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

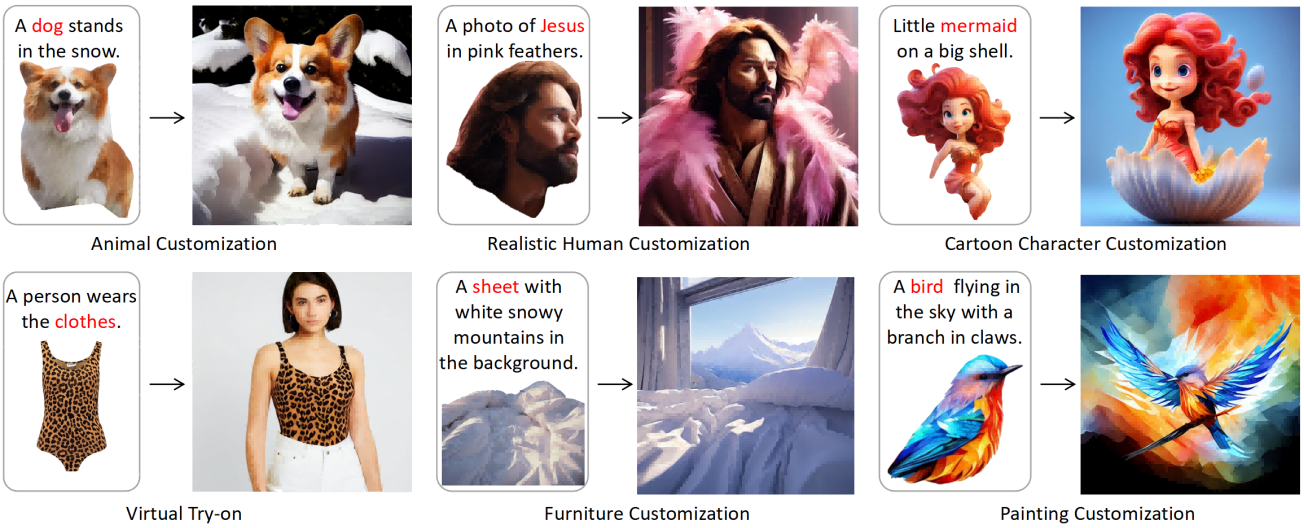

Figure 1. Customizing any object from a single example (text prompt and the reference image). Our CustAny can achieve various customization for general objects with high ID fidelity and flexible text edit-ability, without further fine-tuning.

Abstract

Recent advances in diffusion-based text-to-image models have simplified creating high-fidelity images, but preserving the identity (ID) of specific elements, like a personal dog, is still challenging. Object customization, using reference images and textual descriptions, is key to addressing this issue. Current object customization methods are either object-specific, requiring extensive fine-tuning, or object-agnostic, offering zero-shot customization but limited to specialized domains. The primary issue of promoting zero-shot object customization from specific domains to the general domain is to establish a large-scale general ID dataset for model pre-training, which is time-consuming and labor-intensive. In this paper, we propose a novel pipeline to construct a large dataset of general objects and build the Multi-Category ID-Consistent (MC-IDC) dataset, featuring 315k text-image samples across 10k categories. With the help of MC-IDC, we introduce Customizing Anything (CustAny), a zero-shot framework that maintains ID fidelity and supports flexible text editing for general objects. CustAny features three key components: a general ID extraction module, a dual-level ID injection module, and an ID-aware decoupling module, allowing it to customize any object from a single reference image and text prompt. Experiments demonstrate that CustAny outperforms existing methods in both general object customization and specialized domains like human customization and virtual try-on. Our contributions include a large-scale dataset, the CustAny framework and novel ID processing to advance this field. The official project page is in https://lingjiekong-fdu.github.io.

*Equal contribution
†Corresponding author

1. Introduction

Recent advancements in diffusion-based text-to-image generative models (e.g., SDXL, Improving GAN, PixArt Alpha) [2, 5, 18] have empowered users to generate highly realistic images with minimal technical knowledge. However, generating images of specific objects [7, 11, 27, 28], such as personal pets, remains challenging due to the difficulty of preserving their unique identity. Object customization, a technique that utilizes reference images and textual descriptions to generate images of specific objects, has emerged as a critical area of research. Existing methods [15, 17, 24, 27, 29] for object customization, such as DreamBooth, PortraitBooth and InstantID, have limitations in general-purpose object customization: Object-specific models [4, 24], while powerful, are often inefficient due to the need for extensive fine-tuning for each new object. Conversely, object-agnostic models [11, 27] have limited capabilities and often struggle with diverse object types.

Unlike existing customizing applications tailored to one specific task like face generation or virtual try-ons, this paper focuses on general object customization from a single example, as shown in Fig. 1. Specifically, it addresses zero-shot object customization in various scenarios where no fine-tuning on specific objects is available. Achieving this first requires a large-scale dataset containing various categories of objects for model pre-training. At present, however, there are few publicly released datasets in the field of zero-shot customization for general objects. To address this issue, we propose a novel construction pipeline for general ID dataset and introduce the Multi-Category ID-Consistent (MC-IDC) dataset, the first large-scale dataset for zero-shot object customization across diverse scenarios. It includes over 315K high-quality images across 10K categories, such as human faces, animals, clothing, and tools, supporting a wide range of object types and domains. The dataset features reference-target image pairs with consistent object IDs, including segmentation masks and text captions, making it ideal for ID-consistent generation tasks and advancing research in general object customization.

Based on our MC-IDC, we design a zero-shot general object customization framework which is able to Customize Anything (CustAny) from a single example. The innovation of our framework is mainly reflected in managing and utilizing identity information beyond the narrow focus of previous work. Specifically, the CustAny framework follows three essential steps: ID extraction, ID injection, and ID disentanglement. These steps address three critical questions: (1) How to capture enough identity details to handle a variety of complex customization tasks? (2) How to add this identity information to the model while still enabling text-based edits? (3) How to reduce the impact of irrelevant features from the reference object on the final results?

Formally, CustAny features an ID extraction module, dual-level ID injection, and an ID-aware decoupling module, offering innovative ID handling methods. Unlike prior methods [11, 27, 28] that use a single model for object representations, we leverage multiple pre-trained self-supervised models to capture more detailed ID-related features for high-ID-fidelity customization. To integrate this information into the diffusion model, we use a dual-level ID injection strategy: globally, by merging semantic-level ID features with text descriptions, and locally, by injecting patch tokens through cross-attention modules, preserving both ID fidelity and text editability.

Furthermore, we identify that ID information is often mixed with redundant non-ID elements, like object motion and orientation, which can disrupt both text editing and ID retention. To solve this, we introduce an ID-aware decoupling module during training to help the model separate ID-related properties from non-ID elements, enhancing both ID fidelity and generation diversity. With these advanced ID processing modules, CustAny enables zero-shot customization for various objects while supporting flexible text-based editing. As shown in Fig. 1, CustAny applies to object customization in various scenarios such as human customization, cartoon character personalization, virtual try-on. We conduct extensive experiments showing that CustAny achieves state-of-the-art performance in general object customization and outperforms task-specific methods in areas like human customization and virtual try-on.

In summary, we have these contributions: (1) **Dataset**: We propose a novel construction pipeline for general ID dataset, and present a large-scale dataset, MC-IDC, with 315K images across 10K categories for zero-shot object customization. (2) **CustAny Framework**: We design a zero-shot framework for general object customization in text-to-image generation, focusing on ID fidelity and text-based editing. (3) **ID Processing**: CustAny uses multiple self-supervised models and a dual-level ID injection strategy to integrate ID information while preserving text-editability. We present a module to separate ID from non-ID elements, improving ID fidelity and generation diversity. (4) **Good Results**: CustAny outperforms task-specific methods in general and specialized object customization tasks, including human customization and virtual try-on.

2. Related Work

Diffusion models. Diffusion models have demonstrated their effectiveness for text-to-image generations. The DDPM model pioneered in this field, using a diffusion and denoising process to create a mapping between Gaussian and image distributions. The Latent Diffusion Model (LDM) [22] took this a step further by applying the diffusion model to a latent space rather than pixel space, leading to the creation of text-to-image diffusion models like Stable Diffusion (SD), Midjourney, and DALLE-3 [21]. One

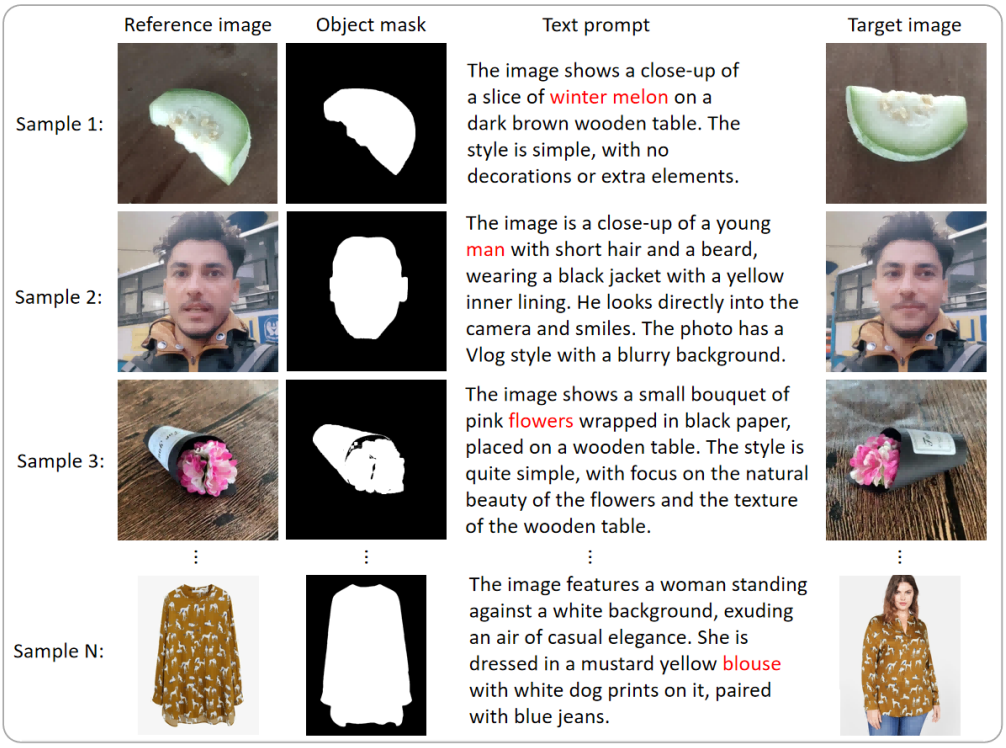

Figure 2. Illustration of our general ID dataset, MC-IDC. In each sample, the reference image with the object mask provides ID information, the text prompt offers semantic-level guidance for generation, and the target image serves as the ground truth.

main research problem for the diffusion model is its structure. Other than the original UNet, DiT [14] and Pixart-α [5] adopt transformer as the backbone structure. Other researchers have also focused on enhancing image fidelity regarding complicated concepts. By manipulating the cross-attention maps according to text prompts, [3] ensures that required objects are sufficiently generated in the images. [8] proposed to process the text prompts to decompose all target attributes, which can be further used to guide the denoising procedure with the help of different objective functions. To enhance content consistency, Tewel et al. [26] adopted an inner localization method by calculating the cross attention. Our method is generally built on diffusion models. While different from these basic models for plain image generations, we focus on the more challenging topic of zero-shot any object customization.

Image customization. Object customization techniques generally fall into two categories: object-specific and object-agnostic (zero-shot) methods. Object-specific approaches, like DreamBooth [24] and DisenBooth [4], fine-tune pretrained diffusion models by incorporating new, object-specific identifiers from a small set of reference images. While effective, these methods require extensive fine-tuning for each new object, which limits their scalability and efficiency. In contrast, object-agnostic or zero-shot approaches such as PhotoMaker [11] and InstantID [27], leverage large-scale training datasets to customize images without the need for additional fine-tuning. Although more efficient, these methods are often domain-specific, focusing on tasks like human customization, facial manipulation, or virtual try-on applications. For example, PhotoMaker trains a person ID extraction network, while InstantID applies ControlNet [29] to control facial expressions based on pretrained embeddings, and MagicCloth [6] uses ControlNet to customize garment details for virtual try-on. These existing methods, however, have limitations when it comes to general-purpose object customization. Object-specific models are powerful but inefficient, requiring time-intensive training for each new item, while object-agnostic models tend to be narrow in focus and are often limited to specific types of objects or applications. In this work, we aim to bridge these gaps by combining the versatility of object-specific methods with the efficiency of zero-shot methods. As illustrated in Fig. 1, our approach targets the ability to customize any object based on a single example image and text description, achieving generalization across diverse categories with minimal fine-tuning. This design marks a significant step toward general object customization across various domains, an area that remains largely unexplored.

3. MC-IDC Dataset

We curate the Multi-Category ID-Consistent (MC-IDC) dataset, the first large-scale dataset for general object customization research. We present the overall illustration of

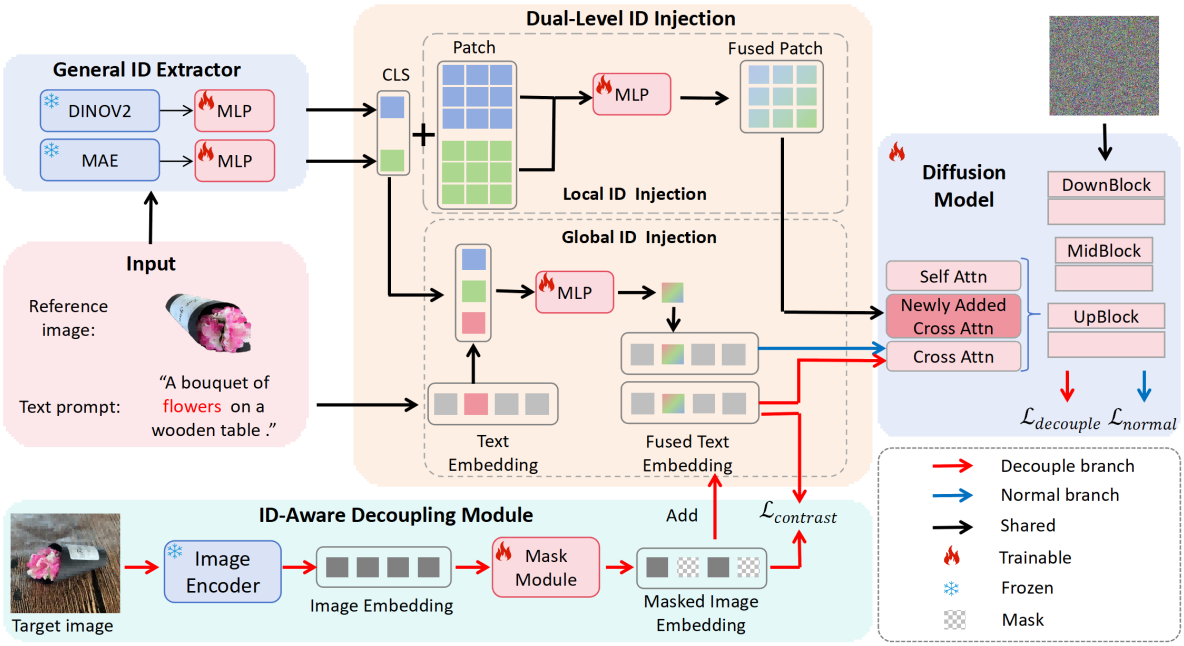

Figure 3. Overview of our CustAny. CustAny is a zero-shot text-to-image customization method for general objects, consisting of general ID extractor, global-local dual-level ID injection, and ID-aware decoupling module.

the MC-IDC as shown in Fig. 2. The dataset's merits and construction process are described below, with additional details in Supplementary.

Construction pipeline. The MC-IDC is constructed by the following four steps: (1) *Data collection*. MC-IDC dataset contains diverse sources including web-crawled images, movies and publicly available datasets. To facilitate the subsequent generation of image pairs, most of the data we collect are video datas or multi-view datas. To ensure the high quality of our dataset, we delete the images with a resolution less than 300×300. (2) *Instance detection and segmentation*. For some publicly available datasets or web-crawled images which do not provide segmentation annotations, we use advanced instance segmentation model [12] for segmentation. For movie dataset, we extract frames with 10fps and do object tracking with [16], aiming to establish the ID connection across frames. (3) *Image pair generation*. For most data samples which are from a clip of video or a group of multi-view data, we randomly select two frames containing the same object and perform a random crop around the object to form the reference-target image pair. For the other data samples from the single image dataset, we perform random data augmentations on the single image two times, and take them as the reference image and the target image respectively. (4) *Text prompt generation*. In each image pair, the target image is densely captioned with the state-of-the-art large vision language model [1], serving as the text prompt to guide the generation.

Merits of MC-IDC: Our MC-IDC dataset contains 315k high-quality samples across 10k+ categories and diverse domains, with each sample featuring reference-target image pairs, segmentation masks, and text captions to support ID-preserving, diverse object customization. Particularly, it has these merits: (1)*Diverse domains and categories:* Our MC-IDC dataset contains around 315k samples across 10k+ categories, including human faces, animals, clothes, and tools. It features images from diverse domains, such as real-world photos, animations, model-generated content, and movies, enabling general ID understanding. Further details are provided in the Supplementary. (2)*High quality:* With an average resolution of 1039×950, our dataset ensures high-quality training and inference for our model. (3) *Reference-target image pair:* Each sample includes a reference image with a segmentation mask, a target image, and a text caption. The reference image provides ID information, while the target image and text guide semantic generation. The reference and target images depict the same object in different states, sharing the same ID and varying non-ID elements like motion and direction. This setup ensures ID preservation while maintaining generation diversity.

4. Method

Overview. Given a reference image and text prompt, our goal is to generate an image that retains the identity (ID) of the object in the reference while modifying non-ID elements, like motions and backgrounds, based on the prompt. Thus we introduce CustAny, a zero-shot text-to-image customization framework for general objects, as in Fig. 3. We first use an ID extractor to capture ID information from the reference image with a segmentation mask. Dual-level ID

injection is then applied to embed the ID into the diffusion model globally and locally, preserving text-editing capabilities. Finally, an ID-aware decoupling module separates ID details from non-ID elements, improving ID fidelity and text-editing accuracy.

4.1. General ID Extraction

To extract comprehensive ID information, we use a combination of self-supervised models, DINOv2 [13] and MAE [9], to extract representations for general objects in the reference image. Specifically, previous methods, such as [11, 27], which target specific domains like faces, typically use a single model like CLIP [19] to extract object features. However, we find that using only one pre-trained model—whether CLIP, DINOv2, or MAE—leaves gaps in capturing features necessary for high-ID-fidelity in general objects (see Fig. 5 and Tab. 2). Essentially, while CLIP as in [7, 27] lacks sufficient detail, DINOv2 (despite its strong details due to contrastive learning) is color-insensitive because it is trained with color-augmented techniques like ColorJitter. MAE, however, retains color sensitivity due to its reconstruction-based training. Thus, combining DINOv2 and MAE allows us to capture both detail and color for more accurate ID extraction.

We combine DINOv2 and MAE, which complement each other, to capture comprehensive information for the challenging task of customizing diverse general objects. To align and prepare these features for injection, we use a two-layer MLP. The extraction process for general object representations is as follows,

$$f^C_{dino}, f^P_{dino} = MLP(F_{dino}(I_{ref} \odot M_{ref})), \quad (1)$$

$$f^C_{mae}, f^P_{mae} = MLP(F_{mae}(I_{ref} \odot M_{ref})), \quad (2)$$

where $\odot$ represents element-wise multiplication; and we have the class tokens and patch tokens f^C_{dino}, f^P_{dino}, f^C_{mae}, f^P_{mae} extracted by DINOv2 and MAE respectively. F_{dino} and F_{mae} are the DINOv2 and MAE backbone. I_{ref} denotes the reference image; and M_{ref} is the segmentation mask of the interested object in the reference image. Consequently, tokens derived from our general ID extractor not only offer intricate details from DINOv2, but also have sufficient colour and structural information from MAE.

4.2. Dual-level ID Injection

We aim to inject maximum extracted ID information into the diffusion UNet without compromising its text-editing ability. Our dual-level ID injection mechanism separately embeds global semantic ID and local fine-grained ID into the UNet, ensuring flexible text edits and high ID fidelity in customizing general objects.

Global ID injection. In text-to-image generation, text prompts guide diffusion models to create diverse images. However, diffusion models cannot reliably generate images with a specific object ID based on text alone. To embed object ID information into the text while preserving editability for other elements like background and motion, we introduce a global ID injection mechanism. Specifically, we merge the semantic-level class tokens (f^C_{dino}, f^C_{mae}) with the class word in the text embedding, such as "dog" or "cat",

$$f^C_{fuse} = MLP(Concat(f^C_{text}, f^C_{dino}, f^C_{mae})), \quad (3)$$

where f^C_{fuse} is fused class token and f^C_{text} is class word embedding. We insert f^C_{fuse} into the class word position in the text embeddings, obtaining the global-level condition c_g with ID-related information, as in Fig. 3. The global-level condition c_g interacts with cross-attention module of UNet, similar to standard text-to-image models [22] as,

$$O_g = Attention(Q_g, K_g, V_g), \quad (4)$$

where $Q_g = Z_g W_q$, $K_g = c_g W_k$, $V_g = c_g W_v$ are the query, key, and value of cross-attention module, in which W_q, W_k, W_v are the weight matrices of the trainable linear projection layers, and Z_g is the query feature containing the latent input information.

Local ID injection. The global-level condition c_g contains limited object details, making it insufficient for complex customization tasks. To provide more ID-related information, we propose the local ID injection mechanism, where we fuse the extracted patch tokens f^P_{dino} and f^P_{mae} using a two-layer MLP to create a unified representation c_l,

$$c_l = MLP(f^P_{dino}) + MLP(f^P_{mae}), \quad (5)$$

where c_l serves as the local-level condition for the diffusion UNet. Then, we add one cross-attention module in each up-block of the diffusion UNet in order to inject the ID-related details into the model without compressing c_l dimensions, feeding the model with as much ID information as possible.

$$O_l = Attention(Q_l, K_l, V_l), \quad (6)$$

where $Q_l = Z_l W_q$, $K_l = c_l W_k$, $V_l = c_l W_v$ are the query, key, and value of cross-attention module, in which W_l, W_l, W_l are the weight matrices of the trainable linear projection layers, and Z_l is query feature containing the latent input information. By dual-level ID injection, we have the model with sufficient ID-related information at both semantic and detail levels, without impairing its text-editing ability.

4.3. Training with ID-Aware Decoupling

The ID-related information injected into the diffusion model is often mixed with non-ID details like motion, direction, and size. Without proper guidance, the model may learn both together, disrupting text editing and limiting generation diversity. For example, in a reference image of "a

standing person" the posture "standing" could be injected, preventing the generation of "a sitting person" from a text prompt, leading to less diverse outputs. To address this, we propose an ID-aware decoupling module, inspired by previous works [4, 9], which includes a decoupling branch and associated losses during training to help the model distinguish ID information from other features.

Decoupling branch and normal branch. In the decoupling branch, we first extract the target image features f_{tar} through CLIP [19] to make an image embedding prior [21]. Next, we mask out the ID information contained in the image embedding by a trainable feature mask m_{id}, and thus the masked image feature $f_{msk} = f_{tar} \odot m_{id}$ contains only non-ID information of the target image, where $\odot$ represents element-wise multiplication. Then, we add the masked image feature f_{msk} to the fused class token f^C_{fuse} and then feed it to the diffusion UNet for generation following the injection way mentioned above. Yet in the normal branch, we inject f^C_{fuse} to the model without f_{msk}. The generation process of two branches can be illustrated as follows,

$$f_{msk} \oplus f^C_{fuse} \Rightarrow I_{tar}, \quad f^C_{fuse} \Rightarrow I_{tar}, \quad f_{msk} \perp f^C_{fuse}, \tag{7}$$

where $\oplus$ denotes element-wise addition, I_{tar} for target image, and $\perp$ indicates that the features on both sides are independent. Through training in both branches, we push non-ID information into masked image feature f_{msk} and retain only ID-related information in the fused class token f^C_{fuse}.

Training strategy. We use three losses in our training stage to ensure that the ID-aware decoupling module takes effect. Specifically, We utilize the following denoising losses to train each branch,

$$\mathcal{L}_{decouple} = \|\epsilon - \epsilon_\theta\big(x_t, c_g, f_{msk}, c_l, t\big)\|^2, \tag{8}$$

$$\mathcal{L}_{normal} = \|\epsilon - \epsilon_\theta\big(x_t, c_g, c_l, t\big)\|^2, \tag{9}$$

where t is the randomly sampled time step; and x_t represents the noisy latent of the target image; ϵ is the ground truth noise and ϵ_θ is the predicted noise. c_g and c_l shall be the global-level and local-level conditions acquired in the dual-level ID injection module. Additionally, we design a contrastive loss, ensuring that the masked image feature f_{msk} captures non-ID information and the fused class token f^C_{fuse} remains free of it, as below,

$$\mathcal{L}_{contrast} = Sim(f^C_{fuse}, f_{msk}), \tag{10}$$

where Sim is instantiated as cosine similarity. In summary, the training loss is the sum of the three loss functions mentioned above,

$$\mathcal{L} = \alpha_1 \mathcal{L}_{normal} + \alpha_2 \mathcal{L}_{decouple} + \alpha_3 \mathcal{L}_{contrast}, \tag{11}$$

where hyperparameters $\alpha_1, \alpha_2, \alpha_3$ are coefficients which are set to 2.0, 1.0, 0.5 respectively.

5. Experiment

Implementation details. We use SD1.5 as the backbone model for compatibility with open communities like Civitai.com, and apply the multi-scale training mode from [18] to handle varied image resolutions. The ID-aware decoupling module uses the CLIP encoder from [23], while MAE-ViT-h/14 and DINOv2-ViT-g/14-reg4 are used to extract ID features. Training involves a 1e-5 learning rate, batch size of 32, and 6 epochs on 32 V100 GPUs, taking about 30 hours. During inference, we perform 50 denoising steps and set the classifier-free guidance scale to 7.

Competitors. To demonstrate CustAny's generality, we compare it with both general object customization methods and task-specific methods in human customization and virtual try-on. We compare CustAny with IP-Adapter [28] for general object customization, PhotoMaker [11] and InstantID [27] for human customization, and MagicClothing [6] for virtual try-on. All methods are zero-shot and use a single reference image, so we exclude object-specific methods like DreamBooth [24] and DisenBooth [4] that require multiple reference images and fine-tuning.

Evaluation dataset. Our evaluation dataset consists of 1,000 text-image samples, covering general objects, human data, and virtual try-on data in a 4:3:3 ratio. None of these samples are included in the training set.

Evaluation metrics. As PhotoMaker [11], we utilize DINO-i [13] and CLIP-i [20] to measure the ID fidelity and use CLIP-t to measure the prompt fidelity. We leverage FID [10] to assess the generation quality. For human customization, we additionally calculate the face similarity (FaceSim) with FaceNet [25] as commonly done in [11, 27]. Further, we introduce DiverSim-i, a novel metric that measures the average DINO similarity across images generated from diverse text prompts. A lower DiverSim-i value indicates a stronger model ability to generate diverse images matching the prompts. Details are in Supplementary.

5.1. Quantitative and Qualitative Analysis

Quantitative results. Our CustAny outperforms previous works on general object customization for all metrics, and achieve comparable or better performance in contrast with specific methods tailored for specialized tasks such as human customization and virtual try-on, as shown in Tab. 1. Specifically, we achieve higher ID fidelity for CLIP-i and DINO-i, better prompt fidelity for CLIP-t, and higher generation quality evaluated by FID. Furthermore, our method exhibits better diversity on various scenarios as presented by DiverSim-i.

Qualitative analysis. The CustAny exhibits outstanding capabilities of high-quality customization for general objects, and even beat task-specialized methods in the specific domains, such as human customization and virtual try-on, as shown in Fig. 4. Our method is capable of generating

Table 1. Quantitative comparison among zero-shot object customization methods in general domain and two specific popular domains namely human customization and virtual try-on.

Domains	Methods	FID↓	CLIP-i↑	CLIP-t↑	DINO-i↑	FaceSim↑	DiverSim-i↓
General objects	IP-Adapter	70.32	77.18	28.03	44.94	-	84.43±0.66
	Ours	**47.09**	**82.16**	**29.27**	**65.13**	-	**74.38±3.99**
Human customization	IP-Adapter	102.69	72.17	29.32	41.44	65.51	-
	PhotoMaker	106.35	71.80	32.13	44.62	64.10	-
	InstanceID	113.18	75.87	**32.89**	49.26	63.26	-
	Ours	**86.40**	**79.60**	30.88	**57.44**	**78.54**	-
Virtual try-on	MagicClothing	126.09	76.53	21.40	29.10	-	89.36±0.40
	IP-Adapter	104.47	81.99	**25.03**	59.39	-	71.28±5.06
	Ours	**50.65**	**83.82**	22.42	**66.24**	-	**71.27±3.63**

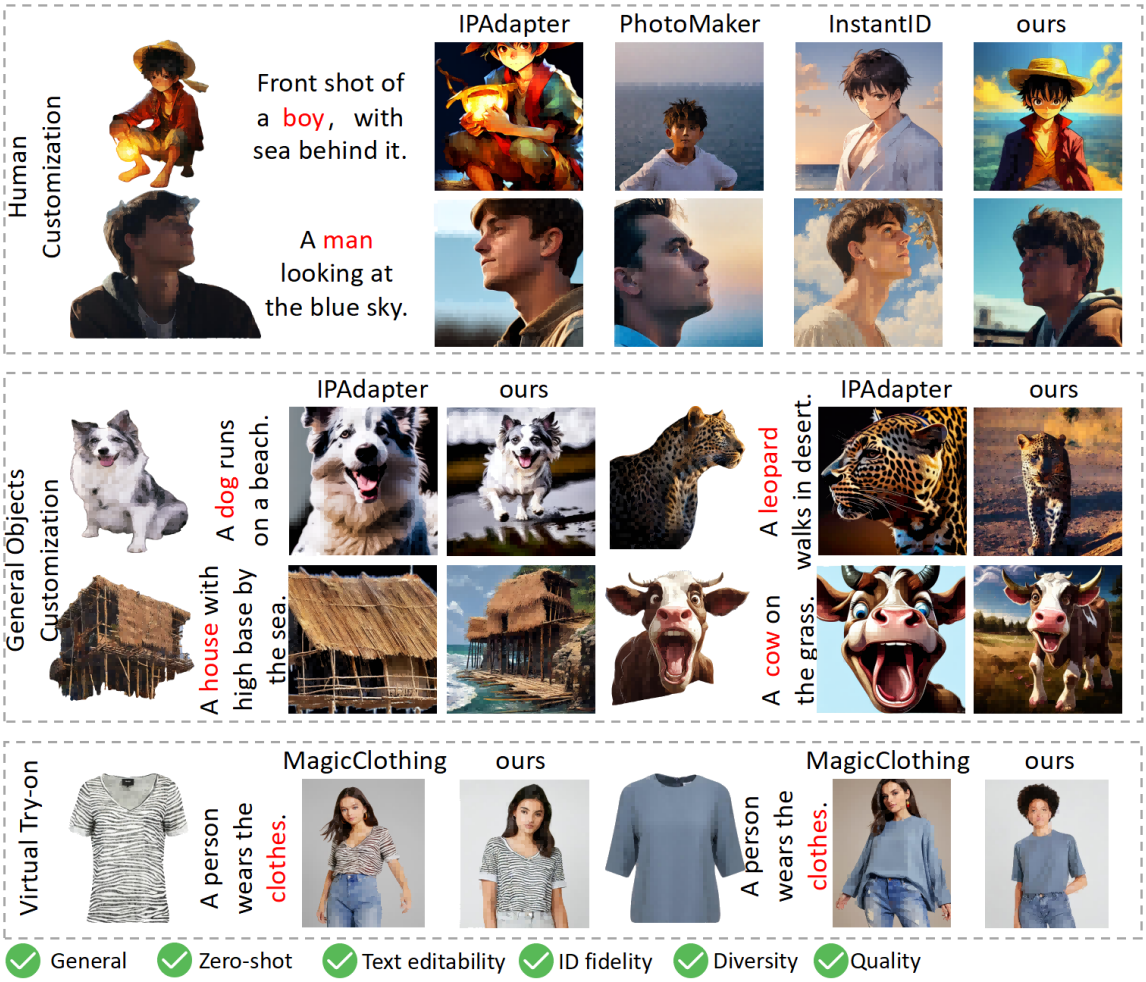

Figure 4. Qualitative results on general domains and two specific specific domains: human customization and virtual try-on. CustAny exhibits great ID-preserving ability with better text controls and more diverse generations on both general objects and specialized domains.

Table 2. Ablation study on the ID extraction methods.

Methods	FID ↓	CLIP-i ↑	DINO-i ↑
CLIP	49.11	79.58	59.45
DINO	48.89	80.82	63.71
MAE	49.00	79.09	59.79
Ours	**47.09**	**82.16**	**65.13**

Table 3. Ablation study on the ID injection methods.

Methods	FID ↓	CLIP-i ↑	DINO-i ↑
Global	49.78	78.76	60.67
Local	48.66	81.24	62.89
Ours	**47.09**	**82.16**	**65.13**

high-ID-fidelity images given one single reference image, and simultaneously editing non-ID elements like posture

Table 4. Ablation study on the ID-aware decoupling module.

Methods	FID ↓	CLIP-i ↑	DINO-i ↑
w.o. decoupling	47.50	81.86	65.12
with decoupling	**47.09**	**82.16**	**65.13**

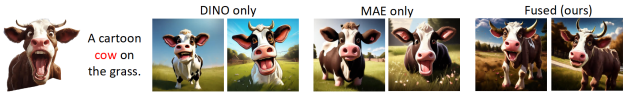

Figure 5. Ablation on different ID extraction methods in the quantitative perspective. Our ID extractor has both the rich details from DINOv2 and the color information from MAE.

Figure 6. Ablation on whether to add the ID-aware decoupling module in the quantitative perspective. The model with the ID-aware decoupling module demonstrates stronger text control ability, as shown by the blue text.

or background in accordance with text prompts, which endows the model with the ability to generate diverse images, such as various scenarios and different postures. Furthermore, benefiting from the ID-aware decoupling module, our model is capable of maintaining the ID fidelity of objects while diversifying the non-ID elements such as motions and directions compared to the reference images, even without corresponding guidance of text prompts. We show more visual results in Supplementary.

Effectiveness of general ID extraction. We compare our proposed general ID extractor with DINOv2 [13], MAE [9] and CLIP [20]. As in Tab. 2, our method performs the best, whether in terms of FID, which measures the quality of generation, or in terms of CLIP-i and DINO-i, which measure ID fidelity. As in Fig. 5, merely using DINOv2 as the extractor fails to maintain color consistency, while solely using MAE as the extractor lacks sufficient details. In contrast, our method combines the advantages of both DINOv2 and MAE, which enables our model to customize images with rich details and the consistent color of the object in reference images.

Benefits of dual-level ID injection. We explore the effectiveness of our dual-level ID injection module compared to only global-level injection and only local-level injection separately. As shown in Tab. 3, by combining global and local injection, CustAny achieves the best generation results both in quality as measured by FID and ID-fidelity as measured by CLIP-i and DINO-i.

Benefits of ID-aware decoupling. We conduct a comparative experiment on whether to add the ID decoupling module to the model during training, aiming to verify its effectiveness. As in Tab. 4, the model trained with the ID-aware decoupling module achieves higher ID-fidelity scores in terms of CLIP-i and DINO-i, which means that the decoupling module can help the model to discern ID information embedded in object representations, thereby generating results with better ID consistency. Further, we visualize the results by models trained with and without decoupling in Fig. 6. The model trained with decoupling exhibit enhanced capabilities in distinguishing ID information from non-ID elements such as motions (open or shut the mouth) and directions (front-facing or side-facing) of objects of interest. The enhanced discrimination ability allows the model to mitigate the influence of non-ID information during generation, thereby better preserving the text editing capabilities.

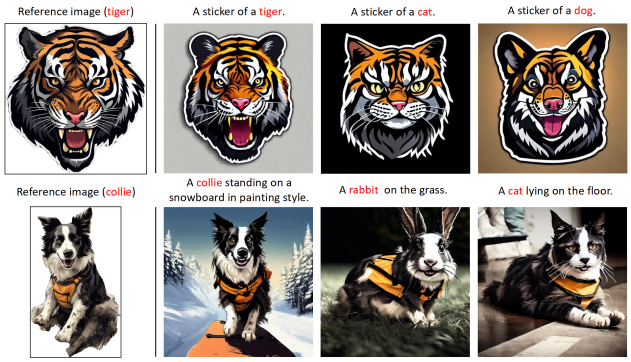

Figure 7. Applications of CustAny on text-image ID mixing, such as merging the text "dog" with the image "tiger".

Figure 8. Applications of CustAny on story generation. Given continuous text prompts and one image example, CustAny can generate corresponding stories.

5.2. Additional Qualitative Applications

Text-image ID mixing. If the category of the interested object in the text prompt and that in the reference image is not the same, our CustAny can merge the two and form a new ID, as shown in Fig. 7

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, as shown in Fig. 8.

6. Conclusion

We introduce CustAny, a zero-shot text-to-image framework for general object customization, integrating general ID extraction, dual-level ID injection, and an ID-aware decoupling module. CustAny achieves high ID fidelity while preserving text editing abilities and outperforms specialized methods in certain domains. Additionally, we create the MC-IDC dataset, the first large-scale general ID dataset, promote the research for object customization.

References

[1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 4

[2] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8, 2023. 2

[3] Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 782–791, 2021. 3

[4] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. In *The Twelfth International Conference on Learning Representations*, 2023. 2, 3, 6

[5] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. 2, 3

[6] Weifeng Chen, Tao Gu, Yuhao Xu, and Chengcai Chen. Magic clothing: Controllable garment-driven image synthesis. *arXiv preprint arXiv:2404.09512*, 2024. 3, 6

[7] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization, 2023. 2, 5

[8] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive text-to-image generation with rich text. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7545–7556, 2023. 3

[9] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners, 2021. 5, 6, 8

[10] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[11] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding, 2023. 2, 3, 5, 6

[12] Feng Liang, Bichen Wu, Xiaoliang Dai, Kunpeng Li, Yinan Zhao, Hang Zhang, Peizhao Zhang, Peter Vajda, and Diana Marculescu. Open-vocabulary semantic segmentation with mask-adapted clip. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7061–7070, 2023. 4

[13] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jegou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision, 2024. 5, 6, 8

[14] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023. 3

[15] Gan Pei, Jiangning Zhang, Menghan Hu, Guangtao Zhai, Chengjie Wang, Zhenyu Zhang, Jian Yang, Chunhua Shen, and Dacheng Tao. Deepfake generation and detection: A benchmark and survey. *arXiv preprint arXiv:2403.17881*, 2024. 2

[16] Jinlong Peng, Changan Wang, Fangbin Wan, Yang Wu, Yabiao Wang, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Yanwei Fu. Chained-tracker: Chaining paired attentive regression results for end-to-end joint multiple-object detection and tracking. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16*, pages 145–161. Springer, 2020. 4

[17] Xu Peng, Junwei Zhu, Boyuan Jiang, Ying Tai, Donghao Luo, Jiangning Zhang, Wei Lin, Taisong Jin, Chengjie Wang, and Rongrong Ji. Portraitbooth: A versatile portrait model for fast identity-preserved personalization. *arXiv preprint arXiv:2312.06354*, 2023. 2

[18] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023. 2, 6

[19] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 5, 6

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6, 8

[21] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 2, 6

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 5

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 6

[24] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir 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*, pages 22500–22510, 2023. 2, 3, 6

[25] Florian Schroff, Dmitry Kalenichenko, and James Philbin. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 815–823, 2015. 6

[26] Yoad Tewel, Yoav Shalev, Idan Schwartz, and Lior Wolf. Zerocap: Zero-shot image-to-text generation for visual-semantic arithmetic. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17918–17928, 2022. 3

[27] Qixun Wang, Xu Bai, Haofan Wang, Zekui Qin, Anthony Chen, Huaxia Li, Xu Tang, and Yao Hu. Instantid: Zero-shot identity-preserving generation in seconds, 2024. 2, 3, 5, 6

[28] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023. 2, 6

[29] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 2, 3