SSR-Encoder: Encoding Selective Subject Representation for Subject-Driven Generation

Yuxuan Zhang[1*], Yiren Song[4], Jiaming Liu[2†], Rui Wang[3*], Jinpeng Yu[6*], Hao Tang[5], Huaxia Li[2], Xu Tang[2], Yao Hu[2], Han Pan[1], Zhongliang Jing[1†]

[1]Shanghai Jiao Tong University, [2]Xiaohongshu Inc., [3]Beijing University of Posts and Telecommunications, [4]National University of Singapore, [5]Carnegie Mellon University, [6]ShanghaiTech University

{zyx153,hanpan,zljing}@sjtu.edu.cn, jmliu1217@gmail.com,
{yicheng,tangshen,xiahou}@xiaohongshu.com, wr_bupt@bupt.edu.cn,
yiren@nus.edu.sg, haotang2@cmu.edu, yujp1@shanghaitech.edu.cn

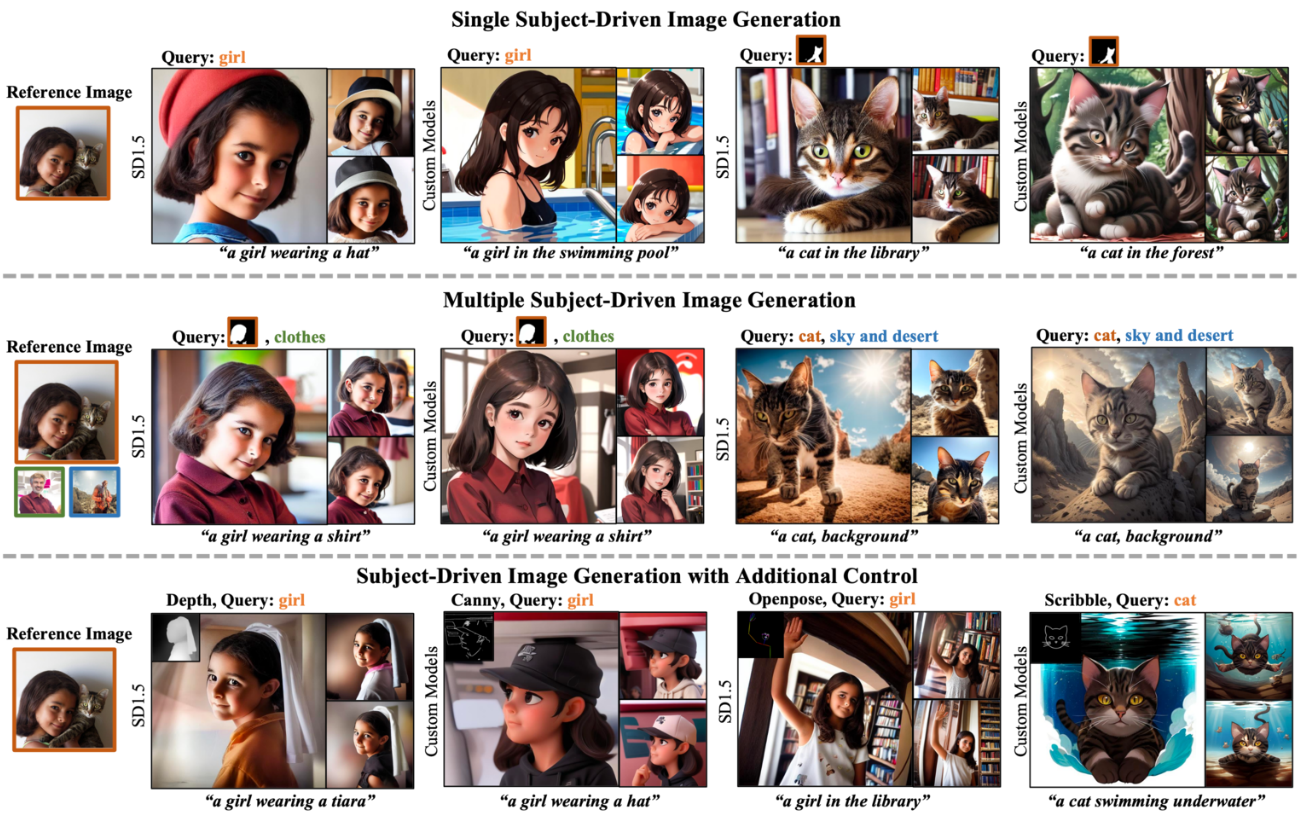

Figure 1. Our *SSR-Encoder* is a model generalizable encoder, which is able to guide any customized diffusion models for single subject-driven image generation (top branch) or multiple subject-driven image generation from different images (middle branch) based on the image representation selected by the text query or mask query without any additional test-time finetuning. Furthermore, our *SSR-Encoder* can also be applied for the controllable generation with additional control (bottom branch).

Abstract

Recent advancements in subject-driven image generation have led to zero-shot generation, yet precise selection and focus on crucial subject representations remain challenging. Addressing this, we introduce the SSR-Encoder, a novel architecture designed for selectively capturing any subject from single or multiple reference images. It responds to various query modalities including text and masks, without necessitating test-time fine-tuning. The SSR-

*Work done during internship at Xiaohongshu Inc.
†Corresponding authors.

Encoder combines a Token-to-Patch Aligner that aligns query inputs with image patches and a Detail-Preserving Subject Encoder for extracting and preserving fine features of the subjects, thereby generating subject embeddings. These embeddings, used in conjunction with original text embeddings, condition the generation process. Characterized by its model generalizability and efficiency, the SSR-Encoder adapts to a range of custom models and control modules. Enhanced by the Embedding Consistency Regularization Loss for improved training, our extensive experiments demonstrate its effectiveness in versatile and high-quality image generation, indicating its broad applicability. Project page: `ssr-encoder.github.io`

1. Introduction

Recent advancements in image generation, especially with the advent of text-to-image diffusion models trained on extensive datasets, have revolutionized this field. A prime example is Stable Diffusion, an open-source model cited as [31], which allows a broad user base to easily generate images from textual prompts. A growing area of interest that has emerged is the subject-driven generation, where the focus shifts from creating a generic subject, like "a cat" to generating a specific instance, such as "the cat". However, crafting the perfect text prompt to generate the desired subject content poses a significant challenge. Consequently, researchers are exploring various strategies for effective subject-driven generation.

Subject-driven image generation aims to learn subjects from reference images and generate images aligning with specific concepts like identity and style. Currently, one prominent approach involves test-time fine-tuning [1, 10, 17, 32], which, while efficient, requires substantial computational resources to learn each new subject. Another approach [7, 16, 30, 36, 39] encodes the reference image into an image embedding to bypass the fine-tuning cost. However, these encoder-based models typically require joint training with the base diffusion model, limiting their generalizability. A concurrent work, IP-adapter [42], tackles both fine-tuning costs and generalizability by learning a projection to inject image information into the U-Net, avoiding the need to fine-tune the base text-to-image model, thereby broadening its application in personalized models.

Despite these advancements, a critical aspect often overlooked is the extraction of the most informative representation of a subject. With images being a complex mixture of subjects, backgrounds, and styles, it's vital to focus on the most crucial elements to represent a subject effectively. To address this, we introduce the *SSR-Encoder*, an image encoder that generates **S**elective **S**ubject **R**epresentations for subject-driven image generation.

Our *SSR-Encoder* firstly aligns patch-level visual embeddings with texts in a learnable manner, capturing detailed subject embeddings guided by token-to-patch attention maps. Furthermore, we propose subject-conditioned generation, which utilizes trainable copies of cross-attention layers to inject multi-scale subject information. A novel Embedding Consistency Regularization Loss is proposed to enhance the alignment between text queries and visual representations in our subject embedding space during training. This approach not only ensures effective token-to-patch alignment but also allows for flexible subject selection through text and mask queries during inference. Our *SSR-Encoder* can be seamlessly integrated into any customized stable diffusion models without extensive fine-tuning. Moreover, the *SSR-Encoder* is adaptable for controllable generation with various additional controls, as illustrated in Fig. 1.

We summarize our main contributions as follows:

- We propose a novel framework, termed as *SSR-Encoder*, for selective subject-driven image generation. It allows selective single- or multiple-subject generation, fully compatible with ControlNets (e.g. canny, OpenPose, etc.), and customized stable diffusion models without extra test-time training.
- Token-to-Patch Aligner and Detail-Preserved Subject Encoder are proposed in our *SSR-Encoder* to learn selective subject embedding. We also present an Embedding Consistency Regularization Loss to enhance token-to-patch text-image alignment in the subject embedding space.
- Our extensive experiments have validated the robustness and flexibility of our approach, showcasing its capability to deliver state-of-the-art (SOTA) results among finetuning-free methods. Impressively, it also demonstrates competitive performance when compared with finetuning-based methods.

2. Related Work

Text-to-image diffusion models. In recent years, text-to-image diffusion models [2, 25, 26, 28, 29, 31, 33, 34, 40, 43] have made remarkable progress, particularly with the advent of diffusion models, which have propelled text-to-image generation to large-scale commercialization. DALLE [28] first achieved stunning image generation results using an autoregressive model. Subsequently, DALLE2 [29] employed a diffusion model as the generative model, further enhancing text-to-image synthesis ability. Imagen [33] and Stable Diffusion [31] trained diffusion models on larger datasets, further advancing the development of diffusion models and becoming the mainstream for image generation large models. DeepFloyd IF [34] utilized a triple-cascade diffusion model, significantly enhancing the text-to-image generation capability, and even generating correct fonts. Stable Diffusion XL [26], a two-stage cascade diffusion model, is the latest optimized version of stable diffusion, greatly improving the generation of high-frequency details, small object features, and overall image color.

Controllable image generation. Current diffusion models can incorporate additional modules, enabling image generation guided by multimodal image information such as edges, depth maps, and segmentation maps. These multimodal inputs significantly enhance the controllability of the diffusion model's image generation process. Methods like ControlNet [44] utilize a duplicate U-Net structure with trainable parameters while keeping the original U-Net parameters static, facilitating controllable generation with other modal information. T2I-adapter [24] employs a lightweight adapter for controlling layout and style using different modal images. Uni-ControlNet [46] differentiates between local and global control conditions, employing separate modules for injecting these control inputs. Paint by Example [41] allows for specific region editing based on reference images. Other methods [3, 5, 9, 13, 23, 45] manipulate the attention layer in the diffusion model's denoising U-Net to direct the generation process. P2P [13] and Null Text Inversion [23] adjust cross-attention maps to preserve image layout under varying text prompts.

Subject-driven image generation. Subject-driven image generation methods generally fall into two categories: those requiring test-time finetuning and those that do not. The differences in characteristics of these methods are illustrated in Table 1. Test-time finetuning methods [1, 6, 10–12, 15, 17, 32, 38] often optimize additional text embeddings or directly fine-tune the model to fit the desired subject. For instance, Textual Inversion [10] optimizes additional text embeddings, whereas DreamBooth [32] adjusts the entire U-Net in the diffusion model. Other methods like Customdiffusion [17] and SVDiff [12] minimize the parameters needing finetuning, reducing computational demands. Finetuning-free methods [16, 18, 21, 30, 36, 37, 39, 42] typically train an additional structure to encode the reference image into embeddings or image prompts without additional finetuning. ELITE [39] proposes global and local mapping training schemes to generate subject-driven images but lack fidelity. Instantbooth [36] proposes an adapter structure inserted in the U-Net and trained on domain-specific data to achieve domain-specific subject-driven image generation without finetuning. IP-adapter [42] encodes images into prompts for subject-driven generation. BLIP-Diffusion [18] enables efficient finetuning or zero-shot setups. However, many of these methods either utilize all information from a single image, leading to ambiguous subject representation, or require finetuning, limiting generalizability and increasing time consumption. In contrast, our SSR-Encoder is both generalizable and efficient, guiding any customized diffusion model to generate images based on the representations selected by query inputs without any test-time finetuning.

Table 1. Comparative Analysis of Previous works. Considering Fine-Tuning free, Model Generalizability, and Selective Representation, SSR-Encoder is the first method offering all three features.

Method	Finetuning -free	Model Generalizable	Selective Representation
Textual Inversion [10]	✗	✓	✗
Dreambooth [32]	✗	✗	✗
LoRA [15]	✗	✓	✗
Custom diffusion [17]	✗	✗	✗
Break-A-Scene [1]	✗	✗	✓
E4T [30]	✗	✗	✗
Instantbooth [36]	✓	✗	✗
ELITE [39]	✓	✗	✗
Taming [16]	✓	✗	✗
IP-adapter [42]	✓	✓	✗
BLIP-diffusion [18]	✓	✗	✓
SSR-Encoder(Ours)	✓	✓	✓

3. The Proposed Method

Selective subject-driven image generation aims to generate target subjects in a reference image with high fidelity and creative editability, guided by the user's specific queries (text or mask). To tackle this, we propose our SSR-Encoder, a specialized framework designed to integrate with any custom diffusion model without necessitating test-time fine-tuning.

Formally, for a given reference image I and a user query q, the SSR-Encoder effectively captures subject-specific information and generates multi-scale subject embeddings c_s. These multi-scale subject embeddings c_s are subsequently integrated into the U-Net model with trainable copies of cross-attention layers. The generation process, conditioned on both subject embeddings c_s and text embedding c_t, allows for the production of desired subjects with high fidelity and creative editability. The overall methodology is illustrated in Fig. 2.

In general, SSR-Encoder is built on text-to-image diffusion models [31][1]. It comprises two key components: the token-to-patch aligner and detail-preserving subject encoder (Sec. 3.1). The subject-conditioned generation process is detailed in Sec. 3.2. Lastly, training strategies and loss functions are presented in Sec. 3.3.

3.1. Selective Subject Representation Encoder

Our Selective Subject Representation Encoder (SSR-Encoder) is composed of two integral parts: Token-to-Patch Aligner and Detail-Preserving Subject Encoder. The details of each component are as follows.

Token-to-patch aligner. Several works [8, 20, 47] have pointed out that CLIP tends to prioritize background regions

[1]Reviewed in the Supplementary.

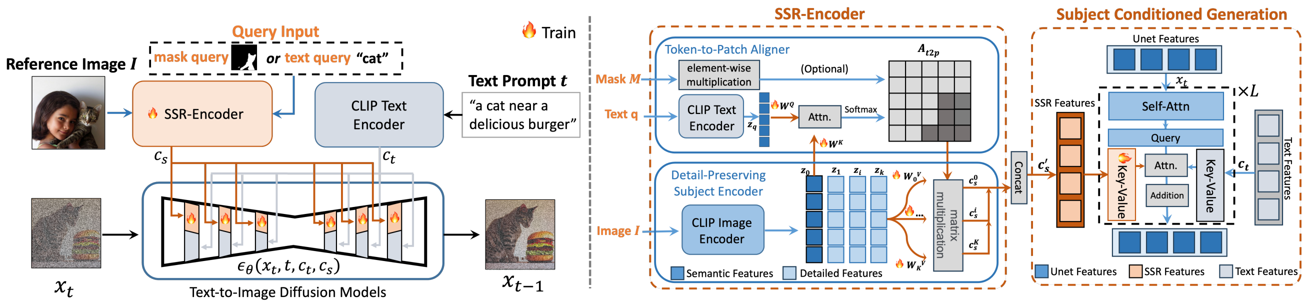

Figure 2. Overall schematics of our method. Given a query text-image pairs (q, I), the SSR-Encoder employs a token-to-patch aligner to highlight the selective regions in the reference image by the query. It extracts more fine-grained details of the subject through the detail-preserving subject encoder, projecting multi-scale visual embeddings via the token-to-patch aligner. Then, we adopt subject-conditioned generation to generate specific subjects with high fidelity and creative editability. During training, we adopt reconstruction loss L_{LDM} and embedding consistency regularization loss L_{reg} for selective subject-driven learning.

over foreground subjects when identifying target categories. Therefore, relying solely on text-image similarity may not adequately capture subject-specific information. To address this issue, we propose the Token-to-Patch (T2P) Aligner, which implements two trainable linear projections to align image patch features with given text token features. Mathematically, given a query text-image pair (q, I), we employ pre-trained CLIP encoders to generate the text query and image reference into query embedding $z_q \in \mathbb{R}^{N_q \times D_q}$ and semantic visual embedding $z_0 \in \mathbb{R}^{N_i \times D_i}$ from the last CLIP layer, respectively, where $N_{(\cdot)}$ and $D_{(\cdot)}$ represent the number of tokens and dimensions for query and image features respectively. We then use the trainable projection layers $\mathbf{W^Q}$ and $\mathbf{W^K}$ to transform them into a well-aligned space. The alignment is illustrated as follows:

$$\begin{aligned} Q &= \mathbf{W^Q} \cdot z_q, \\ K &= \mathbf{W^K} \cdot z_0, \end{aligned} \tag{1}$$

$$A_{t2p} = \text{Softmax}\left(\frac{QK^\top}{\sqrt{d}}\right), \tag{2}$$

where $A_{t2p} \in \mathbb{R}^{N_t \times N_i}$ represents the token-to-patch attention map.

Furthermore, the A_{t2p} matrix serves a dual purpose: similarity identification and region selection. Consequently, our aligner naturally supports **mask-based query**. In practice, we can manually assign a mask M to A_{t2p} for mask-guided generation with null-text query inputs. Following Eq. (2), we can proceed to reweight A_{t2p} using the predefined mask M to highlight selected regions, ensuring our SSR-Encoder focuses solely on the selected valid regions of reference images.

Detail-preserving subject encoder. Following most of the preceding methods [19, 39, 42], we employ a pre-trained CLIP visual backbone to extract image representations from reference images. However, the conventional practice of extracting visual embeddings z_0 from the last CLIP layer does not align with our objective of preserving fine details to the maximum extent. Our preliminary experiments[2] have identified a notable loss of fine-grained details in the semantic image features z_0. Addressing this, we introduce the detail-preserving subject encoder, which extracts features across various layers to preserve more fine-grained details. Formally, the visual backbone processes an image I to produce multi-scale detailed image features $z_I = \{z_k\}_{k=0}^{K}$, where z_0 represents semantic visual embedding used in T2P aligner and z_k represents other detailed visual embeddings at the scale of k in CLIP visual backbone and K refers to the number of target scales. We set K to 6 in all experimental settings.

To fully leverage the benefits of multi-scale representation, we adopt separate linear projections $\mathbf{W_k^V}$ for image feature z_k at different scales. Combining with the token-to-patch attention map A_{t2p}, the subject embeddings $c_s = \{c_s^k\}_{k=0}^{K}$ are computed as per Eq. (3):

$$V_k = \mathbf{W_k^V} \cdot z_k, c_s^k = A_{t2p} \, V_k^\top, \tag{3}$$

where c_s^k denotes subject embedding at scale of k. Our SSR-Encoder now enables to capture multi-scale subject representation $c_s = \{c_s^k\}_{k=0}^{K}$, which are subsequently used for subject-driven image generation via subject-conditioned generation process.

3.2. Subject Conditioned Generation

In our approach, c_s is strategically projected into the cross-attention layers of the U-Net. This is achieved through newly added parallel subject cross-attention layers, each corresponding to a text cross-attention layer in the original U-Net. Rather than disturbing the text embedding c_t, these new layers independently aggregate subject embeddings c_s. Inspired by works like [17, 39, 42, 44], we employ trainable copies of the text cross-attention layers to preserve the efficacy of the original model. The key and value projection layers are then adapted to train specifically for a subject-conditioned generation. To full exploit of both global and local subject representation, we concatenate all c_s^k at the token dimension before projection, i.e. $c_s' = \text{concat}\left(c_s^k, \text{dim} = 0\right)$, where $c_s^k \in \mathbb{R}^{N_q \times D_i}$ represents subject representation at the scale of k. The output

[2]Detailed in the supplementary.

value O of the attention layer is formulated as follows:

$$
\begin{aligned}
O = &\underbrace{\text{CrossAttention}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}, c_t, x_t\right)}_{\text{text condition}} \\
&+ \lambda \underbrace{\text{CrossAttention}\left(\mathbf{Q}, \mathbf{K_S}, \mathbf{V_S}, c_s', x_t\right)}_{\text{subject condition}},
\end{aligned}
\tag{4}
$$

where c_t represents the text embedding and x_t represents the latent. $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ represents query, key, and value projection layers in the original text branch respectively while $\mathbf{K_S}, \mathbf{V_S}$ represents trainable copies of key and value projection layers for concatenated subject embedding c_s. λ is a weight adjustment factor, with a default value of 1.

By our subject-conditioned generation, text-to-image diffusion models can generate target subjects conditioned on both text embeddings and subject embeddings.

3.3. Model Training and Inference

During the training phase, our model processes paired images and texts from multimodal datasets. The trainable components include the token-to-patch aligner and the subject cross-attention layers.

In contrast to CLIP, which aligns global image features with global text features, our token-to-patch aligner demands a more granular token-to-patch alignment. To achieve this, we introduce an Embedding Consistency Regularization Loss L_{reg}. This loss is designed to enhance similarity between the subject embeddings c_s and the corresponding query text embedding z_q, employing a cosine similarity function as demonstrated in Eq. (5):

$$
\begin{aligned}
\overline{c_s} &= \text{Mean}\left(c_s^0, c_s^1, ..., c_s^K\right), \\
\mathcal{L}_{reg} &= \text{Cos}\left(\overline{c_s}, z_q\right) = 1 - \frac{\overline{c_s} \cdot z_q}{|\overline{c_s}||z_q|},
\end{aligned}
\tag{5}
$$

where $\overline{c_s}$ is the mean of subject embeddings and z_q represents the query text embeddings. As illustrated in Fig. 5, our T2P Aligner, trained on a large scale of image-text pairs, can effectively align query text with corresponding image regions. This capability is a key aspect of selective subject-driven generation.

Similar to the original Stable diffusion model, our training objective also includes the same $\mathcal{L}_{LDM}$ loss, as outlined in Eq. (6):

$$
\mathcal{L}_{LDM}(\boldsymbol{\theta}) = \mathbb{E}_{x_0,t,\epsilon}\left[\left\|\epsilon - \epsilon_{\boldsymbol{\theta}}\left(x_t, t, c_t, c_s\right)\right\|_2^2\right], \tag{6}
$$

where $\mathbf{x_t}$ is the noisy latent at time step t, ϵ is the ground-truth latent noise. $\epsilon_{\boldsymbol{\theta}}$ is the noise prediction model with parameters $\boldsymbol{\theta}$.

Thus, our total loss function is formulated as:

$$
\mathcal{L}_{total} = \mathcal{L}_{LDM} + \tau \mathcal{L}_{reg}, \tag{7}
$$

where τ is set as a constant, with a value of 0.01. As depicted in Fig. 6 (in the last column), the inclusion of $\mathcal{L}_{reg}$ significantly enhances the text-image alignment capabilities of the SSR-Encoder. This improvement is evident in the generated images, which consistently align with both the subject prompt and the details of the reference image.

During inference, our method has the ability to decompose different subjects from a single image or multiple images. By extracting separate subject embeddings for each image and concatenating them together, our SSR-Encoder can seamlessly blend elements from multiple scenes. This flexibility allows for the creation of composite images with high fidelity and creative versatility.

4. Experiment

4.1. Experimental Setup

Training data. Our model utilizes the Laion 5B dataset, selecting images with aesthetic scores above 6.0. The text prompts are re-captioned using BLIP2. The dataset comprises 10 million high-quality image-text pairs, with 5,000 images reserved for testing and the remainder for training.

Implementation details. We employed Stable Diffusion V1-5 as the pre-trained diffusion model, complemented by the pre-trained CLIP text encoder. For training, images are resized to ensure the shortest side is 512 pixels, followed by a center crop to achieve a 512×512 resolution, and sent to the stable diffusion. The same image is resized to 224×224 and sent to the SSR encoder. The model training process is divided into two steps. In the first step, the multi-scale strategy is not employed, and the model is trained for 1 million steps on 8H800s GPUs, with a batch size of 16 per GPU and a learning rate of 5e-5. In the second step, the same hyper-parameters are used, and the model parameters obtained from the first step are used as the initialization parameters. The multi-scale strategy is employed in this step to train the model for an additional 100,000 steps. Inference was performed using DDIM as the sampler, with a step size of 30 and a guidance scale set to 7.5.

4.2. Evaluation Metrics

To evaluate our model, we employ several metrics and datasets:

- **Multi-subject bench**: We created a benchmark with 100 images, each containing 2-3 subjects.
- **DreamBench datasets** [32]: This dataset includes 30 subjects, each represented by 4-7 images.

For a comprehensive comparison with state-of-the-art (SOTA) methods, we employed the following metrics: **DINO Scores**[4], **CLIP-I**[27] and **DINO-M Scores** to assess subject alignment, **CLIP-T** [14] for evaluating image-text alignment, **CLIP Exclusive Score (CLIP-ES)** to measure the exclusivity of subject representation, and the **Aes-**

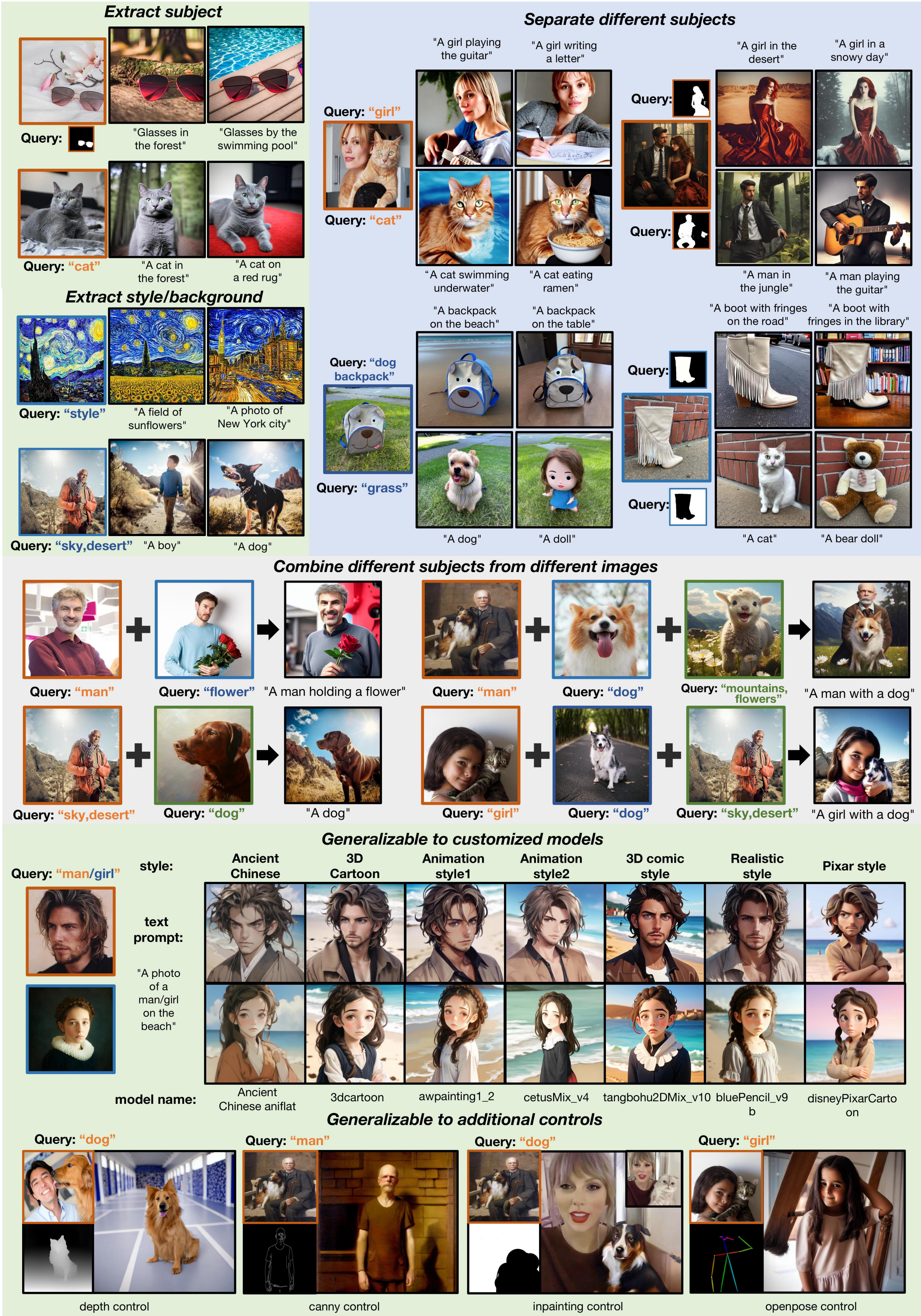

Figure 3. **Qualitative results of SSR-Encoder** in different generative capabilities. Our method supports two query modalities and is adaptable for a variety of tasks, including single- and multi-subject conditioned generation. Its versatility extends to integration with other customized models and compatibility with off-the-shelf ControlNets.

Figure 4. **Qualitative comparison** of different methods. Our results not only excel in editability and exclusivity but also closely resemble the reference subjects in visual fidelity. Notably, the SSR-Encoder achieves this without the need for fine-tuning.

thetic Score [35] to gauge the overall quality of the generated images.

Notably, CLIP-ES is calculated by generating an image I using prompts for subject A from a reference image and evaluating the CLIP-T score with a different subject B and I. A lower CLIP-ES score indicates higher exclusivity. The DINO-M score, specifically designed for multiple subjects, evaluates identity similarity between masked versions of input and generated images, as detailed in [1]. Both CLIP-ES and DINO-M scores are evaluated on the Multi-Subject Bench.

4.3. Comparison Methods

For a comprehensive evaluation of our method, we benchmarked it against a range of state-of-the-art (SOTA) techniques. The methods we compared are categorized based on their approach to fine-tuning. In the fine-tuning-based category, we include **Textual Inversion** [10], **Dreambooth** [32], and **Break-a-Scene** [1]. For fine-tuning-free methods, our comparison encompassed **Reference Only** [22], **Elite** [39], **IP-adapter** [42], and **BLIPDiffusion** [18]. This selection of methods provides a diverse range of approaches for a thorough comparative analysis with our SSR-Encoder.

4.4. Experiment Results

Quantitative comparison. Table 2 presents our quantitative evaluation across two benchmarks: the Multi-Subject Bench and DreamBench. Overall, SSR-Encoder clearly outweighs previous SOTA finetuning-free methods on all of the metrics, including subject alignment, image-text alignment, subject exclusivity, and overall quality. Remarkably, it also outperforms fine-tuning-based methods in image quality and image-text alignment within both benchmarks. Particularly in the Multi-Subject Benchmark, the SSR-Encoder demonstrates outstanding performance in subject exclusivity, markedly outperforming competing methods. This highlights the efficacy of its selective representation capability and editability. While Dreambooth excels in subject alignment within the DreamBench dataset, the SSR-Encoder and Break-A-Scene show comparable performance on the Multi-Subject Bench. This suggests that although Dreambooth is highly effective in capturing detailed subject information, SSR-Encoder achieves a balanced and competitive performance in subject representation.

Qualitative comparison. Fig. 3 displays the high-fidelity outcomes produced by the SSR-Encoder using diverse query inputs, affirming its robustness and zero-shot generative capabilities. The SSR-Encoder demonstrates proficiency in recognizing and focusing on common concepts, ensuring an accurate representation of the selected image subjects. Its seamless integration with other customized models and control modules further solidifies its significant role in the stable diffusion ecosystem.

In qualitative comparisons, as depicted in Fig. 4, Textual Inversion and Reference Only encounter difficulties in maintaining subject identity. Dreambooth, IP-adapter, and BLIP-Diffusion, although advanced, exhibit limitations in effectively disentangling intertwined subjects. Break-A-Scene achieves commendable subject preservation but at the cost of extensive fine-tuning. ELITE, with its focus on local aspects through masks, also faces challenges in consistent identity preservation.

In contrast, our SSR-Encoder method stands out for its fast generation of selected subjects while adeptly preserving their identities. This capability highlights the method's superior performance in generating precise and high-quality subject-driven images, thereby addressing key challenges faced by other current methods.

Ablation study. Our ablation study begins with visualizing the attention maps generated by our Token-to-Patch Aligner, as shown in Fig. 5. These maps demonstrate how different text tokens align with corresponding patches in the reference image, evidencing the Aligner's effectiveness.

Table 2. **Quantitative comparison** of different methods. Metrics that are bold and underlined represent methods that rank 1st and 2nd, respectively. † indicates that the experimental value is referenced from BLIP-Diffusion[18].

Type	Method	CLIP-T ↑ (Multi-subject bench)	CLIP-ES ↓ (Multi-subject bench)	DINO-M ↑ (Multi-subject bench)	CLIP-T ↑ (DreamBench)	DINO ↑ (DreamBench)	CLIP-I ↑ (DreamBench)	Aesthetic Score↑
Finetune-based methods	Textual Inversion	0.240	0.212	0.410	0.255†	0.569†	0.780†	6.029
	Dreambooth	0.298	0.223	**0.681**	0.305†	**0.668**†	0.803†	6.330
	Break-A-Scene	0.285	0.187	0.630	0.287	0.653	0.788	6.234
	Ours(full)	**0.302**	**0.182**	0.556	**0.308**	0.612	**0.821**	**6.563**
Finetune-free methods	BLIP-Diffusion	0.287	0.198	0.514	0.300†	0.594†	0.779†	6.212
	Reference only	0.242	0.195	0.434	0.286	0.542	0.727	5.812
	IP-adapter	0.272	0.201	0.442	0.274	0.608	0.809	6.432
	ELITE	0.253	0.194	0.483	0.298	0.605	0.775	6.283
	Ours(full)	**0.302**	**0.182**	**0.556**	**0.308**	**0.612**	**0.821**	**6.563**

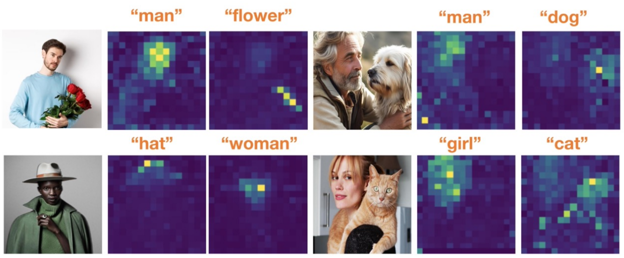

Figure 5. Visualization of attention maps A_{t2p}.

Table 3. **Ablation results on Multi-subject Bench.** Removing each component would lead to a performance drop on different aspects.

Ablation Setups	CLIP-T ↑	CLIP-ES ↓	DINO-M ↑
Text2Image	**0.352**	–	0.318
Ours(w/o multi-scale)	0.257	0.185	0.510
Ours(w/o reg loss)	0.235	0.199	0.552
Ours(full)	0.302	**0.182**	**0.556**

To evaluate the significance of various components, we conducted experiments by systematically removing them and observing the outcomes. Initially, we removed the subject condition, relying solely on the text condition for image generation, to determine if the subject details could be implicitly recalled by the base model. Subsequently, we trained a model without the embedding consistency regularization loss (L_{reg}) to assess its criticality. We also substituted our multi-scale visual embedding with a conventional last-layer visual embedding. The results of these experiments are depicted in Fig. 6.

Our observations reveal that without subject conditioning, the generated subjects failed to correspond with the reference image. Omitting the multi-scale image feature resulted in a loss of detailed information, as evidenced by a significant drop in the DINO-M score. Discarding the embedding consistency regularization loss led to challenges in generating specific subjects from coexisting subjects, adversely affecting the CLIP-ES score. In contrast, the full implementation of our method demonstrated enhanced expressiveness and precision.

Quantitative comparisons, as shown in Table 3, also indicate that our complete method achieves the best results across subject exclusivity and subject alignment. It slightly trails the original Stable Diffusion (SD) model only in text-image alignment. Substituting the multi-scale visual embedding significantly impacts image consistency, while excluding the embedding consistency regularization loss hampers text-image consistency.

Figure 6. **Qualitative ablation.** We ablate our approach by using different model settings. Without the L_{reg}, the model struggles to exclude undesired subjects from reference images. Substituting the multi-scale image feature results in less detailed outputs.

5. Conclusion

In this paper, we introduced the SSR-Encoder, a groundbreaking finetuning-free approach for selective subject-driven image generation. This method marks a significant advancement in the field, offering capabilities previously unattainable in selective subject representation. At its core, the SSR-Encoder consists of two pivotal the token-to-patch aligner and the detail-preserving subject encoder. The token-to-patch aligner effectively aligns query input tokens with corresponding patches in the reference image, while the subject encoder is adept at extracting multi-scale subject embeddings, capturing fine details across different scales. Additionally, the incorporation of a newly proposed embedding consistency regularization loss further enhances the overall performance of the system. Our extensive experiments validate the SSR-Encoder's robustness and versatility across a diverse array of scenarios. The results clearly demonstrate the encoder's efficacy in generating high-quality, subject-specific images, underscoring its potential as a valuable tool in the open-source ecosystem.

References

[1] Omri Avrahami, Kfir Aberman, Ohad Fried, Daniel Cohen-Or, and Dani Lischinski. Break-a-scene: Extracting multiple concepts from a single image. arXiv preprint arXiv:2305.16311, 2023. 2, 3, 7

[2] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, Bryan Catanzaro, et al. ediffi: Text-to-image diffusion models with an ensemble of expert denoisers. arXiv preprint arXiv:2211.01324, 2022. 2

[3] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. arXiv preprint arXiv:2304.08465, 2023. 3

[4] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé J'egou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. 2021 ieee. In CVF International Conference on Computer Vision (ICCV), 2021. 5

[5] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models, 2023. 3

[6] Hong Chen, Yipeng Zhang, 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. 3

[7] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. arXiv preprint arXiv:2307.09481, 2023. 2

[8] Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. arXiv preprint arXiv:2309.16588, 2023. 3

[9] Dave Epstein, Allan Jabri, Ben Poole, Alexei A Efros, and Aleksander Holynski. Diffusion self-guidance for controllable image generation. arXiv preprint arXiv:2306.00986, 2023. 3

[10] 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. 2, 3, 7

[11] Rinon Gal, Moab Arar, Yuval Atzmon, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. Designing an encoder for fast personalization of text-to-image models. arXiv preprint arXiv:2302.12228, 2023.

[12] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. Svdiff: Compact parameter space for diffusion fine-tuning. arXiv preprint arXiv:2303.11305, 2023. 3

[13] 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. 3

[14] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning, 2022. 5

[15] 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. In International Conference on Learning Representations, 2022. 3

[16] Xuhui Jia, Yang Zhao, Kelvin CK Chan, Yandong Li, Han Zhang, Boqing Gong, Tingbo Hou, Huisheng Wang, and Yu-Chuan Su. Taming encoder for zero fine-tuning image customization with text-to-image diffusion models. arXiv preprint arXiv:2304.02642, 2023. 2, 3

[17] 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. 2, 3, 4

[18] Dongxu Li, Junnan Li, and Steven CH Hoi. Blip-diffusion: Pre-trained subject representation for controllable text-to-image generation and editing. arXiv preprint arXiv:2305.14720, 2023. 3, 7, 8

[19] Peipei Li, Rui Wang, Huaibo Huang, Ran He, and Zhaofeng He. Pluralistic aging diffusion autoencoder. In Proceedings of the IEEE/CVF International Conference on Computer Vision, pages 22613–22623, 2023. 4

[20] Yi Li, Hualiang Wang, Yiqun Duan, and Xiaomeng Li. Clip surgery for better explainability with enhancement in open-vocabulary tasks. arXiv preprint arXiv:2304.05653, 2023. 3

[21] Jian Ma, Junhao Liang, Chen Chen, and Haonan Lu. Subject-diffusion: Open domain personalized text-to-image generation without test-time fine-tuning. arXiv preprint arXiv:2307.11410, 2023. 3

[22] Mikubill. sd-webui-controlnet, 2023. GitHub repository. 7

[23] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 6038–6047, 2023. 3

[24] Chong Mou, Xintao Wang, Liangbin Xie, Jian Zhang, Zhongang Qi, Ying Shan, and Xiaohu Qie. T2i-adapter: Learning adapters to dig out more controllable ability for text-to-image diffusion models. arXiv preprint arXiv:2302.08453, 2023. 3

[25] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. arXiv preprint arXiv:2112.10741, 2021. 2

[26] 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. 2

[27] 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. 5

[28] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation, 2021. 2

[29] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. 2

[30] Yuval Atzmon Amit H. Bermano Gal Chechik Daniel Cohen-Or Rinon Gal, Moab Arar. Encoder-based domain tuning for fast personalization of text-to-image models, 2023. 2, 3

[31] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 2, 3

[32] 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. 2022. 2, 3, 5, 7

[33] 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, 2022. 2

[34] 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. Advances in Neural Information Processing Systems, 35:36479–36494, 2022. 2

[35] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 7

[36] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. Instantbooth: Personalized text-to-image generation without test-time finetuning. arXiv preprint arXiv:2304.03411, 2023. 2, 3

[37] Dani Valevski, Danny Wasserman, Yossi Matias, and Yaniv Leviathan. Face0: Instantaneously conditioning a text-to-image model on a face. arXiv preprint arXiv:2306.06638, 2023. 3

[38] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. $p+$: Extended textual conditioning in text-to-image generation. arXiv preprint arXiv:2303.09522, 2023. 3

[39] 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. arXiv preprint arXiv:2302.13848, 2023. 2, 3, 4, 7

[40] Zeyue Xue, Guanglu Song, Qiushan Guo, Boxiao Liu, Zhuofan Zong, Yu Liu, and Ping Luo. Raphael: Text-to-image generation via large mixture of diffusion paths. arXiv preprint arXiv:2305.18295, 2023. 2

[41] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 18381–18391, 2023. 3

[42] 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. 2, 3, 4, 7

[43] Bohan Zeng, Shanglin Li, Yutang Feng, Hong Li, Sicheng Gao, Jiaming Liu, Huaxia Li, Xu Tang, Jianzhuang Liu, and Baochang Zhang. Ipdreamer: Appearance-controllable 3d object generation with image prompts. arXiv preprint arXiv:2310.05375, 2023. 2

[44] Lvmin Zhang and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. arXiv preprint arXiv:2302.05543, 2023. 3, 4

[45] Yuechen Zhang, Jinbo Xing, Eric Lo, and Jiaya Jia. Real-world image variation by aligning diffusion inversion chain. arXiv preprint arXiv:2305.18729, 2023. 3

[46] Shihao Zhao, Dongdong Chen, Yen-Chun Chen, Jianmin Bao, Shaozhe Hao, Lu Yuan, and Kwan-Yee K Wong. Uni-controlnet: All-in-one control to text-to-image diffusion models. arXiv preprint arXiv:2305.16322, 2023. 3

[47] Chong Zhou, Chen Change Loy, and Bo Dai. Extract free dense labels from clip. In European Conference on Computer Vision, pages 696–712. Springer, 2022. 3