# MagicTailor: Component-Controllable Personalization in Text-to-Image Diffusion Models

**Donghao Zhou**[1*], **Jiancheng Huang**[2*], **Jinbin Bai**[3], **Jiaze Wang**[1], **Hao Chen**[1], **Guangyong Chen**[4], **Xiaowei Hu**[5†], **Pheng-Ann Heng**[1]
[1]The Chinese University of Hong Kong
[2]Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences
[3]National University of Singapore
[4]Zhejiang Lab
[5]Shanghai AI Lab
dhzhou@link.cuhk.edu.hk, huxiaowei@pjlab.org.cn

## Abstract

Text-to-image diffusion models can generate high-quality images but lack fine-grained control of visual concepts, limiting their creativity. Thus, we introduce **component-controllable personalization**, a new task that enables users to customize and reconfigure individual components within concepts. This task faces two challenges: *semantic pollution*, where undesired elements disrupt the target concept, and *semantic imbalance*, which causes disproportionate learning of the target concept and component. To address these, we design **MagicTailor**, a framework that uses *Dynamic Masked Degradation* to adaptively perturb unwanted visual semantics and *Dual-Stream Balancing* for more balanced learning of desired visual semantics. The experimental results show that MagicTailor achieves superior performance in this task and enables more personalized and creative image generation.

## 1 Introduction

Text-to-image (T2I) diffusion models [Rombach *et al.*, 2022; Ramesh *et al.*, 2022; Chen *et al.*, 2023] have shown impressive capabilities in generating high-quality images from textual descriptions. While these models can generate images that align well with provided prompts, they struggle when certain visual concepts are hard to express in natural language. To address this, methods like [Gal *et al.*, 2022; Ruiz *et al.*, 2023] enable T2I models to learn specific concepts from reference images, allowing for more accurate integration of those concepts into the generated images. This process, as shown in Fig. 1(a), is referred as personalization.

However, existing personalization methods are limited to replicating predefined concepts and lack flexible and fine-grained control of these concepts. Such a limitation hinders their practical use in real-world applications, restricting their potential for creative expression. Inspired by the observation that concepts often comprise multiple components, a key problem in personalization lies in *how to effectively control and manipulate these individual components.*

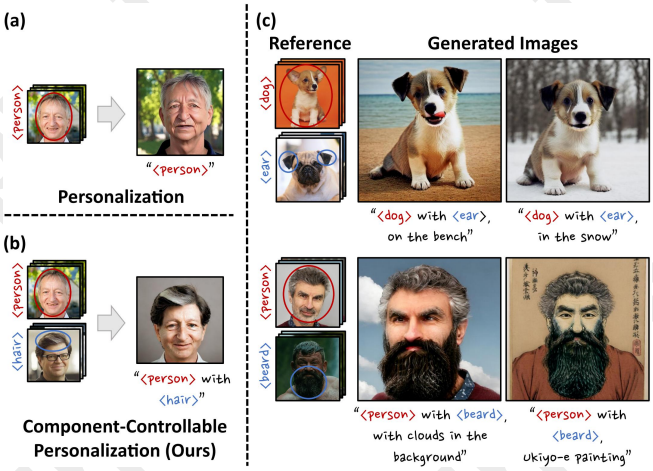


Figure 1: **(a) Personalization:** T2I models learn from reference images and then generate predefined visual concepts. **(b) Component-controllable personalization:** T2I models learn from additional visual references and then enable the integration of specific components into given concepts, further unleashing creativity. **(c) Generated images by MagicTailor:** MagicTailor can effectively achieve component-controllable personalization. Note that red and blue circles indicate the target concept and component, respectively.

In this paper, we introduce **component-controllable personalization**, a new task that enables the reconfiguration of specific components within personalized concepts using additional visual references (Fig. 1(b)). In this approach, a T2I model is fine-tuned with reference images and corresponding category labels, allowing it to learn and generate the desired concept along with the given component. This capability empowers users to refine and customize concepts with precise control, fostering creativity and innovation across various domains, from artworks to inventions.

One challenge of this task is *semantic pollution* (Fig. 2(a)), where unwanted visual elements inadvertently appear in gen-

Project page: https://correr-zhou.github.io/MagicTailor.
The full version is available at arXiv:2410.13370.

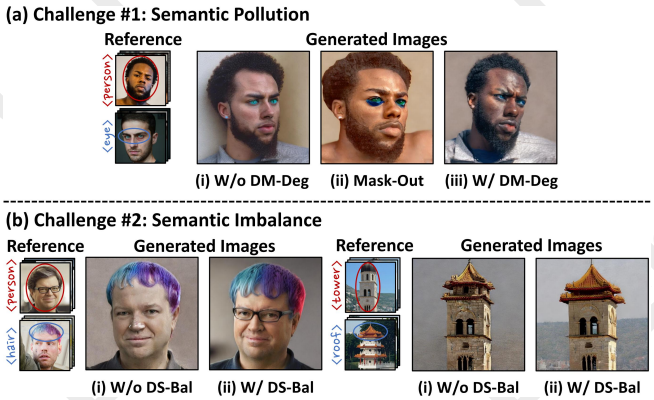


Figure 2: **Major challenges in component-controllable personalization. (a) Semantic pollution:** (i) Undesired elements may interfere with the personalized concept. (ii) A simple mask-out strategy causes unintended results, while (iii) DM-Deg effectively suppresses unwanted semantics. **(b) Semantic imbalance:** (i) Simultaneously learning the concept and component can distort either one. (ii) DS-Bal ensures balanced learning, improving personalization.

erated images, "polluting" the personalized concept. This happens because the T2I model often mixes visual semantics from different regions during training. Masking out unwanted elements in reference images doesn't solve the problem, as it disrupts the visual context and causes unintended compositions. Another challenge is *semantic imbalance* (Fig. 2(b)), where the model overemphasizes certain aspects, leading to unfaithful personalization. This occurs due to the semantic disparity between the concept and component, necessitating a more balanced learning approach to manage concept-level (*e.g.*, person) and component-level (*e.g.*, hair) semantics.

To address these challenges, we propose **MagicTailor**, a novel framework that enables component-controllable personalization for T2I models (Fig. 1(c)). We first use a text-guided image segmenter to generate segmentation masks for both the concept and component and then design *Dynamic Masked Degradation (DM-Deg)* to transform reference images into randomly degraded versions, perturbing undesired visual semantics. This method helps suppress the model's sensitivity to irrelevant details while preserving the overall visual context, effectively mitigating *semantic pollution*. Next, we initiate a warm-up phase for the T2I model, training it on the degraded images using a masked diffusion loss to focus on the desired semantics and a cross-attention loss to strengthen the correlation between these semantics and pseudo-words. To address *semantic imbalance*, we develop *Dual-Stream Balancing (DS-Bal)*, a dual-stream learning paradigm that balances the learning of visual semantics. In this phase, the online denoising U-Net performs sample-wise min-max optimization, while the momentum denoising U-Net applies selective preservation regularization. This ensures more faithful personalization of the target concept and component, resulting in outputs that better align with the intended objective.

In the experiments, we validate the superiority of MagicTailor through various qualitative and quantitative comparisons, demonstrating its state-of-the-art (SOTA) performance in component-controllable personalization. Moreover, detailed ablation studies and analysis further confirm the effectiveness of MagicTailor. In addition, we also show its potential for enabling a wide range of creative applications.

## 2 Methodology

Let $\mathcal{I} = \{(\{I_{nk}\}_{k=1}^{K}, c_n)\}_{n=1}^{N}$ denote a concept-component pair with $N$ samples of concepts and components, where each sample contains $K$ reference images $\{I_{nk}\}_{k=1}^{K}$ with a category label $c_n$. In this work, we focus on a practical setting involving one concept and one component. Specifically, we set $N = 2$ and define the first sample as a concept (*e.g.*, dog) while the second one as a component (*e.g.*, ear). In addition, these samples are associated with the pseudo-words $\mathcal{P} = \{p_n\}_{n=1}^{N}$ serving as their text identifiers. The goal of *component-controllable personalization* is to fine-tune a text-to-image (T2I) model to accurately learn both the concept and the component from $\mathcal{I}$. Using text prompts with $\mathcal{P}$, the fine-tuned model should generate images that integrate the personalized concept with the specified component.

This section begins by providing an overview of the MagicTailor pipeline in Sec. 2.1 and then delves into its two core techniques in Sec. 2.2 and Sec. 2.3.

### 2.1 Overall Pipeline

The overall pipeline of MagicTailor is illustrated in Fig. 3. The process begins with identifying the desired concept or component within each reference image $I_{nk}$, employing an off-the-shelf text-guided image segmenter to generate a segmentation mask $M_{nk}$ based on $I_{nk}$ and its associated category label $c_n$. Conditioned on $M_{nk}$, we design *Dynamic Masked Degradation (DM-Deg)* to perturb undesired visual semantics within $I_{nk}$, addressing *semantic pollution*. At each training step, DM-Deg transforms $I_{nk}$ into a randomly degraded image $\hat{I}_{nk}$, with the degradation intensity being dynamically regulated. Subsequently, these degraded images, along with structured text prompts, are used to fine-tune a T2I diffusion model to facilitate concept and component learning. The model is formally expressed as $\{\epsilon_\theta, \tau_\theta, \mathcal{E}, \mathcal{D}\}$, where $\epsilon_\theta$ represents the denoising U-Net, $\tau_\theta$ is the text encoder, and $\mathcal{E}$ and $\mathcal{D}$ denote the image encoder and decoder, respectively. To promote the learning of the desired visual semantics, we employ the masked diffusion loss, which is defined as:

$$\mathcal{L}_{\text{diff}} = \mathbb{E}_{n,k,\epsilon,t}\left[\left\|\epsilon_n \odot M'_{nk} - \epsilon_\theta(z_{nk}^{(t)}, t, e_n) \odot M'_{nk}\right\|_2^2\right], \quad (1)$$

where $\epsilon_n \sim \mathcal{N}(0, 1)$ is the unscaled noise, $z_{nk}^{(t)}$ is the noisy latent image of $\hat{I}_{nk}$ with a random time step $t$, $e_n$ is the text embedding of the corresponding text prompt, and $M'_{nk}$ is downsampled from $M_{nk}$ to match the shape of $\epsilon$ and $z_{nk}$. Additionally, we incorporate the cross-attention loss to strengthen the correlation between desired visual semantics and their corresponding pseudo-words, formulated as:

$$\mathcal{L}_{\text{attn}} = \mathbb{E}_{n,k,t}\left[\left\|A_\theta(p_n, z_{nk}^{(t)}) - M''_{nk}\right\|_2^2\right], \quad (2)$$

when $A_\theta(p_n, z_{nk}^{(t)})$ is the cross-attention maps between the pseudo-word $p_n$ and the noisy latent image $z_{nk}^{(t)}$ and $M''_{nk}$ is downsampled from $M_{nk}$ to match the shape of $A_\theta(p_n, z_{nk}^{(t)})$.

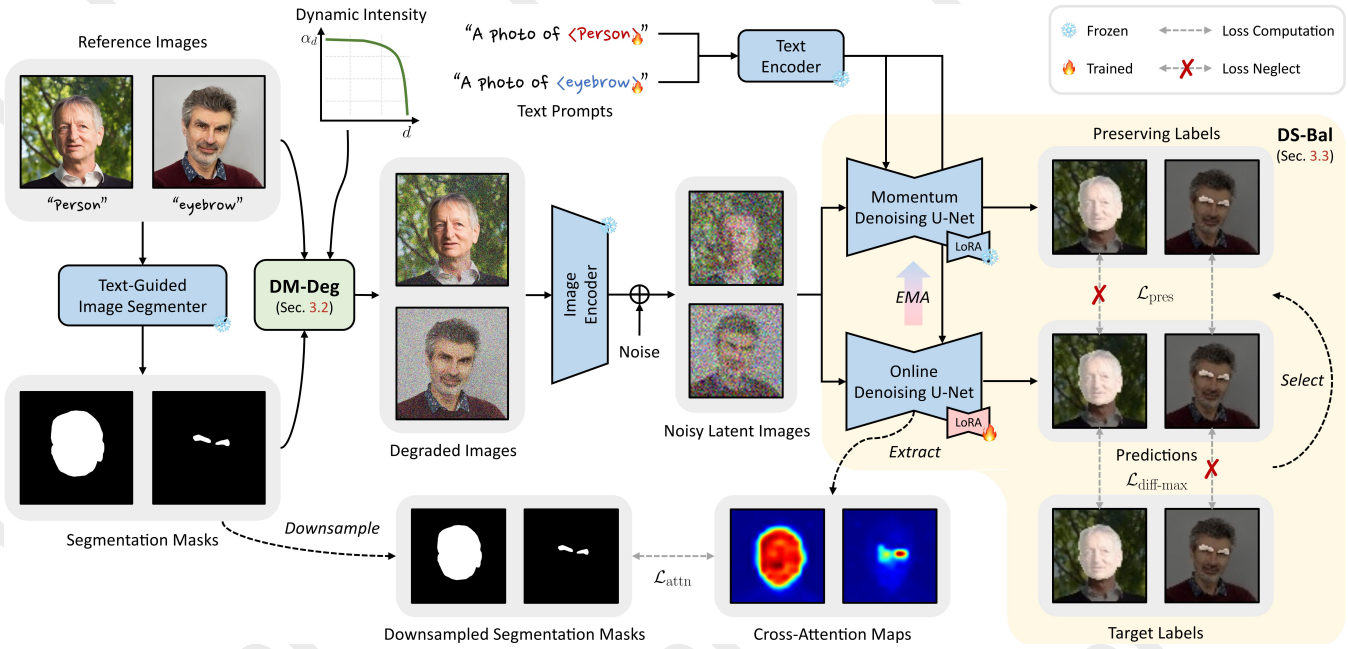


Figure 3: **Pipeline overview of MagicTailor.** This method fine-tunes a T2I diffusion model using reference images to learn both the target concept and component, enabling the generation of images that seamlessly integrate the component into the concept. Two key techniques, *Dynamic Masked Degradation* (*DM-Deg*, see Sec. 2.2) and *Dual-Stream Balancing* (*DS-Bal*, see Sec. 2.3), address *semantic pollution* and *semantic imbalance*, respectively. For clarity, only one image per concept/component is shown, and the warm-up stage is omitted.

Using $\mathcal{L}_{\text{diff}}$ and $\mathcal{L}_{\text{attn}}$, we first warm up the T2I model by jointly learning all samples, aiming to preliminarily inject the knowledge of visual semantics. The loss of the warm-up stage is defined as:

$$\mathcal{L}_{\text{warm-up}} = \mathcal{L}_{\text{diff}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} , \tag{3}$$

where $\lambda_{\text{attn}} = 0.01$ is the loss weight for $\mathcal{L}_{\text{attn}}$. For efficient fine-tuning, we only train the denoising U-Net $\epsilon_\theta$ in a low-rank adaptation (LoRA) [Hu *et al.*, 2021] manner and the text embedding of the pseudo-words $\mathcal{P}$, keeping the others frozen. Thereafter, we employ *Dual-Stream Balancing (DS-Bal)* to address *semantic imbalance*. In this paradigm, the online denoising U-Net $\epsilon_\theta$ conducts sample-wise min-max optimization for the hardest-to-learn sample, and meanwhile the momentum denoising U-Net $\tilde{\epsilon}_\theta$ applies selective preserving regularization for the other samples.

## 2.2 Dynamic Masked Degradation

*Semantic pollution* is a significant challenge for component-controllable personalization. As shown in Fig. 2(a.i), the target concept (*i.e.*, person) can be distorted by the owner of the target component (*i.e.*, eye), resulting in a hybrid person. Masking regions outside the target concept and component can damage the overall context, leading to overfitting and odd compositions (Fig. 2(a.ii)). To address this, undesired visual semantics in reference images must be handled appropriately. We propose *Dynamic Masked Degradation (DM-Deg)*, which dynamically perturbs undesired semantics to suppress their influence on the T2I model while preserving the overall visual context (Fig. 2(a.iii)).

**Degradation Imposition.** In each training step, DM-Deg imposes degradation in the out-of-mask region for each reference image. We use Gaussian noise for degradation due to its simplicity. For a reference image $I_{nk}$, we randomly sample a Gaussian noise matrix $G_{nk} \sim \mathcal{N}(0, 1)$ with the same shape as $I_{nk}$, where the pixel values of $I_{nk}$ range from $-1$ to 1. The degradation is then applied as follows:

$$\hat{I}_{nk} = \alpha_d G_{nk} \odot (1 - M_{nk}) + I_{nk}, \tag{4}$$

where $\odot$ denotes element-wise multiplication, and $\alpha_d \in [0, 1]$ is a dynamic weight controlling the degradation intensity. While previous works [Xiao *et al.*, 2023; Li *et al.*, 2023] have used noise to fully cover the background or enhance data diversity, DM-Deg aims to produce a degraded image $\hat{I}_{nk}$ that retains the original visual context. By introducing $\hat{I}_{nk}$, we can suppress the T2I model from perceiving undesired visual semantics in out-of-mask regions, as these semantics are perturbed by random noise at each training step.

**Dynamic Intensity.** Unfortunately, the T2I model may gradually memorize the introduced noise while learning meaningful visual semantics, leading to noise appearing in generated images (Fig. 4(a)). This behavior is consistent with previous observations on deep networks [Arpit *et al.*, 2017]. To address this, we propose a descending scheme that dynamically regulates the intensity of the imposed noise during training. This scheme follows an exponential curve, maintaining a relatively high intensity in the early stages and decreasing it sharply in later stages. Let $d$ denote the current training step and $D$ denote the total training step. The curve

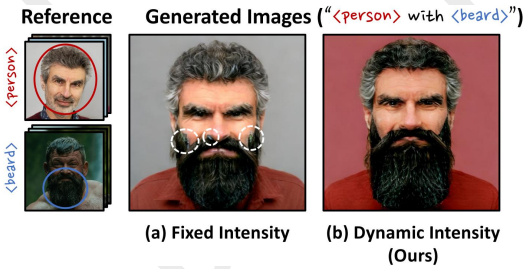


Figure 4: **Motivation of dynamic intensity.** (a) Fixed intensity ($\alpha_d = 0.5$ here) could cause noisy generated images. (b) Our dynamic intensity can mitigate noise memorization.

of dynamic intensity is defined as:

$$\alpha_d = \alpha_{\text{init}}(1 - (\frac{d}{D})^{\gamma}) , \tag{5}$$

where $\alpha_{\text{init}}$ is the initial value of $\alpha_d$ and $\gamma$ controls the descent rate. We empirically set $\alpha_{\text{init}} = 0.5$ and $\gamma = 32$, tuned within the powers of 2. This dynamic intensity scheme effectively prevents semantic pollution and significantly mitigates the memorization of introduced noise, leading to improved generation performance (Fig. 4(b)).

### 2.3 Dual-Stream Balancing

Another key challenge is *semantic imbalance*, which arises from the disparity in visual semantics between the target concept and its component. Specifically, concepts generally possess richer visual semantics than components (*e.g.*, person vs. hair), but in some cases, components may have more complex semantics (*e.g.*, simple tower vs. intricate roof). This imbalance complicates joint learning, leading to overemphasis on either the concept or the component, and resulting in incoherent generation (Fig. 5(a)). To address this, we design *Dual-Stream Balancing (DS-Bal)*, a dual-stream learning paradigm integrated with online and momentum denoising U-Nets (Fig. 3) for balanced semantic learning, aiming to improve personalization fidelity (Fig. 5(b)).

**Sample-Wise Min-Max Optimization.** From a loss perspective, the visual semantics of the concept and component are learned by optimizing the masked diffusion loss $\mathcal{L}_{\text{diff}}$ across all the samples. However, this indiscriminate optimization fails to allocate sufficient learning effort to a more challenging sample, leading to an imbalanced learning process. To address this, DS-Bal uses the online denoising U-Net to focus on learning the hardest-to-learn sample at each training step. Inheriting the weights of the original denoising U-Net, which is warmed up through joint learning, the online denoising U-Net $\epsilon_\theta$ optimizes only the sample with the highest masked diffusion loss as:

$$\mathcal{L}_{\text{diff-max}} = \max_n \mathbb{E}_{k,\epsilon,t}\Big[\big\|\epsilon_n \odot M'_{nk} - \epsilon_\theta(z_{nk}^{(t)}, t, e_n) \odot M'_{nk}\big\|_2^2\Big] , \tag{6}$$

where minimizing $\mathcal{L}_{\text{diff-max}}$ can be considered as a form of min-max optimization [Razaviyayn *et al.*, 2020]. The learning objective of $\epsilon_\theta$ may switch across different training steps

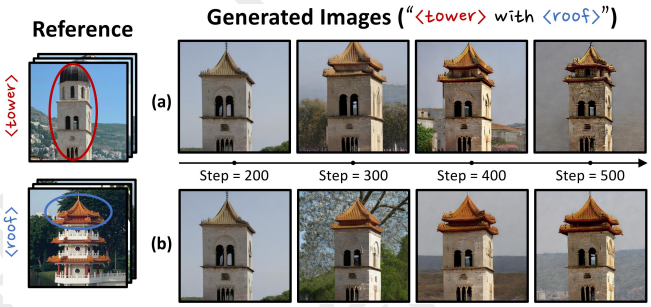


Figure 5: **Learning process visualization.** (a) The vanilla learning paradigm tends to overemphasize the easier one. (b) DS-Bal effectively balances the learning of the concept and component.

and is not consistently dominated by the concept or component. Such an optimization scheme can effectively modulate the learning dynamics of multiple samples and avoid the overemphasis on any particular one.

**Selective Preserving Regularization.** At a training step, the sample neglected in $\mathcal{L}_{\text{diff-max}}$ may suffer from knowledge forgetting. This is because the optimization of $\mathcal{L}_{\text{diff-max}}$, which aims to enhance the knowledge of a specific sample, could inadvertently overshadow the knowledge of the others. In light of this, DS-Bal meanwhile exploits the momentum denoising U-Net $\tilde{\epsilon}_\theta$ to preserve the learned visual semantics of the other sample in each training step. Specifically, we first select the sample that is excluded in $\mathcal{L}_{\text{diff-max}}$, which is expressed as $S = \{n|n = 1, ..., N\} - \{n_{\text{max}}\}$, where $n_{\text{max}}$ is the index of the target sample in $\mathcal{L}_{\text{diff-max}}$ and $S$ is the selected index set. Then, we use $\tilde{\epsilon}_\theta$ to apply regularization for $S$, with the masked preserving loss as:

$$\mathcal{L}_{\text{pres}} = \mathbb{E}_{n\in S,k,t}\Big[\big\|\tilde{\epsilon}_\theta(z_{nk}^{(t)}, t, e_n) \odot M'_{nk} - \epsilon_\theta(z_{nk}^{(t)}, t, e_n) \odot M'_{nk}\big\|_2^2\Big] , \tag{7}$$

where $\tilde{\epsilon}_\theta$ is updated from $\epsilon_\theta$ using EMA [Tarvainen and Valpola, 2017] with the smoothing coefficient $\beta = 0.99$, thereby sustaining the prior accumulated knowledge of $\epsilon_\theta$ in each training step. By encouraging the consistency between the output of $\epsilon_\theta$ and $\tilde{\epsilon}_\theta$ in $\mathcal{L}_{\text{pres}}$, we can facilitate the knowledge maintenance of the other samples while learning a specific sample in $\mathcal{L}_{\text{diff-max}}$. Overall, DS-Bal can be considered a mechanism to adaptively assign target labels $\epsilon_n$ or preserving labels $\tilde{\epsilon}_\theta(z_{nk}^{(t)}, t, e_n)$ to different samples, enabling dynamic loss supervision (Fig. 3). Using a loss weight $\lambda_{\text{pres}} = 0.2$, the total loss of the DS-Bal stage is formulated as:

$$\mathcal{L}_{\text{DS-Bal}} = \mathcal{L}_{\text{diff-max}} + \lambda_{\text{pres}}\mathcal{L}_{\text{pres}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}} . \tag{8}$$

## 3 Experimental Results

### 3.1 Experimental Setup

**Dataset, Implementation, and Evaluation.** For a systematic investigation, we collect a dataset from diverse domains, including characters, animation, buildings, objects, and animals. We use Stable Diffusion (SD) 2.1 [Rombach *et al.*,

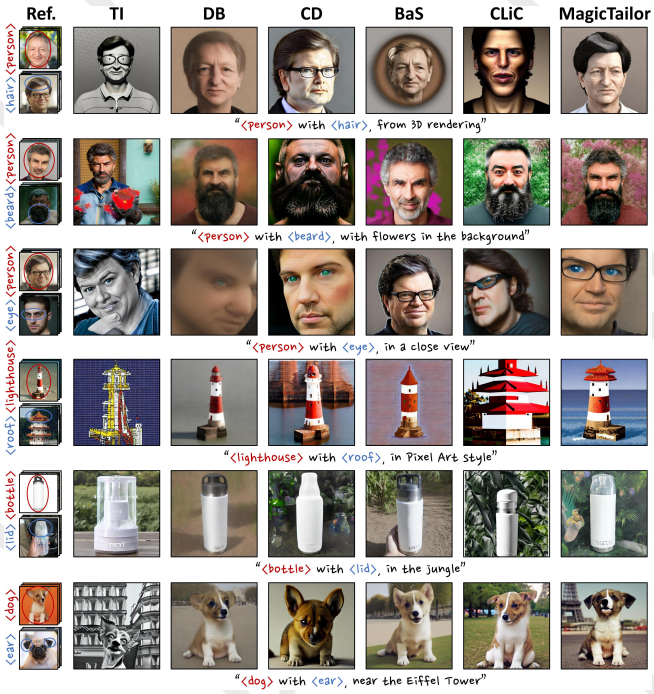


Figure 6: **Qualitative comparisons.** We present images generated by MagicTailor and other methods across various domains. MagicTailor achieves better text alignment, identity fidelity, and generation quality. *Due to space limitations, please zoom in for a better view*. More results are provided in Appendix D.

2022] as the pretrained T2I model. For the warm-up and DS-Bal stages, we set the training steps to 200 and 300, with learning rates of $1 \times 10^{-4}$ and $1 \times 10^{-5}$, respectively. Each concept-component pair requires only about five minutes of training on an A100 GPU. For evaluation, we design 20 text prompts covering a wide range of scenarios and generate 14,720 images for each method. To ensure fairness, all random seeds are fixed during both training and inference. More details of the experimental setup are included in Appendix A.

**Compared Methods.** We compare our MagicTailor with several personalization methods, including Textual Inversion (TI) [Gal *et al.*, 2022], DreamBooth (DB) [Ruiz *et al.*, 2023], Custom Diffusion (CD) [Kumari *et al.*, 2023], Break-A-Scene (BAS) [Avrahami *et al.*, 2023], and CLiC [Safaee *et al.*, 2024]. These methods were selected for their representativeness of personalization frameworks or relevance to learning fine-grained elements. For a fair comparison, we adapt them to our task with minimal modifications, specifically by incorporating the masked diffusion loss (Eq. 1). Apart from method-specific configurations, all methods are implemented using the same setup to ensure consistency.

## 3.2 Qualitative Comparisons

The qualitative results are shown in in Fig. 6. As observed, TI, CD, and CLiC primarily suffer from semantic pollution, where undesired visual semantics significantly distort the personalized concept. Besides, DB and BAS also struggle in this challenging task, with an overemphasis on either the concept or the component due to semantic imbalance, sometimes even causing the target component to be completely absent. An interesting finding is that imbalanced learning can exacerbate semantic pollution, leading to the color and texture of the target concept or component being mistakenly transferred to unintended parts of the generated images. In contrast, MagicTailor effectively generates text-aligned images that accurately represent both the target concept and component. To further demonstrate the performance of MagicTailor, we provide additional comparisons in Appendix B.

| Methods | CLIP-T ↑ | CLIP-I ↑ | DINO ↑ | DreamSim ↓ |
|---|---|---|---|---|
| Textual Inversion [Gal *et al.*, 2022] | 0.236 | 0.742 | 0.620 | 0.558 |
| DreamBooth [Ruiz *et al.*, 2023] | 0.266 | 0.841 | 0.798 | 0.323 |
| Custom Diffusion [Kumari *et al.*, 2023] | 0.251 | 0.797 | 0.750 | 0.407 |
| Break-A-Scene [Avrahami *et al.*, 2023] | 0.259 | 0.840 | 0.780 | 0.338 |
| CLiC [Safaee *et al.*, 2024] | 0.263 | 0.764 | 0.663 | 0.499 |
| MagicTailor (Ours) | **0.270** | **0.854** | **0.813** | **0.279** |

Table 1: **Quantitative comparisons on automatic metrics.** MagicTailor can achieve SOTA performance on all four automatic metrics. The best results are marked in bold.

| Methods | Text Align. ↑ | Id. Fidelity ↑ | Gen. Quality ↑ |
|---|---|---|---|
| Textual Inversion [Gal *et al.*, 2022] | 5.8% | 2.5% | 5.2% |
| DreamBooth [Ruiz *et al.*, 2023] | 15.3% | 14.7% | 12.5% |
| Custom Diffusion [Kumari *et al.*, 2023] | 7.1% | 7.7% | 9.8% |
| Break-A-Scene [Avrahami *et al.*, 2023] | 10.8% | 12.1% | 22.8% |
| CLiC [Safaee *et al.*, 2024] | 4.5% | 5.1% | 6.2% |
| MagicTailor (Ours) | **56.5%** | **57.9%** | **43.4%** |

Table 2: **Quantitative comparisons on the user study.** MagicTailor also outperforms other methods in all aspects of human evaluation.

## 3.3 Quantitative Comparisons

**Automatic Metrics.** We utilize four automatic metrics in the aspects of text alignment (CLIP-T [Gal *et al.*, 2022]) and identity fidelity (CLIP-I [Radford *et al.*, 2021], DINO [Oquab *et al.*, 2023], DreamSim [Fu *et al.*, 2023]). *To precisely measure identity fidelity*, we segment out the concept and component in each reference and evaluation image, and then eliminate the target component from the segmented concept. As we can see in Tab. 1, component-controllable personalization remains a tough task even for SOTA methods of personalization. By comparison, MagicTailor achieves the best results in both identity fidelity and text alignment. It should be credited to the effective framework tailored to this special task.

**User Study.** We further evaluate the methods with a user study. Specifically, a detailed questionnaire is designed to display 20 groups of evaluation images with the corresponding text prompt and reference images. Users are asked to select the best result in each group for three aspects, including text alignment, identity fidelity, and generation quality. Finally, we collect a total of 3,180 valid answers and report the selected rates in Tab. 2. It can be observed that MagicTailor can also achieve superior performance in human preferences, further verifying its effectiveness.

| DM-Deg | DS-Bal | CLIP-T ↑ | CLIP-I ↑ | DINO ↑ | DreamSim ↓ |
|---|---|---|---|---|---|
| | | 0.275 | 0.837 | 0.798 | 0.317 |
| ✓ | | **0.276** | 0.848 | 0.809 | 0.294 |
| | ✓ | 0.270 | 0.845 | 0.802 | 0.304 |
| ✓ | ✓ | 0.270 | **0.854** | **0.813** | **0.279** |

Table 3: **Effectiveness of key techniques.** Our DM-Deg and DS-Bal effectively contribute to a superior performance trade-off.


Figure 7: **Compatibility with different backbones.** We equip MagicTailor with SD 1.5 [Rombach *et al.*, 2022], SD 2.1 [Rombach *et al.*, 2022], and SDXL [Podell *et al.*, 2023]. The results show that MagicTailor can be generalized to multiple backbones, and a better backbone could provide better generation quality.

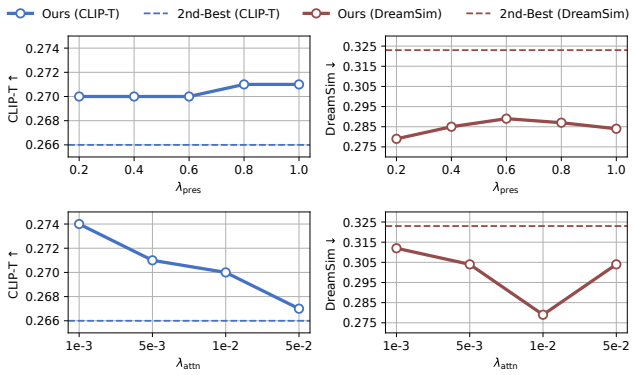


Figure 8: **Robustness on loss weights.** We report CLIP-T [Gal *et al.*, 2022] for text alignment, and DreamSim [Fu *et al.*, 2023] for identity fidelity as it is most similar to human judgments. Second-best results in Table 1 are also presented to highlight our robustness.

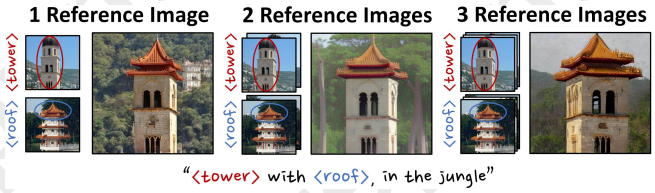


Figure 9: **Performance on different numbers of reference images.** We present qualitative results to show that MagicTailor can still achieve satisfactory performance when provided only 1 or 2 reference image(s) per concept and component.c

## 3.4 Ablation Studies and Analysis

We conduct comprehensive ablation studies and analysis for MagicTailor to verify its capability. More ablation studies and analysis are included in Appendix C.

**Effectiveness of Key Techniques.** In Tab. 3, we investigate two key techniques by starting from a baseline framework described in Sec. 2.1. Even without DM-Deg and DS-Bal, such a baseline framework can still have competitive performance, showing its reliability. On top of that, we introduce DM-Deg and DS-Bal, where the superior performance trade-off indicates their significance. Qualitative results can refer to Fig. 2.

**Compatibility with Different Backbones.** MagicTailor can also collaborate with other T2I diffusion models as it is a model-independent approach. In Fig. 7, we employ MagicTailor in other backbones like SD 1.5 [Rombach *et al.*, 2022] and SDXL [Podell *et al.*, 2023], showcasing MagicTailor can also achieve remarkable results. Notably, we directly use the original hyperparameter values without further selections, showing the generalizability of MagicTailor.

**Robustness on Loss Weights.** In Fig. 8, we analyze the sensitivity of loss weights in Eq. 8 (*i.e.*, $\lambda_{\text{pres}}$ and $\lambda_{\text{attn}}$), since loss weights are often critical for model training. As we can see, when $\lambda_{\text{pres}}$ and $\lambda_{\text{attn}}$ vary within a reasonable range, our MagicTailor can consistently attain SOTA performance, revealing its robustness on these hyperparameters.

**Performance on Different Numbers of Reference Images.** In Fig. 9, we reduce the number of reference images to analyze the performance variation. With fewer reference images, MagicTailor can still show satisfactory results. While more reference images could lead to better generalization ability, one reference image per concept/component is enough to obtain a decent result with our MagicTailor.

**Generalizability to Complex Prompts.** In comparisons, we have used well-categorized text prompts for systemic evaluation. Here we further evaluate MagicTailor's performance on other complex text prompts involving more complicated contexts. As shown in Fig. 11, MagicTailor effectively generates text-aligned images when performing fidelity personalization, showing its ability to handle diverse user needs.

**Generalizability to Difficult Pairs.** We further evaluate MagicTailor's performance on challenging pairs, focusing on two cases: 1) large geometric discrepancy, such as "<person>" in an upper body portrait and "<hair>" in a profile photo, and 2) cross-domain interactions, such as "<person>" and "<ear>" of dogs. As shown in Fig. 12, even facing these hard cases, MagicTailor can still effectively personalize target concepts and components with high fidelity.

## 3.5 Further Applications

**Decoupled Generation.** After learning from a concept-component pair, MagicTailor can also enable decoupled generation. As shown in Fig. 10(a), MagicTailor can generate the target concept and component separately in various and even cross-domain contexts. This should be credited to its remark-


Figure 10: **Further applications of MagicTailor. (a) Decoupled generation:** MagicTailor can also separately generate the target concept and component, enriching prospective combinations. **(b) Controlling multiple components:** MagicTailor shows the potential to handle more than one component, highlighting its effectiveness. **(c) Enhancing other generative tools:** MagicTailor can seamlessly integrate with various generative tools, adding the capability to control components within their generation pipelines.

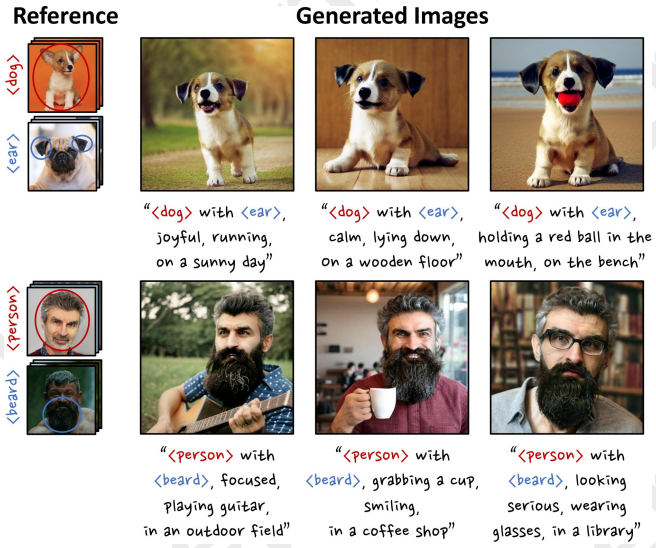


Figure 11: **Generalizability for complex prompts.** We present qualitative results generated with complex text prompts. In addition to those well-categorized text prompts, our MagicTailor can also follow more complex ones to generate text-aligned images.

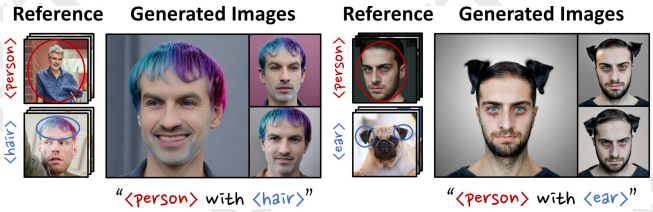


Figure 12: **Generalizability for difficult pairs.** We show the results of two hard cases involving large geometric discrepancy and cross-domain interactions, showing that MagicTailor can effectively handle such challenging scenarios.

able ability to capture different-level visual semantics. Such an ability extends the flexibility of the possible combination between the concept and component.

**Controlling Multiple Components.** In this paper, we focus on personalizing one concept and one component, because such a setting is enough to cover extensive scenarios, and can be further extended to reconfigure multiple components with an iterative procedure. However, as shown in Fig. 10(b), our MagicTailor also exhibits the potential to control two components simultaneously. Handling more components remains a prospective direction of exploring better control over diverse elements for a single concept.

**Enhancing Other Generative Tools.** We demonstrate how MagicTailor enhances other generative tools like ControlNet [Zhang *et al.*, 2023], CSGO [Xing *et al.*, 2024], and InstantMesh [Xu *et al.*, 2024] in Fig. 10(c). MagicTailor can integrates seamlessly, furnishing them with an additional ability to control the concept's component in their pipelines. For instance, working with MagicTailor, InstantMesh can conveniently achieve fine-grained 3D mesh design, exhibiting the practicability of MagicTailor in more creative applications.

## 4 Conclusion

We introduce *component-controllable personalization*, enabling precise customization of individual components within concepts. The proposed *MagicTailor* uses *Dynamic Masked Degradation (DM-Deg)* to suppress unwanted semantics and *Dual-Stream Balancing (DS-Bal)* to ensure balanced learning. Experiments show that MagicTailor sets a new standard in this task, with promising creative applications. In the future, we would like to extend our approach to broader image and video generation, enabling finer control over multi-level visual semantics for creative generation capabilities.

## Contribution Statement

Donghao Zhou and Jiancheng Huang contribute equally. Xiaowei Hu is the corresponding author.

## Acknowledgments

We would like to thank Pengzhi Li, Tian Ye, Jingyu Lin, and Jialin Gao for their valuable discussion and suggestions. This study was supported by the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government via the Hong Kong Centre for Logistics Robotics.

## References

[Arpit *et al.*, 2017] Devansh Arpit, Stanislaw Jastrzebski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *Int. Conf. Mach. Learn.*, pages 233–242, 2017.

[Avrahami *et al.*, 2023] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia*, pages 1–12, 2023.

[Chen *et al.*, 2023] Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, et al. Pixart-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

[Fu *et al.*, 2023] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *arXiv preprint arXiv:2306.09344*, 2023.

[Gal *et al.*, 2022] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An image is worth one word: Personalizing text-to-image generation using textual inversion. In *Int. Conf. Learn. Represent.*, 2022.

[Hu *et al.*, 2021] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021.

[Kumari *et al.*, 2023] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 1931–1941, 2023.

[Li *et al.*, 2023] Zhen Li, Mingdeng Cao, Xintao Wang, Zhongang Qi, Ming-Ming Cheng, and Ying Shan. Photomaker: Customizing realistic human photos via stacked id embedding. *arXiv preprint arXiv:2312.04461*, 2023.

[Oquab *et al.*, 2023] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023.

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

[Radford *et al.*, 2021] 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 *Int. Conf. Mach. Learn.*, pages 8748–8763, 2021.

[Ramesh *et al.*, 2022] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022.

[Razaviyayn *et al.*, 2020] Meisam Razaviyayn, Tianjian Huang, Songtao Lu, Maher Nouiehed, Maziar Sanjabi, and Mingyi Hong. Nonconvex min-max optimization: Applications, challenges, and recent theoretical advances. *IEEE Signal Process. Mag.*, 37(5):55–66, 2020.

[Rombach *et al.*, 2022] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 10684–10695, 2022.

[Ruiz *et al.*, 2023] 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 *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 22500–22510, 2023.

[Safaee *et al.*, 2024] Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, and Ali Mahdavi-Amiri. Clic: Concept learning in context. In *IEEE Conf. Comput. Vis. Pattern Recognit.*, pages 6924–6933, 2024.

[Tarvainen and Valpola, 2017] Antti Tarvainen and Harri Valpola. Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results. In *Adv. Neural Inf. Process. Syst.*, 2017.

[Xiao *et al.*, 2023] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. Fastcomposer: Tuning-free multi-subject image generation with localized attention. *arXiv preprint arXiv:2305.10431*, 2023.

[Xing *et al.*, 2024] Peng Xing, Haofan Wang, Yanpeng Sun, Qixun Wang, Xu Bai, Hao Ai, Renyuan Huang, and Zechao Li. Csgo: Content-style composition in text-to-image generation. *arXiv preprint arXiv:2408.16766*, 2024.

[Xu *et al.*, 2024] Jiale Xu, Weihao Cheng, Yiming Gao, Xintao Wang, Shenghua Gao, and Ying Shan. Instantmesh: Efficient 3d mesh generation from a single image with sparse-view large reconstruction models. *arXiv preprint arXiv:2404.07191*, 2024.

[Zhang *et al.*, 2023] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Int. Conf. Comput. Vis.*, pages 3836–3847, 2023.