Identity Decoupling for Multi-Subject Personalization of Text-to-Image Models

Sangwon Jang[*,1], **Jaehyeong Jo**[*,1], **Kimin Lee**[†,1], **Sung Ju Hwang**[†,1,2]
[*]Equal contribution [†]Equal advising
KAIST[1], DeepAuto.ai[2]
{ sangwon.jang, harryjo97, kiminlee, sjhwang82 }@kaist.ac.kr

Figure 1: Given a few images of multiple subjects (red boxes), MuDI can personalize a text-to-image model (e.g., SDXL [36]) to generate multi-subject images without identity mixing. Some reference images (e.g., Cloud Man and Blue Alien) are created by Sora [31], introducing novel concepts not previously encountered by SDXL.

Abstract

Text-to-image diffusion models have shown remarkable success in generating personalized subjects based on a few reference images. However, current methods often fail when generating multiple subjects simultaneously, resulting in mixed identities with combined attributes from different subjects. In this work, we present MuDI, a novel framework that enables multi-subject personalization by effectively decoupling identities from multiple subjects. Our main idea is to utilize segmented subjects generated by a foundation model for segmentation (Segment Anything) for both training and inference, as a form of data augmentation for training and initialization for the generation process. Moreover, we further introduce a new metric to better evaluate the performance of our method on multi-subject personalization. Experimental results show that our MuDI can produce high-quality personalized images without identity mixing, even for highly similar subjects as shown in Figure 1. Specifically, in human evaluation, MuDI obtains twice the success rate for personalizing multiple subjects without identity mixing over existing baselines and is preferred over 70% against the strongest baseline. Our project page is at https://mudi-t2i.github.io/.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

 https://doi.org/10.52202/079017-3200

Figure 2: **Comparison of multi-subject personalization methods using Corgi and Chow Chow images (red boxes) using SDXL [36].** DreamBooth [42] produces mixed identity dogs, such as a Corgi with Chow Chow ears[1]. Cut-Mix [15] often generates artifacts like unnatural vertical lines. Additionally, using layout conditioning like region control [14] proves ineffective in preventing identity blending in recent advanced diffusion models such as SDXL. In contrast, ours successfully personalizes each dog, avoiding identity mixing and artifacts observed in prior methods.

1 Introduction

Text-to-image diffusion models, trained on large datasets of image and text pairs, have shown great success in generating high-quality images for given text prompts [39, 43, 40, 5]. Building on this success, there is a growing interest in personalizing these text-to-image models. Specifically, given a few images of a single user-defined subject, several methods have been developed to enable these models to generate images of the subject in novel contexts [13, 42, 50, 21]. Furthermore, these personalization methods have been expanded to include the customization of style [45], background [47], and activity [19], offering even greater flexibility and creativity in image generation.

Despite significant progress in personalizing text-to-image models for single subjects, current methods often struggle to handle multiple subjects simultaneously [21, 15]. While successful in rendering each subject individually, these methods suffer from identity mixing during the composition of subjects. For instance, as shown in Figure 2(a), recent works, such as DreamBooth [42], generate images with mixed identities when applied to two dogs. The problem of identity mixing becomes more pronounced with semantically similar subjects that share attributes, such as colors or textures, which leads to greater confusion in maintaining distinct identities.

To address identity mixing in multi-subject personalization, Han et al. [15] proposed to use Cut-Mix [55], an augmentation technique that presents the models with cut-and-mixed images of the subjects during personalization. However, using Cut-Mix-like images inevitably often results in the generation of unnatural images with stitching artifacts, such as vertical lines that separate the subjects. Moreover, Cut-mix remains unsuccessful in decoupling similar subjects (see Figure 2(b)). There are alternative approaches [8, 25, 14] that rely on pre-defined conditioning, e.g., bounding boxes or ControlNet [56] to separate the identities spatially. However, such auxiliary inputs like sketch [56] could be difficult to obtain, and we have observed that the layout conditioning is ineffective for recent diffusion models such as SDXL [36] (see Figure 2(c) and Appendix B.11).

In this work, we propose MuDI, a multi-subject personalization framework that effectively addresses identity mixing, even for highly similar subjects. Our key idea is to leverage the segmented subjects obtained by a foundation model for image segmentation (Segment Anything Model (SAM) [20]), enabling the decoupling of the identities among different subjects. Specifically, we extract segmentation maps of the user-provided subjects using SAM and utilize them for both training and inference. For training, we introduce a data augmentation method that randomly composes segmented subjects, which allows efficient personalization by removing identity-irrelevant information. Additionally, we utilize the segmented subjects to initialize the generation process. Instead of starting from Gaussian noise, we begin with a mean-shifted random noise created from segmented subjects. We find that this provides a helpful hint for the model to separate the identities and further reduces subject missing during generation. Notably, our approach significantly mitigates identity mixing as shown in Figure 2, without relying on preset auxiliary conditions such as bounding boxes or sketches.

We evaluate the effectiveness of the proposed framework using a new dataset composed of subjects prone to identity mixing, which includes a diverse range of categories from animals to objects and scenes. To facilitate this evaluation, we introduce a new metric specifically designed to assess the

[1]Custom Diffusion [21] also results in identity mixing and we provide the examples in Figure 20.

fidelity of multiple subjects in the images, taking into account the degree of identity mixing. In our experiments, MuDI successfully personalizes the subjects without mixed identities, significantly outperforming DreamBooth [42], Cut-Mix [15], and Textual Inversion [13], in both qualitative and quantitative comparisons. Further human study with side-by-side comparisons of MuDI over other methods shows that human raters prefer our method by more than 70% over the strongest baseline.

2 Related work

Text-to-image personalization Personalized text-to-image diffusion models have shown impressive abilities to render a single user-specific subject in novel contexts from only a few images. Two representative classes of personalization methods have been proposed by Gal et al. [13] and Ruiz et al. [42]. Textual Inversion [13] optimizes new text embedding for representing the specified subjects and has been improved for learning on extended embedding spaces [50, 1]. On the other hand, DreamBooth [42] fine-tunes the weights of the pre-trained model to bind new concepts with unique identifiers and has been developed by recent works [21, 15, 48] for efficiently fine-tuning the models.

However, existing methods fall short of synthesizing multiple user-defined subjects together, suffering from identity mixing. Han et al. [15] introduce Cut-Mix to address identity mixing by augmenting Cut-Mix-like images during training but fail to separate similar subjects and generates stitching artifacts. Other lines of work [25, 14] compose personalized subjects using layout conditioning, which manipulates cross-attention maps with user-defined locations. Yet such conditioning based on cross-attention maps is ineffective for recent diffusion models such as SDXL [36]. In this work, we develop a novel framework that allows the personalization of multiple subjects without identity mixing even for subjects with similar appearances.

Modular customization Recent works [21, 14, 35] explore a different scenario for personalizing multiple subjects, namely *modular customization*, where the subjects are independently learned by models and users mix and match the subjects during inference to compose them. Custom Diffusion [21] merges individually fine-tuned models by solving constrained optimization and Mix-of-Show [14] introduces gradient fusion to merge single-concept LoRAs [18]. When handling multiple subjects, these works also suffer from identity mixing, and they rely on preset spatial conditions such as ControlNet [56] and region control [14] to address the problem. Notably, our method can be applied to this scenario to decouple subjects' identities without using such conditions.

3 Preliminaries

Text-to-image diffusion models Diffusion models [17, 46] generate samples from noise by learning to reverse the perturbation, i.e., denoise, which can be modeled by a diffusion process. To be specific, at each step of the diffusion, the model predicts the random noise $\boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I})$ that has been used to corrupt the sample. Text-to-image diffusion models [43, 40] incorporate text conditions for the generation. Given the dataset $\mathcal{D}$ consisting of the image-text pairs $(\boldsymbol{x}, \boldsymbol{c})$, text-to-image diffusion models parameterized by the noise prediction model $\boldsymbol{\epsilon}_\theta$ can be trained with the following objective:

$$\mathcal{L}_{DM}(\theta; \mathcal{D}) = \mathbb{E}_{(\boldsymbol{x},\boldsymbol{c})\sim\mathcal{D},\boldsymbol{\epsilon}\sim\mathcal{N}(0,\mathbf{I}),t\sim\mathcal{U}(0,T)} \left[\left\| \boldsymbol{\epsilon}_\theta(\boldsymbol{x}_t; \boldsymbol{c}, t) - \boldsymbol{\epsilon} \right\|_2^2 \right], \tag{1}$$

where $\boldsymbol{\epsilon}$ is the random noise, time t is sampled from the uniform distribution $\mathcal{U}(0, T)$, and $\boldsymbol{x}_t = \alpha_t \boldsymbol{x} + \sigma_t \boldsymbol{\epsilon}$ for the coefficients α_t and σ_t that determine the noise schedule of the diffusion process.

Personalizing text-to-image models Given a few images of a single specific subject, DreamBooth [42] fine-tunes the weights of the diffusion model with a unique identifier for the subject, i.e., "a [identifier] [class noun]". The model weights are updated to learn the subject while preserving the visual prior, which can be achieved by minimizing the objective:

$$\mathcal{L}_{DB}(\theta) = \underbrace{\mathcal{L}_{DM}(\theta; \mathcal{D}_{ref})}_{\text{personalization loss}} + \lambda \underbrace{\mathcal{L}_{DM}(\theta; \mathcal{D}_{prior})}_{\text{prior preservation loss}}, \tag{2}$$

where $\mathcal{L}_{DM}$ is the loss defined in Eq. (1), $\mathcal{D}_{ref}$ is the dataset consisting of reference images of the subject, $\mathcal{D}_{prior}$ is the dataset consisting of class-specific prior images, and λ is a coefficient for the prior preservation loss. Similar to personalizing a single subject, existing works [21, 15] jointly train for multiple subjects by combining the images from the set of user-specified subjects to construct $\mathcal{D}_{ref}$ and using different identifiers for each subject.

https://doi.org/10.52202/079017-3200

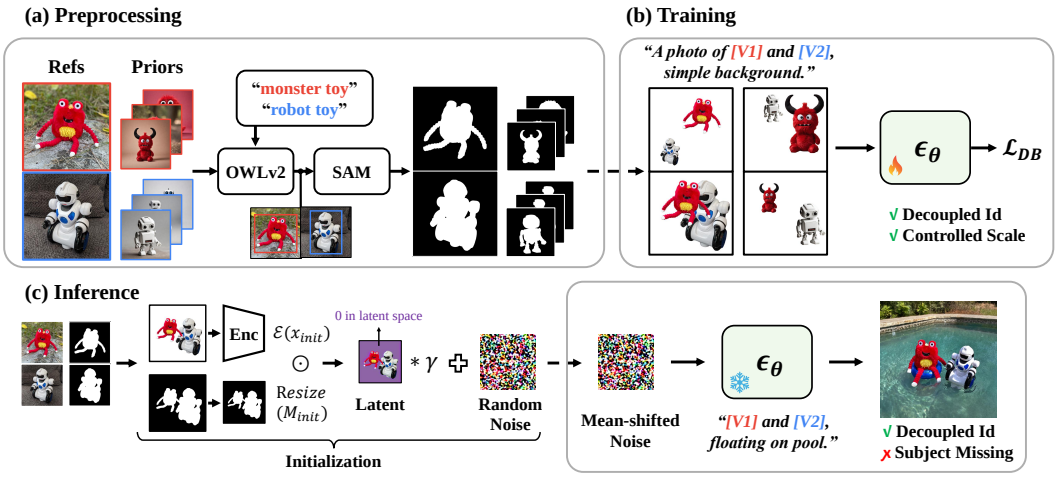


Figure 3: **Overview of MuDI**. (a) We automatically obtain segmented subjects using SAM [20] and OWLv2 [28] in the preprocessing stage. (b) We augment the training data by randomly positioning segmented subjects with controllable scales to train the diffusion model ϵ_θ. We refer to this data augmentation method as Seg-Mix. (c) We initialize the generation process with mean-shifted noise created from segmented subjects, which provides a signal for separating identities without missing.

4 MuDI: Multi-subject personalization for decoupled identities

In this section, we present MuDI: **Mu**lti-subject personalization for **D**ecoupled **I**dentities, which leverages segmented subjects to separate identities. In Section 4.1, we describe our training method, which augments training data through random compositions of segmented subjects. We also introduce a simple inference method that initializes noise for sample generation based on subject segmentation in Section 4.2. Finally, we present a new metric to evaluate the multi-subject fidelity in Section 4.3.

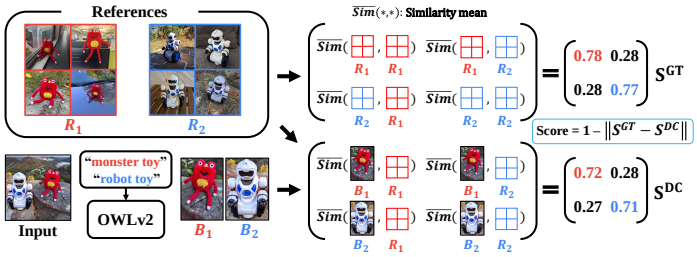
4.1 Training

Personalization with augmentation To address identity mixing in multi-subject personalization, we introduce a new data augmentation method for training the pre-trained text-to-image model called *Seg-Mix*. We aim to mitigate identity mixing by leveraging segmented subjects during personalizing text-to-image models. By isolating each subject from the background, Seg-Mix enables the model to learn to distinguish between different identities effectively. We integrate Seg-Mix with DreamBooth [42], which personalizes text-to-image models using unique identifiers (see Eq. (2)).

To implement Seg-Mix, we preprocess reference images by automatically extracting segmentation maps of user-provided subjects using the Segment Anything Model (SAM) [20]. Specifically, this process begins with the extraction of subject bounding boxes using the OWLv2 [28], an object detection model with an open vocabulary. Subsequently, SAM segments the subjects based on these bounding boxes, as illustrated in Figure 3(a). After the preprocessing step, we create augmented images by randomly positioning the resized segmented subjects, as illustrated in Figure 3(b). We provide the detailed procedures of our method in Algorithm 1. These augmented images are paired with a simple prompt "A photo of $[V_1]$ and $[V_2]$, simple background.", which is designed to explicitly remove identity-irrelevant information. We also apply this augmentation to the prior dataset by creating images from segmented prior subjects. Using augmented datasets, we fine-tune text-to-image models based on the DreamBooth objective function in Eq.(2).

One of the key advantages of Seg-Mix is its ability to train models without identity-irrelevant artifacts, due to the removal of backgrounds. This process also mitigates unnatural artifacts, such as stitching artifacts observed in previous methods like Cut-Mix [15]. Moreover, by allowing subjects to overlap during Seg-Mix, which is different from Cut-Mix, we prevent attributes from leaking to neighboring identities and enhance interactions among the subjects, which we analyze in Appendix B.2.

Descriptive class Intuitively, for two similar subjects in the same category, separating them solely with unique identifiers is a challenging task and prone to identity mixing. In single-subject personalization, Chae et al. [7] observed that adding detailed descriptions in front of the class nouns helps in capturing the visual characteristics of rare subjects. Inspired by this observation, we adopt



Metrics	Spearman	AUROC
DS [12]	0.219	0.606
DINOv2 [33]	0.156	0.559
D&C-DS	**0.618**	**0.850**
D&C-DINO	0.445	0.741

Figure 4: **(Left) Overview of Detect-and-Compare**. We calculate the mean similarities between detected subjects and reference images to evaluate multi-subject fidelity. Specifically, we compare $\boldsymbol{S}^{GT}$ and $\boldsymbol{S}^{DC}$. We provide pseudo-code in Algorithm 3. **(Right) Correlation between metrics and human evaluation**. We report the Spearman's rank correlation coefficient and AUROC.

specific class nouns (e.g., Weimaraner instead of dog) or add detailed descriptions in front of general class nouns (e.g., white robot toy instead of toy). Instead of manually selecting appropriate classes, we leverage GPT4-v [32] to automatically obtain these specific class nouns or descriptions. We empirically validate that this simple modification improves the preservation of the details for multiple subjects leading to the decoupling of the identities of highly similar subjects.

4.2 Inference

It has been observed that initial noises for the generation affect the overall quality of generated images [27, 44], a finding that holds for personalized models with Seg-Mix as well. Motivated by this observation, we propose a novel inference method to improve identity decoupling without additional training or computational overhead. As illustrated in Figure 3(c), we first create an image $\boldsymbol{x}_{init}$ of segmented subjects following Seg-Mix and extract its latent embedding from VAE encoder $\mathcal{E}$. We then add this latent embedding to a random Gaussian noise $\boldsymbol{\epsilon}$, scaled by a coefficient γ as $\boldsymbol{z}_T = (\mathcal{E}(\boldsymbol{x}_{init}) \odot Resize(\boldsymbol{M}_{init})) * \gamma + \boldsymbol{\epsilon}$ where $Resize(\boldsymbol{M}_{init})$ denotes the resized version of segmentation mask $\boldsymbol{M}_{init}$. This *mean-shifted* noise $\boldsymbol{z}_T$ encodes coarse information about the subjects and their layout, serving as a good starting point in sample generation. We analyze the effect of coefficient γ in Appendix B.3 and validate the diversity of generated images from the initialization in Appendix B.4. The proposed inference method is summarized in Algorithm 2. Additionally, instead of using randomly composed initial latent, we explore utilizing Large Language Models (LLMs) [9] to generate the layouts of bounding boxes for each subject aligned with the given prompt. Such an LLM-guided initialization enhances the ability to render complex interactions between subjects (see Figure 24 in Appendix B.5 for supporting results).

We remark that our initialization method also addresses the issue of subject dominance [49], where certain subject dominates the generation while other subjects are ignored. By providing information through the initial composition, our inference method guides the model to consider all subjects without additional computation. In Section 5.3 and Appendix B.6, we validate that our inference method alleviates subject dominance, playing a crucial role when rendering many subjects simultaneously.

4.3 New metric for multi-subject fidelity

Existing metrics designed for measuring subject fidelity, such as CLIP-I [37] or DINOv2 [33], are not suitable for evaluating multiple subjects because they do not account for identity mixing. Therefore, we introduce a new metric, called *Detect-and-Compare* (D&C), for evaluating multi-subject fidelity.

First, we utilize OWLv2 [28] to detect the subjects in the generated image, with text queries as the supercategories of the subjects. For the detected subjects $\{B_i\}_{i=1}^{N}$ and the reference subjects $\{R_j\}_{j=1}^{M}$, we construct similarity matrices by measuring the similarities between the subjects using subject fidelity metrics such as DreamSim [12] or DINOv2 [33]. Specifically, we first construct the *D&C similarities* matrix $\boldsymbol{S}^{DC}$, where ij-th entry represents the similarity between detected subject B_i and reference R_j (see Figure 4). Similarly, we construct the *ground-truth similarities* $\boldsymbol{S}^{GT}$, where ij-th entry represents the similarity between reference objects R_i and R_j. Since a mixed-identity subject yields high similarities to multiple references, we compare $\boldsymbol{S}^{DC}$ and $\boldsymbol{S}^{GT}$ to account for identity mixing. Notably, the difference between $\boldsymbol{S}^{DC}$ and $\boldsymbol{S}^{GT}$ yields a matrix where diagonal entries denote similarities to the corresponding subject, while off-diagonal entries indicate similarities to other subjects which represent identity mixing. The closer $\boldsymbol{S}^{DC}$ is to $\boldsymbol{S}^{GT}$, the more accurately

https://doi.org/10.52202/079017-3200

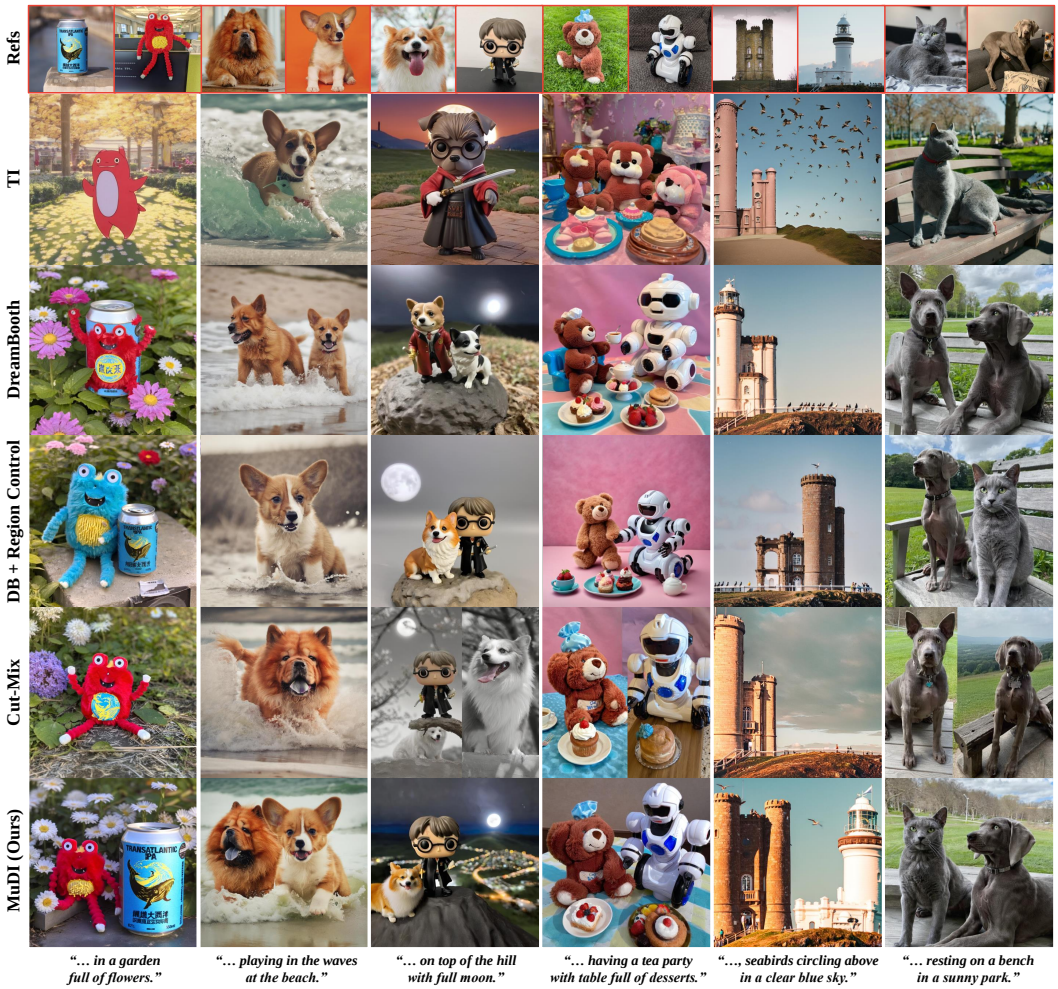


Figure 5: **Qualitative comparison** of Textual Inversion (TI) [13], DreamBooth (DB) [42], DB with region control [14], Cut-Mix [15], and MuDI. Images in the same column are generated with the same random seed. We provide more examples in Figure 19.

the detected subjects resemble the references, resulting in successful identity decoupling. Therefore, we define the *D&C score* as $1 - \|\boldsymbol{S}^{GT} - \boldsymbol{S}^{DC}\|_F^2$. We illustrate an overview of D&C in Figure 4.

To validate that our D&C is capable of measuring multi-subject fidelity, we compare it with previous single-subject fidelity metrics extended to multi-subject settings. These extended metrics compute the mean similarity to all the reference images [24]. For 1600 images generated from various models, we assess the correlation between each metric and human evaluations using Spearman's rank correlation coefficient and the Area under the Receiver Operating Characteristic (AUROC) (see Appendix A.3 for details). Table in Figure 4 shows that D&C with DreamSim (D&C-DS) exhibits the highest correlation with human evaluation. We provide qualitative examples in Figure 14 and Figure 18. These demonstrate that D&C can capture multi-subject fidelity and is suitable as an evaluation metric.

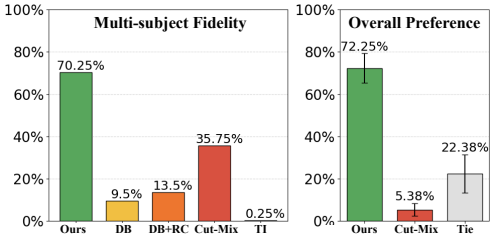
5 Experiments

5.1 Experimental setup

Dataset We construct a new dataset to evaluate the performance of identity decoupling for multi-subject personalization methods. It consists of 8 pairs of similar subjects that are prone to identity mixing. We collected images from the DreamBench dataset [42] and the CustomConcept101 dataset [21], consisting of diverse categories including animals, objects, and scenes. For each pair of subjects, we generate 5 evaluation prompts using ChatGPT [30], which describe scenes involving the subjects with simple actions and backgrounds. We provide more details of the dataset in Appendix A.2.

https://doi.org/10.52202/079017-3200

Figure 6: **(Left) Human evaluation results** on multi-subject fidelity and overall preference. **(Right) Quantitative results** on multi-subject fidelity and text fidelity. † denotes the text fidelity score considering the permutation of the subjects in the prompt to avoid position bias.



Method	Multi-Subject Fidelity		Text Fidelity	
	D&C-DS↑	D&C-DINO↑	ImageReward†↑	CLIPs†↑
TI [13]	0.116	0.132	-0.149	0.227
DB [42]	0.371	0.388	0.579	0.255
DB+Region [14]	0.340	0.379	0.349	0.245
Cut-Mix [15]	0.432	0.460	-0.287	0.225
MuDI (Ours)	**0.637**	**0.610**	**0.770**	**0.263**

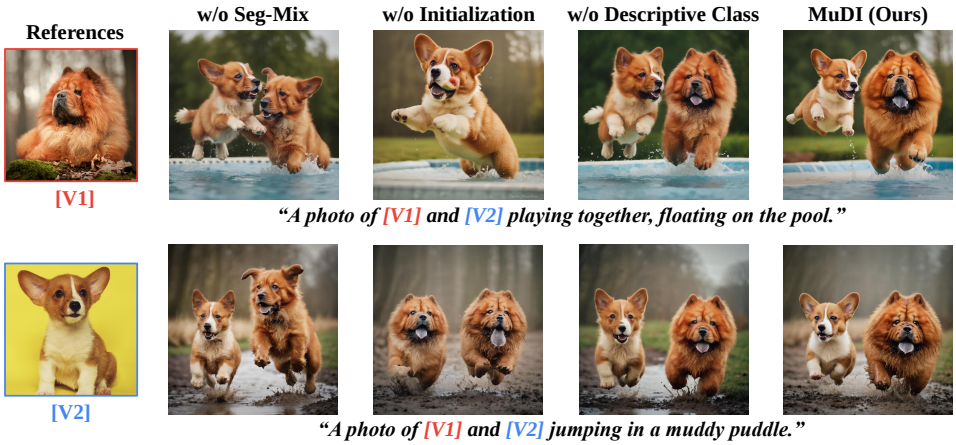


Figure 7: **Ablation Studies** on MuDI. While our method successfully personalizes Corgi and Chow Chow, ablating Seg-Mix results in mixed identity dogs. Inference without our initialization generates images of the subject missing. Training without descriptive class fails to catch subject details.

Implementation details For all experiments, we use Stable Diffusion XL (SDXL) [36] as the pre-trained text-to-image diffusion model and employ a LoRA [18] with a rank of 32 for U-Net [41] module. We also present experiments with other Stable Diffusion models [40] in Appendix B.12 and B.13. For all methods, we pair the reference images with comprehensive captions obtained through GPT-4v [32] which effectively mitigates overfitting to the background and shows better text alignment. We evaluate 400 generated images for each method, across 8 combinations with 5 evaluation prompts and 10 images of fixed random seeds. We provide more details in Appendix A.1.

Baselines We evaluate our method against multi-subject personalization methods: *DreamBooth* [42], DreamBooth with region control [14], DreamBooth using Cut-Mix [15] augmentation, namely *Cut-Mix*, and *Textual Inversion* [13]. Note that we exclude Custom Diffusion [21] from the baselines due to its low quality when applied to SDXL (see Appendix B.1). For both Cut-Mix and Seg-Mix, we use a fixed augmentation probability of 0.3, and we do not use Unmix regularization [15] as it degrades the image quality for SDXL (see Appendix B.10). We describe further details in Appendix A.1.

5.2 Main results

Qualitative comparison As shown in Figure 5, our approach successfully generates the subjects avoiding identity mixing, even for similar subjects such as two dogs (2nd column). On the contrary, DreamBooth results in mixed identities, and using region control proves ineffective for separating identities, as it seldom succeeds and frequently fails. Cut-Mix also falls short of decoupling the identities while producing stitching artifacts. Textual Inversion fails to preserve the subjects' details.

Human evaluation We conduct human evaluations to assess the quality of images generated by the baselines and our method. We first ask human raters to evaluate the multi-subject fidelity via binary feedback. Additionally, we provide reference images of each subject along with two anonymized images: one from MuDI and the other from Cut-Mix. Human raters are asked to indicate which one is better, or tie based on three criteria: (1) similarity to the subjects in the reference images, (2) alignment with the given text, and (3) overall image fidelity. We provide more details in Appendix A.4.

https://doi.org/10.52202/079017-3200

Figure 8: **Personalizing more than two subjects.** (a) MuDI successfully personalizes more than two subjects without identity mixing. (b) Success rates when varying the number of subjects.

(a) Qualitative samples **(b) Success by # of subj.**

(a) Simple prompt **(b) "[V1] and [V2] in <place>"** **(c) "[V1] and [V2] in <style> style"**

Figure 9: **Diverse backgrounds generated by MuDI**. Our Seg-Mix does not have a bias with backgrounds due to the training prompt "A photo of $[V_1]$ and $[V_2]$, simple background". (a) Inference with simple prompts. (b) Inference with various places. (c) Inference with various styles.

As shown in Figure 6 (Left), MuDI significantly outperforms prior works in multi-subject fidelity, achieving twice the success rate in preventing identity mixing compared to Cut-Mix. Due to this, raters strongly prefer images generated by MuDI in side-by-side evaluations. These results confirm that MuDI effectively decouples the identities of highly similar subjects without stitching artifacts.

Quantitative results We evaluate multi-subject personalization methods on two key aspects: *multi-subject fidelity*, which measures the preservation of subject details for multiple subjects, and *text fidelity*, which assesses how well the generated images align with the given text prompt. We use our D&C scores to evaluate multi-subject fidelity. For text fidelity, we report the results of *ImageReward* [54] and CLIP score (*CLIPs*) [37]. To avoid position bias, we calculate scores for the two different orders and average them, for example "$[V_1]$ and $[V_2]$" and "$[V_2]$ and $[V_1]$."

As shown in the Table of Figure 6 (Right), our framework achieves the highest scores in both multi-subject and text fidelity, significantly outperforming previous methods. These results are consistent with qualitative assessments and human evaluations, where MuDI preserves subject details effectively without identity mixing, unlike prior methods. The superior text fidelity also indicates that our method generates images that closely follow the given prompt without mixing the subjects.

5.3 Ablation studies

Necessity of Seg-Mix To validate that Seg-Mix is crucial for decoupling the subjects' identities, we compare MuDI against its variant without it. As shown in Table 1, ablating Seg-Mix results in low multi-subject fidelity due to identity mixing. Figure 7 demonstrates that the attributes of the Corgi and Chow Chow are completely mixed without Seg-Mix. In particular, we show in Figure 26 that using additional spatial conditioning, e.g., ControlNet [56], without Seg-Mix still suffers from identity mixing.

Table 1: Results on ablation studies.

	Multi-Subject Fidelity	
Method	D&C-DS↑	D&C-DINO↑
w/o Seg-Mix	0.475	0.481
w/o Initialization	0.477	0.480
w/o Desc. Class	0.556	0.558
MuDI (Ours)	**0.637**	**0.610**

Figure 10: **Images generated using FLUX [22] as a pre-trained text-to-image diffusion model**. (Top row) DreamBooth produces mixed-identity teddy bears while MuDI generates distinct bears. (Bottom row) MuDI can personalize many subjects without identity mixing on diverse backgrounds.

Importance of our initialization We show in Figure 7 that the inference initialization improves identity separation and alleviates subject dominance. Table 1 validates that images generated without initialization result in lower subject fidelity. We empirically find that our initialization provides significant benefits in three scenarios: (1) personalizing unusual subjects that pre-trained models struggle to generate (e.g., the cloud man of Figure 1 bottom-right), (2) personalizing more than two subjects, and (3) using complex prompts, which we explain in detail in Appendix B.6.

Descriptive class We show in Figure 7 that using descriptive classes to represent the subjects improves the preservation of the subjects' detail, and Table 1 further shows that this method enhances subject fidelity. Despite the improvement, relying only on descriptive classes may occasionally lead to some subjects being ignored. This is effectively addressed by applying our initialization which results in significantly improved outcomes.

More than two subjects Figure 8(b) shows the success rates of MuDI, DreamBooth [42], and Cut-Mix [15] as the number of subjects varies. Our method achieves significantly high success rates, while previous approaches [42, 15] fail to personalize even two subjects effectively. In particular, our method shows over 50% success for generating four objects together (see Figure 8(a)). However, we observe that the performance of MuDI decreases as the number of personalized subjects increases, particularly for highly similar subjects. We provide further details in Appendix B.9.

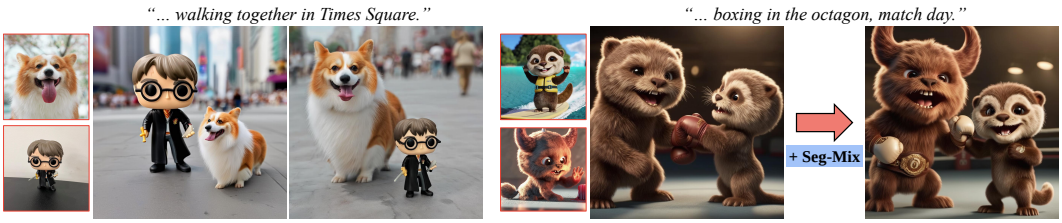
Diverse background As shown in Figure 9, our Seg-Mix does not have a bias with white backgrounds and can generate diverse backgrounds. This is because the prompt "A photo of $[V_1]$ and $[V_2]$, simple background" is used during training for the image of segmented subjects composed on a white background. This effectively disentangles the background from the identities through the text "simple background", preventing overfitting.

Model agnostic Notably, MuDI is a model-agnostic personalization method as it is based on data augmentation during training that does not require model-specific techniques, such as utilizing attention maps [3, 53] or choosing where to fine-tune [21, 48]. We validate this by using FLUX [22] as the pre-trained text-to-image model based on DiT [34], which is different from SDXL [36] based on UNet [41] backbone. As shown in Figure 10 top row, DreamBooth produces mixed identity teddy bears while MuDI successfully generates distinct bears without identity mixing. We show in Figure 10 bottom row that MuDI can personalize multiple subjects using FLUX in diverse backgrounds.

5.4 Other use cases

Controlling relative size Our framework offers an intuitive way to control the relative size between the personalized subjects. By resizing the segmented subjects according to user intents in Seg-Mix,

https://doi.org/10.52202/079017-3200



(a) Controlling Relative Size **(b) Modular Customization with Seg-Mix**

Figure 11: Examples of other use cases of our method. **(a) Controlling relative size with Seg-Mix.** We visualize samples generated by MuDI using size-controlled Seg-Mix. **(b) Modular customization.** Applying Seg-Mix after merging LoRAs significantly improves identity decoupling.

we find that personalized models generate subjects with the desired relative sizes. This showcases another benefit of our method unlike previous methods [42, 21, 15], which often result in inconsistent relative sizes due to a lack of size information during fine-tuning. As shown in Figure 11(a), our method allows the model to be personalized to generate either a larger dog compared to the toy or vice versa, by setting their relative sizes during Seg-Mix. The generated images show a consistent relative size which we provide more examples in Figure 34. Additionally, controlling the relative size of the segmented subjects during inference initialization can further improve the size consistency.

Modular customization The proposed Seg-Mix can also be applied to modular customization, where the subjects are independently learned in advance by single-subject LoRAs [18]. We then efficiently combine these LoRAs to generate multi-subject images. To integrate Seg-Mix with modular customization, we first generate images for each subject using their respective single-subject LoRA, which serve as reference images. Next, we merge the single-subject LoRAs using an existing method such as gradient fusion [14]. After merging, we apply Seg-Mix with the generated single-subject images for 200-300 iterations. This approach effectively reduces identity mixing and avoids the need for training from scratch by reusing the single-subject LoRAs. We illustrate the process of using Seg-Mix in Figure 35.

Figure 11(b) shows samples generated by gradient fusion [14], a modular customization method, applied to two-subject personalization. Without spatial conditioning, it produces mixed identities for the characters of the otter and the monster (left). However, if we fine-tune fused model with Seg-Mix only for a few iterations, the fine-tuned model produces a high-quality image of clearly separated subjects (Figure 11(b), right). We note that it is important to incorporate Kullback-Leibler (KL) divergence as regularization [11] in fine-tuning in order to prevent saturation and overfitting.

5.5 Iterative training

To further improve the quality, we investigate a fully automatic iterative training (IT) method [45], which fine-tunes the personalized model using high-quality samples obtained from an earlier training stage. Specifically, we first generate multi-subject images with MuDI and select high-quality images based on the D&C score, which closely aligns with the human evaluation. These selected images are then used to fine-tune the personalized model, with KL regularization [11] added to Eq. (2). By applying IT to the images of Corgi and Chow Chow, the D&C-DS score is improved from 0.613 to 0.672, achieving a higher success rate (see Figure 37). We provide further details in Appendix C.3.

6 Conclusions

In this work, we present MuDI, a novel personalizing framework for multiple subjects that addresses identity mixing. We leverage segmented subjects automatically obtained from the foundation model for image segmentation for both training and inference, through data augmentation for training pre-trained models and initializing the generation process. We experimentally validate our approach on a new dataset comprising combinations of subjects prone to identity mixing, for which ours successfully prevents mixing even for highly similar subjects. We describe the limitations and societal impacts of our work in Appendix D. We hope that our work can serve as a starting point to develop personalizing methods for multiple concepts in more challenging scenarios.

https://doi.org/10.52202/079017-3200

7 Acknowledgements

We thank Juyong Lee, and Jaewoo Lee for providing valuable feedback. This work was supported by National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. RS-2023-00256259), Institute for Information & communications Technology Promotion(IITP) grant funded by the Korea government(MSIT) (No.RS-2019-II190075 Artificial Intelligence Graduate School Program(KAIST)), Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (No. RS-2024-00509279 Global AI Frontier Lab), Artificial intelligence industrial convergence cluster development project funded by the Ministry of Science and ICT(MSIT, Korea)&Gwangju Metropolitan City, and KAIST-NAVER Hypercreative AI Center.

References

[1] Yuval Alaluf, Elad Richardson, Gal Metzer, and Daniel Cohen-Or. A neural space-time representation for text-to-image personalization. *Association for Computing Machinery Transactions on Graphics*, 42(6):243:1–243:10, 2023.

[2] Moab Arar, Andrey Voynov, Amir Hertz, Omri Avrahami, Shlomi Fruchter, Yael Pritch, Daniel Cohen-Or, and Ariel Shamir. Palp: Prompt aligned personalization of text-to-image models. *arXiv:2401.06105*, 2024.

[3] 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*, 2023.

[4] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, Tero Karras, and Ming-Yu Liu. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv:2211.01324*, 2022.

[5] 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*, 2023.

[6] Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. Training diffusion models with reinforcement learning. In *International Conference on Learning Representations*, 2024.

[7] Daewon Chae, Nokyung Park, Jinkyu Kim, and Kimin Lee. Instructbooth: Instruction-following personalized text-to-image generation. *arXiv:2312.03011*, 2023.

[8] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *Conference on Computer Vision and Pattern Recognition*, 2024.

[9] Jaemin Cho, Abhay Zala, and Mohit Bansal. Visual programming for step-by-step text-to-image generation and evaluation. In *Advances in Neural Information Processing Systems*, 2023.

[10] Kevin Clark, Paul Vicol, Kevin Swersky, and David J. Fleet. Directly fine-tuning diffusion models on differentiable rewards. In *International Conference on Learning Representations*, 2024.

[11] Ying Fan, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. Reinforcement learning for fine-tuning text-to-image diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

[12] 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. In *Advances in Neural Information Processing Systems*, 2023.

[13] 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 *International Conference on Learning Representations*, 2023.

 https://doi.org/10.52202/079017-3200

[14] Yuchao Gu, Xintao Wang, Jay Zhangjie Wu, Yujun Shi, Yunpeng Chen, Zihan Fan, Wuyou Xiao, Rui Zhao, Shuning Chang, Weijia Wu, Yixiao Ge, Ying Shan, and Mike Zheng Shou. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

[15] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris N. Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. In *International Conference on Computer Vision*, 2023.

[16] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations*, 2023.

[17] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems*, 2020.

[18] 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. In *International Conference on Learning Representations*, 2022.

[19] Siteng Huang, Biao Gong, Yutong Feng, Xi Chen, Yuqian Fu, Yu Liu, and Donglin Wang. Learning disentangled identifiers for action-customized text-to-image generation. In *Conference on Computer Vision and Pattern Recognition*, 2024.

[20] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloé Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross B. Girshick. Segment anything. In *International Conference on Computer Vision*, 2023.

[21] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Conference on Computer Vision and Pattern Recognition*, 2023.

[22] Black Forest Labs. Flux, 2024. URL https://github.com/black-forest-labs/flux/.

[23] Kimin Lee, Hao Liu, Moonkyung Ryu, Olivia Watkins, Yuqing Du, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, and Shixiang Shane Gu. Aligning text-to-image models using human feedback. *arXiv.2302.12192*, 2023.

[24] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation. In *International Conference on Machine Learning*, 2023.

[25] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable image synthesis with multiple subjects. In *Advances in Neural Information Processing Systems*, 2023.

[26] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

[27] Jiafeng Mao, Xueting Wang, and Kiyoharu Aizawa. Guided image synthesis via initial image editing in diffusion model. In *Association for Computing Machinery International Conference on Multimedia*, 2023.

[28] Matthias Minderer, Alexey A. Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, 2023.

[29] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, Jian Zhang, Zhongang Qi, and Ying Shan. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. In *Association for the Advancement of Artificial Intelligence*, 2024.

[30] OpenAi. Chatgpt, 2023. URL https://chat.openai.com/.

[31] OpenAi. Sora: Video generation models as world simulators, 2024. URL https://openai.com/sora.

[32] OpenAi. Gpt-4v(ision) technical work and authors, 2024. URL https://openai.com/contributions/gpt-4v.

[33] 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 G. Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024.

[34] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision*, 2023.

[35] Ryan Po, Guandao Yang, Kfir Aberman, and Gordon Wetzstein. Orthogonal adaptation for modular customization of diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2024.

[36] 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. In *International Conference on Learning Representations*, 2024.

[37] 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. In *International Conference on Machine Learning*, 2021.

[38] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems*, 2023.

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

[40] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Conference on Computer Vision and Pattern Recognition*, 2022.

[41] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2015.

[42] 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 *Conference on Computer Vision and Pattern Recognition*, 2023.

[43] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems*, 2022.

[44] Dvir Samuel, Rami Ben-Ari, Simon Raviv, Nir Darshan, and Gal Chechik. Generating images of rare concepts using pre-trained diffusion models. In *Association for the Advancement of Artificial Intelligence*, 2024.

[45] Kihyuk Sohn, Lu Jiang, Jarred Barber, Kimin Lee, Nataniel Ruiz, Dilip Krishnan, Huiwen Chang, Yuanzhen Li, Irfan Essa, Michael Rubinstein, Yuan Hao, Glenn Entis, Irina Blok, and Daniel Castro Chin. Styledrop: Text-to-image synthesis of any style. In *Advances in Neural Information Processing Systems*, 2023.

[46] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. In *International Conference on Learning Representations*, 2021.

 https://doi.org/10.52202/079017-3200

[47] Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E. Jacobs, Bharath Hariharan, Yael Pritch, Neal Wadhwa, Kfir Aberman, and Michael Rubinstein. Realfill: Reference-driven generation for authentic image completion. *arXiv:2309.16668*, 2023.

[48] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-locked rank one editing for text-to-image personalization. In Erik Brunvand, Alla Sheffer, and Michael Wimmer, editors, *Association for Computing Machinery Special Interest Group on Computer Graphics and Interactive Techniques*, 2023.

[49] Hazarapet Tunanyan, Dejia Xu, Shant Navasardyan, Zhangyang Wang, and Humphrey Shi. Multi-concept t2i-zero: Tweaking only the text embeddings and nothing else. *arXiv:2310.07419*, 2023.

[50] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: extended textual conditioning in text-to-image generation. *arXiv:2303.09522*, 2023.

[51] Bram Wallace, Meihua Dang, Rafael Rafailov, Linqi Zhou, Aaron Lou, Senthil Purushwalkam, Stefano Ermon, Caiming Xiong, Shafiq Joty, and Nikhil Naik. Diffusion model alignment using direct preference optimization. *arXiv:2311.12908*, 2023.

[52] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo. ELITE: encoding visual concepts into textual embeddings for customized text-to-image generation. In *International Conference on Computer Vision*, 2023.

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

[54] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems*, 2023.

[55] Sangdoo Yun, Dongyoon Han, Sanghyuk Chun, Seong Joon Oh, Youngjoon Yoo, and Junsuk Choe. Cutmix: Regularization strategy to train strong classifiers with localizable features. In *International Conference on Computer Vision*, 2019.

[56] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *International Conference on Computer Vision*, 2023.

Appendix

Organization The Appendix is organized as follows: In Section A, we describe the details of the experiments and our framework. We provide additional experimental results in Section B, and further discussion on other use cases of MuDI in Section C. Lastly, in Section D, we discuss the limitations and societal impacts of our work.

A Experimental details

A.1 Implementation details

Training details In our experiment, we use Stable Diffusion XL (SDXL) [36] as the pre-trained text-to-image diffusion model. We employ LoRA [18] with a rank of 32 for the U-Net [41] module, instead of training the full model weights. We do not train the text encoder.

For all methods, we pair the reference images with comprehensive captions obtained through GPT-4v [32], instead of using a simple prompt like "A photo of a [V]". This effectively mitigates overfitting to the background and shows better text alignment for the baselines and our method.

We construct the prior dataset $\mathcal{D}_{ref}$ in Eq. (2) by generating from the pre-trained text-to-image models using a prompt *"A photo of <class>, simple background, full body shot."*. Since the generated prior images may contain more than one subject, we select images that contain a single subject.

We determine the training iterations of Seg-Mix on each combination in the dataset based on the difficulty of personalizing the subjects individually using DreamBooth [42]. For example, combinations including highly detailed subjects, such as "can" in Figure 5, require from 1400 to 1600 training iterations, while combinations containing subjects easy to learn, e.g., "dog," require about 1200 iterations. We use a fixed augmentation probability of 0.3 for both Cut-Mix and our Seg-Mix with the same number of training iterations for a fair comparison. To prevent subject overfitting, we use 1000 training iterations for DreamBooth.

Algorithm 1 MuDI training (Seg-Mix)

```
# preprocess_data: list of all reference images & masks pair ([(imgs_0, masks_0), ...])
# max_m: max margin from both ends of the image (if large, allow overlap)
# scales: control the relative size else random resizing

def create_seg_mix(imgs, masks, out_size=(1024,1024), max_m=1, scales=None):
    imgs, masks = random.choice(preprocess_data, 2, replace=False) # sample 2 refs
    # randomly(or relative) resize each image & mask pair
    imgs, masks = resize(imgs, masks, out_size, scales)
    out, out_mask = np.zeros((*out_size, 3)), np.zeros(out_size) # blank image, mask
    if random.random() < 0.5: # random order swap
        imgs, masks = imgs[::-1], masks[::-1]
    # random margin from ends of the image
    m = [random.randint(0, max_margin) _ for in range(2)]
    out, out_mask = paste_left(out, out_mask, imgs[0] * masks[0], m[0])
    out, out_mask = paste_right(out, out_mask, imgs[1] * masks[1], m[1])
    return out, out_mask

def train_loss(seg_mix_prob=0.3, **kwargs):
    img, mask, class, prompt = dataloader.next()
    if random.random() < seg_mix_prob: # do augmentation
        # sample another class for seg-mix
        new_class = random.choice(class_list - class)
        new_img, new_mask = sample_img_mask(new_class)
        imgs, masks = [img, new_img], [mask, new_mask]
        img, mask = create_seg_mix(imgs, masks, **kwargs)
    # DreamBooth training (Eq. 1)
    loss = LDM_loss(img, prompt)
    return loss.mean()
```


https://doi.org/10.52202/079017-3200

Algorithm 2 MuDI inference (Initialization)

```
# class_list: classes in prompt
# gamma: guidance strength of our initialization
# kwargs: same arguments from Algorithm 1

def latent_initialize(class_list, gamma=1.0, **kwargs):
    # sample reference images
    imgs, masks = zip(*[sample_img_mask(cls) for cls in class_list])
    # create_seg_mix from Algorithm 1
    out, mask = create_seg_mix(imgs, masks, **kwargs)
    # encode image to latent
    out_latent = encoder(out)
    # resize to latent size
    out_mask = resize(out_mask)
    # segmented latent
    init_latent = out_mask * out_latent
    noise = torch.rand_like(init_latent)
    # forward process with gamma scaling
    init_latent = add_noise(init_latent * gamma, strength=1, noise=noise)
    return init_latent

def inference(prompt, class_list, gamma=1.0, **kwargs):
    # execute computation only once
    init_latent = latent_initialize(class_list, gamma, **kwargs)
    # existing inference pipeline
    img = inference_pipe(prompt, init_latent=init_latent, **kwargs)
    return img
```

Figure 12: **Cut-Mix with and without negative prompt**. We observe that using the negative prompt "A dog and a dog" leads to reduced artifacts but results in over-saturation as shown in the first row.

MuDI training details We provide a pseudocode of our MuDI training in Algorithm 1. For Seg-Mix, we randomly rescale the segmented subjects and randomly choose the location of the subjects to create random compositions of the segmented subjects. Note that the margin from the boundaries of the image is also set as random, where larger margins allow for the possibility of subjects overlapping.

In particular, when creating the prior dataset $\mathcal{D}_{ref}$ for training, we use a descriptive class (Section 4.1) to serve as the prior class. We automatically create segmentation masks for the images in the prior dataset using OWLv2 [28] and SAM [20], similar to segmenting the reference images described in Section 4.1. We illustrate the segmentation of the prior dataset in Figure 3 (denoted as Priors). Note that we select images that contain a single subject and also result in a single segmentation mask. We generated 50 images for the prior dataset.

After the preprocessing step, the training of MuDI takes almost the same duration as DreamBooth [42], taking about 90 minutes to personalize two subjects on a single RTX 3090 GPU. We use AdamW optimizer [26] with $\beta_1 = 0.9$, $\beta_2 = 0.999$, weight decay of 0.0001, and a learning rate of 1e-4, following the setting of DreamBooth [42], and set the batch size to 2.

Figure 13: **Dataset.** We introduce a new dataset comprising eight combinations of similar subjects. For each combination, we visualize one image per subject (red boxes) and three images generated by DreamBooth. The score below the subjects denotes the DreamSim [12] similarity score between 0 and 1, where a larger value indicates higher similarity. The bottom-most two combinations have the highest similarity which makes them challenging to personalize without mixing the identities.

MuDI inference details We provide a pseudocode of our inference initialization in Algorithm 2. We first create images of randomly composed segmented subjects using the reference images and the extracted segmentation masks. The composition can be either random, manually set, or obtained by using LLM as described in Section 4.2. The images are then encoded into a latent with the VAE of the SDXL [36], which is scaled by a factor of γ. The scaled latent is perturbed by the forward noising process from time 0 to T, which results in the initial latent for the inference process. We control the magnitude of the γ scale and the relative size between the segmented subjects to address identity mixing as well as subject dominance.

Inference using negative prompt While Han et al. [15] propose using negative prompts to reduce stitching artifacts, we observe that this produces over-saturated samples. As shown in the top row of Figure 12, using negative prompts results in low-quality images. Therefore, we opt not to use negative prompts in the evaluation of Cut-Mix [15]. For our framework, we use a simple negative prompt *"sticker, collage."* that alleviates sticker-like artifacts caused by over-training with the segment-and-mixed images.

A.2 Dataset

We introduce a new dataset to facilitate evaluation for multi-subject personalization, comprising 8 combinations of similar subjects prone to identity mixing. We collected images from the datasets widely used, namely the DreamBench dataset [42] and the CustomConcept101 dataset [21]. We construct the dataset to comprise diverse categories including animals, objects, and scenes. We visualize the subjects and the identity-mixed samples from DreamBooth in Figure 13.

https://doi.org/10.52202/079017-3200

Algorithm 3 Detect & Compare (D&C)

Input: Embeddings of reference subjects $(R_1, R_2, ...R_N)$, Embeddings of boxes $(B_1, B_2, ...B_M)$
Output: D&C score ($0 \sim 1$, it depends on similarity between references)

if M != N **then** ▷ Count Error, return 0
  return 0
else
  $\boldsymbol{S}^{GT}, \boldsymbol{S}^{DC} = [0]_{N \times N}, [0]_{N \times N}$ ▷ Initialize square matrix
  for $i = 1, 2, \dots N$ **do**
    for $j = 1, 2, \dots N$ **do**
      $\boldsymbol{S}^{GT}_{ij} = \text{mean}(\text{matmul}(R_i, R_j))$ ▷ This can be pre-calculated before, and has symmetric property
      $\boldsymbol{S}^{DC}_{ij} = \text{mean}(\text{matmul}(B_i, R_j))$
    end for
  end for
end if
$\boldsymbol{S}^{DC}$ = row-wise-sort($\boldsymbol{S}^{DC}$) ▷ Sort the rows based on the maximum value of each column sequentially
score $= 1 - \|\boldsymbol{S}^{GT} - \boldsymbol{S}^{DC}\|_F^2$
return score

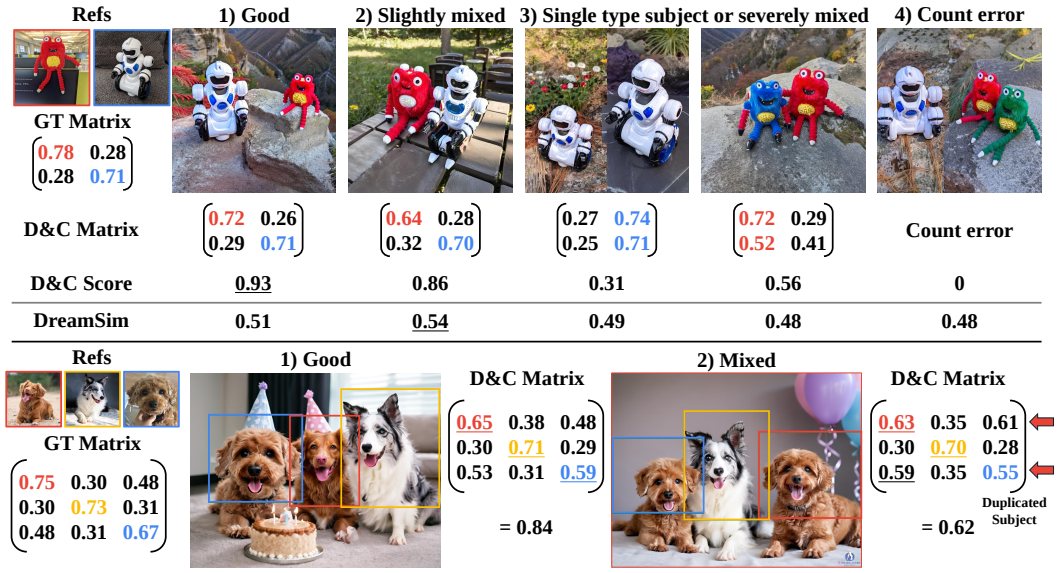


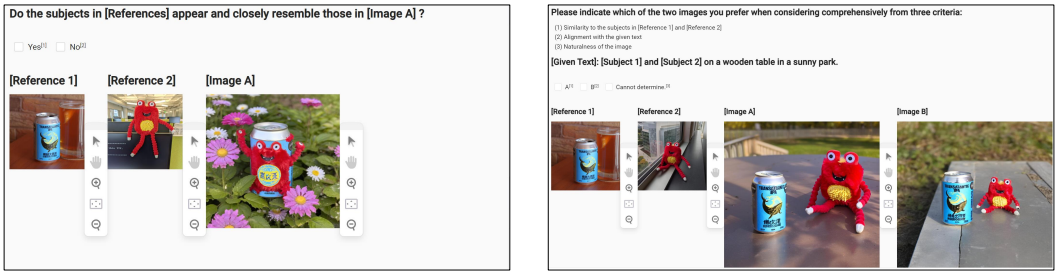
Figure 14: **(Top row)** We visualize the D&C-DS scores and DreamSim scores for various cases. **(Bottom row)** We provide examples of D&C-DS scores for images of three subjects.

A.3 Detect-and-Compare

We summarize the process of measuring the D&C score in Algorithm 3.

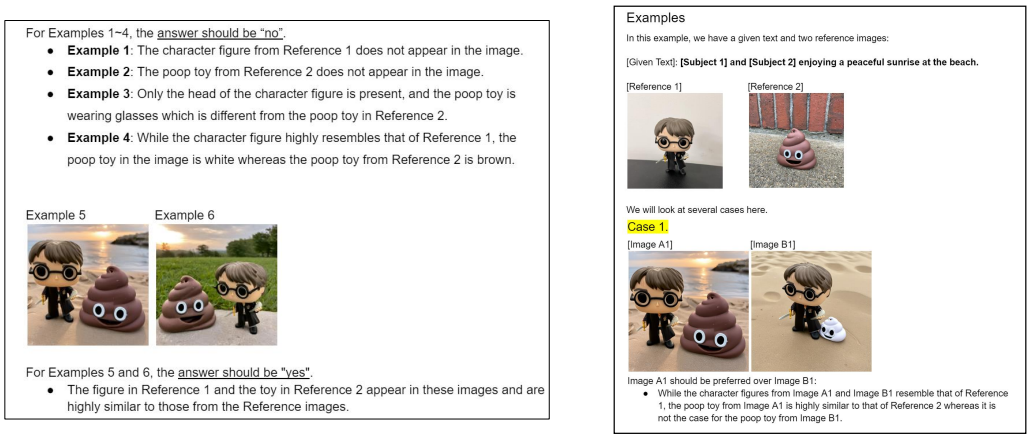
Correlation with human evaluation In the table of Figure 4, we assess the correlation between human evaluation and the metrics (D&C scores and DreamSim scores). We generate a total of 1600 images, 400 images from Textual Inversion [13], DreamBooth [42], Cut-Mix [15], and our MuDI, respectively, and ask human raters to evaluate the multi-subject fidelity via binary feedback. We then measure the correlation between human evaluation results and the scores using Spearman's rank correlation coefficient and Area under the Receiver Operating Characteristic (AUROC). Note that the Spearman's rank correlation was computed using the normalized sum of all human evaluation answers (e.g., if 3 out of 5 raters answered "good," the score is 0.6). The AUROC was computed based on the majority voting results (0 or 1).

Examples In the top row of Figure 14, we categorize the generated images from DreamBooth [42], Cut-Mix [15], and MuDI into four cases, and provide the D&C matrix, D&C score, and DreamSim score for each image. For successful images, the difference between $\boldsymbol{S}^{GT}$ and $\boldsymbol{S}^{DC}$ is significantly small and results in high D&C scores. In cases where the identities are severely mixed or show two



(a) Multi-Subject Fidelity **(b) Overall Preference**

Figure 15: A screenshot of questionnaires from our human evaluation on (a) multi-subject fidelity and (b) overall preference.



(a) Labeling instruction for multi-subject fidelity **(b) Labeling instruction for overall preference**

Figure 16: A screenshot of labeling instruction from our human evaluation on (a) multi-subject fidelity and (b) overall preference.

identical subjects, the difference becomes considerably larger, and the D&C scores decrease. For example, the fourth generated image features two monster toys, where one is blue and the other is red. The blue monster toy resembles both the reference monster toy and the robot toy, leading to a significant difference in the second row of $\boldsymbol{S}^{GT}$ and $\boldsymbol{S}^{DC}$. However, DreamSim [12] extended to multi-subject settings, which compute the mean similarity to all the reference images [24], fails to distinguish between these cases effectively.

We also provide an example of D&C matrices and scores for three subjects in the bottom row of Figure 14. Our D&C can be easily applied to evaluate identity mixing for many subjects.

Qualitative comparison of D&C and DreamSim Additionally, we analyze the alignment of D&C to the human evaluation by comparing with DreamSim in Figure 18. We sort 24 images generated by MuDI based on the D&C-DS scores and DreamSim scores, respectively, and compare the ranking with the human evaluation. D&C perfectly aligns with the human evaluation, giving lower scores to failed images with mixed identities. However, the single-subject metric DreamSim fails to align with human evaluation, giving high scores to images with mixed identities or the wrong number of subjects. This indicates that single-subject metrics are ill-suited to be used for evaluating multi-subject fidelity. The qualitative comparison agrees with the quantitative analysis in Figure 4, where we show that D&C achieves a high correlation with the human evaluation of multi-subject fidelity.

A.4 Evaluation details

We evaluate 400 generated images for each method, across 8 combinations with 5 evaluation prompts and 10 images of fixed random seeds.

https://doi.org/10.52202/079017-3200

Monster toy & Can

[V1] and [V2] on a wooden table in a sunny park.
[V1] and [V2] at the beach under a bright sun.
[V1] and [V2] in a classroom with colorful chairs.
[V1] and [V2] in a garden full of flowers.
[V1] and [V2] placed on a tree trunk with fallen leaves.

Cat & Dog

[V1] and [V2] walking in a bustling city street corner.
[V1] and [V2] sitting on a snowy path in a quiet forest.
[V1] and [V2] watching a beautiful sunrise at the beach.
[V1] and [V2] resting on a bench in a sunny park.
[V1] and [V2] playing in a garden full of flowers.

Teddy bear & Robot toy

[V1] and [V2] having a tea party with table full of desserts in a colorful room.
[V1] and [V2] on a wooden bench in a sunny park.
[V1] and [V2] sitting on a cliff at the top of a mountain.
[V1] and [V2] resting on a bed inside a cozy house.
[V1] and [V2] in the woods with fallen leaves.

Monster toy & Robot toy

[V1] and [V2] on a wooden table in a sunny park.
[V1] and [V2] at the beach under a bright sun.
[V1] and [V2] in a classroom with colorful chairs.
[V1] and [V2] in a garden full of flowers.
[V1] and [V2] placed on a tree trunk with fallen leaves.

Dog & Action figure

[V1] and [V2] watching a beautiful sunset at the beach.
[V1] and [V2] in a garden full of flowers.
[V1] and [V2] in a bustling street corner.
[V1] and [V2] on a snowy path in a quiet forest.
[V1] and [V2] on top of the hill with full moon.

Castle & Lighthouse

[V1] is next to [V2], people watching fireworks.
[V1] is next to [V2], dawn's early light softening the rugged landscape.
[V1] is next to [V2], seabirds circling above in a clear blue sky.
[V1] is next to [V2], with a beautiful sunset casting a warm glow.
[V1] is next to [V2], stars twinkling in the night sky.

Teddy bear & Teddy bear

[V1] and [V2] enjoying a picnic.
[V1] and [V2] dancing together in a disco party.
[V1] and [V2] sitting on a stone bench.
[V1] and [V2] on a book shelf.
[V1] and [V2] in a jungle.

Dog & Dog

[[V1] and [V2] playing in the waves at the beach.
[V1] and [V2] running down the stairs.
[V1] and [V2] walking in a bustling city street corner.
[V1] and [V2] watching a sunrise at the cliff.
[V1] and [V2] sitting on a wooden bench at a sunny park.

Figure 17: Our evaluation prompts for each concept.

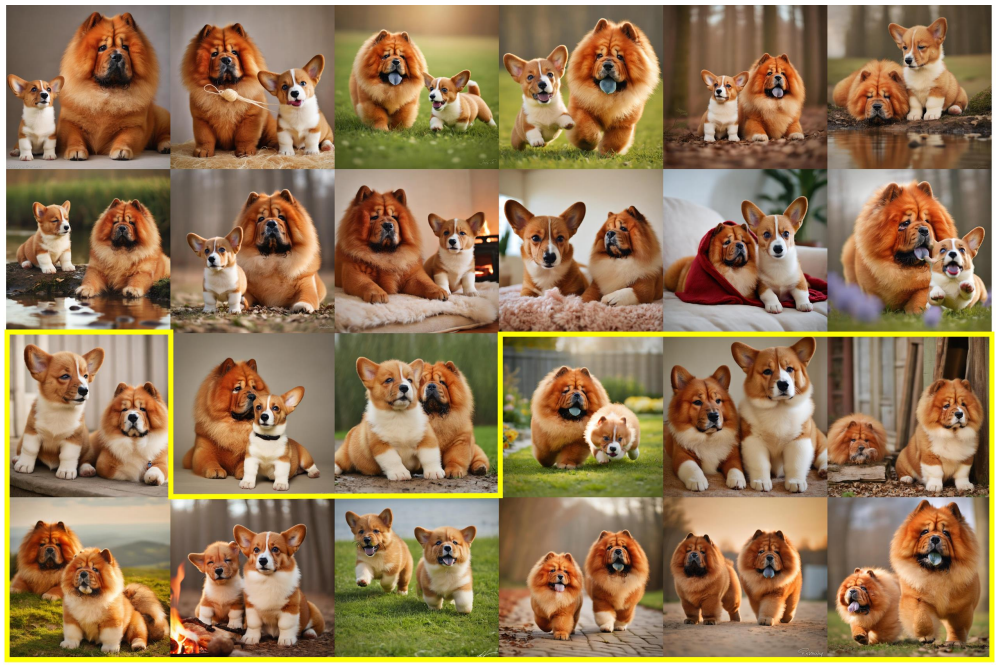
Evaluation prompts To evaluate the personalization methods, we generate 5 evaluation prompts for each combination in the dataset using ChatGPT [30]. Each prompt describes a scene of the subjects with simple action such as *"... in a classroom with colorful chairs,"* or *"... walking in a bustling city street corner."*. We avoid using complex prompts as models fail to generate images that align with such prompts, regardless of the identity mixing. We note that the evaluation prompts were unseen during training. Details of our evaluation prompts for each concept are in Figure 17.

Quantitative evaluation For evaluating the text fidelity, we made fair comparisons by using descriptive classes for both evaluation and text prompts. To avoid positional bias from the order of subjects in the prompts, we measured the scores for all possible orders of the subjects, for example, "monster toy and can" and "can and monster toy", and reported the average of the scores. In the case of ImageReward [54], the score differences between the different subject orders were not small.

Human evaluation Our human evaluation was conducted in two main aspects: (1) multi-subject fidelity and (2) overall preference. For multi-subject fidelity, a random subset containing an equal number of instances from each method was created and provided to human raters for binary feedback. For overall preference, images from both Cut-Mix and ours were provided in random order, with all images generated from the same seed. We provide reference images of each subject along with two anonymized images, i.e., one from MuDI and the other from Cut-Mix We ask human raters to evaluate which image they prefer based on three criteria: (1) similarity to the subjects in the reference images, (2) alignment with the given text, and (3) image fidelity. If both images fail to depict the subjects in the reference images, raters are instructed to select "cannot determine". We provide screenshots of questionnaires and labeling instructions in Figure 15 and Figure 16, respectively.

A.5 Additional generated examples

We provide additional non-curated generated examples in Figure 19.



(a) D&C-DS

(b) D&C-DS (Max)

Figure 18: **Qualitative comparison of D&C-DS and DreamSim [12].** We sort 24 images generated by MuDI based on (a) D&C-DS and (b) DreamSim. The highest-scored image is placed at the top left, with scores decreasing progressively towards the bottom right. Note that for DreamSim similarity, we take the average of the similarities to all the reference images. The yellow boxes indicate failed images evaluated by human raters, for instance, mixed identity dogs or a dog is missing. We observe that D&C score perfectly aligns with the human evaluation, yielding the failed images lower scores than the successful images. On the other hand, DreamSim does not align with human evaluation where the failed images are ranked high.

https://doi.org/10.52202/079017-3200

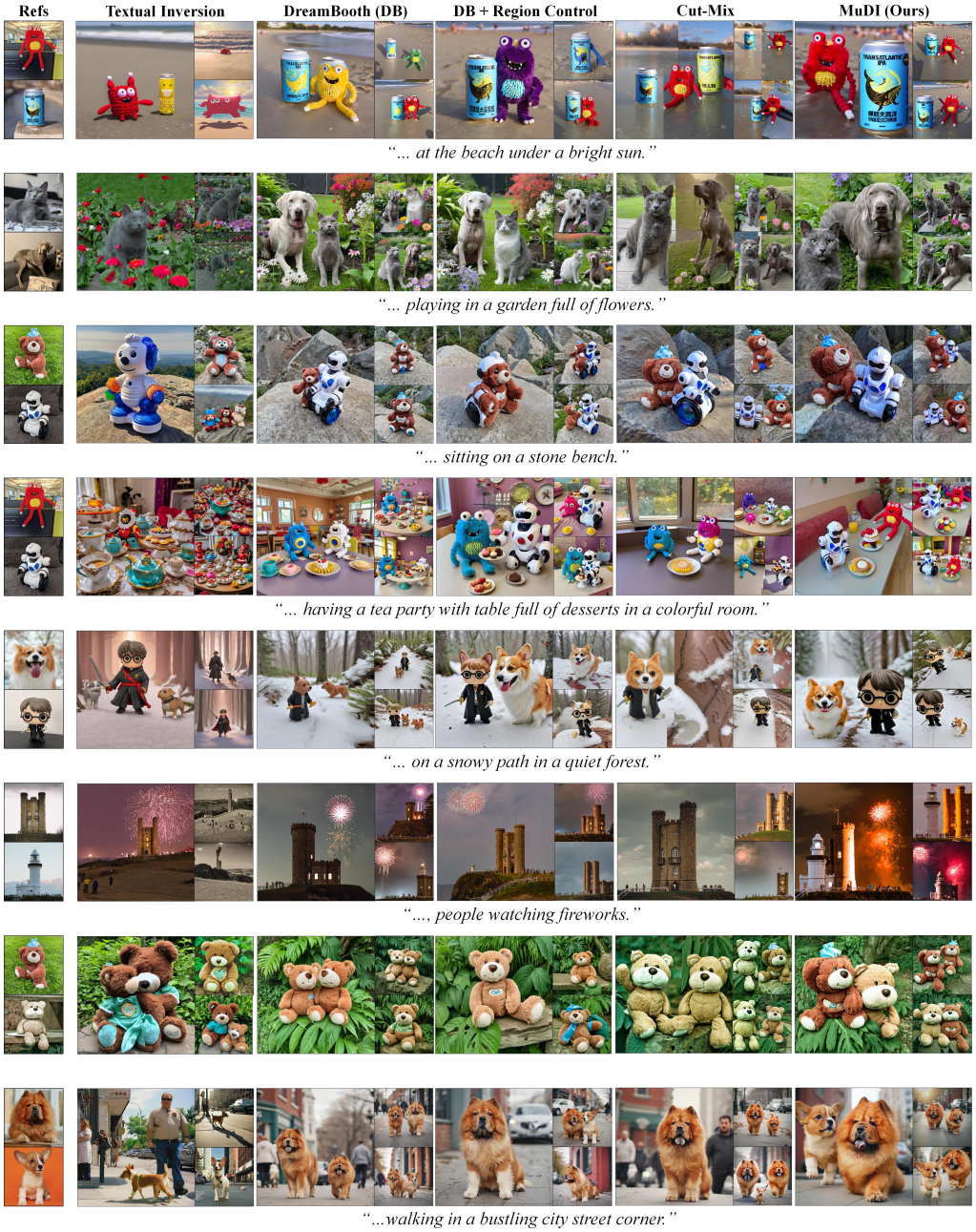
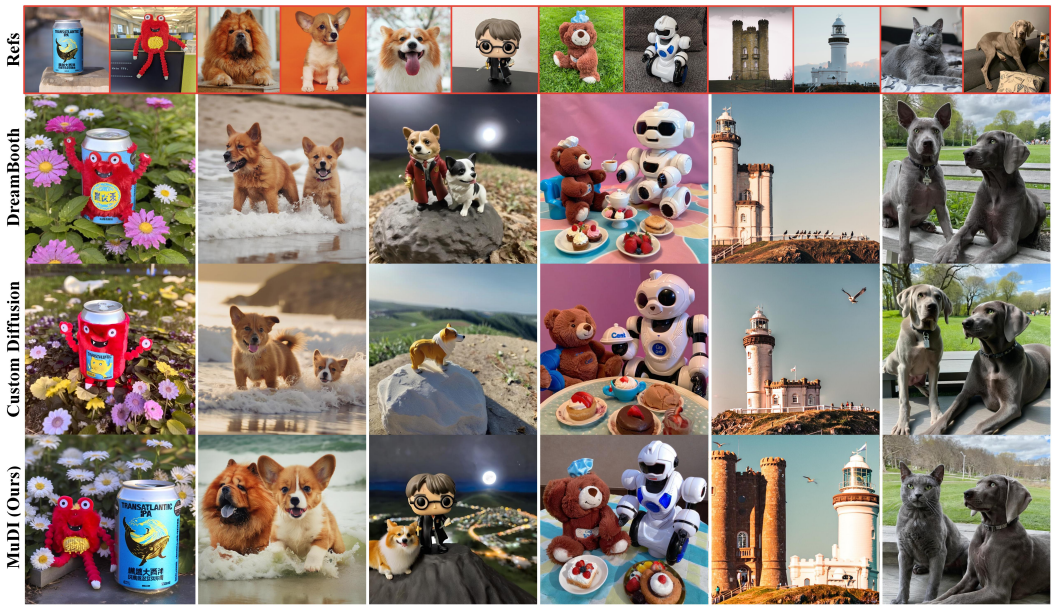


Figure 19: **Qualitative comparison** of images generated by Textual Inversion [13], DreamBooth (DB) [42], DreamBooth with region control [14], Cut-Mix [15], and our MuDI. We visualize non-curated images generated with the same random seed.

Figure 20: **Qualitative comparison** of images generated by DreamBooth [42], Custom Diffusion [21], and MuDI. Similar to DreamBooth, Custom Diffusion results in identity mixing.



	Multi-Subject Fidelity		Text Fidelity	
Method	D&C-DS↑	D&C-DINO↑	ImageReward†↑	CLIPs†↑
Textual Inversion [13]	0.116	0.132	-0.149	0.227
DreamBooth [42]	0.371	0.388	0.579	0.255
Custom Diffusion [21]	0.353	0.389	0.144	0.243
Cut-Mix [15]	0.432	0.460	-0.287	0.225
MuDI (Ours)	**0.637**	**0.610**	**0.770**	**0.263**

Table 2: **Quantitative results** on multi-subject fidelity and text fidelity. † denotes the text fidelity score considering the permutation of the subjects in the prompt to avoid position bias.

B Additional experimental results

B.1 Comparison with Custom Diffusion

Here, we provide the results of Custom Diffusion [21] that uses SDXL [36] as the pre-trained text-to-image diffusion model. Due to GPU constraints, we fine-tune the weights of LoRA [18] instead of directly fine-tuning the model weights. We evaluate two different models, one that uses a high rank (i.e., rank 128) and the other that uses the same rank as ours (i.e., rank 32). However, we do not observe significant differences between them.

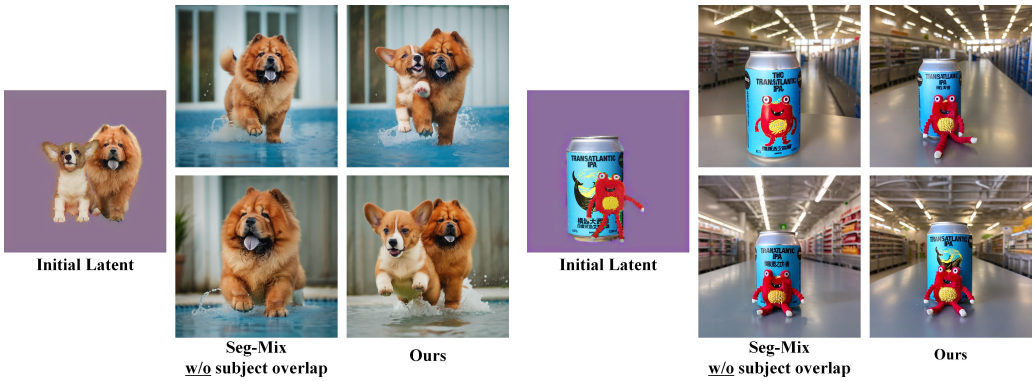
As shown in Figure 20, Custom Diffusion demonstrates degradation in the subject fidelity compared to DreamBooth [42]. The quantitative results in Table 2 similarly show that Custom Diffusion results in lower multi-subject fidelity as well as lower text fidelity compared to DreamBooth and MuDI. Due to the degradation, we exclude Custom Diffusion from our baseline in the main experiments.

B.2 Importance of subject overlap in Seg-Mix

As described in Section 4.1, our data augmentation method, Seg-Mix, allows subjects to be overlapped when randomly positioning the segmented subjects (see Figure 3 upper right). This differs from Cut-Mix [15] which is restricted to augmenting images of non-overlapped subjects. Here, we verify that training the text-to-image models with images of overlapped subjects is crucial for generating interaction between the subjects. In Figure 21, we qualitatively compare our MuDI with its variant that is trained with Seg-Mix which does not allow subject overlap during data augmentation, namely

https://doi.org/10.52202/079017-3200

Figure 21: **Ablation study on Subject Overlap for our Seg-Mix**. We compare MuDI against its variant trained with Seg-Mix which does not allow subject overlap during data augmentation, i.e., Seg-Mix w/o subject overlap. MuDI successfully personalizes the subjects with natural interaction. However, Seg-Mix without subject overlap results in identity mixing and subject ignorance.



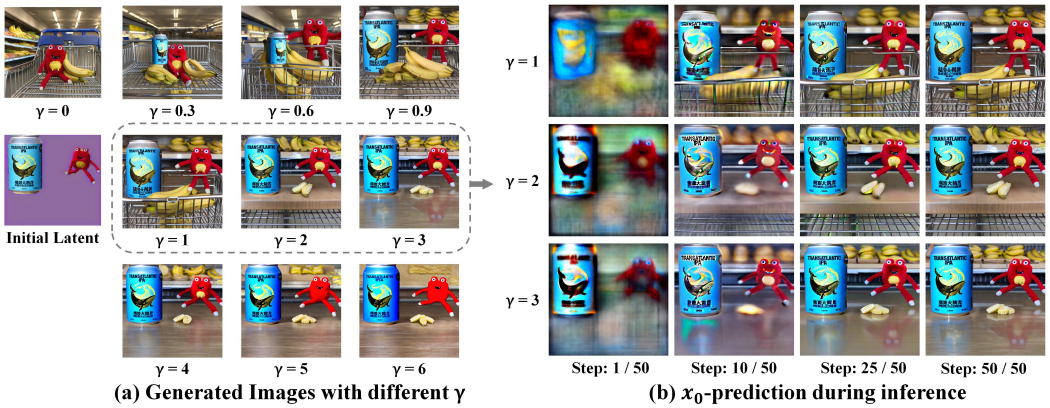


Figure 22: **Analysis on γ-scaling for inference initialization. (a) Generated Images for varying γ.** A larger scale γ results in more information preserved from the initial latent. **(b) x_0-prediction through inference steps.** The image except for the fine details is determined in the first 10 steps. Thus providing information from the start via inference initialization plays a critical role in generating successful multi-subject composition.

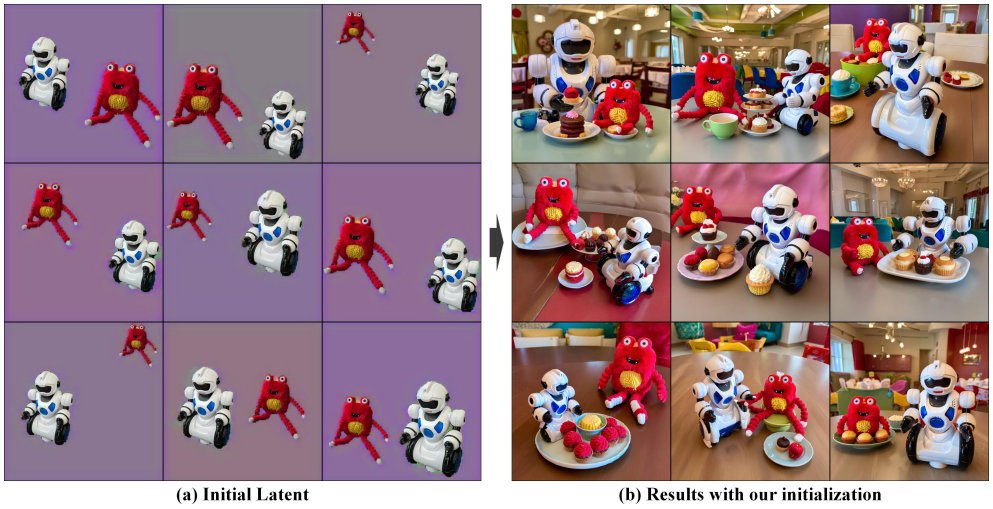
Seg-Mix w/o subject overlap. Our MuDI successfully personalizes the subjects distinctly with natural close-distance interaction, for example, two dogs playing in the pool. In contrast, Seg-Mix w/o subject overlap produces mixed identities for neighboring subjects (e.g., monster toy in the can) or subject ignorance (e.g., generating only the Chow Chow while ignoring the Corgi).

B.3 Analysis on γ-scaling for inference initialization

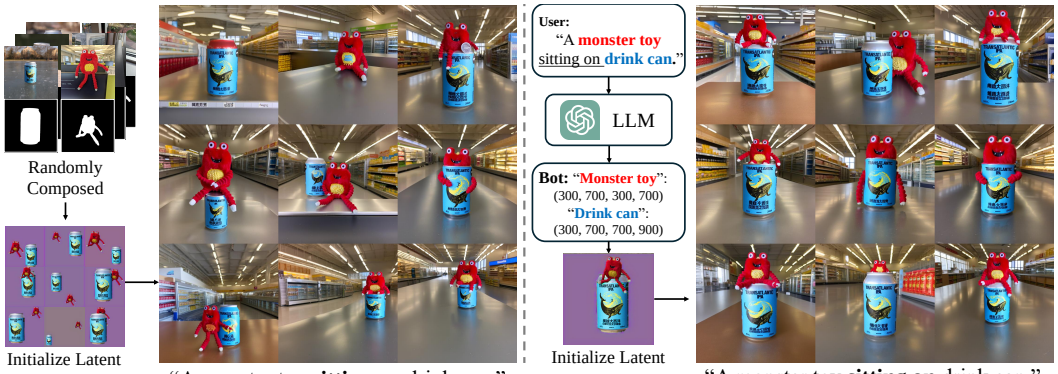
Here, we analyze the effect of γ-scaling for our initialization by varying the magnitude of γ for generating samples. As demonstrated in Figure 22(a), without inference initialization (i.e., $\gamma = 0$) it results in an image independent of the initial latent, while the larger scale of γ yields images with layouts of subjects similar to the initial latent. Empirically, we observe that γ exceeding 4 produces a highly saturated image with the same posture and layout as the initial latent.

In particular, we validate the reason for the effectiveness of our initialization by investigating the predicted clean image (i.e., $\boldsymbol{x}_0$) through the inference steps. As shown in Figure 22(b), we observe that the overall image except for the fine details is determined within the first 10 steps of the inference. Therefore, our initialization is critical in generating successful multi-subject composition by providing information at the start. Using a larger scale γ yields more information that strongly affects the final image but also fixes the fine details such as postures and layouts.

Figure 23: **Diversity of images generated by MuDI**. MuDI can generate images of personalized subjects in diverse postures not restricted to the initial latent of our inference initialization. Each visualized initial latent arranged in a 3×3 grid corresponds to the generated image of the same position in the 3×3 grid. All the images are generated from the same random seed using the prompt *"... having a tea party with a table full of desserts in a colorful room."*



(a) Initial Latent **(b) Results with our initialization**



(a) Our Initialization (Random) **(b) LLM-guided Initialization**

Figure 24: **Examples of LLM-guided initialization for interactions.** Latents and images located at the same position in each 3× 3 grid are paired. All images are generated from the same random seed. **(a) Our initialization (Random).** Prompts describing interactions, for example, "monster toy sitting on a drink can," may not fit the randomly created initial layouts. Even though our initialization prevents identity mixing, the generated images may fail to reflect the interaction. **(b) LLM-guided initialization.** Instead of randomly positioning the segmented subjects, we utilize LLM to automatically generate prompt-aligned layouts for the inference initialization. We find that LLM-guided initialization enables the generation of complex interactions between subjects.

B.4 Diversity of images generated by MuDI

In Figure 23, we demonstrate that MuDI is able to generate images of personalized subjects in diverse postures not restricted by the initial latent of our inference initialization. In particular, the generated subjects integrate smoothly with the background and exhibit natural interactions between the subjects.

B.5 LLM-guided initialization for interactions

Generating complex interactions involving relations like "[V_1] toy sitting on [V_2] can" can be challenging for personalized subjects. However, our initialization method can significantly assist this by providing a well-aligned layout reflecting the prompt, such as placing the toy above the can.

https://doi.org/10.52202/079017-3200

Figure 25: We describe three scenarios where our inference initialization provides significant advantages. For each scenario, we visualize images generated with and without initialization on the left and the right columns respectively, where the image pairs (left and right) are generated using the same random seed. **(a) personalizing unusual subjects** that pre-trained models struggle to generate, such as cloud man. **(b) personalizing more than two subjects**, in particular for similar subjects. **(c) using complex prompts** like *"... as astronaut, floating on the moon, crater, space shuttle..."*.

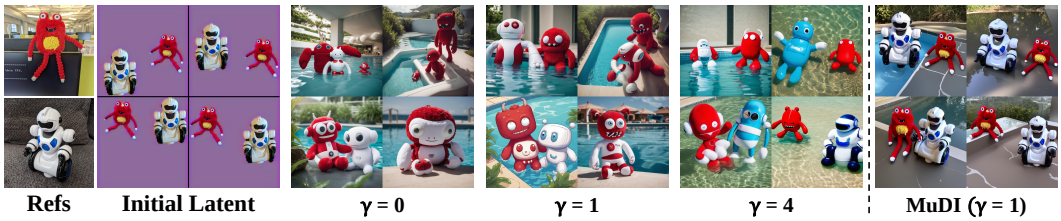
Figure 26: Even with the strong spatial conditioning of ControlNet [56], existing methods [42, 15] suffer from identity mixing. The generated images show a monster toy with the robot-like body or a robot toy with the color of the monster toy.

Inspired by Cho et al. [9], we utilize Large Language Models (LLMs) to generate prompt-aligned layouts of the segmented subjects. The generated layouts are used instead of randomly created layouts for the inference initialization. Such LLM-guided initialization enhances the ability to render complex interactions between subjects which we visualize in Figure 24.

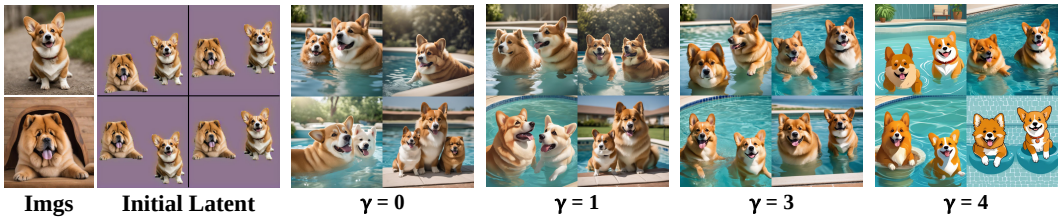
B.6 Importance of inference initialization

We explain in detail the scenarios where our inference initialization provides significant benefits. First, as shown in Figure 25(a), unusual subjects that pre-trained models struggle to generate (e.g., the cloud man) are frequently ignored during generation without initialization. However, our initialization alleviates the ignorance of unusual subjects by guiding the model to consider all subjects starting from the initial latent. Furthermore, initialization is crucial when personalizing more than two subjects as demonstrated in Figure 25(b). Generating images of more than two subjects without initialization often results in some subjects missing. It is almost impossible to compose many subjects together in an image without initialization. Lastly, we observe that initialization plays an important role when the given prompt is complex as shown in Figure 25(c). Personalized diffusion models fail to generate images of the subjects when the prompt describes uncommon or highly detailed scenes, resulting in subjects of mixed identities or some subjects missing. The inference initialization mitigates this problem by providing information on the subjects through the initial latent for which the model can focus more on rendering the scene described by the prompts. Our approach allows us to create images of personalized subjects, in particular new characters, in novel scenes.

Figure 27: **Inference initialization for pre-trained text-to-image models.** **(a)** For the unseen subjects, using initialization without personalization fails to preserve the details of the subjects, even initializing with a high gamma value (i.e., $\gamma = 4$). **(b)** For the known subjects, where the images are generated by the model, initialization still results in identity mixing.



(a) Unseen Subjects (Reference images)



(b) Known Subjects (Self-generated images)

B.7 Multiple subject composition with ControlNet

We validate that leveraging ControlNet [56] for previous approaches, for example, DreamBooth [42] and Cut-Mix [15], fails to address identity mixing. Figure 26 demonstrates that DreamBooth and Cut-Mix produce mixed identity toys even with the spatial conditioning of ControlNet. We note that other types of layout conditioning based on cross-attention maps, for instance, the region control [14], do not alleviate identity mixing when using SDXL as the pre-trained model, as explained in Section B.10.

B.8 Inference initialization for pre-trained text-to-image model

In Figure 27, we provide examples of inference initialization applied to the pre-trained text-to-image model. When the initial latents are created from subjects that were not seen by the pre-trained model (monster toy and robot toy in Figure 27(a)), the model fails to generate the details of the subjects. Only the layouts are preserved when using a high γ scale for the initialization. When the initial latent is created from known subjects (Corgi and Chow Chow in Figure 27(b)), the model results in identity mixing, even with a high $gamma$ scale. Therefore, we can observe that our inference initialization can only be effectively used to address identity mixing when the model is fine-tuned by our Seg-Mix.

B.9 More than two subjects

Qualitative comparison In Figure 28, we provide a qualitative comparison of previous approaches [42, 15] and our MuDI on personalizing three subjects. As DreamBooth [42] suffers from identity mixing even for two subjects, it fails to generate a composition of the three personalized subjects. Cut-Mix [15] also produces mixed-identity dogs and often generates images of some subjects missing. In contrast, MuDI can successfully generate high-quality images of the dogs and the cat that align with the given prompts.

Number of personalized subjects We analyze the performance of MuDI with respect to the number of subjects in Figure 29. We used two types of datasets composed of five subjects where the first category consists of five objects (monster toy, drink can, robot toy, Harry Potter toy, and teddy bear), while the second category consists of five animals (four types of dogs and one type of cat). After fine-tuning the pre-trained text-to-image model for each category, we generated 600 images composing N subjects for $N \in \{2, 3, 4, 5\}$. The images were generated using the prompts "[V_1], [V_2], ... [V_N] are in the jungle," for the objects and "[V_1], [V_2], ... [V_N] are playing together in the pool," for the animals. We consider all combinations and permutations of the subjects' order in the prompts uniformly. In particular, we observe that adding "N *objects:*" or "N *animals:*" at the start of the prompt achieves a higher success rate. We use a γ scale of 2 for cases with two or three subjects,

https://doi.org/10.52202/079017-3200

Figure 28: **Qualitative comparison** of personalizing three subjects. DreamBooth [42] suffers from severe identity mixing, especially for the two dogs. Cut-Mix [15] also fails to generate three subjects often ignoring some subjects and producing stitching artifacts. In contrast, MuDI successfully generates three personalized subjects without identity mixing that align with the given prompts.

(a) Success rate of MuDI according to number of subjects

(b) Qualitative samples

Figure 29: **Analysis of the number of personalized subjects. (a) Success rate of MuDI according to number of subjects.** MuDI shows a higher success rate compared to the baseline, DreamBooth. **(b) Qualitative samples.** MuDI generates images with 4 and 5 subjects without identity mixing.

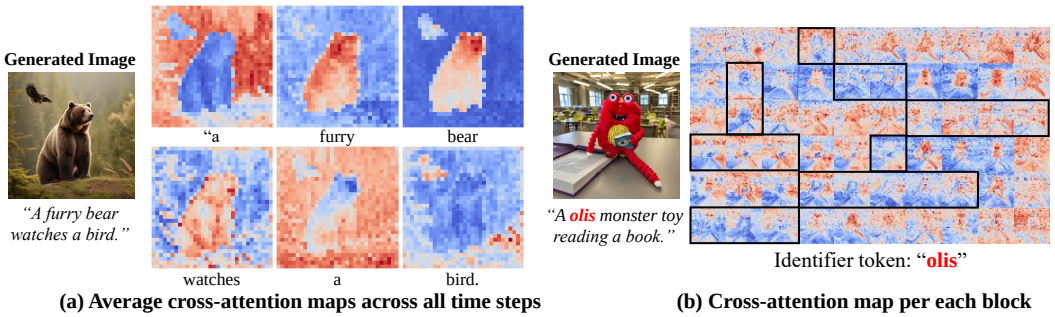
and a γ scale of 3 for cases with more than three subjects. The success rate was measured by first filtering the images using the D&C scores and then evaluating the success by humans.

We report the success rate of the generated images with respect to the number of subjects in Figure 29(a). DreamBooth completely fails to personalize more than two subjects for both categories. In contrast, MuDI achieves a significantly high success rate with the objects (Category I), showing over 50% success for generating four subjects together without identity mixing. While MuDI shows a relatively lower success rate with the animals (Category II) due to the high similarities of the subjects, MuDI can generate high-quality images of five animals with lively actions that align perfectly with the background. We visualize the successful images generated by MuDI in Figure 29(b). However, we observe that the performance of MuDI decreases as the number of personalized subjects increases, particularly for highly similar subjects.

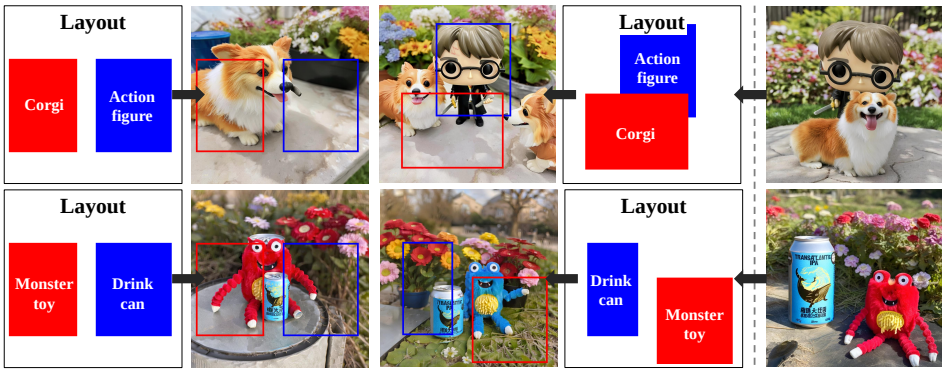
Empirical findings We end this section by providing empirical findings for personalizing multiple subjects. First, during training, we find it to be sufficient to augment images by composing only pairs of subjects, rather than composing three or more subjects together. Furthermore, when personalizing more than three subjects, a higher augmentation probability is required during training compared to the case with two subjects. Also, a higher γ value is needed during inference to generate all the

https://doi.org/10.52202/079017-3200

Figure 30: **Visualization of cross-attention maps in SDXL**. **(a)** The token for the bear demonstrates a high value in the region corresponding to the bird, and the token for the bird takes a high value in an irrelevant location (right bottom). Note that this figure can be compared to Figure 4 of Hertz et al. [16]. **(b)** The cross-attention maps of the identifier token (e.g., olis) do not show consistent results with the corresponding subject (i.e., monster toy in this example). We highlight the maps with black rectangles that have low values for the subject compared to the subject-irrelevant regions.



(a) Average cross-attention maps across all time steps **(b) Cross-attention map per each block**



(a) Regionally controllable sampling (Mix-of-show) **(b) MuDI**

Figure 31: **Comparison of MuDI and regionally controllable sampling [14] for SDXL. (a)** Regionally controllable sampling often results in missing subject or identity mixing. **(b)** MuDI prevents identity mixing as well as subject missing.

subjects together. Lastly, the prompt is crucial for generating multiple subjects. Prompts describing a detailed background or challenging actions may likely yield unsuccessful images. Notably, adding detailed descriptions like "*N objects:*" at the start of the prompt results in a higher success rate.

B.10 Analysis on cross-attention maps of SDXL

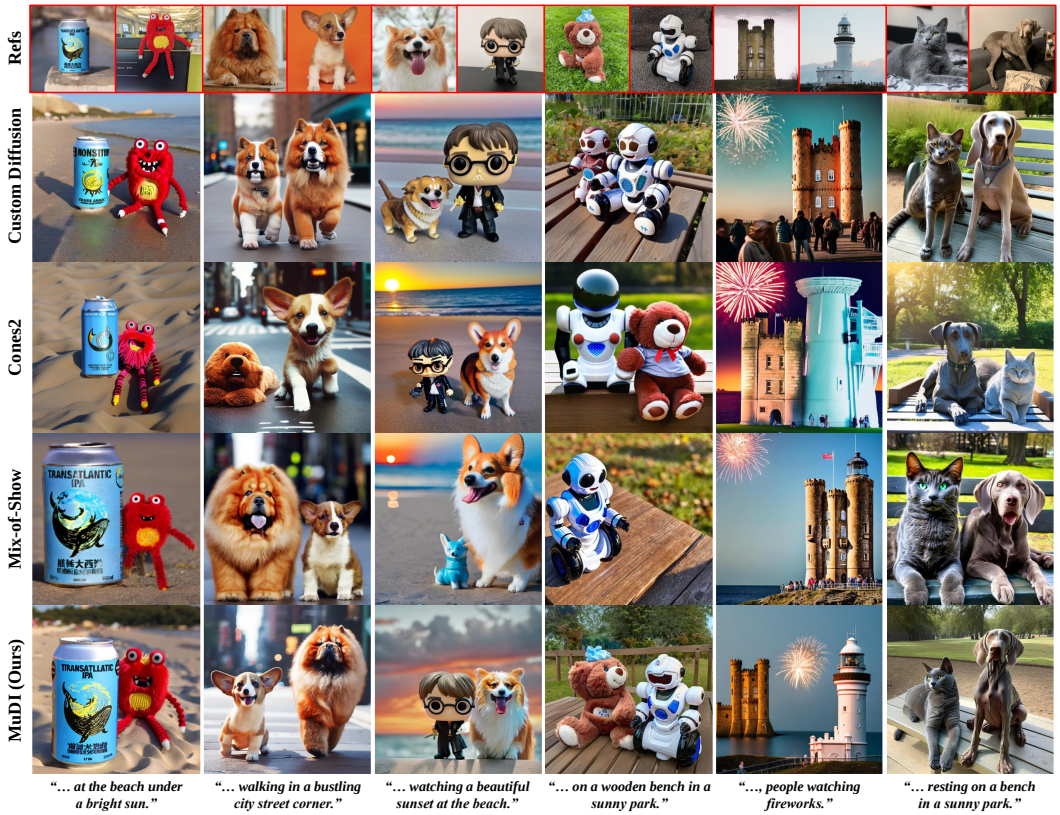
Cross-attention maps have been widely used in prior works on image editing [16], layout-guided generation [25, 14], single-subject personalization [52, 3], and zero-shot multi-subject personalization [53] due to their controllability on the relation between the spatial layouts and the words in the prompt [16]. While the cross-attention maps worked successfully on previous diffusion models like Stable Diffusion (SD) [40], it is not the case for recent diffusion models such as Stable Diffusion XL (SDXL) [36]. The architectural design of SDXL, where an additional text condition is added to the time embedding [4], significantly reduces the consistency of the cross-attention maps which we demonstrate in Figure 30. Therefore, previous approaches based on the cross-attention maps [25, 14] are not directly applicable when using SDXL as a pre-trained text-to-image diffusion model. For example, Han et al. [15] propose Unmix regularization, a technique that utilizes cross-attention maps to reduce stitching artifacts of the generated images, which we observe it to be ineffective for SDXL.

B.11 Regionally controllable sampling for SDXL

To address identity mixing, Gu et al. [14] propose regionally controllable sampling, i.e., region control, which leverages multiple regional prompts and the corresponding cross-attention maps during inference. However, manipulating the cross-attention map is not effective in preventing

https://doi.org/10.52202/079017-3200

Figure 32: **Qualitative comparison** of images generated by Custom Diffusion [21], Cones2 [25], Mix-of-Show [14], and MuDI that use Stable Diffusion v2 [40] as a pre-trained text-to-image model. $*$ denotes that it used ControlNet [56] to generate images.



Method	Multi-Subject Fidelity		Text Fidelity		Speed
	D&C-DS↑	D&C-DINO↑	ImageReward†↑	CLIPs†↑	Time
Custom Diffusion [21]	0.469	0.497	0.588	0.234	1x
Cones2 [25]	0.408	0.429	0.607	**0.254**	9.2x
Mix-of-Show [14]	0.367	0.364	0.470	0.240	1.3x
Mix-of-Show* [14]	0.688	**0.666**	0.061	0.223	1.5x
MuDI (Ours)	**0.692**	0.661	**0.683**	0.250	1.1x

Table 3: **Quantitative comparison using Stable Diffusion v2 [40]** as a pre-trained text-to-image model. $*$ indicates using ControlNet [56]. † denotes the text fidelity score considering the permutation of the subjects in the prompt to avoid position bias.

identity mixing for recent diffusion models like SDXL [36]. We empirically observe that region control is highly likely to produce images with some subjects missing or having mixed identities, as demonstrated in Figure 31(a).

On the other hand, our MuDI can prevent both subject missing or identity mixing without using cross-attention maps, even in cases involving highly similar subjects or overlapping layouts. Notably, our initialization does not require additional computational overhead, in contrast to regional control which requires 1.6 times the inference time due to the high number of cross-attention blocks.

B.12 Comparison with existing works using layout conditioning

We compare our MuDI with existing works on multi-subject composition using layout conditioning [14, 25]. We use Stable Diffusion v2 [40] as the pre-trained text-to-image diffusion model. We create the layouts required for the baselines, Cones2 [25] and Mix-of-Show [14], from the random initial latent of our inference initialization. For a fair comparison, we did not use any image-based

Figure 33: **Qualitative results using Stable Diffusion v1.5 [40]** as a pre-trained text-to-image model. Similar to the case of using SDXL [36] as a pre-trained model, DreamBooth [42] produces images of mixed-identity toys. Our Seg-Mix effectively addresses identity mixing but often generates images with the robot toy missing. In contrast, our MuDI, which leverages both Seg-Mix and our inference initialization, successfully personalizes the subjects distinctly without identity mixing. Note that the images of the same positions in the 3×3 grid are generated using the same random seed.

conditioning such as ControlNet [56] and T2I-adapter [29] for Mix-of-Show. We additionally report the results of Mix-of-Show using Canny edge ControlNet, namely Mix-of-Show∗. Note that Mix-of-Show demonstrates significant performance degradation when the size of the bounding boxes consisting of the layouts is not sufficiently large. Therefore, we manually set the bounding boxes in the layout to be sufficiently large, for example, as the leftmost layout in Figure 31.

As shown in Table 3, MuDI achieves the highest D&C-DS scores as well as the ImageReward. Cones2 demonstrates low subject fidelity on unseen subjects, for example, the monster toy, as it is trained solely on text embeddings. Mix-of-Show frequently generates images with some subjects missing, and results in low D&C scores. When used with ControlNet, Mix-of-Show generally shows higher subject fidelity but often produces blurry backgrounds or images that are not aligned with the given prompts. We also report the relative inference time compared to the normal inference time using a pre-trained model for generating images with two subjects, as shown in Table 3. We use the official codes available for each method and measure the inference time using a single RTX 3090 GPU. MuDI achieves significantly faster inference speed compared to the methods based on cross-attention maps, namely Cones2 and Mix-of-Show.

Additionally, we provide a qualitative comparison in Figure 32. Custom diffusion [21] results in identity mixing for similar subjects, while Cones2 produces significantly low subject fidelity for unseen subjects such as the monster toy. Mix-of-Show generates images with some subjects missing In contrast, our MuDI successfully generates multi-subject images without identity mixing.

B.13 Stable Diffusion v1.5 as a pre-trained model

In Figure 33, we provide qualitative results of DreamBooth [42] and MuDI using Stable Diffusion v1.5 [40] as the pre-trained text-to-image diffusion model. Similar to the case when using SDXL [36] as the pre-trained model, DreamBooth results in identity mixing. Our Seg-Mix effectively addresses identity mixing but often generates images of a subject missing. In contrast, our MuDI which leverages both Seg-Mix and inference initialization successfully personalizes the subjects without identity mixing or subject ignorance.

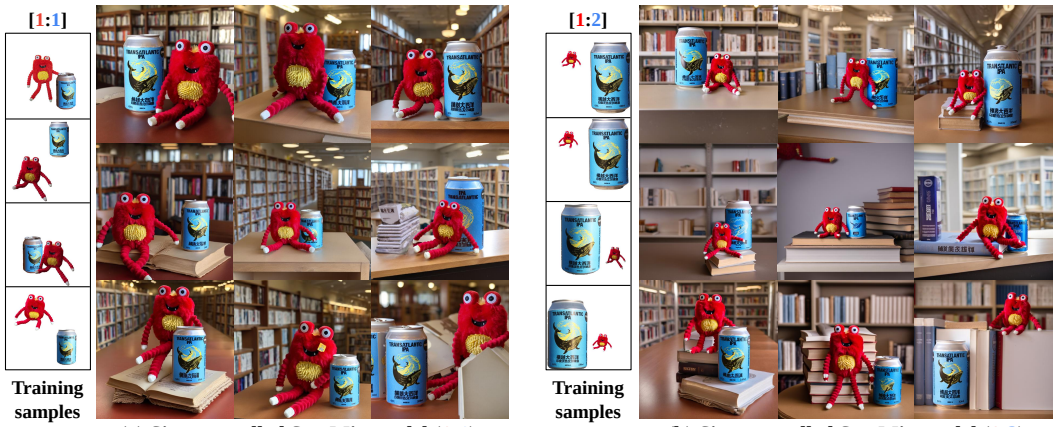
C Other use cases

C.1 Relative size control

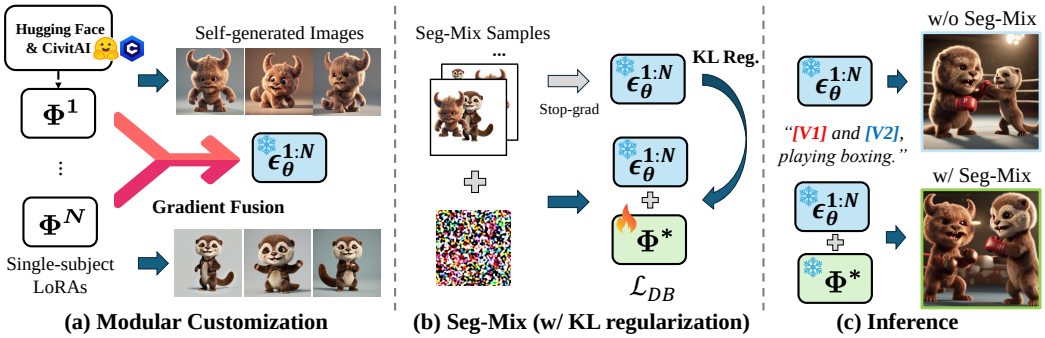
MuDI enables control of relative size between the personalized subjects. During training, we can augment the images of segmented subjects with fixed relative sizes according to user intents, instead of random relative sizes. This corresponds to setting the argument *scales* of the function *create_seg_mix* in Algorithm 1. As shown in Figure 34, we can personalize models to generate the toy to be larger

https://doi.org/10.52202/079017-3200

Figure 34: **Examples of relative size control using Seg-Mix**. During training, we augment the images of segmented subjects with fixed relative sizes. **(a)** When the relative size of the toy and the can is equal (i.e., toy:can=1:1), the generated samples display a toy and a can of similar size. **(b)** When we set the relative size of the toy to be smaller than the can (i.e., toy:can=1:2), the generated samples display a relatively small toy compared to the can. Note that the images of the same positions in the 3 grid are generated using the same random seed.



(a) Size controlled Seg-Mix model (1:1) **(b) Size controlled Seg-Mix model (1:2)**



(a) Modular Customization **(b) Seg-Mix (w/ KL regularization)** **(c) Inference**

Figure 35: **Modular customization with Seg-Mix**. **(a)** We first generate single-subject images using the pre-trained LoRAs (Φ^i), and then merge the LoRAs using the gradient fusion [14] to obtain a fused model $\epsilon_\theta^{1:N}$. **(b)** We use the self-generated images to train additional LoRA for identity decoupling via Seg-Mix. We add KL regularization to the training objective to prevent overfitting and saturation. We only train for 200-300 iterations. **(c)** While the fused model results in mixed-identity characters, our Seg-Mix fine-tuning effectively addresses identity mixing.

than the can or vice versa. We observe a consistent relative size of the personalized subjects in the generated images.

C.2 Modular customization

Our Seg-Mix can also be applied to modular customization, i.e., when we possess single-subject LoRAs that have been independently fine-tuned to each subject. Instead of re-training the models each time for new combinations of subjects, we can efficiently merge the pre-trained models that are independently fine-tuned for each subject, avoiding the need for training from the beginning.

We first generate images of each subject using their corresponding LoRA, which are subsequently utilized as reference images. We then merge the single-subject fine-tuned LoRAs [18] using an existing method such as gradient fusion [14] to obtain a fused model (Figure 35(a)). While the fused model can successfully generate each subject individually, composing multiple subjects results in severe identity mixing. Therefore, we apply Seg-Mix with the generated single-subject images for 200-300 iterations (Figure 35(b)). When applying Seg-Mix, we apply a new LoRA to fine-tune the fused model and set the seg-mix probability to 1. This is because the fused model has already

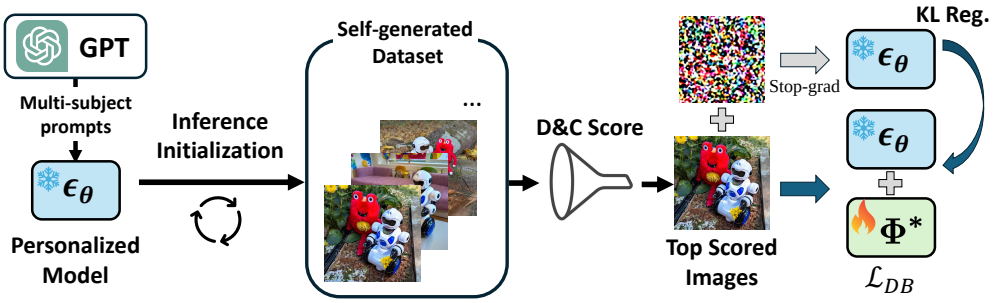


Figure 36: **Illustration of Iterative Training for MuDI.**

been trained with each subject and only needs to be trained with the composition of the subjects. In particular, we add KL regularization [11] to the personalization objective (Eq. (2)) in order to prevent overfitting and saturation. Our approach effectively reduces identity mixing as shown in Figure 35(c).

Notably, we find that for certain subjects, using self-generated images instead of the original reference images for Seg-Mix fine-tuning achieved superior performance, especially alleviating posture overfitting. For example, the reference images for the characters from Sora [31] (e.g., the otter) are obtained from the video frames that have highly limited postures. Instead of using the reference images directly, we can generate diverse images of the subjects using the single-subject LoRAs, and use them when applying Seg-Mix.

C.3 Iterative training

We present an iterative training (IT) method [45] for MuDI to further improve the image quality. The key idea is to additionally fine-tune the personalized model using high-quality images generated with MuDI. We introduce a fully automatic training based on our MuDI and the D&C score which closely aligns with human evaluation for the multi-subject fidelity.

To be specific, we generate 200 multi-subject images with MuDI using simple prompts created by ChatGPT [30], for example, "[$V1$] dog and [$V2$] dog watching birds from a window sill.". The top 50 images based on the D&C-DS score are used to fine-tune the personalized model using LoRA [18]. We fine-tune the model with the KL regularization [11] added to the personalization objective $\mathcal{L}_{\mathcal{DB}}$ of Eq. (2). In particular, we observe that the KL regularization is crucial for preventing saturation and preserving the image quality of the self-generated images. Empirically, setting the KL regularization weight as 1.0 results in a good trade-off between preventing saturation and multi-subject fidelity. We provide an overview of the iterative training framework in Figure 36.

As shown in Figure 37, IT considerably improves MuDI on personalizing highly similar subjects. We believe using other RL-based fine-tuning methods such as Direct Preference Optimization (DPO) [38, 51] instead of our supervised fine-tuning approach, would be more robust against oversaturation and better to reflect human preferences. Additionally, combining different reward models with RL methods could further improve how well the system aligns with human preferences, which we leave for future work.

https://doi.org/10.52202/079017-3200

(a) Seg-Mix (base) (b) Seg-Mix + IT

(c) Seg-Mix + Initialization (MuDI) (d) MuDI + IT

Figure 37: **Qualitative comparison** of our iterative training (IT). Images at the same position in each 3×3 grid are generated from the same random seed. **(a) Seg-Mix training** without initialization does not perfectly address identity mixing. **(b) Iterative training** without initialization shows improvement compared to the Seg-Mix training. **(c) MuDI** addresses identity mixing and subject missing, but occasionally fails to decouple highly similar subjects. **(d) MuDI with iterative training** successfully personalizes multiple subjects that are highly similar.

Figure 38: **Personalizing 11 concepts together with MuDI using a single LoRA [18].** We use descriptive classes for each dog and cat, for example, Weimaraner or Mudi, which enhances the ability to personalize multiple subjects that are highly similar.

https://doi.org/10.52202/079017-3200

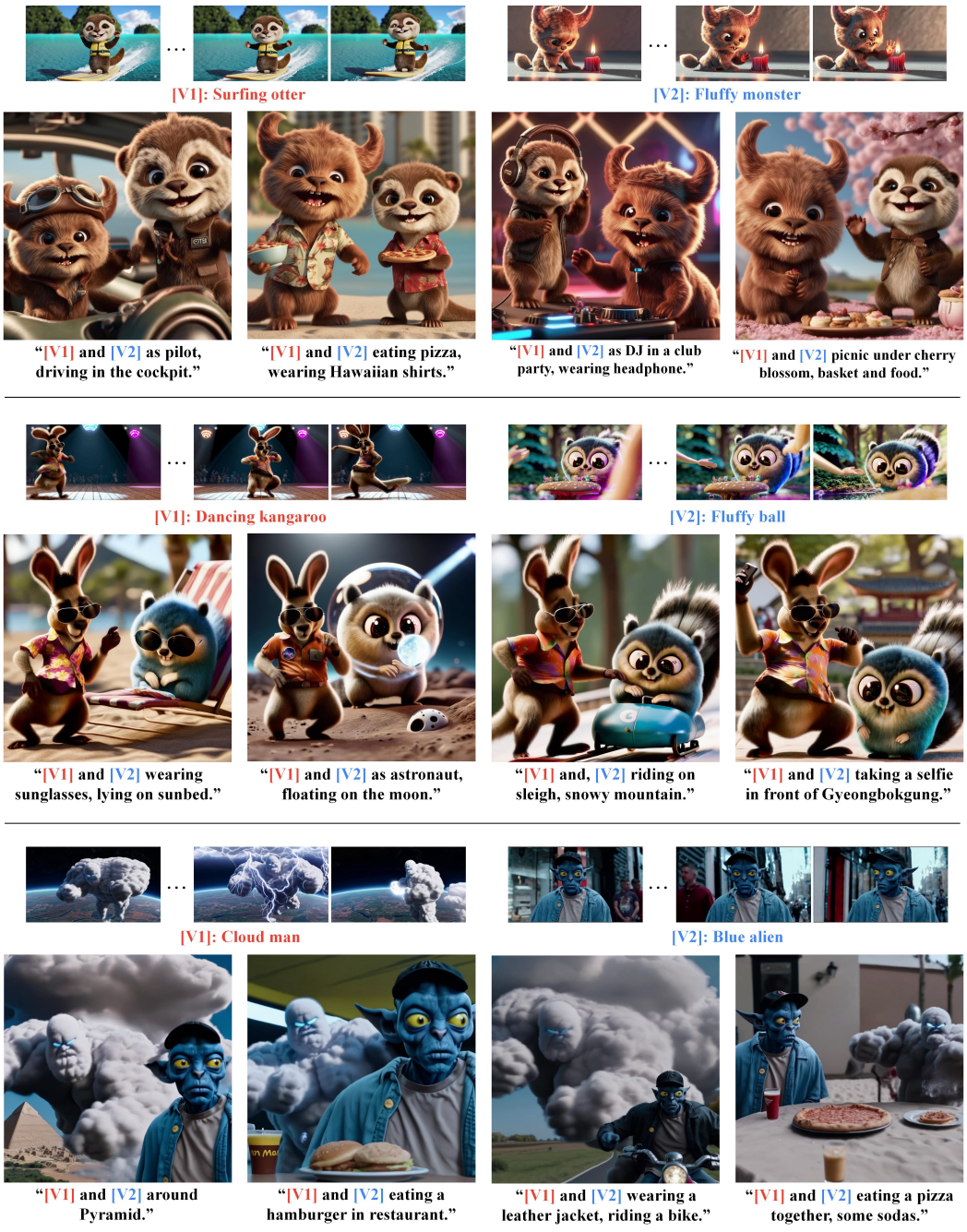
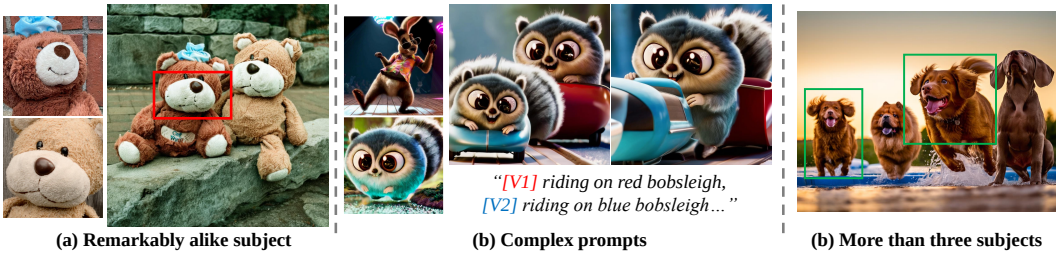


Figure 39: **Generated samples from MuDI by personalizing characters from Sora [31].**



(a) Remarkably alike subject **(b) Complex prompts** **(b) More than three subjects**

Figure 40: **Limitations. (a) Remarkably alike subjects** are challenging to decouple perfectly as they lie very close in the latent space. For example, two brown teddy bears can be easily mixed up as they have highly similar designs and colors. **(b) Complex prompts** that describe unusual or detailed scenes bring additional difficulty in preserving the details of the subjects. In this case, the subjects can be easily ignored during the generation. **(c) More than three subjects.** MuDI significantly mitigates identity mixing but often duplicates the same subjects in the generated images.

D Limitations and societal impacts

Limitations We find that decoupling the identities of remarkably alike subjects is still challenging even for our method, for example, two brown teddy bears in our dataset (see Figure 40(a)). Such subjects are very close in the image latent space which may be difficult to separate with the current text-to-image pre-trained models. Furthermore, we observe that our method faces difficulties when the given prompt is complex. For example, we show in Figure 40(b) that the generated images of personalized characters with the prompt "[V1] riding on red bobsleigh, [V2] riding on blue bobsleigh." do not display the kangaroo character. This issue could be alleviated by optimizing to the specific prompt [2]. Lastly, although our framework effectively alleviates identity mixing for several subjects, we find that subject dominance becomes stronger as the number of personalized subjects increases. For instance, MuDI may duplicate the same subjects in the generated images, as in Figure 40(c) Adjusting the γ scale in our initialization can address subject dominance but may yield image saturation. We believe our iterative training framework may potentially address these limitations and can be further developed by applying recent RLHF approaches [23, 11, 6, 51, 10].

Societal impacts Our method allows for the synthesis of realistic images of multiple personalized subjects. However, there is a risk that our framework can be misused to generate harmful content for the public or to include subjects that are sensitive to privacy. To prevent this, it is necessary to apply measures such as watermarking to the generated images to prevent misuse, as well as protective watermarking specifically for privacy-sensitive subjects.

https://doi.org/10.52202/079017-3200

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and the introduction reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations of the work in the Appendix.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: The paper does not include theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We describe the details of the experiments in the main paper as well as the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-3200

Answer: [Yes]

Justification: We have submitted our code for the experiments as supplementary materials.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We discuss the training and test details in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We visualize the error bars of our human evaluation results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We describe the computer resources used in our experiments in the Appendix.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics https://neurips.cc/public/EthicsGuidelines?

   Answer: [Yes]

   Justification: We conform NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss both potential positive and negative societal impacts in the Appendix.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-3200

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper do not pose such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We properly credit the original owners of assets used in the paper.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We discussed details of the dataset/code/model in the Appendix.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: We include information of the human evaluation in the Appendix.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [No]

    Justification: Our work does not require IRB approval as it involves human raters evaluating image preferences from a benchmark dataset and the ratings are collected anonymously. The study does not pose any potential risks to the participants.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-3200