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]A*STAR

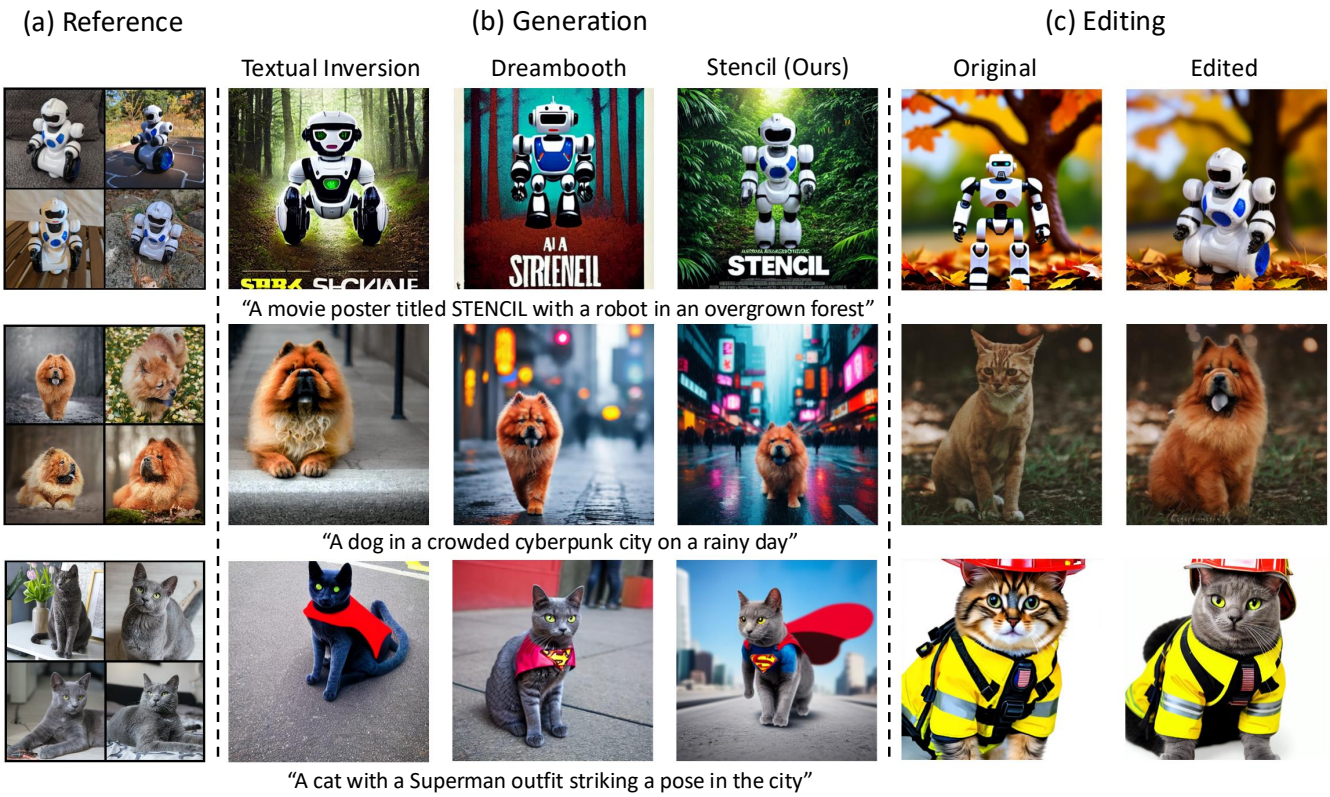

Fig. 1: **Overview of *Stencil***. Given a few (a) reference images, Stencil achieves (b) subject-driven generation and (c) subject editing with high textual and subject fidelity in just 100 fine-tuning steps.

ABSTRACT

The emergence of text-to-image diffusion models marked a revolutionary breakthrough in generative AI. However, training a text-to-image model to consistently reproduce the same subject remains a challenging task. Existing methods often require costly setups, lengthy fine-tuning processes and struggle to generate diverse, text-aligned images. Moreover, the increasing size of diffusion models over the years highlights a scalability challenge for previous fine-tuning methods, as tuning on these large models is even more costly. To address these limitations, we present Stencil. Stencil leverages a large diffusion model to contextually guide a smaller fine-tuned model during generation. This allows us to combine the superior generalization capabilities of large models with the efficient fine-tuning of small models. Stencil excels at generating high-fidelity, novel renditions of the subject and can do so in just 30 seconds, nearly $\times$20 faster than DreamBooth, delivering state-of-the-art performance and setting a new benchmark in subject-driven generation.

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. Subject-driven generation builds on T2I models to enable customization of generated subject characteristics, and remains an active area of research.

The emergence of transformer-based T2I diffusion models [2] marks a shift towards increasingly larger architectures. This has made previous fine-tuning methods [3, 4, 5] for

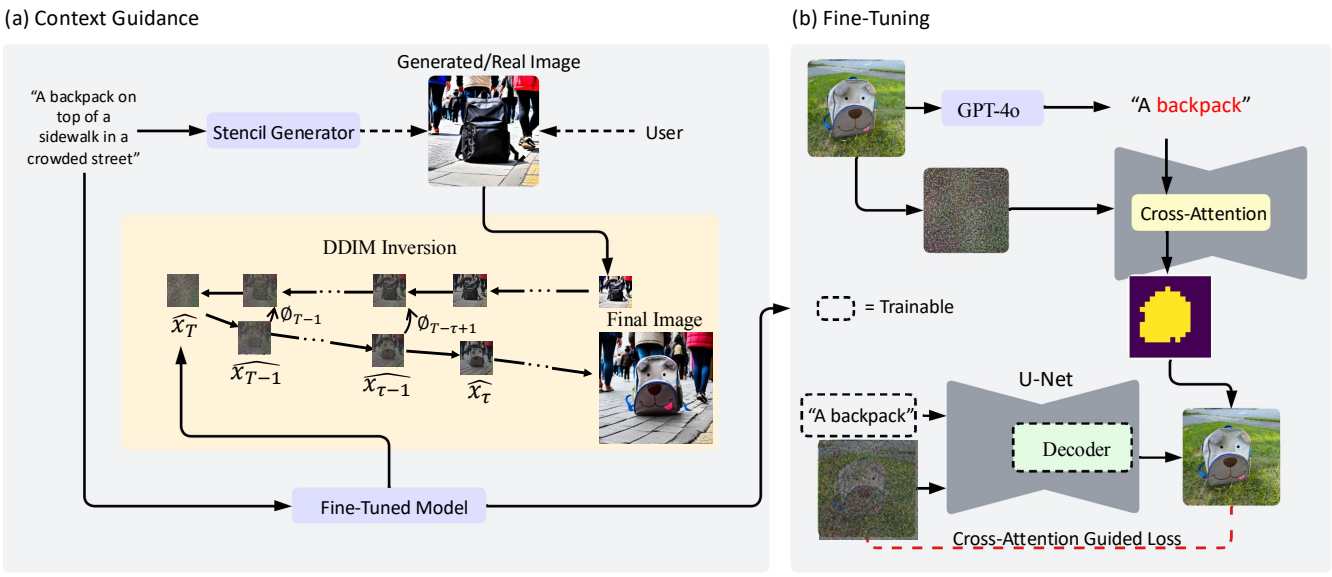

Fig. 2: **Stencil Framework.** In (a) Context Guidance, we take an input image either generated by a large pre-trained model (Stencil Generator) or provided by the user for generation and editing tasks, respectively. The image is inverted and fed into a smaller fine-tuned model to refine subject representations in alignment with reference images. In (b) Fine-Tuning, we obtain the cross-attention map for the subject token via a single denoising step of the noisy latent of the reference images. The cross-attention map is applied to the loss function to guide the U-Net to focus precisely on learning pixels where the subject is present.

subject-driven generation, designed for smaller U-Net architectures, impractical for use in the latest models due to their prohibitive computational costs. This underscores the need for more efficient subject-drive generation techniques to harness the generation capabilities of the latest T2I models.

To address these challenges, we propose Stencil. Stencil employs context guidance, where we use a large pre-trained diffusion model to guide the generation process of a smaller fine-tuned model during inference. This approach enables the generation of diverse and high-quality images comparable to those from large diffusion models without the significant cost of training them. During fine-tuning, Stencil employs the Cross-Attention Guided Loss Function to mask the loss of irrelevant background and foreground elements. This approach simplifies optimization and reduces fine-tuning by guiding the model to focus on relevant subject pixels instead of processing reference images uniformly. Stencil supports image editing (Fig. 1) and achieves state-of-the-art (SOTA) results in generation while being $\times 20$ faster than DreamBooth, making it an incredibly cheap and effective model. We summarize our main contributions as follows:

- We propose Stencil, a novel fine-tuning method for subject-driven generation. Stencil uses context guidance to achieve high textual and subject fidelity at low costs.
- We propose the Cross-Attention Guided Loss, where we apply a mask to the loss function to guide the model to learn subject representation from the most relevant pixels.
- Our extensive experiments have validated the robustness of our approach, achieving state-of-the-art results.

2. RELATED WORKS

Recent methods for subject-driven generation can be divided into two camps - those that fine-tune the diffusion model on the reference images during test-time [3, 6, 4, 5], and those that train an additional structure to encode the reference images [7, 8, 9, 10, 11]. In this paper, we focus on the former. Textual Inversion [6] optimizes token embeddings within text prompts to better capture subject representation. DreamBooth [3] fine-tunes the diffusion U-Net to bind the appearance of a subject with specific class tokens. Custom Diffusion [4] proposes to enhance efficiency by limiting fine-tuning to the cross-attention layers of the U-Net. Despite these advancements, we observe a problematic trend: as we shift towards larger model architectures [2, 12], applying these fine-tuning methods on the latest models become increasingly time consuming and computationally inefficient. To address this issue, we propose context guidance (Sec. 3.4). Furthermore, we observe that existing methods optimize every pixel of the reference image, including irrelevant foreground and background details, which can complicate the optimization task. To overcome this, we introduce the Cross-Attention Guided Loss Function (Sec. 3.3).

3. METHOD

Fig. 2 provides an illustration of our method framework. Stencil consists of a fine-tuning stage (Sec. 3.3) followed by context guidance during inference (Sec. 3.4). Additional fine-tuning techniques are discussed in Sec. 3.2.

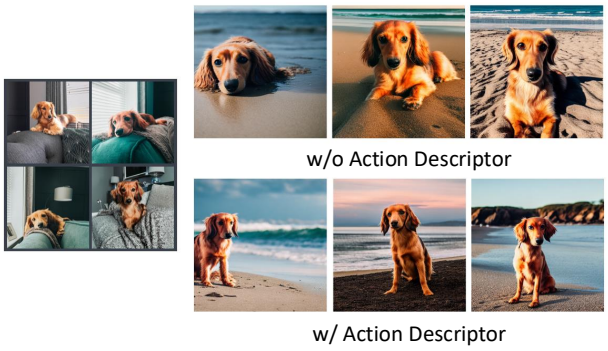

Fig. 3: **Promoting Diversity with Action Descriptors.** Vanilla fine-tuning can result in over-fitting to the input image layout. Without using action descriptors such as 'sitting' or 'lying down' during fine-tuning, the dog above retains the same pose as in the reference images. Action descriptors prevent the model from binding the subject token to a specific layout, encouraging diversity.

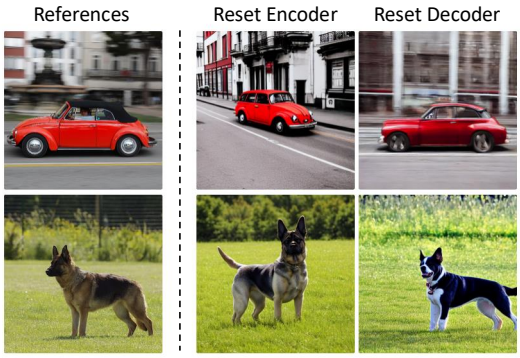

Fig. 4: **Understanding Encoder and Decoder Learning.** We fine-tune the entire U-Net on a single reference image. Subsequently, we reset either the encoder or the decoder by replacing their parameters with the pre-trained ones. We observed that resetting the encoder preserves the object's appearance but causes a loss of layout, whereas resetting the decoder preserves the layout but loses fine-grain image details.

3.1. Preliminaries

Text-to-Image Latent Diffusion Models. Diffusion models consist of a forward and reverse diffusion process. During the forward diffusion process, Gaussian noise $\epsilon_t \sim \mathcal{N}(0,1)$ is iteratively applied to the original image x_0 over t time-steps to convert an image into pure noise. Each intermediate sample x_t, where $t \in \{0,\ldots,T\}$, satisfies:

$$x_t = \sqrt{\alpha_t}\, x_0 + \sqrt{1-\alpha_t}\, \epsilon \tag{1}$$

where $0 = \alpha_T < \cdots < \alpha_0 = 1$ are hyper-parameters of the diffusion schedule. The reverse diffusion process tries to remove the noise that was added in the forward process by training a denoising U-Net network $f_\theta(x_t, t, \psi(P))$, conditioned on the text embedding $\psi(P)$, to predict the noise residual ϵ_t added to the sample at time-step $t-1$.

Latent diffusion models [13] reduce compute complexity by applying the diffusion process on a lower-dimensional latent space z_t. The overall loss is computed as:

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

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:

$$M = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right) \tag{3}$$

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. The attention output $\widehat{\phi}(z_t) = MV$, which is a weighted average of the values of V, is used to update the spatial features $\phi(z_t)$ that are propagated to the subsequent layers of the U-Net.

3.2. Decoupling Layout and Appearance

We observe that poor text-image alignment can easily cause the model to over-fit to the layout of the reference images during fine-tuning. Specifically, we find that the prompt, 'A [subject token] [action descriptor]' is a lot less prone to over-fitting than 'A [subject token]' (Fig. 3). We conclude that this is due to language drift, where a prompt without an action descriptor can cause the model to bind the subject tokens to both the subject's appearance and its layout. Action descriptors can help disentangle appearance and layout by associating them to two or more seperate tokens, hence enhancing generalization capabilities. Thus, we propose to leverage a vision-language model (VLM) to generate captions P that adhere to this format. This approach also removes the reliance on class identifier tokens, as used in DreamBooth [3], thereby speeding up fine-tuning by eliminating the need to train on rare tokens and their semantic representations.

Additionally, we show that as spatial features propagate through the U-Net, higher-frequency information is captured. The shallower layers learn the structure, whereas the deeper layers learn the finer appearances of the image (Fig.4). Since subject-driven generation concerns the latter, we propose to exclusively fine-tune the U-Net decoder blocks. This significantly reduces the parameters that require fine-tuning.

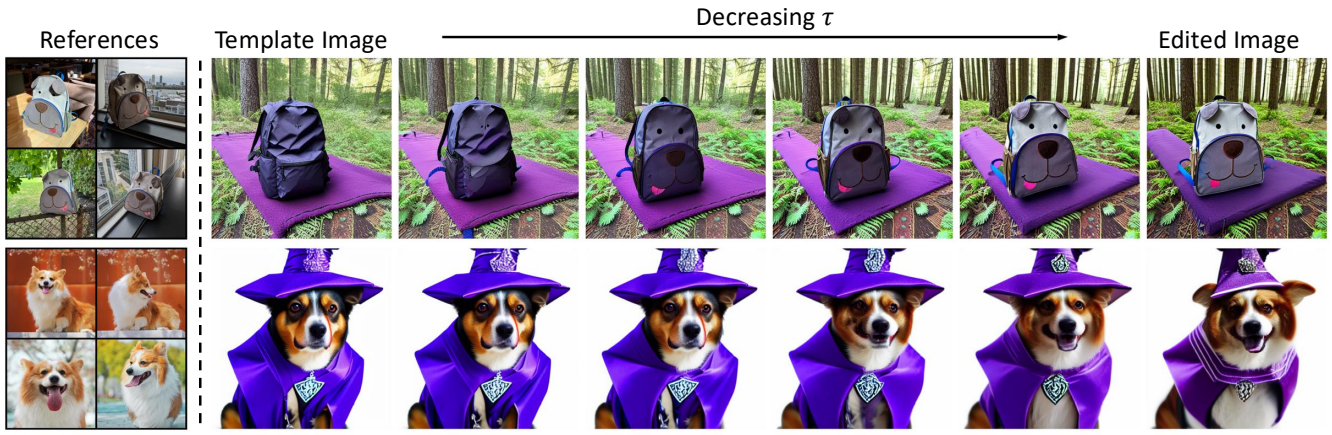

Fig. 5: **Modifying τ** We can control deviation from the template image by changing τ. Decreasing τ will allow the subject to drift away and gradually align more closely with the reference images. Increasing τ preserves more template features.

3.3. Cross-Attention Guided Loss Function

Fine-tuning images uniformly can complicate the optimization task and lead to slow convergence. To address this, we leverage the cross-attention map of the subject token S to guide the U-Net to focus on learning the subject. Specifically, we add t time-step noise to the reference images. We then perform a single forward pass of the noisy latent, conditioned on P (See Sec.3.2), through the fixed U-Net backbone. During this forward pass, we save the cross-attention map of S across all heads and layers. We then up-sample all the them to the spatial resolution of the latent image, compute the mean values, and normalize them to get an average cross-attention map, $\widehat{M}_S$. We define our new loss function as,

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

where $1_{\widehat{M}_S > p_t}$ denotes a binary mask obtained from the cross-attention map and threshold hyper-parameter p_t. Hence, the loss is applied only to the subject, rather than the entire image. Effectively, we are guiding the U-Net to learn the subject representation by telling it what to focus on.

3.4. Context Guidance

For inference, we utilize two seperate diffusion models to perform subject-driven generation: a large pre-trained model (stencil generator) as well as a smaller fine-tuned model (See Sec. 3.3). The stencil generator produces high-fidelity template images, while the fine-tuned model refines the subject representations of those images. This allows us to leverage the efficiency of fine-tuning smaller models and the high-quality image generation of larger SOTA diffusion models.

Initially, the stencil generator f_θ generates a template image I conditioned on the target prompt $\mathcal{P}_T$. We then perform null-text inversion [14] of I using the fine-tuned model $\hat{f}_\theta$ 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 denoise $\hat{\mathbf{x}}_t$ with $\hat{f}_\theta$. However, instead of injecting $\emptyset_t$ at every time-step which would result in an almost perfect reconstruction of I, we halt at time-step τ. We denote this operation as

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

We show empirically in Fig. 5 that this allows the subject appearance to drift away from the template subject and towards the reference subject while maintaining faithfulness to the original image.

4. EXPERIMENTS

4.1. Experiment Setup

We use Stable Diffusion V1-5 [13] as our base diffusion model, Stable Diffusion 3 Medium [2] as our stencil generator and GPT-4o [15] as our VLM. Reference images are resized to 512x512 resolution, center-cropped, and normalized. For the cross-attention-guided loss function, we set the threshold p_t to 0.2. 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 [16], with a step size of 50 and a guidance scale set to 7.5. For context guidance, we set τ to 3.

4.2. Evaluation Metrics

We evaluate our model on the DreamBench dataset [3], consisting of 30 subjects each represented by 4-7 reference images. Each subject is associated with 25 prompts. To assess the subject consistency, we compute the DINO scores, corresponding to the average pairwise cosine similarities between the ViT-S/16 DINO embeddings of generated and real images. To assess text-to-image alignment, we compute the

Fig. 6: **Qualitative Results**. Our results demonstrate close-to-perfect faithfulness to the references (bottom right corner)

Table 1: **Quantitative comparison on Dreambench.** The **Bold** and Underline represent first and second-ranked methods.

Type	Method	Base Model	Subject Consistency (↑)	Text Alignment (↑)
Fine-tuning	**Textual Inversion** [7]	SDv1.5	0.569	0.255
	DreamBooth [3]	SDv1.5	<u>0.668</u>	0.305
	Custom Diffusion [4]	SDv1.5	0.643	0.305
	Stencil (Ours)	SDv1.5	**0.671**	**0.328**
Fine-tuning Free	**ELITE** [9]	SDv1.4	0.621	0.293
	BLIP-Diffusion [7]	SDv1.5	0.594	0.300
	IP-Adapter [11]	SDXL	0.613	0.292
	Kosmos-G [10]	SDv1.5	0.618	0.250
	Emu2 [10]	SDXL	0.563	0.273
	λ-eclipse [10]	Kv2.2	0.613	0.307
	SSR-Encoder [8]	SDv1.5	0.612	<u>0.308</u>

CLIP-T scores, corresponding to the average cosine similarity between prompt and image CLIP embeddings.

Method	Text-Alignment	Subject Consistency
Stencil	0.764	0.782
DreamBooth	0.173	0.153
Undecided	0.062	0.064

Table 2: User Study comparing Stencil to Dreambooth

5. EXPERIMENT RESULTS

5.1. Quantitative Evaluation

Table. 1 presents our quantitative evaluations. Stencil outperforms all previous methods in both subject fidelity and text-to-image alignment while being the most cost-effective model to train. To our best knowledge, Stencil is the new SOTA. Notably, Stencil performs significantly better than the rest at producing semantically accurate images. This further validates context guidance as a cheap and effective technique to enhance generation capabilities of base models.

5.2. Qualitative Evaluation

Fig. 6 showcases images generated by Stencil. Compared to other methods, Stencil excels at maintaining the subject's appearance and generating diverse layouts. This is because we do not train the stencil generator on the reference images, enabling it to generate completely unseen image structures. Table. 2 presents results from our User Study.

6. CONCLUSION

In this paper, we introduced Stencil, an efficient fine-tuning approach for subject-driven generation. Stencil incorporates two key innovations: first, a cross-attention guided loss function that directs the network's learning toward the subject, enabling faster convergence; and second, context guidance, where a large pre-trained diffusion model generates an image template that a smaller fine-tuned model refines to achieve precise subject alignment. This method also makes Stencil the only fine-tuning method that is scalable, as it can continually benefit from advancements in T2I diffusion models to further enhance image quality and text-to-image alignment at no additional cost.

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] 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.

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

[5] 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.

[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.

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

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

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

[14] 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.

[15] 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.

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