STENCIL: SUBJECT-DRIVEN GENERATION WITH CONTEXT GUIDANCE

Gordon Chen[1,2], Ziqi Huang[1], Cheston Tan[2], Ziwei Liu[1]

[1]Nanyang Technological University, [2]CFAR, IHPC, A*STAR

stencil.github.io

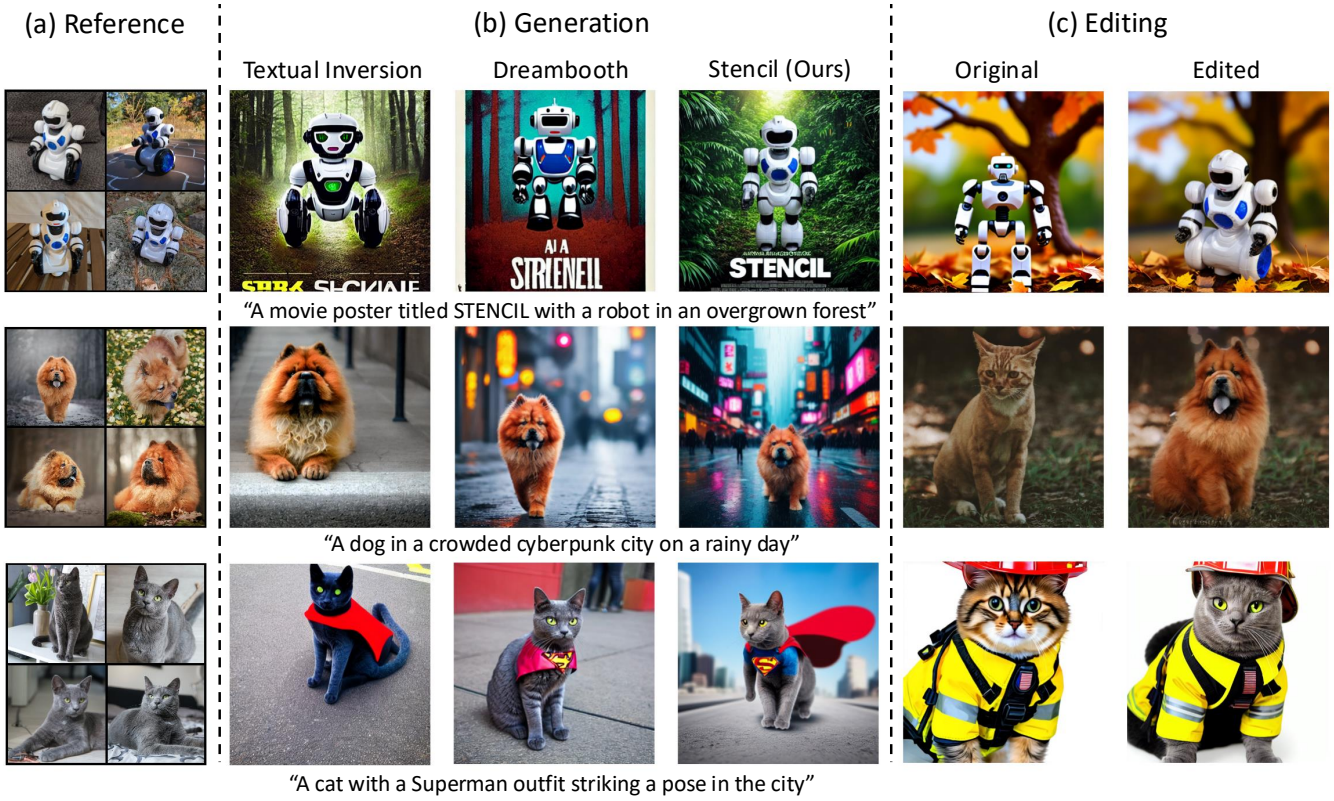

Fig. 1: **Overview of *Stencil***. Given a few (a) reference images, Stencil achieves (b) subject-driven generation and (c) subject editing with high textual and subject fidelity in just 100 fine-tuning steps.

ABSTRACT

Recent text-to-image diffusion models can produce impressive visuals from textual prompts, but they struggle to reproduce the same subject consistently across multiple generations or contexts. Existing fine-tuning based methods for subject-driven generation face a trade-off between quality and efficiency. Fine-tuning larger models yield higher-quality images but is computationally expensive, while fine-tuning smaller models is more efficient but compromises image quality. To this end, we present Stencil. Stencil resolves this trade-off by leveraging the superior contextual priors of large models and efficient fine-tuning of small models. Stencil uses a small model for fine-tuning while a large pre-trained model provides contextual guidance during inference, injecting rich priors into the generation process with minimal overhead. Stencil excels at generating high-fidelity, novel renditions of the subject in less than a minute, delivering state-of-the-art performance and setting a new benchmark in subject-driven generation. Supplementary materials are available at IEEE SigPort.

Index Terms— Computer Vision, Diffusion Models, Image Editing, Subject-Driven Generation

1. INTRODUCTION

Text-to-image (T2I) diffusion models [1] have demonstrated remarkable success in producing high-quality, text-aligned images. Recently, subject-driven generation has emerged as

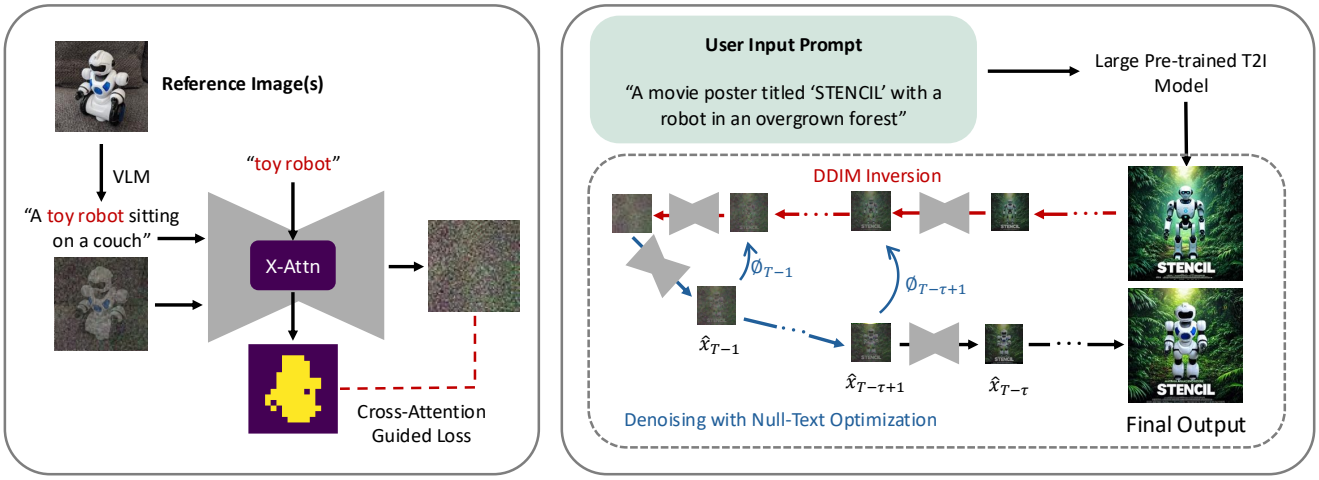

Fig. 2: ***Stencil*** **Framework. (a) Lightweight Subject Fine-Tuning.** We fine-tune a lightweight text-to-image diffusion model on the reference image(s) of the subject. The Cross-Attention Guided Loss is applied so that gradients are computed only in regions influenced by the subject token (e.g. "*toy robot*"). **(b) Context Guidance.** Given a user prompt, we draft an image with a large frozen text-to-image model. The draft is inverted into the latent space of the lightweight fine-tuned model and refined via null-text optimisation, producing the final image that preserves both the prompt context and the personalized subject.

a pivotal research area, enabling users to personalize subjects by providing reference images as input to T2I models. However, subject-driven generation remains a difficult task, with scalability being a primary challenge. Fine-tuning a small-scale diffusion model is efficient, but often results in degraded image quality. In contrast, fine-tuning a large-scale diffusion model [2, 3] yields results with superior image quality, but is computationally expensive. This underscores the need for methods that are both efficient and high-quality.

To this end, we propose Stencil. Stencil resolves this trade-off by leveraging a novel technique called Context Guidance. Stencil fine-tunes a lightweight model, while a large pre-trained model injects its rich contextual priors at inference to guide generation. This collaboration between two models enables efficient and high-fidelity subject-driven outputs that neither model can achieve independently. Furthermore, we introduce the Cross-Attention Guided Loss, which leverages the cross-attention mechanism to focus fine-tuning on subject-relevant regions of the reference images, excluding irrelevant details from the loss computation, reducing optimization complexity and enabling faster convergence of the subject. Stencil achieves state-of-the-art (SOTA) results in generation while being the most cost-effective framework. Our main contributions are as follows:

• We propose Context Guidance, a novel technique that combines the efficiency of fine-tuning small diffusion models with the expressiveness of large ones.

• We propose the Cross-Attention Guided Loss function, where we only optimize subject-relevant areas of the reference images to improve fine-tuning stability and efficiency.

• Our extensive experiments have validated the robustness of our approach, achieving SOTA results.

2. RELATED WORKS

Recent methods for subject-driven generation can be divided into two categories - those that fine-tune the diffusion model on the reference images during test-time [4, 5, 6, 7, 8, 9], and those that train an additional structure to encode the reference images [10, 11, 12, 13, 14, 15, 16]. In this paper, we focus on the former. Textual Inversion [5] optimizes token embeddings within text prompts to capture subject representation. DreamBooth [4] fine-tunes the denoising U-Net to bind the appearance of a subject with specific class tokens. Custom Diffusion [6] proposes to enhance efficiency by limiting fine-tuning to the cross-attention layers of the U-Net. Empirical results show that fine-tuning the U-Net typically yield the best performance, yet the issue of efficiency is often overlooked. Scaling to higher-quality outputs requires fine-tuning larger diffusion models, making each iteration significantly more computationally expensive. To bridge this gap, we propose Context Guidance. Moreover, existing fine-tuning approaches rely on the traditional Mean Squared Error (MSE) for loss computation, which treats every pixel equally and can lead to entanglement between the subject and the rest of the reference image. Hence, to address this, we propose the Cross-Attention Guided Loss, which prevents irrelevant details being baked into the learned subject's representation.

3. METHOD

Fig. 2 illustrates Stencil's method framework, consisting of an initial fine-tuning component with the Cross-Attention Guided Loss (Sec. 3.2) followed by Context Guidance at inference time (Sec. 3.3). Additional implementation details are discussed in *Supplementary Materials*.

3.1. Preliminaries

Text-to-Image Latent Diffusion Models. Diffusion models are a class of generative models that progressively transform pure Gaussian noise x_T into a target image x_0 through iterative denoising steps. Each model consists of a denoising network $f_\theta(x_t, t, \psi(P))$, traditionally a U-Net, conditioned on the text embedding $\psi(P)$, to predict the noise residual ϵ_t of x_t at time-step t, enabling the reconstruction of a slightly de-noised sample x_{t-1}. Latent diffusion models [17] reduce compute complexity by applying the diffusion process on a lower-dimensional latent space z_t. The overall loss function for training such a denoising network is computed as:

$$L = E_{z_0,\epsilon,t,\psi(P)}\left[\|\epsilon - f_\theta(z_t, t, \psi(P))\|_2^2\right] \tag{1}$$

Cross-Attention Mechanism. In cross-attention, the deep spatial features $\phi(z_t)$ are linearly projected to a query $Q = \ell_Q\phi(z_t)$, key $K = \ell_K\psi(P)$, and value $V = \ell_V\psi(P)$ matrix via learned projections ℓ_Q, ℓ_K, ℓ_V respectively. The attention map is formulated as:

$$M = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \tag{2}$$

where d is the latent projection dimension of the keys and queries. The entry M_{ij} defines the weight of the j-th token on the pixel i. Intuitively, cross-attention maps bind each text token to specific regions of the image, which guides the placement of textual elements in the generated image [18]. The attention output $\widehat{\phi}(z_t) = MV$, updates $\phi(z_t)$ and is propagated to the subsequent layers of the U-Net.

Inversion with Null-Text Optimization. Inversion is a core technique allowing us to retrieve the noise vector corresponding to a given image, such that the forward diffusion process on the noise vector reproduces the given image. However, due to the stochasticity of diffusion, the reconstructed image may look different from the original image. Null-text optimization [19] solves this problem by leveraging DDIM inversion to establish a pivot trajectory. At each denoising time-step, the unconditional embeddings $\emptyset_t$ are optimized to minimize the deviations from the pivot trajectory, enabling perfect reconstructions of the original image. The objective is:

$$\min_{\emptyset_t} \left\| z^*_{t-1} - \text{DDIM}(\bar{z}_t, \hat{\epsilon}_t, t)\right\|_2^2, \quad \hat{\epsilon}_t = f_\theta(\bar{z}_t, t, \emptyset_t, \psi(P)) \tag{3}$$

3.2. Cross-Attention Guided Loss Function

Fine-tuning with the traditional MSE loss fails to distinguish between subject and background pixels in the reference images. This not only complicates optimization and slows convergence but also risks embedding irrelevant background features into the learned subject representation. The Cross-Attention Guided Loss effectively addresses this problem by regulating the learned subject token representation to concentrate on the subject-relevant pixels in the reference images.

Specifically, we use a vision-language model (VLM) to generate a caption C for each reference image and use the cross-attention map of the subject token S in C to guide the learning. We first add t time-step noise to the reference images. We then perform a single forward pass of the noisy latent, conditioned on C, through the frozen U-Net. During this forward pass, we extract the cross-attention maps of S from all heads and layers, up-sample them to match the latent resolution, compute the mean and normalize the values to obtain an average cross-attention map, $\widehat{M}_S$. The Cross-Attention Guided Loss is defined as:

$$L = E_{z_0,\epsilon,t,\psi(C)}\left[\|1_{\widehat{M}_S > p_t} \cdot (\epsilon - f_\theta(z_t, t, \psi(C)))\|_2^2\right] \tag{4}$$

where $1_{\widehat{M}_S > p_t}$ is a binary mask (with the same dimensions as the latent image) indicating pixels where the subject token's attention weight exceeds the threshold p_t. A value of 1 marks subject-relevant pixels, while 0 indicates background. We apply this mask element-wise to the loss, ensuring it is computed only over subject regions. This guides the U-Net to focus on learning the subject representation.

3.3. Context Guidance

At inference, we utilize a small fine-tuned T2I model $\hat{f}$ (Sec. 3.2) for generation as well as a large pre-trained T2I model F for Context Guidance. This allows us to generate images with rich contextual priors injected from F, without the computational cost of fine-tuning on the same large-scale models that are capable of producing them. We first draft a high-fidelity image I using F conditioned on the user's target prompt $\mathcal{P}_T$. We then perform DDIM inversion with null-text optimization of I using $\hat{f}$ to obtain the inverted latent $\hat{\mathbf{x}}_T$ and the optimized unconditional embeddings $\emptyset_t$ at each time-step t. We then proceed to de-noise $\hat{\mathbf{x}}_T$ with $\hat{f}$. However, instead of injecting $\emptyset_t$ at every time-step which would result in an almost perfect reconstruction of I, we halt injections after the initial denoising steps. We denote this operation as:

$$\epsilon_t = \begin{cases} \hat{f}_\theta(\mathbf{z}_t, t, \psi(P), \emptyset_t) & \text{if } \tau > T - t \\ \hat{f}_\theta(\mathbf{z}_t, t, \psi(P)) & \text{otherwise} \end{cases} \tag{5}$$

We show empirically in Fig. 3 that this operation allows the subject appearance to drift away from I towards the learned subject representation of $\hat{f}$ while maintaining the contextual priors of I.

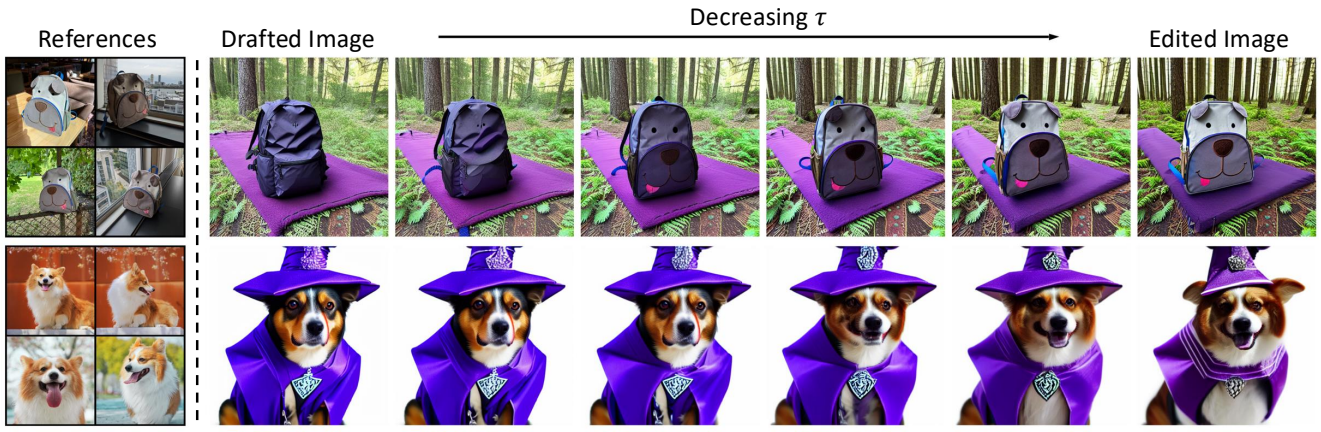

Fig. 3: **Modifying τ.** Decreasing τ will shift the subject's appearance towards the learned representation of the reference subject while still preserving the contextual priors of the drafted image. This is enabled by the Cross-Attention Guided Loss, which updates the subject token representation while minimizing the perturbation to the fine-tuned models remaining priors.

Fig. 4: **Qualitative Results**. Our generated results feature subjects which closely resemble the true subjects (bottom right).

4. EXPERIMENTS

Experiment Setup. We use Stable Diffusion V1-5 [17] as our base diffusion model. At inference, we use Stable Diffusion 3 Medium [2] to provide Context Guidance. We use GPT-4o [20] to generate captions for each reference image. Reference images are resized to 512x512 resolution, center-cropped, and normalized. We set the threshold p_t to 0.2 for our Cross-Attention Guided Loss. Fine-tuning is then performed in batches of 6 on a single A100 GPU for 100 iterations at a learning rate of 2e-5. Inference was performed with DDIM sampling [21], with a step size of 50 and a guidance scale set to 7.5. We set $\tau = 60$ for Context Guidance.

Evaluation Metrics. We evaluate Stencil on the DreamBench dataset [4], consisting of 30 subjects each represented by 4-7 reference images and tested across 25 prompts. We assess the image quality along two key dimensions: *subject consistency*, which measures how closely the generated subject resembles the true subject using DINO scores (computed as the average pairwise cosine similarity between ViT-S/16 DINO embeddings of the generated and reference images), and *text alignment*, which evaluates how accurately the generated image reflects the user prompt using CLIP-T scores (computed as the average cosine similarity between CLIP embeddings of the prompt and the generated image).

5. EXPERIMENT RESULTS

5.1. Quantitative Evaluation

Table 1 presents our quantitative evaluations. Stencil outperforms all previous methods in both text and subject fidelity.

Table 1: **Quantitative comparison on DreamBench. Bold** entries indicate the top-performing method for each evaluation metric. Experimental results are referenced from the original papers as well as [13].

Type	Method	Base Model	Subject Consistency (↑)	Text Alignment (↑)	GPU Hours (↓)
Fine-tuning Free	**ELITE**	SDv1.4	0.621	0.293	336
	BLIP-Diffusion	SDv1.5	0.594	0.300	2304
	IP-Adapter	SDXL	0.613	0.292	672
	Kosmos-G	SDv1.5	0.618	0.250	12300
	Emu2	SDXL	0.563	0.273	-
	λ-eclipse	Kv2.2	0.613	0.307	74
	SSR-Encoder	SDv1.5	0.612	0.308	-
Fine-tuning	**Textual Inversion**	SDv1.5	0.569	0.255	1
	DreamBooth	SDv1.5	0.668	0.305	0.2
	Custom Diffusion	SDv1.5	0.643	0.305	0.2
	Stencil (Ours)	SDv1.5	**0.671**	**0.328**	**0.1**

Method	Subject Consistency	Text Alignment
Stencil (Ours)	0.782	0.764
DreamBooth	0.153	0.173
Undecided	0.064	0.062

Table 2: **User Study.** We compare Stencil to DreamBooth. Values indicate user preferences in decimal values.

To our best knowledge, Stencil is the new SOTA among open-source methods. Its Context Guidance mechanism enables Stencil to noticeably outperform the others at producing semantically accurate images. Stencil is also the most cost-effective framework. Despite the added inference overhead from Context Guidance, it achieves the lowest end-to-end GPU time, thanks to the Cross-Attention Guided Loss which enables much faster fine-tuning convergence.

5.2. Qualitative Evaluation

Fig. 4 showcases images generated by Stencil. Compared to other methods, Stencil excels at generating diverse layouts while maintaining subject fidelity. We attribute this to Context Guidance, which introduces unseen image structures into the generation process. Table 2 presents results from our user study consisting of 30 participants, where Stencil decisively outperforms DreamBooth in both text alignment and subject consistency.

5.3. Discussions and Limitations

A detailed discussion of Stencil's applications and limitations are provided in our *Supplementary Materials*. Stencil supports a wide range of applications, including age progression/regression, expression editing, accessorization, perspective-conditioned generation, pose manipulation, and style transfer.

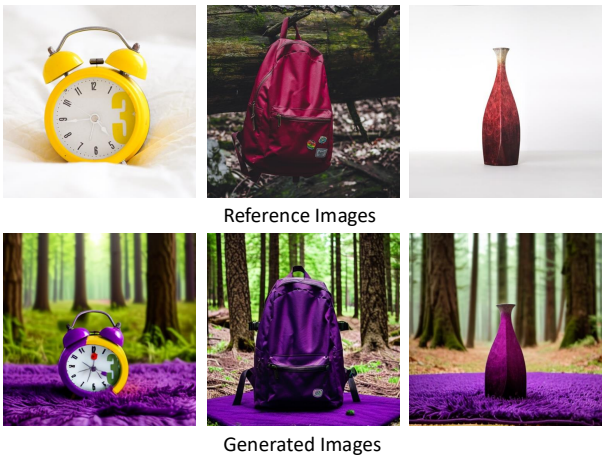

Fig. 5: **Failure Case.** Stencil can sometimes inherit the limitations of its base diffusion model.

As illustrated in Fig. 5, Stencil inherits certain limitations from its underlying diffusion models, most notably the tendency for local features to be applied globally. Additionally, some subjects (e.g., animals) are easier to learn than others (e.g., human faces), due to differences in training distribution.

6. CONCLUSION

In this paper, we introduced Stencil, an efficient fine-tuning approach for subject-driven generation. Stencil incorporates two key innovations: first, the Cross-Attention Guided Loss function that directs the network's learning towards subject pixels, enabling faster and more stable convergence; and second, Context Guidance, where we inject rich contextual priors from a large pre-trained diffusion model into the generation process. Experimental results further validate Stencil's robustness and state-of-the-art performance. We hope our work inspires future research in subject-driven generation.

7. REFERENCES

[1] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al., "Photorealistic text-to-image diffusion models with deep language understanding," *NeurIPS*, 2022.

[2] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al., "Scaling rectified flow transformers for high-resolution image synthesis," in *International Conference on Machine Learning*, 2024.

[3] 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," *arXiv preprint arXiv:2307.01952*, 2023.

[4] 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 *CVPR*, 2023.

[5] 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.

[6] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu, "Multi-concept customization of text-to-image diffusion," in *CVPR*, 2023.

[7] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski, "Break-a-scene: Extracting multiple concepts from a single image," in *SIGGRAPH Asia 2023 Conference Papers*, 2023.

[8] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu, "Reversion: Diffusion-based relation inversion from images," in *SIGGRAPH Asia 2024 Conference Papers*, 2024.

[9] Maxwell Jones, Sheng-Yu Wang, Nupur Kumari, David Bau, and Jun-Yan Zhu, "Customizing text-to-image models with a single image pair," in *SIGGRAPH Asia 2024 Conference Papers*, 2024.

[10] Dongxu Li, Junnan Li, and Steven Hoi, "Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing," *NeurIPS*, 2024.

[11] Yuxuan Zhang, Yiren Song, Jiaming Liu, Rui Wang, Jinpeng Yu, Hao Tang, Huaxia Li, Xu Tang, Yao Hu, Han Pan, et al., "Ssr-encoder: Encoding selective subject representation for subject-driven generation," in *CVPR*, 2024.

[12] Yuxiang Wei, Yabo Zhang, Zhilong Ji, Jinfeng Bai, Lei Zhang, and Wangmeng Zuo, "Elite: Encoding visual concepts into textual embeddings for customized text-to-image generation," in *CVPR*, 2023.

[13] Maitreya Patel, Sangmin Jung, Chitta Baral, and Yezhou Yang, "lambda-eclipse: Multi-concept personalized text-to-image diffusion models by leveraging clip latent space," *arXiv preprint arXiv:2402.05195*, 2024.

[14] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang, "Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models," *arXiv preprint arXiv:2308.06721*, 2023.

[15] Moab Arar, Rinon Gal, Yuval Atzmon, Gal Chechik, Daniel Cohen-Or, Ariel Shamir, and Amit H. Bermano, "Domain-agnostic tuning-encoder for fast personalization of text-to-image models," in *SIGGRAPH Asia 2023 Conference Papers*, 2023.

[16] Rinon Gal, Or Lichter, Elad Richardson, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or, "Lcm-lookahead for encoder-based text-to-image personalization," in *European Conference on Computer Vision*. Springer, 2024, pp. 322–340.

[17] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer, "High-resolution image synthesis with latent diffusion models," in *CVPR*, 2022.

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or, "Prompt-to-prompt image editing with cross attention control," *arXiv preprint arXiv:2208.01626*, 2022.

[19] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or, "Null-text inversion for editing real images using guided diffusion models," in *CVPR*, 2023.

[20] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al., "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[21] Jiaming Song, Chenlin Meng, and Stefano Ermon, "Denoising diffusion implicit models," *arXiv preprint arXiv:2010.02502*, 2020.