Object Fusion via Diffusion Time-step for Customized Image Editing with Single Example

Xue Song[1], Zhongqi Yue[2], Jiequan Cui[3], Hanwang Zhang[2], Jingjing Chen[1, 4*]
[1] College of Computer Science and Artificial Intelligence, Fudan University
[2]College of Computing and Data Science, Nanyang Technological University
[3]School of Computer Science and Information Engineering, Hefei University of Technology
[4]Institute of Trustworthy Embodied AI, Fudan University
xuesong21@m.fudan.edu.cn, yuez0002@gmail.com, jiequancui@gmail.com,
hanwangzhang@ntu.edu.sg, chenjingjing@fudan.edu.cn

Abstract

We tackle the task of customized image editing using a text-conditioned Diffusion Model (DM). The goal is to fuse the subject in a reference image (*e.g.*, sunglasses) with a source one (*e.g.*, a boy), while retaining the fidelity of them both (*e.g.*, the boy wearing the sunglasses). An intuitive approach, called LoRA fusion, first separately trains a DM LoRA for each image to encode its details. Then the two LoRAs are linearly combined by a weight to generate a fused image. Unfortunately, even through careful grid search or learning the weight, this approach still trades off the fidelity of one image against the other. We point out that the evil lies in the overlooked role of diffusion time-step in the generation process, *i.e.*, a smaller time-step controls the generation of a more fine-grained attribute. For example, a large LoRA weight for the source may help preserve its fine-grained details (*e.g.*, face attributes) at a small time-step, but could overpower the reference subject LoRA and lose the fidelity of its overall shape at a larger time-step. To address this deficiency, we propose *TimeFusion*, which learns a time-step-specific LoRA fusion weight that resolves the trade-off, *i.e.*, generating the source and reference subject in high fidelity given their respective prompt. Then we can customize image editing using this weight and a target prompt.

Introduction

Image editing modifies an image I by altering user-defined visual attributes while retaining its fidelity, *i.e.*, preserving other attributes. For example, editing an image of a boy with the target prompt "wearing sunglasses" should only add sunglasses without altering the boy or his background. Recent efforts (Hertz et al. 2022; Orgad, Kawar, and Belinkov 2023; Tumanyan et al. 2023; Cao et al. 2023; Pan et al. 2023; Wallace, Gokul, and Naik 2023; Wu and De la Torre 2023; Kawar et al. 2023; Hertz, Aberman, and Cohen-Or 2023) utilize a text-conditioned Diffusion Model (DM) (Ho, Jain, and Abbeel 2020), whose reverse process progressively transforms random noise into an image aligning with a given text prompt. The general paradigm involves two steps: first, calibrating the reverse process to reconstruct I and retain fidelity; then, modifying it by introducing the target prompt to complete the edit. In particular, the main differences in existing methods lie in the reconstruction step. One approach, known as DDIM inversion (Song, Meng, and Ermon 2020), attempts to identify a noise initialization from which running the reverse process yields I. However, the inversion is prone to errors, leading to fidelity loss (Tumanyan et al. 2023; Wallace, Gokul, and Naik 2023). Therefore, we base our study on an improved technique (Hu et al. 2021) that involves learning a set of Low-Rank Adaptation (LoRA) layers injected into the DM (or learning a LoRA for short), which enables the LoRA-guided reverse process to reconstruct I.

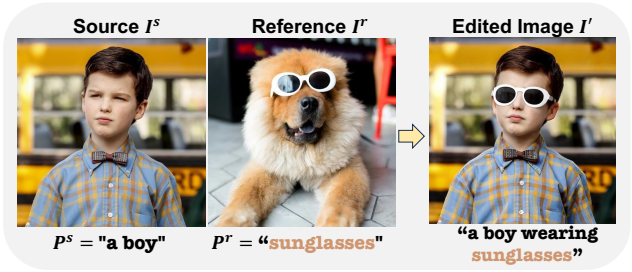

Figure 1: Customized image editing, which fuses the source image with the subject in a reference image, while retaining the fidelity of them both.

Most existing image editing methods are text-based, which oftentimes lack customizability. For example, as shown in Figure 1, a user may want to edit a source image I^s so that the boy wears the exact sunglasses in a reference image I^r to produce the edited I'. In this case, it is impractical to fully specify the appearance of the sunglasses using textual control. We aim to bridge this gap by exploring **customized image editing**, where the goal is to fuse the subject in I^r[1] with an image I^s, while retaining the fidelity of both.

To address this, a natural extension to the above editing paradigm is LoRA fusion (Ryu 2023): first, learn a source and reference LoRA to reconstruct I^s and the subject in I^r, respectively; then linearly combine their guidance in the

*Corresponding author

[1]We segment the subject with SAM (Kirillov et al. 2023), requiring the user to click its location in I^r based on the prompt P^r. Note that this can be easily automated with a text-conditioned segmentation model such as (Rasheed et al. 2024).

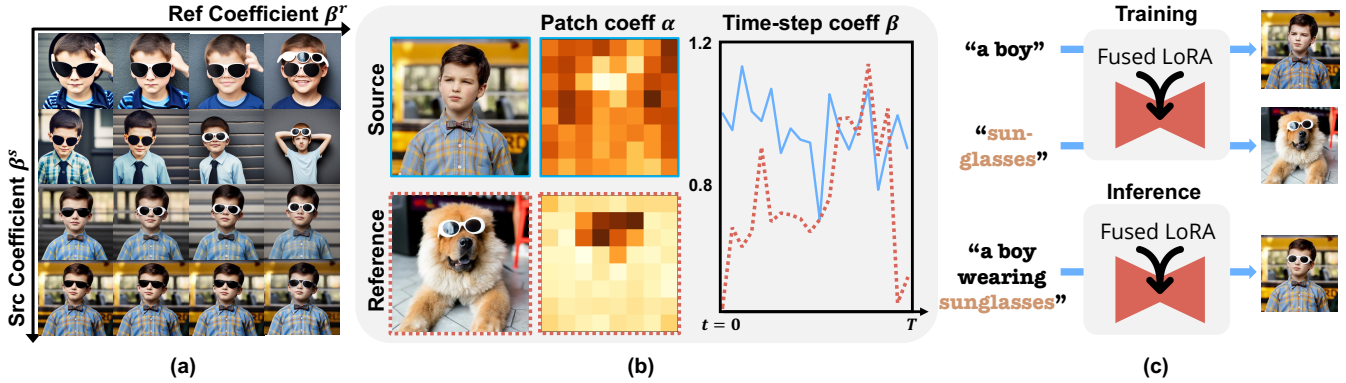

Figure 2: (a) Failure of current LoRA fusion in customized image editing, where no combination of coefficients maintains the fidelity of both images. (b) The proposed learnable time-step-specific coefficient and patch-specific one (visualized values are averaged across the LoRA injection layers). (c) The high-level training and inference pipeline for the proposed TimeFusion.

reverse process using a fusion coefficient (β^s, β^r), *e.g.*, a larger β^s increases the guidance strength from the source LoRA. However, as shown in Figure 2(a), no combination of coefficients maintains the fidelity of both I^s and I^r.

Therefore, the crux of customized image editing lies in finding a more precise method for fusing LoRAs that accommodates the visual attributes of both I^s and I^r. This motivates us to improve LoRA fusion by considering the diffusion time-step. The motivation is based on two key observations: 1) Diffusion time-step is theoretically and empirically linked to visual attributes (Yue et al. 2024a,b), where the reverse process at a smaller time-step is responsible for generating a more fine-grained attribute. For example, as we will later show in Figure 4, providing LoRA guidance at smaller time-steps modifies more fine-grained attributes. 2) Thus, a time-step-specific fusion coefficient provides the required precision, *e.g.*, using a large β^s at a small time-step to preserve the fine-grained facial details in I^s, while increasing β^r at a larger time-step to maintain the more coarse-grained shape of the sunglasses in I^r.

Building on the above analysis, we propose a time-step specific LoRA fusion strategy called **TimeFusion**. Specifically, we first learn a source and reference LoRA using a standard technique (Avrahami et al. 2023). Then our coefficients for LoRA fusion consist of two parts, as shown in Figure 2(b): 1) the aforementioned time-step specific ones (β^s_t, β^r_t) for each diffusion time-step $t \in \{1, \dots, T\}$. 2) a patch-level one $\boldsymbol{\alpha}^s, \boldsymbol{\alpha}^r \in \mathbb{R}^{8\times8}$ for the 8×8 latent patches in each layer where LoRA is injected (layer index omitted for simplicity). The patch-level coefficient acts as a modifier to the time-step one, further refining the fusion by considering spatial information, *e.g.*, applying a large source LoRA coefficient on a background patch to faithfully reconstruct I^s. Overall, during the reverse process at time-step t, the LoRA fusion coefficient for a patch at spatial location (i, j) will be $(\boldsymbol{\alpha}^s_{i,j}\beta^s_t, \boldsymbol{\alpha}^r_{i,j}\beta^r_t)$. The paradigm of TimeFusion is summarized in Figure 2(c). We train the coefficients so that the fused LoRA reconstructs I^s and I^r according to their respective prompt, *i.e.*, accommodating the visual attributes of both images. In inference, using the learned LoRA fusion, we achieve customized image editing by simply supplying the target prompt. Our contributions include:

- We tackle the challenging task of customized image editing with a text-conditioned DM, by improving the current LoRA fusion.
- Motivated by the connection between diffusion time-step and visual attribute, we propose TimeFusion, a novel time-step-specific LoRA fusion strategy.
- Extensive qualitative and quantitative experiments demonstrate the superiority of our TimeFusion over existing works in customized image editing.

Problem Formulations

Text-Conditioned Diffusion Model (DM)

A text-conditioned DM uses a *forward process* that incrementally adds noise to input data to learn a *reverse process*, where the model is trained to reconstruct clean data from noisy one and its text description. In this work, we focus on Stable Diffusion (SD) (Rombach et al. 2022), where each input data $\mathbf{z}_0$ is an image feature.

Forward Process. It progressively adds Gaussian noise to image $\mathbf{z}_0$ in T time-steps, producing noisy images $\mathbf{z}_1, \dots, \mathbf{z}_T$. Recent works (Yue et al. 2024a,b) show that the forward process connects diffusion time-step with the visual attributes of an image. In a nutshell, an increasing time-step t causes a large overlap between noisy image distributions, essentially collapsing different images into similar ones by losing the visual attributes that differentiate them. In particular, the theory suggests that fine-grained attributes (*i.e.*, those affecting local appearances, like expression) become lost at a smaller time-step compared to coarse-grained ones (*i.e.*, those affecting global appearances, like background). This pattern of attribute loss has interesting implications in the following reverse process.

Reverse Process. It corresponds to a learned Gaussian transition $p_\theta(\mathbf{z}_{t-1}|\mathbf{z}_t, P)$ conditioned on a text prompt P and parameterized by θ. The term is computed in two steps: first reconstructing $\mathbf{z}_0$ as $\mathbf{z}'_0$ by a learnable denoising network $d(\mathbf{z}_t, P, t; \theta)$ with parameter θ, then computing

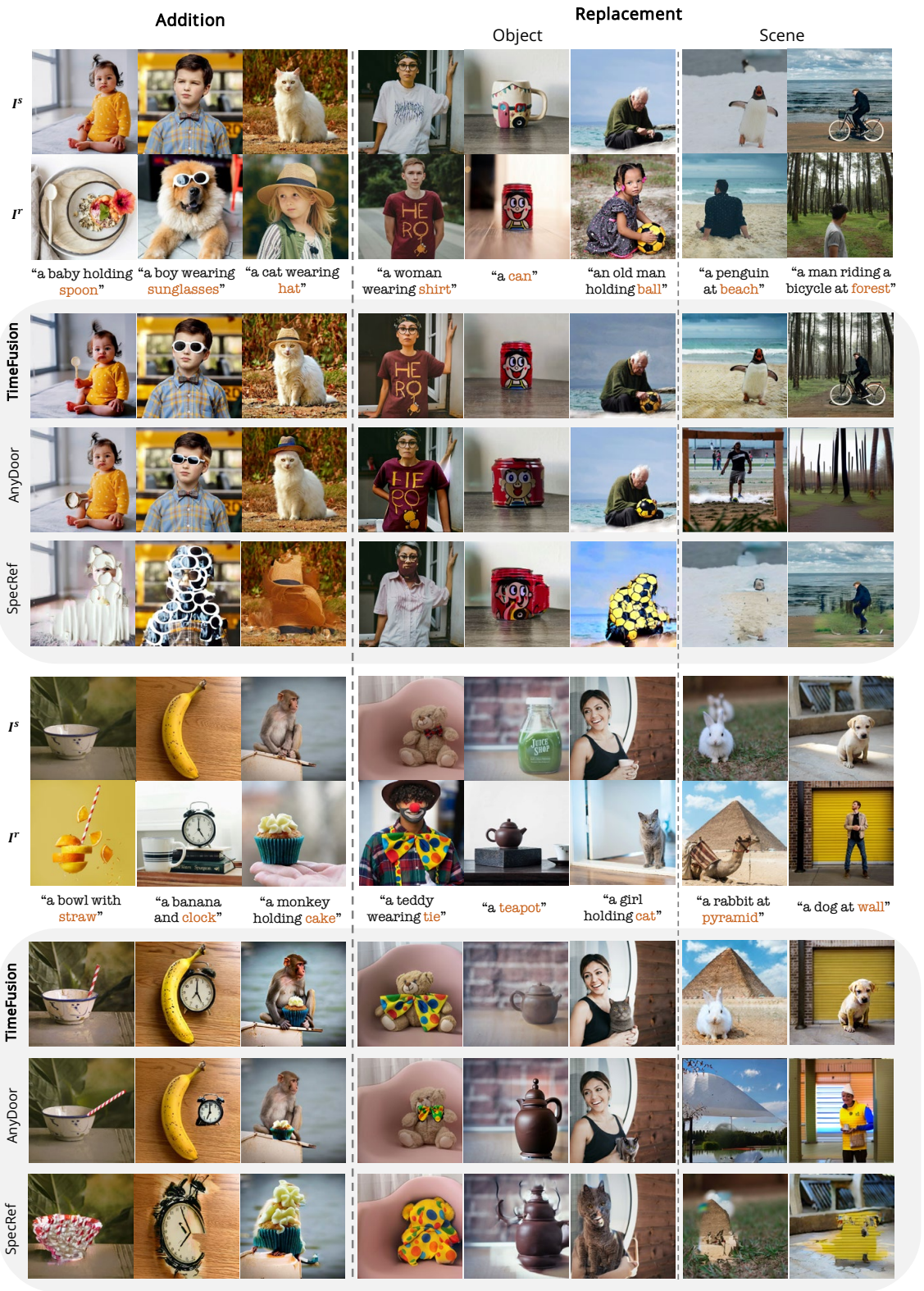

Figure 3: Comparison of customized image editing results between our TimeFusion and existing SOTAs (Chen et al. 2023; Chen and Huang 2023). We compare 3 tasks, including addition, object and scene replacement. The source prompt is omitted, and the reference subject prompt is highlighted in orange inside the target prompt. For fairness, examples are chosen based on their best visual quality from various random seeds.

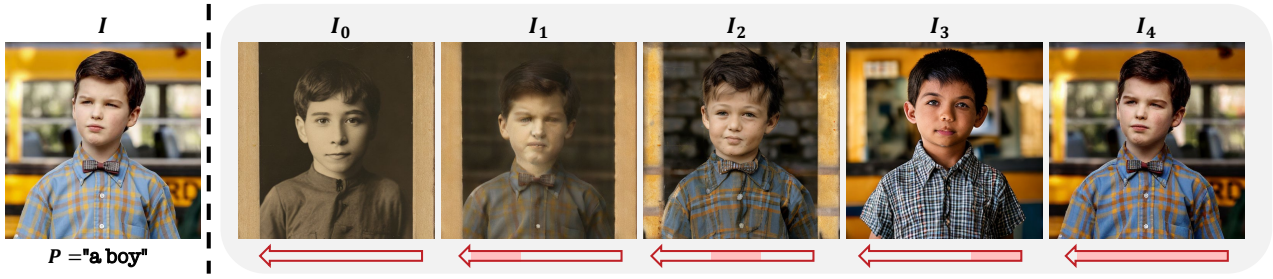

Figure 4: Effect of applying LoRA guidance at a subset of diffusion time-steps (highlighted by a red bar), ranging from no guidance in I_0 to full guidance in I_4. The LoRA is learned to reconstruct I given P.

$q(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{z}'_0)$, which has a closed-form solution. In training, we minimize the reconstruction error:

$$\mathcal{L}(\theta, \mathbf{z}_0, P) = \sum_{t=1}^{T} \mathop{\mathbb{E}}_{q(\mathbf{z}_t|\mathbf{z}_0)} \|\mathbf{z}_0 - d(\mathbf{z}_t, P, t; \theta)\|^2, \quad (1)$$

where $\mathbf{z}_t$ is sampled from the forward process at a random time-step t. In particular, due to the progressive loss of attributes going from $\mathbf{z}_0$ to $\mathbf{z}_T$ in the forward process, the reverse process must correspondingly make up for the lost attribute in each step to accurately reconstruct $\mathbf{z}_0$. Hence as fine-grained attributes are lost at smaller time-steps, the reverse process is responsible for generating them at smaller time-steps correspondingly.

Low-Rank Adaptation (LoRA)

Training a DM with a randomly initialized θ by minimizing Equation 1 can be extremely expensive. Hence the common approach is to initialize θ from a DM pre-trained on diverse data (*e.g.*, SD), and fine-tune it to a downstream task. LoRA is the most mainstream fine-tuning method (Hu et al. 2021; Song et al. 2024; Gu et al. 2024).

LoRA refers to a set of low-rank matrices, each of which is injected into a layer of DM for weight updating. Specifically, let $W \in \mathbb{R}^{m\times n}$ denote the pre-trained weight matrix of a layer in DM. The LoRA matrice injected to this layer is given by $\Delta W = AB$, where $A \in \mathbb{R}^{m\times r}, B \in \mathbb{R}^{r\times n}$, such that the rank of ΔW equals a small $r \ll \min(m, n)$. After injection, the weight of this layer becomes $W + \Delta W$.

LoRA Training. In fine-tuning, the original DM weights are frozen, and only the injected LoRA is trained. We denote the LoRA matrices as θ^l, and the weight of DM after LoRA injection as $\theta \oplus \theta^l$. Given a downstream dataset $\mathcal{D}$ where each sample $(\mathbf{z}_0, P)$ comprises of an image $\mathbf{z}_0$ and its prompt description P, the objective of fine-tuning is given below:

$$\min_{\theta^l} \sum_{(\mathbf{z}_0,P)\in\mathcal{D}} \mathcal{L}(\theta \oplus \theta^l, \mathbf{z}_0, P). \quad (2)$$

LoRA in Image Editing. This is a special case of the above fine-tuning, where $\mathcal{D}$ contains only the image I for editing and its prompt P (*e.g.*, "a boy"), *i.e.*, we learn a LoRA θ^l to reconstruct I. To generate an edited image, one can run this reverse process parameterized by $\theta \oplus \theta^l$ with a modified prompt P' (*e.g.*, "a smiling boy"). In particular, we can visualize the effect of LoRA guidance (by ΔW) at different ranges of time-steps. In Figure 4, we compare the image generated by original SD (I_0), by injecting LoRA at a subset of time-steps (I_1, I_2, I_3) and at all time-steps (I_4). It is clear that the guidance controls a more fine-grained attribute at a smaller time-step, *e.g.*, I_3 (guiding large time-steps) mainly retains the overall background of I, while I_1 (guiding small ones) mainly alters the local face attributes of I_0.

LoRA Fusion. Without loss of generality, one can fuse two LoRAs trained on different datasets by a tuple of tunable strength coefficients (β_1, β_2). After injecting the fused LoRAs, the weight of a DM layer becomes $W + \beta_1\Delta W_1 + \beta_2\Delta W_2$, where ΔW_i and β_i denote the corresponding low-rank matrix in the i-th LoRA and its coefficient, respectively. However, the current way of fusing LoRAs does not have the required precision to tackle the customized image editing task (Figure 2(a)).

TimeFusion

We aim to tackle the customized image editing task: given a source image I^s and its text prompt P^s, a reference image I^r containing a subject described by a prompt P^r, the goal is to fuse the subject in I^r with I^s according to a target prompt P', while retaining their fidelity.

Our TimeFusion is an extension to LoRA fusion, where the fusion coefficient additionally depends on diffusion time-step and spatial location in the feature map. It consists of three steps: 1) learn a LoRA to reconstruct I^s and I^r, respectively; 2) initialize time-step-specific coefficients and patch-specific coefficients for LoRA fusion; 3) learn the coefficients to retain the reconstruction capability of each individual LoRA after fusing them. We detail each step below:

Step 1. We aim to learn a LoRA θ^s to reconstruct I^s, and a LoRA θ^r to reconstruct the subject in I^r. For pre-processing, we use SAM (Kirillov et al. 2023) to get the subject mask in I^r by asking the user to click the subject location based on P^r. After getting the subject mask, we train the LoRAs by:

$$\min_{\theta^s} \mathcal{L}(\theta\oplus\theta^s, \mathbf{z}_0^s, P^s), \qquad \min_{\theta^r,[\mathrm{V}]} \mathcal{L}(\theta\oplus\theta^r, \hat{\mathbf{z}}_0^r, \text{“[V]”}), \quad (3)$$

where θ is the pre-trained weight of SD, $\mathbf{z}_0^s$ denotes the image latent of I^s, $\hat{\mathbf{z}}_0^r$ denotes the image latent of I^r after applying the subject mask (*i.e.*, $\mathcal{L}$ is only evaluated inside the mask), and [V] denotes a learnable token embedding following standard practice (Avrahami et al. 2023; Gal et al. 2022).

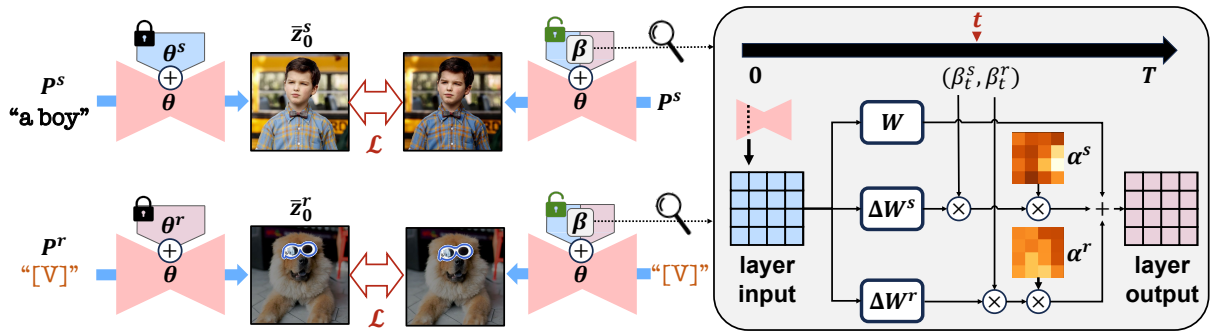

Figure 5: The overall pipeline of learning the time-step-specific and patch-specific coefficients (Steps 2 and 3). We highlight the reference object (segmented by SAM) with a blue border. The lock and unlock icon denotes frozen and trainable parameters, respectively. On the right, we detail our LoRA fusion strategy on an example layer in the SD U-Net, where $\times$ denotes the element-wise product. We omit the channel dimensions in layer input and output for simplicity.

Step 2. For *time-step-specific coefficients*, we define β_t^s, β_t^r for each $t \in \{1, \dots, T\}$. In practice, we find it unnecessary to learn a unique coefficient at each t. Instead, we sequentially group the time-steps into K splits of equal size (*e.g.*, the first split being $1, \dots, T/K$), and share the coefficient value inside each split. For *patch-specific coefficients*, we define $\boldsymbol{\alpha}^s, \boldsymbol{\alpha}^r \in \mathbb{R}^{8\times8}$ for each layer with LoRA injection (layer index omitted for simplicity). Note that the spatial dimension of the feature map at different layers in SD can be different, *i.e.*, ranging from 8×8 to 64×64. To keep our coefficients simple, we fix the number of patches to 8×8 by varying the patch size at different layers (*e.g.*, 8 when the spatial dimension is 64×64). Overall, at time-step t, the LoRA fusion coefficient for a patch at location (i, j) (out of the 8×8 patches) is $(\boldsymbol{\alpha}_{i,j}^s \beta_t^s, \boldsymbol{\alpha}_{i,j}^r \beta_t^r)$, as shown in Figure 5 right. We denote the set of all coefficients as $\boldsymbol{\beta}$, and the DM weight after injecting the fused LoRA as $\theta \oplus \theta^{\boldsymbol{\beta}}$. We initialize all coefficients in $\boldsymbol{\beta}$ as 1 and then train them.

Step 3. As illustrated in Figure 5 left, we learn $\boldsymbol{\beta}$ by the following objective:

$$\min_{\boldsymbol{\beta}} \mathcal{L}\left(\theta \oplus \theta^{\boldsymbol{\beta}}, \bar{\mathbf{z}}_0^s, P^s\right) + \mathcal{L}\left(\theta \oplus \theta^{\boldsymbol{\beta}}, \bar{\mathbf{z}}_0^r, \text{"[V]"}\right), \quad (4)$$

where [V] is the token embedding learned in Step 1, $\bar{\mathbf{z}}_0^s, \bar{\mathbf{z}}_0^r$ denotes the reconstructed image latent by LoRA θ^s and θ^r, respectively. Overall, we are training the fusion weight, such that the fused LoRAs can accommodate all visual attributes of I^s and I^r to accurately reconstruct their latents.

Generating Edited Image. After training, we replace the subject name in the target prompt P' by the learned token [V] (*e.g.*, "a boy wearing sunglasses"→"a boy wearing [V]" for subject "sunglasses"). Then we run the reverse process parameterized by $\theta \oplus \theta^{\boldsymbol{\beta}}$ given the modified P' to get the edited image. We highlight two points: 1) This scheme can be extended to do object replacement (*e.g.*, replacing a mug with "can" by P' = "[V]" with [V] being the learned token for "can") or scene replacement (*e.g.*, changing a "dog" background by P' = "a dog at [V]" with [V] being the learned token for a background). 2) Once the LoRAs and their fusion coefficients are learned, they can directly generalize to different editing prompts, leveraging the prior knowledge of SD.

Experiment

Implementation Details. Following prior editing works (Kawar et al. 2023; Song et al. 2024; Zhang et al. 2023), we adopt Stable Diffusion (Rombach et al. 2022) as our DM. When learning LoRA θ^s to reconstruct I^s, the learning rate is set as 1e-4 and the optimization iteration is 800. We use the approach in (Avrahami et al. 2023) to learn the subject in a single image I^r. For time-step-specific coefficients, we equally divide the time-steps into 20 splits, *i.e.*, $K = 20$. When learning with Equation 4, we set the learning rate as $5e^{-2}$ and training iterations as 100. Experiments are conducted on an NVIDIA A100 GPU with a batch size of 1.

Dataset. To evaluate the effectiveness of our TimeFusion in handling various objects, we collected images separately as source and reference ones from the widely used website, *i.e.*, Unsplash (https://unsplash.com/). In particular, the number of source images is 18 and the main subjects include humans, animals, and objects. For the reference images, there are objects, animals, and scenes totaling 20 images. During the fusion of source and reference images, each source image is paired with 2-3 different reference images. Finally, we obtain 50 source-reference pairs with corresponding prompts for customized image editing, including 20, 18, and 12 samples for object addition, object replacement, and scene replacement, respectively.

Qualitative Evaluation

Our Results. Our TimeFusion supports both addition and replacement editing. Especially, the replacement involves object and scene replacement. The qualitative results are shown in Figure 3. Overall, our generated images largely preserve the fidelity of both the source images and the reference subjects while achieving high alignment with target prompts, demonstrating the superiority of TimeFusion's fusion and editing capability. For example, we could make a baby holding a spoon and replace a bottle with a teapot.

Comparison with SOTAs. The closest work SpecRef (Chen and Huang 2023) is based on DDIM inversion. Yet it falls short in customized image editing due to the error-prone inversion process. AnyDoor (Chen et al. 2023) adopts a more

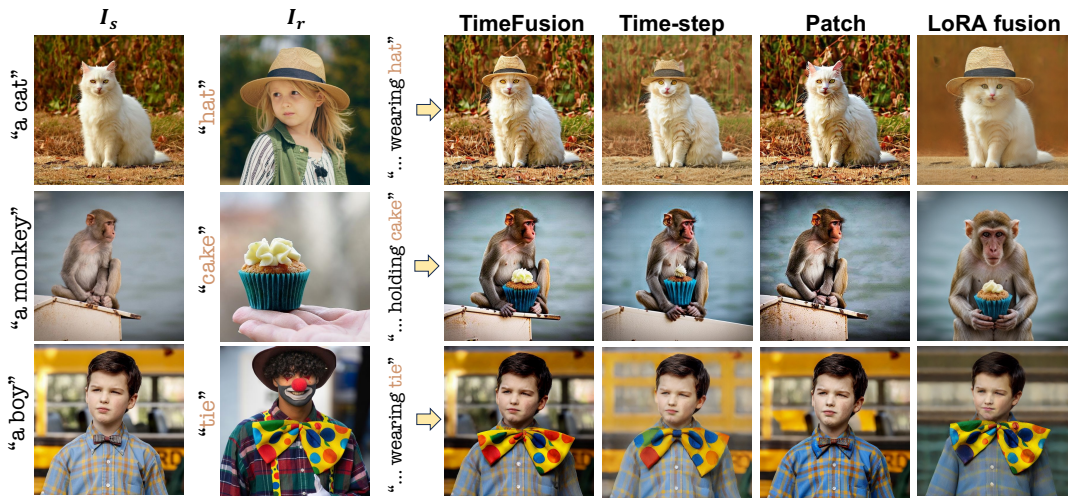

Figure 6: Qualitative comparison by ablating the use of fusion coefficients. "Time-step" and "Patch" denote using only the corresponding coefficients. "LoRA fusion" denotes its standard approach with a single set of coefficients.

	CLIP-T ↑	CLIP-I ↑	DINO ↑
SpecRef	0.266	0.694	0.778
AnyDoor	0.288	0.780	0.780
TimeFusion	0.316	0.828	0.806

Table 1: Quantitative evaluation of our TimeFusion against SOTAs. See main text for metric explanation.

Method	β_t	$\boldsymbol{\alpha}$	CLIP-T ↑	CLIP-I ↑	DINO ↑
Time-step only	✓		0.316	0.820	0.795
Patch only		✓	0.297	0.871	0.768
LoRA fusion			0.321	0.797	0.788
TimeFusion	✓	✓	0.316	0.828	0.806

Table 2: Ablation of using only time-step-specific coefficients, only patch-specific ones, and standard LoRA fusion with a single set of coefficients.

restrictive setting. It additionally requires the user to provide the editing mask in the source image for inserting the reference subject. Even under this easier task, we observe that it often loses the reference subject fidelity (*e.g.*, the sunglasses, the hat, or the can in the first row of Figure 3). Furthermore, some fused subjects look unnatural in the source image (*e.g.*, the clock or the cat in the last row of Figure 3). We also highlight a limitation of AnyDoor's setting. While a user-provided editing mask reduces the complexity of the task, it may accidentally lead to inferior results in some cases, *e.g.*, changing a cup to a can in the first row of Figure 3 will require editing the reflection on the table, but this can be easily overlooked by the user when providing the mask, causing the reflection unedited.

Quantitative Evaluation

Similarity Metrics. Since customized image editing requires image-level alignment with the source image and subject-level alignment with the reference subject, we consider the metrics in both concept customization (Gu et al. 2024; Shah et al. 2023) and image editing (Kawar et al. 2023; Song et al. 2024): 1) text alignment (CLIP-T) which measures the CLIP similarity between the target prompt and the edited image; 2) source alignment (CLIP-I) which measures the CLIP similarity between the source image and edited image; 3) reference alignment (DINO), which computes the cosine similarity between ViT S/16 DINO embeddings (Caron et al. 2021) of reference and edited images. It is worth noting that as SD generates an image that matches the CLIP text embedding of the given prompt, the direct text-to-image results from SD (with no control) will have the highest CLIP-T score. Moreover, making no edits to the source image will lead to the highest CLIP-I score. However, none of these two situations are expected in customized image editing. Therefore, the key is to obtain a balanced CLIP-I and CLIP-T score, instead of having a high score in only one of them. The results are summarized in Table 1, where our TimeFusion achieves the best scores across the three metrics compared with SpecRef and AnyDoor.

User Study. We further evaluate our proposed TimeFusion through an extensive human perceptual evaluation study. It consists of 50 source-reference pairs with corresponding prompts. 51 AMT evaluators participated in this study to rate the editing quality of the 50 samples. Each sample includes a source image, a reference image containing a specified subject, a target prompt, and three edited images generated by TimeFusion, AnyDoor, and SpecRef, which are randomly shuffled. They are required to choose the best result among the three. Finally, we collected 2,550 answers. The results show that TimeFusion, AnyDoor, and SpecRef separately receive 89.6%, 10.2%, and 0.2% of the preferences, highlighting the superiority of TimeFusion.

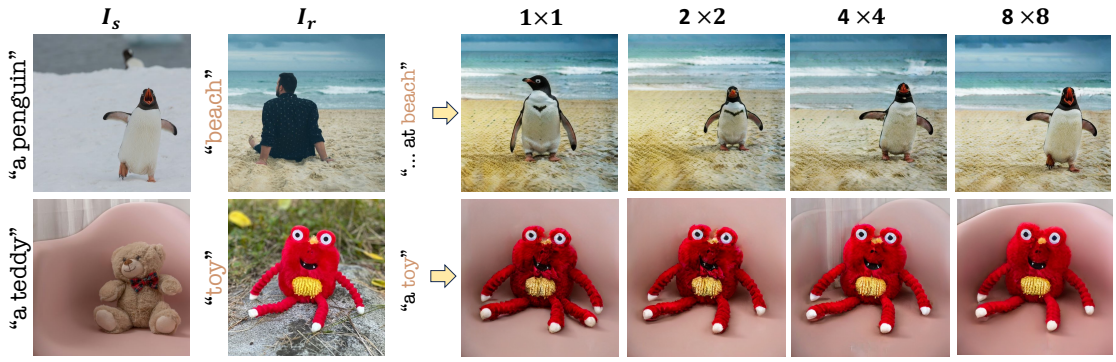

Figure 7: Ablation on the size of patch-level coefficients. TimeFusion uses 8×8.

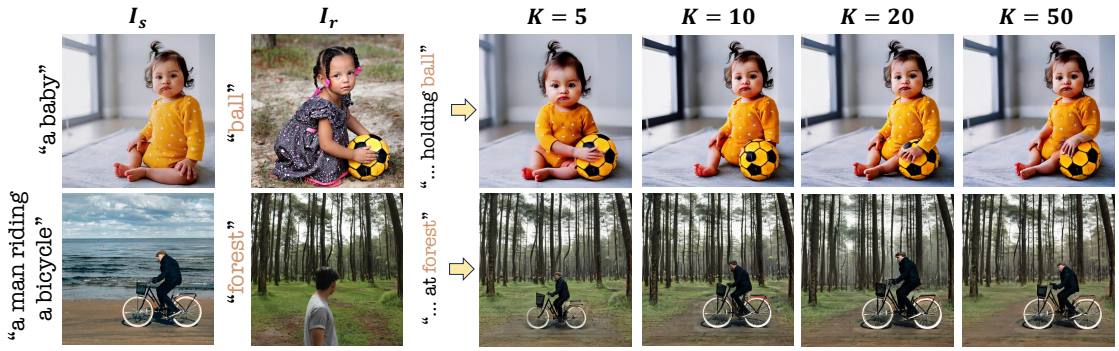

Figure 8: Ablation on the number of splits K for the time-step-specific coefficients.

Ablation Analysis

LoRA Coefficients. We ablate the use of time-step-specific coefficient and patch-specific one in Figure 6 and Table 2. The "LoRA fusion" coefficients are obtained by grid-search, and the rest are learned by Equation 4. With only time-step coefficients, we observe a loss of spatial information, *e.g.*, the backgrounds of the cat and boy are blurry. Only using patch coefficients barely makes any edit in Figure 6, hence it is the best in CLIP-I but the worst in the other two metrics in Table 2. The LoRA fusion has little control (only a single set of coefficients) of the edited image, hence it has the highest CLIP-T score, yet alters the source image significantly (lowest CLIP-I score). Finally, since time-step coefficient β_t provides the fusion precision and patch coefficient $\boldsymbol{\alpha}$ refines spatial information, TimeFusion equipped with both coefficients gains a balanced score and achieves the best results.

Patch-level Coefficient. We examine the effect of the patch-level coefficient's size in Figure 7. It could be observed that a small size provides only coarse spatial information, hence it fails to maintain the fidelity of either the source image or reference subject, *e.g.*, the posture and position of the penguin and the appearance of the toy are not well preserved. As the size of the patch-level coefficient increases from 1×1 to 8×8, the edited images progressively achieve better fidelity.

Time-step Coefficient. We ablate the number of splits K for the time-step coefficient. In Figure 8, as K increases from 5 to 20, we observe consistent improvements in fidelity, *e.g.*, the baby's posture at $K = 5$ and the baby's leg at $K = 10$ are not preserved. This is because when K is small, a long range of time-steps shares the same time-step coefficients, yet corresponds to multiple visual attributes. Hence the fusion may be imprecise to lose fidelity. However, further improving K to 50 leads to a slight loss of fidelity, *e.g.*, the shade of the baby's clothes. We conjecture that the reason is inadequate learning for coefficients: as the optimization iteration of fusion is fixed to 100, each time-step coefficient is only expected to be trained twice when $K = 50$. Yet increasing the iteration increases the training time. Hence we choose $K = 20$ to balance the editing results and compute.

Conclusion

We presented TimeFusion, which enables customized image editing with the text-conditioned Stable Diffusion (SD). The crux is to first learn two SD LoRAs for encoding the visual attributes of the source image and reference subject, respectively, and then fuse them in a precise way to maintain the fidelity of both. Motivated by the connection between diffusion time-step and visual attributes, we extend the current LoRA fusion with learnable time-step-specific and patch-specific coefficients. We show that the additional coefficients enable LoRA fusion to simultaneously accommodate the visual attributes of the user-provided source and reference images. Hence we significantly improve the fidelity in customized image editing, compared to other methods.

Acknowledgments

This research was supported by NSFC project (No. 62232006) and by National Research Foundation, Singapore, NRF-NRFI10-2024-0004.

References

Avrahami, O.; Aberman, K.; Fried, O.; Cohen-Or, D.; and Lischinski, D. 2023. Break-a-scene: Extracting multiple concepts from a single image. In *SIGGRAPH Asia 2023 Conference Papers*, 1–12.

Cao, M.; Wang, X.; Qi, Z.; Shan, Y.; Qie, X.; and Zheng, Y. 2023. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 22560–22570.

Caron, M.; Touvron, H.; Misra, I.; Jégou, H.; Mairal, J.; Bojanowski, P.; and Joulin, A. 2021. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, 9650–9660.

Chen, S.; and Huang, J. 2023. SpecRef: A Fast Training-free Baseline of Specific Reference-Condition Real Image Editing. In *2023 International Conference on Image Processing, Computer Vision and Machine Learning (ICICML)*, 369–375. IEEE.

Chen, X.; Huang, L.; Liu, Y.; Shen, Y.; Zhao, D.; and Zhao, H. 2023. Anydoor: Zero-shot object-level image customization. *arXiv preprint arXiv:2307.09481*.

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*.

Gu, Y.; Wang, X.; Wu, J. Z.; Shi, Y.; Chen, Y.; Fan, Z.; Xiao, W.; Zhao, R.; Chang, S.; Wu, W.; et al. 2024. Mix-of-show: Decentralized low-rank adaptation for multi-concept customization of diffusion models. *Advances in Neural Information Processing Systems*, 36.

Hertz, A.; Aberman, K.; and Cohen-Or, D. 2023. Delta denoising score. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2328–2337.

Hertz, A.; Mokady, R.; Tenenbaum, J.; Aberman, K.; Pritch, Y.; and Cohen-Or, D. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*.

Ho, J.; Jain, A.; and Abbeel, P. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33: 6840–6851.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Kawar, B.; Zada, S.; Lang, O.; Tov, O.; Chang, H.; Dekel, T.; Mosseri, I.; and Irani, M. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6007–6017.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; Dollár, P.; and Girshick, R. 2023. Segment Anything. *arXiv:2304.02643*.

Orgad, H.; Kawar, B.; and Belinkov, Y. 2023. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7053–7061.

Pan, Z.; Gherardi, R.; Xie, X.; and Huang, S. 2023. Effective real image editing with accelerated iterative diffusion inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15912–15921.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. GLaMM: Pixel Grounding Large Multimodal Model. *The IEEE/CVF Conference on Computer Vision and Pattern Recognition*.

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.

Ryu, S. 2023. Merging loras. https://github.com/cloneofsimo/lora.

Shah, V.; Ruiz, N.; Cole, F.; Lu, E.; Lazebnik, S.; Li, Y.; and Jampani, V. 2023. Ziplora: Any subject in any style by effectively merging loras. *arXiv preprint arXiv:2311.13600*.

Song, J.; Meng, C.; and Ermon, S. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*.

Song, X.; Cui, J.; Zhang, H.; Chen, J.; Hong, R.; and Jiang, Y.-G. 2024. Doubly Abductive Counterfactual Inference for Text-based Image Editing. *arXiv preprint arXiv:2403.02981*.

Tumanyan, N.; Geyer, M.; Bagon, S.; and Dekel, T. 2023. Plug-and-play diffusion features for text-driven image-to-image translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1921–1930.

Wallace, B.; Gokul, A.; and Naik, N. 2023. Edict: Exact diffusion inversion via coupled transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22532–22541.

Wu, C. H.; and De la Torre, F. 2023. A latent space of stochastic diffusion models for zero-shot image editing and guidance. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 7378–7387.

Yue, Z.; Wang, J.; Sun, Q.; Ji, L.; Chang, E. I.; Zhang, H.; et al. 2024a. Exploring Diffusion Time-steps for Unsupervised Representation Learning. *arXiv preprint arXiv:2401.11430*.

Yue, Z.; Zhou, P.; Hong, R.; Zhang, H.; and Sun, Q. 2024b. Few-shot Learner Parameterization by Diffusion Time-steps. *arXiv preprint arXiv:2403.02649*.

Zhang, Z.; Han, L.; Ghosh, A.; Metaxas, D. N.; and Ren, J. 2023. Sine: Single image editing with text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6027–6037.