IMAGE-ALCHEMY: ADVANCING SUBJECT FIDELITY IN PERSONALIZED TEXT-TO-IMAGE GENERATION

Amritanshu Tiwari, Cherish Puniani, Kaustubh Sharma, Ojasva Nema *
Indian Institute of Technology Roorkee, India
{amritanshu_t@mfs, cherish_p@me, kaustubh_s@ee, ojasva_n@mt}.iitr.ac.in

ABSTRACT

Recent advances in text-to-image diffusion models, particularly Stable Diffusion, have enabled the generation of highly detailed and semantically rich images. However, personalizing these models to represent novel subjects based on a few reference images remains challenging. This often leads to catastrophic forgetting, over-fitting, or large computational overhead. We propose a two-stage pipeline that addresses these limitations by leveraging LoRA-based fine-tuning on the attention weights within the U-net of Stable Diffusion XL model. Next, we exploit the unmodified SDXL to generate a generic scene, replacing the subject with its class label. Then we selectively insert the personalized subject through a segmentation-driven Img2Img pipeline that uses the trained LoRA weights. The framework isolates the subject encoding from the overall composition, thus preserving SDXL's broader generative capabilities while integrating the new subject in a high-fidelity manner. Our method achieves a DINO similarity score of 0.789 on SDXL, outperforming existing personalized text-to-image approaches.

Figure 1: Image-Alchemy : Illustration of personalized generation for the subject "iklan", the token chosen for the given person. The left panel shows four reference images used to fine-tune our model, and the right panel demonstrates four diverse outputs, highlighting how the learned subject adapts seamlessly to various prompts.

1 INTRODUCTION

Deep generative models (DGMs), particularly latent diffusion models have played a significant role in revolutionizing high resolution image synthesis. The state-of-the-art latent diffusion based image generation models, including Stable Diffusion(Rombach et al. (2022)), Dall-e (Ramesh et al. (2021)), have been able to produce convincing images of generic objects and scenes, but they struggle with representing a new, specific subject. To distill the information of a new subject into a

*Equal contribution

diffusion model's existing knowledge, one needs to update the vocabulary for a new token (Gal et al. (2022)), and train the model again with the images of that subjet through expensively large computational resources, which is not feasible to do for a new object like a pet or a person. Naive solutions, such as fine-tuning all model parameters on a handful of subject images, often result in catastrophic forgetting, where the model's broader knowledge is overwritten by the new concept. Few shot fine-tuning techniques including Dreambooth (Ruiz et al. (2023)), Hyperdreambooth (Ruiz et al. (2024)) and Textual inversion (Gal et al. (2022)) either have a high computational burden, or low image fidelity, limiting their adaptability. Furthermore, ensuring that the new subject fits seamlessly into a scene becomes difficult as the fine-tuning process significantly alters the model's prior distribution.

In this work, we propose a two-stage pipeline that aims to preserve the original generative strengths of the latent diffusion-based (Rombach et al. (2022)) image generation model Stable Diffusion XL (Podell et al. (2023)) while introducing new subjects learned in a fast, lightweight manner. Our work include:

1. **Token selection:** We look for combination of letters in the CLIP (Radford et al. (2021)) tokenizer that carry little to no prior association with any word. Prompting the model with these tokens yields varied or unrelated images, indicating that the model does not strongly link them to an existing concept.
2. **LoRA-based fine-tuning:** We fine-tune only the attention layers of the Unet (Ronneberger et al. (2015)) of SDXL (Podell et al. (2023)) with LoRA (Hu et al. (2021)) on 4–5 subject images, such that the model overfits on the new object, and store the LoRA safetensors separately.
3. **Two-Stage Generation:**
   (a) **Stage 1:** Generate a generic image using the base (unmodified) SDXL by substituting the subject with its class label (e.g., "person") in the prompt.
   (b) **Stage 2:** Segmentation blurring using grounded SAM (Ren et al. (2024)) + Img2Img process to replace the segmented subject region with the newly learned concept using the LoRA safetensors, ensuring minimal interference with the overall scene.
4. **Experimental Analysis:** We demonstrate that this approach efficiently combines the model's compositional power with high-fidelity subject insertion, with the entire pipeline taking only about 7–8 minutes to complete on the SDXL model.

Our empirical evaluation demonstrates that the proposed pipeline offers a fast, lightweight, and transferable solution to personalization in diffusion models. By isolating subject insertion from the broader generative distribution, we effectively mitigate challenges such as token selection, catastrophic forgetting, and excessive overfitting. We envision this approach as a stepping stone toward more robust, theoretically principled, and practically efficient methods for embedding novel concepts in deep generative models. This aligns with the workshop's emphasis on bridging theoretical insights and practical efficacy, contributing to the advancement of DGMs in real-world applications.

2 Related Work

2.1 Text-to-image generation

Generative models have significantly evolved, with diffusion models (Ho et al. (2020), Nichol & Dhariwal (2021) , Rombach et al. (2022) , Saharia et al. (2022a) , Saharia et al. (2021) , Sohl-Dickstein et al. (2015) , Song et al. (2022) , Song & Ermon (2020a) , Song & Ermon (2020b)), Generative Adversarial Networks (GANs)(Brock et al. (2019) , Goodfellow et al. (2014) , Karras et al. (2021) , Karras et al. (2019) , Karras et al. (2020)), and transformer-based architectures driving advancements in text-to-image synthesis. Diffusion models, such as DALL·E 2 (Ramesh et al. (2022)), Imagen (Saharia et al. (2022b)), Stable Diffusion (Rombach et al. (2022)), and MidJourney[1] , have demonstrated remarkable capabilities by iteratively denoising a latent representation to refine random noise into coherent images.

[1] https://www.midjourney.com

2.2 PERSONALIZED IMAGE GENERATION

Several techniques have been developed to introduce personalization into diffusion-based image generation. DreamBooth (Ruiz et al. (2023)) refines personalization by fine-tuning text-to-image models on a few user-provided images, allowing for the synthesis of subject-specific visuals while maintaining identity consistency. Textual Inversion (Gal et al. (2022)) introduces a novel approach by learning new embeddings that encode personalized concepts within the latent space of a diffusion model, eliminating the need for full model fine-tuning. HyperDreamBooth (Ruiz et al. (2024)) improves these methods by incorporating hypernetworks, which improve adaptability and efficiency when fine-tuning diffusion models. Infusion (Zeng et al. (2024)) prevents concept overfitting by categorizing it into concept-agnostic and concept-specific types but struggles to preserve image diversity. Cones (Liu et al. (2023)) introduces concept neurons, focusing on interpretability over fine-tuning effectiveness, limiting its use for general customization. InstantBooth (Shi et al. (2023)) eliminates test-time fine-tuning, it may struggle with capturing intricate identity details and handling complex concepts due to its reliance on global embeddings. Diffusion models have also been widely adopted for image editing, where methods like MyStyle (Nitzan et al. (2022)) performed well in maintaining identity during image editing but needed extensive fine-tuning and computational power, restricting scalability. ForgeEdit (Zhang et al. (2024)) enable fine-grained local modifications while maintaining global coherence. InstructPix2Pi (Brooks et al. (2023)) extends this capability by integrating natural language instructions into the editing process, allowing users to specify precise changes using textual descriptions.

2.3 SEGMENTATION FOR ENHANCED CONTROL

Segmentation models have played a crucial role in refining the control mechanisms for generative models. Grounded SAM (Ren et al. (2024)), which combines the capabilities of Grounding DINO (Liu et al. (2024)) and the Segment Anything Model (SAM) (Kirillov et al. (2023)), allows for open-set object detection and segmentation guided by textual prompts. This integration enhances the applicability in tasks requiring region-specific transformations, such as targeted image editing, compositional generation, and object-focused synthesis.

2.4 EFFICIENT ADAPTATION USING LORA

Fine-tuning large-scale diffusion models for personalized or domain-specific tasks is computationally expensive. Low-Rank Adaptation (LoRA) (Hu et al. (2021)) addresses this challenge by decomposing weight updates into low-rank matrices, significantly reducing the number of trainable parameters while maintaining expressive power. Applied to diffusion models, LoRA enables efficient adaptation with limited computational resources, facilitating rapid personalization and domain adaptation without requiring full model retraining. This technique has proven particularly useful for adapting generative models to new tasks while preserving efficiency.

3 METHODOLOGY

3.1 TOKEN SELECTION

In text-to-image diffusion models like SDXL, each prompt token is first mapped to a token embedding via CLIP's text encoder. If an existing token already has a strong semantic association (e.g., "dog," "cup," or any common concept), using it to introduce a new subject can cause conflicts in the model's learned representation. Consequently, we seek a "rare" token—one that the model does not strongly associate with any specific concept, and for which the CLIP tokenizer is not well trained on—so that the newly introduced subject can be learned without disrupting the model's learned representation by a substantial margin.

Vocabulary Scan and Generation Test. We begin by scanning the CLIP tokenizer vocabulary for apparently gibberish tokens, typically short words (4–5 letters) that are unlikely to appear in

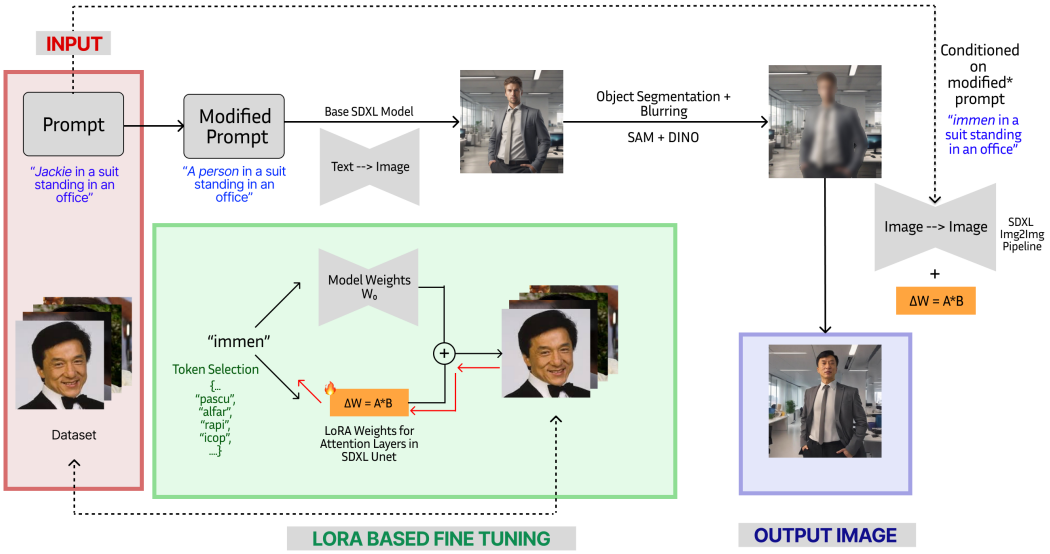

Figure 2: Overall Pipeline of Image-Alchemy

conventional usage [2]. For each candidate token, we perform multiple text-to-image generations using the unmodified SDXL under different random seeds. We then inspect the resulting images:

- If the token consistently yields visually similar or thematically consistent outputs, it suggests the model associates it with an existing concept.
- If the token leads to highly varied or incoherent images, we infer that the model does not possess a strong prior for that token.

By comparing the generated images visually and through SSIM, we select the tokens that exhibit the greatest variability and least recognizable pattern.

Choosing a token with little to no prior allows the model to map our new subject's features onto an essentially "blank slate." When we train the model on a small set of subject images, the learned LoRA parameters effectively overwrite the token embedding and attention pathways tied to that placeholder. Because the base model had no strong associations with that token, the new concept is incorporated without disrupting other learned representations.

3.2 LoRA BASED FINE TUNING

Having selected a suitable placeholder token (3.1), we next fine-tune the Stable Diffusion XL (SDXL) model on a small set of subject images. Instead of modifying all model parameters, as shown in Multi-Concept Customization of Text-to-Image Diffusion ((Kumari et al., 2023)) that during fine-tuning of diffusion models, the weight changes mostly occur in the cross-attention and self-attention layers, therefore we apply LoRA (Low-Rank Adaptation) to only a subset of the Unet of the Stable Diffusion-XL model weights. The Low Rank adaptation ensures prevention of catastrophic forgetting in the model's learned distribution. [3]

3.2.1 INTENTIONAL OVERFITTING

Unlike conventional fine-tuning, our goal here is to ensure the model memorizes the subject's unique features. We leverage the small training set (4–5 images) to push the LoRA parameters toward capturing fine details of the subject, resulting in overfitting. While overfitting is generally discouraged, we circumvent this problem by isolating the subject insertion to a later stage (3.4).

[2]For more details and our proposed tokens , see appendix

[3]For theoretical proof of the concept, see appendix

3.2.2 STORING LORA WEIGHTS SEPARATELY

The LoRA approach appends low-rank parameter matrices (ΔW) to the original model weights. After training, we store these LoRA parameters (in .safetensors format) separately. This ensures non-destructive adaptation where the base model remains unchanged and flexible deployment such that a user can load or unload the specialized LoRA weights as and when required.

3.3 BASE IMAGE GENERATION AND SEGMENTATION

The segmentation pipeline integrates Grounding DINO's zero-shot detection capabilities with SAM's prompt-driven segmentation in a two-step framework, resulting in mask generation. This methodology employs cross-modal alignment to achieve semantic segmentation and spatial refinement mechanisms to achieve pixel level boundaries in output masks.

3.3.1 DETECTION PHASE:

Grounding DINO uses a Swin Transformer (Liu et al. (2021)) image encoder and a BERT-based (Devlin et al. (2019)) text encoder to project visual and textual features into a shared d-dimensional space ($d = 256$). Cross-modality fusion is achieved through attention mechanisms:

$$A_{i,j} = \text{softmax}\left(\frac{\phi(f_i)^T \psi(e_j)}{\sqrt{d}}\right) \tag{1}$$

where ϕ and ψ are linear projections of image features (f_i) and text embeddings (e_j). The model outputs N bounding boxes ($B \in \mathbb{R}^{N\times 4}$) with confidence scores ($s \in [0,1]^N$), filtered using a threshold τ to remove low-confidence proposals.

3.3.2 SEGMENTATION PHASE:

SAM processes the detected boxes using a ViT-H/16 (Dosovitskiy et al. (2021)) image encoder ($Z \in \mathbb{R}^{256\times H/16\times W/16}$) and encodes box coordinates into sparse embeddings ($E_b \in \mathbb{R}^{N\times 256}$). The mask decoder computes:

$$M = \sigma\left(\text{MLP}\left([Z \oplus (Z \odot E_b)]\right)\right) \tag{2}$$

where $\oplus$ denotes concatenation and σ is the sigmoid activation. The optimal mask is selected from multiple hypotheses using:

$$M^* = \arg\max_{M_i}\left(\text{IoU}(M_i, B) - \lambda|M_i|\right) \tag{3}$$

where λ is a hyperparameter balancing mask precision and area coverage.

3.4 IMG2IMG PIPELINE

Once the base image is generated (3.3) and the subject region is blurred, we load the LoRA-adapted SDXL model from Section 3.2 and construct a prompt containing the placeholder token. For example, if the original user prompt was "A photo of Rahul sitting on a chair," we replace "Rahul" with the placeholder token (e.g., "immen") to obtain "A photo of immen sitting on a chair." We then apply an Img2Img process: the blurred image serves as the initial input, and the prompt is applied as a condition, as the learned LoRA weights guide the diffusion model to fill in the blurred region with the newly introduced subject.

3.4.1 REVERSE DIFFUSION WITH BLURRED REGIONS

Latent diffusion models can be viewed as a stochastic differential equation (SDE):

$$dx = -\nabla_x \log p_t(x)\, dt + \sqrt{\beta(t)}\, dW_t, \tag{4}$$

where x is the latent variable, and $-\nabla_x \log p_t(x)$ guides denoising. In blurred areas, gradient cues ($-\nabla_x \log p_t(x)$) are weaker because there is less high-frequency structure for the model to preserve. Consequently, the model relies more heavily on its learned priors—in our case, the subject-specific LoRA weights tied to the placeholder token. This mechanism enables accurate reconstruction of the newly introduced subject in regions designated for replacement.

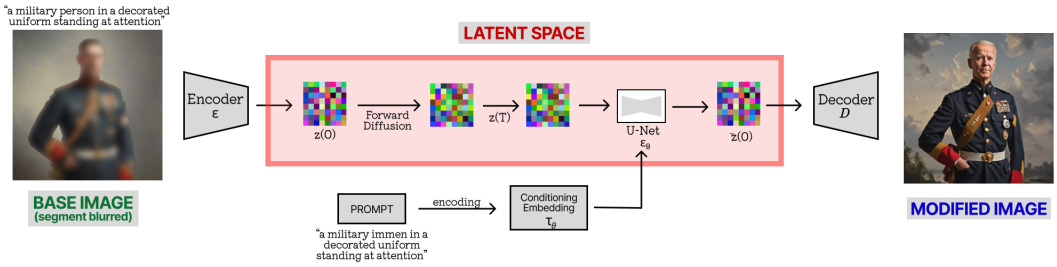

Figure 3: **Illustration of our Img2Img pipeline.** The base image is first segmented and blurred (left), then encoded into the latent space. The learned LoRA token (e.g., "immen") is combined with the prompt and fed to the U-Net for iterative refinement, after which the decoder produces the final, personalized image (right).

3.4.2 Composition and Style Preservation

Isolating the subject-insertion step after a generic image has already been produced removes the risk of catastrophic forgetting. Rather than relying on the fine-tuned model to generate the entire scene, we exploit SDXL's original, unaltered capabilities for layout and background details. Consequently, the overfitted LoRA parameters only dominate where the subject must appear. This approach balances high-fidelity personalization with preserved compositional strength, which ensures integration of the new subject into the generated image effectively. Figure 3 illustrates how isolating subject insertion preserves composition while ensuring high-fidelity personalization.

Figure 4: Column 1 has the prompts used to generate the images, Column 2 portrays the images generated using normal fine-tuning technique, the Column 3 contains the corresponding final output images of our pipeline.

4 Experiments

4.1 Direct Fine-Tuning and Catastrophic Forgetting

We fine-tune SDXL on 4–5 reference images of a specific subject (e.g., "Rahul") using a standard approach similar to DreamBooth, without prior preservation loss. Figure 4 illustrates that the model loses its broader generative capabilities or produces images that do not align accurately with the given prompt. However, our model effectively preserves its generative diversity while ensuring

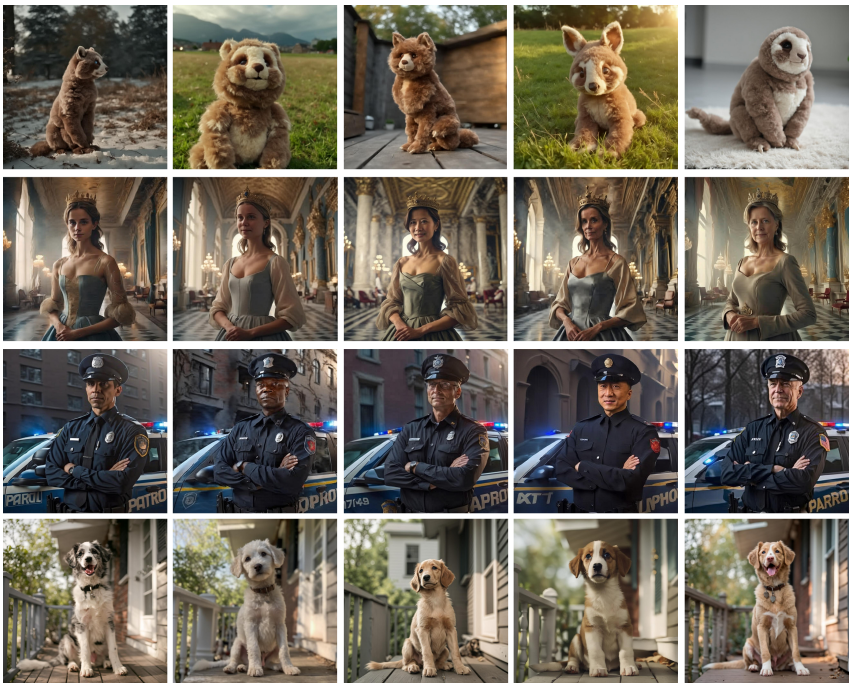

Figure 5: **Seamless Subject Integration.** Row 1 displays a combination of bear and cat images, merging their characteristics while preserving the overall quality and context.Row 2 presents images of various women, each integrated smoothly while keeping their unique identities.Row 3 features images of different men, with each image tailored to an individual while maintaining coherence.Row 4 showcases images of dogs, each distinctly personalized to a different dog, while ensuring a consistent style and quality..

that the generated images remain closely aligned with the input prompt, demonstrating a balanced trade-off between adaptability and fidelity.

4.2 Subject Diversity

We started by evaluating the performance of our model in various subject categories [4], including humans (both male and female) and animals (such as dogs and cats) to evaluate its adaptability and strength. To further examine its ability to generalize, after training the model on images of a specific subject, we provided a prompt for a different entity while using the learned token of the original subject. This led to impressive infusion effects, where characteristics of the trained subject were seamlessly integrated into the newly generated content as shown in Fig 5.

5 Results

5.1 Embedding-Based Similarities

We compute embeddings for both reference images and generated outputs using **CLIP** (CLIP-I) (Radford et al. (2021)) and **DINO** (Caron et al. (2021)), by measuring cosine similarity in the subject region. Higher values indicate better subject preservation.

These metrics are used to evaluate how well the generated outputs preserve the subject from the reference images. **CLIP-I** measures cosine similarity between the embeddings of reference and generated images, but it is not considered a reliable metric for evaluating frameworks like ours because it fails to recognize rare or unique tokens (e.g., ”mccre wearing a leather jacket and jeans”) that hold no meaning in its text encoder, leading to arbitrary similarity scores. **DINO** evaluates

[4]For the dataset details, see appendix

subject preservation by computing cosine similarity. DINO generally performs well in capturing visual features and is less reliant on text-based understanding compared to CLIP. **CLIP-T** measures cosine similarity between textual descriptions of the generated and reference images. It measures how well the generated image aligns with textual prompts.

Table 1: CLIP and DINO embedding similarities (cosine) between final outputs and reference images. Higher is better.

METHOD	DINO	CLIP-I	CLIP-T
Real	0.834	0.763	NA
Dreambooth(Stable Diffusion)	0.668	**0.803**	0.305
Textual Inversion(Stable Diffusion)	0.569	0.780	0.255
Custom Diffusion	0.643	0.798	0.256
Subject Diffusion	0.711	0.787	0.293
Ours (Two-Stage)	**0.789**	0.557	**0.334**

5.2 Image Quality Metrics

We assess image fidelity using metrics such as **NIQE** (Mittal et al. (2013)) which quantifies naturalness by measuring statistical distances between multivariate Gaussian models of test image features and those of pristine natural images, using MSCN coefficients from local patches, **BRISQUE** (Mittal et al. (2012)) evaluates distortions through SVR-based analysis of locally normalized luminance statistics' deviation from natural image models in the spatial domain and, **MANIQA** (Yang et al. (2022)) a transformer-based IQA using multi-head self-attention to learn hierarchical quality features through spatial and channel attention weighting of image patches, where applicable.

Table 2: Blind Image quality metrics

METHOD	NIQE ↑	MANIQA ↑	BRISQUE ↓
Real Images	7.837	0.5264	23.644
Unmodified SDXL	5.681	0.3930	45.047
Ours (Two-Stage)	5.727	0.3571	46.302

Table 2 suggests that our two-stage pipeline successfully personalizes the subject while minimally impacting the broader scene quality.The difference between SDXL's baseline generation and our personalized outputs remains small, suggesting minimal degradation in overall fidelity. As expected, real images achieve the best metric scores.

6 Conclusion

In conclusion, our framework for personalized text-to-image generation addresses key challenges in adapting diffusion models to novel subjects while maintaining their broader generative capabilities. By leveraging LoRA-based fine-tuning on SDXL's attention layers and employing a segmentation-driven Img2Img process, the method effectively isolates subject personalization from scene composition. This approach mitigates issues such as catastrophic forgetting and token selection conflicts, while significantly reducing computational costs. The experimental results demonstrate that this pipeline achieves high-fidelity subject integration with minimal disruption to the original model's versatility. The modular nature of LoRA weights ensures flexible deployment without altering the base model, making this solution practical for real-world applications. This work represents a significant step forward in personalized image synthesis, offering a lightweight and efficient framework that bridges theoretical advancements with practical utility. Future research could explore extending this approach to multiple and more complex subjects,further enhancing its adaptability and robustness in diverse use cases.

REFERENCES

Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis, 2019. URL https://arxiv.org/abs/1809.11096.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions, 2023. URL https://arxiv.org/abs/2211.09800.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers, 2021. URL https://arxiv.org/abs/2104.14294.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019. URL https://arxiv.org/abs/1810.04805.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale, 2021. URL https://arxiv.org/abs/2010.11929.

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, 2022. URL https://arxiv.org/abs/2208.01618.

Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014. URL https://arxiv.org/abs/1406.2661.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. URL https://arxiv.org/abs/2106.09685.

Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4396–4405, 2019. doi: 10.1109/CVPR.2019.00453.

Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8107–8116, 2020. doi: 10.1109/CVPR42600.2020.00813.

Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-free generative adversarial networks, 2021. URL https://arxiv.org/abs/2106.12423.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023. URL https://arxiv.org/abs/2304.02643.

Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion, 2023. URL https://arxiv.org/abs/2212.04488.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows, 2021. URL https://arxiv.org/abs/2103.14030.

Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept neurons in diffusion models for customized generation, 2023. URL https://arxiv.org/abs/2303.05125.

Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012. doi: 10.1109/TIP.2012.2214050.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

Alex Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

Yotam Nitzan, Kfir Aberman, Qiurui He, Orly Liba, Michal Yarom, Yossi Gandelsman, Inbar Mosseri, Yael Pritch, and Daniel Cohen-or. Mystyle: A personalized generative prior, 2022. URL https://arxiv.org/abs/2203.17272.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis, 2023. URL https://arxiv.org/abs/2307.01952.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. URL https://arxiv.org/abs/2102.12092.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL https://arxiv.org/abs/2204.06125.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, Zhaoyang Zeng, Hao Zhang, Feng Li, Jie Yang, Hongyang Li, Qing Jiang, and Lei Zhang. Grounded sam: Assembling open-world models for diverse visual tasks, 2024. URL https://arxiv.org/abs/2401.14159.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022. URL https://arxiv.org/abs/2112.10752.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation, 2015. URL https://arxiv.org/abs/1505.04597.

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, 2023. URL https://arxiv.org/abs/2208.12242.

Nataniel Ruiz, Yuanzhen Li, Varun Jampani, Wei Wei, Tingbo Hou, Yael Pritch, Neal Wadhwa, Michael Rubinstein, and Kfir Aberman. Hyperdreambooth: Hypernetworks for fast personalization of text-to-image models, 2024. URL https://arxiv.org/abs/2307.06949.

Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement, 2021. URL https://arxiv.org/abs/2104.07636.

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models, 2022a. URL https://arxiv.org/abs/2111.05826.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022b. URL https://arxiv.org/abs/2205.11487.

Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning, 2023. URL https://arxiv.org/abs/2304.03411.

Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics, 2015. URL https://arxiv.org/abs/1503.03585.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models, 2022. URL https://arxiv.org/abs/2010.02502.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution, 2020a. URL https://arxiv.org/abs/1907.05600.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models, 2020b. URL https://arxiv.org/abs/2006.09011.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. Maniqa: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1191–1200, 2022.

Weili Zeng, Yichao Yan, Qi Zhu, Zhuo Chen, Pengzhi Chu, Weiming Zhao, and Xiaokang Yang. Infusion: Preventing customized text-to-image diffusion from overfitting, 2024. URL https://arxiv.org/abs/2404.14007.

Shiwen Zhang, Shuai Xiao, and Weilin Huang. Forgedit: Text guided image editing via learning and forgetting, 2024. URL https://arxiv.org/abs/2309.10556.

A Appendix

A.1 Blurring Techniques

We integrated selective blurring of the main object in the base image before passing it to the img-to-img SDXL pipeline. This step was designed to enhance the quality of edits by focusing modifications in proximity to the primary object of interest. Two distinct blurring techniques were evaluated for this purpose:

- **Gaussian Blur:** In this method, we applied a Gaussian blur to the region identified by the mask generated by the Grounded Segment Anything Model (SAM). The blur was applied uniformly across the masked region with kernel_size=151, $\sigma = 100$ for the Gaussian kernel.

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right)$$

- **Exponential Decay Gaussian Blur:** This technique introduced a more nuanced approach by incorporating an exponential decay factor into the Gaussian blur. The degree of blurring decreased exponentially with increasing distance from the main object, as determined by the nearest zero value in the mask generated by Grounded SAM. (kernel_size=151, $\sigma = 100$, $\lambda = 5$)

$$G'(x, y) = G(x, y) \cdot e^{-\lambda d}$$

The experimental results indicate that the second technique consistently produces superior results. The key reason behind this improvement is that an exponential decay in blurring ensures a smoother transition between the blurred and unblurred regions. Unlike the fixed Gaussian blur, which applies a uniform blur within the segmented area and often leads to unnatural transitions, the decay-based approach maintains spatial coherence by maintaining the details in the background of the object and smoother transitions in the vicinity of the object.

A.2 TOKENS

We found that the tokens used by DreamBooth for image personalization show strong biases in the CLIP text encoder of the Stable Diffusion XL (SDXL) model. For example, the token "sks" often led to the generation of a firearm when used for subject personalization. This indicates that the token had a strong existing meaning in the model, which can harm subject adaptation.

To address this problem, we propose a new set of tokens for subject personalization in SDXL. These tokens are chosen for their limited use in the text encoder, ensuring that the model has little to no prior understanding of them. This allows for more accurate and unbiased subject learning, resulting in better image personalization without interference from existing associations. Using these tokens encode and represent the personalized subject more effectively, producing more reliable and controllable results. Such tokens are listed below:

immen	pasqu	iklan	rapi
bhar	ellu	ffin	icop
aben	mmor	psal	phyl
rrrr	wozni	geaux	koval
ayles	mccre	fortn	prote
pascu	lisam	percu	alfar
insom	offro	syour	redon

A.3 DATASET

We used data from various sources to strengthen our fine-tuning at different stages of our process. We selected images that are commercially licensed and copyright-free to meet legal and ethical standards and for object-specific fine-tuning, we included the dataset from the DreamBooth paper. Figure 6 illustrates a sample dataset instance used for metric evaluation and pipeline validation.

A.4 CATASTROPHIC FORGETTING MITIGATION THROUGH LOW RANK ADAPTATION

In our setup, the base Stable Diffusion XL model is parameterized by $\theta \in \mathbb{R}^D$, and we learn a small set of LoRA parameters $\Delta\theta \in \mathbb{R}^d$ (where $d \ll D$) that modify only the attention layers. Let $p_\theta(\mathbf{x} \mid \mathbf{y})$ denote the model's distribution over images $\mathbf{x}$ given prompt $\mathbf{y}$. After LoRA adaptation, the parameters become $(\theta, \Delta\theta)$, inducing a (slightly) modified distribution $p_{\theta+\Delta\theta}(\mathbf{x} \mid \mathbf{y})$.

We introduce the LoRA update as:

$$W_{\text{LoRA}} = W + \alpha \cdot UV^\top, \tag{5}$$

where $W \in \mathbb{R}^{d_1 \times d_2}$ is a pre-trained attention weight matrix, $U \in \mathbb{R}^{d_1 \times r}$ and $V \in \mathbb{R}^{r \times d_2}$ define a rank-r update (with $r \ll \min(d_1, d_2)$), and α is a scaling constant (α = 1 in our case). Only U, V are optimized, leaving W fixed. Consequently, the effective update $\Delta W = \alpha \cdot UV^\top$ occupies a low-dimensional subspace.

Bound on Distributional Shift. We can characterize the distributional shift via a divergence measure, e.g. the Kullback-Leibler (KL) divergence:

$$D_{\text{KL}}\big(p_{\theta+\Delta\theta}(\mathbf{x} \mid \mathbf{y}) \,\big\|\, p_\theta(\mathbf{x} \mid \mathbf{y})\big). \tag{6}$$

Under assumptions of Lipschitz continuity in θ-space, we can bound equation 6 by a function of $\|\Delta\theta\|$ (or $\|\Delta W\|$). Since ΔW is rank-r , we have:

$$\|\Delta W\|_F \;=\; \|\alpha \cdot UV^\top\|_F \;\leq\; \alpha\,\|U\|_F\,\|V^\top\|_F \;=\; \alpha\,\|U\|_F\,\|V\|_F,$$

which is significantly smaller than a full $d_1 \times d_2$ dimensional update if r is controlled. Thus,

$$D_{\text{KL}}\big(p_{\theta+\Delta\theta}, p_\theta\big) \;\leq\; \kappa \cdot \|\Delta W\|_F,$$

for some Lipschitz constant $\kappa > 0$. Therefore, if $\|\Delta W\|_F$ is kept small, the change in the global distribution $p_\theta(\mathbf{x} \mid \mathbf{y})$ remains bounded, preserving much of the original knowledge.

Basically, *distributional shift* from p_θ to $p_{\theta+\Delta\theta}$ depends on $\|\Delta W\|_F$, which in turn is bounded by $\alpha\,\|U\|_F\,\|V\|_F$. Thus, a small-rank update constrains how *strongly* the new subject can overwrite existing representations, mitigating forgetting.

Figure 6: An instance from the datasets used to fine tune Stable Diffusion XL model.

When Might LoRA Fail? If the new subject is extremely complex or demands rewriting fundamental visual features, a low-rank subspace might be insufficient, leading to partial adaptation. In that case, there is incomplete personalization or minor forgetting for other concepts. However, for typical tasks (like injecting a single person or object), this rank-limited approach is both sufficiently expressive and far safer than full-model fine-tuning.