VQA-Diff: Exploiting VQA and Diffusion for Zero-Shot Image-to-3D Vehicle Asset Generation in Autonomous Driving

Yibo Liu[1,2,⋆], Zheyuan Yang[1,3,*], Guile Wu[1], Yuan Ren[1], Kejian Lin[1], Bingbing Liu[1], Yang Liu[1], and Jinjun Shan[2]

[1] Huawei Noah Ark's Lab, Toronto ON L3R 5A4, Canada
[2] York University, Toronto ON M3J 1P3, Canada
[3] University of Toronto, Toronto ON M5S 1A1, Canada
buaayorklau@gmail.com, andrewzheyuan.yang@mail.utoronto.ca, {guile.wu, yuan.ren3, liu.bingbing, yang.liu9}@huawei.com, jjshan@yorku.ca

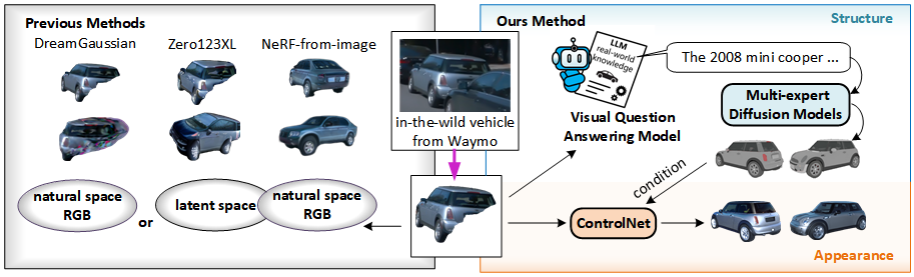


Fig. 1: Previous methods learn to generate novel views using image RGB information in a natural space or a latent space, resulting in poor zero-shot prediction capability to handle in-the-wild vehicle observations with occlusion or tricky viewing angles. Our method, VQA-Diff, tackles this problem by exploiting the robust zero-shot prediction ability of the Visual Question Answering (VQA) model and the rich structure and appearance generation ability of Diffusion Models. This helps to create consistent and photorealistic multi-view renderings of any unseen vehicle in the wild.

Abstract. Generating 3D vehicle assets from in-the-wild observations is crucial to autonomous driving. Existing image-to-3D methods cannot well address this problem because they learn generation merely from image RGB information without a deeper understanding of in-the-wild vehicles (such as car models, manufacturers, *etc.*). This leads to their poor zero-shot prediction capability to handle real-world observations with occlusion or tricky viewing angles. To solve this problem, in this work, we propose VQA-Diff, a novel framework that leverages in-the-wild vehicle images to create photorealistic 3D vehicle assets for autonomous driving. VQA-Diff exploits the real-world knowledge inherited from the Large

⋆ Equal contribution. Work done during an internship with Huawei Noah Ark's Lab.

Language Model in the Visual Question Answering (VQA) model for robust zero-shot prediction and the rich image prior knowledge in the Diffusion model for structure and appearance generation. In particular, we utilize a multi-expert Diffusion Models strategy to generate the structure information and employ a subject-driven structure-controlled generation mechanism to model appearance information. As a result, without the necessity to learn from a large-scale image-to-3D vehicle dataset collected from the real world, VQA-Diff still has a robust zero-shot image-to-novel-view generation ability. We conduct experiments on various datasets, including Pascal 3D+, Waymo, and Objaverse, to demonstrate that VQA-Diff outperforms existing state-of-the-art methods both qualitatively and quantitatively.

Keywords: 3D Vehicle Assets Generation· Visual Question Answering · Diffusion Models

1 Introduction

Photorealistic 3D vehicle asset generation from in-the-wild images is important to autonomous driving as it benefits many downstream tasks, such as training data augmentation and developing sim2real technology [40, 46]. It aims to create novel renderings of the vehicle given a single RGB image captured in the wild. Although some image-to-3D methods [24, 31, 38] may be used for photorealistic 3D vehicle asset generation, these methods only learn to generate novel views from RGB information in the natural space and latent space without deeply understanding the characteristics of real-world vehicles. As a result, they can merely learn from in-the-wild observations of vehicles collected in datasets [7, 44] that cannot cover all car models, manufacturers, occlusion and viewing angles observed in the real world. As illustrated in Fig. 1, when a given image does not fall into the distribution of the training data (*e.g.* an image with occlusion or a tricky viewing angle), these methods cannot achieve effective zero-shot prediction with correct geometry and appearance to unseen observations.

Recently, the Visual Question Answering (VQA) model [19] has shown impressive zero-shot prediction ability thanks to the broad real-world knowledge inherited from the Large Language Model (LLM) [41] and the rich image prior gained from the extensive visual training data [9, 33]. In the application of autonomous driving, by setting questions to ask the VQA model about the car model, manufacturer, production year, and major features, we can obtain a detailed description of the vehicle even from an occluded in-the-wild observation as shown in Fig. 1. However, this zero-shot prediction ability only applies to image-to-text transformation and is thus not directly suitable for image-to-novel-views conversion. On the other hand, Diffusion Models (DMs) [18, 35] are powerful generative models designed for various text-to-image and image-to-image tasks. A pretrained DM has abundant image prior and real-world knowledge to generate photorealistic images based on a prompt or image. Despite the high fidelity of generated images, most DMs cannot control object poses for 3D asset generation. [24] introduces the pose embedding into the original Stable Diffusion [35]

and employs it to rig the object pose in the output. However, as shown in the results of Zero123XL [7, 24] in Fig. 1, relying merely on DMs cannot address the challenge of rendering novel views of in-the-wild vehicles because of out-of-distribution observations.

In this work, we develop a novel generative model, named VQA-Diff, which inherits the merits from both VQA and DMs, to tackle the problem of 3D vehicle asset generation from in-the-wild observations. As illustrated in Fig. 1, instead of directly learning image-to-3D or image-to-multi-view mappings, VQA-Diff utilizes text with encoded broad real-world knowledge inherited from LLM as an intermediary to bridge the VQA model with DMs for 3D asset generation. Considering that the zero-shot prediction of VQA only applies to image-to-text, we design multi-expert DMs to convert text into multi-view structures. This multi-expert DMs design facilitates learning better image quality and vehicle structure compared to using a single DM. Then, we employ the structures given by multi-expert DMs as the controlling condition and the raw image as the driving subject to generate the photorealistic appearance for multi-views with an edge-to-image ControlNet [18, 51]. In this way, VQA-Diff does not have to learn the generation from a large-scale image-to-3D vehicle dataset collected from the real world but maintains robust zero-shot image-to-novel-view generation ability.

The **contributions** of this work are threefold:

- We introduce a novel generative model, dubbed VQA-Diff, for creating photorealistic novel renderings from one single in-the-wild vehicle image. VQA-Diff utilizes the robust zero-shot prediction ability in the VQA model and the rich structure and appearance generation ability in DMs for 3D asset generation.
- We design a multi-expert DMs strategy to learn vehicle structures, which generates multi-views with better image quality and consistency compared to using a single DM.
- We conduct both qualitative and quantitative experiments on three datasets, including Pascal 3D+ [44], Waymo [37], and Objaverse [7], to demonstrate the superiority of our method over the existing state-of-the-art methods.

2 Related Work

Novel views synthesis from multi-view images. Gaussian Splatting (GS) [14] and Neural Radiance Fields (NeRFs) [28] are currently the most popular and widely used solutions for 3D reconstruction from multi-view images. Despite their different rendering strategies, their standard use-case is encoding/learning representation of a scene given multi-view images with associated camera poses and then rendering novel views. Many follow-up works on GS and NeRFs are dedicated to improving rendering quality [2,3,17], reducing training time [29,30], and extending to dynamic scenes [1, 5, 32, 42, 47]. Moreover, there has been some work [27, 43, 49, 53] focusing on reducing the number of required images.

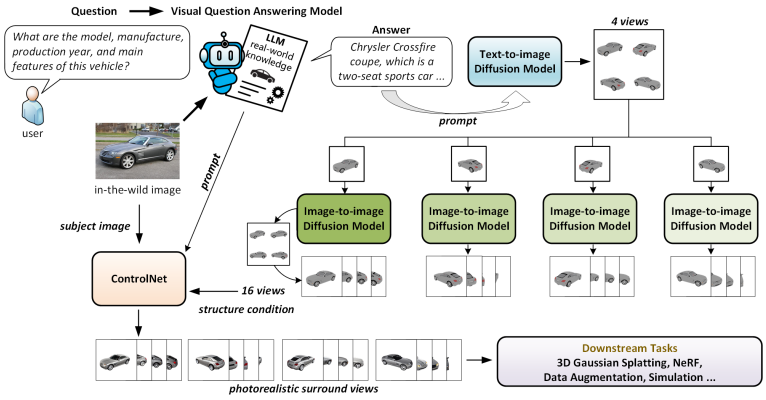


Fig. 2: The framework of the proposed VQA-Diff. The VQA model first generates a prompt containing detailed key information regarding the model, manufacturer, production year, and main features of the vehicle. Then, multi-expert DMs adopt the prompt to create multi-view structures of the vehicle. Finally, the subject-driven structure-controlled generation with ControlNet renders the multi-view structures into photorealistic novel views with controllable poses. The photorealistic novel views can be utilized in various downstream tasks, including the creation of 3D assets with the GS/NeRF representation and training data augmentation. It can also be applied in a simulation environment for autonomous driving.

Novel views synthesis from a single image. A common solution, as adopted in [10, 13, 31, 34, 48, 54], for synthesizing novel views from a single image is to learn the generalizable backbones of NeRFs and GS by encoding each object/scene with a latent code. Among them, NeRF-from-Image (NFI) [31] develops a framework to learn the generation of shape, pose, and appearance of vehicles on Pascal 3D+ [44], which is a dataset containing posed single views of vehicles collected from the real world. Unlike these methods leveraging 3D representation, Zero123 [24] introduces pose embeddings into the original 2D Stable Diffusion (SD) [35], aiming to generate novel views directly from an image-to-image perspective. Some work [23, 26, 38, 39] utilizes the diffusion prior in assisting 3D content generation through Score Distillation Sampling (SDS). For example, DreamGaussian (DG) [38] is the first work introducing diffusion prior from Zero123 [24] into 3D content generation with the GS representation. The aforementioned methods learn the image-to-novel-view mapping merely using the image RGB information in the natural space and latent space. They fail to deeply understand the characteristics of vehicles with real-world knowledge. Thus, their performance is limited to training samples included in datasets [7,44]. Unfortunately, existing datasets [7,44] do not contain sufficient car models, various viewing angles, and complex occlusion cases. The incomplete representation of vehicles in existing datasets will result in a failure of the previous methods in developing their zero-shot prediction capability for novel view rendering. We intend to learn from a deeper understanding of the vehicles to achieve better

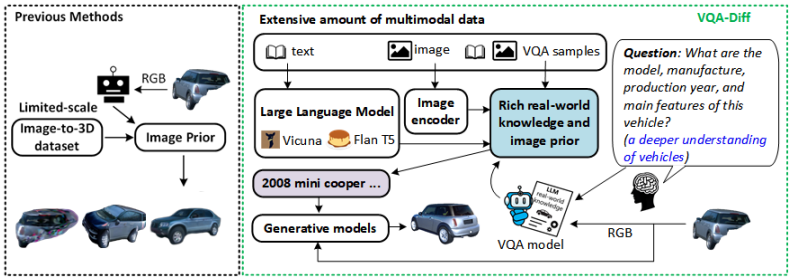


Fig. 3: A comparison of the processes for dealing with the image-to-novel-view problem of previous methods and the proposed VQA-Diff.

novel views rendering performance. Particularly, we propose to transfer the robust zero-shot prediction ability of the VQA model [19] into the generation of novel views in this work.

3 Methodology

The overview framework of VQA-Diff is depicted in Fig. 2. There are three components: the VQA model, multi-expert DMs, and the ControlNet. We render novel views from a single image with three steps, the VQA processing, the structure generation, and the appearance generation. The VQA processing introduced in Sec. 3.1 transforms the in-the-wild observation into a detailed prompt containing key information. Then, the multi-expert DMs proposed in Sec. 3.2 generate the consistent multi-view structures of the vehicle. Finally, as introduced in Sec. 3.3, an edge-to-image ControlNet is utilized to render the multi-view structures into photorealistic novel views.

3.1 VQA Processing

Motivation for VQA processing. Considering the complex structures and appearances of vehicle observations in autonomous driving, the model must have a robust zero-shot prediction ability to render novel views. As aforementioned, it is tricky to develop the zero-shot prediction ability through learning an image-to-novel-view mapping due to the lack of an ideal large-scale dataset. Thus, instead of only focusing on the image modal and developing the zero-shot prediction ability from scratch, we opt to introduce the modal of text and integrate the zero-shot prediction ability of the VQA model into this problem. In particular, the text in the VQA [19, 22] model is encoded with rich real-world knowledge inherited from the LLMs [41, 52], which are trained with a tremendous amount of text samples. Furthermore, the image encoder [9, 33] of the VQA model also gained rich image prior from extensive image samples. The VQA model utilizes the strengths of LLMs and the image encoder through the design of bridging modules. These modules are trained using a comprehensive set of VQA samples (*e.g.*

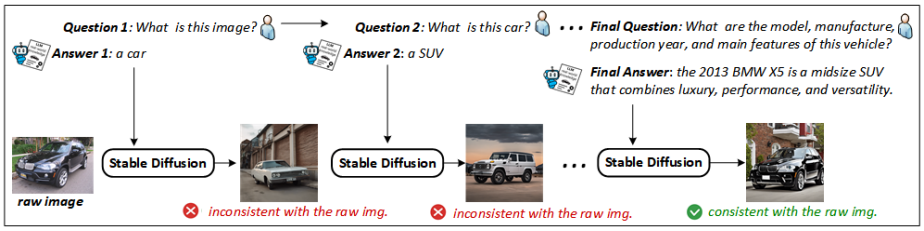


Fig. 4: A illustration of the question design. We tune the question based on the feedback from Stable Diffusion [35].

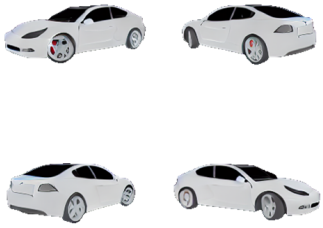
BLIP-2 [19] is trained with 254 million VQA samples). Thus, VQA models can robustly extract information from an image with a deeper understanding than the models that merely gain image prior from scale-limited image-to-3D datasets (*e.g.* Objaverse-XL [7] contain around 10 million objects and Pascal 3D+ [44] only covers 8500 instances). Leveraging the better/deeper image-understanding ability of the VQA models, we can extract useful and detailed vehicle information to boost the novel view generation of vehicles. A comparison of the processes for handling the image-to-novel-view problem of previous methods [24, 31, 38] and the proposed VQA-Diff is presented in Fig. 3.

Question Design. Although the VQA model [19] inherits real-world knowledge from LLM [41], designing the question is important as the answer is the prompt for the following generative models and the prompt is crucial for text-guided generation [11]. Inspired by [11], we design the question based on the feedback of an SD model [35]. As shown in Fig. 4, we first set a simple question "What is this image?" for the VQA model. The given answer is "a car". If we input this rough description into a pretrained SD, the output image is an old-fashioned sedan, which is inconsistent with the raw image. Thus, in the second step, we make the question more specific and ask the VQA model "What car is it?". This time the generated answer is "an SUV". Again we input the answer into the SD and obtain an image of a Jeep-like full-size SUV, which is still inconsistent with the raw vehicle. We keep adjusting the question to make the output image of the SD more and more consistent with the raw image. Finally, the designed question in this work is "What are the model, manufacture, production year, and main features of this vehicle?" With this question, the output of SD is exactly the same vehicle as in the one in the raw image. More discussions on the question design can be found in the supplementary material.

3.2 Multi-expert DMs for Structures Generation

The geometry of a vehicle is determined if the key information including model, manufacturer, production year, and main features are given. Thus, the VQA model deals with the disocclusion of geometry by providing a detailed and accurate description. Since structure and appearance generation are separately handled for novel view rendering, VQA-Diff does not have to learn the genera-

(a) Images of Tesla Model 3 generated by a pretrained Stable Diffusion [35].



(b) The multi-view structures of the Tesla Model 3 generated by our method. Note that our model does not learn the geometry of this car from ShapeNetV2 [6].

Fig. 5: An illustration of transferring real-world knowledge and image prior of a pretrained DM [35] into multi-view structure generation.

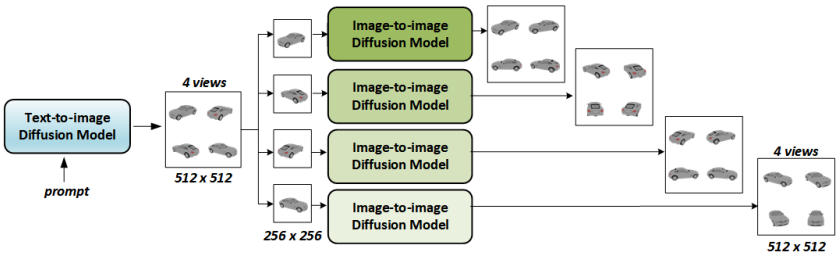
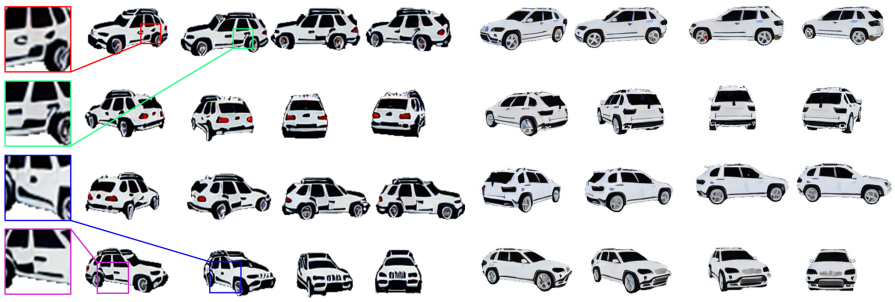


Fig. 6: An illustration of the proposed multi-expert DMs.

tion of geometry and texture simultaneously as in previous methods [24, 31, 38]. Instead, our model only learns to transform the prompt into structures at this stage. Inspired by the good generalizability of the previous work on shape completion of vehicles [8, 25, 50] trained on ShapeNetV2 [6], our model learns vehicle structures from the ShapeNetV2 dataset to transfer the zero-shot prediction of the VQA model to structures.

Motivation for adopting DM. ShapeNetV2 [6] does not include the car models developed in recent years, such as the Tesla Model 3. To increase the variety of our model in asset creation, we utilize a pretrained SD model [35] that has sufficient prior knowledge of vehicle structures. For example, when provided with a text prompt about the Tesla Model 3, the pretrained SD model can generate diverse and accurate structures of the vehicle, as shown in Fig. 5a. However, the lack of control over vehicle poses in the output of SD hinders the utilization of the images as 3D vehicle assets. Thus, we fine-tune a pretrained SD model on the ShapeNetV2 dataset to control vehicle poses in the output while maintaining the model's capability of generating structures for various vehicles.

Design of Multi-expert DMs. Fig. 6 shows an architecture of the multi-expert DMs design for vehicle structure learning. In particular, to transfer the zero-shot prediction of the VQA model, we first train a text-to-image DM that



(a) Multi-view structures generated by one single DM.

(b) Multi-view structures generated by multi-expert DMs.

Fig. 7: Comparison of multi-view structures generated by one single DM and multi-expert DMs. The same prompt regarding a BMW X5 is provided for the two methods.

can generate a 512 × 512 image consisting of four 256 × 256 sub-images based on a text prompt, each of which contains a rendering of the vehicle from a different fixed camera pose. Due to multiple anchor views of these sub-images, the VQA model can learn structures from a wide variety of perspectives for a single car. For each of the sub-images, we employ an image-to-image DM to generate a 512 × 512 image consisting of four 256 × 256 sub-images, resulting in a total of 16 surrounding views of the vehicle, with fixed and controllable camera poses equally spaced around the object. The design enables the model to capture the correlated local structures among the anchor views. Fig. 5b shows the multi-view structure of the Tesla Model 3 generated by our method. As can be seen, despite the absence of this vehicle in ShapeNetV2, the VQA model successfully generates consistent images for the Tesla Model 3 in multiple views.

Multi-expert DMs vs One Single DM. An alternative to the multi-expert DMs is to create the 16 multi-views with one single text-to-image DM. However, we experimentally found that it results in a less effective learning outcome. A visual comparison between a single text-to-image DM and the multi-expert DMs trained with the same experiment setup is presented in Fig. 7, where Fig. 7a exhibits inconsistent structural details generated by a single DM. In addition, the overall image quality of a single DM is inferior to that of the multi-expert DMs shown in Fig. 7b. A quantitative ablation study is presented in Sec. 4.4.

3.3 Appearance Generation

Appearance information extraction. The structure generation will lead to image appearances in ShapeNet style. To render photorealistic appearances resembling the original vehicle, we need to extract the appearance information. However, this task is challenging due to potential occlusion and tricky viewing angles in the raw images. Therefore, we still apply the VQA model [19] to tackle this problem. Yet instead of employing the entire VQA framework, we only utilize its multimodal encoder to encode both the raw image (segmented by

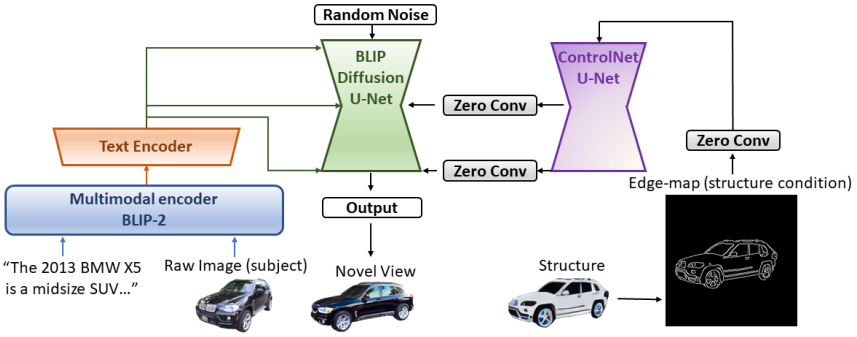


Fig. 8: An illustration of the subject-driven structure-controlled generation.

SAM [16] to eliminate the effect of the background) and the prompt. This allows us to robustly extract the appearance information from in-the-wild observations.
Novel views rendering. To utilize the extracted appearance information, we follow previous subject-image-driven generation methods [18, 36] and transform the output of the multimodal encoder into text embeddings of a text-to-image DM [35]. Thus, the output of the text-to-image DM is controlled by the raw image. Furthermore, considering that only the geometry of the multi-view structures is beneficial, we extract the geometry information through the Canny edge transformation. To control the generation based on the geometry information, we attach an edge-to-image ControlNet [51] to the UNet of the text-to-image DM. In this way, the text-to-image DM [18, 35] generates photorealistic novel views, which are controlled by the VQA result, the raw image, and the vehicle structure prior. Fig. 8 presents an illustration of the appearance generation.

4 Experiments

In this section, we present qualitative and quantitative results on three datasets, Pascal 3D+ [44], Waymo [37], and Objaverse [7]. We curated 20 vehicles with diverse structures and appearances from each of the datasets.
Implementation Details To train the multi-expert DMs, we create 16 renderings for each instance in the car taxonomy of ShapeNetV2 [6]. Each 3D model is normalized into a cube of $[-0.5, 0.5]^3$ and rendered in Blender. The virtual cameras are equally spaced around the object with a distance of 1.5 to the object center and an elevation of 5°. For the best prompt accuracy, we empirically apply BLIP-2 [19] to the first image to generate the prompt. Every fourth image, starting from the first one, is selected as an anchor view. These anchor views are then used with the prompts to train the text-to-image DM. Each anchor view, in conjunction with its three adjacent views, is used to train one image-to-image DM. We adopt the framework of SD v1.5 [35] to build up the multi-expert DMs and fine-tune the pretrained text-to-image SD and image-to-image DM [4] with our datasets. Each DM is trained for 50 epochs with the Adam optimizer [15], using

Fig. 9: Comparison with state-of-the-art methods (NFI [31], Zero123XL [24], and DG [38]) on the Pascal 3D+ dataset [44].

a learning rate of 1e-5 and a batch size of one. For the appearance generation, we adopt the pre-trained BLIP-2 multimodal encoder [19], the text encoder and U-Net of BLIP-Diffusion [18], and an edge-to-image ControlNet [51] to construct the subject-driven structure-controlled generation mechanism.

Metrics. We follow previous work [31, 38] to report FID [12] and the CLIP Similarity [33]. In addition, we report the Image-Text Contrastive (ITC) score proposed in BLIP-2 [19] and the VQA score proposed in [21] as metrics to evaluate the matching between the generated views and the VQA results.

Competitors. We compare our method with three state-of-the-art methods, including NFI [31], Zero123XL [7,24] and DG [38]. Particularly, NFI is pretrained on Pasacal 3D+ [44], and Zero123XL and DG are pretrained on Objaverse [7]. Our method only learned vehicle structures from ShapeNetV2 [6] while inheriting prior knowledge from DMs [18, 35] and the VQA model [19]. Note that the superiority of our method is not derived from training on ShapeNetV2 [6] but from the method itself. Please refer to the supplementary material for details.

4.1 Results on Pascal 3D+

The qualitative and quantitative comparisons of our approach against the state-of-the-art methods are presented in Fig. 9 and Table 1. As seen in Fig. 9, our method visually yields the best quality. Take the first Dodge Ram 1500 pick-up truck as an example, since the observation is obtained from a challenging view,

Table 1: Comparison with the state-of-the-art methods on Pascal 3D+ [44].

Method	ITC score ↑	CLIP similarity ↑	FID ↓	VQA score ↑
Ground Truth	0.401	–	–	–
DG [38]	0.268	0.704	269.24	0.519
Zero123XL [24]	0.270	0.750	193.27	0.506
NFI [31]	0.284	0.784	129.43	0.599
Ours	0.380	0.856	117.49	0.903

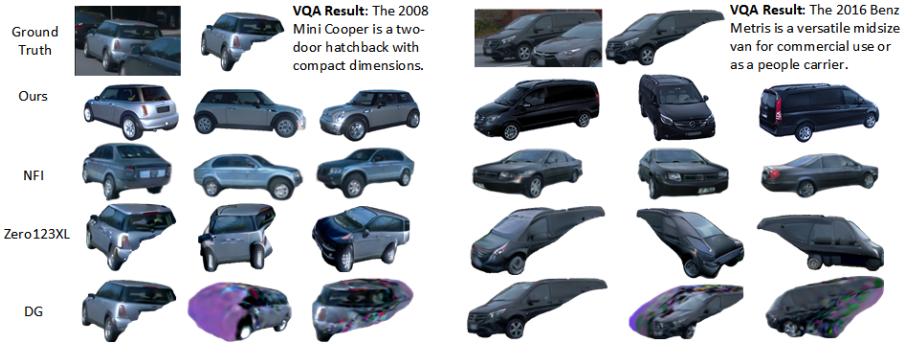


Fig. 10: Comparison with state-of-the-art methods (NFI [31], Zero123XL [24], and DG [38]) on the Waymo dataset [37].

the trunk appears as a small part. As a result, NFI [31] ignores the curve between the body and trunk and mistakes it as an SUV. Zero123XL [7,24] presents a decent appearance at the original viewing angle, but its geometry estimation fails at other views due to a lack of zero-prediction capability. In particular, it takes the trunk as a small part compared to the body thus the geometry rendered from other views looks unrealistic. DG [38] shows perfect rendering at the original view as it employs the raw image to train the GS [14]. However, it utilizes SDS from Zero123XL to reconstruct unseen views, resulting in unsuccessful novel view rendering as Zero123XL fails to provide correct geometry and appearance information. In comparison, our method successfully recognizes the vehicle's model and manufacturer, and thus, it generates the correct geometry of a pick-up truck while rendering photorealistic appearances resembling the raw image. Moreover, as shown in Table 1, the fidelity of our approach outperforms the other state-of-the-art methods quantitatively. Specifically, our method achieves the best ITC score close to the ground truth and the highest VQA score. This indicates that the novel views generated by our approach keep the most semantic information from the perspective of Vision-Language [19, 20, 22].

4.2 Results on Waymo

The qualitative and quantitative results are presented in Fig. 10 and Table 2 respectively. In this experiment, none of the competitors are pretrained on

Table 2: Comparison with the state-of-the-art methods on Waymo [37].

Method	ITC score ↑	CLIP similarity ↑	FID ↓	VQA score ↑
Ground Truth	0.422	–	–	–
DG [38]	0.282	0.819	350.61	0.644
Zero123XL [24]	0.306	0.835	251.59	0.589
NFI [31]	0.298	0.829	188.23	0.194
Ours	0.418	0.840	163.40	0.854

Fig. 11: Comparison with state-of-the-art methods (NFI [31], Zero123XL [24], and DG [38]) on the Objaverse dataset [7].

Waymo [37], ensuring a fair comparison in terms of zero-shot prediction. As seen in Fig. 10, NFI [31] is pretrained on Pascal 3D+ [44], where occlusion exists in some instances. Consequently, NFI is able to generate a vehicle, albeit with incorrect geometry. Zero123XL [7,24] and DG [38] cannot handle occlusion at all. They interpret occluded parts as non-existing and even fail to generate car-like objects. In contrast, our method can generate complete and accurate structures along with high-fidelity appearances, even from occluded observations. As shown in Table 2, our method yields the best performance.

4.3 Results on Objaverse

The qualitative and quantitative results are presented in Fig. 11 and Table 3 respectively. Zero123XL [7, 24] and DG [38] have a certain advantage in this test as they are pretrained on Objaverse [7]. For NFI [31] and our method, it is still a test for zero-shot prediction. As depicted in Fig. 11, the inputs consist of observations from relatively challenging viewing angles compared to a simple side view. Take the Range Rover Evoque as an example, as the color dark orange is not common in Pascal 3D+ [44], NFI fails to create the correct appearance while the structure is complete and close to the raw image. The novel views generated by Zero123XL, while exhibiting a similar appearance to the raw image, suffer from severe structural defects, such as an additional wheel. Since DG [38] relies on Zero123XL to reconstruct unseen views, the novel views given by DG are

Table 3: Comparison with the state-of-the-art methods on Objaverse [7].

Method	ITC score ↑	CLIP similarity ↑	FID ↓	VQA score ↑
Ground Truth	0.429	–	–	–
DG [38]	0.269	0.761	296.39	0.516
Zero123XL [24]	0.319	0.812	175.18	0.366
NFI [31]	0.289	0.772	146.06	0.669
Ours	0.412	0.872	114.75	0.838

Table 4: Ablation study of the multi-expert DMs.

Method	ITC score ↑	CLIP similarity ↑	FID ↓
Single DM (50 epochs)	0.272	0.781	192.55
Single DM (100 epochs)	0.283	0.806	189.04
Multi-expert DMs (50 epochs)	0.333	0.835	122.91

messy. In comparison, our method creates novel views with precise structures and photorealistic appearances akin to the raw image. Consequently, our method outperforms the other state-of-the-art methods as presented in Table 3.

4.4 Ablation Study

We conducted an experiment to compare the effectiveness of multi-expert DMs with a single text-to-image DM, from which all 16 views are generated. We take the ShapeNet renderings of a BMW X5 as the ground truth and provide the corresponding prompt to our multi-expert DMs and the single DM. Table 4 shows the quantitative comparison. As seen, by adopting multi-expert, the model can learn structure generation more efficiently while achieving a better performance. The analysis of this experiment and ablation study on the structure training dataset and ControlNet are provided in the supplementary material.

4.5 Extended Diverse 3D Vehicle Assets Generation

In addition to creating 3D vehicle assets from in-the-wild observations, we introduce complementary solutions, as depicted in Fig. 12, for generating extended diverse vehicle assets to build a comprehensive asset bank. In particular, by collaborating with text-guided style transfer models [4,18], text-to-image generative models [35], and object inpainting models [45], we can alter the colors of vehicles, generate vehicles merely through designed prompts, and create intricate vehicles by providing fine-grained prompts (such as spoiler, wheel, shape and part). More details and results are presented in the supplementary material.

4.6 Limitation

Despite the promising results, our method still has limitation. Our method is elaborately designed for vehicle asset generation, so when extending it to more

Fig. 12: Evaluation on generating extended diverse 3D vehicle assets through collaboration with the text-guided-style-transfer model [4, 18], text-to-image generation model [35], and text and shape guided object inpainting model [45].

generic objects, it may not always perform well (please refer to the supplementary material for the success and failure cases). We conjecture that the main reason is that vehicles are a specific type of object whose structures can be determined by specifying the model, make, and production year, which allows VQA model to provide geometry constraints by generating precise prompts regarding key information. However, it is challenging for the VQA model to constrain the structure of a generic object (such as a teddy bear). With the ongoing development of VQA, extending the proposed method to more objects shows promise.

5 Conclusion

In this work, we propose a new VQA-Diff framework to facilitate novel view rendering from in-the-wild vehicles for autonomous driving. The core idea is to integrate the robust zero-shot prediction ability of the VQA model into novel view rendering. Thus, VQA-Diff can deal with in-the-wild observations with occlusions and a challenging viewing angle. We first design the VQA processing to extract information about the vehicle from a deeper understanding than the previous image-to-3D methods. Then, we develop the multi-expert DMs for learning to generate vehicle structures based on the VQA result from a synthesis dataset. Finally, we apply the subject-driven structure-controlled generation mechanism to transform the multi-view structures into photorealistic novel view renderings. Qualitative and quantitative evaluations on Pascal 3D+, Waymo, and Objaverse show the superiority of our approach over the state-of-the-art methods.

Acknowledgments. The authors gratefully acknowledge the support of MindSpore, CANN, and Ascend AI Processor.

References

1. Bansal, A., Zollhoefer, M.: Neural pixel composition for 3d-4d view synthesis from multi-views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 290–299 (2023)
2. Barron, J.T., Mildenhall, B., Tancik, M., Hedman, P., Martin-Brualla, R., Srinivasan, P.P.: Mip-nerf: A multiscale representation for anti-aliasing neural radiance fields. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 5855–5864 (2021)
3. Barron, J.T., Mildenhall, B., Verbin, D., Srinivasan, P.P., Hedman, P.: Mip-nerf 360: Unbounded anti-aliased neural radiance fields. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5470–5479 (2022)
4. Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023)
5. Cao, A., Johnson, J.: Hexplane: A fast representation for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 130–141 (2023)
6. Chang, A.X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al.: Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012 (2015)
7. Deitke, M., Schwenk, D., Salvador, J., Weihs, L., Michel, O., VanderBilt, E., Schmidt, L., Ehsani, K., Kembhavi, A., Farhadi, A.: Objaverse: A universe of annotated 3d objects. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 13142–13153 (2023)
8. Duggal, S., Wang, Z., Ma, W.C., Manivasagam, S., Liang, J., Wang, S., Urtasun, R.: Mending neural implicit modeling for 3d vehicle reconstruction in the wild. In: Proc. of IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 1900–1909 (2022)
9. Fang, Y., Wang, W., Xie, B., Sun, Q., Wu, L., Wang, X., Huang, T., Wang, X., Cao, Y.: Eva: Exploring the limits of masked visual representation learning at scale. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19358–19369 (2023)
10. Gao, J., Shen, T., Wang, Z., Chen, W., Yin, K., Li, D., Litany, O., Gojcic, Z., Fidler, S.: Get3d: A generative model of high quality 3d textured shapes learned from images. Advances In Neural Information Processing Systems **35**, 31841–31854 (2022)
11. Hao, Y., Chi, Z., Dong, L., Wei, F.: Optimizing prompts for text-to-image generation. Advances in Neural Information Processing Systems **36** (2024)
12. Heusel, M., Ramsauer, H., Unterthiner, T., Nessler, B., Hochreiter, S.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
13. Jang, W., Agapito, L.: Codenerf: Disentangled neural radiance fields for object categories. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12949–12958 (2021)

14. Kerbl, B., Kopanas, G., Leimkühler, T., Drettakis, G.: 3d gaussian splatting for real-time radiance field rendering. ACM Transactions on Graphics **42**(4) (2023)
15. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. In: Proc. of the International Conference on Learning Representations (Poster) (2015)
16. Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., Xiao, T., Whitehead, S., Berg, A.C., Lo, W.Y., et al.: Segment anything. arXiv preprint arXiv:2304.02643 (2023)
17. Lee, B., Lee, H., Sun, X., Ali, U., Park, E.: Deblurring 3d gaussian splatting. arXiv preprint arXiv:2401.00834 (2024)
18. Li, D., Li, J., Hoi, S.: Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. Advances in Neural Information Processing Systems **36** (2024)
19. Li, J., Li, D., Savarese, S., Hoi, S.: BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML (2023)
20. Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022)
21. Lin, Z., Pathak, D., Li, B., Li, J., Xia, X., Neubig, G., Zhang, P., Ramanan, D.: Evaluating text-to-visual generation with image-to-text generation. arXiv preprint arXiv:2404.01291 (2024)
22. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
23. Liu, M., Xu, C., Jin, H., Chen, L., Xu, Z., Su, H., et al.: One-2-3-45: Any single image to 3d mesh in 45 seconds without per-shape optimization. arXiv preprint arXiv:2306.16928 (2023)
24. Liu, R., Wu, R., Van Hoorick, B., Tokmakov, P., Zakharov, S., Vondrick, C.: Zero-1-to-3: Zero-shot one image to 3d object. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 9298–9309 (2023)
25. Liu, Y., Zhu, K., Wu, G., Ren, Y., Liu, B., Liu, Y., Shan, J.: Mv-deepsdf: Implicit modeling with multi-sweep point clouds for 3d vehicle reconstruction in autonomous driving. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8306–8316 (2023)
26. Long, X., Guo, Y.C., Lin, C., Liu, Y., Dou, Z., Liu, L., Ma, Y., Zhang, S.H., Habermann, M., Theobalt, C., et al.: Wonder3d: Single image to 3d using cross-domain diffusion. arXiv preprint arXiv:2310.15008 (2023)
27. Long, X., Lin, C., Wang, P., Komura, T., Wang, W.: Sparseneus: Fast generalizable neural surface reconstruction from sparse views. In: European Conference on Computer Vision. pp. 210–227. Springer (2022)
28. Mildenhall, B., Srinivasan, P.P., Tancik, M., Barron, J.T., Ramamoorthi, R., Ng, R.: Nerf: Representing scenes as neural radiance fields for view synthesis. Communications of the ACM **65**(1), 99–106 (2021)
29. Müller, T., Evans, A., Schied, C., Keller, A.: Instant neural graphics primitives with a multiresolution hash encoding. ACM transactions on graphics (TOG) **41**(4), 1–15 (2022)
30. Niedermayr, S., Stumpfegger, J., Westermann, R.: Compressed 3d gaussian splatting for accelerated novel view synthesis. arXiv preprint arXiv:2401.02436 (2023)
31. Pavllo, D., Tan, D.J., Rakotosaona, M.J., Tombari, F.: Shape, pose, and appearance from a single image via bootstrapped radiance field inversion. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4391–4401 (2023)

32. Pumarola, A., Corona, E., Pons-Moll, G., Moreno-Noguer, F.: D-nerf: Neural radiance fields for dynamic scenes. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10318–10327 (2021)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
34. Rebain, D., Matthews, M., Yi, K.M., Lagun, D., Tagliasacchi, A.: Lolnerf: Learn from one look. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 1558–1567 (2022)
35. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10684–10695 (2022)
36. Ruiz, N., Li, Y., Jampani, V., Pritch, Y., Rubinstein, M., Aberman, K.: 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. pp. 22500–22510 (2023)
37. Sun, P., Kretzschmar, H., Dotiwalla, X., Chouard, A., Patnaik, V., Tsui, P., Guo, J., Zhou, Y., Chai, Y., Caine, B., et al.: Scalability in perception for autonomous driving: Waymo open dataset. In: Proc. of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2446–2454 (2020)
38. Tang, J., Ren, J., Zhou, H., Liu, Z., Zeng, G.: Dreamgaussian: Generative gaussian splatting for efficient 3d content creation. arXiv preprint arXiv:2309.16653 (2023)
39. Tang, J., Wang, T., Zhang, B., Zhang, T., Yi, R., Ma, L., Chen, D.: Make-it-3d: High-fidelity 3d creation from a single image with diffusion prior. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV). pp. 22819–22829 (October 2023)
40. Wang, J., Manivasagam, S., Chen, Y., Yang, Z., Bârsan, I.A., Yang, A.J., Ma, W.C., Urtasun, R.: Cadsim: Robust and scalable in-the-wild 3d reconstruction for controllable sensor simulation. In: 6th Annual Conference on Robot Learning (2022), https://openreview.net/forum?id=Mp3Y5jd7rnW
41. Wei, J., Bosma, M., Zhao, V., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. In: International Conference on Learning Representations
42. Wu, G., Yi, T., Fang, J., Xie, L., Zhang, X., Wei, W., Liu, W., Tian, Q., Wang, X.: 4d gaussian splatting for real-time dynamic scene rendering. arXiv preprint arXiv:2310.08528 (2023)
43. Wu, R., Mildenhall, B., Henzler, P., Park, K., Gao, R., Watson, D., Srinivasan, P.P., Verbin, D., Barron, J.T., Poole, B., et al.: Reconfusion: 3d reconstruction with diffusion priors. arXiv preprint arXiv:2312.02981 (2023)
44. Xiang, Y., Mottaghi, R., Savarese, S.: Beyond pascal: A benchmark for 3d object detection in the wild. In: IEEE winter conference on applications of computer vision. pp. 75–82. IEEE (2014)
45. Xie, S., Zhang, Z., Lin, Z., Hinz, T., Zhang, K.: Smartbrush: Text and shape guided object inpainting with diffusion model. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 22428–22437 (2023)
46. Yang, Z., Manivasagam, S., Chen, Y., Wang, J., Hu, R., Urtasun, R.: Reconstructing objects in-the-wild for realistic sensor simulation. ICRA (2023)
47. Yang, Z., Yang, H., Pan, Z., Zhu, X., Zhang, L.: Real-time photorealistic dynamic scene representation and rendering with 4d gaussian splatting. arXiv preprint arXiv:2310.10642 (2023)

48. Yang, Z., Liu, Y., Wu, G., Cao, T., Ren, Y., Liu, Y., Liu, B.: Learning effective nerfs and sdfs representations with 3d generative adversarial networks for 3d object generation: Technical report for iccv 2023 omniobject3d challenge. arXiv preprint arXiv:2309.16110 (2023)
49. Yu, A., Ye, V., Tancik, M., Kanazawa, A.: pixelnerf: Neural radiance fields from one or few images. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 4578–4587 (2021)
50. Yu, X., Rao, Y., Wang, Z., Liu, Z., Lu, J., Zhou, J.: Pointr: Diverse point cloud completion with geometry-aware transformers. In: Proceedings of the IEEE/CVF international conference on computer vision. pp. 12498–12507 (2021)
51. Zhang, L., Rao, A., Agrawala, M.: Adding conditional control to text-to-image diffusion models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 3836–3847 (2023)
52. Zhang, S., Roller, S., Goyal, N., Artetxe, M., Chen, M., Chen, S., Dewan, C., Diab, M., Li, X., Lin, X.V., et al.: Opt: Open pre-trained transformer language models. arXiv preprint arXiv:2205.01068 (2022)
53. Zhou, Z., Tulsiani, S.: Sparsefusion: Distilling view-conditioned diffusion for 3d reconstruction. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 12588–12597 (2023)
54. Zou, Z.X., Yu, Z., Guo, Y.C., Li, Y., Liang, D., Cao, Y.P., Zhang, S.H.: Triplane meets gaussian splatting: Fast and generalizable single-view 3d reconstruction with transformers. arXiv preprint arXiv:2312.09147 (2023)