# Supplementary Material of CustAny: Customizing Anything from A Single Example

Lingjie Kong[1,*], Kai Wu[2,*], Chengming Xu[2], Xiaobin Hu[2], Wenhui Han[2], Jinlong Peng[2]
Donghao Luo[2], Mengtian Li[3], Jiangning Zhang[2], Chengjie Wang[2], Yanwei Fu[1,†]
[1]School of Data Science, Fudan University [2]Youtu Lab, Tencent
[3]Shanghai Film Academy, Shanghai University
ljkong22@m.fudan.edu.cn wukaiwork@outlook.com chengmingxu@tencent.com
xiaobin.hu@tum.de whhan18@gmail.com jeromepeng@tencent.com
michaelluo@tencent.com mtli@shu.edu.cn 186368@zju.edu.cn
jasoncjwang@tencent.com yanweifu@fudan.edu.cn

## 1. Additional Analysis about MC-IDC

**Main categories**. We record several main categories that appear most frequently in MC-IDC, as shown in Tab. 1.

Table 1. Sample numbers of main categories in MC-IDC.

| Categories | Images |
|---|---|
| man | 46,720 |
| woman | 26,670 |
| clothes | 20,040 |
| girl | 3,498 |
| panda | 3,007 |
| train | 2,286 |
| car | 1,974 |
| boy | 1,855 |
| dog | 1,820 |

**Data sources**. The data sources of MC-IDC can be divided into three categories: public datasets, web-crawled images, and movies. We detail the statistical information about various data sources in Tab. 2.

**The impact of single image datasets on diversity.** Our construction pipeline ensures both the consistency of object IDs and the diversity of the dataset: (1) As described in Tab. 2, the proportion of reference-target image pairs from single image datasets only accounts for 20% of the total. Most of the image pairs are from video data with stronger diversity. (2) For single image datasets, we adopt augmentation as mentioned in the main paper, which can increase the diversity of the dataset in terms of orientation, size, color, etc.

**Statistical analysis on the dataset.** We present the statistical data of MC-IDC from two aspects: text diversity and image domains. In terms of text diversity, we calculate the compression ratio, homogenization score and ngram diversity score, with the values being 3.717, 0.246 and 2.538 respectively. Regarding image domains, our dataset includes real-world content, animations, model-generated content, and movies, and their proportions are 0.362, 0.239, 0.227, and 0.172 respectively.

## 2. Additional Details about Experiment Setup

**Categories in evaluation dataset**. The evaluation dataset can be divided into general objects, human data, and virtual try-on data. The human data and the virtual try-on data each contain 300 different samples. General objects in the evaluation dataset consist of 50 categories, each of which contains 8 diverse samples. We summarize the 50 categories in Tab. 3.

**Text prompts for calculating DiverSim-i**. We use diverse text prompts describing different scenarios to guide the generation, and calculate DiverSim-i among the generated images. We record the text prompts in Tab. 4.

## 3. More Visual Results

**Various applications**. Our CustAny exhibits outstanding performance on various applications, such as story generation in Fig. 1, virtual try-on in Fig. 2, and ID-consistent generation in Fig. 3. We also show the visual results of the same reference picture under different text prompts in Fig. 4. Our CustAny can ensure both the ID fidelity and the generating diversity simultaneously.

**Additional visual comparisons with IP-Adapter in the virtual try-on domain**. IP-Adapter has a higher CLIP-t score than ours only in the virtual try-on domain. Subjectively speaking, however, the performance of IP-Adapter is inferior to that of our model, as shown on the left side in Fig. 5.

*Equal contribution
†Corresponding author

Table 2. Details about data sources of MC-IDC.

| Source | Dataset | Type | Image pair numbers |
|---|---|---|---|
| Public datasets | HumanFace [5] | Video | 55,830 |
| | VOS [4] | Video | 55,823 |
| | VIPSEG [3] | Video | 27,983 |
| | MVImgNet [6] | Multi-view image | 53,909 |
| | VITON [2] | Multi-view image | 20,000 |
| | LVIS [1] | Single image | 8,003 |
| Web-crawled images | - | Single image | 55,829 |
| Movies | - | Video | 38,405 |

Table 3. Categories of general objects in the evaluation dataset.

| | | | | |
|---|---|---|---|---|
| Winter melon | Cabbage | Vessel | Pillow | Screw driver |
| Pants | Computer mouse | Lipstick | Rice cooker | Toy figure |
| Clothing | Pineapple | Can | Plush toy | Grape |
| Toilet paper | Paper box | Skirt | Pawpaw | Ginger |
| Bowl | Train | Bottle | Cantaloupe | Sanitary napkin |
| Soccer | Bag | Umbrella | Hammer | Book |
| Flower | Shoe | Towel | Ashcan | Telephone |
| Faucet | Flowerpot | Motorcycle | Mug | Kiwi |
| Pot | Grapefruit | Jug | Car | Basket |
| Balloons | Tomato | Flashlight | Bagged snacks | Toy duck |

**Additional results on complex words**. Our method performs well when facing complex words in text prompts, which is shown on the right side in Fig. 5.

## References


[1] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 2

[2] Xintong Han, Zuxuan Wu, Zhe Wu, Ruichi Yu, and Larry S Davis. Viton: An image-based virtual try-on network. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7543–7552, 2018. 2

[3] Jiaxu Miao, Xiaohan Wang, Yu Wu, Wei Li, Xu Zhang, Yunchao Wei, and Yi Yang. Large-scale video panoptic segmentation in the wild: A benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21033–21043, 2022. 2

[4] Ning Xu, Linjie Yang, Yuchen Fan, Dingcheng Yue, Yuchen Liang, Jianchao Yang, and Thomas Huang. Youtube-vos: A large-scale video object segmentation benchmark. *arXiv preprint arXiv:1809.03327*, 2018. 2

[5] Jianhui Yu, Hao Zhu, Liming Jiang, Chen Change Loy, Weidong Cai, and Wayne Wu. Celebv-text: A large-scale facial text-video dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14805–14814, 2023. 2

[6] Xianggang Yu, Mutian Xu, Yidan Zhang, Haolin Liu, Chongjie Ye, Yushuang Wu, Zizheng Yan, Chenming Zhu, Zhangyang Xiong, Tianyou Liang, et al. Mvimgnet: A large-scale dataset of multi-view images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9150–9161, 2023. 2

Table 4. Text prompts for calculating DiverSim-i.

| Scenarios | Text prompts |
|---|---|
| Snow | Original text prompt + "The scene of the picture is in the snow."<br>Original text prompt + "The background of the picture is in the snow." |
| Grass | Original text prompt + "The scene of the picture is on the grass."<br>Original text prompt + "The background of the picture is on the grass." |
| Beach | Original text prompt + "The scene of the picture is on the beach."<br>Original text prompt + "The background of the picture is on the beach." |
| Jungle | Original text prompt + "The scene of the picture is in the jungle."<br>Original text prompt + "The background of the picture is in the jungle." |
| Eiffel Tower | Original text prompt + "The scene of the picture is beside the Eiffel Tower."<br>Original text prompt + "The background of the picture is beside the Eiffel Tower." |


Figure 1. Additional visual results of story generation. Our CustAny can generate diverse images under the guidance of text prompts, while maintaining the same identity as the object of interest in the reference image, thereby enabling the creation of a cohesive narrative.

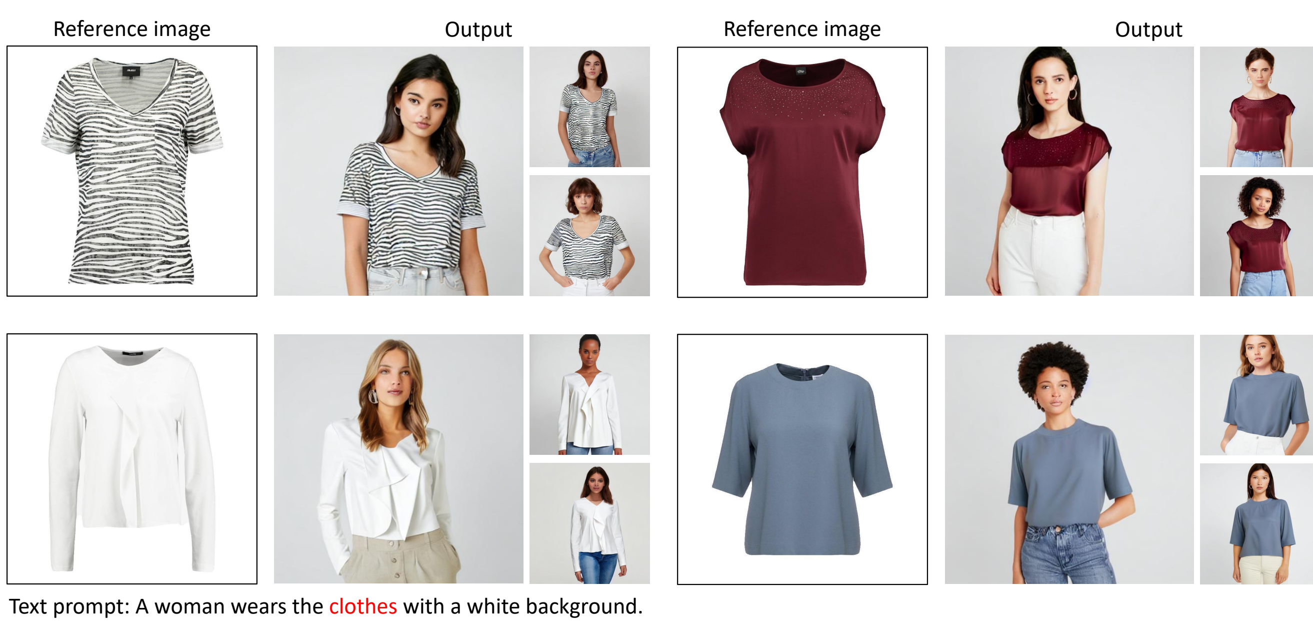


Figure 2. Additional visual results of virtual try-on. Given a piece of clothing, the CustAny can generate images of the clothing worn on a person.


Figure 3. Additional visual results of ID-consistent generation. The CustAny can generate multiple ID-consistent images with diverse non-ID elements such as motions and orientations.

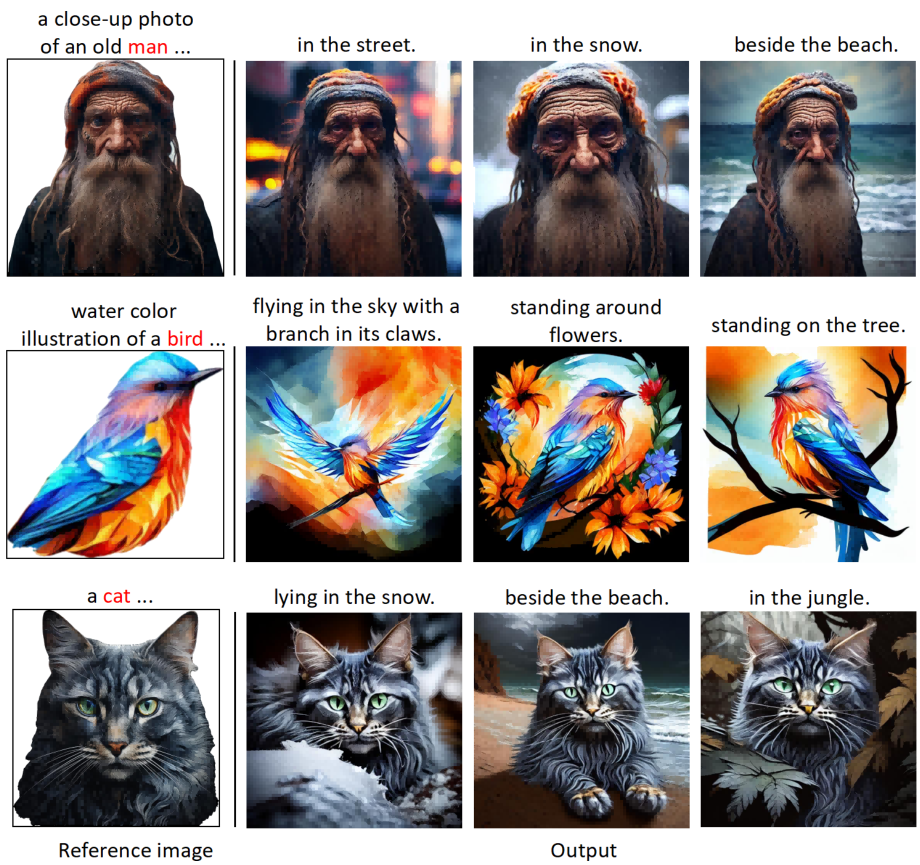


Figure 4. Additional results: the same reference in different text prompts.


Figure 5. Left: virtual try-on. Right: complex word.On the left side is the additional comparison between our method and the IP-Adapter, and on the right side is the performance of our method when facing complex words, such as cases with two words.