

# Supplementary Materials: CustomNet: Object Customization with Variable-Viewpoints in Text-to-Image Diffusion Models

Anonymous Authors

## 1 MORE DETAILS AND DISCUSSIONS

### 1.1 Additional Dataset Details

In Fig. 1, we showcase some data samples from two distinct dataset construction pipelines. Pipeline (a) (the upper part of the figure) is the synthetic datasets construction pipeline, and pipeline (b) (the bottom) is the real-world datasets construction pipeline. The synthetic datasets are constructed by simply blending the objects and background images. Diverse 3D assets are rendered into multi-view images and are utilized to construct such synthetic data pairs. However, training our model only with synthetic datasets will result in inharmonious customization since the synthesized data is inharmonious and unnatural. On the contrary, the real-world dataset construction pipeline uses the Zero-1-to-3 to generate an object with a novel view from a natural image in the real world. Thus the learning targets are naturally enough for harmonious customization. With our dataset construction pipelines, we can train the unified framework in an end-to-end manner and achieve more harmonious customization.

We perform joint training with both synthetic and real-world data with sampling ratio 5% : 95%, respectively. Though the ablation experiments in the main paper show that training with only synthetic data will generate inharmonious results, we still use 5% synthetic datasets during training due to their superior 3D consistency between the object and the target customized image. By training with these datasets, the model enhances its comprehension of the complex 3D geometry of objects.

It is worth mentioning that our datasets use text prompts as conditions. These text prompts may also contain some description of the objects, which is some high-level semantic information that also helps with model performance. In Toss [5], they introduce text prompts to the task of novel view synthesis (NVS) from just a single RGB image, which also demonstrates the benefits of the additional textual description in improving the generation quality.

### 1.2 More Style Control Details

We further fine-tune the text branch of our CustomNet to enhance the control over textual style. More specifically, the instruction-based editing dataset proposed in InstructPix2Pix [1] is utilized to achieve this goal. This dataset provides paired data in a specific format: Each pair consists of an original image, a prompt (which often serves as a style guidance for editing the image), and the corresponding edited image. This format is particularly useful as it allows us to align the text prompts with the desired image editing results. The edited images serve as targets during this fine-tuning process, guiding the model to learn the desired style transformations. As shown in Figs. 2 3 and 4, this fine-tuning process significantly enhances the textual capabilities of our model. As a result, CustomNet can control the style of the objects in the image, and simultaneously manage their viewpoints.



(a) Data samples from synthetic dataset construction pipeline.



(b) Data samples from real-world dataset construction pipeline.

**Figure 1: Data samples from our constructed synthetic datasets and real-world datasets. (a) shows data samples from the synthetic dataset construction pipeline. The synthetic datasets simply blend the objects and background images. It can utilize diverse multi-view 3D datasets but result in inharmonious results. (b) shows data samples from real-world dataset construction pipelines. It uses the Zero-1-to-3 to generate the object’s multi-view conditions, as the target images come from the real world, which is naturally harmonious.**

### 1.3 Difference between Different Customization Methods

To achieve image customization with diffusion models, some encoder-based methods have been developed to achieve efficient zero-shot customization. Usually, the reference object image is first encoded into embedding, which is crucial for extracting the visual information within the image. Initial attempts, such as Paint-by-Example [7] and GLIGEN [4], utilized the pretrained CLIP image encoder to extract single visual embedding. However, this approach often fails when dealing with input objects that possess complex appearances. To address this, ELITE [6] introduced multi-layer features of CLIP for local feature enhancements. Meanwhile, BLIP-Diffusion [3] took a different approach by first training a Q-former. This Q-former extracts the image embedding sequence from the object image through multimodal representation learning. IP-Adapter [8], on the other hand, opts for a simpler method, projecting the CLIP image embedding into a sequence of features to train the diffusion model with an image construction target. While these subsequent methods

have made strides in improving object identity preservation, they still fail to keep object identity in customization compared with our method.

Different from above methods, (1) our CustomNet addresses identity preservation by concatenating the VAE latent of an object with the noisy latent in the channel dimension. This approach demonstrates a strong capability for identity preservation. (2) Furthermore, we utilize viewpoints as an additional condition to guide the image toward diverse generations. (3) Moreover, we design dataset construction pipelines to handle complex real-world images. Through both quantitative and qualitative experiments, our CustomNet outperforms other encoder-based method and achieves harmonious results.

#### 1.4 Classifier-free Guidance Details

We apply classifier-free guidance concerning two conditions: image ( $x, R, x_L$ , which is related to object viewpoint and location) and text ( $T$  for textual description). For simplification, we use  $C_I$  and  $C_T$  to represent the two conditions respectively. When sampling, we set two guidance scales ( $S_I, S_T$ ) to control their influence respectively as follows:

$$\begin{aligned} \hat{\epsilon}_\theta(z_t, C_I, C_T) = & \epsilon_\theta(z_t, \emptyset_I, \emptyset_T) \\ & + S_I \cdot (\epsilon_\theta(z_t, C_I, \emptyset_T) - \epsilon_\theta(z_t, \emptyset_I, \emptyset_T)) \\ & + S_T \cdot (\epsilon_\theta(z_t, C_I, C_T) - \epsilon_\theta(z_t, C_I, \emptyset_T)) \end{aligned} \quad (1)$$

where  $\emptyset_*$  is set the  $*$  condition to null. The conditional probability of our model is as follows:

$$P(z|C_I, C_T) = \frac{P(z, C_I, C_T)}{P(C_I, C_T)} = \frac{P(C_T|C_I, z)P(C_I|z)P(z)}{P(C_I, C_T)} \quad (2)$$

The log probability is s:

$$\begin{aligned} \log(P(z|C_I, C_T)) = & \log(P(C_T|C_I, z)) + \log(P(C_I|z)) \\ & + \log(P(z)) - \log(P(C_T, C_I)) \end{aligned} \quad (3)$$

The derivative of the log probability is the score [2] of the diffusion model:

$$\begin{aligned} \nabla_z \log(P(z|C_I, C_T)) = & \nabla_z \log(P(z)) \\ & + \nabla_z \log(P(C_I|z)) \\ & + \nabla_z \log(P(C_T|C_I, z)) \end{aligned} \quad (4)$$

## 2 MORE QUANTITATIVE AND QUALITATIVE RESULTS

### 2.1 Quantitative Comparison to Inpainting Methods

We further report the metrics, including **DINO-I**, **CLIP-I** following BLIP-Diffusion [3], to conduct a quantitative comparison of various inpainting methods. As Tab. 1 shows, CustomNet achieves the best metrics when given a referenced object image and background image compared with other inpainting methods.

### 2.2 More Results of CustomNet

We showcase the application of our model to real-world object customization tasks. These results, which are presented in Figures 2, 3, and 4, provide a comprehensive view of the model's capabilities. In these figures, a diverse range of objects have been customized using our model, which demonstrates the model's ability to handle

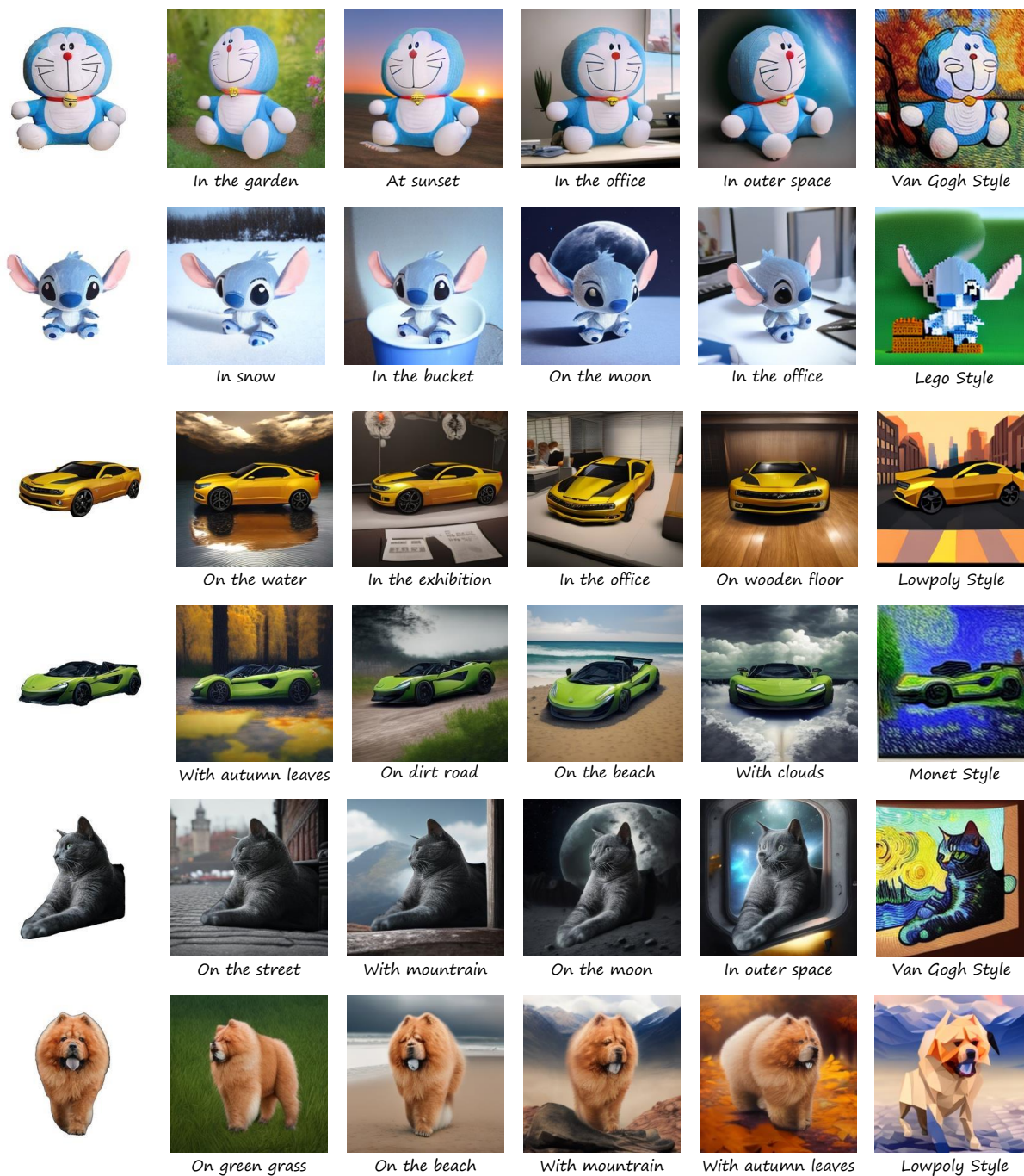
**Table 1: Quantitative Comparison. We compute DINO-I, CLIP-I following [3] to compare different inpainting models.**

Method	DINO-I $\uparrow$	CLIP-I $\uparrow$
Paint-by-Example [7]	0.5070	0.7234
GLIGEN [4]	0.5242	0.7489
<b>CustomNet (Ours)</b>	<b>0.7603</b>	<b>0.8107</b>

complex real-world scenarios. These results show the potential of our model for practical customization applications.

## REFERENCES

- [1] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.
- [2] Aapo Hyvärinen and Peter Dayan. 2005. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research* 6, 4 (2005).
- [3] Dongxu Li, Junnan Li, and Steven CH Hoi. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. *arXiv:2305.14720* (2023).
- [4] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. 2023. Gligen: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 22511–22521.
- [5] Yukai Shi, Jianan Wang, He Cao, Boshi Tang, Xianbiao Qi, Tianyu Yang, Yukun Huang, Shilong Liu, Lei Zhang, and Heung-Yeung Shum. 2023. Toss: High-quality text-guided novel view synthesis from a single image. *arXiv preprint arXiv:2310.10644* (2023).
- [6] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. 2023. Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation. *arXiv preprint arXiv:2302.13848* (2023).
- [7] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. 2023. Paint by example: Exemplar-based image editing with diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recog.*
- [8] Hu Ye, Jun Zhang, Sibio Liu, Xiao Han, and Wei Yang. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721* (2023).



**Figure 2: More real-world object customized results of the proposed CustomNet.**



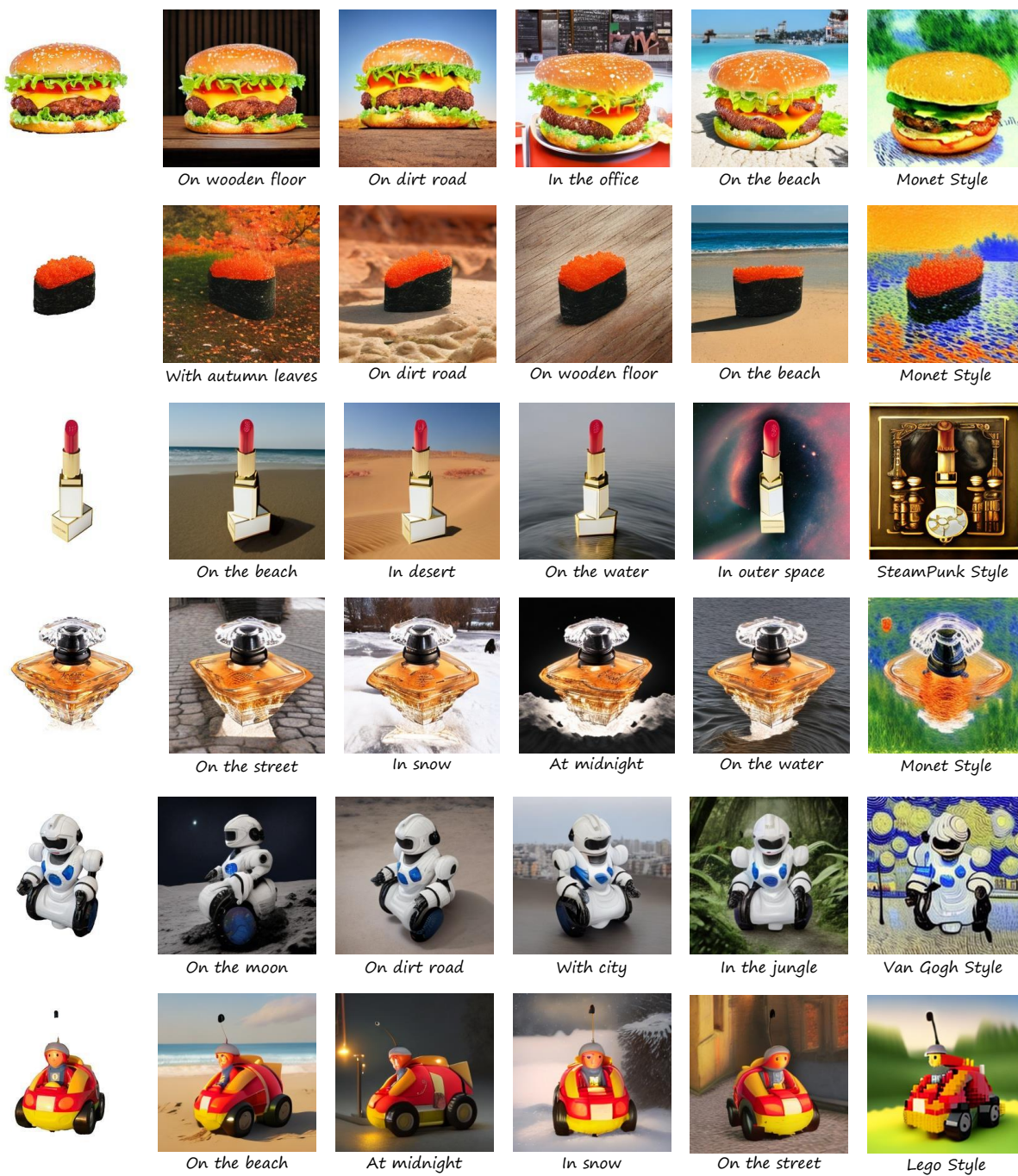
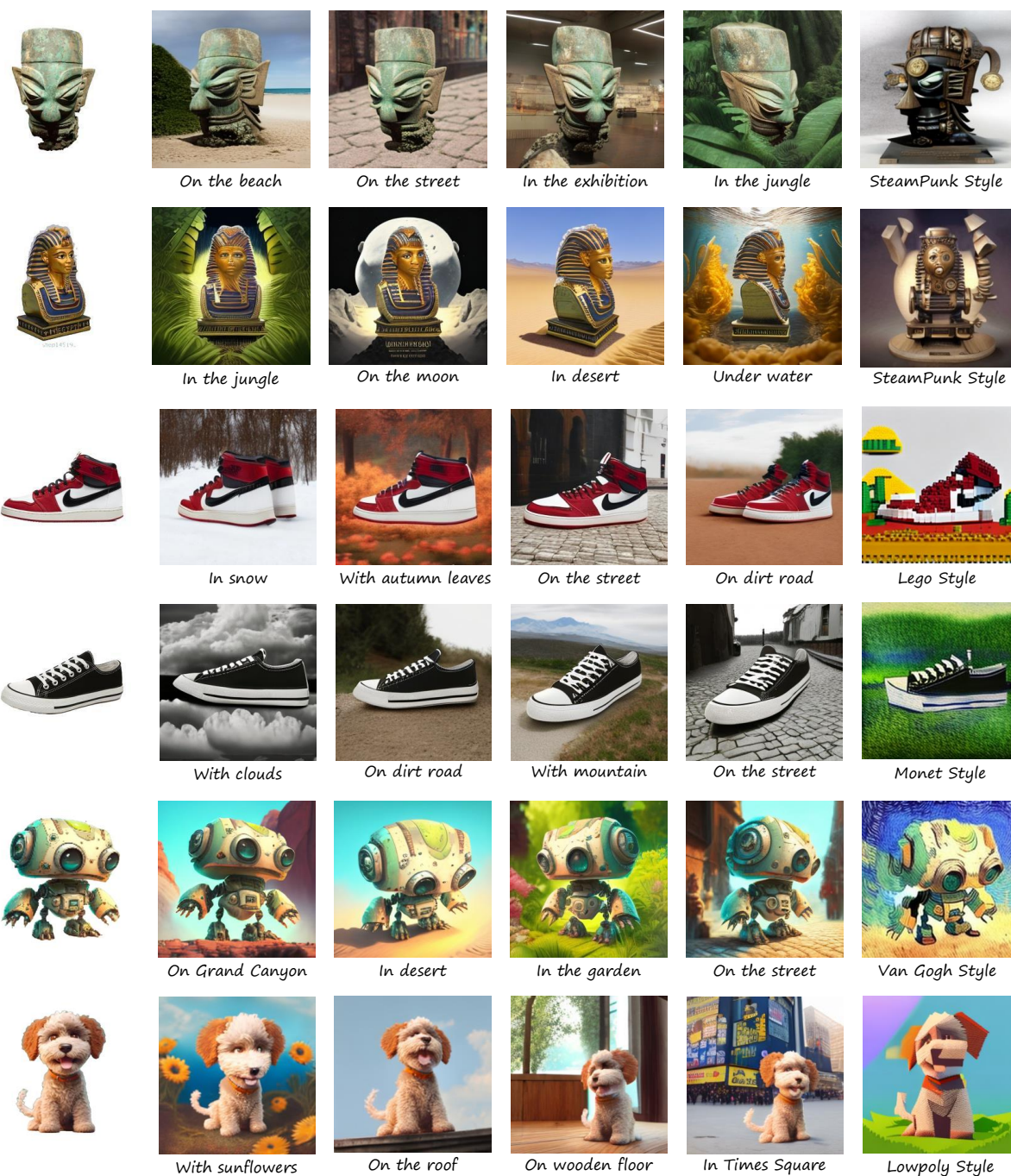


Figure 3: More real-world object customized results of the proposed CustomNet.



**Figure 4: More real-world object customized results of the proposed CustomNet.**