One-Shot Identity Swapping via Attention-Guided Diffusion without Training

Yunseo Kyung Yoonsang Oh Jaehee Won Sungmin Jeon

Seounl National University
{kys2148, 5yoondori, wonjaehe_e, jsm5792}@snu.ac.kr

Abstract

In this project, we propose a novel image interpolation and composition framework that enables object identity extraction and transfer between two real images. Specifically, our method integrates multiple diffusion-based techniques by (i) applying Textual Inversion to image A to learn a compact text embedding that represents the target object's identity, (ii) performing Null-Text Inversion (NTI) on image B to recover an accurate reverse diffusion trajectory, and (iii) injecting the learned text embedding from A into the NTI-guided generation process of B, thereby synthesizing the object from A naturally within the contextual scene of B. Our approach enables object replacement and identity interpolation across real images using text conditioning alone, without requiring explicit spatial masks or additional model fine-tuning.

1. Introduction

Recent diffusion-based generative models have significantly advanced image synthesis, evolving from DDPM to DDIM and further to Latent Diffusion Models (LDMs), thereby achieving high-resolution, high-fidelity image generation with strong text-conditioned control. Along this line of progress, text-guided image editing has rapidly developed. However, seamlessly combining two real images or replacing object identities without explicit spatial masks or model fine-tuning remains an open challenge. Most existing generation and editing methods primarily focus on single-image manipulation, or rely on additional training procedures or carefully designed masks to support multi-image composition.

To address these limitations, we propose a new framework that integrates NTI-based inversion with attention steering and text embedding optimization to enable the natural fusion of two real images using prompt manipulation alone. Our approach represents the object identity of image A as a learned text embedding and injects it into the reverse diffusion process of image B. As a result, we can synthesize images that preserve the scene structure of B while transferring the object characteristics of A, without requiring explicit masking or model fine-tuning. For our code, visit our page.[1]

2. Related Work

Diffusion Models. Diffusion models are probabilistic generative models that synthesize images through a gradual denoising process. Ho et al. (2020) introduced Denoising Diffusion Probabilistic Models (DDPM), establishing the foundational framework for diffusion-based generation. Building on this, Song et al. (2021) proposed Denoising Diffusion Implicit Models (DDIM), which enable deterministic sampling, improving inference efficiency and enabling inversion. In addition, classifier-free guidance (CFG) by Ho and Salimans has become a core technique for controlling the strength of text conditioning in diffusion models. Based on these advances, Rombach et al. (2022) proposed Latent Diffusion Models (LDMs), which perform diffusion in a learned latent space, achieving both high image quality and computational efficiency. Stable Diffusion, built upon the LDM architecture, has since become a de facto standard for a wide range of image generation and editing tasks.

Latent-Space Image Interpolation. In parallel, prior works have explored image interpolation directly in the latent space of diffusion models. One representative approach is DID (Denoise–Interpolate–Denoise), which iteratively performs interpolation and denoising between the latent representations of two input images across diffusion timesteps (Wang & Golland, 2023). This recursive generation process allows the synthesized results to remain visually close to each input image, producing smooth intermediate transitions. However, because DID primarily relies on linear blending in latent space, it lacks fine-grained control over semantic attributes. As a result,

[1] *https://github.com/jsm5792/DL-project/tree/main*.

performing targeted image editing or compositional manipulation remains challenging compared to models explicitly designed for image editing.

Diffusion-Based Image Composition. Another active line of research focuses on image composition using diffusion models. Representative methods include Paint by Example (Yang et al., 2022) and ControlCom (Zhang et al., 2023), which enable the synthesis of composite images by conditioning on reference images. Despite their effectiveness, these approaches typically require users to provide explicit spatial masks to specify target regions, which limits usability and flexibility in practical scenarios.

Text-Guided Image Editing. Text-guided image editing has rapidly expanded with the emergence of large-scale text–image pretrained models. Early work such as GLIDE (Nichol et al., 2022) demonstrated the feasibility of diffusion-based editing via text-conditioned inpainting. Subsequent approaches have largely focused on global editing through prompt modification; however, they often struggle to preserve the structural consistency of real images. Prompt-to-Prompt (P2P) editing (Hertz et al., 2023) addressed this limitation by exploiting the cross-attention layers of diffusion models, which establish correspondences between textual tokens and spatial image regions. By controlling attention associated with specific words in the prompt, P2P enables precise local edits using text alone.

Inversion for Real Image Editing. DDIM inversion leverages the reversibility of deterministic sampling to map images back to latent noise trajectories. However, when applied to text-conditioned models with classifier-free guidance, errors accumulate and limit reconstruction fidelity. Null-Text Inversion (NTI) by Mokady et al. (2023) resolves this issue by optimizing the null-text embeddings at each diffusion step, enabling training-free and high-fidelity inversion of real images. NTI further allows Prompt-to-Prompt editing to be robustly applied to real images.

Personalization and Identity Representation. Another line of research explores personalizing text-to-image models for specific users or objects. Textual Inversion (Gal et al., 2022) proposes a lightweight personalization method that learns a pseudo-token embedding representing a specific object or style from a small set of images. In contrast, DreamBooth (Ruiz et al., 2022) fine-tunes the entire model to strongly inject object identity, achieving high fidelity at the cost of substantial training overhead and potential catastrophic forgetting. Imagic (Kawar et al.) introduces an optimization-based framework for real image editing, but requires direct optimization of model parameters or latent variables and is therefore not training-free.

Limitations of Textual Inversion for Interpolation. Textual Inversion has been widely adopted due to its efficiency, particularly for concept recreation and text-guided scene composition. Prior work has combined DID with techniques such as textual inversion and pose guidance to achieve more natural latent-space interpolation results (Wang & Golland, 2023). Nevertheless, since DID fundamentally generates intermediate images through latent blending, even with textual inversion, it remains significantly more difficult to achieve precise, prompt-driven editing compared to NTI-based methods. Furthermore, such approaches often suffer from limited reconstruction fidelity to the original input images.

Our Contribution. Within this context, we (1) reinterpret Textual Inversion not as generic concept learning, but as lightweight embedding optimization for extracting object identity from one or a few images, (2) inject the learned embedding directly into the NTI-based inversion trajectory of a real image, and (3) combine this process with cross-attention steering to enable mask-free, training-free, and prompt-driven fusion of two real images. Unlike prior works that treat Textual Inversion, NTI, and P2P as independent techniques, our approach unifies them into a single pipeline, enabling a novel application: cross-image identity fusion.

3. Background

In this section, we review the core concepts of diffusion models and image editing techniques that form the foundation of our approach. In particular, DDPM and DDIM, Latent Diffusion Models, Classifier-Free Guidance, Null-Text Inversion, and Prompt-to-Prompt editing constitute the key methodological components of this work.

3.1. DDPM and DDIM

Diffusion models are probabilistic generative models that combine a forward diffusion process, which gradually maps a high-dimensional data distribution $q(x_0)$ into a simple Gaussian distribution, with a reverse denoising process that recovers data samples from noise.

Ho et al. (2020) formalized this framework by defining the forward process as a fixed Markov chain governed by Gaussian transitions parameterized by a predefined variance schedule.

$$q(x_t \mid x_{t-1}) = \mathcal{N}\left(x_t ; \sqrt{1-\beta_t}\, x_{t-1},\, \beta_t \mathbf{I}\right) \quad (1)$$

parameterized by a variance schedule $\beta_0, \ldots, \beta_T \in (0,1)$

By accumulating this process, a closed-form expression can be derived that directly relates a noisy sample at an arbitrary timestep to the original data sample.

$$x_t = \sqrt{\bar{\alpha}_t} x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon,\ \epsilon \sim N(0,I) \qquad (2)$$
$$\bar{\alpha}_t = \prod_{i=1}^{t}(1-\beta_i)$$

Starting from pure Gaussian noise $x_T \sim \mathcal{N}(0,I)$, the reverse process is assumed to follow Gaussian transitions that progressively denoise the sample.

$$x_{t-1} = \mu_\theta(x_t,t) + \sigma_t z, \quad z \sim N(0,I) \qquad (3)$$

In practice, a convolutional neural network based on the U-Net architecture is trained to predict the noise $\epsilon_\theta(x_t,\ t)$ added during the forward process, and the predicted noise is then used to estimate the mean of the reverse transition.

$$\mu_\theta(x_t,\ t) = \frac{1}{\sqrt{\alpha_t}}\left(x_t - \frac{\beta_t}{\sqrt{1-\bar{\alpha}_t}}\,\epsilon_\theta(x_t,\ t)\right) \qquad (4)$$

DDIM extends this formulation by observing that the stochastic reverse process admits a family of alternative reverse processes that share the same marginals $q(x_t|\ x_0)$. By selecting a special case that removes stochasticity by setting $\sigma_t = 0$, DDIM enables deterministic sampling, ensuring that the same initial noise always leads to the same generated image. This property significantly improves inference efficiency and makes inversion feasible.

$$x_{t-1} =$$
$$\sqrt{\alpha_{t-1}}\left(\frac{x_t - \sqrt{1-\alpha_t}\,\epsilon_\theta(x_t,t)}{\sqrt{\alpha_t}}\right) + \sqrt{1-\alpha_{t-1}}\,\epsilon_\theta(x_t,t) \qquad (5)$$

3.2. Latent Diffusion Model and Stable Diffusion

Make Latent Diffusion Models (LDMs) perform the diffusion process not in pixel space, but in the latent space of a pretrained autoencoder, enabling substantially more efficient high-resolution image generation. Rather than preserving exact pixel-level details, the latent representation retains perceptually important information such as structure, style, color, and shape in a compressed form. Standard DDPM or DDIM processes are then applied in this latent space.

Stable Diffusion is a high-performance text-conditional image generation model built upon the LDM architecture. It incorporates multiple cross-attention modules within the U-Net-based noise prediction network, allowing textual conditions to directly influence the image generation process. Token embeddings extracted by a CLIP-based text encoder (Radford et al., 2021) are injected into intermediate feature maps via cross-attention, enabling fine-grained alignment between text and image. Editing techniques such as NTI and Prompt-to-Prompt are also built upon this Stable Diffusion framework..

3.3. Classifier-Free Guidance (CFG)

Stable Diffusion employs Classifier-Free Guidance (CFG) during all sampling steps to control the strength of text conditioning. CFG plays a crucial role in enhancing semantic alignment between text and image, as well as emphasizing specific objects or stylistic attributes.

Instead of relying on an external classifier, CFG combines conditional and unconditional noise predictions produced by a single diffusion model.
During sampling, the model predicts noise both with and without text conditioning for the same noisy latent.

$$\epsilon_{\text{cond}} = \epsilon_\theta(x_t,t,c) \qquad (6)$$
$$\epsilon_{uncond} = \epsilon_\theta(x_t,t,\emptyset) \qquad (7)$$

These two predictions are linearly combined to form a guidance signal that amplifies the influence of the text condition.

$$\epsilon_{cfg} = \epsilon_{uncond} + \omega(\epsilon_{cond} - \epsilon_{uncond}) \qquad (8)$$

3.4. Prompt-to-Prompt (P2P) Cross-Attention Control

Prompt-to-Prompt (P2P) is a text-guided image editing method that enables precise, mask-free editing by directly manipulating cross-attention maps in text-conditioned diffusion models.

At each diffusion timestep, the U-Net projects image features $\phi(x_t)$ as queries Q_t and text embeddings $\psi(P)$ as keys and values (K_t, V_t). The resulting cross-attention map M_t is defined such that each entry $M_{t,ij}$represents the degree of attention assigned by the *i*-th spatial location to the *j*-th text token at timestep *t*.

$$M_t = \text{Softmax}\left(\frac{Q_t K_t^T}{\sqrt{d}}\right) \qquad (9)$$

The P2P pipeline consists of three main steps.

1. attention maps are collected for both the original prompt and the edited prompt through forward passes of the U-Net
2. these attention maps are selectively mixed according to predefined editing rules such as word replacement, prompt refinement, or attention re-weighting
3. the diffusion reverse process is executed using the modified attention maps.

By controlling attention rather than pixels, P2P allows users to modify specific regions of an image using text alone, addressing the common issue where naive prompt changes lead to global and uncontrollable image alterations. This is achieved without explicit spatial masks or additional training..

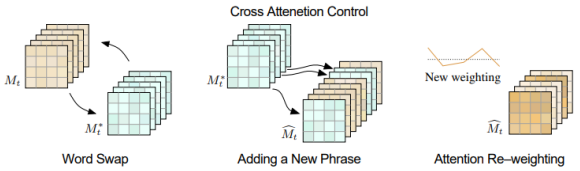

Figure 1. Different types of cross-attention map control in Prompt-to-Prompt (P2P) [9].

3.5. Null-Text Inversion (NTI)

Since real images do not lie on the native generation trajectory of diffusion models, Prompt-to-Prompt editing cannot be directly applied to them. To bridge this gap, an inversion process is required to map a real image back to a corresponding diffusion trajectory.

Owing to the deterministic characteristics of DDIM, one may naturally consider using DDIM inversion to map a generated image back to its corresponding diffusion trajectory. A single step of DDIM sampling can be inverted by reversing the direction of time, effectively increasing the timestep instead of decreasing it. However, DDIM inversion is not a structurally exact inverse of the sampling process and therefore cannot perfectly reconstruct the original image. This limitation becomes particularly pronounced in text-conditioned models such as Stable Diffusion, which rely on classifier-free guidance. When CFG is applied, the sampling process becomes inherently non-invertible, causing errors to accumulate along the inversion trajectory. As a result, DDIM inversion alone fails to faithfully reconstruct even images originally generated by the model, and the reconstruction error is further amplified when applied to real images.

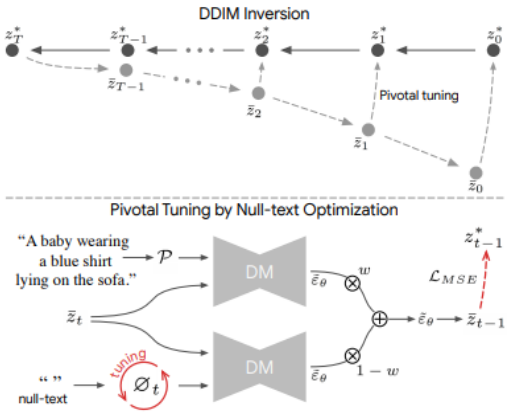

Figure 2. Illustration of DDIM inversion with tuning (top) and the NTI pipeline (bottom).

Mokady et al. (2023) addressed these challenges through pivotal inversion and Null-Text Optimization. Specifically, an initial DDIM inversion is first performed **without classifier-free guidance** by setting the guidance scale to ω = 1, thereby minimizing the influence of the text condition and obtaining a relatively stable reference trajectory. Then, during the guided reverse process, the model weights are kept fixed while the unconditional (null-text, " ") embeddings $\emptyset_t$ are optimized at each timestep so that the reconstructed latent trajectory aligns with the pivot.

$$\min_{\emptyset_t} \|z^*_{t-1} - z_{t-1}(z_t, \emptyset_t, C)\|_2^2 \qquad (10)$$

This procedure enables high-fidelity reconstruction of real images without model fine-tuning and serves as a crucial preprocessing step that allows Prompt-to-Prompt editing to be reliably applied to real images. Figure 2 provides an overview of the NTI process.

3.6. Textual Inversion

Textual Inversion is a lightweight personalization technique that extends the text embedding space of text-to-image diffusion models, allowing a specific object or style to be represented as a single pseudo-token using a small number (1–5) of user-provided images. Unlike fine-tuning–based approaches such as DreamBooth, this method does not modify the model architecture or pretrained weights; instead, it optimizes only the embedding vector corresponding to the newly introduced token in the text encoder's embedding table.

The embedding optimization is performed by minimizing the noise prediction error at a randomly sampled diffusion timestep *t*, given the latent representation $z_0 = E(\boldsymbol{x})$ of an input image $\boldsymbol{x}$:

$$||\varepsilon - \epsilon_\theta\left(z_t, t, c_\theta\left(y(S^*)\right)\right)||_2^2 \qquad (11)$$

Here, S^* denotes the newly introduced pseudo-token, and the text condition is constructed using neutral template prompts such as *"a photo of S^*"*.

The optimized embedding $v*$ functions as a single token that captures the visual characteristics of the target object through the cross-attention mechanism of the diffusion model, enabling consistent identity preservation across diverse scenes and prompts. In this work, we employ this mechanism not for conventional concept recreation, but as a means of extracting the identity of an object from a real image (image A), which is then injected into an NTI-based inversion trajectory to enable cross-image identity fusion.

Figure 3. Examples of Textual Inversion applied to personalized text-to-image generation [11].

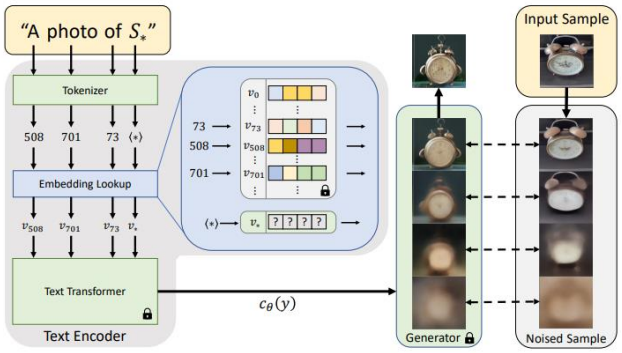

Figure 4. Text embedding and inversion process for Textual Inversion [11].

4. Method

Our method enables training-free identity swapping between two real images by integrating three key components: (1) *Textual Inversion with Attention-based Masking* to learn object identity, (2) *Null-text Inversion* to extract structural information, and (3) *Soft NTI combined with Prompt-to-Prompt control* to inject the learned identity into the preserved structure. Figure 1 provides an overview of our three-stage pipeline.

Problem Formulation. Given a source image I_A depicting an object in a particular scene layout and a reference image I_B containing a target object with distinct visual characteristics, our goal is to synthesize a new image $\hat{I}$ that preserves the structural composition (pose, background, spatial arrangement) of I_A while replacing the object's identity with that of I_B. Unlike existing approaches that require explicit spatial masks or model fine-tuning, our method operates solely through text-conditioned diffusion control, making it fully training-free.

4.1. Stage 1. Identity Learning via Masked Textual Inversion

To capture the visual identity of the target object in I_B, we adopt a modified Textual Inversion approach. Traditional Textual Inversion optimizes a pseudo-token embedding $v*$ across the entire image. However, for identity extraction from a single image, this often leads to learning irrelevant background features. To address this, we introduce attention-based automatic masking.

We first encode I_B into latent space:

$$z_B = \epsilon(I_B) \qquad (12)$$

We add noise at a fixed intermediate timestep t_mask = 400 to obtain:

$$z_{noisy} = \sqrt{\overline{\alpha}_{t_{mask}}} * z_B + \sqrt{1 - \overline{\alpha}_{t_{mask}}}\,\epsilon \qquad (13)$$

Using a neutral prompt such as "a photo of a {word}", where {word} is the object category (e.g., "cat"), we perform a single forward pass through the UNet to populate the cross-attention maps. We then extract the attention map $A \in \mathbb{R}^{16\times16}$ corresponding to the token index of {word} from the 16×16 resolution cross-attention layer. This map is upsampled via bilinear interpolation to 64×64 to match the latent resolution, normalized to [0,1], and thresholded at $\theta = 0.3$ to produce a binary mask:

$$M = \mathbb{1}_{[A>\theta]}(A) \qquad (14)$$

This mask is then used to spatially weight the diffusion loss during embedding optimization:

$$\mathcal{L} = \|\epsilon_\theta(z_t, t, c_\tau) - \epsilon\|^2 \odot M \qquad (15)$$

where $\epsilon_{pred} = \epsilon_\theta(z_t, t, c_\tau)$ is the noise predicted by the UNet conditioned on the text embedding containing the placeholder token τ, and ϵ is the ground-truth noise. By focusing the optimization exclusively on the masked object region, the learned embedding $v*$ encodes only the identity-relevant features of the target object while remaining agnostic to the background.

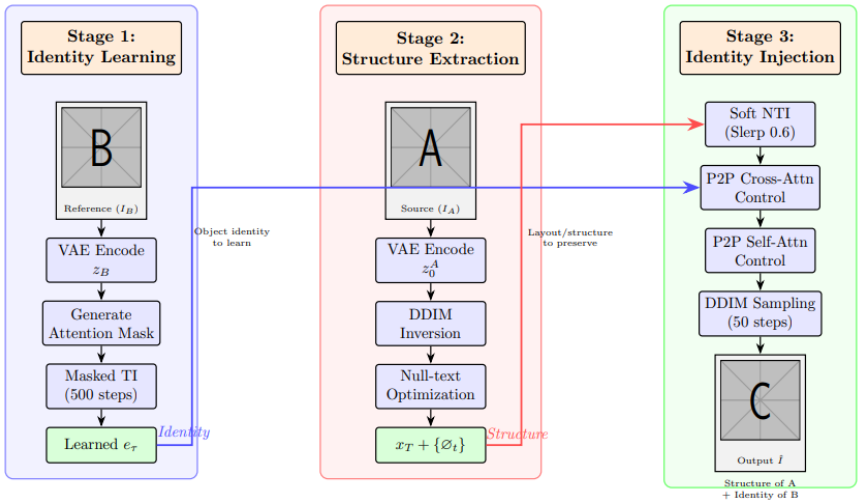

Figure 5. Overall pipeline of training-free identity swapping. **Stage 1.** learns the identity of reference object IB through masked textual inversion. **Stage 2**. extracts the structure of source image IA via null-text inversion. **Stage 3.** synthesizes the final image by injecting the learned identity into the preserved structure using Prompt-to-Prompt attention control

We train for $N_{TI} = 500$ steps with learning rate $\eta_{TI} = 1 \times 10^{-3}$ using the Adam optimizer, applying gradient masking to ensure that only the placeholder token embedding is updated.

4.2. Stage 2. Structure Extraction via Null-text Inversion

To preserve the structural composition of the source image I_A, we employ Null-text Inversion (NTI), which enables high-fidelity reconstruction of real images through optimization of unconditional embeddings. We first apply DDIM inversion without classifier-free guidance (CFG scale ω = 1) to map I_A from pixel space to a noisy latent trajectory $\{z_0^A, z_1^A, \dots, z_T^A\}$ where $z_0^A = VAE_{enc}(I_A)$.

To enable accurate reconstruction under CFG during generation, we then optimize the null-text embeddings $\emptyset_t$ at each timestep t by minimizing the reconstruction error:

$$\min_{\emptyset_t} \|z_t^A - \mathcal{F}(z_{t+1}^A, \emptyset_t, C_A)\|_2^2 \qquad (16)$$

where C_A is the text embedding for the source prompt describing I_A, and $\mathcal{F}$ denotes the DDIM backward step.

This optimization is performed for 10 inner iterations per timestep with a learning rate of 1e-5. The resulting optimized embeddings $\{\emptyset_0, \emptyset_1, \dots, \emptyset_{T-1}\}$ and inverted latent $x_T = z_T^A$ encode the complete structural information of I_A while remaining compatible with text-guided generation under CFG.

4.3. Stage 3. Identity Injection with Soft NTI and Prompt-to-Prompt Control.

Pure NTI provides perfect reconstruction but lacks the flexibility needed for semantic editing. To enable identity substitution while preserving structure, we introduce Soft NTI, which interpolates between the inverted latent x_T and random Gaussian noise ε_{rand}:

$$z_{start} = Slerp(x_T, \varepsilon_{rand}, 1 - \lambda)$$
$$= \frac{\sin((1-\lambda)\theta)}{\sin(\theta)} \cdot x_T + \frac{\sin(\lambda\theta)}{\sin(\theta)} \cdot \varepsilon_{rand} \qquad (17)$$

where Slerp denotes spherical linear interpolation and λ ∈ [0,1] controls the balance between structural preservation and editing flexibility. We empirically set λ = 0.6, which retains 60% of the original structure while allowing sufficient variation for identity transfer.

To perform the actual identity swap, we construct a target prompt p_B by replacing the object word in the source prompt p_A with the learned placeholder token. For example, if $p_A =$ "$a\ cat\ sitting\ next\ to\ a\ mirror$,then $p_B =$ "$a\ <sks-cat>\ sitting\ next\ to\ a\ mirror$".

During DDIM sampling, we apply Prompt-to-Prompt (P2P) cross-attention control to inject the identity while maintaining the spatial layout:

Cross-attention replacement. For the first 40% of denoising steps (controlled by cross_replace_steps = 0.4),

we replace the attention map of the target word in p_B with the corresponding attention map from p_A. This ensures that the object is placed in the same spatial location as in the source image.

Self-attention injection. For the first 20% of denoising steps (controlled by self_replace_steps = 0.2), we inject the self-attention features from the original DDIM trajectory of I_A. This preserves the overall scene structure and layout while allowing the diffusion process to adapt colors and textures according to the new identity.

Throughout generation, we use the optimized null-text embeddings $\{\emptyset_t\}$ from Stage 2 in the unconditional branch of CFG, ensuring that the structural information is properly preserved. The final output $\hat{I}$ is decoded from the resulting latent z_0 using the VAE decoder.

Implementation Details. We use Stable Diffusion v1.4 with the DDIM scheduler, setting the number of inference steps to 50 and classifier-free guidance scale to 7.5. All experiments are conducted on a single NVIDIA GPU(with Colab environment). The complete pipeline requires no model fine-tuning and takes approximately 2-3 minutes per image pair on an A100 GPU. Our key hyperparameters are: textual inversion steps $N_{TI} = 500$, learning rate $\eta_{TI} = 1 \times 10^{-3}$, attention mask threshold $\theta_{mask} = 0.3$, soft NTI strength $\lambda = 0.6$, cross-attention control ratio $r_{cross} = 0.4$, and self-attention control ratio $r_{self} = 0.2$.

5. Experiments

Experimental Setup. All images used in our experiments were collected from publicly available online sources, including Pinterest and general web image repositories, and were used solely for academic research purposes. Our implementation is based on the official open-source codebases released by the authors of Stable Diffusion, DDIM, and Null-Text Inversion, with minor modifications for integrating the proposed pipeline. We implement our method using Stable Diffusion v1.4 with the DDIM scheduler, conducting all experiments on a single NVIDIA A100 GPU in a Colab environment.

We evaluate our method on real-world image pairs to demonstrate its effectiveness in training-free identity swapping. Our primary test case consists of a source image I_A depicting a cat sitting next to a mirror in a bathroom setting, and a reference image I_B containing a different cat with distinct visual characteristics, including unique coat patterns, facial features, and collar design. The source image presents a particularly challenging scenario due to the presence of reflective surfaces (the mirror) and complex spatial constraints, making it an ideal test case for evaluating both identity transfer fidelity and structural preservation capabilities.

To determine optimal hyperparameters, we perform an extensive grid search over key parameters: soft NTI strength $\lambda \in \{0.6, 0.7, 0.8, 0.9\}$, cross-attention control ratio $r_{cross} \in \{0.4, 0.7, 0.9\}$, and self-attention control ratio $r_{self} \in \{0.2, 0.3\}$. This yields a total of 24 parameter combinations, each of which we evaluate through qualitative visual inspection, focusing on identity fidelity, background preservation, and the absence of structural artifacts. We note that the optimal configuration varies depending on the specific characteristics of each image pair, as different combinations of source structure complexity and reference identity features may require different balance points. For the particular test case presented in Appendix B (cat sitting next to mirror), we identify $\lambda = 0.6$, $r_{cross} = 0.4$, and $r_{self} = 0.2$ as the configuration that produces the best visual quality. For the textual inversion stage, we use $N_{TI} = 500$ training steps with learning rate $\eta_{TI} = 1 \times 10^{-3}$, and set the attention mask threshold to $\theta_{mask} = 0.3$ across all experiments.

Comparison with Baseline. We compare our approach against ControlNet (Zhang et al., 2023), a widely-adopted method for controllable image generation that has demonstrated strong performance across various conditional generation tasks. While ControlNet was originally designed for structure-preserving generation through learned conditioning adapters, it can be adapted for identity transfer by using the source image as structural guidance. We generate results using both methods under identical experimental conditions — using the same source and reference images, equivalent inference steps ($T = 50$), and comparable guidance scales — to ensure fair and meaningful comparison. It is important to note that for tasks involving identity swapping, there exists no ground truth image that perfectly combines the structure of I_A with the identity of I_B. Therefore, evaluation necessarily relies on visual quality assessment and subjective judgment of whether the method successfully achieves the intended goal.

Qualitative Results. The experimental results presented in the Appendix provide visual comparisons across three distinct test cases: cat identity swapping (cat sitting next to a mirror), car identity swapping, and snowman identity swapping. Each case presents unique challenges in terms of object complexity, background structure, and lighting conditions. Notably, the hyperparameter configuration that produces the best visual quality varies across these test cases, confirming that optimal settings are image-pair dependent.

Our method demonstrates successful identity transfer across all three scenarios. In the cat case, although the mirror reflection is unfortunately lost and the bathroom background undergoes some changes, the distinctive coat pattern, facial features, and collar details from the reference image are faithfully reproduced. In the car case, the specific vehicle model characteristics are transferred while preserving the original scene composition. The snowman case illustrates a particularly clear example: the source snowman wearing a beanie and scarf is successfully replaced with the reference snowman wearing a top hat and plaid scarf , resulting in a composite image where the new snowman identity is clearly recognizable.

Overall, our results demonstrate that training-free identity swapping is achievable with reasonable quality. The transferred identities are visually recognizable and distinct from the original objects, while the scene structure — including object pose, approximate positioning, and general environmental context — remains largely intact. For comparison, the ControlNet baseline (not shown for brevity) consistently produces more generic and averaged object appearances that fail to capture the specific visual characteristics from the reference images, suggesting that our textual inversion-based approach provides superior identity fidelity even at the cost of some structural flexibility.

6. Limitations

While our method achieves compelling identity transfer results, several limitations remain. The soft NTI approach, which interpolates noise to enable semantic editing, inevitably introduces structural deviations from the source image. Background elements may exhibit blurring or subtle alterations, and visual artifacts occasionally occur where reference image characteristics unintentionally blend with the scene. This trade-off between identity fidelity and structural preservation is inherent to our noise-based editing strategy. Additionally, extreme pose differences between source and reference objects can lead to unnatural compositions, and our current implementation does not support selective editing when multiple objects of the same category appear in the source image.

7. Conclusion

We presented a training-free approach for identity swapping between real images by creatively integrating three diffusion-based techniques: masked textual inversion for learning object identity, null-text inversion for extracting scene structure, and prompt-to-prompt attention control for spatially-aware synthesis. Our method demonstrates that effective identity transfer is achievable without model fine-tuning or explicit spatial masks, opening practical applications in interactive image editing and creative design tools. The modular nature of our pipeline suggests promising future directions, including multi-object editing, fine-grained attribute control, and integration with 3D-aware models for pose-invariant identity transfer across diverse visual domains.

References

[1] J. Ho, A. Jain, and P. Abbeel. *Denoising Diffusion Probabilistic Models.* In Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] J. Song, C. Meng, and S. Ermon. *Denoising Diffusion Implicit Models.* In International Conference on Learning Representations (ICLR), 2021.

[3] J. Ho and T. Salimans. *Classifier-Free Diffusion Guidance.* arXiv preprint arXiv:2207.12598, 2021.

[4] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. *High-Resolution Image Synthesis with Latent Diffusion Models.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022.

[5] C. J. Wang and P. Golland. *Interpolating between Images with Diffusion Models.* arXiv preprint arXiv:2307.12560, 2023.

[6] B. Yang, S. Gu, B. Zhang, T. Zhang, X. Chen, X. Sun, D. Chen, and F. Wen. Paint by Example: Exemplar-Based Image Editing With Diffusion Models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[7] B. Zhang, Y. Duan, J. Lan, Y. Hong, H. Zhu, W. Wang, and L. Niu. *ControlCom: Controllable Image Composition using Diffusion Model.* arXiv preprint arXiv:2308.10040, 2023.

[8] A. Nichol, P. Dhariwal, P. Ramesh, P. Shyam, B. Mishkin, et al. *GLIDE: Towards Photorealistic Image Generation and Editing with Text-Guided Diffusion Models.* In International Conference on Machine Learning (ICML), 2022.

[9] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or. *Prompt-to-Prompt Image Editing with Cross Attention Control.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[10] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or. *Null-text Inversion for Editing Real Images using Guided Diffusion Models.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[11] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. Bermano, G. Chechik, and D. Cohen-Or. *An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion.* arXiv preprint arXiv:2208.01618, 2022.

[12] R. Ruiz, B. Li, J. Zhang, A. Torralba, E. Shechtman, and D. Jacobs. *DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation.* In European Conference on Computer Vision (ECCV), 2022.

[13] O. Kawar, B. Dolhansky, F. Cole, D. Sun, J. Kopf, and M. Laptev. *Imagic: Text-Based Real Image Editing with Diffusion Models.* In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023.

[14] Y. Radford, J. Kim, C. Hallacy, et al. *Learning Transferable Visual Models From Natural Language Supervision.* In International Conference on Machine Learning (ICML), 2021.

Appendix

A. Algorithm

Algorithm 1 Stage 1: Identity Learning via Masked Textual Inversion

Require: Reference image I_B containing target object
Require: Target word w (e.g., "cat")
Require: Placeholder token τ (e.g., "<sks-cat>")
Require: Number of training steps N_{TI} (e.g., 500)
Require: Learning rate η (e.g., 1×10^{-3})
Ensure: Learned embedding e_τ encoding identity of target object

/* Step 1.1: Initialize Placeholder Token */
Initialize τ embedding: $e_\tau \leftarrow e_w$ ▷ Copy from target word
Add τ to tokenizer vocabulary

/* Step 1.2: Encode Reference Image */
$z_B \leftarrow \text{VAE}_{\text{enc}}(I_B)$ ▷ Encode to latent space

/* Step 1.3: Generate Attention-based Mask */
Set $t_{\text{mask}} \leftarrow 400$ ▷ Fixed timestep for stable attention
Sample $\epsilon \sim \mathcal{N}(0, I)$
Add noise: $z_{\text{noisy}} \leftarrow \sqrt{\bar{\alpha}_{t_{\text{mask}}}} z_B + \sqrt{1 - \bar{\alpha}_{t_{\text{mask}}}} \epsilon$
$c \leftarrow \text{TextEncoder}(\text{"a photo of a } w\text{"})$
$\epsilon_\theta(z_{\text{noisy}}, t_{\text{mask}}, c)$ ▷ Forward pass populates attention maps
Extract cross-attention: $A \leftarrow \text{CrossAttention}[:, :, \text{index}(w)]$
Upsample: $M \leftarrow \text{Interpolate}(A, 64 \times 64)$
Normalize: $M \leftarrow \frac{M - \min(M)}{\max(M) - \min(M)}$
Binarize: $M \leftarrow \mathbb{1}_{M > \theta_{\text{mask}}}$ ▷ $\theta_{\text{mask}} = 0.3$

/* Step 1.4: Masked Diffusion Training */
for $i = 1$ **to** N_{TI} **do**
  Sample timestep: $t \sim \mathcal{U}(1, T)$
  Sample noise: $\epsilon \sim \mathcal{N}(0, I)$
  Add noise: $z_t \leftarrow \sqrt{\bar{\alpha}_t} z_B + \sqrt{1 - \bar{\alpha}_t} \epsilon$
  Encode prompt with τ: $c_\tau \leftarrow \text{TextEncoder}(\text{"a photo of a } \tau\text{"})$
  Predict noise: $\epsilon_{\text{pred}} \leftarrow \text{UNet}(z_t, t, c_\tau)$
  Compute masked loss: $\mathcal{L} \leftarrow \text{MSE}(\epsilon_{\text{pred}}, \epsilon) \odot M$
  Compute gradient: $g \leftarrow \nabla_{e_\tau} \mathcal{L}$ ▷ Only w.r.t. e_τ
  Update embedding: $e_\tau \leftarrow e_\tau - \eta \cdot g$
end for

return Learned embedding e_τ

Algorithm 2 Stage 2: Structure Extraction via Null-text Inversion

Require: Source image I_A whose structure to preserve
Require: Source prompt p_A describing I_A
Require: Number of DDIM steps T (e.g., 50)
Require: Number of inner optimization steps N_{inner} (e.g., 10)
Require: Learning rate η (e.g., 1×10^{-5})
Ensure: Inverted latent x_T and optimized null-text embeddings $\{\varnothing_0, \ldots, \varnothing_{T-1}\}$

/* Step 2.1: Encode Source Image */
$z_0^A \leftarrow \text{VAE}_{\text{enc}}(I_A)$ ▷ Encode to latent
$c_A \leftarrow \text{TextEncoder}(p_A)$ ▷ Encode text prompt

/* Step 2.2: DDIM Inversion (Forward Process) */
for $t = 0$ **to** $T - 1$ **do**
  Predict noise: $\epsilon_t \leftarrow \text{UNet}(z_t^A, t, c_A)$
  Compute $\hat{x}_0$: $\hat{x}_0 \leftarrow \frac{z_t^A - \sqrt{1-\bar{\alpha}_t}\epsilon_t}{\sqrt{\bar{\alpha}_t}}$
  Deterministic step: $z_{t+1}^A \leftarrow \sqrt{\bar{\alpha}_{t+1}} \hat{x}_0 + \sqrt{1 - \bar{\alpha}_{t+1}} \epsilon_t$
end for
$x_T \leftarrow z_T^A$ ▷ Inverted noisy latent

/* Step 2.3: Null-text Optimization (Backward Process) */
Initialize: $\{\varnothing_0, \ldots, \varnothing_{T-1}\} \leftarrow \text{TextEncoder}(\text{""})$ ▷ Empty prompt
for $t = T - 1$ **down to** 0 **do**
  /* Optimize $\varnothing_t$ to reconstruct z_t^A */
  for $j = 1$ **to** N_{inner} **do**
    $\epsilon_{\text{uncond}} \leftarrow \text{UNet}(z_{t+1}^A, t, \varnothing_t)$
    $\epsilon_{\text{cond}} \leftarrow \text{UNet}(z_{t+1}^A, t, c_A)$
    Classifier-free guidance: $\epsilon \leftarrow \epsilon_{\text{uncond}} + \gamma(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$ ▷ $\gamma = 7.5$
    Compute $\hat{x}_0$: $\hat{x}_0 \leftarrow \frac{z_{t+1}^A - \sqrt{1-\bar{\alpha}_{t+1}}\epsilon}{\sqrt{\bar{\alpha}_{t+1}}}$
    DDIM backward: $\bar{z}_t \leftarrow \sqrt{\bar{\alpha}_t} \hat{x}_0 + \sqrt{1 - \bar{\alpha}_t} \epsilon$
    Reconstruction loss: $\mathcal{L} \leftarrow \|\bar{z}_t - z_t^A\|^2$
    Update: $\varnothing_t \leftarrow \varnothing_t - \eta \nabla_{\varnothing_t} \mathcal{L}$
  end for
end for

return Inverted latent x_T, Optimized embeddings $\{\varnothing_0, \ldots, \varnothing_{T-1}\}$

Algorithm 3 Stage 3: Identity Injection via Soft NTI + Prompt-to-Prompt

Require: Inverted latent x_T from Stage 2
Require: Optimized null-text embeddings $\{\varnothing_0, \ldots, \varnothing_{T-1}\}$ from Stage 2
Require: Learned identity embedding e_τ from Stage 1
Require: Source prompt p_A, Target word w, Placeholder token τ
Require: Soft NTI strength $\lambda \in [0, 1]$ (e.g., 0.6)
Require: Cross-attention control ratio r_{cross} (e.g., 0.4)
Require: Self-attention control ratio r_{self} (e.g., 0.2)
Require: Guidance scale γ (e.g., 7.5)
Ensure: Synthesized image $\hat{I}$ with target identity

/* Step 3.1: Prepare Target Prompt */
$p_B \leftarrow p_A.\text{replace}(w, \tau)$ ▷ e.g., "a <sks-cat> sitting next to a mirror"
$c_A \leftarrow \text{TextEncoder}(p_A)$
$c_B \leftarrow \text{TextEncoder}(p_B)$ ▷ Uses learned e_τ

/* Step 3.2: Soft Null-text Inversion */
Sample random noise: $\epsilon_{\text{rand}} \sim \mathcal{N}(0, I)$
Spherical interpolation: $z_{\text{start}} \leftarrow \text{Slerp}(x_T, \epsilon_{\text{rand}}, 1 - \lambda)$
▷ $\lambda = 0.6$ preserves 60% structure, allows 40% flexibility

/* Step 3.3: DDIM Sampling with P2P Control */
Initialize: $z_T \leftarrow z_{\text{start}}$
for $t = T - 1$ **down to** 0 **do**

  /* Cross-Attention Control (First 40% steps) */
  if $t > (1 - r_{\text{cross}}) \times T$ **then**
    Replace: $\text{Attn}_{\text{cross}}(z_t, c_B, w) \leftarrow \text{Attn}_{\text{cross}}(z_t, c_A, w)$
    ▷ Use source structure attention for target word
  end if

  /* Self-Attention Control (First 20% steps) */
  if $t > (1 - r_{\text{self}}) \times T$ **then**
    Inject: $\text{Attn}_{\text{self}}(z_t) \leftarrow \text{Attn}_{\text{self}}(z_t^A$ from Stage 2)
    ▷ Preserve spatial structure
  end if

  /* Noise Prediction with Optimized Null-text */
  $\epsilon_{\text{uncond}} \leftarrow \text{UNet}(z_t, t, \varnothing_t)$ ▷ Use optimized $\varnothing_t$
  $\epsilon_{\text{cond}} \leftarrow \text{UNet}(z_t, t, c_B)$
  Classifier-free guidance: $\epsilon \leftarrow \epsilon_{\text{uncond}} + \gamma(\epsilon_{\text{cond}} - \epsilon_{\text{uncond}})$

  /* DDIM Backward Step */
  Compute $\hat{x}_0$: $\hat{x}_0 \leftarrow \frac{z_t - \sqrt{1-\bar{\alpha}_t}\epsilon}{\sqrt{\bar{\alpha}_t}}$
  $z_{t-1} \leftarrow \sqrt{\bar{\alpha}_{t-1}} \hat{x}_0 + \sqrt{1 - \bar{\alpha}_{t-1}} \epsilon$
end for

/* Step 3.4: Decode to Image */
$\hat{I} \leftarrow \text{VAE}_{\text{dec}}(z_0)$

return Synthesized image $\hat{I}$

B. Results

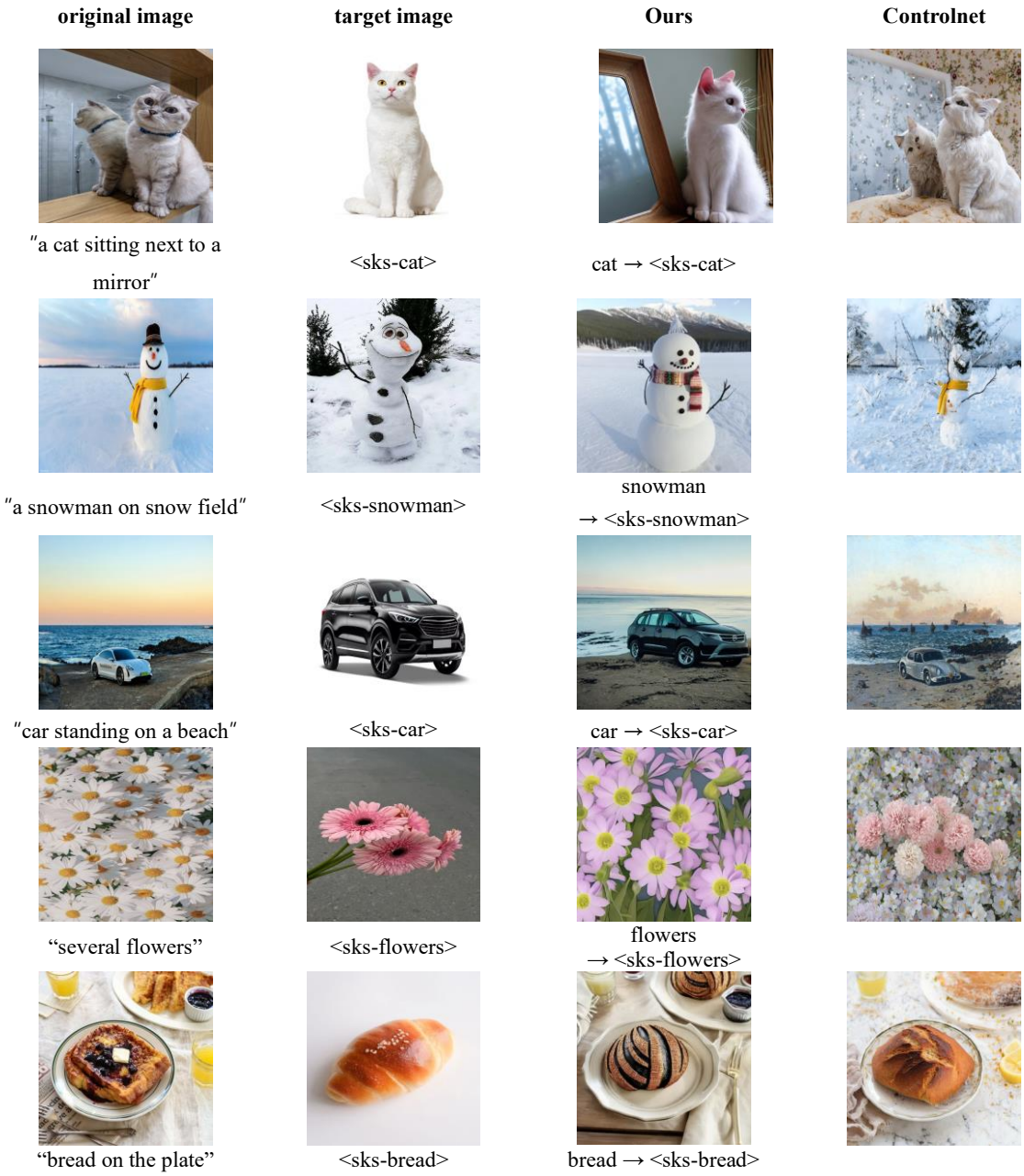

Figure 6. Qualitative comparison of identity swapping results. Each row shows, from left to right, the source image (A), the reference image providing the target identity (B), the result produced by our method, and the result generated by ControlNet. Additional results on flowers and bread are provided in the Appendix.