DemoCaricature: Democratising Caricature Generation with a Rough Sketch

Dar-Yen Chen Ayan Kumar Bhunia Subhadeep Koley Aneeshan Sain
Pinaki Nath Chowdhury Yi-Zhe Song
SketchX, CVSSP, University of Surrey, United Kingdom.
{d.chen, a.bhunia, s.koley, a.sain, p.chowdhury, y.song}@surrey.ac.uk

https://democaricature.github.io

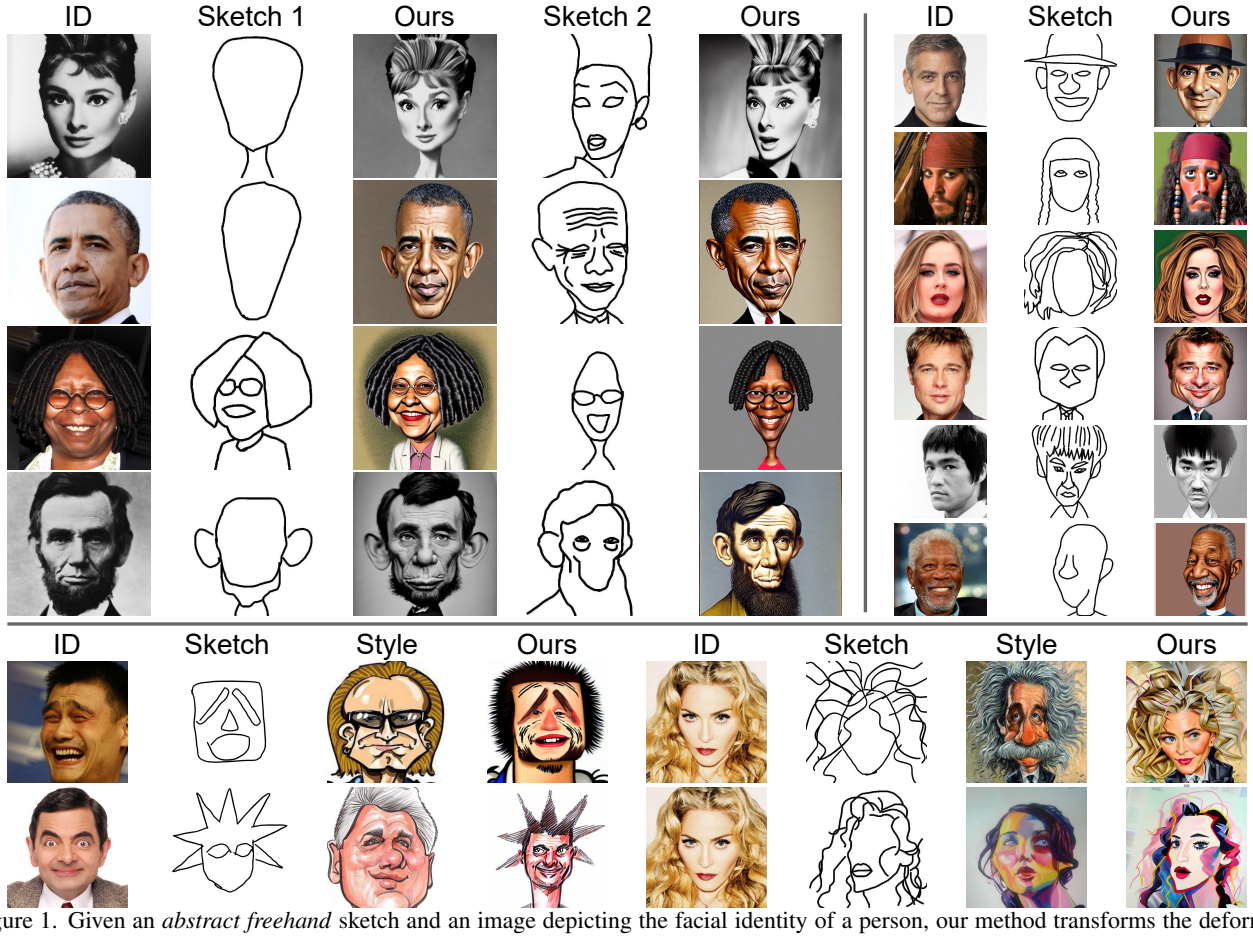

Figure 1. Given an *abstract freehand* sketch and an image depicting the facial identity of a person, our method transforms the deformed sketch into a plausible-looking caricature while maintaining *identity-fidelity* and imitating the *exaggerations* portrayed in the input sketch. Additionally, it can seamlessly transmit the *look-and-feel* of a given *style-image* into the output caricature.

Abstract

In this paper, we democratise caricature generation, empowering individuals to effortlessly craft personalised caricatures with just a photo and a conceptual sketch. Our objective is to strike a delicate balance between abstraction and identity, while preserving the creativity and subjectivity inherent in a sketch. To achieve this, we present Explicit Rank-1 Model Editing alongside single-image personalisation, selectively applying nuanced edits to cross-attention layers for a seamless merge of identity and style. Additionally, we propose Random Mask Reconstruction to enhance robustness, directing the model to focus on distinctive identity and style features. Crucially, our aim is not to replace artists but to eliminate accessibility barriers, allowing enthusiasts to engage in the artistry.

1. Introduction

Ever wondered when you would finally decide to get that personalised caricature created, perhaps during a holiday? Look no further, this paper is for you – we strive to democratise caricature [6, 26, 27] generation for everyone! With a portrait of yourself and a conceptual sketch of how you envision your caricature, we will automatically generate a high-fidelity caricature that unmistakably captures

your essence [27]. Our aim however is not to replace artists; after all, the realm of art may be one that AI will never entirely conquer – so when you find yourself in Paris next, *do* get your caricature expertly crafted by a skilled artist!

We commence our study by asking the fundamental question that arises when scrutinising a caricature – Is this me? (or Obama or Mr. Bean for that matter in Fig. 1?) Indeed, the core challenge in caricature generation is navigating the delicate balance of infusing abstraction [20] into the process to achieve that distinctive caricature appearance, while still preserving the essential identity cues that unmistakably represent the intended person [10]. Over and above all, how do we seamlessly inject your individuality [8] and creativity [5] into the art generation process, ensuring the resulting caricature is genuinely *your own*, rather than one dictated *solely* by AI?

Our solution lies in your sketch! A single rough sketch [2, 3, 33] is all it takes to encapsulate *your* vision for *your* caricature, as illustrated in Fig. 1. The scientific challenge is clear: regardless of your artistic skill or the lack of it, how can we design a system to adeptly generate a plausible caricature while still preserving your identity [27]? And one more thing, if there is a specific caricature style you prefer, we would like to accommodate that preference as well.

We most certainly are not pioneers in caricature generation [6, 26, 27]; our motivation primarily draws from prior art in this field. However, our set of challenges notably surpasses the technical capabilities of previous systems [6, 27], particularly those primarily deformation-based [16, 57], which tend to prioritise style creation over identity. Crucially, these prior systems often fall short in including "you" in the solution. This deficiency results in generated caricatures lacking expressiveness and missing interesting features like local abstraction [20], hairstyle variation [66], and view changes – all of which can be easily injected into our system with just your single sketch!

Our approach to modelling the delicate balance between identity and style relies on the interaction between a novel single-image Text-to-Image (T2I) personalisation module and a sketch-specific T2I-Adapter [41]. The former ensures identity, while the latter allows for sketch-controlled caricature generation. This, of course, is not trivial. Latest single-image T2I personalisation approaches [13, 50, 61] often grapple with overfitting during single-image fine-tuning, resulting in a highly specialised yet inflexible model that lacks generalisation beyond training data. This makes them especially challenging for them to adapt to the highly exaggerated and subjective human sketches, which are often Out-of-Distribution (OOD). This challenge is further exacerbated, as we face the task of merging concepts of identity and style. If done blindly, this would lead to a blending of features, resulting in caricatures that lack distinction or skew towards one aspect at expense of the other.

We thus propose Explicit Rank-1 Model Editing for single-sketch personalisation, enabling effective learning and the fusion of identity and style. By incorporating an explicit mechanism, it independently manipulates the explicit editing of identity and style in the cross-attention layers [41], with minimal extra parameters while maintaining the integrity of textual contexts. This provides a more subjective and fine-grained control over desired concepts, mitigating the overfitting typically encountered. Furthermore, we introduce Random Mask Reconstruction to enhance the robustness of distorted shapes. It achieves this by masking random patches of the input image, compelling the model to focus on crucial identity and style features over local variations. This capability importantly allows the model to better handle exaggerated caricature sketches while emphasising the essential learned features.

Our contributions are: *(i)* we democratise caricature generation, enabling individuals to easily create personalised caricatures, from a photo and a conceptual sketch. *(ii)* we address the delicate balance between abstraction and identity via Explicit Rank-1 Editing, offering nuanced control by selectively applying rank-1 edits to cross-attention layers. *(iii)* we enhance system robustness with Random Mask Reconstruction, enabling effective handling of distorted shapes while emphasising essential identity and style.

2. Related Work

Deep Caricature Synthesis: Caricature synthesis aims to *exaggerate* or *distort* specific facial features for a *stylised* yet *recognisable* portrayal of a subject [4, 14, 16]. Such methods typically involve a deformation stage, followed by image-to-image translation. Introduced as a GAN [15]-based framework [6] involving facial landmarks to guide deformations, it was enhanced by automating control point prediction for warping and embedding a discriminator, acting as an *identity* classifier to help in its preservation [57]. A few subsequent works include diversifying caricature generation to multiple facial exaggeration types [16], leveraging SENet [25] and spatial transformer modules to produce high-fidelity warps based on dense warping field [14], and leveraging StyleGAN [28] with GAN inversion [44, 60] to propose *shape exaggeration* blocks for additional control [27]. Towards spatial manipulation within caricature synthesis, while Semantics CariGAN [9] leveraged semantic shape transformations for caricature-control from warped semantic maps, a segmentation-guided dual-domain synthesis framework [4] combined few-shot GAN [47] with RepurposingGAN [62]. Addressing the limitations of the deformation-based pipeline, we strive to enhance creative freedom in caricature synthesis via freehand sketches.

Denoising Diffusion Probabilistic Models (DDPM): Recently, DDPMs [22] have emerged as the de facto choice for generative modelling, thanks to their high-fidelity im-

age synthesis potential [34]. Earlier works [21, 22, 42, 48] have significantly improved text-to-image (T2I) models, such as Imagen [53], DALL-E2 [46], and Stable Diffusion (SD) [48] – further enhanced by training on diverse image-caption pair datasets [54, 55]. Harnessing the prior knowledge of *pretrained* T2I models, research progressed to guide generation under additional conditions [68, 69, 71]. For instance, ControlNet [70] and T2I-Adapter [41] introduced content semantics adapters for targeted tasks such as pose, depth map, and sketch-conditional synthesis [65], which enhances the flexibility of the generation process.

T2I Personalisation: With a limited set of reference images, T2I personalisation aims to adapt pretrained T2I models [48, 53] to specific concepts, while retaining its generalisability. Among the proposed strategies [12, 58, 64], while Textual Inversion [13] optimises text embeddings to capture new concepts, DreamBooth [50] personalises the output by fine-tuning the whole Stable Diffusion [48] and Imagen [53] models. Research on Parameter-Efficient Fine-Tuning (PEFT) methods [18], such as LoRA [24, 52] and SVDiff [17], focuses on reducing the computational burden during model training. Additionally, CustomDiffusion [37] fine-tunes only cross-attention layers, while Perfusion [61] introduces Rank-1 Model Editing (ROME) [40] to optimise the Value-pathway in the cross-attention mechanism. InstantBooth [56] enables personalised inference with single images. FastComposer [67] uses a novel image encoder for concept embeddings, while HyperDreamBooth [51] achieves efficient fine-tuning with a hypernetwork. However, resource-intensive training may limit their application [56, 67]. We thus offer a rapid and universal single-image method, that extends personalisation beyond individual identity images to include reference style images, facilitating synthesis of stylised caricatures, while costing minimal iterations and parameter overhead.

3. Revisiting Text-to-Image Diffusion Models

Overview: Diffusion models [11, 48, 59], rely on two stochastic processes, termed as *forward* and *backward* diffusion [22]. The *forward process* involves iteratively adding Gaussian noise to a clean image $x_0 \in \mathbb{R}^{h\times w\times 3}$ over t time-steps, producing a noisy image $x_t \in \mathbb{R}^{h\times w\times 3}$ as: $x_t = \sqrt{\bar{\alpha}_t}\, x_0 + (\sqrt{1-\bar{\alpha}_t})\epsilon$, where, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$ is the added noise, $\{\alpha_t\}_1^T$ represents a predefined noise schedule [22] with $\alpha_t = \prod_{s=1}^{T} \alpha_s$ and time-step t value is sampled from a Uniform distribution $t \sim U(0, T)$. With sufficiently large T, x_T approximates isotropic Gaussian noise. The *backward process* entails training a modified UNet [49] denoiser $F_\theta(\cdot)$. It takes the noisy input x_t and the corresponding time-step t to estimate the input noise $\epsilon_t \approx F_\theta(x_t, t)$. Once trained with a standard MSE loss [22], F_θ can reverse the effect of forward diffusion. During inference, F_θ is applied iteratively for T time-steps on a randomly sampled 2D Gaussian noise image x_T to get a cleaner image x_{t-1} at each time-step t, thus eventually resulting in one of the cleanest images x_0 resembling the original target distribution [22]

Text-Conditioned Diffusion Model: Diffusion models can generate images conditioned on different signals (*e.g.*, class labels [23], textual prompts [46, 48], etc.). Given a textual prompt p, the initial step involves its conversion to the word-embedding space $\mathcal{W}$ on applying a word-embedding function $\mathbf{W}$. Subsequently, the transformed prompt is passed through a CLIP [45] text encoder denoted by $\mathbf{T}(\cdot)$, which produces the text encoding as $\mathbf{t_p} = \mathbf{T}(\mathbf{W}(p)) \in \mathbb{R}^{77\times d}$ in the text encoding space $\mathcal{T}$. This $\mathbf{t_p}$ controls the diffusion process via cross-attention, thus allowing $F_\theta(x_t, t, \mathbf{t_p})$ to perform p controlled denoising on x_t.

Stable Diffusion: Latent Diffusion Models (*i.e.*, Stable Diffusion) [48] perform forward and backward denoising in the *latent* space for [48]. In its *two-stage* approach, Stable Diffusion (SD) [48] first trains a *variational autoencoder* (VAE) [31], comprising an encoder $E(\cdot)$ and a decoder $D(\cdot)$ in sequence. $E(\cdot)$ converts the input image $x_0 \in \mathbb{R}^{h\times w\times c}$ to its latent representation $z_0 \in \mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times d}$ [48]. The forward process adds Gaussian noise to z_0 over t time-steps, producing a noisy latent $z_t = \sqrt{\bar{\alpha}_t}\, z_0 + (\sqrt{1-\bar{\alpha}_t})\epsilon$. Later, a UNet [49] denoiser $\epsilon_\theta(\cdot)$ is trained to perform conditional denoising based on textual prompt p directly in the latent space [48] with loss objective as:

$$\mathcal{L}_{sd} = \mathbb{E}_{z_t, t, \epsilon, p}(||\epsilon - \epsilon_\theta(z_t, t, \mathbf{t_p})||_2^2) \tag{1}$$

SD incorporates the text conditioning using $\mathbf{t_p}$ into the denoising process via *cross-attention* [48] as:

$$\begin{cases} Q = W_Q z_t;\ K = W_K \mathbf{t_p};\ V = W_V \mathbf{t_p} \\ \texttt{Attention}(Q, K, V) = \texttt{SoftMax}(\frac{QK^T}{\sqrt{d}}) \cdot V \end{cases} \tag{2}$$

where W_Q, W_K, and W_V are the learnable projection matrices. W_K and W_V linearly projects the text ecoding $\mathbf{t_p} \in \mathbb{R}^{77\times 768}$ to form the *"Key"* and *"Value"* vectors. Whereas, W_Q projects the intermediate noisy latents to form the *"Query"* maps [48]. The cross-attention map is produced as $\texttt{SoftMax}(\frac{QK^T}{\sqrt{d}}) \cdot V$. Essentially, the cross-attention map indicates the correspondence between the textual prompt and the spatial regions of the image [48].

T2I-Adapter: Moving beyond conventional *textual conditioning* [48], T2I-Adapter [41] enables a myriad of different *spatial conditioning signals* [36] (*e.g.*, segmentation masks, scribbles, sketches, key pose, depth maps, colour palates, etc., or their weighted combinations) to guide the T2I image generation process of SD [48]. In practice, T2I-adapter [41] trains a lightweight network (comprising one convolutional and four residual blocks) that extracts deep features from spatial conditioning signals at four different scales. Those extracted *conditioning-features* are then added with intermediate features of SD's UNet decoder at each scale [41] to influence denoising with the given condition [41].

4. Problem Definition and Challenges

Given a reference portrait photo $\mathcal{I}_p$ depicting a specific identity p and a free-hand abstract sketch $\mathcal{S}$ as the query, we aim to generate a caricature $\mathcal{C}_p^s$, which should retain the identity [10] captured in $\mathcal{I}_p$, while reflecting the sketch's ($\mathcal{S}$) influence on its shape. Notably, $\mathcal{I}_p$ may represent an individual *not* encountered previously, and the free-hand sketch $\mathcal{S}$ may depict random or highly *unconstrained* deformations [6] or shape exaggerations [57], to be reflected in $\mathcal{C}_p^s$.

Complexity stems from the delicate balance between preserving identity [10] and introducing sketch-guided shape deformations [9]. Learning the *unique* identity from a single reference image is non-trivial, given the risk of overfitting [13, 50] on limited data. Adding to it is the complexity of generating an exaggerated shape in *accordance* with the sketch [1, 35]. Addressing these challenges requires a robust model capable of learning and generalising [61] from a single reference image while optimising the trade-off between identity preservation and shape exaggeration.

5. Sketch for Caricature Generation

Overview: Our approach diverges from the prevalent use of pre-trained StyleGAN's [28–30] latent space in facial image editing [43] tasks. Instead, we opt for a pre-trained text-to-image stable diffusion (SD) model [48], known for its generalisation [38] and adaptability [13] across diverse and wild scenarios. Our problem being inherently multi-modal [59], where $\mathcal{I}_p$ is a *real photo*, $\mathcal{S}$ is a black-and-white sparse abstract [20] *line drawing*, and the output caricature ($\mathcal{C}_p^s$) typically extends *beyond real photo* modality, SD [48] becomes an ideal fit as it excels in handling such scenarios which are less encountered [59] in real life.

Our personalised text-to-image (T2I) framework involves fine-tuning the SD model to capture identity in the reference photo $\mathcal{I}_p$ and generate the *same* identity in various contexts [13]. Consequently, we leverage an off-the-shelf T2I-Sketch-Adapter [41] to spatially condition the identity-adapted SD model. This process effectively integrates shape guidance from sketch, aligning $\mathcal{C}_p^s$ with the intended shape.

Our caricature generation pipeline extends further to include style [27] adaptation, by acquiring low-level style features from a single style-reference image $\mathcal{I}g$ characterised by a specific style g. The resulting output caricature $\mathcal{C}_{p|g}^s$, now concurrently preserves the identity, style, and shape derived from $\mathcal{I}_p$, $\mathcal{I}_g$, and $\mathcal{S}$, respectively.

5.1. Baseline off-the-shelf Solutions

Given the recent rise of personalised T2I frameworks [48, 50, 53], one can naively finetune it using a single reference identity image, and further generate a sketch-conditioned shape-exaggerated output caricature plugging an off-the-shelf T2I-Sketch-Adapter [13]. Among such frameworks, Textual Inversion [13] aims to learn a new pseudo word embedding $\mathbf{v}_*$ (representing the concept) in $\mathcal{W}$ space by directly optimising the LDM loss as in Eq. (1) against reference images. Whereas, Perfusion [61] further adjusts visual representations through ROME [40], modifying the Value-pathway activation according to the component of all words that are aligned with the target concept.

Such a naive solution would however suffer from a few challenges. Firstly, training from a single reference identity ($\mathcal{I}_p$) or style ($\mathcal{I}_g$) image easily leads to overfitting [61] in word embeddings, thereby compromising on generalisability to multiple contexts. Secondly, integrating homologous concepts like identity and style, encounters a substantial degree of semantic overlap [61]. This results in detrimental interference (see Fig. 5), causing the concepts to overshadow each other in sketch+style guided caricature generation. Lastly, being trained on a single reference image only, it fails to generalise towards imbibing the exaggerated [57] shape guidance from *diverse* sketches [20].

Accordingly, we propose three key solutions: *(i) Explicit* Rank-1 Model Editing (Sec. 5.2), that edits only at the concept index. Besides preventing potential interference, it also refines the optimisation scope, rendering the adaptation process more effective. Secondly, we implement Random Mask Reconstruction (Sec. 5.3) to enable training with locally masked images, directing the model's focus *away* from local variations and emphasising on key features. This enhances the model's resilience to diverse facial shape constraints [32] crucial for caricature synthesis. Thirdly, we incorporate additional regularisation (Sec. 5.4) using superclass on word embeddings and text encoding, to counter overfitting, which ensures the model's attention mechanism remains less burdened by the identity [61], allowing more *free*-form shape exaggeration, while preserving identity.

5.2. Explicit Rank-1 Model Editing

Rank-1 Model Editing (ROME) [40] in NLP considers transformer [63] feed-forward layers as memory storage. It utilises learnable outputs to edit this memory, aligning it with the target concept. ROME differentially modifies only the knowledge related to the target, preserving rest of the pre-trained model's memory completely. In T2I [41], textual context is integrated via cross-attention layers, using *'Key'* and *'Value'* pathways akin to feed-forward layers in transformers [61]. Our contribution, Explicit Rank-1 Model Editing (Explicit ROME), refines T2I models by applying modifications to the textual encoding locally, specifically *at the position of the concept index* while *leaving* other textual contexts untouched.

Given a reference identity photo $\mathcal{I}_p$ and a textual prompt p = `'a photo of a P*'`, we convert p to a series of word embedding vectors p_w through word-embedding layer $\mathbf{W}$ where the word embedding corresponding to *concept token* `P*` is replaced with a learnable pesudo word embedding

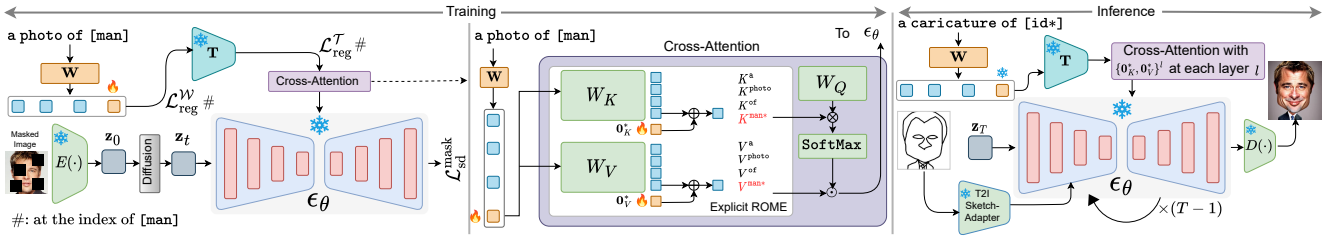

Figure 2. Within cross-attention layers, Explicit ROME (Sec. 5.2) edits the concept entry with trainable target output $\mathbf{o}^*$ that encapsulates the identity features. We also employ a dynamic masking method (Sec. 5.3), selectively occluding latent regions during training to enhance model robustness. Additional regularisation (Sec. 5.4) is applied to word embeddings and text encoding through superclass. During inference, a frozen T2I-sketch-adapter [41] provides shape guidance, resulting in an output caricature with the desired identity and shape. A similar training pipeline is used for the style image as well. We use Eq. (4) to perform sketch+style guided caricature generation.

vector $\mathbf{v}^* \in \mathbb{R}^{768}$ for SD v1.5 [48]. It is initialised from the word embedding of its corresponding superclass word, like 'man' or 'woman' based on the gender of the identity photo $\mathcal{I}_p$. Position of the concept token is denoted as $\mathtt{c_i}$. Next, we use a CLIP-text encoder to obtain textual encoding (in $\mathcal{T}$ space) as $\mathbf{t_p} = \mathbf{T}(p_w) \in \mathbb{R}^{77\times768}$. This $\mathbf{t_p}$ influences the intermediate feature map of SD-UNet through *'Key'* and *'Value'* pathways which we edit via *Explicit Rank-1 Model Editing* in the next stage.

Considering $W \in \mathbb{R}^{320\times768}$ (for SD v1.5 [48]) from Eq. (2) as the embedding matrix for either *'Key'* as W_K or *'Value'* as W_V and $\mathbf{t_p}$ as the textual encoding, the standard output $h = Wt_p$ is edited by *Explicit ROME* as:

$$h[\mathtt{c_i}] \leftarrow h[\mathtt{c_i}] + s \cdot \Phi(\mathbf{t_p}[\mathtt{c_i}], i^*) \cdot \mathbf{o}^* \tag{3}$$

where $\Phi(\cdot,\cdot)$ is the cosine similarity function and $\mathbf{o}^*$ is a learnable vector of size $\mathbb{R}^{320}$ (for SD v1.5 [48]). The target input i^* is initialised from CLIP [45] text encoding $\mathbf{t_p}$ at $\mathtt{c_i}$ index, and at every step is updated through the exponential moving average [61] as $i^* \leftarrow 0.98 \cdot i^* + \mathbf{t_p}[\mathtt{c_i}]$. The input i^* serves as a prototype for gauging the alignment of various contexts with the learned identity. The scale s allows modulating the degree of personalisation during inference, offering more control over results. The similarity $\Phi(\mathbf{t_p}[\mathtt{c_i}], i^*)$ represents how closely the input matches i^*, after the interaction of learnable pseudo word embedding $\mathbf{v}^*$ and other word embeddings. It can adjust what level of identity features in the caricature should be embedded depending on various contexts (freehand sketch in our case). Instead of complex Mahalanobis distance [61] based formulation we utilise the cosine distance to calculate the similarity, which is an intuitive and effective choice according to the text encoder CLIP [45]. In all the cross-attention layers' *'Key'* and *'Value'* pathways, we explicitly apply Explicit ROME as in Eq. (3) only at the index position of the concept token $\mathtt{c_i}$. This aligns visual features with the target concept, therefore preserving other textual contexts, and consequently ensuring generalisability without compromise.

Similar to adapting to a reference identity photo $\mathcal{I}_p$, one can adapt it for a specific style-image $\mathcal{I}_g$ as well, taking superclass word for style images as 'comics', 'illustration' etc. In particular, Eq. (3) can be extended to combine multiple independently trained concepts as follows:

$$h[\mathtt{c_i}] \leftarrow h[\mathtt{c_i}] + \sum_{j=1}^{J} s_j \cdot \Phi(\mathbf{t_p}[\mathtt{c_i}], i_j^*) \cdot \mathbf{o}_j^* \tag{4}$$

This equation independently treats each concept at its respective j^{th} index, preserving unique elements without *unintentional blending*. This ensures easier integration of multiple concepts in caricature synthesis [61], addressing the challenge of blending homologous identity and style [27].

To sum up, our method has the following trainable parameters: *(i)* a single pseudo word embedding $\mathbf{v}^* \in \mathbb{R}^{768}$. *(ii)* the tuple $\{\mathbf{o}_\mathbf{K}^*, \mathbf{o}_\mathbf{V}^*\}^l$ at each cross-attention layer l for *'Key'* and *'Value'* pathway respectively. Every $\mathbf{o}^*$ has a dimension of 320, thus making our Explicit ROME overall $30\times$ lesser learnable parameters than Perfusion [61].

5.3. Random Mask Reconstruction

One of the major challenges of caricature synthesis is *recreation* of the reference-style [27], while *maintaining* the subject's unique identity [6, 26, 27, 61]. To ensure the seamless reproduction of style and identity in the output caricature, we introduce random mask reconstruction (RMR) loss. We hypothesise that random masking of the reference images would shift the model's focus from local spatial regions, enforcing it to understand the global concepts (*i.e.*, style and identity). Given a random masked image, we pass it through the encoder $E(\cdot)$, to obtain a masked latent image z_0^m which after forward diffusion becomes z_t^m. This upon passing through UNet-denoiser ϵ_θ conditioned on $\mathbf{t_p}$, the modified SD objective becomes:

$$\mathcal{L}_{\text{sd}}^{\text{mask}} = \mathbb{E}_{\mathbf{z}_t,t,\mathbf{t_p},\epsilon}(||(\epsilon - \epsilon_\theta(\mathbf{z}_t^m, t, \mathbf{t_p})) \odot M||_2^2) \tag{5}$$

where M is the equivalent latent space binary mask with size same as z_t. It is used to impose $\mathcal{L}_{\text{sd}}^{\text{mask}}$ on the unmasked areas only. In practice, we obtain M via bilinear downscaling from a randomly sampled mask [19] in the pixel space.

5.4. Concept Regularisation

Any marked deviation of the concept word embedding $\mathbf{v}^*$, risks dominating of the text encoder and attention mecha-

Figure 3. **Qualitative comparison with GAN-based deformation models.** These visual results illustrate our method's higher fidelity and shape flexibility in caricature synthesis compared to existing method *viz.* StyleCariGAN [27], CariGANs [6], and WarpGAN [57].

nism by the *concept*, thus losing generalisability to sketch-based deformations. We thus apply l_2 regularisation [13] on the concept word embedding against its superclass word embedding S_c^w in $\mathcal{W}$ space to prevent overfitting of the text encoder. Furthermore, we impose cosine distance-based regularisation loss between text encodings $\mathbf{t_p}$ (using $\mathbf{v}^*$) and $\mathbf{t_p^{s_c}}$ (using S_c^w) from CLIP textual encoder $\mathbf{T}$, at the position of concept token $\mathtt{c_i}$. Therefore, the regularisation losses in $\mathcal{W}$ and $\mathcal{T}$ spaces become:

$$\mathcal{L}_{\text{reg}}^{\mathcal{W}} = l_2(\mathbf{v}^*, S_c^w)\ ;\ \mathcal{L}_{\text{reg}}^{\mathcal{T}} = 1 - \Phi(\mathbf{t_p}[\mathtt{c_i}], \mathbf{t_p^{s_c}}[\mathtt{c_i}]) \quad (6)$$

Finally, the overall training loss becomes $\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{sd}}^{mask} + \lambda_1 \mathcal{L}_{\text{reg}}^{\mathcal{W}} + \lambda_2 \mathcal{L}_{\text{reg}}^{\mathcal{T}}$. Please see Fig. 2 for a summarised overview of training and inference pipelines.

6. Experiments

Datasets. We use the WebCaricature dataset [26] to source identities and styles. To validate our approach via a quantitative comparison and a user study, we curate a test dataset encompassing 20 identities, 4 styles, and 12 distinctive edge maps as shapes. These edge maps are extracted from caricature images of WebCaricature [26], leading to 960 unique caricature pairs for evaluation. For a fair assessment of our method, the carefully selected identities encompass a wide-spectrum of race, gender, and age, thereby upholding diversity and inclusiveness in our evaluation. Analysing qualitative results, we incorporate amateur freehand sketches, incorporating real user interpretation into the assessment.

Implementation Details. Our implementation is based on Stable Diffusion v1.5 [48]. We train using AdamW [39] optimiser, with a batch size of 16, learning rates 0.2 and 0.002 for target outputs and embeddings respectively. Fine-tuning consists of 40 and 100 steps for identities and styles, respectively. We conduct all experiments on a single NVIDIA GTX 4090 GPU, taking 1 minute for identity and 2 minutes for style fine-tuning. For inference, results are sampled with 50 steps along with a classifier-free guidance [21] scale of 9. We use the prompts "`a caricature of` [`id`$*$]" and "`a caricature of` [`id`$*$] `in the style of` [`style`$*$]" to generate caricatures.

6.1. Qualitative Evaluation

Fig. 1 shows the efficacy of our proposed method in generating caricatures while faithfully conforming to specifications of identity [26], style [27], and sketch shape [20]. Given a subject, our method demonstrates its robust caricature synthesis [27] potential in the above half of Fig. 1. It moves beyond the traditional confines of feature scaling [6, 27, 57] to a paradigm where features can be adjusted and exaggerated with ease of using sketch-based guidance. Such flexibility reaches into the domain of fine-grained facial feature manipulation [7] adjusting shape, features (mouth, ears, nose), expressions, as well as hairstyles, while also attending to accessories and novel perspectives. From simple one-stroke outlines to intricate details, our model demonstrates adaptability to varying sketch complexities. Remarkably, it achieves this without reliance on identity-tailored components [51], capturing the subtle essence of human faces from merely a *single* reference image with only a *few* fine-tuning steps. Our framework addresses the challenge of identity preservation while applying exaggeration and distortion, exemplifying a robust resistance to overfitting. When constrained by a sketch, the model seamlessly integrates identity into the shape, ensuring recognisability without apparent visual artefacts, while maintaining the prior knowledge of the SD [48] model.

Lower half of Fig. 1 illustrates our model's ability to harmonise two conflicting concepts: identity and style, each derived from separate human likenesses. The objective is to unify them within a single synthesised caricature face. Diffusion backbones [48] usually struggle with such duality, yet our model overcomes this, rendering caricatures with high fidelity to both identity and style elements.

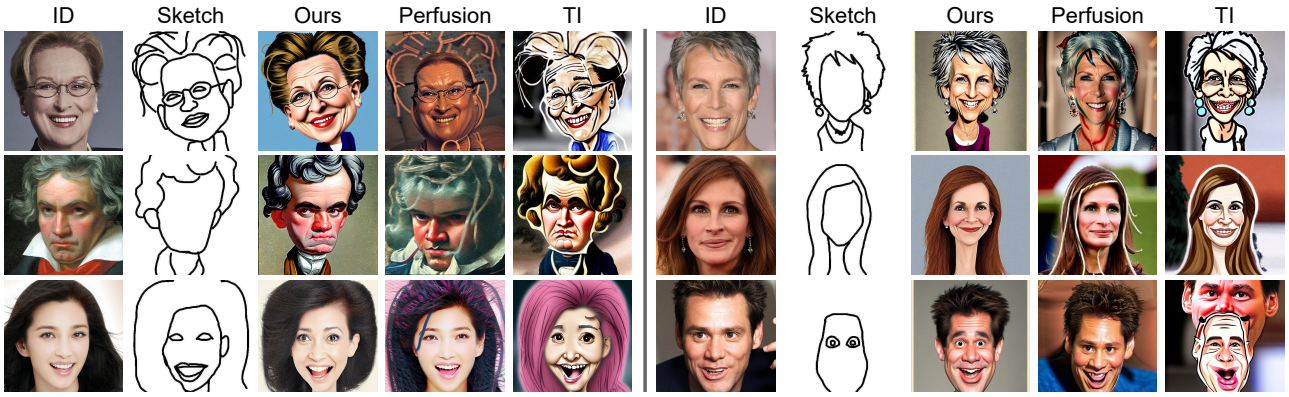

Figure 4. **Comparison with T2I personalisation approaches.** Our framework is stronger in single-image personalisation caricature synthesis against Perfusion [61] and TI [13].

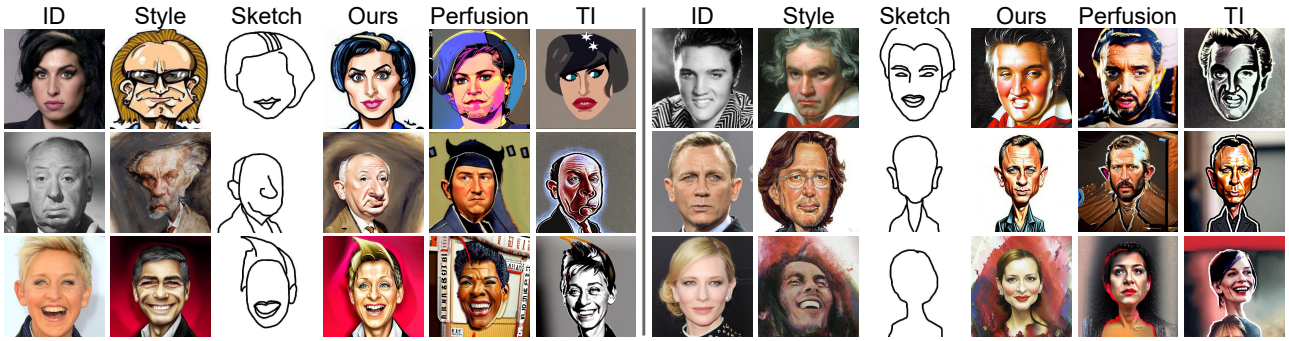

Figure 5. **Comparison with T2I personalisation approaches with style reference.** Demonstrates our model's robustness in generating stylised caricatures with faithful identity and style, surpassing other methods like Perfusion [61] and TI [13].

6.2. Comparison with SOTA

We benchmark our caricature synthesis against three state-of-the-art (SOTA) deformation-based models *viz.* **StyleCariGAN** [27], **CariGANs** [6], and **WarpGAN** [57]. These models however do not support caricature synthesis with combined conditioning on identity [6], style [27], and shape [26] like ours. We extend our comparison to advanced SD-based [48] personalisation models, like **Textual Inversion (TI)** [13] and **Perfusion** [61] as well.

CariGANs [6] and **WarpGAN** [57] which rely on landmarks and control point manipulation, clearly show distortions and artefacts in Fig. 3. By leveraging deep feature-map modulation from StyleGAN [28], **StyleCariGAN** [27] delivers higher-fidelity caricatures, yet it is limited to predefined scale-based exaggeration [27], ignoring shape information. On the other hand **TI** [13] and **Perfusion** [61] fail to preserve identity in caricatures due to overfitting caused by single-image personalisation (Fig. 4). Furthermore, lacking an effective interaction-control mechanism, they suffer from (Fig. 5) identity and style ambiguity, thus deviating from corresponding references. Our Explicit ROME strategy circumvents these pitfalls, ensuring targeted editing at corresponding positions without disrupting other text and concept encodings in the cross-attention mechanism [41], as verified by our superior qualitative results.

Table 1. **Quantitative comparison.** Quantitative metrics of various approaches and our framework ablative design, reflecting the precise quantitative edge our model holds over existing methods.

Methods	ID ↑	Style ↑	Shape ↑
TI [13]	0.634	0.553	0.633
Perfusion [61]	0.536	0.549	0.676
Ours (w/o rand mask)	0.659	0.567	0.694
Ours (w/o explicit)	0.664	0.530	0.661
Ours (Mahalanobis)	0.666	0.574	0.663
Ours-full	0.671	0.576	0.654

Now, for quantitative evaluation (Tab. 1) , we use CLIP-Score [45] on *ID*, *Style*, and *Shape*. It measures *ID*/*style*-fidelity as the similarity between generated caricatures and ID/style images using a pre-trained CLIP [45] encoder, and *shape* fidelity as the same between edgemaps of generated caricatures and conditioning sketches. Notably, at the same level of shape similarity, our results have the highest identity and style similarity at 0.671 and 0.576, which is 3× and 10× faster than **Perfusion** [61] and **TI** [13] respectively. Notably, this was achieved within three minutes of fine-tuning for identity and style.

Human Study. We conduct a thorough human study to judge the efficacy of our method from end-users' perspective. Specifically, each of the 15 users were shown 20 tuples, each containing {ID, input sketch, style image, output caricature} from *all* competing methods, and asked to

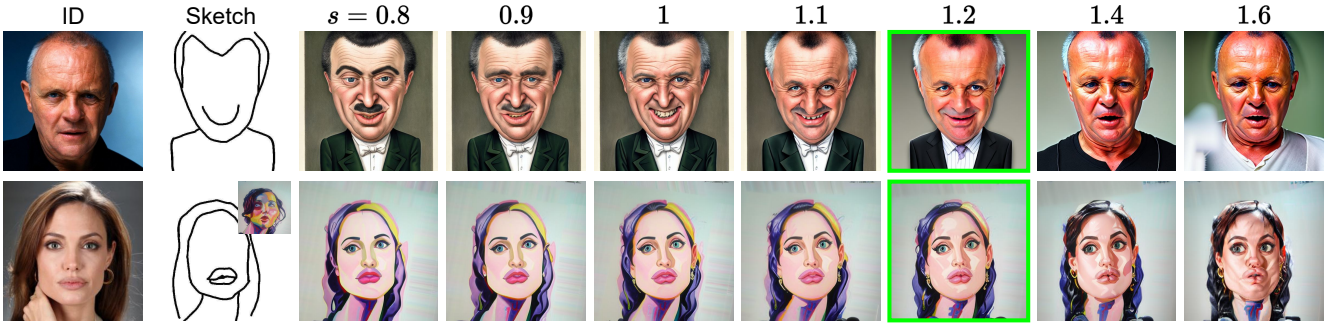

Figure 6. **Identity Scale Adaptability.** Our method provides a dynamic adjustment of the identity scale s, exemplifying flexibility.

Figure 7. Qualitative results of our ablation study

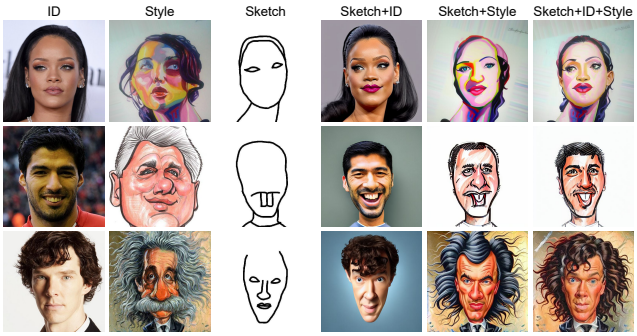

Figure 8. Our model's capacity to integrate various modalities.

rate the caricatures on a discrete scale of $[1, 5]$ (*worst* to *best*) based on fidelity to input *sketch-shape*, *style*, and *ID* – resulting in a total of 300 responses per method. The final score for each method is calculated from the mean of all its responses. Our method with high shape-fidelity and identity-preservation, garners an impressive overall score of 4.1 (Tab. 2) surpassing others. Although, the users preferred **TI** [13] over **Perfusion** [61] in terms of identity-preservation, they both score lower compared to ours.

Table 2. **Human Study Scores.**

Methods	ID ↑	Style ↑	Shape ↑	Overall ↑
TI [13]	3.3	2.5	2.9	2.9
Perfusion [61]	2.8	3.1	2.4	2.7
Ours	4.4	3.8	4.2	4.1

6.3. Ablation Study

Design Choices. Our ablation experiments are depicted in Fig. 7 and Tab. 1. *(i)* To judge the impact of explicit editing we exclude it for an experiment, to observe that caricatures lose defining visual characteristics, dropping scores to 0.006 and 0.046 in identity and style similarities, respectively, thus proving its significance. *(ii)* Removing Random mask reconstruction (Sec. 5.3) results in 0.659 (0.553) for ID (style), validating its role in reinforcing robustness against local distortions in personalisation [61]. *(iii)* The replacement from the Mahalanobis distance to applying cosine similarity on the Euclidean distance alleviates the need for cumbersome pre-cached uncentered covariance estimation [61], leading to a more streamlined training process. More importantly, it causes an apparent improvement in visual quality, and a slight increase ($0.05/0.002$ in ID/style) in the similarities as well, thus replacing cosine similarity with a more efficient choice.

Modalities. While the fourth column of Fig. 8 validates our model's precision in preserving identity, the fifth displays our integration of styles with shapes. Finally, the sixth column highlights our Sketch+ID+Style result, achieving high fidelity to input ID [26], sketch [20] and style [27].

Impact of Identity Scale (s). Fig. 6 shows the influence of identity scale s on the generated caricatures. Evidently, a higher s tends to *retain* a higher proportion of identity traits in a caricature and vice-versa. Around the sweet spot below 1.40, users can freely choose this balance as per their own subjective tastes to obtain coherent personalised caricatures [61]. In all our experiments we had set s as 1.2, empirically.

7. Conclusion

In conclusion, our work marks a significant leap in democratising caricature generation, offering individuals an effortless means to craft personalised artworks with minimal input – just a photo and a conceptual sketch. By navigating the delicate balance between abstraction and identity, our proposed Explicit Rank-1 Model Editing and Random Mask Reconstruction, empower users to seamlessly merge their unique identity and desired artistic style in the caricature synthesis process. We emphasise that our intention is not to replace the irreplaceable touch of artists but to remove accessibility barriers, allowing enthusiasts to engage in the creative realm of caricature art. More generally, our contribution underscores the potential for AI to harmoniously collaborate with human creativity, ensuring that art remains a captivating and inclusive expression for all.

References

[1] Hmrishav Bandyopadhyay, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Tao Xiang, Timothy Hospedales, and Yi-Zhe Song. Sketchinr: A first look into sketches as implicit neural representations. In *CVPR*, 2024. 4

[2] Hmrishav Bandyopadhyay, Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. What sketch explainability really means for downstream tasks. In *CVPR*, 2024. 2

[3] Hmrishav Bandyopadhyay, Subhadeep Koley, Ayan Das, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Doodle your 3d: From abstract freehand sketches to precise 3d shapes. In *CVPR*, 2024. 2

[4] Dena Bazazian, Andrew Calway, and Dima Damen. Dual-Domain Image Synthesis Using Segmentation-Guided GAN. In *CVPRW*, 2022. 2

[5] Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Subhadeep Koley, Rohit Kundu, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Doodle It Yourself: Class Incremental Learning by Drawing a Few Sketches. In *CVPR*, 2022. 2

[6] Kaidi Cao, Jing Liao, and Lu Yuan. CariGANs: Unpaired photo-to-caricature translation. *ACM TOG*, 2018. 1, 2, 4, 5, 6, 7

[7] Shu-Yu Chen, Wanchao Su, Lin Gao, Shihong Xia, and Hongbo Fu. DeepFaceDrawing: Deep Generation of Face Images from Sketches. *ACM TOG*, 2020. 6

[8] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. What Can Human Sketches Do for Object Detection? In *CVPR*, 2023. 2

[9] Wenqing Chu, Wei-Chih Hung, Yi-Hsuan Tsai, Yu-Ting Chang, Yijun Li, Deng Cai, and Ming-Hsuan Yang. Learning to Caricature via Semantic Shape Transform. *IJCV*, 2021. 2, 4

[10] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *CVPR*, 2019. 2, 4

[11] Prafulla Dhariwal and Alexander Quinn Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 3

[12] Ziyi Dong, Pengxu Wei, and Liang Lin. DreamArtist: Towards Controllable One-Shot Text-to-Image Generation via Positive-Negative Prompt-Tuning, 2023. 3

[13] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit Haim Bermano, Gal Chechik, and Daniel Cohen-or. An Image is Worth One Word: Personalizing Text-to-Image Generation using Textual Inversion. In *ICLR*, 2023. 2, 3, 4, 6, 7, 8

[14] Julia Gong, Yannick Hold-Geoffroy, and Jingwan Lu. AutoToon: Automatic Geometric Warping for Face Cartoon Generation. In *WACV*, 2020. 2

[15] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. In *NeurIPS*, 2014. 2

[16] Zheng Gu, Chuanqi Dong, Jing Huo, Wenbin Li, and Yang Gao. CariMe: Unpaired Caricature Generation with Multiple Exaggerations. *IEEE T-MM*, 2021. 2

[17] Ligong Han, Yinxiao Li, Han Zhang, Peyman Milanfar, Dimitris Metaxas, and Feng Yang. SVDiff: Compact Parameter Space for Diffusion Fine-Tuning. In *ICCV*, 2023. 3

[18] Haoyu He, Jianfei Cai, Jing Zhang, Dacheng Tao, and Bohan Zhuang. Sensitivity-Aware Visual Parameter-Efficient Fine-Tuning. In *CVPR*, 2023. 3

[19] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Doll'ar, and Ross B Girshick. Masked Autoencoders Are Scalable Vision Learners. In *CVPR*, 2021. 5

[20] Aaron Hertzmann. Why Do Line Drawings Work? A Realism Hypothesis. *Perception*, 2020. 2, 4, 6, 8

[21] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPSW*, 2021. 3, 6

[22] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020. 2, 3

[23] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded Diffusion Models for High Fidelity Image Generation. *JMLR*, 2022. 3

[24] 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 *ICLR*, 2022. 3

[25] Jie Hu, Li Shen, and Gang Sun. Squeeze-and-Excitation Networks. In *CVPR*, 2018. 2

[26] Jing Huo, Wenbin Li, Yinghuan Shi, Yang Gao, and Hujun Yin. WebCaricature: a benchmark for caricature recognition. In *BMVC*, 2018. 1, 2, 5, 6, 7, 8

[27] Wonjong Jang, Gwangjin Ju, Yucheol Jung, Jiaolong Yang, Xin Tong, and Seungyong Lee. Stylecarigan: caricature generation via stylegan feature map modulation. In *SIGGRAPH*, 2021. 1, 2, 4, 5, 6, 7, 8

[28] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *CVPR*, 2019. 2, 4, 7

[29] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *ICCV*, 2020.

[30] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Alias-Free Generative Adversarial Networks. In *NeurIPS*, 2021. 4

[31] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014. 3

[32] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Picture that sketch: Photorealistic image generation from abstract sketches. In *CVPR*, 2023. 4

[33] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. How to Handle Sketch-Abstraction in Sketch-Based Image Retrieval? In *CVPR*, 2024. 2

[34] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Text-

to-Image Diffusion Models are Great Sketch-Photo Matchmakers. In *CVPR*, 2024. 3

[35] Subhadeep Koley, Ayan Kumar Bhunia, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. You'll Never Walk Alone: A Sketch and Text Duet for Fine-Grained Image Retrieval. In *CVPR*, 2024. 4

[36] Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. It's All About Your Sketch: Democratising Sketch Control in Diffusion Models. In *CVPR*, 2024. 3

[37] Nupur Kumari, Bingliang Zhang, Richard Zhang, Eli Shechtman, and Jun-Yan Zhu. Multi-concept customization of text-to-image diffusion. In *CVPR*, 2023. 3

[38] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your diffusion model is secretly a zero-shot classifier. In *ICCV*, 2023. 4

[39] Ilya Loshchilov and Frank Hutter. Decoupled Weight Decay Regularization. In *ICLR*, 2019. 6

[40] Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and Editing Factual Associations in GPT. In *NeurIPS*, 2022. 3, 4

[41] 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. 2, 3, 4, 5, 7

[42] Alexander Quinn 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. In *ICML*, 2022. 3

[43] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. Styleclip: Text-driven manipulation of stylegan imagery. In *ICCV*, 2021. 4

[44] Jingtan Piao, Keqiang Sun, Quan Wang, Kwan-Yee Lin, and Hongsheng Li. Inverting Generative Adversarial Renderer for Face Reconstruction. In *CVPR*, 2021. 2

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 3, 5, 7

[46] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 2022. 3

[47] Esther Robb, Wen-Sheng Chu, Abhishek Kumar, and Jia-Bin Huang. Few-Shot Adaptation of Generative Adversarial Networks. *arXiv preprint arXiv:2010.11943*, 2020. 2

[48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 3, 4, 5, 6, 7

[49] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 3

[50] 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. 2, 3, 4

[51] 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, 2023. 3, 6

[52] Simo Ryu. Low-rank Adaptation for Fast Text-to-Image Diffusion Fine-tuning. https://github.com/cloneofsimo/lora, 2022. 3

[53] 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. In *NeurIPS*, 2022. 3, 4

[54] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs, 2021. 3

[55] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5B: An open large-scale dataset for training next generation image-text models. In *NeurIPSW*, 2022. 3

[56] Jing Shi, Wei Xiong, Zhe Lin, and Hyun Joon Jung. InstantBooth: Personalized Text-to-Image Generation without Test-Time Finetuning, 2023. 3

[57] Yichun Shi, Debayan Deb, and Anil K Jain. Warpgan: Automatic caricature generation. In *CVPR*, 2019. 2, 4, 6, 7

[58] Kihyuk Sohn, Nataniel Ruiz, Kimin Lee, Daniel Castro Chin, Irina Blok, Huiwen Chang, Jarred Barber, Lu Jiang, Glenn Entis, Yuanzhen Li, et al. StyleDrop: Text-to-Image Generation in Any Style. *arXiv preprint arXiv:2306.00983*, 2023. 3

[59] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion. In *NeurIPS*, 2023. 3, 4

[60] Ayush Tewari, Mohamed Elgharib, Florian Bernard, Hans-Peter Seidel, Patrick Pérez, Michael Zollhöfer, and Christian Theobalt. Pie: Portrait image embedding for semantic control. *ACM TOG*, 2020. 2

[61] Yoad Tewel, Rinon Gal, Gal Chechik, and Yuval Atzmon. Key-Locked Rank One Editing for Text-to-Image Personalization. In *SIGGRAPH*, 2023. 2, 3, 4, 5, 7, 8

[62] Nontawat Tritrong, Pitchaporn Rewatbowornwong, and Supasorn Suwajanakorn. Repurposing GANs for One-shot Semantic Part Segmentation. In *CVPR*, 2021. 2

[63] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *Advances in Neural Information Processing Systems*, 2017. 4

[64] Andrey Voynov, Qinghao Chu, Daniel Cohen-Or, and Kfir Aberman. P+: Extended Textual Conditioning in Text-to-Image Generation. *arXiv*, 2023. 3

[65] Tengfei Wang, Ting Zhang, Bo Zhang, Hao Ouyang, Dong Chen, Qifeng Chen, and Fang Wen. Pretraining is All You Need for Image-to-Image Translation. *arXiv preprint arXiv:2205.12952*, 2022. 3

[66] Tianyi Wei, Dongdong Chen, Wenbo Zhou, Jing Liao, Zhentao Tan, Lu Yuan, Weiming Zhang, and Nenghai Yu. HairCLIP: Design Your Hair by Text and Reference Image. In *CVPR*, 2022. 2

[67] Guangxuan Xiao, Tianwei Yin, William T. Freeman, Frédo Durand, and Song Han. FastComposer: Tuning-Free Multi-Subject Image Generation with Localized Attention. *arXiv*, 2023. 3

[68] Xingqian Xu, Jiayi Guo, Zhangyang Wang, Gao Huang, Irfan Essa, and Humphrey Shi. Prompt-Free Diffusion: Taking "Text" out of Text-to-Image Diffusion Models. *arXiv preprint arXiv:2305.16223*, 2023. 3

[69] 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. 3

[70] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023. 3

[71] 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. In *NeurIPS*, 2023. 3