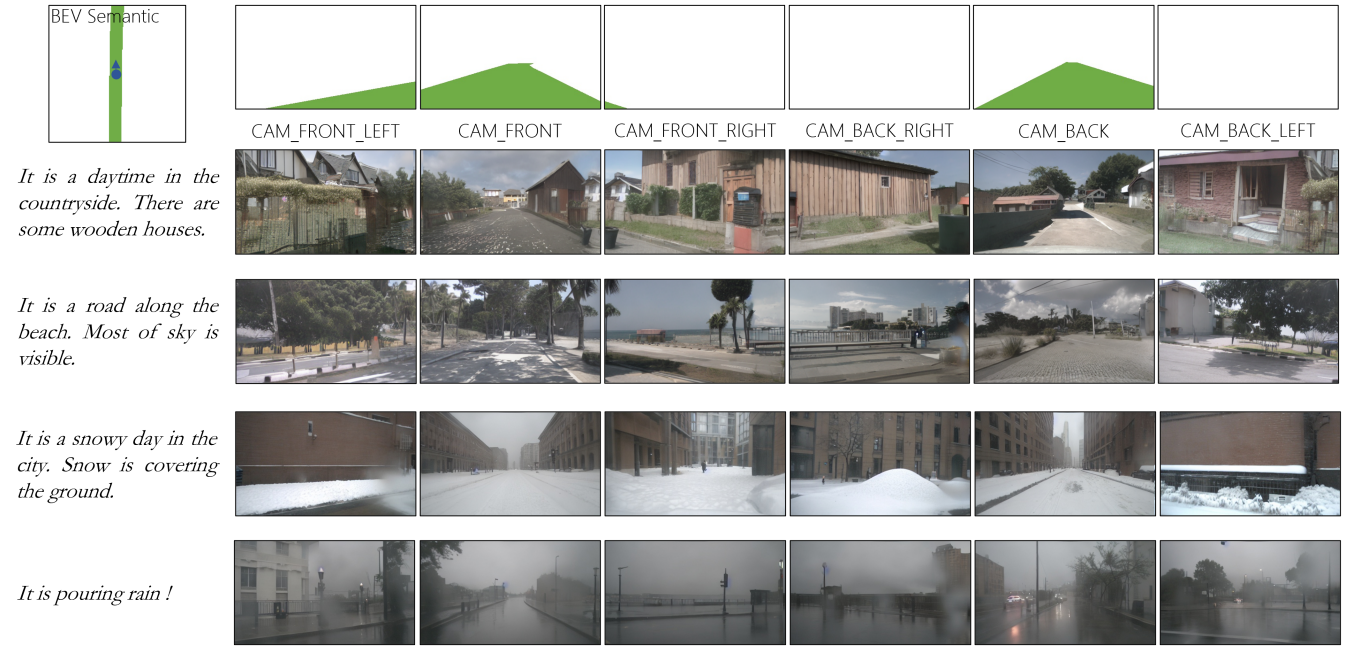


Figure 4: We provide four generated examples with fixed BEV while text prompt changes. Our MVPbev can generalize to various prompts, yielding diverse results with consistency in both semantic and textual aspect. Notably, our method can even generate results with unseen weather (e.g. "snowy day") that SOTA MagicDrive[3] can't achieve (see Conclusion section in their paper).

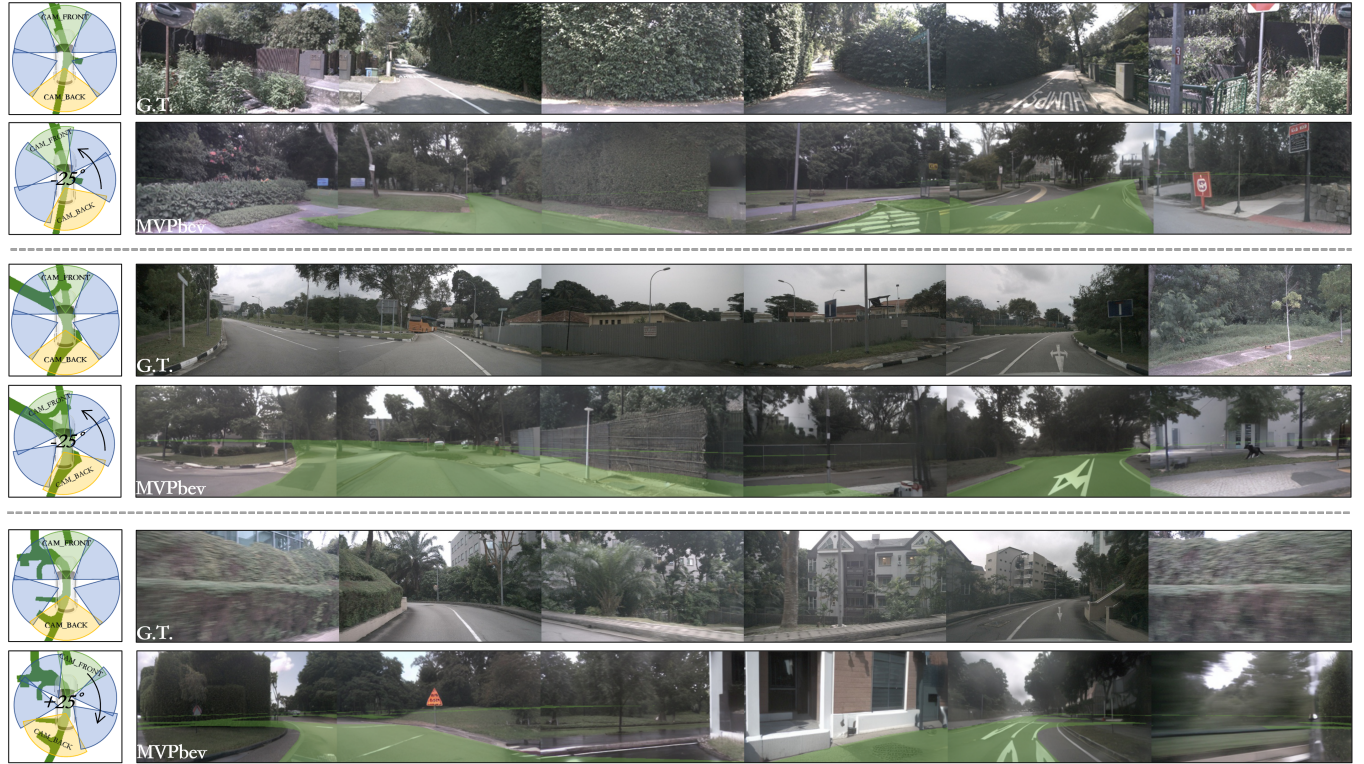


Figure 5: We provide three examples that show how our MVPbev generalize to different camera poses. The projected BEV semantics are overlaid in generated results.

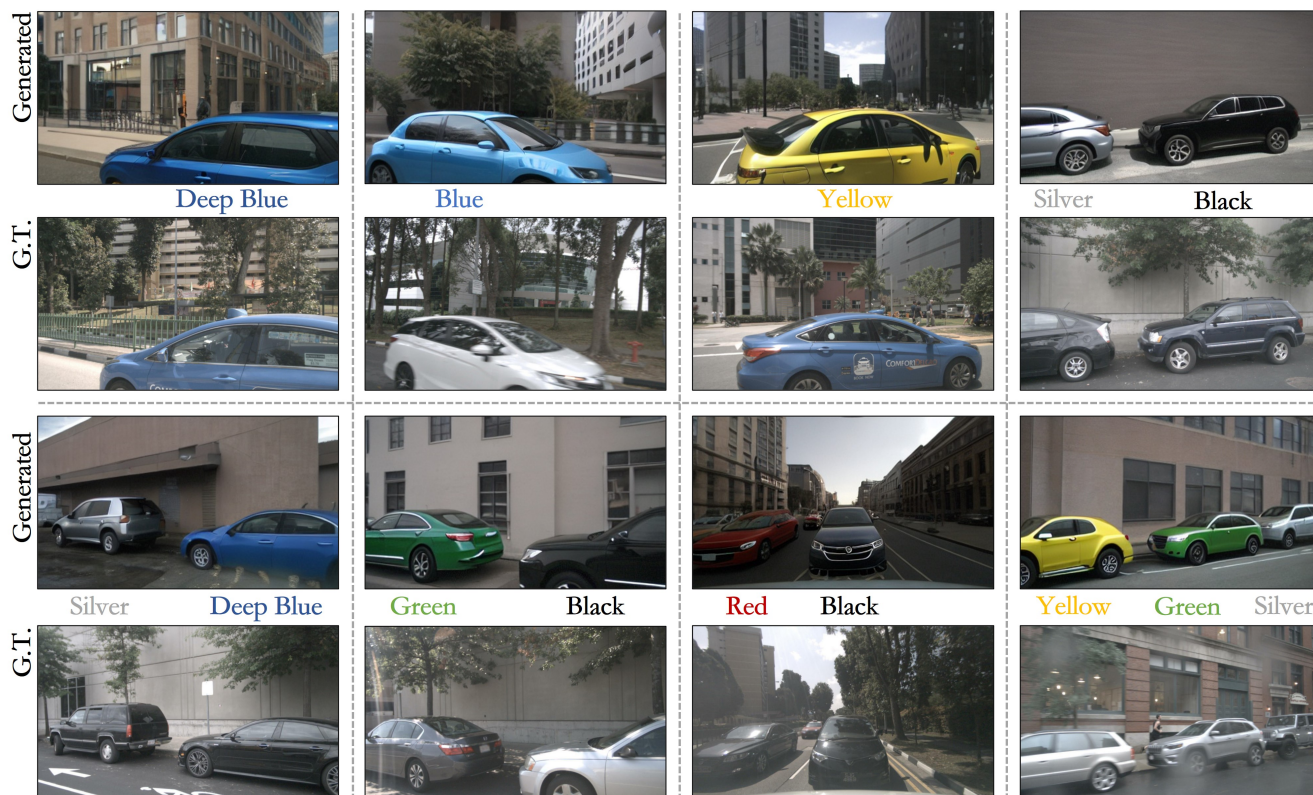


Figure 6: We provide more results in instance color control. Our training-free method can handle multiple instance control, ensuring controlled instances are aligned with their control signals (i.e. multiple instance-level prompts and paired masks), and generate natural instances, integrating well with backgrounds.

REFERENCES

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuscenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11621–11631.

[2] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. 2022. Masked-attention mask transformer for universal image segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1290–1299.

[3] Ruiyuan Gao, Kai Chen, Enze Xie, Lanqing Hong, Zhenguo Li, Dit-Yan Yeung, and Qiang Xu. 2023. Magicdrive: Street view generation with diverse 3d geometry control. *arXiv preprint arXiv:2310.02601* (2023).

[4] Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[5] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems* 32 (2019).

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2021. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752* [cs.CV]

[7] Shitao Tang, Fuyang Zhang, Jiacheng Chen, Peng Wang, and Yasutaka Furukawa. 2023. MVDiffusion: Enabling Holistic Multi-view Image Generation with Correspondence-Aware Diffusion. *arXiv* (2023).

[8] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. 2022. Diffusers: State-of-the-art diffusion models. <https://github.com/huggingface/diffusers>.

[9] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models.