2024 IEEE International Conference on Robotics and Automation (ICRA 2024)
May 13-17, 2024. Yokohama, Japan

From Bird's-Eye to Street View: Crafting Diverse and Condition-Aligned Images with Latent Diffusion Model

Xiaojie Xu, Tianshuo Xu, Fulong Ma and Yingcong Chen

***Abstract*— We explore Bird's-Eye View (BEV) generation, converting a BEV map into its corresponding multi-view street images. Valued for its unified spatial representation aiding multi-sensor fusion, BEV is pivotal for various autonomous driving applications. Creating accurate street-view images from BEV maps is essential for portraying complex traffic scenarios and enhancing driving algorithms. Concurrently, diffusion-based conditional image generation models have demonstrated remarkable outcomes, adept at producing diverse, high-quality, and condition-aligned results. Nonetheless, the training of these models demands substantial data and computational resources. Hence, exploring methods to fine-tune these advanced models, like Stable Diffusion, for specific conditional generation tasks emerges as a promising avenue. In this paper, we introduce a practical framework for generating images from a BEV layout. Our approach comprises two main components: the Neural View Transformation and the Street Image Generation. The Neural View Transformation phase converts the BEV map into aligned multi-view semantic segmentation maps by learning the shape correspondence between the BEV and perspective views. Subsequently, the Street Image Generation phase utilizes these segmentations as a condition to guide a fine-tuned latent diffusion model. This finetuning process ensures both view and style consistency. Our model leverages the generative capacity of large pretrained diffusion models within traffic contexts, effectively yielding diverse and condition-coherent street view images.**

I. INTRODUCTION

The emerging era of autonomous driving hinges on the adoption of sophisticated technologies and representations to ensure optimal navigation and decision-making. Among these, the bird's-eye view (BEV) holds a unique position. By offering a top-down, map-like representation, the BEV provides invaluable insights into the immediate environment, capturing pertinent obstacles and hazards.

While BEV perception [1], [2], [3] has been a focal point in recent studies, promising to bridge the transformation between street-level views and overhead perspectives, BEV generation—specifically the synthesis of realistic street-view images from a predefined BEV semantic layout—offers untapped potential.

At its core, BEV generation [4] translates a semantic layout, which captures a traffic scenario, into tangible street-view images. This translation facilitates an enhanced visualization of traffic scenarios in a real-world setting, making the abstract more accessible. One of the most compelling applications of BEV generation is the intuitive interface it offers for traffic scene visualization and modification. BEV generation allows human operators and system designers to modify a layout effortlessly, producing corresponding street-view images via generative models. This not only streamlines the training of autonomous systems but also serves as an effective testing and validation tool.

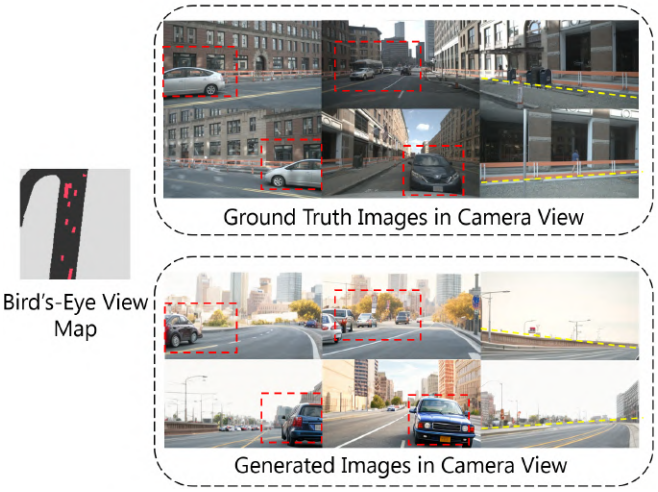

Fig. 1. From a bird's-eye view semantic map, our framework is capable of generating high-quality and varied camera view images. In terms of map elements, our results closely match the ground truth images. The red boxes (seen in the left four images) represent vehicles, while the yellow lines (in the right two images) delineate road contours.

BEVGen [4] represents a pioneering effort in addressing the BEV generation problem. Within its BEV representations, map components are bifurcated into two categories: vehicles and roads. The model employs an autoregressive transformer [5] with a spatial attention design to comprehend the relationship between camera and map perspectives. While BEVGen sets a baseline by generating multi-view images consistent with its map perspective, it doesn't consistently ensure condition coherence due to its implicit encoding mechanism.

In contrast to the previous method, our proposed method disentangles the view transformation and image generation processes. The view transformation phase focuses on learning the shape correspondence between map and camera perspectives. Here, to project the BEV map onto camera views using camera parameters, we assign height from a prior distribution to each BEV map segment. Using this

The authors Xiaojie Xu, Tianshuo Xu and Fulong Ma are with The Hong Kong University of Science and Technology(Guangzhou), Nansha District, Guangzhou, Guangdong, China. {xxu763, txu647, fmaaf}@connect.hkust-gz.edu.cn

The corresponding author Yingcong Chen are with The Hong Kong University of Science and Technology(Guangzhou), Nansha District, Guangzhou, Guangdong, China and The Hong Kong University of Science and Technology, Clear Water Bay, Kowloon, Hong Kong. yingcongchen@ust.hk

projection as a preliminary estimate, a convolutional network is employed for shape refinement, achieving a more precise camera view segmentation. This refined segmentation acts as the conditional information for the image generator. For image synthesis, we resort to a conditional latent diffusion model [6], chosen for its standout performance in conditional image generation tasks. Initially trained on diverse datasets, the diffusion model is fine-tuned using our driving scene imagery. Notably, the fine-tuning procedure encode camera viewpoint explicitly, ensuring that various views yield plausible outcomes (e.g., reasonable orientation of vehicles and roads). Leveraging precise transformed segmentation as condition and the generative ability of the diffusion model, our framework delivers high-quality, diverse, and condition-coherent results.

Our contributions are summarized as the following:

- We develop a novel framework for street-view image generation from a BEV layout, leveraging a large, pre-trained latent diffusion model. This encompasses view transformation, street-view adaptation, and conditional generation.
- We explore the methodology of encoding viewpoint for multi-view images and incorporating them into generative diffusion models, through which our method can produce diverse and flexible scenes that match the desired view and layout.
- We investigate the potential of utilizing large generative models for the task of BEV image generation and conduct a thorough comparison with other methods that are trained from scratch. Our method is efficient and effective, achieving high-quality and diverse results. Our experimental results demonstrate that our approach outperforms or matches existing methods in terms of visual quality and condition consistency.

II. RELATED WORK

Conditional image generation: The field of conditional image generation has seen notable advancements recently, with models predominantly conditioned on text [7], [8] or speech [9] inputs. Varied formats, such as class conditions [10], sketches [11], style [12], and distinct human poses [13], can convey the envisaged image specifications. Furthermore, several scholars have explored methodologies with high level representations, including generating images from semantic masks [14] or translating intricate constructs like scene graphs [15] and bounding boxes [16] into equivalent semantic masks. Diverging from these mentioned paradigms, our emphasis lies on the bird's-eye view map. Though akin to a semantic segmentation map, it offers a perspective distinct from the resulting image, which is seldom explored in earlier studies.

Image diffusion models: Originally proposed by Sohl-Dickstein et al. [17], Image diffusion models have found recent applications in image generation [18]. The Latent Diffusion Models(LDM) [6] execute diffusion in the latent image space [19], optimizing computational efficiency. Text-to-image diffusion models, by encoding textual inputs into latent vectors using pretrained language models like CLIP [20], set new benchmarks in image generation. Glide [21] stands out as a text-driven diffusion model for both image creation and editing. Stable Diffusion scales up the concept of latent diffusion [6], and Imagen [8] takes a distinct approach by diffusing pixels through a pyramid structure, bypassing latent imagery. We employ Stable Diffusion as our foundational pretrained model. Through fine-tuning, we adapt it to various viewpoints and driving scenes.

BEV perception and generation: Recent growth in large 3D datasets in autonomous driving [22], [23], [24] has propelled studies on map-view perception. Given the disparity between the coordinate frames of inputs and outputs, this domain poses challenges. While inputs derive from calibrated cameras, outputs are rasterized onto a map. A prevalent method assumes a mostly planar scene, simplifying image-to-map transformations via homography [25]. However, this can create artifacts for dynamic entities like vehicles. As a solution, some studies [26], [27] utilize depth and semantic maps to present objects in BEV. Alternatively, other methods [2], [3] bypass explicit geometric modeling to generate map-view predictions directly from images.

As its counterpart, generating from a BEV map layout remains relatively unexplored. BEVGen [4] pioneered this domain, employing an auto-regressive transformer to encode the connection between image and BEV representations. In contrast to BEVGen, our approach leverages a large, pretrained diffusion model as the backbone and finetunes it using driving scene images.

III. METHOD

The objective of BEV generation is to generate multiple camera-view images from a semantic BEV layout. Earlier studies have represented the BEV layout in either rasterized [2] or vectorized forms [28]. In this work, we favor the rasterized representation due to its aptness for creating from projections of 3D bounding boxes onto local street maps [2], or directly from driving simulation frameworks [29]. Consequently, the BEV layout is denoted by $B \in \mathbb{R}^{H_b \times W_b \times c}$, where c represents the number of map element categories, such as vehicles and roads.

Given the BEV map B and n camera views $(K_i, R_i, t_i)_{i=1}^{n}$, where K_i, R_i, t_i denotes the intrinsics, extrinsics rotation and extrinsics translation of the i_{th} camera, our goal is to generate n corresponding images in camera view $\mathcal{I} = \{\mathbf{I}^i \in \mathbb{R}^{H \times W \times 3} \mid i = 1, ..., n\}$.

As depicted in Fig. 2, our pipeline operates in two stages. Initially, the BEV's semantic information is projected into the camera view leveraging camera parameters, under a height assumption. This shape is subsequently refined using a CNN. In the succeeding stage, a pre-trained UNet undertakes the backward diffusion process [6], where Gaussian noise is progressively eliminated. This UNet receives the polished semantic information coupled with the prompt as conditioning inputs. Furthermore, to ensure accurate viewpoints across various camera perspectives, we fine-tune the network.

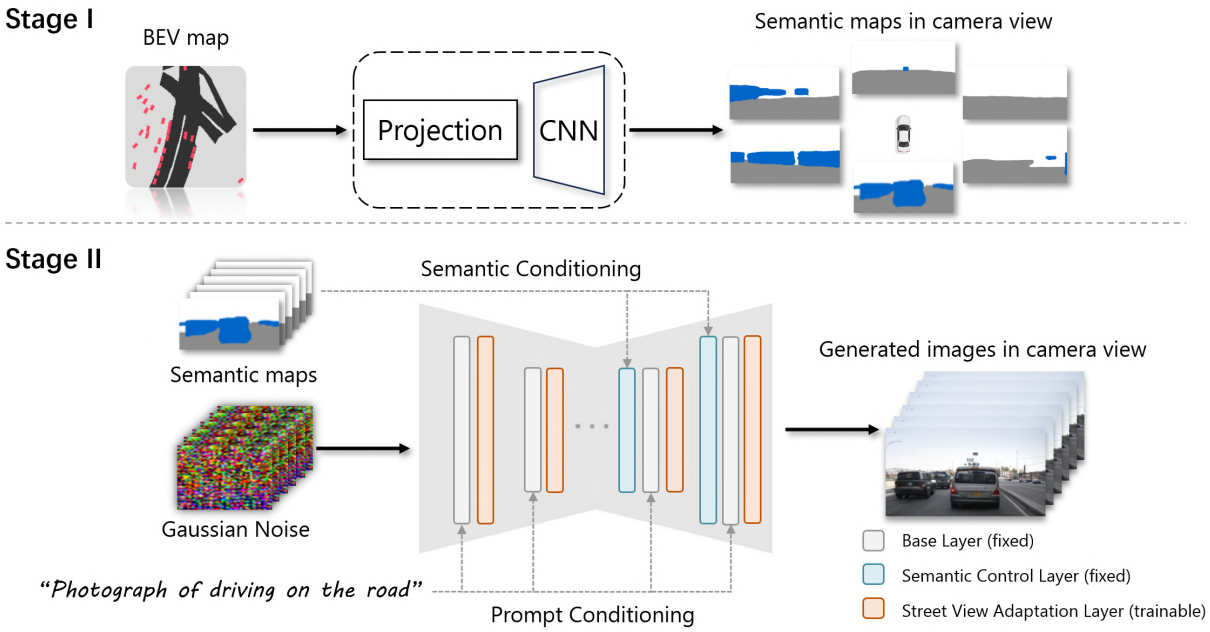

Fig. 2. Our two-staged pipeline. Initially, a BEV map is projected and refined to produce semantic maps from the camera's perspective. These semantic maps, paired with the prompt, are then fed into a pretrained U-Net for iterative denoising. We've incorporated street-view adaptation layers into the network to ensure style and viewpoint alignment.

A. Stage I: Neural View Transformation

Taking inspiration from [30], we treat the BEV-to-camera view transformation as an image translation task, where the input and output share a pronounced spatial correspondence. We decompose this transformation into two phases: initial setup using camera parameters and shape refinement via a neural network.

Initial projection with camera parameters: For any world coordinate $X \in \mathbb{R}^3$, the perspective transformation describes its corresponding image coordinate $x \in \mathbb{R}^3$ in the view of the i_{th} camera by

$$x = K_i R_i (X - t_i) \tag{1}$$

in homogeneous coordinates.

The lack of precise height data renders the world coordinates of BEV map data ambiguous, necessitating height estimation. While Inverse Perspective Mapping (IPM) techniques [31] operate under the premise of a flat ground, this assumption can introduce distortions for objects of varied heights, such as buildings and vehicles. Given our focus on roads and vehicles, we retain this simplified assumption for roads.

For vehicles, we posit that their height adheres to a predetermined distribution. Practically speaking, each vehicle on the BEV map is allocated a height randomly sampled from $\mathcal{U}(1.5, 2)$, offering a plausible initial height approximation. With the estimated heights for roads and vehicles in place, the BEV map is projected into camera views using Equ. 1, given the camera parameters.

Shape refinement network: Through height estimation and projection, we obtain preliminary semantic maps in the camera view. Nonetheless, this simplistic initialization fails to preserve the intricate shapes of map elements accurately. Given that vehicles are rendered on the BEV map using their true 3D bounding boxes as described in [2], our projection approach results in the vehicle appearing as a cube from the camera's viewpoint. Hence, a shape-refinement post-processing step is imperative.

Fig. 3. The impact of shape refinement on the final image generation is evident. Without refinement, the resulting image (left) resembles a cube. In contrast, the refined version (right) exhibits a more natural form.

The initial projection yields a low-resolution estimate. To address this, we employ an enhanced UNet architecture with residual connections [32]. This network bridges the shape discrepancy between the estimated and the true semantic maps. Functioning as an upsampling module, it outputs high-resolution semantic maps with finer geometry. These refined maps subsequently serve as conditional inputs to the image generator. The contribution of this network to the final image generation outcome is illustrated in Fig. 3.

B. Stage II: Street Image Generation

We utilize Stable Diffusion, which is a strong pretrained image generator based on latent diffusion [6] framework, as our generative backbone. In this section we discuss how the conditional generation mechanism works and how to adapt

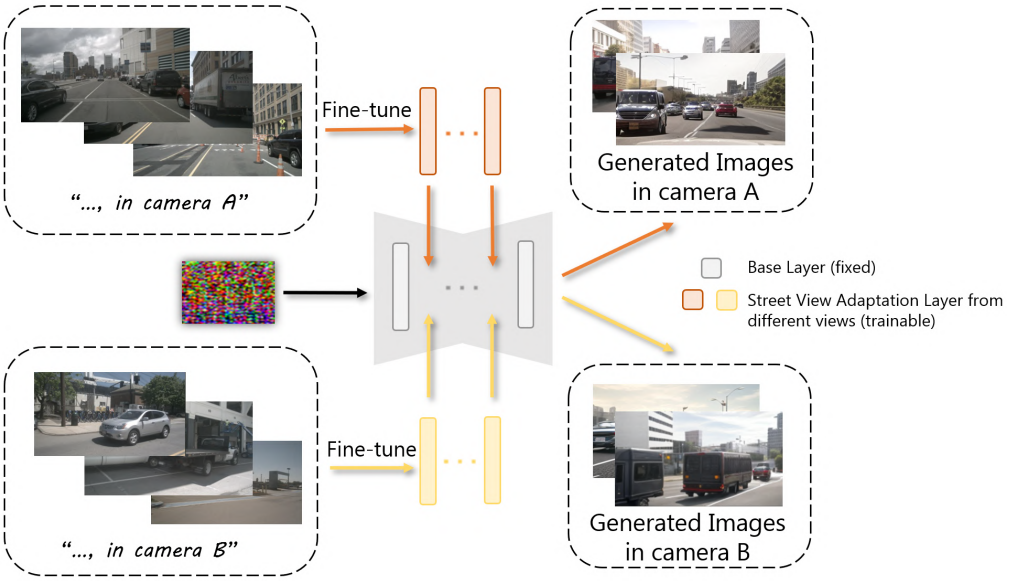

Fig. 4. We incorporate viewpoints into our foundational diffusion model by integrating specific views into the text prompts, resulting in distinct View Adaptation Layers. During sampling from the model, we can generate images from a designated camera by invoking its learned novel prompt.

the large pretrained model to our driving domain.

Conditional generation with latent diffusion model: Diffusion models can be conceptualized as a uniformly weighted sequence of denoising autoencoders, given by $\epsilon_\theta(x_t, t); t = 1 \dots T$, These autoencoders aim to predict a denoised version of their input x_t, where x_t represents a noisy variant of the original input x. This leads to the following objective:

$$L_{DM} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right] \tag{2}$$

with t uniformly sampled from $\{1, \dots, T\}$.

As a large text-to-image diffusion model, latent diffusion introduces CLIP [20] encoder $\mathcal{T}_\theta$ that projects the text prompt y to an intermediate representation $\mathcal{T}_\theta(y)$, which is then mapped to the intermediate layers of the UNet via a cross-attention layer implementing

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right) \cdot V \tag{3}$$

, with $Q = W_Q^{(i)} \cdot \varphi_i(z_t), K = W_K^{(i)} \cdot \tau_\theta(y), V = W_V^{(i)} \cdot \tau_\theta(y)$. In this context, $\varphi_i(z_t)$ symbolizes a flattened representation of the U-Net at an intermediate stage.

Our generation task encompasses more than just using the prompt as conditional information. The semantic data transformed from the BEV map serves as a superior control mechanism, given that the resultant image should align spatially with these semantic maps in pixel space. This necessitates a more precise conditioning mechanism for our objective.

Drawing inspiration from ControlNet [33], which employs zero convolution and a trainable duplicate of the original neural network, our approach manipulates the input conditions of neural network blocks. This strategy allows for a more nuanced control over the entire neural network's behavior. We integrate the pretrained ControlNet layers, designed for semantic segmentation, into our architecture (as depicted in Fig. 2). These layers act as conditioning controllers for the image generation process. Even though these semantic control layers were trained on a broader dataset [34], they exhibit robust generalization capabilities in our driving scenarios.

Street-view adaption: Our street view adaptation module serves a dual purpose. Firstly, it emulates the driving scene's image style found in our dataset [22]. Secondly, it encapsulates the viewpoints associated with various cameras.

While fine-tuning the diffusion model using Equ. 2 aids in capturing a realistic style specific to driving scenarios, it's crucial to remember that street scenes, when viewed from different camera perspectives, can vary significantly. For instance, when viewed through our front camera, a vehicle directly ahead should align with our car's driving direction. In contrast, the same vehicle observed from a side camera would appear at an angle. Likewise, driveable areas typically extend more prominently when viewed from the front and rear cameras but appear more constrained from the side angles.

Informed by these insights, we fine-tune our image generator for specific viewpoints. The mechanism for view encoding is detailed in Fig. 4. Taking a cue from DreamBooth [36], which hones in on a personalized concept (e.g., a particular dog) as a unique prompt, we treat the viewpoint as an abstract concept and introduce a view-specific loss to optimize the diffusion model. This ensures that the viewpoint is distinct from foundational concepts like cars or streets within the prompts. The training loss is articulated as:

$$\begin{aligned}\mathbb{E}_{\mathbf{x},\mathbf{c},\boldsymbol{\epsilon},\boldsymbol{\epsilon}',t}[w_t\|\hat{\mathbf{x}}_\theta(\alpha_t\mathbf{x} + \sigma_t\boldsymbol{\epsilon}, \mathbf{c}) - \mathbf{x}\|_2^2 + \\ \lambda w_{t'}\|\hat{\mathbf{x}}_\theta(\alpha_{t'}\mathbf{x}_{view} + \sigma_{t'}\boldsymbol{\epsilon}', \mathbf{c}_{view}) - \mathbf{x}_{view}\|_2^2]\end{aligned} \tag{4}$$

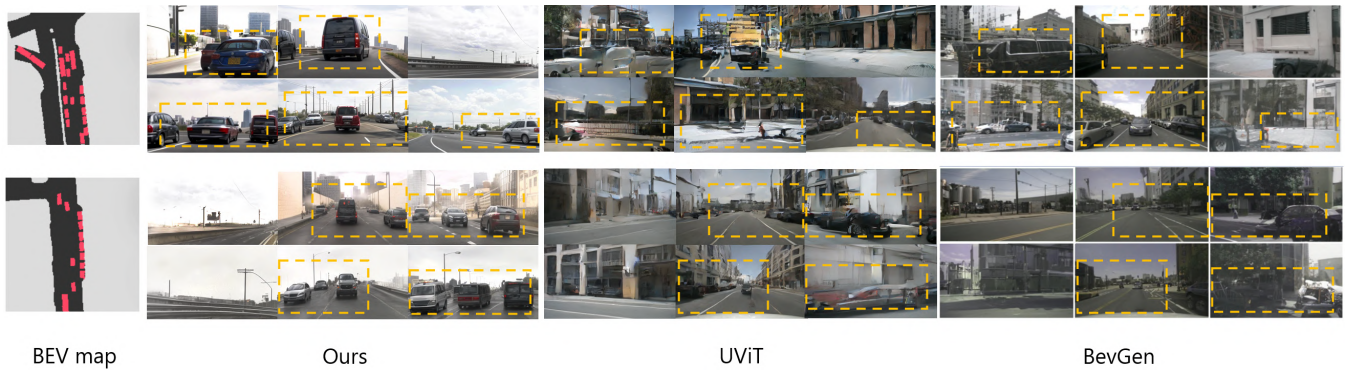

Fig. 5. We compare our method (left) with UViT (middle) [35] and BevGen(right) [4]. Our results demonstrate greater stability and more effective use of conditional information, especially in the highlighted yellow regions where the condition should take effect. For best results, it is recommended to zoom in.

, where $\mathbf{x}_\theta$ represents the base model, σ_t and $\sigma_{t'}$ refer to distinct Gaussian noises, and $\mathbf{c}$ and $\mathbf{c}_{view}$ signify the prompt, either with or without the explicit inclusion of the viewpoint.

Rather than fine-tuning the entire network, we leverage the Low Rank Adaptation (LoRA) [37] technique to achieve rapid training and enhanced flexibility. The base model is shared across views.

IV. EXPERIMENTS AND RESULTS

A. Dataset

The nuScenes dataset [22] is a comprehensive collection encompassing 1,000 diverse street-view scenes, captured under varied weathers, times of day, and traffic conditions. Spanning over 20 seconds, each scene consists of 40 frames, amounting to a total of 40,000 samples within the entire dataset. Designed to provide a 360° perspective around the ego-vehicle, the data is derived from six distinct camera views, capturing images from the side, front, and back of the vehicle. Every camera view comes with calibrated intrinsics (K) and extrinsics (R, t) for each timestep. Furthermore, objects, including vehicles, are consistently tracked across frames and annotated using 3D bounding boxes derived from LiDAR data. The dataset is organized into 700 training, 150 validation, and 150 testing scenes.

Following [2], the semantic mask of the vehicle in BEV is rendered with a resolution of (200,200). This is achieved by orthographically projecting the 3D box annotations onto the ground plane, which corresponds to a (100m,100m) region in the real-world context. The road masks are formulated using the NuScenes map devkit, which integrates both lanes and road segments.

B. Implementations

Shape refinement network: The shape refinement network is a convnet comprising three down-sampling blocks and four up-sampling blocks. It accepts inputs with a resolution of (56,100) and produces outputs with a resolution of (224,400). Given that the original nuScenes dataset does not include image semantic labels, we employ SegFormer [38] to generate pseudo labels. We train the network for 10 epoches with a learning rate 1e-7.

Pretrained Stable Diffusion and control module: We utilize the pretrained Stable Diffusion model "RealisticVision" available on HuggingFace [39]. The control module is adapted from ControlNet [33], which was originally trained on the ADE20K dataset [34] and captioned using BLIP [40].

Street view adaptation module: For each camera view, we use a set of 100 images to train the respective adaptation module. Our foundational prompts for regularization include "road", "car", and "street background". To specify viewpoints, we use alphanumeric designations (e.g., cam0) to prevent any overlap with existing concepts within the pretrained CLIP text encoder [20]. During finetuning, the image resolution is set at (400, 224). The training extends over 5000 steps with a batch size of 4 and a learning rate set at 1e-4. The rank for LoRA [37] is set to 16.

C. Results

Qualitative result: We juxtapose our approach against BEVGen [4] and a from-scratch trained latent diffusion model using a transformer architecture, specifically UViT [35]. Notably, our strategy involves finetuning a pre-trained, expansive model, while the other two approaches train their models from the ground up. The results can be observed in Fig. 5.

Our method showcases superior stability in image quality, and its conditioning mechanism proves to be effective. Both UViT and BevGen employ cross attention to manage conditional information. However, their models occasionally falter due to the absence of explicit spatial relationships between the semantic and the resultant generated images. This makes it challenging for their conditioning mechanisms to consistently function effectively. Concerning image quality and diversity, methods that are trained from scratch tend to be closely tied to specific datasets, often risking overfitting. In particular, the UViT-based diffusion model faces challenges when trained with a limited dataset.

In Fig. 6, we showcase additional illustrations underscoring the diversity of our generated outcomes. Our methodology effortlessly facilitates the generation of images under various weather scenarios, significantly enhancing the model's adaptability.

Fig. 6. Leveraging the robust generative capabilities of the large pretrained Stable Diffusion model, our generated outcomes display remarkable diversity. This figure presents results under varying weather conditions, all derived from a consistent BEV input. The red region illustrates the road variations in the driving scene image due to changing weather conditions. For best results, it is recommended to zoom in.

Quantitative result: In Table. I, we juxtapose our approach with the benchmark BEVGen and a transformer-driven diffusion model. Utilizing the Frechet Inception Distance (FID) [41], akin to BEVGen, we evaluate the congruence between the generated images and the training dataset. While our outputs are visually appealing and consistent, our FID score lags behind BEVGen. This can be attributed to our reliance on limited data for fine-tuning, hence the visual style largely remains anchored to the foundational diffusion model. For a more equitable comparison, we trained a UViT-based latent diffusion model from scratch, which yielded an even less favorable FID score. This suggests that the scope of the training dataset might be insufficient, complicating the task of cultivating a robust diffusion model from scratch.

Further, we assessed our methodology using a pretrained BEV segmentation model [2]. To gauge the congruity between the predicted and actual BEV segmentation maps, we employed the mean Intersection over Union (mIOU). The findings reveal that in the context of roads, our model stands shoulder to shoulder with the baseline. Given that roads are consistently obscured, it poses a challenge for our refinement model to assimilate an accurate road contour. Conversely, for vehicles, our method substantially outperforms the baseline, underscoring the potency of our segment-focused conditioning and viewpoint encoding techniques.

Method	FID↓	Road mIOU↑	Vehicle mIOU↑
BEVGen [4]	25.54	50.20	5.89
UViT [35]	79.22	37.69	9.16
Ours	48.65	47.45	17.70

TABLE I

QUANTITATIVE COMPARISION BETWEEN BEVGEN, UVIT AND OUR METHOD.

D. Ablation Studies

In our research, we carried out ablation studies, specifically honing in on two of our core innovations: the shape refinement process and the street view adaptation technique. The detailed results of these studies can be found in Table. II. The shape refinement process is pivotal in ensuring that map elements are accurately positioned. When the shape within the camera's perspective aligns more semantically, it resonates more effectively with the given prompt. On the other hand, the street view adaptation module plays a crucial role as a style encoder. Its primary function is to make sure that the generated images bear a strong resemblance to those in the training dataset. Moreover, this module greatly assists the image generator by enabling it to achieve proper and accurate orientations for the various map elements.

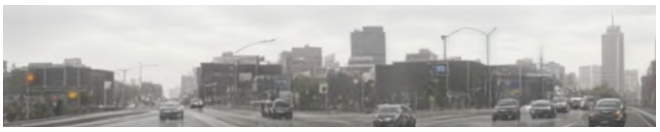

Fig. 7. Creating consistently aligned multi-view images poses a challenge for large pretrained diffusion models, given their typical training on standard datasets.

Method	FID↓	Road mIOU↑	Vehicle mIOU↑
Base diffusion	82.25	46.76	11.82
+ shape refinement	78.13	47.92	15.69
+ view adaptation	48.65	47.45	17.70

TABLE II

ABLATION STUDY ON OUR CORE DESIGN: SHAPE REFINEMENT AND STREET VIEW ADAPTATION.

V. LIMITATIONS AND FUTURE WORKS

In the specific setup we've devised, the integration of multiple cameras has the capability to produce a comprehensive panoramic image that boasts a significantly extended aspect ratio. This is a departure from traditional images and poses a unique challenge. Ideally, the most efficient approach would be to directly generate a panoramic or multi-view image, as this would inherently uphold and maintain the consistency of the view throughout the image. But herein lies the challenge: the vast majority of large-scale image diffusion models available today have been fundamentally trained to cater to standard, more conventional aspect ratios. As a result, these models, when applied to our specific need, fall short. This limitation is clearly demonstrated in Fig. 7. These models face considerable difficulty when tasked with rendering high-quality images that demand a broad and expansive field-of-view.

Recognizing this gap, our future endeavors will be centered around delving deeper and exploring more robust and effective techniques that can leverage these large image diffusion models to seamlessly produce multi-view images.

VI. CONCLUSION

We introduced an innovative framework for generating street-view images from a BEV layout by harnessing the power of a robust, pretrained latent diffusion model. Our methodology integrates view transformation, street-view adaptation, and conditional generation. When compared to baseline models trained from scratch, our model excels in terms of image quality, conditioning precision, and diversity.

REFERENCES

[1] J. Philion and S. Fidler, "Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d," in *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*. Springer, 2020, pp. 194–210.

[2] B. Zhou and P. Krähenbühl, "Cross-view transformers for real-time map-view semantic segmentation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 13 760–13 769.

[3] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Y. Qiao, and J. Dai, "Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers," in *European conference on computer vision*. Springer, 2022, pp. 1–18.

[4] A. Swerdlow, R. Xu, and B. Zhou, "Street-view image generation from a bird's-eye view layout," *arXiv preprint arXiv:2301.04634*, 2023.

[5] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in neural information processing systems*, vol. 30, 2017.

[6] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 10 684–10 695.

[7] M. D. M. Reddy, M. S. M. Basha, M. M. C. Hari, and M. N. Penchalaiah, "Dall-e: Creating images from text," *UGC Care Group I Journal*, vol. 8, no. 14, pp. 71–75, 2021.

[8] C. Saharia, W. Chan, S. Saxena, L. Li, J. Whang, E. L. Denton, K. Ghasemipour, R. Gontijo Lopes, B. Karagol Ayan, T. Salimans *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," *Advances in Neural Information Processing Systems*, vol. 35, pp. 36 479–36 494, 2022.

[9] L. Chen, S. Srivastava, Z. Duan, and C. Xu, "Deep cross-modal audio-visual generation," in *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, 2017, pp. 349–357.

[10] A. Brock, J. Donahue, and K. Simonyan, "Large scale gan training for high fidelity natural image synthesis," *arXiv preprint arXiv:1809.11096*, 2018.

[11] P. Isola, J.-Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 1125–1134.

[12] L. A. Gatys, A. S. Ecker, and M. Bethge, "Image style transfer using convolutional neural networks," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 2414–2423.

[13] L. Ma, X. Jia, Q. Sun, B. Schiele, T. Tuytelaars, and L. Van Gool, "Pose guided person image generation," *Advances in neural information processing systems*, vol. 30, 2017.

[14] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 2337–2346.

[15] H. Dhamo, A. Farshad, I. Laina, N. Navab, G. D. Hager, F. Tombari, and C. Rupprecht, "Semantic image manipulation using scene graphs," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 5213–5222.

[16] Z. Li, J. Wu, I. Koh, Y. Tang, and L. Sun, "Image synthesis from layout with locality-aware mask adaption," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 13 819–13 828.

[17] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*. PMLR, 2015, pp. 2256–2265.

[18] P. Dhariwal and A. Nichol, "Diffusion models beat gans on image synthesis," *Advances in neural information processing systems*, vol. 34, pp. 8780–8794, 2021.

[19] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2021, pp. 12 873–12 883.

[20] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[21] A. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. McGrew, I. Sutskever, and M. Chen, "Glide: Towards photorealistic image generation and editing with text-guided diffusion models," *arXiv preprint arXiv:2112.10741*, 2021.

[22] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, "nuscenes: A multimodal dataset for autonomous driving," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 11 621–11 631.

[23] P. Sun, H. Kretzschmar, X. Dotiwalla, A. Chouard, V. Patnaik, P. Tsui, J. Guo, Y. Zhou, Y. Chai, B. Caine *et al.*, "Scalability in perception for autonomous driving: Waymo open dataset," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 2446–2454.

[24] B. Wilson, W. Qi, T. Agarwal, J. Lambert, J. Singh, S. Khandelwal, B. Pan, R. Kumar, A. Hartnett, J. K. Pontes *et al.*, "Argoverse 2: Next generation datasets for self-driving perception and forecasting," *arXiv preprint arXiv:2301.00493*, 2023.

[25] S. Sengupta, P. Sturgess, L. Ladickỳ, and P. H. Torr, "Automatic dense visual semantic mapping from street-level imagery," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2012, pp. 857–862.

[26] Z. Wang, B. Liu, S. Schulter, and M. Chandraker, "A parametric top-view representation of complex road scenes," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 325–10 333.

[27] S. Schulter, M. Zhai, N. Jacobs, and M. Chandraker, "Learning to look around objects for top-view representations of outdoor scenes," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 787–802.

[28] Y. Liu, T. Yuan, Y. Wang, Y. Wang, and H. Zhao, "Vectormapnet: End-to-end vectorized hd map learning," in *International Conference on Machine Learning*. PMLR, 2023, pp. 22 352–22 369.

[29] Q. Li, Z. Peng, L. Feng, Q. Zhang, Z. Xue, and B. Zhou, "Metadrive: Composing diverse driving scenarios for generalizable reinforcement learning," *IEEE transactions on pattern analysis and machine intelligence*, vol. 45, no. 3, pp. 3461–3475, 2022.

[30] A. Saha, O. Mendez, C. Russell, and R. Bowden, "Translating images into maps," in *2022 International conference on robotics and automation (ICRA)*. IEEE, 2022, pp. 9200–9206.

[31] H. A. Mallot, H. H. Bülthoff, J. Little, and S. Bohrer, "Inverse perspective mapping simplifies optical flow computation and obstacle detection," *Biological cybernetics*, vol. 64, no. 3, pp. 177–185, 1991.

[32] A. Chaurasia and E. Culurciello, "Linknet: Exploiting encoder representations for efficient semantic segmentation," in *2017 IEEE visual communications and image processing (VCIP)*. IEEE, 2017, pp. 1–4.

[33] L. Zhang and M. Agrawala, "Adding conditional control to text-to-image diffusion models," *arXiv preprint arXiv:2302.05543*, 2023.

[34] B. Zhou, H. Zhao, X. Puig, S. Fidler, A. Barriuso, and A. Torralba, "Scene parsing through ade20k dataset," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017, pp. 633–641.

[35] F. Bao, S. Nie, K. Xue, Y. Cao, C. Li, H. Su, and J. Zhu, "All are worth words: A vit backbone for diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 669–22 679.

[36] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2023, pp. 22 500–22 510.

[37] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[38] E. Xie, W. Wang, Z. Yu, A. Anandkumar, J. M. Alvarez, and P. Luo, "Segformer: Simple and efficient design for semantic segmentation with transformers," *Advances in Neural Information Processing Systems*, vol. 34, pp. 12 077–12 090, 2021.

[39] S. M. Jain, "Hugging face," in *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*. Springer, 2022, pp. 51–67.

[40] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International Conference on Machine Learning*. PMLR, 2022, pp. 12 888–12 900.

[41] M. Heusel, H. Ramsauer, T. Unterthiner, B. Nessler, and S. Hochreiter, "Gans trained by a two time-scale update rule converge to a local nash equilibrium," *Advances in neural information processing systems*, vol. 30, 2017.