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

Supplementary Material

A. Overview

This material includes both supplementary quantitative and qualitative experimental results, along with additional information such as implementation details, and discussions to complement the main paper.

B. Extended 3D Vehicle Assets Generation

In this section, we introduce solutions to generate extended 3D vehicle assets with the proposed method.

Cooperating with Text-guided Style Transfer Models. While the proposed method effectively addresses the challenge of rendering novel views from in-the-wild observations, it is also desired to generate diverse appearances to construct a comprehensive 3D asset bank. Thus, we propose to use the text-guided style transfer models [2, 5] to transform the paint of the raw vehicle, and then generate extended 3D vehicle assets through the proposed method. An illustration of this solution is presented in Fig. 1.

Cooperating with Text-to-image Models. We cannot guarantee that the data collected on the road covers all the wanted car models. Hence, to build a complete 3D vehicle assets bank, we also want to generate vehicle assets without requiring in-the-wild observations. Therefore, we propose to use text-to-image generative models, such as Stable Diffusion (SD) [11], to produce reference images for VQA-Diff directly. Thus, the proposed method can generate 3D vehicle assets based on prompts. An illustration of this solution is depicted in Fig. 2.

Cooperating with Text and Shape Guided Object Inpainting Models. VQA-Diff can cooperate with text and shape-guided object inpainting models [15] to generate intricate vehicles by providing fine-grained prompts, such as

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

Fig. 1: We propose applying text-guided style transfer models [2,5] to alter the vehicle's paint in the raw image, utilizing prompts like "painted brown". Then, we use the proposed method to generate photorealistic renderings of the same vehicle model with different paintings.

spoiler, wheel, shape, and part. The results are presented in Fig. 3. In particular, we adopt the inpainting model to change the details of the raw vehicle by providing a fine-grained prompt, such as adding a rear spoiler, and specifying the desired inpainting region. Simultaneously, we modify the VQA result with fine-grained words (*e.g.*, "with a rear spoiler") to guide the generation of the desired geometry with VQA-Diff. Finally, the newly obtained reference image and the multi-views are utilized to create intricate vehicles with desired details.

Remark. Note that the reference images shown in Fig. 1 and Fig. 2, while photorealistic, are not 3D vehicle assets, as their object poses are not controllable. In contrast, the renderings created by our method have controlled object poses. In addition, we want to emphasize that generating vehicle assets from real-world observations holds particular significance in constructing a simulation environment for autonomous driving [14, 16]. Because the appearances of vehicles generated in this manner closely resemble those found in data collected on the road.

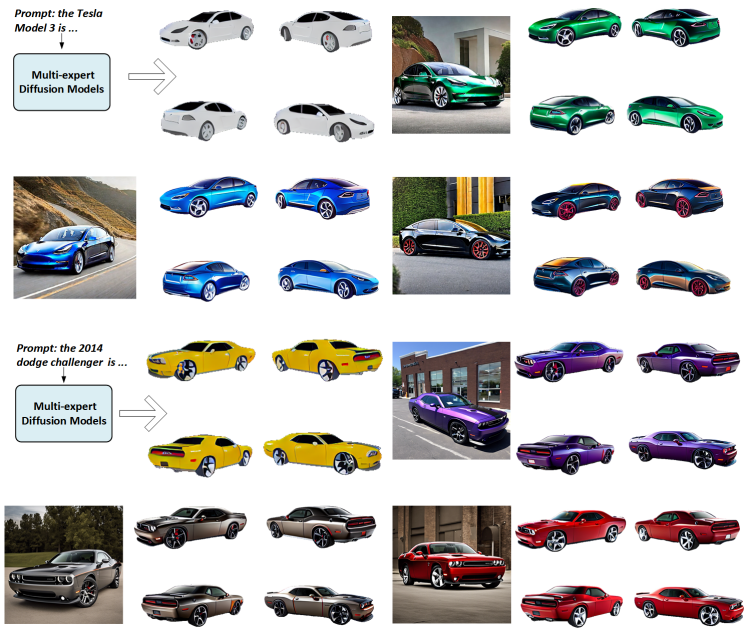

Fig. 2: We propose applying text-to-image models, such as Stable Diffusion [11], to generate reference images for our method. Consequently, our approach can produce photorealistic renderings of vehicles without relying on in-the-wild observations. In this example, the reference images are generated by Stable Diffusion [11] from the manually designed prompts. Note that the multi-expert DMs did not learn the structures of a Tesla Model 3 as ShapeNetV2 [3] does not encompass this model.

C. Effect of Number of DMs

In this section, we explore the effect of adopting different DM schemes as a supplementary experiment to the main paper. Specifically, we compare the proposed multi-expert DMs with a single DM. In the single DM scheme, the model learns to generate all desired multi-view structures in one image. "Single DM (16)" and "Single DM (9)" imply that there are either 16 views or 9 views within a single image. All the models are obtained through fine-tuning a pretrained Stable Diffusion v1.5 [11] on the ShapeNetV2 [3] for 50 epochs with a learning rate of 1e-5 and a batch size of one. We input the same prompt regarding a BMW X5 into all the models and compare the generated images with the multi-view images of the BMW X5 in the training data. The qualitative and quantitative comparisons are presented in Fig. 4 and Table 1. As seen in Table 1, the multi-expert DMs yield better performance than the single DM schemes. Moreover, as pointed out by the red arrows in Fig. 4, there are inconsistent parts in the views generated by the single DM schemes. We provide our analysis in the following. The SD Model [11] is a Latent Diffusion Model (LDM). The learning of the LDM is supervised by both text embeddings (derived from the prompts) and feature maps (extracted

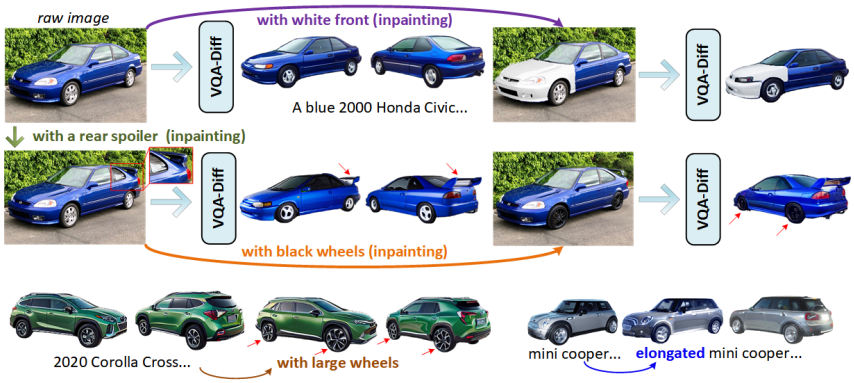

Fig. 3: We propose applying text and shape guided object inpainting models, such as SmartBrush [15], to generate intricate vehicles with desired details. Utilizing fine-grained prompts, the inpainting model creates a new reference image with desired details. Then, VQA-Diff generates intricate vehicle assets based on the new reference image and fine-grained prompt.

Table 1: Ablation study of the multi-expert DMs.

Method	ITC score ↑	CLIP similarity ↑	FID ↓
Single DM (16)	0.272	0.781	192.55
Single DM (9)	0.311	0.811	150.31
Multi-expert DMs	0.333	0.835	122.91

from the raw images). Given that the prompts provided to the models are the same, we analyze the $4 \times 64 \times 64$ feature map encoded by VAE [10]. Considering that the resolution of the feature map is fixed in VAE, as seen in Fig. 5, the representation quality of a single view degrades as the number of views increases. This indicates that the single DM has to learn from a supervision with a worse quality and the performance will be inferior to our method.

D. Impact of Structure Training Dataset and ControlNet

In this section, we study the impact of the structure training dataset and ControlNet. We compare all the methods on Waymo [12]. First, to demonstrate

Table 2: Ablation study of the structure training dataset and ControlNet.

Method	ITC score ↑	CLIP similarity ↑	FID ↓
NFI [9] (ShapeNet [3])	0.210	0.744	428.35
NFI [9] (ShapeNet [3] + ControlNet [5, 17])	0.261	0.761	267.47
Ours (ShapeNet [3])	0.403	0.831	186.98
Ours (ShapeNet [3] + ControlNet [5, 17])	0.418	0.840	163.40

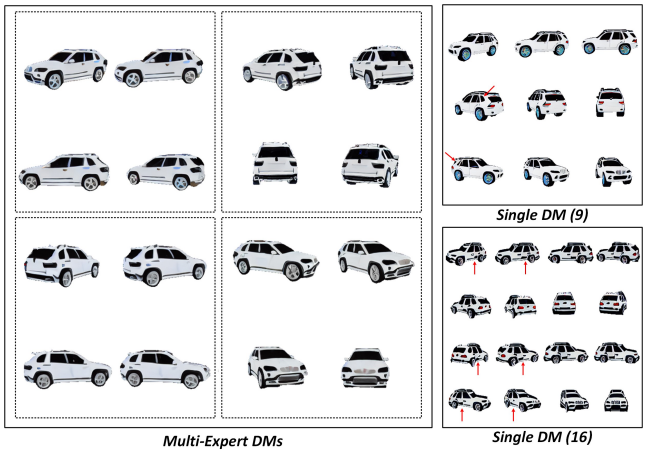

Fig. 4: A visual comparison of the generated multi-view structures using different DM schemes.

that the superiority of VQA-Diff is not brought by training on ShapeNetV2 [3] but the method itself, we compare the NFI [9] trained on ShapeNetV2, which is denoted by NFI (ShapeNet), with our method without adopting ControlNet, which is denoted by Ours (ShapeNet). As presented in Table 2, although both models are trained on ShapeNetV2, our method outperforms NFI as our method generates the correct geometry using the VQA result. The appearance of novel views in our method is created by ControlNet [5, 17]. Thus, for a fair and comprehensive comparison, we also apply the ControlNet to the result of NFI (ShapeNet), which is represented by NFI (ShapeNet+ControlNet). As seen in Table 2, although the performance is boosted by applying the ControlNet, NFI (ShapeNet+ControlNet) is still inferior to Ours (ShapeNet) because the geometry is wrong. Fig. 6 shows the visual comparison.

E. Impact of Question Design and VQA Model

Fig. 7 shows the evaluation of the generated images during the question design process. In particular, we employ the CLIP txt2txt score [10, 13] to evaluate the matching between the VQA prediction and the ground truth caption. In particular, the original image is used to update the prompt, and we compute the txt2txt score each time we obtain an answer from the VQA model. As seen, the score increases and the image quality improves until reaching convergence, where the txt2txt score is high and the generated image refers to the same vehicle as the raw image. Compared to the question design solution proposed in VQA-Diff, one might consider an alternative method: iteratively using SD outputs as inputs to VQA, where the SD outputs are used to update the prompt. We present the results of this solution in Fig. 8. As seen, from the second step, SD predicts wrong images as inputs to VQA, so using SD outputs as inputs to VQA yields worse

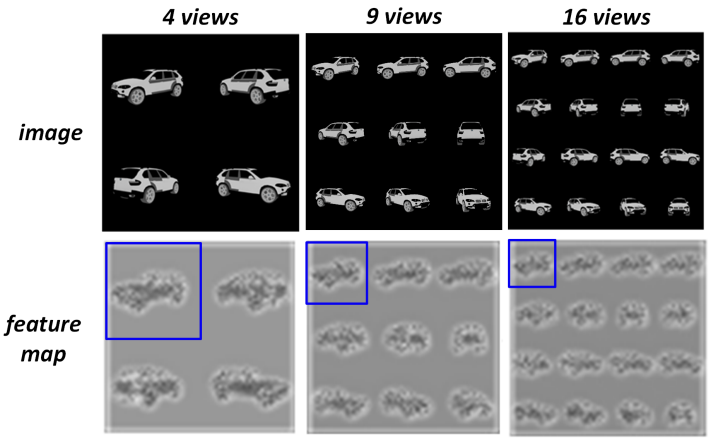

Fig. 5: An illustration of how the number of views affects the feature map. The representations of an individual view are denoted by blue boxes.

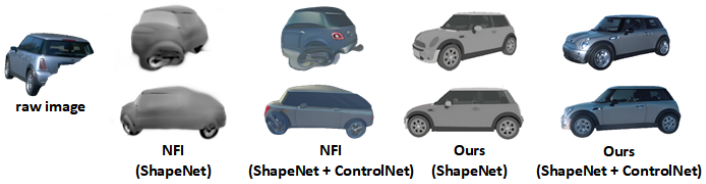

Fig. 6: A visual comparison of NFI [9] and our method trained on ShapeNetV2 [3] w/ and w/o ControlNet [5, 17].

score and image quality. The effect of adopting different VQA models in VQA-Diff is illustrated in Fig. 9. Although there are multiple SOTA VQA models, such as LLaVA [7], Qwen VL [1], and BLIP-2 [6], we chose BLIP-2 because we found it capable of recognizing more vehicle cues. As seen in Fig. 9, LLaVa fails to provide key information while QWen VL fails to provide production year, resulting in inferior renderings, although they might outperform BLIP-2 in other VQA tasks. We conjecture this is attributed to pretraining data of VQA models.

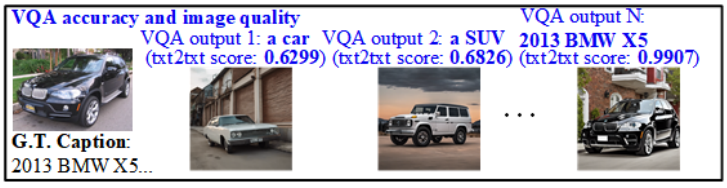

Fig. 7: Evaluation of the generated images during our question design process.

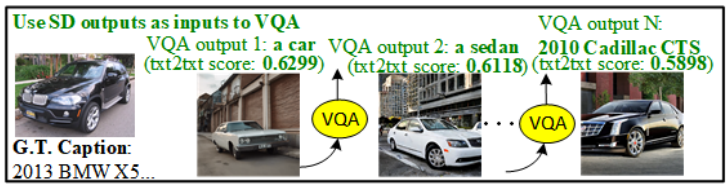

Fig. 8: An illustration depicting the effect of an alternative question design solution.

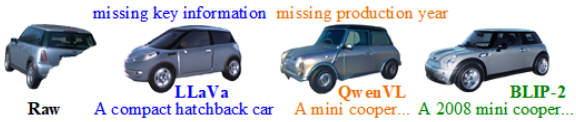

Fig. 9: Comparison of adopting different VQA models [1, 6, 7] in VQA-Diff.

Fig. 10: Success and failure cases of other categories.

F. Application to More Generic Objects

As mentioned in the main paper, extending the proposed method to generic objects is challenging. We present success and failure cases of other categories in Fig. 10. For the success case, as shown in Fig. 10, VQA-Diff can be extended to

generate some motorbike assets and shows better results compared with those generated by Zero123XL [4, 8]. However, we also found that VQA predictions are less robust for motorbikes compared to cars, which we conjecture is due to the pretraining data of VQA. Moving on to the failure case, as shown in Fig. 10, our approach does not show good result for the teddy bear generation. In particular, we tried to follow the same prompt engineering process for dealing with vehicles to obtain geometry constraints via VQA. Unfortunately, VQA cannot provide key information about the teddy bear (such as model, manufacturer, etc.), so it cannot constrain fine-grained structures. Consequently, although the final description of the teddy bear includes many details and the color of the final rendering is close to the raw image, the structure of the final rendering is still inconsistent with the raw image.

References

1. Bai, J., Bai, S., Yang, S., Wang, S., Tan, S., Wang, P., Lin, J., Zhou, C., Zhou, J.: Qwen-vl: A frontier large vision-language model with versatile abilities. arXiv preprint arXiv:2308.12966 (2023)
2. 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)
3. 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)
4. 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)
5. 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)
6. 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)
7. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems **36** (2024)
8. 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)
9. 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)
10. 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)
11. 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)

12. 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)
13. 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)
14. 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
15. 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)
16. Yang, Z., Manivasagam, S., Chen, Y., Wang, J., Hu, R., Urtasun, R.: Reconstructing objects in-the-wild for realistic sensor simulation. ICRA (2023)
17. 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)