# CoRe: Context-Regularized Text Embedding Learning for Text-to-Image Personalization

**Feize Wu[1*], Yun Pang[1*], Junyi Zhang[1*], Lianyu Pang[1*], Jian Yin[1], Baoquan Zhao[1], Qing Li[2], Xudong Mao[1†]**

[1]Sun Yat-sen University
[2]The Hong Kong Polytechnic University

## Abstract

Recent advances in text-to-image personalization have enabled high-quality and controllable image synthesis for user-provided concepts. However, existing methods still struggle to balance identity preservation with text alignment. Our approach is based on the fact that generating prompt-aligned images requires a precise semantic understanding of the prompt, which involves accurately processing the interactions between the new concept and its surrounding context tokens within the CLIP text encoder. To address this, we aim to embed the new concept properly into the input embedding space of the text encoder, allowing for seamless integration with existing tokens. We introduce Context Regularization (CoRe), which enhances the learning of the new concept's text embedding by regularizing its context tokens in the prompt. This is based on the insight that appropriate output vectors of the text encoder for the context tokens can only be achieved if the new concept's text embedding is correctly learned. CoRe can be applied to arbitrary prompts without requiring the generation of corresponding images, thus improving the generalization of the learned text embedding. Additionally, CoRe can serve as a test-time optimization technique to further enhance the generations for specific prompts. Comprehensive experiments demonstrate that our method outperforms several baseline methods in both identity preservation and text alignment.

**Code** — https://github.com/pangy9/CoRe
**Extended version** — https://arxiv.org/abs/2408.15914


Figure 1: CoRe enables text-aligned personalized generations, allowing for high visual variability of the user-provided concept.

## Introduction

Text-to-image personalization involves adapting a pre-trained diffusion model to generate novel images based on user-provided concepts and text prompts. The goal of personalization techniques is to produce high-quality images that not only accurately preserve the concept's identity but also align well with the text prompt. However, balancing the trade-off between identity preservation and text alignment remains a core challenge in personalization of diffusion models.

In this work, we focus on improving text alignment for text-to-image personalization. Our investigation is based on the fact that a precise semantic understanding of the prompts is the premise for aligning the generated images with the prompts. The semantic understanding of the prompts is managed by the CLIP text encoder, which involves processing the text embeddings of all tokens and their interactions. Therefore, we aim to learn a proper text embedding for the new concept, which not only accurately represents the concept but also seamlessly integrates with existing tokens.

Instead of investigating the text embedding of the new concept itself (Gal et al. 2022), we shift our focus to the context tokens surrounding the new concept in prompts. Here, the term *text embedding* refers to the input to the CLIP text encoder. Moreover, for clarity and following the terminology used in (Lu et al. 2022), we refer to the embeddings before and after the CLIP text encoder also as the *input embedding* and *output embedding*, respectively. As illustrated

[*]These authors contributed equally.
[†]Corresponding author.
Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

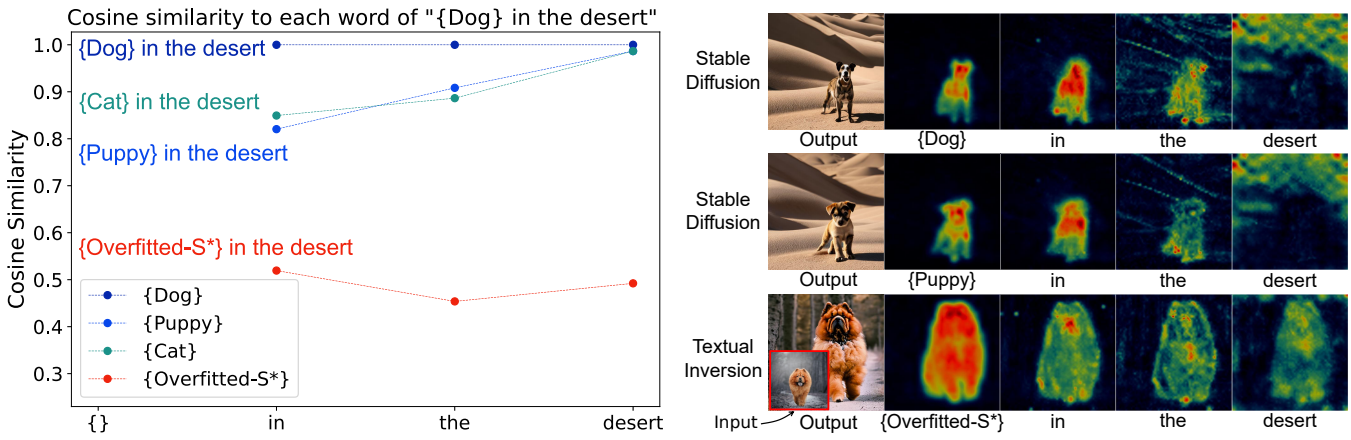


Figure 2: For the four similar prompts ("{} in the desert"), we show the cosine similarity between the output embeddings of each token (left), and the cross-attention map visualization of each token (right). Replacing "dog" with "puppy" or "cat" results in similar output embeddings and attention maps for other tokens. In contrast, using the overfitted $S_*$ by Textual Inversion significantly alters the output embeddings and attention maps. A more comprehensive analysis is provided in the Appendix.

in Figure 2, consider a scenario where the object token in a prompt is switched from "dog" to "cat"; the output embeddings of the other tokens largely remain consistent. However, using an overfitted text embedding by Textual Inversion (Gal et al. 2022) significantly alters the output embeddings of its context tokens. This alteration occurs because the overfitted embedding adversely affects the output of the context tokens within the text encoder. As shown, these inappropriate output embeddings subsequently lead to incorrect allocations in the attention maps for the context tokens.

Based on these observations, we introduce a new method named Context Regularization (CoRe), which enhances text embedding learning for a new concept by regularizing its context tokens. CoRe can improve the compatibility of the new concept's text embedding, thereby facilitating a more precise semantic understanding of the prompt. As indicated in Figure 2, replacing the object token "dog" with a compatible embedding of the new concept should yield similar output embeddings for the context tokens. Therefore, we propose a regularization strategy that encourages the output embeddings of the context tokens to align with those from a reference prompt containing a super-category token. As the attention maps play a crucial role in generation, we also impose constraints on the attention maps for the context tokens. We avoid imposing constraints directly between the new concept and its super-category, due to the typically substantial differences between them.

As our proposed CoRe is applied only to the output embeddings and attention maps, without the need for generating images, it can be used with arbitrary prompts. Therefore, we construct a regularization prompt set that covers a broad range of prompts to improve the generalization of the new concept's text embedding. During training, a prompt is randomly selected from this set for regularization purposes. Moreover, at test time, CoRe can serve as a test-time optimization approach to further refine the generations for specific prompts.

We demonstrate the effectiveness of our method by comparing it with four state-of-the-art personalization methods through both qualitative and quantitative evaluations. Our method shows superior performance in identity preservation and text alignment compared to the baselines, especially for prompts requiring high visual variability. Moreover, in addition to personalizing general objects, our method also works well for face personalization, generating more identity-preserved face images compared to three recent face personalization methods.

## Related Work

**Text-to-Image Generation.** Text-to-image generation involves creating visual images from textual prompts, a task that has seen significant advances with diffusion models (Sohl-Dickstein et al. 2015; Ho, Jain, and Abbeel 2020; Nichol and Dhariwal 2021). To achieve high-resolution text-to-image generation, various methods have been developed, such as DALL-E 2 (Ramesh et al. 2022), Imagen (Saharia et al. 2022), and LDM (Rombach et al. 2022).

**Text-to-Image Personalization.** Text-to-image personalization focuses on adapting pre-trained diffusion models to incorporate new concepts with a few user-provided images. Early methods invert the new concept into the text embedding (Gal et al. 2022), the entire U-Net (Ruiz et al. 2023), or a few parameters within the U-Net (Kumari et al. 2023). Recent advancements in text-to-image personalization have significantly improved identity preservation (Voynov et al. 2023; Alaluf et al. 2023; Zhou et al. 2023) and text alignment (Tewel et al. 2023; Qiu et al. 2024). Some works (Arar et al. 2024; Huang et al. 2024b) enhance text alignment by optimizing generations for specific prompts at test time. Moreover, tuning-free approaches (Wei et al. 2023; Shi et al.

2023; Li, Li, and Hoi 2023; Ye et al. 2023) focus on accelerating the personalization process. Additionally, many studies (Xiao et al. 2023; Wang et al. 2024a; Li et al. 2023; Wang et al. 2024b) concentrate on the personalized synthesis of widely-interested human faces.

**Text Embedding Learning.** Customizing a concept by inverting it into the text embedding space was first introduced in Textual Inversion (Gal et al. 2022). XTI (Voynov et al. 2023) extends this space to be more expressive by using multiple tokens, assigning one token per attention layer. NeTI (Alaluf et al. 2023) further expands the text embedding space to depend on both the denoising timestep and the U-Net layer. AttnDreamBooth (Pang et al. 2024b) suggests optimizing the text embedding for the new concept with very few steps, as it is prone to overfitting. A concurrent work, ClassDiffusion (Huang et al. 2024a), also utilizes a super-category token to guide the learning of text embeddings for the new concept. Our work differs in several aspects. First, we use context tokens to indirectly regularize the text embedding learning without including the super-category token, whereas concurrent work focuses on narrowing the gap between the new concept and its super-category. Second, we further regularize the attention maps of the context tokens. Third, we construct a regularization prompt set to cover a broad range of prompts, making the learned text embeddings more generalizable.

**Cross-Attention Control.** The control of the cross-attention maps has demonstrated the effectiveness in image synthesis for diffusion models (Chefer et al. 2023). Several studies (Jin et al. 2023; Nam et al. 2024; Ma et al. 2023; Zhang et al. 2024) have also investigated controlling the attention maps for text-to-image personalization. Custom Diffusion (Kumari et al. 2023) illustrates how incorrect attention maps can lead to failed compositions involving multiple concepts. ViCo (Hao et al. 2023) proposes regularizing the attention maps to focus on meaningful regions. Break-A-Scene (Avrahami et al. 2023) enhances the generation of multiple concepts by using segmentation masks to guide the learning of the attention maps.

## Preliminaries

**Latent Diffusion Models.** In Latent Diffusion Model (LDM) (Rombach et al. 2022), an encoder $\mathcal{E}$ transforms an image $x$ into a latent representation $z = \mathcal{E}(x)$ in lower-dimensional space, and a decoder $\mathcal{D}$ reconstructs the image from this latent code, i.e., $\mathcal{D}(\mathcal{E}(x)) \approx x$. Furthermore, a denoising diffusion probabilistic model (Ho, Jain, and Abbeel 2020) is utilized to generate latent codes within the autoencoder's latent space. To create images from textual descriptions, the model relies on a conditioning input vector $c(y)$, which is derived from the given text prompt $y$. The training objective of LDM is expressed as follows:

$$\mathcal{L}_{\text{diffusion}} = E_{z\sim\mathcal{E}(x),y,\varepsilon\sim\mathcal{N}(0,1),t}\left[\|\varepsilon - \varepsilon_\theta\left(z_t, t, c(y)\right)\|_2^2\right], \quad (1)$$

where the denoising network $\varepsilon_\theta$ is used to remove the noise added to the latent code given the noised latent $z_t$, the timestep $t$ and the conditioning vector $c(y)$.

**Textual Inversion.** Given several image examples of a target concept, Textual Inversion (TI) (Gal et al. 2022) learns the concept by inverting it into the text embedding space. TI introduces a new token $S_*$ and a corresponding embedding $v_*$. During the learning process, $v_*$ is optimized to minimize the diffusion loss (Eq. 1) as follows:

$$v_* = \arg\min_v E_{z,y,\epsilon,t}\left[\|\varepsilon - \varepsilon_\theta(z_t, t, c(y, v))\|_2^2\right], \quad (2)$$

where $c(y, v)$ denotes the the conditioning vector using the optimized text embedding $v$.

**DreamBooth.** DreamBooth (Ruiz et al. 2023) fine-tunes the entire U-Net of the diffusion model to learn the target concept. It employs a rarely used token to represent the concept and fixes its text embedding during optimization. Since the entire U-Net and possibly the text encoder are fine-tuned, DreamBooth usually achieves better identity preservation than Textual Inversion.

## Method

### Text Embedding Learning with CoRe

To achieve text-aligned generations, we aim to learn an appropriate text embedding for the new concept that is compatible with and seamlessly integrates into existing tokens. This is because text-aligned generations depend on a precise semantic understanding of the prompt, which in turn depends on the correct interactions between the new concept and the other tokens. Instead of directly improving the new concept's embedding, we focus on constraining the context tokens surrounding the new concept. Our method derives from two key insights. First, proper output embeddings of the context tokens can only be achieved if the new concept's input embedding is correctly learned; otherwise, it adversely impacts the output embeddings of the context tokens within the text encoder. Second, when replacing the object token in a prompt with another, the output embeddings and attention maps of the context tokens should largely remain consistent. We verify these insights through experiments illustrated in Figure 2. For instance, in the prompt "dog in the desert", replacing "dog" with an overfitted embedding by Textual Inversion significantly alters the output embeddings and attention maps of the other tokens. In contrast, replacing "dog" with "cat" maintains the consistency of the output embeddings and attention maps.

Based on these insights, we propose Context Regularization (CoRe) that enhances the text embedding learning for the new concept by regularizing its context tokens. For a training prompt containing the new concept, we construct a reference prompt by replacing the new concept token with a super-category token. We then enforce a similarity constraint on the output embeddings and attention maps of the context tokens between these two prompts. It is important to note that our context regularization can be used with arbitrary prompts because it is applied only to the output embeddings and attention maps, without the need for generating images. Therefore, we construct a regularization prompt set designed to cover a broad range of prompts, with details provided in the Appendix.

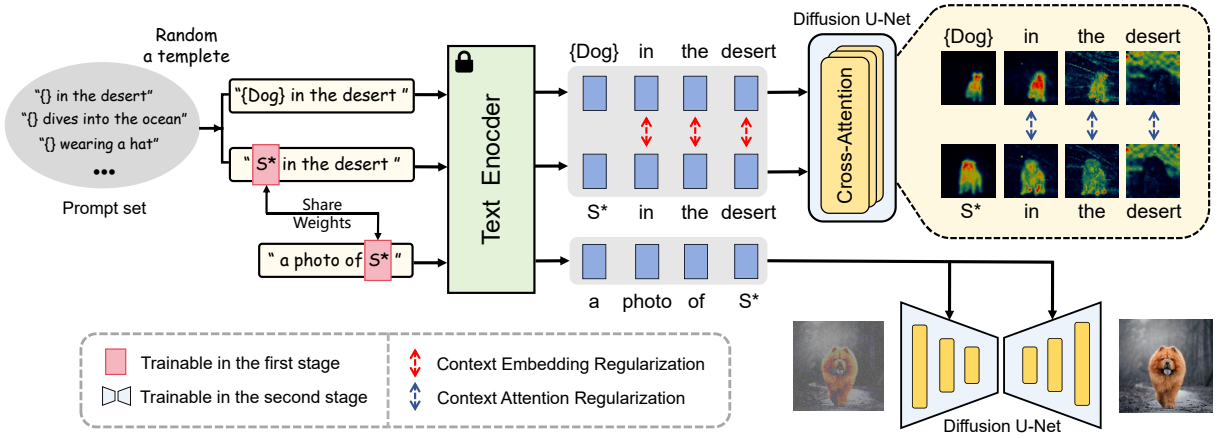


Figure 3: Overview of the proposed CoRe. Our method enhances the text embedding learning for $S_*$ by regularizing its context tokens. Specifically, we randomly select a regularization prompt (e.g., "$S_*$ in the desert") and a reference prompt (e.g., "Dog in the desert") from the prompt set. During training, the proposed context embedding regularization and context attention regularization are applied together with the diffusion loss, which encourages the representations of the context tokens surrounding $S_*$ to align with those in the reference prompt. These regularization terms make the text embedding of $S_*$ more compatible with existing tokens.

**Context Embedding Regularization.** Formally, we randomly select a prompt template (e.g., "a {} in the jungle") from the regularization prompt set, and fill it with the new concept token and its super-category token, respectively, producing a pair of prompts $y_*$ and $y_{\text{cat}}$ (e.g., "a $S_*$ in the jungle" and "a [super-category] in the jungle"). The input embeddings of these two prompts, $\{v_i\}$ and $\{v'_i\}$, are then fed into the text encoder $E$, producing corresponding output embeddings $\{E(v_i)\}$ and $\{E(v'_i)\}$. We minimize the average cosine distance between these two sets of output embeddings:

$$\mathcal{L}_{\text{emb}} = \frac{1}{m}\sum_{i=1}^{m}\big(1 - \cos(E(v_i), E(v'_i))\big), \tag{3}$$

where $\{1, 2 \dots m\}$ denotes the index set after removing the indices corresponding to $S_*$ and the super-category,and $\cos(\cdot,\cdot)$ denotes the cosine similarity. Note that we avoid imposing constraints between $S_*$ and [super-category] due to the observed significant degradation in identity preservation, as the new concept and its super-category usually exhibit substantial differences.

**Context Attention Regularization.** As illustrated in Figure 2, the overfitted embedding of the new concept subsequently results in incorrect attention maps for the context tokens. Therefore, we utilize attention maps to further regularize the text embedding learning for the new concept. We introduce an additional regularization term that enforces a similarity constraint between the attention maps of the two prompts, $y_*$ and $y_{\text{cat}}$. Formally, the output embeddings $\{E(v_i)\}$ and $\{E(v'_i)\}$ for $y_*$ and $y_{\text{cat}}$ are fed into the 16 different cross-attention layers of the U-Net, generating 16 attention maps $\{M_i^{1:16}\}$ and $\{M_i'^{1:16}\}$, respectively. We minimize the average squared difference between the mean values of these attention maps as follows:

$$\mathcal{L}_{\text{attn}} = \frac{1}{m}\sum_{i=1}^{m}\big(\mu(M_i^{1:16}) - \mu(M_i'^{1:16})\big)^2, \tag{4}$$

where $\mu(M^{1:16})$ denotes the mean of all values across the 16 attention maps, and $\{1, 2 \dots m\}$ denotes the index set after removing the indices corresponding to $S_*$ and the super-category.

Overall, the full optimization objective of our method is defined as:

$$v_* = \arg\min_{v_k} \mathcal{L}_{\text{diffusion}} + \lambda_{\text{emb}}\mathcal{L}_{\text{emb}} + \lambda_{\text{attn}}\mathcal{L}_{\text{attn}}. \tag{5}$$

**Embedding Rescaling.** As identified in (Alaluf et al. 2023; Pang et al. 2024a), during optimization, the scale of the new concept's text embedding tends to become excessively large, leading to significant degradation in text alignment. Inspired by (Alaluf et al. 2023), we propose rescaling the norm of the text embedding during optimization to mitigate this issue. Specifically, after one optimization step, we reset the norm of the updated embedding to match the norm from the previous step. The rescaled embedding is given by:

$$v_*^s = \frac{v_*^s}{\|v_*^s\|}\|v_*^{s-1}\|, \tag{6}$$

where $s$ denotes the $s$-th optimization step. In practice, we apply this rescaling strategy only during the intermediate phase of the optimization, as we empirically find that rescaling at the beginning or end phases can lead to degraded identity preservation, likely due to the information loss introduced by rescaling.

## Embedding-to-Identity Training Strategy

Solely optimizing the text embedding is insufficient to capture the concept identity. Inspired by (Roich et al. 2022;


Figure 4: Qualitative comparison. We present personalization results of our method and four baseline methods, including Custom Diffusion (Kumari et al. 2023), NeTI (Alaluf et al. 2023), OFT (Qiu et al. 2024), and AttnDreamBooth (Pang et al. 2024b). Our method demonstrates superior performance in both text alignment and identity preservation compared to these baselines, especially for the prompts that require high visual variability of the concept.

Pang et al. 2024b), we propose a two-stage training strategy. Initially, we employ CoRe to learn a text embedding for the new concept that is compatible with existing tokens. This yields an editable embedding but provides a coarse depiction of the concept identity. In the second stage, we freeze the text embedding and fine-tune all layers of the U-Net to precisely capture the concept identity.

### Test-Time Optimization

At test time, our proposed method, CoRe, can optionally serve as a test-time optimization technique to enhance generation for specific prompts. Specifically, given a prompt for generation, we refine the output embeddings and attention maps associated with this prompt by performing a few additional optimization steps using CoRe. This refinement is done without employing the diffusion loss. Moreover, our test-time optimization technique can be effectively applied to models trained using other method, such as TI (Gal et al. 2022) and AttnDreamBooth (Pang et al. 2024b). Note that in our experiments, this test-time optimization strategy is not applied when comparing with the baselines to ensure a fair comparison.

## Experiments

**Datasets.** For a comprehensive evaluation, we collect 24 concepts from previous studies (Gal et al. 2022; Ruiz et al. 2023; Kumari et al. 2023). Following (Tewel et al. 2023), we categorize these concepts into two groups: animate objects (e.g., "cat" and "child doll") and inanimate objects (e.g., "clock" and "berry bowl"). Accordingly, we use two sets of prompts for these two groups, respectively. Some


Figure 5: Face personalization results of our method and three baseline methods, including Cross Initialization (CI) (Pang et al. 2024a), PhotoMaker (PM) (Li et al. 2023), and Face2Diffusion (FD) (Shiohara and Yamasaki 2024). Our method achieves more identity-preserved face generations compared to the baselines, especially when the input image is a side face.

| Methods | CLIP-T↑ | CLIP-I↑ | DINO↑ |
|---|---|---|---|
| Custom Diffusion | 0.2537 | 0.6706 | 0.5163 |
| NeTI | 0.2386 | **0.7104** | 0.5623 |
| OFT | 0.2397 | 0.7018 | 0.5612 |
| AttnDreamBooth | 0.2547 | 0.6918 | 0.5641 |
| Ours | **0.2568** | 0.7054 | **0.5842** |

Table 1: Quantitative comparison. CLIP-I and DINO evaluate identity preservation by measuring the similarity between the generated and input images. CLIP-T evaluates text alignment by measuring the similarity between the generated images and the text prompts.

| Baselines | Prefer Baseline | Prefer Ours |
|---|---|---|
| Custom Diffusion | 14.3% | **85.7%** |
| NeTI | 23.7% | **76.3%** |
| OFT | 28.4% | **71.6%** |
| AttnDreamBooth | 35.3% | **64.7%** |

Table 2: User study. For each paired comparison, our method is preferred over the baselines.

prompts are shared across all concepts, including background change, concept color change, and artistic style, while others are specific to animate objects, such as action and outfit change.

**Evaluation Setup.** We compare our method against four recent baseline methods: Custom Diffusion (Kumari et al. 2023), NeTI (Alaluf et al. 2023), OFT (Qiu et al. 2024), and AttnDreamBooth (Pang et al. 2024b). For quantitative evaluation, we employ a set of 20 prompts, detailed in the Appendix, using the following metrics: (1) identity preservation, measured by the visual similarity between the generated and input images in the CLIP-I (Radford et al. 2021) and DINO (Caron et al. 2021) feature spaces; and (2) text alignment, measured by the CLIP-T similarity between the generated images and the prompts. Following (Zeng et al. 2024), the CLIP-I and DINO scores are exclusively calculated on foreground-masked images to eliminate background variations and better reflect concept identity similarity. Additionally, prompts involving stylization or outfit change are excluded from the CLIP-I and DINO score calculations because these modifications can significantly alter the concept's appearance. The implementation details of our method and the baselines are provided in the Appendix.

## Results

**Qualitative Evaluation.** In Figure 4, we present a visual comparison of personalized generations for various concepts. We employ a set of complex prompts for evaluation, such as depicting the pets in a human-like posture and dressing (e.g., "$S_*$ dressed as Spider-Man swings between tall buildings"), complex spatial relationships (e.g., "$S_*$ inside a box, floating on the water"), and composition of multiple changes (e.g., "A steampunk $S_*$ with gears and pipes, exploring a retro factory"). As observed, Custom Diffusion fails to generate text-aligned images and sometimes discards the new concept in the generation. NeTI and OFT struggle to accurately adapt the given concept in new scenes. AttnDreamBooth achieves improved personalized generations, but still fails to generate identity-preserved and text-aligned images, especially for prompts requires high visual variability (e.g., "A cat$_*$ dressed as Spider-Man"). In contrast, the generations by our method accurately preserve the concept identity and align with the complex prompts. Additional qualitative results are provided in the Appendix.

Although our method is primarily designed for personalizing general objects, it also performs well in personalizing human faces. Figure 5 shows our personalization results on human faces compared with three specialized face personalization methods, including Cross Initialization (Pang et al. 2024a), PhotoMaker (Li et al. 2023), and Face2Diffusion (Shiohara and Yamasaki 2024). Our method demonstrates superior identity preservation compared to these baselines.

**Quantitative Evaluation.** We quantitatively evaluate each method using 24 concepts and 20 text prompts, generating 32 samples per prompt for each concept. To ensure a fair and unbiased evaluation, these concepts were selected from multiple datasets (Gal et al. 2022; Ruiz et al. 2023; Kumari et al. 2023), covering 8 animal toys/animals, 8 figurines, and 8 inanimate objects. The results are presented in Table 1. Note that prompts requiring high visual variability are excluded from quantitative evaluation due to the limitations of quantitative metrics in accurately assessing the quality of generated images for these prompts, for two main reasons. First, such prompts significantly alter the concept's appearance, which makes them unsuitable for mea-


Figure 6: Ablation study. We compare models trained without Context Embedding Regularization (w/o CER), without Context Attention Regularization (w/o CAR), and without embedding rescaling strategy (w/o Rescale). All submodules are essential for achieving identity-preserved and text-aligned personalized generations.

suring identity similarity to the input images. Second, methods that neglect to incorporate the new concept in generations tend to achieve high text alignment scores, as these scores are calculated without considering the new concept. Consequently, using relatively simple prompts, our method achieves slightly higher CLIP-T scores than AttnDreamBooth. In terms of CLIP-I and DINO scores, our method outperforms AttnDreamBooth, likely due to the insufficient text embedding learning in AttnDreamBooth. NeTI achieves the highest CLIP-I score but ranks lowest in text alignment, indicating a tendency to overfit the new concept. Overall, the results demonstrate that our method achieves a superior balance between identity preservation and text alignment compared to the baselines.

**User Study.** We conduct a paired human preference study to compare CoRe with the baseline methods. In each question, we present two generated images, one from our method and one from a baseline, using the same prompt. Participants are asked to evaluate the generated images based on identity preservation and text alignment. We collect 1200 responses from 60 participants. As shown in Table 2, our method is clearly preferred over the baselines, indicating its superiority in identity preservation and text alignment.

## Ablation Study

In this section, we ablate each sub-module of our method to demonstrate its contribution. Figures 6 shows the results of the ablation study. As shown, the absence of the contextual embedding regularization leads to degradation in both identity preservation and text alignment. The model without the context attention regularization tends to generate images that are similar to the input, indicating a potential overfit to the concept. Additionally, without applying the embedding rescaling strategy, the model exhibits slight degradation in both text alignment and identity preservation. Additional ablation study results can be found in the Appendix.

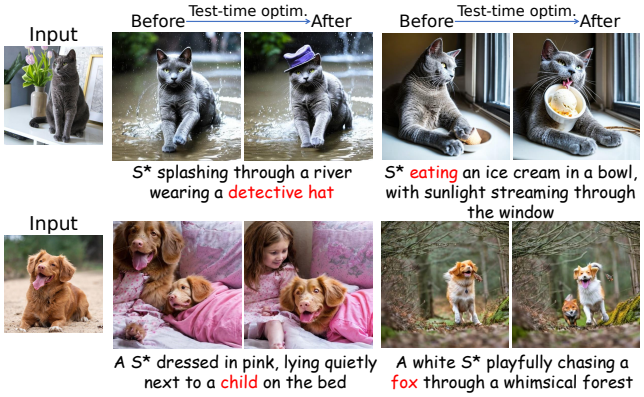


Figure 7: Serving as a test-time optimization technique, CoRe enables previously omitted words to be reflected in the generated images.

## Test-Time Optimization

In this section, we evaluate the effectiveness of CoRe for Test-Time Optimization (TTO). Given a specific prompt for generation, we perform an additional 10 optimization steps using CoRe to refine the output embeddings and attention maps for this prompt. As illustrated in Figure 7, this strategy helps to better align the generations with the prompts, allowing previously omitted words to be reflected in the new images. For example, in the second row, TTO effectively replaces the unintended "dog" with the correct "child", and retrieves the missing "fox". It is noteworthy that our TTO approach is also effective for other methods, with further details provided in the Appendix.

## Conclusions and Limitations

In conclusion, we proposed a personalization method named CoRe that enhances the text embedding learning for the new concept by regularizing context tokens. This method is based on the insight that appropriate output embeddings of context tokens are achievable only when the new concept's text embedding is correctly learned. Our experimental results demonstrate that CoRe outperforms the baseline methods. As shown in Figure 7, our method still faces challenges with difficult compositions involving the learned concept and other objects, which is partly inherited from the pretrained model. CoRe can serve as a test-time optimization technique to enhance the generation of such difficult compositions.

## Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62176223 and No. 62302535), Guangdong Basic and Applied Basic Research Foundation (No. 2023A1515012897 and No. 2023A1515011639), and Zhuhai Basic and Applied Basic Research Foundation (No. 2320004002745).

## References


Alaluf, Y.; Richardson, E.; Metzer, G.; and Cohen-Or, D. 2023. A Neural Space-Time Representation for Text-to-Image Personalization. *arXiv preprint arXiv:2305.15391*.

Arar, M.; Voynov, A.; Hertz, A.; Avrahami, O.; Fruchter, S.; Pritch, Y.; Cohen-Or, D.; and Shamir, A. 2024. PALP: Prompt Aligned Personalization of Text-to-Image Models. *arXiv preprint arXiv:2401.06105*.

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*.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *ICCV*.

Chefer, H.; Alaluf, Y.; Vinker, Y.; Wolf, L.; and Cohen-Or, D. 2023. Attend-and-Excite: Attention-Based Semantic Guidance for Text-to-Image Diffusion Models. In *SIGGRAPH*.

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. *arXiv preprint arXiv:2208.01618*.

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. In *NeurIPS*.

Huang, J.; Liew, J. H.; Yan, H.; Yin, Y.; Zhao, Y.; and Wei, Y. 2024a. ClassDiffusion: More Aligned Personalization Tuning with Explicit Class Guidance. *arXiv preprint arXiv:2405.17532*.

Huang, M.; Mao, Z.; Liu, M.; He, Q.; and Zhang, Y. 2024b. RealCustom: Narrowing Real Text Word for Real-Time Open-Domain Text-to-Image Customization. In *CVPR*, 7476–7485.

Jin, C.; Tanno, R.; Saseendran, A.; Diethe, T.; and Teare, P. 2023. An Image is Worth Multiple Words: Learning Object Level Concepts using Multi-Concept Prompt Learning. *arXiv preprint arXiv:2310.12274*.

Kumari, N.; Zhang, B.; Zhang, R.; Shechtman, E.; and Zhu, J.-Y. 2023. Multi-concept customization of text-to-image diffusion. In *CVPR*.

Li, D.; Li, J.; and Hoi, S. C. H. 2023. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. *arXiv preprint arXiv:2305.14720*.

Li, Z.; Cao, M.; Wang, X.; Qi, Z.; Cheng, M.-M.; and Shan, Y. 2023. PhotoMaker: Customizing Realistic Human Photos via Stacked ID Embedding. *arXiv preprint arXiv:2312.04461*.

Lu, Y.; Liu, J.; Zhang, Y.; Liu, Y.; and Tian, X. 2022. Prompt Distribution Learning. In *CVPR*.

Ma, J.; Liang, J.; Chen, C.; and Lu, H. 2023. Subject-Diffusion:Open Domain Personalized Text-to-Image Generation without Test-time Fine-tuning. *arXiv preprint arXiv:2307.11410*.

Nam, J.; Kim, H.; Lee, D.; Jin, S.; Kim, S.; and Chang, S. 2024. DreamMatcher: Appearance Matching Self-Attention for Semantically-Consistent Text-to-Image Personalization. *arXiv preprint arXiv:2402.09812*.

Nichol, A.; and Dhariwal, P. 2021. Improved Denoising Diffusion Probabilistic Models. In *ICML*.

Pang, L.; Yin, J.; Xie, H.; Wang, Q.; Li, Q.; and Mao, X. 2024a. Cross Initialization for Personalized Text-to-Image Generation. In *CVPR*.

Pang, L.; Yin, J.; Zhao, B.; Wu, F.; Wang, F. L.; Li, Q.; and Mao, X. 2024b. AttnDreamBooth: Towards Text-Aligned Personalized Text-to-Image Generation. *arXiv preprint arXiv:2406.05000*.

Qiu, Z.; Liu, W.; Feng, H.; Xue, Y.; Feng, Y.; Liu, Z.; Zhang, D.; Weller, A.; and Schölkopf, B. 2024. Controlling text-to-image diffusion by orthogonal finetuning. In *NeurIPS*.

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 *ICML*.

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*.

Roich, D.; Mokady, R.; Bermano, A. H.; and Cohen-Or, D. 2022. Pivotal tuning for latent-based editing of real images. *TOG*.

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *CVPR*.

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 *CVPR*.

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. In *NeurIPS*.

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*.

Shiohara, K.; and Yamasaki, T. 2024. Face2Diffusion for Fast and Editable Face Personalization. In *CVPR*, 6850–6859.

Sohl-Dickstein, J.; Weiss, E. A.; Maheswaranathan, N.; and Ganguli, S. 2015. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*.

Tewel, Y.; Gal, R.; Chechik, G.; and Atzmon, Y. 2023. Key-Locked Rank One Editing for Text-to-Image Personalization. In *SIGGRAPH*.

Voynov, A.; Chu, Q.; Cohen-Or, D.; and Aberman, K. 2023. $P+$: Extended Textual Conditioning in Text-to-Image Generation. *arXiv preprint arXiv:2303.09522*.

Wang, Q.; Bai, X.; Wang, H.; Qin, Z.; and Chen, A. 2024a. InstantID: Zero-shot Identity-Preserving Generation in Seconds. *arXiv preprint arXiv:2401.07519*.

Wang, Q.; Jia, X.; Li, X.; Li, T.; Ma, L.; Zhuge, Y.; and Lu, H. 2024b. StableIdentity: Inserting Anybody into Anywhere at First Sight. *arXiv preprint arXiv:2401.15975*.

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. *arXiv preprint arXiv:2302.13848*.

Xiao, G.; Yin, T.; Freeman, W. T.; Durand, F.; and Han, S. 2023. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. *arXiv preprint arXiv:2305.10431*.

Ye, H.; Zhang, J.; Liu, S.; Han, X.; and Yang, W. 2023. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*.

Zeng, Y.; Patel, V. M.; Wang, H.; Huang, X.; Wang, T.-C.; Liu, M.-Y.; and Balaji, Y. 2024. JeDi: Joint-Image Diffusion Models for Finetuning-Free Personalized Text-to-Image Generation. In *CVPR*.

Zhang, X.; Wei, X.-Y.; Wu, J.; Zhang, T.; Zhang, Z.; Lei, Z.; and Li, Q. 2024. Compositional Inversion for Stable Diffusion Models. In *AAAI*.

Zhou, Y.; Zhang, R.; Sun, T.; and Xu, J. 2023. Enhancing Detail Preservation for Customized Text-to-Image Generation: A Regularization-Free Approach. *arXiv preprint arXiv:2305.13579*.