Supplementary Material

Experimental Details

The training process was conducted on a single machine equipped with a Quadro RTX 8000 GPU, and detailed training configurations are provided in the supplementary material. For a fair evaluation, we used the official codes and implementations provided on Hugging Face, following the default parameter settings.

For **DreamBooth** [12], we set the batch size to 1 and the learning rate to 5×10^{-6}. For **Custom Diffusion** [9], we used a batch size of 2 and a learning rate of 1×10^{-5}. For **Textual Inversion** [6], the learning rate was set to 5.0×10^{-4}, with a batch size of 1 and gradient accumulation steps of 4. For **NeTI** [1], we used a batch size of 2, a learning rate of 1×10^{-3}, and gradient accumulation steps of 4. For **AttnDreamBooth** [11], Stage 1 used a learning rate of 1×10^{-3} with a batch size of 8; Stage 2 used 2×10^{-5} with batch size 8; and Stage 3 used 2×10^{-6}. The official implementation used a batch size of 8 and gradient accumulation steps of 1, but due to memory limitations, we used a batch size of 4 and accumulation steps of 2 in Stage 3. For **DisenBooth** [4], we used a learning rate of 1×10^{-4} and a batch size of 1. For **Cones2** [10], the learning rate was set to 5×10^{-6} and the batch size to 1.

Comparison With Other Methods

Fine-grained Detail Preservation

NeTI, Textual Inversion and Custom Diffusion require less storage compared to DreamBooth, making them more efficient. However, in preserving fine-grained details, DreamBooth remains superior, as illustrated in Fig. 1. The results indicate that MINDiff enables the model to maintain DreamBooth-level subject fidelity while ensuring that the text prompt is faithfully reflected. This is evident from the appearance of the backpack, the whale graphic on the can, and the number '3' on the clock, which demonstrate the superior subject fidelity of our model.

Qualitative Results on SDXL

Figure 2 shows the results of LoRA-based [7] DreamBooth using the Stable Diffusion XL 1.0 backbone. These results demonstrate the compatibility of MINDiff with various versions of Stable Diffusion. While maintaining a similar level of image fidelity, MINDiff achieves improved text alignment.

Mask-Based Editing Comparison

Table 1 summarizes key differences between MINDiff and existing mask-based editing approaches [2, 3, 5, 8]. Prior methods typically blend a reconstructed and an edited representation using a spatial mask, following a formulation such as $A \odot (1 - m) + B \odot m$, where A denotes the reconstructed representation, B the edited representation, and m is a binary mask indicating the target region for editing.

In contrast to blending-based approaches, MINDiff introduces a structurally distinct mechanism by directly subtracting a mask-weighted suppression term—derived from an auxiliary attention branch—from the output of the main attention operation. This enables spatial control over the subject's semantic influence, thereby mitigating overfitting in personalization models.

PFB-Diff also incorporates masking within the cross-attention mechanism. However, it applies masks to the attention scores (*i.e.*, $QK^\top$) to control the appearance of specific tokens. By contrast, MINDiff modulates the full attention output (*i.e.*, $\mathrm{softmax}(QK^\top)V$) across all channels, providing a more direct and global form of suppression.

Clarifying mask application. As shown in Tab. 1, the compared methods differ in where the mask is applied. Although the table lists "Feature map" as the main mask application level for PFB-Diff, it also performs pixel-space blending in the early stages of the diffusion process. DiffEdit automatically generates a mask by comparing noise predictions under different prompts, while MINDiff derives its mask from the cross-attention map of the subject token. PFB-Diff and our method share certain similarities. Both operate across multiple levels of the architecture, allowing for more seamless integration of mask-based control.

Mask Analysis

Temporal Evolution of the Mask

To better illustrate the effect of Mask-Integrated Negative Attention Diffusion (MINDiff), we visualize the generated masks by dividing the denoising process into six stages in Fig. 3. This mask represents the pre-inversion state. When

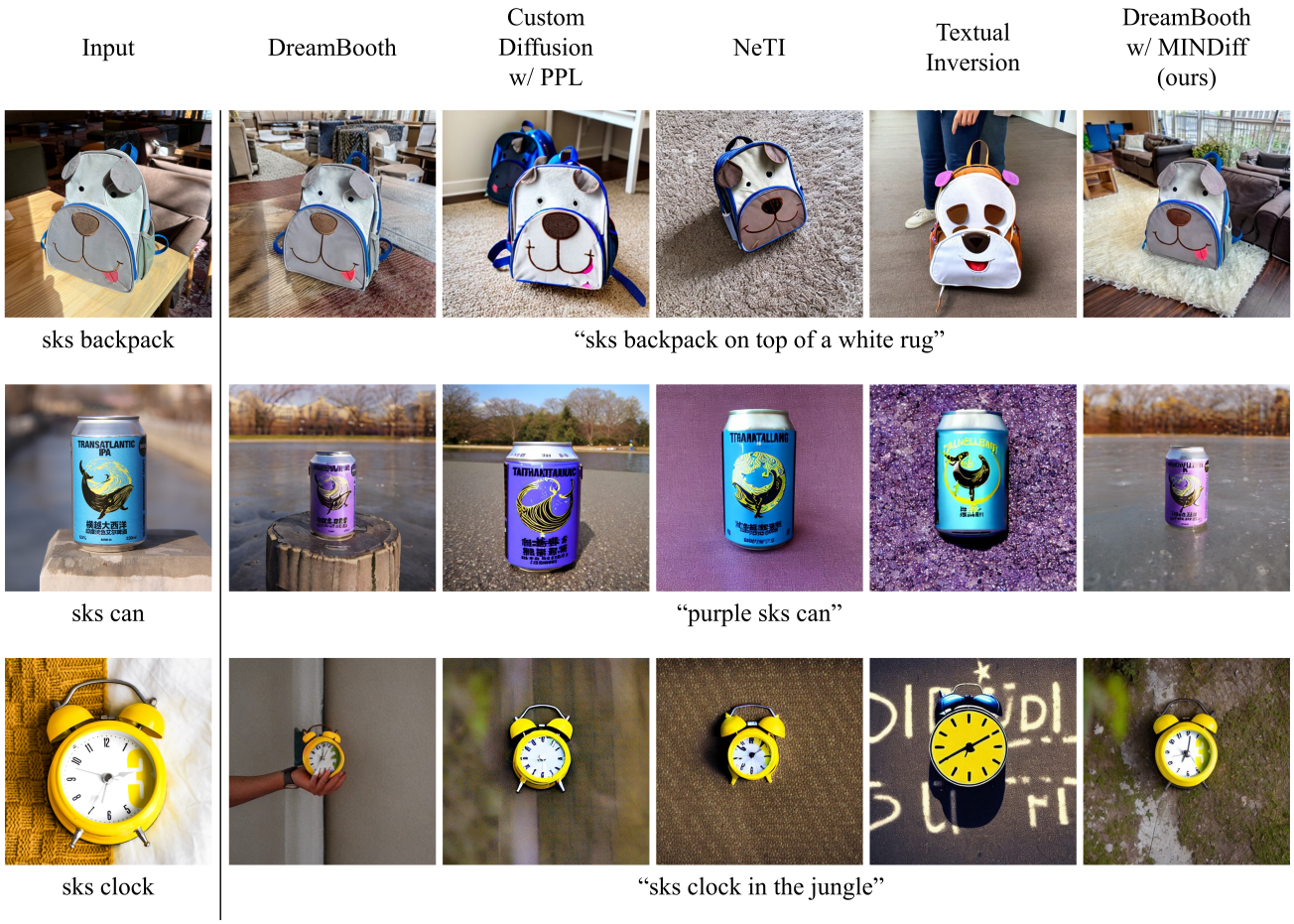

Figure 1. Qualitative comparison of fine-grained detail preservation across personalization methods: DreamBooth, Custom Diffusion with PPL, NeTI, Textual Inversion, and DreamBooth with MINDiff (ours). All images are generated using the same random seed for fair comparison. NeTI uses the highest truncation value, and MINDiff applies a suppression scale of $\lambda = 0.6$. Results show that MINDiff retains DreamBooth-level visual fidelity while improving text alignment, whereas other methods tend to distort subject identity. All results are based on Stable Diffusion 1.4.

Table 1. Comparison With Mask-Based Editing Approaches

Method	**Application Level**	**Mask Source**	**Mechanism**	**Hierarchical**
Blended Diffusion	Pixel	User-provided	Blending	✗
Blended Latent Diffusion	Latent	User-provided	Blending	✗
DiffEdit	Latent	Auto	Blending	✗
PFB-Diff	Feature map	User-provided	Blending	✓
MINDiff	Attention	Auto	**Attention Suppression**	✓

applying Negative Attention, this mask is inverted for Background Masking. This comparison demonstrates how the suppression of subject influence evolves over time, ensuring better alignment with the text prompt while reducing overfitting.

Threshold Sensitivity

To analyze how different threshold values affect mask generation, we modify the threshold—initially set as the mean attention value—by applying scaling factors ranging from 0.6 to 1.0. As shown in Fig. 4, this analysis demonstrates how variations in the mask range influence the generated results.

Figure 2. Qualitative results of MINDiff applied to LoRA-based DreamBooth using the Stable Diffusion XL 1.0 backbone. The first column shows input images, and the remaining columns show generated results conditioned on different text prompts. MINDiff maintains subject fidelity while improving text alignment, demonstrating its applicability to LoRA-based model.

As the scale value decreases, the selected region becomes more similar to the input image, leading to an expansion of the similar area. At scale 0.7, a tray appears, and at scale 0.6, a cup is generated, while these objects do not emerge at other scales. When the extracted mask region expands and includes part of the background—such as at scale 0.6 and 0.7—the entire background tends to follow the input image, even when it is not fully covered.

Distributional Evidence for the Need for Controllability

Prior personalization methods typically report average scores over evaluation sets to summarize subject fidelity and text alignment. However, these scalar metrics do not reflect the inherent variability introduced by different prompts, subjects, and initial noise conditions. In Fig. 5, we visualize the distribution of 3,000 samples used for quantitative evaluation. This reveals considerable variability across con-

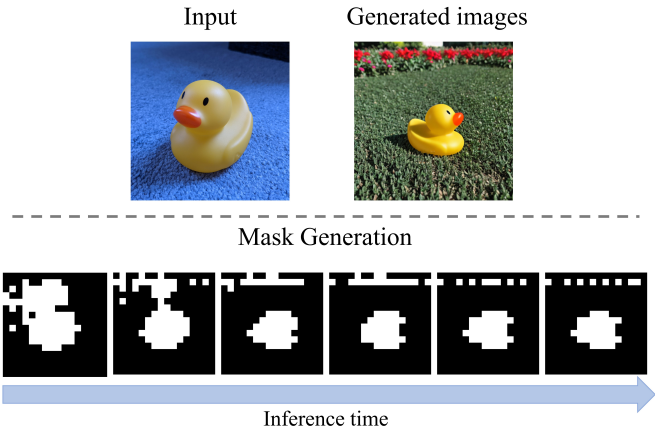

Figure 3. The mask comparison is presented by dividing the denoising process into six timesteps in MINDiff. The figure follows a left-to-right flow, with arrows indicating the progression of inference time. The text prompt used for this visualization is “a sks toy in a flower garden”.

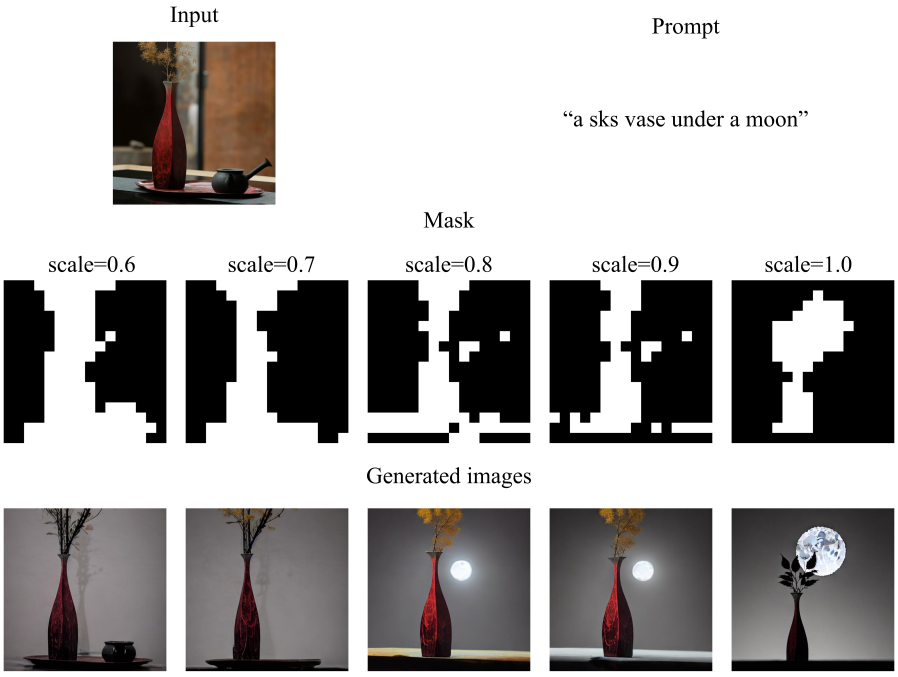

Figure 4. Mask generation results with varying threshold values. This is the result of scaling the initial threshold from 0.6 to 1.0 All results are generated with the same seed.

ditions, even under a fixed model, highlighting the importance of user-controllable mechanisms to steer generation toward desired outcomes.

Comparison of Attention: Full vs. Subject Prompt

We visualize cross-attention outputs at three denoising steps—5, 25, and 45—to compare attention behavior between the full prompt and the isolated subject prompt. As shown in Figure 6, the left panel corresponds to the full prompt (“a sks dog in the snow”), while the right panel shows the attention for the subject-only prompt (“a sks dog”). Although the attention values differ, the spatial layouts remain largely aligned across steps. This demonstrates that the subtraction-based suppression mechanism preserves the latent structure while effectively removing the subject’s influence from undesired regions.

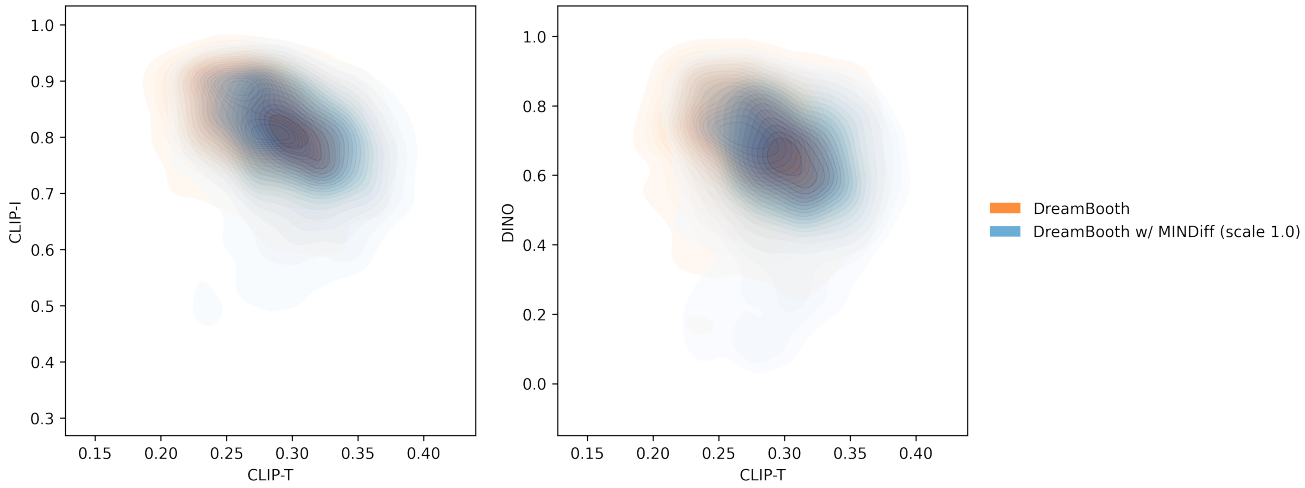

Figure 5. 2D KDE visualization of CLIP-T vs. CLIP-I (left) and CLIP-T vs. DINO (right) scores over the same 3,000 evaluation samples used in quantitative experiments. These results highlight the inherent trade-off variation across subject–prompt pairs, and demonstrate that MINDiff enables directional control over this distribution via scale adjustment.

Failure Cases from Improper Lambda Tuning

Figure 7 shows failure cases observed when the λ value is not appropriately set. When λ is insufficient, the model tends to disregard the text prompt, often omitting key elements such as the "blue house" in the background. Conversely, when λ is excessively high, the model prioritizes text alignment at the cost of subject fidelity, occasionally failing to generate the target subject entirely. These results highlight the importance of selecting an appropriate λ to balance subject preservation and prompt adherence.

Inference Latency and Memory Usage

We evaluated the inference performance of MINDiff on a system equipped with a Quadro RTX 8000 GPU and Stable Diffusion 1.4 (fp16). As shown in Tab. 2, MINDiff introduces only a marginal overhead compared to vanilla DreamBooth. The average GPU inference latency increased slightly, from 3055.71 ms to 3209.44 ms, while peak memory usage remained nearly unchanged (3269.06 MB vs. 3271.56 MB). Furthermore, as summarized in Tab. 3, both latency and throughput across varying batch sizes (1–8) show minimal differences between DreamBooth and DreamBooth with MINDiff. These results indicate that MINDiff adds negligible computational overhead and scales efficiently under increased workloads.

Additional Generation Results

Figure 8 shows the results of artistic rendering, where MINDiff effectively captures and represents the essence of each artistic style. Figure 9 showcases diverse icon designs, demonstrating the model's ability to generate stylistically diverse outputs.

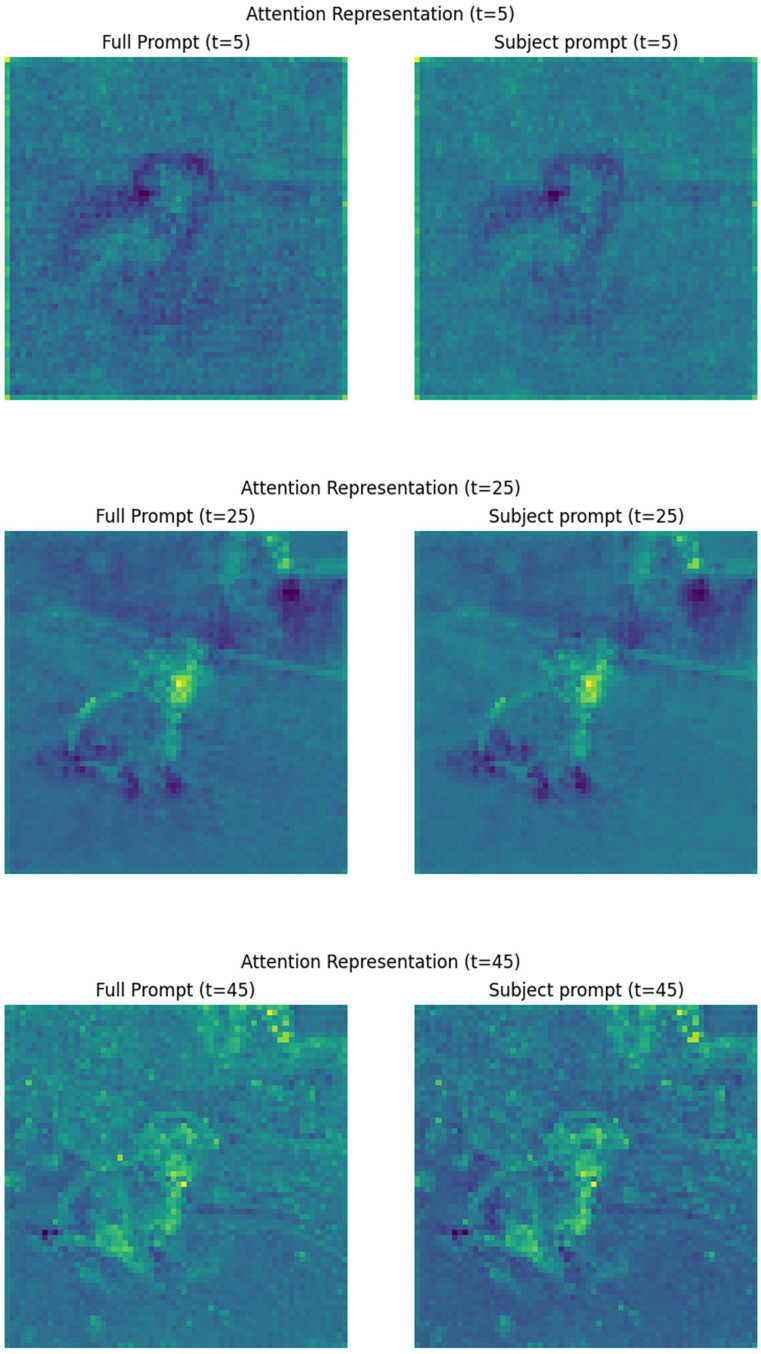

Figure 6. Cross-attention maps extracted at different denoising steps (5, 25, and 45 from top to bottom). The left column shows attention from the full prompt ("a sks dog in the snow"), and the right column corresponds to the subject-only prompt ("a sks dog"). While the attention values differ, the spatial layouts remain largely consistent, supporting the validity of our subtraction-based suppression mechanism.

Figure 7. Failure cases observed with different λ values. Each row is generated using the same seed. The text prompt used is "a sks *class* with a blue house in the background", where *class* represents the broader category to which the subject belongs. Some images fail to generate the subject, while others do not faithfully incorporate the text prompt, highlighting the challenge of selecting an appropriate λ.

Table 2. Inference latency and peak GPU memory usage (batch size = 1, repeat = 30)

Metric	DreamBooth	DreamBooth w/ MINDiff
Inference latency (ms, CPU)	3030.23	3181.53
Inference latency (ms, GPU)	3055.71	3209.44
Peak GPU memory (MB)	3269.06	3271.56

Table 3. Per-image inference latency across different batch sizes (ms)

Batch Size	DreamBooth	DreamBooth w/ MINDiff
1	3061.46	3137.04
2	2694.44	2748.96
4	2426.00	2474.99
8	2301.62	2358.18

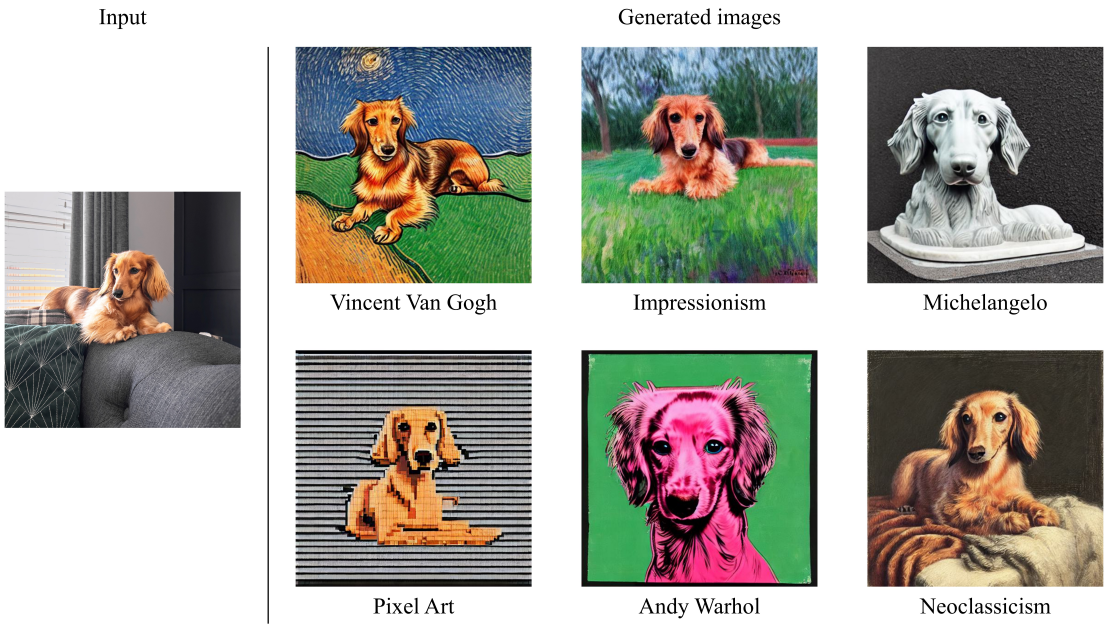

Figure 8. Artistic style generation using MINDiff. The model successfully generates images in diverse artistic styles, including Van Gogh, Impressionism, Michelangelo, Pixel Art, Andy Warhol, and Neoclassicism.

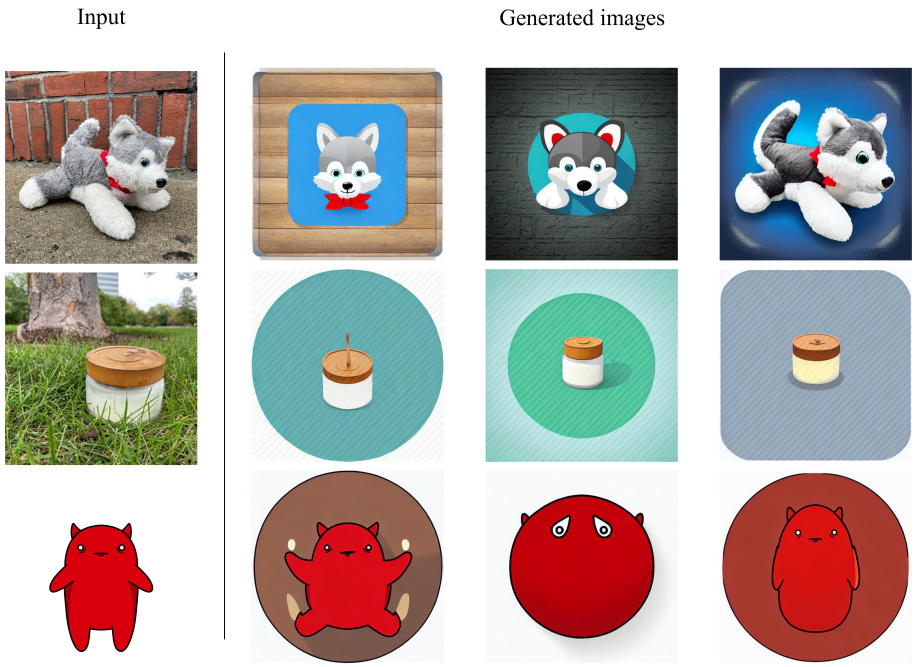

Figure 9. Icon generation using MINDiff.

References

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

[2] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 18208–18218, 2022. 1

[3] Omri Avrahami, Ohad Fried, and Dani Lischinski. Blended latent diffusion. *ACM transactions on graphics (TOG)*, 42 (4):1–11, 2023. 1

[4] Hong Chen, Yipeng Zhang, Simin Wu, Xin Wang, Xuguang Duan, Yuwei Zhou, and Wenwu Zhu. Disenbooth: Identity-preserving disentangled tuning for subject-driven text-to-image generation. *arXiv preprint arXiv:2305.03374*, 2023. 1

[5] Guillaume Couairon, Jakob Verbeek, Holger Schwenk, and Matthieu Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. *arXiv preprint arXiv:2210.11427*, 2022. 1

[6] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. *arXiv preprint arXiv:2208.01618*, 2022. 1

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3, 2022. 1

[8] Wenjing Huang, Shikui Tu, and Lei Xu. Pfb-diff: Progressive feature blending diffusion for text-driven image editing. *Neural Networks*, 181:106777, 2025. 1

[9] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1931–1941, 2023. 1

[10] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Customizable image synthesis with multiple subjects. *Advances in neural information processing systems*, 36:57500–57519, 2023. 1

[11] Lianyu Pang, Jian Yin, Baoquan Zhao, Feize Wu, Fu Lee Wang, Qing Li, and Xudong Mao. Attndreambooth: Towards text-aligned personalized text-to-image generation. *Advances in Neural Information Processing Systems*, 37: 39869–39900, 2024. 1

[12] Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Yael Pritch, Michael Rubinstein, and Kfir Aberman. Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22500–22510, 2023. 1