# Decoupled Textual Embeddings for Customized Image Generation

**Yufei Cai**[1,2,3], **Yuxiang Wei**[3], **Zhilong Ji**[4], **Jinfeng Bai**[4], **Hu Han**[1,2], **Wangmeng Zuo**[3,5]

[1]Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Harbin Institute of Technology
[4]Tomorrow Advancing Life
[5]Pengcheng Lab
caiyufei23@mails.ucas.ac.cn, yuxiang.wei.cs@gmail.com, zhilongji@hotmail.com, jfbai.bit@gmail.com
hanhu@ict.ac.cn, wmzuo@hit.edu.cn

## Abstract

Customized text-to-image generation, which aims to learn user-specified concepts with a few images, has drawn significant attention recently. However, existing methods usually suffer from overfitting issues and entangle the subject-unrelated information (*e.g*., background and pose) with the learned concept, limiting the potential to compose concept into new scenes. To address these issues, we propose the DETEX, a novel approach that learns the disentangled concept embedding for flexible customized text-to-image generation. Unlike conventional methods that learn a single concept embedding from the given images, our DETEX represents each image using multiple word embeddings during training, *i.e*., a learnable image-shared subject embedding and several image-specific subject-unrelated embeddings. To decouple irrelevant attributes (*i.e*., background and pose) from the subject embedding, we further present several attribute mappers that encode each image as several image-specific subject-unrelated embeddings. To encourage these unrelated embeddings to capture the irrelevant information, we incorporate them with corresponding attribute words and propose a joint training strategy to facilitate the disentanglement. During inference, we only use the subject embedding for image generation, while selectively using image-specific embeddings to retain image-specified attributes. Extensive experiments demonstrate that the subject embedding obtained by our method can faithfully represent the target concept, while showing superior editability compared to the state-of-the-art methods. Our code will be available at https://github.com/PrototypeNx/DETEX.

## Introduction

Recently, diffusion models (Ramesh et al. 2022; Saharia et al. 2022; Rombach et al. 2022; Nichol et al. 2022) have demonstrated remarkable superiority in text-to-image generation. Benefiting from the large-scale pretraining, these models can generate diverse and photo-realistic images based on textual descriptions, showing great potential in various tasks, such as artistic and video creation (Saharia et al. 2022; Ramesh et al. 2022; Zhang et al. 2023).

Besides text-to-image generation, substantial efforts have been devoted to customized image generation (Gal et al. 2022), which aims to learn user-specified concept from a small set of images describing target concept (typically 3-5 images). Existing methods for customized text-to-image generation (Kumari et al. 2023; Ruiz et al. 2023; Gal et al. 2022) usually aligned the target concept with a user-specified word by finetuning the word embedding or model parameters. However, due to the limited training examples, the learned concept inevitably contains subject-unrelated information (*e.g*., image background, subject pose and position), resulting in degenerated editability. Although some studies (Avrahami et al. 2023; Wei et al. 2023) employed a subject mask to filter the background information, the gain was unsatisfactory, as the irrelevant disturbances (*e.g*., blank background and pose) still entangled with the learned concept. Disenbooth (Chen et al. 2023) decomposed the target concept into a subject embedding and an additional irrelevant embedding to exclude the irrelevant information. However, there lacks explicit supervision to effectively facilitate the decoupling between the subject concept and the irrelevant information.

To address the above issues, we propose DETEX, a method that learns disentangled concept embedding for flexible customized text-to-image generation. Following the principles of Custom Diffusion (Kumari et al. 2023), we adapt the pretrained Stable Diffusion (Rombach et al. 2022) model to new concept by finetuning both model parameters and the word embedding. Instead of learning a single concept embedding to represent all given images, our DETEX represents the target concept and subject-unrelated information using separate word embeddings, *i.e*., an image-shared subject embedding and several image-specific subject-unrelated embeddings. Specifically, we introduce a learnable image-shared subject embedding to represent the target concept. To decouple irrelevant attributes from the subject embedding, for each input image, we additionally introduce several image-specific subject-unrelated embeddings to capture the irrelevant information. Here, we consider irrelevant information from two main aspects: pose (view) and background. The corresponding attribute words (*i.e*., `[B] background` and `[P] pose/view`) are incorporated with the embeddings to facilitate them capturing the irrelevant information.

To effectively decouple the target concept with unrelated pose and background information, we propose a joint training strategy. Specifically, we encourage different embed-

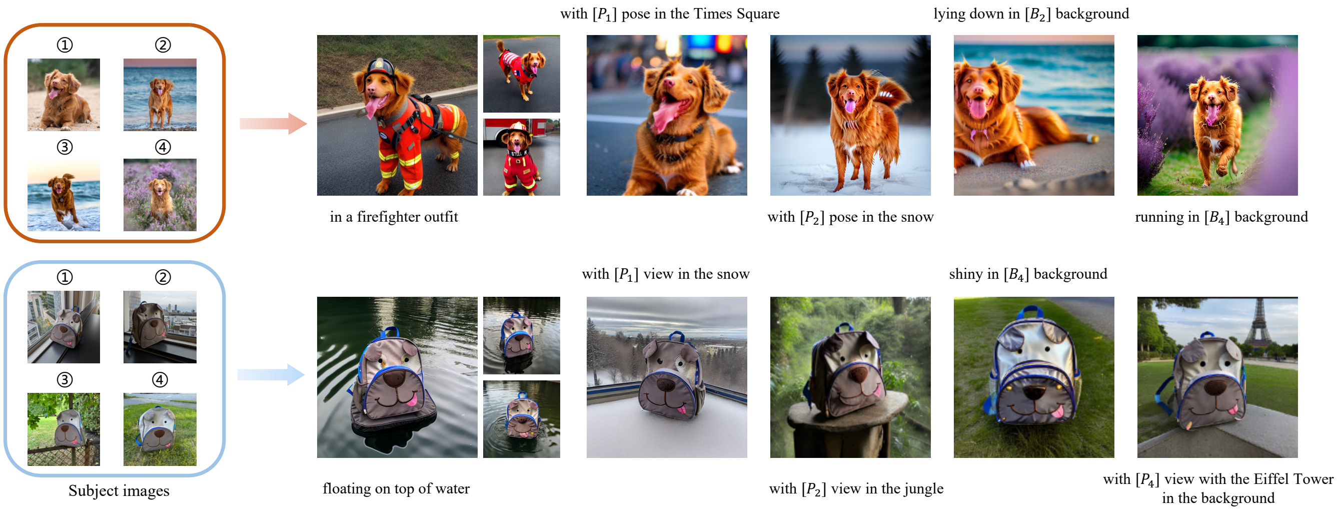

with [$P_1$] pose in the Times Square
lying down in [$B_2$] background
①
②
③
④
in a firefighter outfit
with [$P_2$] pose in the snow
running in [$B_4$] background
with [$P_1$] view in the snow
shiny in [$B_4$] background
Subject images
floating on top of water
with [$P_2$] view in the jungle
with [$P_4$] view with the Eiffel Tower in the background

Figure 1: Customized images generated by our DETEX. The identifier [$B_i$] ([$P_i$]) indicates the learned background (pose) of the $i$-th input image. Our DETEX can generate high-quality and diverse images when composing the learned concept into a new scene, while allowing users to selectively retain specific attribute information for controllable generation.

ding to reconstruct the corresponding information representation based on the given images and their subject masks. However, directly optimizing the image-specific subject-unrelated embeddings is insufficient to capture specific attributes efficiently. To address this limitation, we propose an attribute mapper for each attribute which projects each image as the corresponding subject-unrelated embedding. The CLIP (Radford et al. 2021) image encoder serves as the feature extractor in this process. During inference, only the subject embedding is utilized for image generation. Our experimental results show that our learned subject embedding can faithfully represent the target concept, and possess excellent editability compared to the state-of-the-art (SOTA) methods. Furthermore, we can selectively retain specific irrelevant attribute information by keeping corresponding unrelated embeddings, allowing users to flexibly control the image generation (see Fig. 1).

Our contributions can be summarized as follows:

- We propose a customized image generation method that utilizes multiple tokens to alleviate the issue of overfitting and entanglement between the target concept and unrelated information.
- Our method enables more precise and efficient control over preserving input image content in the generated results during inference by selectively utilizing different tokens. This enhances the controllability of personalized generation.
- Extensive experiments show that our method outperforms the SOTA methods in terms of editing flexibility. Furthermore, our method exhibits stronger editing potential, especially when the number of input images is extremely limited.

In conclusion, this work proposes a novel solution to customized generation in text-to-image models, addressing the challenges of overfitting and controllability.

## Related Work

**Text-to-Image Diffusion Models.** Diffusion models (Ho, Jain, and Abbeel 2020; Dhariwal and Nichol 2021) are parametric neural networks that learn the data distribution by progressive denoising. It has achieved remarkable performance in the field of image synthesis. Subsequent works have focused on exploring diffusion-based conditional generation. Classifier-free guidance (Ho and Salimans 2022) based on the derivation of conditional probabilities is a direct conditional generation approach that jointly optimizes over a large amount of data, resulting in reliable and detail-preserving generation effects. Benefiting from the advancements in language models and multi-modal models such as CLIP (Radford et al. 2021) and BERT (Devlin et al. 2018), further work has been devoted to diffusion-based text-to-image generation. Some work (Avrahami, Lischinski, and Fried 2022; Kim, Kwon, and Ye 2022) use text as a condition to control image synthesis. By leveraging the powerful information extraction capability of language models and the classifier-free guidance techniques, current approaches (Ramesh et al. 2022; Saharia et al. 2022; Nichol et al. 2022) promote semantic alignment between cross-modal information. Through large-scale training on extensive data samples, they have achieved impressive text-to-image synthesis results. Latent diffusion models (Rombach et al. 2022) compress data into a low-dimensional latent space, finding an optimal balance between computational complexity reduction and detail preservation. Text-to-image models based on LDM, such as Stable Diffusion, exhibit more efficient text-driven synthesis capabilities, reaching high levels of diversity and generality.

Despite the impressive performance of existing large-scale text-to-image models, they solely rely on natural language prompts guidance and cannot perform customized image generation. Specifically, it is extremely challenging to

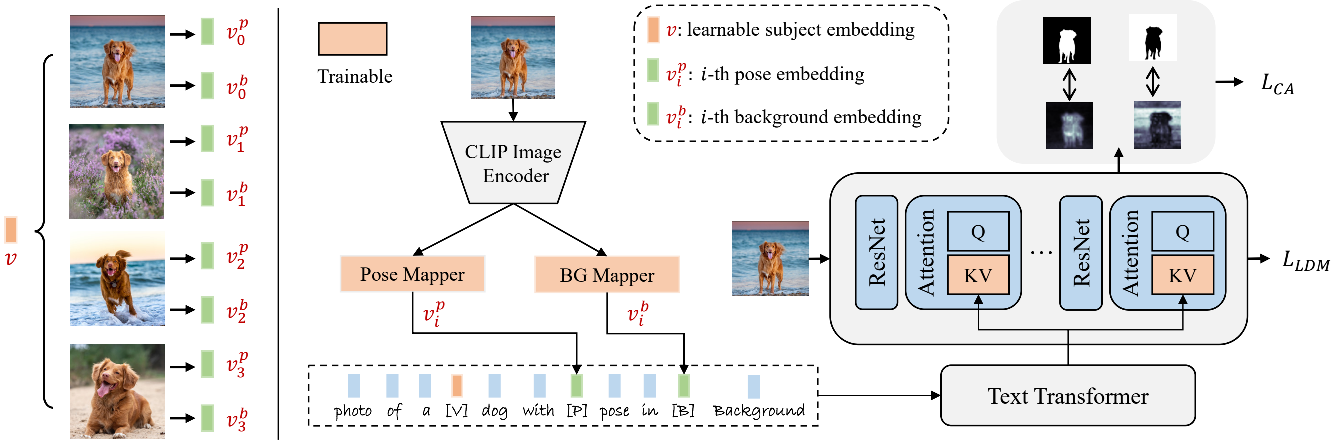


Figure 2: Framework of our DETEX. Left: Our DETEX represents each image with multiple decoupled textual embeddings, *i.e.*, an image-shared subject embedding $v$ and two image-specific subject-unrelated embeddings (pose $v_i^p$ and background $v_i^b$). Right: To learn target concept, we initialize the subject embedding $v$ as a learnable vector, and adopt two attribute mappers to project the input image as the pose and background embeddings. During training, we jointly finetune the embeddings with the K, V mapping parameters in cross-attention layer. A cross-attention loss is further introduced to facilitate the disentanglement.

consistently maintain the identity of a specific concept in the images generated by such models. This has led to the emergence of customized image generation tasks.

**Customized Image Generation.** Current customized generation methods primarily rely on fine-tuning. Textual Inversion (Gal et al. 2022) only fine-tunes word embeddings to learn the target concepts. However, it lacks the ability to fit new concepts effectively and struggles to capture fine-grained details of concepts in various conditions. DreamBooth (Ruiz et al. 2023) fine-tunes the full weights of the U-net and text encoder to achieve more effective concept-fitting ability, but it introduces language drifts and information forgetting, leading to poorer editing flexibility and controllability. Subsequent work attempts to find a more compact and efficient parameter space for fine-tuning to alleviate model overfitting. Custom Diffusion (Kumari et al. 2023) only fine-tunes the weights of the cross-attention layers, while SVDiff (Han et al. 2023) fine-tunes the spectral space of weight matrices. Later, non-fine-tuning methods for customized generation emerged. Cones (Liu et al. 2023) focuses on identifying the effective concept neurons related to the target concept, while ViCo (Hao et al. 2023) proposes a plug-in image attention module to adjust the diffusion process. Other works (Wei et al. 2023; Shi et al. 2023; Li, Hou, and Loy 2023) explore achieving customized generation without finetuning. These encoder-based approaches propose an additional module pretrained on additional datasets to reduce the time cost of generation significantly. DisenBooth (Chen et al. 2023) explores the decoupling of identity-irrelevant information from the target concept during customized fintune. It decomposes text embeddings into identity-related and identity-irrelevant parts to generate new images. In this paper, we model the pose and background information as independent image-specific words, achieving finer decoupling at the word embedding stage. Our approach enables more flexible editing ability and precise control over retaining inherent content from input images during inference.

# Proposed Method

## Preliminary

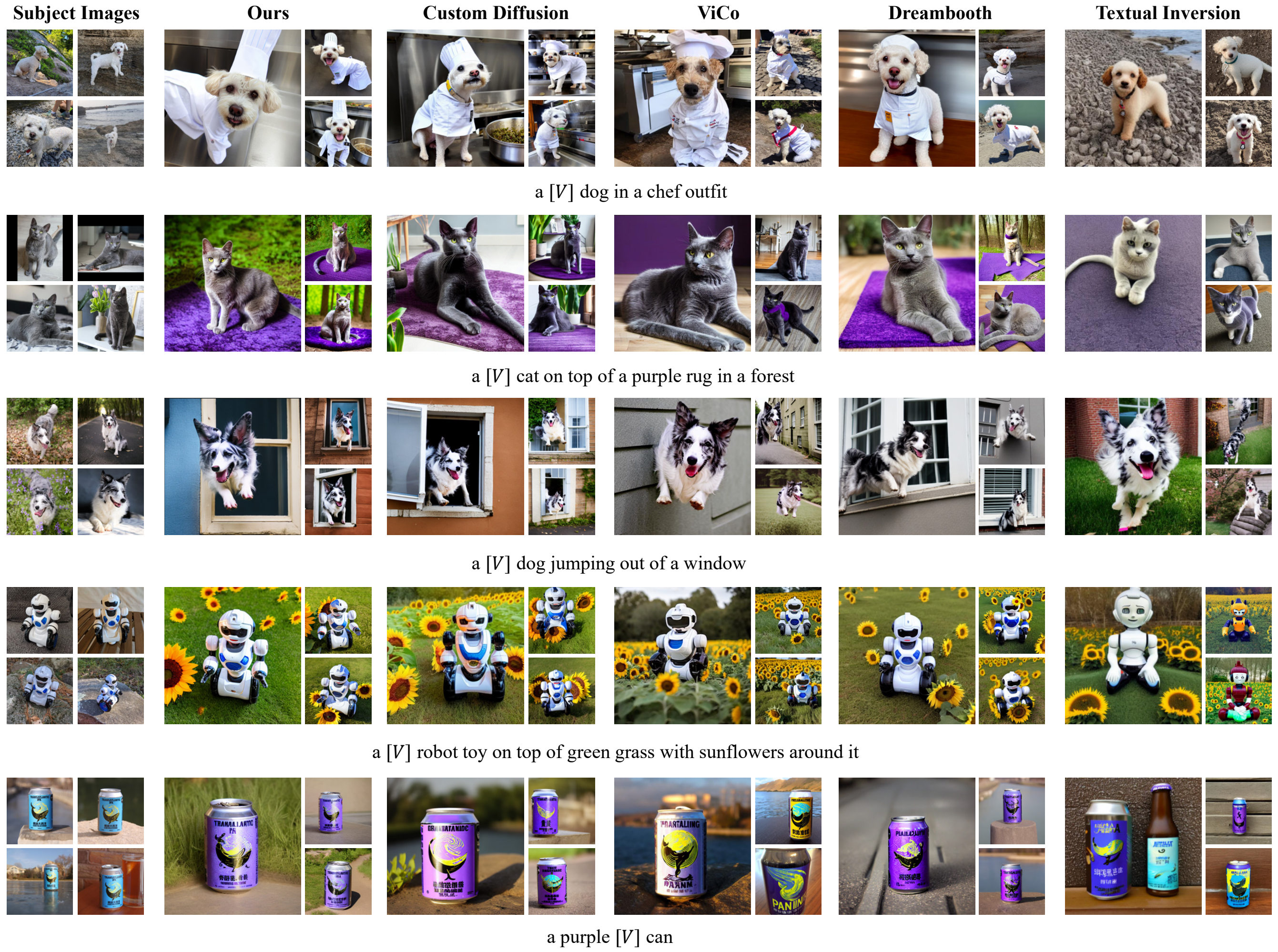
In this work, we employ the pretrained Stable Diffusion (SD) (Rombach et al. 2022) as our text-to-image model, and adapt it for customized text-to-image generation. In the following, we will give a brief introduction of the SD and our baseline method for customized text-to-image generation, *i.e.*, Custom Diffusion (Kumari et al. 2023).

**Stable Diffusion.** Stable Diffusion (Rombach et al. 2022) is trained on large-scale data and comprises two components. First, an autoencoder ($\mathcal{E}(\cdot)$, $\mathcal{D}(\cdot)$) is trained to map an image $x$ to a lower dimensional latent space by the encoder $z = \mathcal{E}(x)$, and then reconstructed back to the image by the decoder $D(\mathcal{E}(x)) \approx x$. Then, the conditional diffusion model $\epsilon_\theta(\cdot)$ is trained on the latent space to generate latent codes based on text condition $y$. To train the diffusion model, a simple mean-squared loss is adopted,

$$L_{LDM} = \mathbb{E}_{z\sim\mathcal{E}(x),y,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|_2^2\right], \quad (1)$$

where $\epsilon$ denotes unscaled noise, $t$ is time step, $z_t$ is latent noised to time $t$, $\tau_\theta(\cdot)$ represents the pretrained CLIP text encoder (Radford et al. 2021). During inference, a random Gaussian noise $z_T$ is iteratively denoised to $z_0$, ultimately yielding the final image through the decoder $x' = \mathcal{D}(z_0)$.

**Customized Text-to-Image Generation.** Given a small set of images $\{x_i\}_{i=1}^N$ describing the user-specified concept, Custom Diffusion (Kumari et al. 2023) first introduces a learnable word embedding $v$ to represent the target concept. A pseudo-word `[V]` is further introduced to the vocabulary and $v$ is associated as its word embedding. With `[V]`, we



a [$V$] dog in a chef outfit

a [$V$] cat on top of a purple rug in a forest

a [$V$] dog jumping out of a window

a [$V$] robot toy on top of green grass with sunflowers around it

a purple [$V$] can

Figure 3: Visual comparisons. Our DETEX demonstrates superior concept editability and subject fidelity compared to Textual Inversion (Gal et al. 2022), Dreambooth (Ruiz et al. 2023), Custom Diffusion (Kumari et al. 2023), and ViCo (Hao et al. 2023).

can generate new images with desired concept, such as "`A [V] swimming`". During training, Custom Diffusion simultaneously optimizes $v$ and the parameters of K, V mapping in cross-attention layers by minimizing Eqn. 1 over the given images. To incorporate $v$ into the generation, the input text prompt $y$ is formulated as "`Photo of a [V] [class]`", where `[class]` is the category prior to the given subject. Furthermore, a prior preserving loss $L_{pr}$ is adopted by Custom Diffusion to preserve the prior characteristics of pretrained SD model,

$$L_{pr} = \mathbb{E}_{z \sim \mathcal{E}(x_{pr}), y_{pr}, \epsilon \sim \mathcal{N}(0,1), t} \left[ \| \epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y_{pr})) \|_2^2 \right], \quad (2)$$

The regularization images $\{x_{pr}\}$ are retrieved from LAION-400M (Schuhmann et al. 2021) dataset or generated by original SD model, while the regularization prompt $y_{pr}$ is set as "`Photo of a [class]`".

## Decoupled Textual Embeddings

Custom Diffusion learns a novel subject embedding to represent all images. However, due to the limited training examples, the learned subject embedding inevitably contains subject-unrelated information (*e.g.*, image background, subject pose, and position), thereby limiting its ability in composing novel scenes. To address this issue, we propose a novel approach DETEX, which learns the decoupled textual embedding for flexible customized text-to-image generation. Following the principles of Custom Diffusion (Kumari et al. 2023), we adapt the pretrained Stable Diffusion (Rombach et al. 2022) model to learn new concept by finetuning both model parameters and the image-shared subject embedding $v$. To decouple irrelevant information from the subject embedding, we introduce multiple image-specific subject-unrelated embeddings. Here, we consider two primary sources of irrelevant information: pose (view) and background. For each input image $x_i$, we introduce two image-specific subject-unrelated embeddings $v_i^p$ ([$P_i$]) and $v_i^b$ ([$B_i$]) designated to represent the pose and background information, respectively. [$P_i$] and [$B_i$] are the corresponding pseudo-

| Method | CLIP-T (↑) | CLIP-I (↑) | CLIP-I (FG) (↑) | DINO-I (↑) | DINO-I (FG) (↑) |
|---|---|---|---|---|---|
| Textual Inversion (Gal et al. 2022) | 0.2877 | 0.6833 | 0.6793 | 0.5203 | 0.5178 |
| Custom Diffusion (Kumari et al. 2023) | 0.3143 | 0.7870 | 0.7553 | 0.6602 | 0.5970 |
| SVDiff (Han et al. 2023) | 0.3167 | 0.7782 | 0.7449 | 0.6297 | 0.5672 |
| ViCo (Hao et al. 2023) | 0.2908 | 0.7774 | 0.7445 | 0.6337 | 0.5642 |
| Dreambooth (Ruiz et al. 2023) | 0.3113 | **0.7897** | 0.7499 | **0.6636** | 0.5981 |
| Dreambooth+LoRA (Hu et al. 2021) | 0.3123 | 0.7872 | 0.7498 | 0.6498 | 0.5710 |
| Ours | **0.3249** | 0.7824 | **0.7688** | 0.6504 | **0.6121** |

Table 1: Quantitative comparisons with existing methods. For subject fidelity metrics, we compute on both the original input images (CLIP-I and DINO-I) and foreground-only input images (CLIP-I (FG) and DINO-I (FG)).

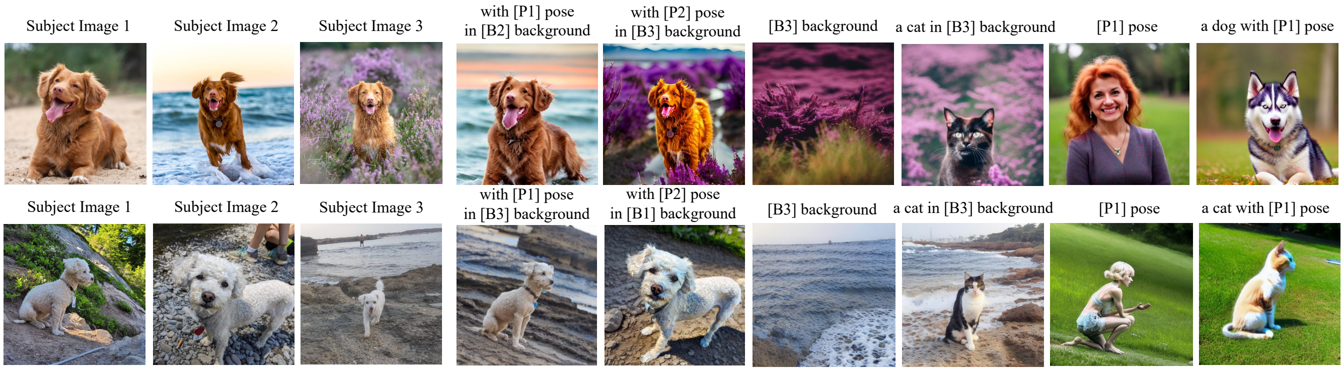


Figure 4: Visualization of decoupled textual embeddings. The unrelated tokens [P] and [B] can be used with various kinds combinations. The specific attributes are well maintained in the corresponding generation results.

words. These embeddings are incorporated with specified attribute words to capture the corresponding information, which takes the form of "`Photo of a [V] class with [P] pose/view in [B] background`".

Since the pose and background may vary significantly across different images, directly optimizing the corresponding subject-unrelated embeddings is usually non-trivial. To facilitate the embedding learning, we propose an embedding mapper for each attribute that projects each image as the corresponding subject-unrelated embedding, as shown in Fig. 2. Specifically, we leverage a pretrained CLIP (Radford et al. 2021) image encoder to extract features for image $x_i$. Then, the pose mapper projects the CLIP feature to subject-unrelated pose embedding,

$$v_i^p = M_i^P(E_I(x_i)), \quad (3)$$

where $E_I(\cdot)$ is the pretrained CLIP image encoder. $M_i^P$ denotes the pose mapper, which is implemented as a small Multilayer Perceptron (MLP). Similarly, we can obtain the subject-unrelated background embedding $v_i^b$ through the background mapper $M_i^B$.

## Joint Training Strategy

To effectively decouple the subject concept and unrelated information, we further propose a cross-attention loss alongside a joint training strategy. Intuitively, if the pose and background embeddings are well disentangled, they should focus on the respective image regions during image generation. Therefore, we first introduce a cross-attention loss that encourages different embedding to reconstruct the related information. In particular, we employ a subject mask that highlights the target concept and the background region. During training, we constrain the cross attention map of $[P]$ word to align with the subject region, while the attention map of $[B]$ word aligns with the background region,

$$L_{CA} = \mathbb{E}_{i,z,t}[||A(P_i, z_t) - m_i|| + ||A(B_i, z_t) - \overline{m_i}||], \quad (4)$$

where $A(P_i, z_t)$ and $A(B_i, z_t)$ are the cross attention maps w.r.t., $B_i$ and $P_i$, respectively. $m_i$ denotes the subject mask corresponding to input image $x_i$ and $\overline{m_i} = 1 - m_i$. In conclusion, our overall training objective is,

$$L = L_{LDM} + \lambda_{pr} L_{pr} + \lambda_{CA} L_{CA}, \quad (5)$$

where $\lambda_{pr}$ and $\lambda_{pr}$ are weights balancing different losses. In our implementation, we set $\lambda_{pr} = 1.0$ and $\lambda_{CA} = 0.01$.

Additionally, we present a joint training strategy to further enhance the disentanglement. Specifically, we introduce a random image background filtering mechanism with a probability of $\gamma$ during training to prevent the subject embedding and pose embedding from entangling the background information. In this case, we set the input text prompt $y'$ as "`Photo of a [V] class with [P] pose/view`", while using the background-masked image $\{x_i'\}_{i=1}^N$ as model input. To achieve this, we pre-process the input image set $\{x_i\}_{i=1}^N$ using the subject mask $\{m_i\}_{i=1}^N$, resulting in foreground subject images with a blank back-

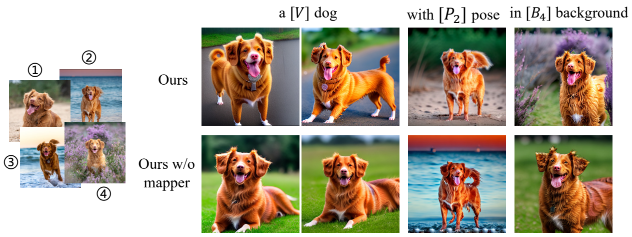


Figure 5: Visual comparisons on the effect of attribute mappers. With the attribute mappers, our method can effectively disentangle the subject, pose, and background information. Our learned subject embedding can be used to generate target concept with various pose and background. These embeddings independently control the specific attributes.

ground. With this, our method can effectively disentangle the subject-unrelated pose and background information, obtaining a well-editable subject concept for flexible customized text-to-image generation.

# Experiments

## Experimental Details

**Datasets.** We conduct our experiment on the DreamBench (Ruiz et al. 2023) dataset. It consists of 30 subjects with different categories (*e.g.*, animals, toys, and wearable items), and each subject has $4 \sim 7$ images. We randomly filter out some images to ensure that each subject only keeps 4 images for training. There are 25 editing prompts for each subject. For evaluation, we randomly generate 8 images for each subject-prompt pair, obtaining 6,000 images in total.

**Metrics.** Following Dreambooth (Ruiz et al. 2023), we evaluate our method with three metrics: CLIP-T, CLIP-I and DINO-I. CLIP-T calculates the feature similarity between the CLIP visual feature of the generated image and the CLIP textual feature of the corresponding prompt text. We omit the placeholder and keep the class name when calculating the textual feature. CLIP-I calculates the CLIP visual similarity between the generated and target concept image. DINO-I calculates the feature similarity between the ViTS/16 DINO (Caron et al. 2021) embeddings of generated and concept images.

**Implementation Details.** We employ Stable Diffusion v1-4 as our pretrained text-to-image model. The training process is conducted on RTX 3090 using AdamW (Loshchilov and Hutter 2018) optimizer with a batch size of 4 for 600 steps. The learning rate is set as 1e-5, and the drop probability $\gamma$ is set as 0.5. Our embedding mappers are implemented as the 3-layer MLPs, and each has a size of 6.7 MB. The cross attention loss (Eqn. 4) is calculated at the resolution of 32 × 32, and we average the attention maps along the head dimension. During testing, we generate images with 50 DDIM (Song, Meng, and Ermon 2020) steps, and the scale of classifier-free guidance is 6.

| Method | Ours vs. CD | Ours vs. DB | Ours vs. ViCo | Ours vs. SVDiff |
|---|---|---|---|---|
| Text-alignment | 58.30 | 60.65 | 71.45 | 54.95 |
| Image-alignment | 67.90 | 54.15 | 79.55 | 75.25 |

Table 2: User study. The numbers indicate the percentage (%) of volunteers who favor the results of our method over those of the competing methods based on the given question.

## Qualitative Evaluation

We first qualitatively compare our DETEX with existing methods, including Textual Inversion (Gal et al. 2022), Dreambooth (Ruiz et al. 2023), Custom Diffusion (Kumari et al. 2023), SVDiff (Han et al. 2023), and ViCo (Hao et al. 2023). The comparisons are illustrated in Fig. 3. We can see, our method shows superior ability in generating images with higher subject fidelity and text alignment. For subject fidelity, our subject embedding $v$ faithfully captures the target concept, and generates image with concise details (*e.g.*, the appearance of the robot on the $4th$ row and the pattern on the bottle body on the $5th$ row). For text alignment, our method exhibits flexible editability, allowing it to compose into new scene. For instance, with the prompt "`on top of a purple rug in a forest`" ($2nd$ row), our method can generate images that align well with text, whereas the competitors may overlook the "`in a forest`". Furthermore, we can selectively retain specific irrelevant attributes by keeping corresponding unrelated embeddings, thereby allowing users to flexibly control the image generation (see Fig. 1). More qualitative results are provided in the *Supple*.

## Quantitative Evaluation

We further conduct the quantitative evaluation to validate the effectiveness of our DETEX. As shown in Table 1, our methods achieves better text alignment (*i.e.*, CLIP-T) compared to the state-of-the-art methods, demonstrating its superior editability. Although existing methods like Dreambooth and Custom Diffusion exhibit higher image alignment scores (*i.e.*, CLIP-I and DINO-I), these methods tend to entangle the subject-irrelevant information (*e.g.*, background) with the learned concept and inevitably contain these information in the generated images (*e.g.*, the plant of Custom Diffusion on the $2nd$ row). Consequently, a direct comparison of CLIP-I and DINO-I metrics may be unfair since they are calculated on full images. To alleviate this issue, we additionally calculate CLIP-I and DINO-I metrics solely within the foreground-subject region, denoted as CLIP-I(FG) and DINO-I(FG), respectively. As shown in Table 1, our method has better CLIP-I(FG) and DINO-I(FG), highlighting the ability to faithfully capture the target concept. Furthermore, it is worth noting that the image alignment of existing methods consistently decreases after filtering out the background. In contrast, our method shows less decrease in values, thereby supporting our earlier analysis.

**User Study.** We performed a user study to compare our approach with existing methods. Given a subject, users were presented with two synthesized images, asked to select the

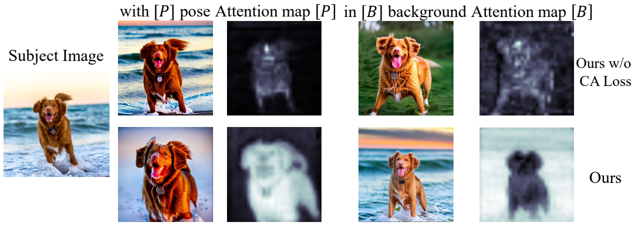


Figure 6: Visual comparisons on the effect of cross attention loss. Without the cross-attention loss, the learned pose is entangled with background information.

| Method | CLIP-T | CLIP-I(FG) | DINO-I(FG) |
|---|---|---|---|
| w/o Mapper | 0.3295 | 0.7563 | 0.5789 |
| w/o Attn Loss | 0.3287 | 0.7548 | 0.5766 |
| w/o Joint Training | **0.3327** | 0.7403 | 0.5625 |
| Ours | 0.3301 | **0.7695** | **0.5797** |

Table 3: Ablation study. With the proposed attribute mapper, attention loss $L_{CA}$, and joint training strategy, our DETEX achieves the best subject fidelity while maintaining a comparable text alignment performance.

better one from two aspects: i) Text alignment: “Which image is more consistent with the text?”; ii) Image alignment: “Which image better represents the objects in target images?”. For each evaluated view, we employ 20 users, each user is asked to answer 400 randomly selected questions, resulting in 8000 responses in total. As shown in Table 2, our method receives more preference than other methods.

## Ablation Study

We have conducted the ablation studies to evaluate the effects of various components in our method, including the multiple textual word embeddings, attribute mappers, the cross attention loss, and the joint training strategy.

**Effect of Decoupled Textual Embeddings.** We first visualize the textual word embeddings learned by our method, and the results are illustrated in Fig. 4. Our DETEX can combine the pose and background from different images with the target concept. Meanwhile, the unrelated tokens can be separately used to represent the specific attributes. They can also be combined with non-target concept. All the results above demonstrate our method decouples the irrelevant information from the subject embedding successfully.

**Effect of Attribute Mappers.** We have conducted the ablation to evaluate the effect of the attribute mappers. Specifically, we remove the mapper and optimize the pose and background embeddings directly. As shown in Fig. 5, without the attribute mappers, the subject embedding tends to entangle with pose information, while the pose embedding entangles with background information. In contrast, with the mappers, our method can disentangle each information successfully. From Table 3, our method with attribute mappers achieves better on both text and image alignment scores.

**Effect of Cross Attention Loss.** We also study the effect of introduced cross-attention loss and compare it with the variant that is trained without $L_{CA}$. As shown in Fig. 6, without the cross-attention loss, the subject-unrelated embeddings `[P]` and `[B]` fail to exhibit independent control over the pose and background, respectively. Neither `[P]` nor `[B]` has significant attention at the target region. After applying the attention constraint, our method performs better in the disentanglement of unrelated information, leading to higher text and image alignment scores (as shown in Table 3).

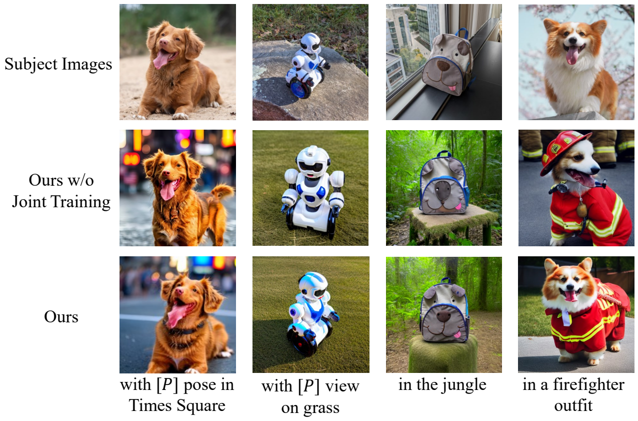


Figure 7: Visual comparisons on the effect of the joint training strategy. Without the joint training, the learned concept tends to have inconsistent details with the given image (*e.g.*, the backpack in the $2rd$ row).

**Effect of the Joint Training Strategy.** The joint training strategy effectively facilitates the decoupling between learned embeddings, resulting in better consistency between the learned subject embedding and the target concept. To demonstrate this, we additionally train our method without the joint training strategy (*i.e.*, $\gamma$ = 0). As shown in Fig. 7, without the joint training strategy, the learned pose and background embeddings are entangled, and the details of subject embedding (`[V]`) are also inconsistent with the target concept (*e.g.*, the backpack in $3rd$ column). From Table 3, with the joint training strategy, better decoupling between the learned embeddings is achieved, resulting in improved maintenance of target concept consistency.

## Conclusion

In this paper, we proposed a novel method, namely DETEX, for flexible customized text-to-image generation. In particular, our DETEX adopts multiple decoupled textual embeddings to represent the subject information and the subject-unrelated pose and background information separately. Our proposed joint training strategy and the cross-attention loss further facilitate the decoupling between the subject embedding and the irrelevant embeddings, while improving the consistency between the learned subject and the target concept. Quantitative and qualitative evaluations demonstrate that our method outperforms the SOTA methods in terms of both editing flexibility and subject fidelity. Next, we will try to investigate the ability to adaptively identify meaningful attributes of the given subject (*e.g.*, face attributes), thus facilitating a more flexible generation process.

## Acknowledgments

This work was supported in part by National Key R&D Program of China under Grant No. 2020AAA0104500, and the National Natural Science Foundation of China under Grant No.s 62176249 and U19A2073.

## References

Avrahami, O.; Aberman, K.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. Break-A-Scene: Extracting Multiple Concepts from a Single Image. *arXiv preprint arXiv:2305.16311*.

Avrahami, O.; Lischinski, D.; and Fried, O. 2022. Blended Diffusion for Text-driven Editing of Natural Images. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Caron, M.; Touvron, H.; Misra, I.; Jegou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging Properties in Self-Supervised Vision Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*.

Chen, H.; Zhang, Y.; Wang, X.; Duan, X.; Zhou, Y.; and Zhu, W. 2023. DisenBooth: Disentangled Parameter-Efficient Tuning for Subject-Driven Text-to-Image Generation. *arXiv preprint arXiv:2305.03374*.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dhariwal, P.; and Nichol, A. 2021. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34: 8780–8794.

Gal, R.; Alaluf, Y.; Atzmon, Y.; Patashnik, O.; Bermano, A. H.; Chechik, G.; and Cohen-or, D. 2022. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *The Eleventh International Conference on Learning Representations*.

Han, L.; Li, Y.; Zhang, H.; Milanfar, P.; Metaxas, D.; and Yang, F. 2023. Svdiff: Compact parameter space for diffusion fine-tuning. *arXiv preprint arXiv:2303.11305*.

Hao, S.; Han, K.; Zhao, S.; and Wong, K.-Y. K. 2023. ViCo: Detail-Preserving Visual Condition for Personalized Text-to-Image Generation. *arXiv preprint arXiv:2306.00971*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising Diffusion Probabilistic Models. *Neural Information Processing Systems*.

Ho, J.; and Salimans, T. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*.

Hu, E. J.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W.; et al. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Kim, G.; Kwon, T.; and Ye, J. C. 2022. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2426–2435.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1931–1941.

Li, X.; Hou, X.; and Loy, C. C. 2023. When StyleGAN Meets Stable Diffusion: a W+ Adapter for Personalized Image Generation. *arXiv preprint arXiv:2311.17461*.

Liu, Z.; Feng, R.; Zhu, K.; Zhang, Y.; Zheng, K.; Liu, Y.; Zhao, D.; Zhou, J.; and Cao, Y. 2023. Cones: Concept neurons in diffusion models for customized generation. *arXiv preprint arXiv:2303.05125*.

Loshchilov, I.; and Hutter, F. 2018. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Nichol, A. Q.; Dhariwal, P.; Ramesh, A.; Shyam, P.; Mishkin, P.; Mcgrew, B.; Sutskever, I.; and Chen, M. 2022. GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models. In *International Conference on Machine Learning*, 16784–16804.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 10684–10695.

Ruiz, N.; Li, Y.; Jampani, V.; Pritch, Y.; Rubinstein, M.; and Aberman, K. 2023. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22500–22510.

Saharia, C.; Chan, W.; Saxena, S.; Li, L.; Whang, J.; Denton, E. L.; Ghasemipour, K.; Gontijo Lopes, R.; Karagol Ayan, B.; Salimans, T.; et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35: 36479–36494.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Shi, J.; Xiong, W.; Lin, Z.; and Jung, H. J. 2023. Instantbooth: Personalized text-to-image generation without test-time finetuning. *arXiv preprint arXiv:2304.03411*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising Diffusion Implicit Models. In *International Conference on Learning Representations*.

Wei, Y.; Zhang, Y.; Ji, Z.; Bai, J.; Zhang, L.; and Zuo, W. 2023. ELITE: Encoding Visual Concepts into Textual Embeddings for Customized Text-to-Image Generation.

Zhang, Y.; Wei, Y.; Jiang, D.; Zhang, X.; Zuo, W.; and Tian, Q. 2023. ControlVideo: Training-free Controllable Text-to-Video Generation. *arXiv preprint arXiv:2305.13077*.