It's All About *Your* Sketch: Democratising Sketch Control in Diffusion Models

Subhadeep Koley[1,2] Ayan Kumar Bhunia[1] Deeptanshu Sekhri[1] Aneeshan Sain[1]
Pinaki Nath Chowdhury[1] Tao Xiang[1,2] Yi-Zhe Song[1,2]
[1]SketchX, CVSSP, University of Surrey, United Kingdom.
[2]iFlyTek-Surrey Joint Research Centre on Artificial Intelligence.
{s.koley, a.bhunia, d.sekhri, a.sain, p.chowdhury, t.xiang, y.song}@surrey.ac.uk
https://subhadeepkoley.github.io/StableSketching

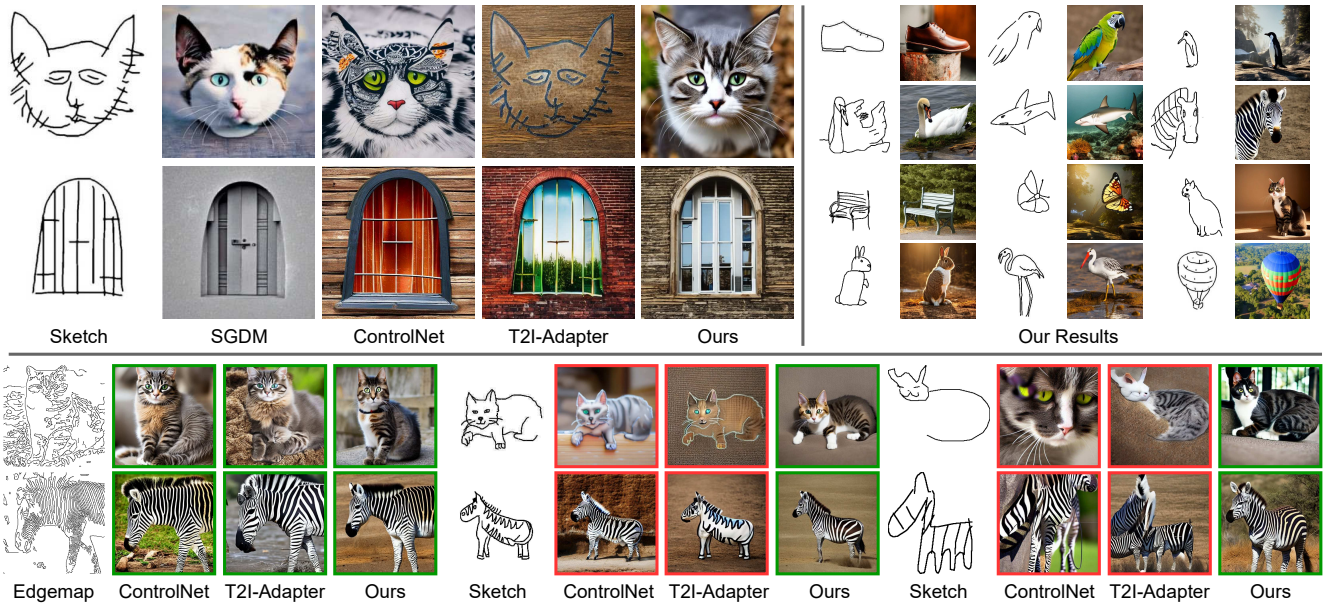

Figure 1. *Top-left*: Comparison of images generated by our method with SGDM [81], ControlNet [90], and T2I-Adapter [55]. *Top-right*: A set of photos generated by our method. *Bottom*: While existing methods [55, 90] generate realistic images from *pixel-perfect edgemaps*, they perform sub-optimally for *freehand abstract sketches*. *(Best view when zoomed in.)*

Abstract

This paper unravels the potential of sketches for diffusion models, addressing the deceptive promise of direct sketch control in generative AI. We importantly democratise the process, enabling amateur sketches to generate precise images, living up to the commitment of "what you sketch is what you get". A pilot study underscores the necessity, revealing that deformities in existing models stem from spatial-conditioning. To rectify this, we propose an abstraction-aware framework, utilising a sketch adapter, adaptive time-step sampling, and discriminative guidance from a pre-trained fine-grained sketch-based image retrieval model, working synergistically to reinforce fine-grained sketch-photo association. Our approach operates seamlessly during inference without the need for textual prompts; a simple, rough sketch akin to what you and I can create suffices! We welcome everyone to examine results presented in the paper and its supplementary. Contributions include democratising sketch control, introducing an abstraction-aware framework, and leveraging discriminative guidance, validated through extensive experiments.

1. Introduction

This paper is dedicated to unlocking the full potential of *your* sketches to control diffusion models [24, 25, 61]. Diffusion models [16, 24, 25, 61] have made a significant impact, empowering individuals to unleash their visual creativity – consider prompts like "astronauts riding a horse on Mars" and other "creative" ones of your own! While prevailing in text-to-image generation [16, 61, 64], recent works [55, 81, 90] have started to question the expressive power of text as a conditioning modality. This shift has led to an exploration of sketches – a modality that offers a degree of fine-grained control that is unparalleled by text [13, 70], resulting in generated content of closer resemblance. The promise is "what you sketch is what you get".

This promise is, however, deceptive. Current works (*e.g.*, ControlNet [90], T2I-Adapter [55]) predominantly focus on curated edgemap-like sketches – you better sketch like a trained artist, otherwise "what you get" will literally be reflecting deformities captured in your ("half-decent") sketch (Fig. 1). The primary goal of this paper is to democratise sketch control in diffusion models, empowering real am-

ateur sketches to generate photo-precise images, ensuring that "what you get" aligns with your *intended sketch*, regardless of how well you drew it! To achieve this, we draw insights from the sketch community [37, 38, 65, 67, 87] and introduce, for the first time, an awareness of sketch abstraction (as a result of varying drawing skills) into the generative process. This novel approach permits sketches of different abstraction levels to guide the generation process while maintaining output fidelity.

We conduct a pilot study to reaffirm the necessity of our research (Sec. 4). In which, we identify that the deformed output of existing sketch-conditional diffusion models stems from their *spatial-conditioning* approach – they directly translate sketch contours into the output photo domain, therefore producing deformed output. Conventional means of controlling the influence of spatial sketch-conditioning on the final output via weighing factors [55, 81] or sampling tricks [90], however, require careful tuning. Reducing output deformity by assigning less weight to the sketch-conditioning often makes the output more coherent with the textual description, thus reducing its fidelity to the guiding sketch; yet, assigning higher weight to the textual prompt introduces lexical ambiguity [71]. On the contrary, avoiding lexical ambiguity by assigning a higher weight to the guiding sketch almost always produces deformed and non-photorealistic outputs [55, 81, 90]. Last but not least, the sweet spot between the conditioning weights is different for different sketch instances (as seen in Fig. 2).

As such, our goal is to craft an effective sketch-conditioning strategy that not only operates without *any* textual prompts during inference but is also *abstraction-aware*. At the core of our work is a sketch adapter that transforms an input sketch into its *equivalent textual embedding*, directing the denoising process of the diffusion model via cross-attention. Through the use of a smart time-step sampling strategy, we ensure the adaptability of the denoising process to the abstraction level of the input sketch. Additionally, by capitalising on the pre-trained knowledge of an off-the-shelf [66] fine-grained sketch-based image retrieval (FG-SBIR) model, we incorporate discriminative guidance into our system for fine-grained sketch-photo association. Unlike widely used external classifier-guidance [16], our proposed discriminative guidance mechanism does not require any specifically trained classifier capable of classifying *both* noisy and real data. Lastly, even though our inference pipeline *does not* rely on textual prompts, we use synthetically generated textual prompts during training to learn the sketch adapter with the limited sketch-photo paired data.

Our contributions are: *(i)* we democratise sketch control, enabling real amateur sketches to generate accurate images, fulfilling the promise of "what you sketch is what you get". *(ii)* we introduce an abstraction-aware framework that overcomes limitations of text prompts and spatial-conditioning. *(iii)* we leverage discriminative guidance through a pre-trained FG-SBIR model for fine-grained sketch-fidelity. Extensive experiments validate the effectiveness of our method in addressing existing limitations in this domain.

2. Related Works

Diffusion Models for Vision Tasks. Diffusion models [24, 25, 74] have now become the gold-standard for different controllable image generation frameworks like DALL-E [57], Imagen [64], T2I-Adapter [55], ControlNet [90], etc. Besides image generation, several methods like Dreambooth [63], Imagic [32], Prompt-to-Prompt [22], SDEdit [52], SKED [54] extend it for realistic image editing. Beyond image generation and editing, diffusion model is also used in several downstream vision tasks like recognition [43], semantic [2] and panoptic [84] segmentation, image-to-image translation [79], medical imaging [15], image correspondence [78], retrieval [39], etc.

Sketch for Visual Content Creation. Following its success in sketch-based image retrieval (SBIR) [3, 11, 66], sketches are now being used in other downstream tasks like saliency detection [6], augmented reality [50, 51], medical image analysis [35], object detection [14], class-incremental learning [4], etc. Apart from the plethora of sketch-based $2D$ and $3D$ image generation and editing frameworks [21, 36, 47, 54, 55, 60, 81, 82, 90], sketches are also getting significant traction in other visual content creation tasks like animation generation [73] and inbetweening [72], garment design [12, 46], caricature generation [10], CAD modelling [44, 88], anime editing [28], etc.

Sketch-to-Image (S2I) Generation. Prior GAN-based S2I models typically leverage either contextual loss [49], multi-stage generation [19], etc. or performs latent mapping [36, 60] on top of pre-trained GANs. Among diffusion-based frameworks, PITI [82] trains a dedicated encoder to map the guiding sketch to the pre-trained diffusion model's latent manifold, SDEdit [52] sequentially adds noise to the guiding sketch and iteratively denoise it based on a text prompt, while SGDM [81] trains an MLP that maps the latent feature of the noisy images to the guiding sketches in order to force the intermediate noisy images to closely follow the guidance sketches. Among more recent multi-conditional (*e.g.*, depth map, colour palate, key pose, etc.) frameworks, ControlNet [90] learns to control a frozen diffusion model by creating a trainable copy of its UNet encoders and connects it with the frozen model with *zero-convolution* [90], while T2I-Adapter [55] learns an encoder to extract features from the guidance signal (*e.g.*, sketch) and conditions the generation process by adding the guidance features with the intermediate UNet features at each scale. While existing methods can generate photorealistic images from precise edgemaps, they struggle with abstract freehand sketches (see Fig. 1). Furthermore, it is

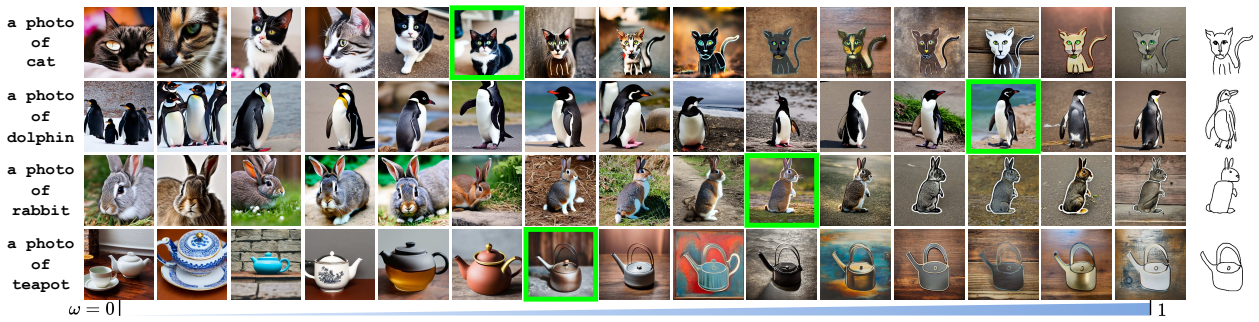

Figure 2. Images generated by T2I-Adapter [55] for different sketch-guidance factors ($\omega \in [0, 1]$). Determining the optimum ω to obtain an ideal balance (green-bordered) between *photorealism* and *sketch-fidelity* requires manual intervention and is sample-specific. A high value of ω works well for less deformed sketches, while the same for an abstract sketch produces deformed outputs and vice-versa.

noteworthy that almost all of the diffusion-based S2I models [52, 55, 81, 82, 90] rely heavily on highly-engineered and detailed textual prompts.

3. Revisiting Diffusion Model (DM)

Overview. Diffusion models comprises two complementary random processes *viz.* *"forward"* and *"reverse"* [25] diffusion. Forward diffusion process iteratively adds Gaussian noise of varying magnitude to a clean training image $\mathbf{x}_0 \in \mathbb{R}^{h\times w\times 3}$ for t time-steps to yield a noisy image $\mathbf{x}_t \in \mathbb{R}^{h\times w\times 3}$ as:

$$\mathbf{x}_t = \sqrt{\bar{\alpha}_t}\mathbf{x}_0 + (\sqrt{1-\bar{\alpha}_t})\epsilon \tag{1}$$

where, $\epsilon \sim \mathcal{N}(0, \mathbf{I})$, $t \sim U(0, T)$, and $\{\alpha_t\}_1^T$ is a pre-defined noise schedule with $\bar{\alpha}_t = \prod_{i=1}^{t} \alpha_i$ [25]. Reverse diffusion process trains a modified denoising UNet [62] $\mathcal{F}_\theta(\cdot)$, that estimates the input noise $\epsilon \approx \mathcal{F}_\theta(\mathbf{x}_t, t)$ from the noisy image $\mathbf{x}_t$ at each time-step t. $\mathcal{F}_\theta$ being trained with an l_2 loss [25] can reverse the effect of the forward diffusion procedure. During inference, starting from a random $2D$ noise $\mathbf{x}_T$ sampled from a Gaussian distribution, $\mathcal{F}_\theta$ is applied iteratively (for T time-steps) to denoise $\mathbf{x}_t$ at each time-step t to get a cleaner image $\mathbf{x}_{t-1}$, eventually leading to a cleanest image $\mathbf{x}_0$ of the original target distribution [25].

The unconditional denoising diffusion process could be made "conditional" by influencing the $\mathcal{F}_\theta$ with auxiliary conditioning signals d (*e.g.*, textual description [58, 61, 64], etc.). Thus, $\mathcal{F}_\theta(\mathbf{x}_t, t, d)$ could perform denoising on $\mathbf{x}_t$ while being guided by d via cross-attention [61].

Latent Diffusion Model. Unlike standard diffusion models [16, 25], *Latent Diffusion Model* [61] (*a.k.a.* Stable Diffusion–SD) performs denoising diffusion on the latent space for faster and more stable training [61]. SD first *trains an autoencoder* (consists of an encoder $\mathcal{E}(\cdot)$ and a decoder $\mathcal{D}(\cdot)$ in series) to convert the input image $\mathbf{x}_0 \in \mathbb{R}^{h\times w\times 3}$ to its latent representation $\mathbf{z}_0 = \mathcal{E}(\mathbf{x}_0) \in \mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times d}$. Later, SD *trains a modified denoising UNet* [62] $\epsilon_\theta(\cdot)$ to perform denoising directly on the latent space. The textual prompt d upon passing through a CLIP textual encoder [56] $\mathbf{T}(\cdot)$ produces the corresponding token-sequence that influences the intermediate feature maps of the UNet via cross-attention [61]. SD trains with an l_2 loss as:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{\mathbf{z}_t, t, d, \epsilon}(||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathbf{T}(d))||_2^2) \tag{2}$$

During inference, SD *discards* $\mathcal{E}(\cdot)$, directly sampling a noisy latent $\mathbf{z}_T$ from a Gaussian distribution [61]. It then estimates noise from $\mathbf{z}_T$ iteratively for T iterations via ϵ_θ (conditioned on d) to obtain a clean latent $\hat{\mathbf{z}}_0$. The frozen decoder generates the final image as: $\hat{\mathbf{x}}_0 = \mathcal{D}(\hat{\mathbf{z}}_0)$ [61].

4. What's wrong with Sketch-to-Image DM?

Recent controllable image generation methods like ControlNet [90], T2I-Adapter [55], etc. offer extreme photorealism, supporting different conditioning signals (*e.g.*, depth map, label mask, edgemap, etc.) However, conditioning the same from sparse *freehand sketches* is often sub-optimal (Fig. 1).

Sketch *vs.* Other Conditional Inputs. Sparse and binary freehand sketches while good for providing *fine-grained* spatial cues [6, 14, 89], often depict significant shape-deformity [17, 23, 65] and hold far less contextual information [79] than other *pixel-perfect* conditioning signals like depth maps, normal maps, or pixel-level segmentation masks. Hence, conditioning from freehand sketches is non-trivial and needs to be handled uniquely unlike the rest of the pixel-perfect conditioning signals.

Sketch *vs.* Text Conditioning: A Trade-off. Previous S2I diffusion models [55, 81, 90] exhibit two major challenges. *Firstly*, quality of generated outputs being highly dependent on precise and accurate textual prompts [90], inconsistencies or lack of suitable prompts can negatively impact (Fig. 3) the results [55, 90]. *Secondly*, ensuring a balance between the influence of sketch and text-conditioning on the final output requires manual intervention, which can be challenging. Adjusting the weighting of these factors often results in a *trade-off* between output's coherence with the text and fidelity to the sketch [55]. In some cases, giving higher weight to text can lead to lexical ambiguity [71], while prioritising sketch tends to produce distorted and non-photorealistic results [55, 81]. Achieving photorealistic output from existing S2I DMs [55, 81] thus demands *meticu-*

lous fine-tuning of these weights, where the optimal balance varies for different sketch instances as seen in Fig. 2.

Figure 3. Passing null prompt (*i.e.*, " ") in existing [55, 81, 90] sketch-conditioned DMs significantly distorts the output quality.

Problems with Spatial-Conditioning for Sketches. We identify that the deformed and non-photorealistic (*e.g.*, edge-bleeding in Fig. 2) outputs of existing sketch-conditional DMs [55, 81, 90] are primarily a consequence of their *spatial-conditioning* approach. T2I-Adapter [55] directly integrates the *spatial features* of the conditioning-sketch into the UNet encoder's feature maps, while ControlNet [90] applies this to skip connections and middle blocks. SGDM [81], on the other hand, projects the latent features of noisy images to *spatial* edgemaps guiding the denoising process towards following the edgemaps. Additionally, these models are trained and tested with *synthetically-generated* [7, 76, 83] edgemaps/contours rather than *real* freehand sketches. Instead, we aim to devise an effective conditioning strategy for *real* freehand sketches while ensuring that the output faithfully captures an end-users' *semantic intent* [36] without any deformities.

5. Proposed Methodology

Overview. We aim to eliminate *spatial sketch-conditioning* by converting the input sketch into an *equivalent fine-grained textual embedding*, thereby preserving users' semantic-intent without *pixel-level* spatial alignment. Consequently, our method would alleviate issues pertaining to spatial distortions (*e.g.*, deformed shapes, edge-bleeding, etc.) while maintaining *fine-grained fidelity* to the input sketch. We introduce three salient designs (Fig. 4) – *(i)* fine-grained discriminative loss for maintaining the *fine-grained* sketch-photo correspondence (Sec. 5.2). *(ii)* guiding our training process with textual prompts (*not* used during inference), as a means of *super-concept* preservation (Sec. 5.3). Finally, *(iii)* unlike the *uniform* time-step (t) sampling of prior arts [81, 90], we introduce a *sketch-abstraction-aware* t-sampling (Sec. 5.4). For a highly abstract sketch, a higher probability is assigned to larger t and vice-versa.

5.1. Sketch Adapter

Aiming to mitigate the evident disadvantages (Sec. 4) of direct *spatial-conditioning* approach of existing sketch-conditional diffusion models (*e.g.*, ControlNet [90], T2I-Adapter [55], etc.), we take a parallel approach to "sketch-condition" the generation process via cross-attention. In that, instead of treating the input sketches *spatially*, we encode them as a sequence of feature vectors [42] as an

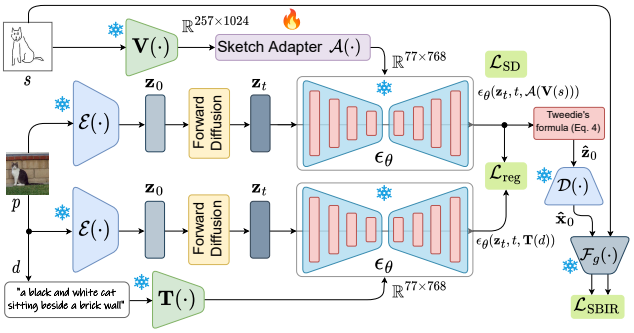

Figure 4. Our overall training pipeline. *(More in the text.)*

equivalent fine-grained textual embedding. Direct spatial-conditioning enforces the model to *remember* the contextual information rather than *understanding* it [85]. This results in a direct translation of the strong sketch features (*e.g.*, stroke boundaries) into the output photo. To overcome this, we aim to increase the hardness of the problem by compressing the spatial sketch input to a *bottlenecked-representation* via sketch adapter.

In particular, given a sketch s, we use a pre-trained CLIP [56] ViT-L/14 image encoder $\mathbf{V}(\cdot)$ to generate its patch-wise sketch embedding $\mathbf{s} = \mathbf{V}(s) \in \mathbb{R}^{257\times1024}$. Our sketch adapter $\mathcal{A}(\cdot)$ consists of 1-dimensional convolutional and vanilla attention [80] modules followed by FC layers. The convolutional and FC layers handle the dimension mismatch between text and sketch-embedding (*i.e.*, $\mathbb{R}^{257\times1024} \rightarrow \mathbb{R}^{77\times768}$), whereas the attention module tackles the large sketch-text domain gap. The patch-wise sketch embedding $\mathbf{s}$ upon passing through $\mathcal{A}(\cdot)$ generates the equivalent textual embedding as $\hat{\mathbf{s}} = \mathcal{A}(\mathbf{s}) \in \mathbb{R}^{77\times768}$. Now replacing the textual conditioning in Eq. (2) with our sketch adapter conditioning, the modified loss objective becomes:

$$\mathcal{L}_{\text{SD}} = \mathbb{E}_{\mathbf{z}_t,t,s,\epsilon}(||\epsilon - \epsilon_\theta(\mathbf{z}_t, t, \mathcal{A}(\mathbf{V}(s)))||_2^2) \qquad (3)$$

Once trained, the sketch adapter efficiently converts an input sketch s into its equivalent textual embedding $\hat{\mathbf{s}}$, which through cross-attention controls the denoising process of SD [61]. Nonetheless, conditioning solely via the proposed sketch adapter poses multiple challenges – *(i)* sparse free-hand sketches and pixel-perfect photos depict a huge domain gap. The standard l_2 loss [61] of a text-to-image diffusion model is not enough to ensure a *fine-grained* matching between sketch and photo. *(ii)* training a robust sketch adapter from the *limited* available sketch-photo pairs is difficult. Consequently, during training, we aim to use *pseudo texts* as a learning signal to guide the training of our sketch adapter. Please note, our inference pipeline *does not* involve any textual prompts. *(iii)* the sketch adapter treats *all* sketch samples *equally* regardless of their *abstraction levels*. While this equal treatment might suffice for dense pixel-level conditioning, it might be inadequate for sparse sketches, as different sketches depicting different abstraction levels are not *semantically-equal* [5, 86].

5.2. Fine-Grained Discriminative Learning

To ensure a *fine-grained matching* between sparse freehand sketches and pixel-perfect photos, we utilise a pre-trained fine-grained (FG) SBIR model [66] $\mathcal{F}_g(\cdot)$. A photo sits close to its paired sketch in a pre-trained FG-SBIR model's *discriminative* latent embedding space compared to other unpaired ones [66]. Previous attempts at guiding the diffusion process with external discriminative models include classifier-guidance [16] that require a pre-trained fixed-class classifier capable of classifying *both* noisy and real data [16] to guide the denoising procedure [16]. However, as our frozen FG-SBIR model is not trained on *noisy* data, it requires a *clean* image at each t, to perform in an *off-the-shelf* manner. Now, for each t, as the denoiser estimates that noise $\epsilon_t \approx \epsilon_\theta(\mathbf{z}_t, t, \mathcal{A}(\mathbf{V}(s)))$, which was added to $\mathbf{z}_0$ to get $\mathbf{z}_t$ during forward diffusion, we can use Eq. (1) to recreate $\mathbf{z}_0$ from ϵ_t. Specifically, we utilise Tweedie's formula [34] to estimate [1, 40, 85] the clean latent image $\hat{\mathbf{z}}_0$ from the t^{th}-step noisy latent $\mathbf{z}_t$ in a single-step for efficient training as:

$$\hat{\mathbf{z}}_0(\mathbf{z}_t) := \frac{\mathbf{z}_t - \sqrt{1-\bar{\alpha}_t}\,\epsilon_\theta(\mathbf{z}_t, t, \mathcal{A}(\mathbf{V}(s)))}{\sqrt{\bar{\alpha}_t}} \tag{4}$$

$\hat{\mathbf{z}}_0$ upon passing through SD's [61] frozen VAE decoder $\mathcal{D}(\cdot)$ approximates the clean image $\hat{\mathbf{x}}_0$ (Sec. 3). To learn the sketch adapter $\mathcal{A}$, we use a discriminative SBIR loss that calculates cosine similarity $\delta(\cdot,\cdot)$ between s and $\hat{\mathbf{x}}_0$ as:

$$\mathcal{L}_{\text{SBIR}} = 1 - \delta\left(\mathcal{F}_g(s) \cdot \mathcal{F}_g(\hat{\mathbf{x}}_0)\right) \tag{5}$$

5.3. Super-concept Preservation Loss

An inherent complementarity exists between *sketch* and *text* [13]. A textual caption of an image can correspond to multiple *plausible* photos in the embedding space. Adding a sketch *with it* however, narrows down the scope to a *particular* image [13, 70] (*i.e.*, fine-grained). We posit that a textual description being less fine-grained than a sketch [13, 75, 85], acts as a *super-concept* of the corresponding sketch. Although we *do not* use any textual prompt during inference, we aim to use them during training of our sketch adapter. Text-to-image diffusion models being trained on a large corpus of text-image pairs [61], implicitly hold superior text-to-image generation capability (although *not* fine-grained [18]). We thus aim to use this super-concept knowledge from textual descriptions to distil the large-scale text-to-image knowledge of a pre-trained SD to train our sketch-adapter with *limited* sketch-photo paired data.

As our sketch-photo (s, p) dataset [69] lacks paired textual captions, we use a pre-trained state-of-the-art image captioner [45] to synthetically generate caption d for every ground truth photo p. Then, at each t, the noise predicted through *text-conditioning* $(\mathbf{T}(d))$ acts as a reference to calculate a regularisation loss to learn the sketch adapter $\mathcal{A}$ as:

$$\mathcal{L}_{\text{reg}} = ||\epsilon_\theta(\mathbf{z}_t, t, \mathbf{T}(d)) - \epsilon_\theta(\mathbf{z}_t, t, \mathcal{A}(\mathbf{V}(s)))||_2^2 \tag{6}$$

5.4. Abstraction-aware Importance Sampling

Existing literature [26, 27, 55, 85] indicates that during the denoising process, high-level semantic structures of the output image tend to manifest in the early stages, while finer appearance details emerge later. Synthetic pixel-perfect conditioning signals (*e.g.*, depth map [59], key pose [8], edgemap [7], etc.) exhibit minimal subjective abstraction [23]. In contrast, human-drawn freehand sketches exhibit varying abstraction levels, influenced by factors like skill, style, and subjective interpretation [65, 67]. Thus, uniform time-step sampling [27] for abstract sketches may compromise output generation quality and sketch-fidelity. Hence, we propose adjusting the time-step sampling procedure based on the input sketch's abstraction level [87]. For highly abstract sketches, we skew the sampling distribution to emphasise the later t values that govern the high-level semantics in the output. Instead of sampling the time-step from uniform distribution $t \sim U(0, T)$, we sample from:

$$\mathcal{S}_\omega(t) = \frac{1}{T}\left(1 - \omega\cos\frac{\pi t}{T}\right) \tag{7}$$

where, $\mathcal{S}_\omega(\cdot)$ is our *abstraction-aware t-sampling function*, where increasing or decreasing $\omega \in (0, 1]$, controls the skewness of this sampling probability density function. Pushing ω towards 1 increases the probability of sampling a larger t value (Fig. 5). We aim to make this skewness-controlling ω value sketch-abstraction specific.

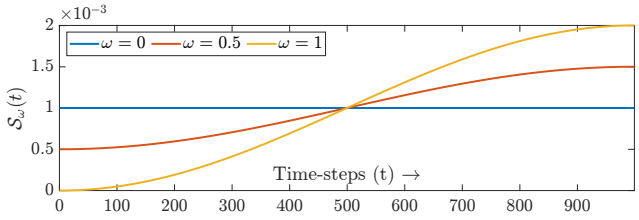

Figure 5. Abstraction-aware t-sampling function for different ω.

Now the question remains as to how we can quantify the abstraction level of a freehand sketch. Taking inspiration from [87], we design a CLIP [56]-based (a generic classifier) sketch classifier with a MagFace [53]-based loss where the l_2-norm of a sketch feature $\mathbf{a} \in [0, 1]$, denotes how closely it sits from its respective class-centre. While $\mathbf{a} \rightarrow 1$ represents edgemap-like *less abstract* sketches, $\mathbf{a} \rightarrow 0$ denotes *highly-abstract and deformed* ones. We posit that edgemaps being less deformed (*i.e.*, easier to classify), will implicitly stay close to their respective class centres in the latent space. Whereas, freehand sketches being highly abstract and deformed (*i.e.*, harder to classify), will be placed away from their corresponding class centres. We thus train the sketch classifier with sketches and synthesised [9] edgemaps of the associated photos from Sketchy [69], using our classification loss:

$$\mathcal{L}_{\text{abs}} = -\log\frac{e^{s\cos(\theta_{y_i} + m(\mathbf{s}_i))}}{e^{s\cos(\theta_{y_i} + m(\mathbf{s}_i))} + \sum_{j \neq y_i} e^{s\cos\theta_j}} + \lambda_g g(\mathbf{s}_i) \tag{8}$$

where s is a global scalar value, θ_{y_i} is the cosine similarity between extracted global visual feature (from CLIP [56] visual encoder) of the i^{th} sketch sample $\mathbf{s}_i = \mathbf{V}(s_i) \in \mathbb{R}^d$ with l_2-normalisation, and j^{th} class centre $w_j \in \mathbb{R}^d$ computed from ground truth class labels by CLIP [56] text encoder. $m(\mathbf{s}_i)$ is the magnitude-aware margin parameter $m(\mathbf{s}_i) = \frac{(u_m - l_m)}{(u_a - l_a)l_a + l_m}$, where l_m, u_m denotes the lower and upper bounds of the margin, and l_a, u_a denotes that of the feature magnitude. $g(\mathbf{s}_i)$ is a hyper-parameter (λ_g)-controlled regularisation term (see [53] for more details). With the trained classifier, given a sketch s, the *scalar abstraction score* $\mathbf{a} \in [0, 1]$ is given by the l_2-norm of the extracted sketch feature $\mathbf{V}(s)$. To keep parity with ω, we complement $\mathbf{a}$ to get the sketch instance-specific $\omega \leftarrow (1 - \mathbf{a})$, followed by empirically clipping ω in the range $[0.2, 0.8]$.

In summary, we train the sketch adapter $\mathcal{A}(\cdot)$ using sketch-abstraction-aware t-sampling with a total loss of $\mathcal{L}_{\text{total}}$=$\lambda_1\mathcal{L}_{\text{SD}}$+$\lambda_2\mathcal{L}_{\text{SBIR}}$+$\lambda_3\mathcal{L}_{\text{reg}}$. During inference, we compute the abstraction score of the input sketch, taking l_2-norm of classifier feature. Based on the abstraction level, we perform t-sampling. The input sketch passing through $\mathcal{A}$ controls the diffusion procedure and generates the output.

6. Experiments

Dataset and Implementation Details. We train and evaluate our model on the Sketchy dataset [69] containing $12,500$ images from 125 categories with at least 5 sketches per image with *fine-grained* association. For training and evaluation, we split this dataset in 90:10. We use Stable Diffusion v1.5 [61] in all experiments with a CLIP [56] embedding dimension $d = 768$. The sketch adapter is trained with a learning rate of 10^{-4}, keeping the SD model, FG-SBIR backbone, and CLIP encoders frozen. We train our model for 50 epochs using AdamW [48] optimiser with 0.09 weight decay, and batch size of 8. Values of $\lambda_{1,2,3}$ are set to 1, 0.5, and 0.1, empirically.

Evaluation Metrics. Following [36, 55, 90], we quantitatively evaluate the generation quality and sketch-fidelity with four metrics – *Frechèt Inception Distance-InceptionV3 (FID-I) [31] and CLIP (FID-C)* [41] calculates the similarity between generated and real images using pre-trained InceptionV3 [77] and CLIP [56] ViT-B/32 models respectively. Lower values of FID-I and FID-C depict better generation quality. We measure the output image's fidelity to the input sketch using *Fine-Grained Metric (FGM)* [36] which computes the cosine similarity between them via a pre-trained FG-SBIR model [66], where a higher value denotes better fine-grained correspondence. Additionally, we also perform a human study to collect *Mean Opinion Score (MOS)* [29]. Here, we asked 25 *non-artist* users to draw 40 sketches each, and rate the generated photos on a discrete scale (interval=0.5) of $[1, 5]$ (worst to best) based on *output photorealism* and *sketch-fidelity*. For each method, we compute the final MOS by averaging all its 1000 MOS values.

Competitors. We compare against different diffusion and GAN-based state-of-the-art (SoTA) S2I models and two baselines. *(i) Sketch-only Baselines:* To alleviate the necessity of text, **B-Classification** first trains a prompt learning-based sketch classifier [33] that classifies every sketch into one of the predefined classes. From predicted class labels, it forms a textual prompt (*i.e.*, "`a photo of [CLASS]`") to generate images using a frozen text-to-image SD model [61]. Given the input sketches, **B-Captioning** first generates detailed captions using a pre-trained image captioner [45] from their paired photos, which are then used to generate images from a frozen SD model [61]. *(ii) SoTAs:* Among diffusion-based SoTAs, we compare with **ControlNet** [90], **T2I-Adapter** [55], **SGDM** [81], and **PITI** [82]. We also compare qualitatively against two GAN-based S2I paradigms *viz.* **Pix2Pix** [30] and **CycleGAN** [91]. While we train ControlNet [90], T2I-Adapter [55], and PITI [82] on the entire Sketchy [69] train set, we train pix2pix [30], and CycleGAN [91] individually for each of the depicted classes (Fig. 6) from scratch with Sketchy [69] sketch-photo pairs. We only perform a qualitative comparison with SGDM [81] by taking the results directly from the paper, as their model weights/code are unavailable. Notably, for diffusion-based SoTAs [55, 82, 90], we use an *additional* fixed textual prompt "`a photo of [CLASS]`", replacing `[CLASS]` with class-labels of respective input sketches.

6.1. Performance Analysis & Discussion

Result Analysis. Among GAN-based methods, pix2pix [30] and CycleGAN [91] depict visible deformities (Fig. 6) mostly due to their weaker [16] GAN-based generator, compared to an internet-scale pre-trained SD model [61]. Among diffusion-based SoTAs, although SGDM [81] generates plausible colour schemes and styles, outputs exhibit substantial deformations (Fig. 1). A similar observation can be made for PITI [82], where generated images look non-photorealistic with pronounced edge-adherence (Fig. 6). Whereas, edge-bleeding (Fig. 6) is quite frequent for T2I-Adapter [55]. ControlNet [90] surpasses PITI [82], SGDM [81], and T2I-Adapter [55] in terms of photorealism but mostly follows the input sketch boundaries (Fig. 6). Contrarily, images generated by our method are more photorealistic with fewer deformities, capturing semantic-intent without transmitting edge boundaries in the output. Quantitative results presented in Tab. 1 show B-Caption to surpass B-Classification (by 0.11 FGM) thanks to the comparatively higher [45] generalisation potential of the captioning model [45] than the generic sketch classifier [33]. Nonetheless, our method exceeds these baselines both in terms of generation quality and sketch-fidelity with an FID-C of 16.20 and FGM of 0.81. Due to its superior conditioning strategy, ControlNet [90] achieves the lowest FID-I

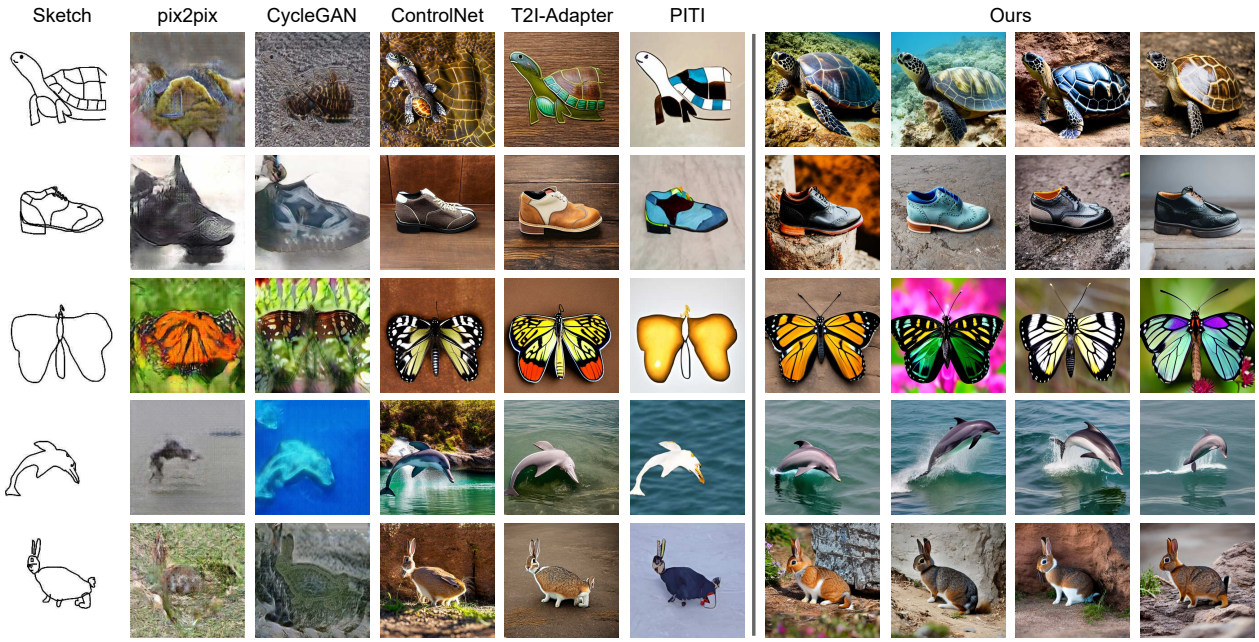

Figure 6. Qualitative comparison with SoTA sketch-to-image generation models on Sketchy [69]. For ControlNet [90], T2I-Adapter [55], and PITI [82], we use the fixed prompt "`a photo of` [`CLASS`]", with [`CLASS`] replaced with corresponding class-labels of the input sketches.

among all prior SoTAs (Tab. 1). Although less pronounced in terms of FID-I/FID-C, our method offers the highest *fine-grained sketch-fidelity* with 23.45% FGM improvement of over ControlNet [90]. Finally, thanks to the photorealistic generation quality and fine-grained sketch correspondence, our method surpasses competitors in terms of MOS value from user-study with an average 1.36 ± 0.2 point improvement. Notably, unlike ours, image generation via diffusion-based competitors needs textual prompts, the absence of which results in much worse output quality (Fig. 3).

Table 1. Benchmarks on the Sketchy [69] dataset.

Methods	FID-I ↓	FID-C ↓	FGM ↑	MOS ↑ $\mu \pm \sigma$
ControlNet [90]	26.68	21.22	0.62	3.68±0.2
T2I-Adapter [55]	26.94	18.92	0.56	3.11±0.6
PITI [82]	84.71	25.85	0.23	2.64±0.3
B-Classification	28.93	19.01	0.36	3.13±0.2
B-Captioning	28.31	18.81	0.47	3.21±0.4
Proposed	**25.07**	**16.20**	**0.81**	**4.52±0.1**

Generalisation Potential. As our method alleviates the direct spatial influence of input sketches in the denoising process, it enables generalisation across multiple dimensions. Fig. 7 shows that our sketch-adapter trained on Sketchy, generalises well on random sketch samples from TU-Berlin [17] and QuickDraw [20] datasets, on synthetically generated [7] edgemaps, and to different stroke-styles. Furthermore, as our sketch adapter does not distort the original text-to-image pre-training of the frozen SD model, the same adapter could be used to perform sketch-conditional generation from other versions of the SD model (Fig. 8).

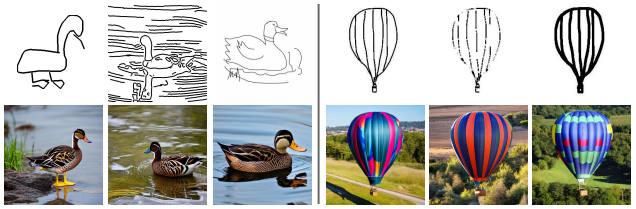

Figure 7. Examples showing generalisation potential across different *datasets* (left) and *stroke-styles* (right).

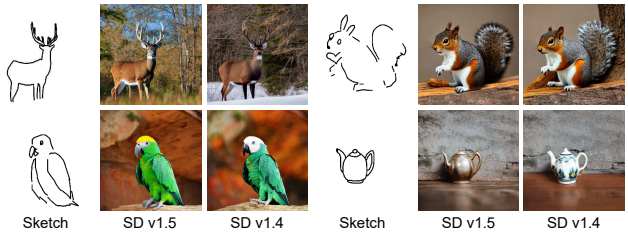

Figure 8. Illustration of *cross-model* generalisation. Our method trained with SD v1.5 [61], performs well on other unseen SD variants (*e.g.*, v1.4) without further fine-tuning.

Robustness and Sensitivity. Amateur freehand sketching often introduces irrelevant and noisy strokes [5]. We thus demonstrate our model's resilience to such strokes by progressively adding them during inference, and assessing its performance. On the other hand, to judge our model's stability against partially-complete sketches, we render input sketches at $\{25, 50, 75, 100\}\%$ prior to generation. As our method is devoid of *direct* spatial-conditioning, outputs remain relatively stable (Fig. 9) even for spatially distorted sketches (*e.g.*, noisy or partially-complete).

Fine-grained Semantic Editing. Harnessing the large-scale pre-training of the frozen SD model [61], our method

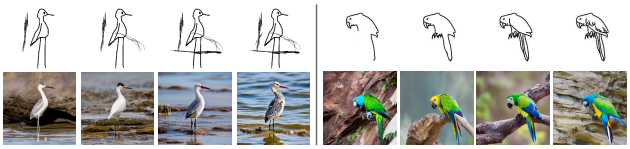

Figure 9. Examples depicting the effect of adding *noisy strokes* (left) and generation from *partially-completed* sketches (right).

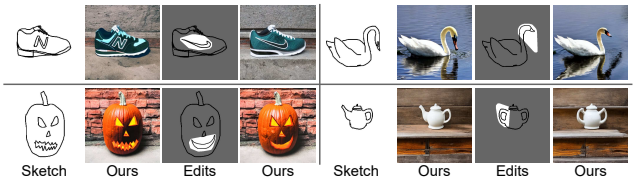

Figure 10. Our method seamlessly transfers local semantic edits on input sketches into output photos. *(Best view when zoomed in.)*

enables *fine-grained* semantic editing. Here, fixing the generation seed, and performing local semantic edits in the sketch-domain produces seamless edited images (Fig. 10).

6.2. Ablation on Design

[i] Importance of Sketch Adapter. Our sketch adapter (Sec. 5.1) converts an input sketch to its corresponding *textual equivalent embedding*. To judge its efficacy, we replace it with simple convolutional and FC-layers converting the $\mathbb{R}^{257\times1024}$ sketch embedding to equivalent $\mathbb{R}^{77\times768}$ textual embedding. Although less pronounced in FID scores, the FGM score plummets substantially (49.38%) in case of **w/o Sketch adapter** (Tab. 2), indicating the significance of the proposed adapter in maintaining high sketch-fidelity.

[ii] Why Discriminative Learning? Fine-grained discriminative loss (Eq. 5) helps the conditioning process by distilling knowledge learned inside a pre-trained FG-SBIR model. As seen in Tab. 2, a noticeable FGM drop (44.44%) for **w/o Discriminative learning** indicates that fine-grained sketch-conditioning is incomplete without explicit discriminative learning via $\mathcal{L}_{\text{SBIR}}$.

[iii] Does Abstraction-aware Importance Sampling help? Unlike existing sketch-conditional DMs, we take freehand sketch abstraction into account via *abstraction-aware* t-sampling. Omitting it results (Tab. 2) in a sharp increase in FID-I scores (26.64%). We hypothesise that in absence of the proposed adaptive t-sampling, the system treats *all* sketches *equally*, regardless of their abstraction level, resulting in sub-optimal performance.

[iv] Impact of Super-concept Preservation. Although our inference procedure *does not* use any textual prompt, we employ them during our training process to facilitate the preservation of *super-concepts*. Eliminating this again destabilises the system causing an additional 15.06% and 17.28% decline in FID-C and FGM scores (Tab. 2). This justifies our incorporation of synthetic text prompts during training, as it aligns well with the original text-to-image generation objective of the pre-trained SD model [61]. Visual ablation results are presented in Fig. 11.

Figure 11. Visual ablation of different design components.

Table 2. Ablation on design.

Methods	FID-I ↓	FID-C ↓	FGM ↑
w/o Sketch adapter	29.23	20.34	0.41
w/o Discriminative learning	29.14	19.97	0.45
w/o Super-concept preservation	27.21	18.64	0.67
w/o Abs.-aware t-sampling	31.75	23.17	0.55
Ours (SD v1.4)	26.12	17.09	0.77
Ours-full	**25.07**	**16.20**	**0.81**

6.3. Failure Cases & Future Works

Despite showcasing superior generation quality without significant deformations, our method has a few limitations. For Instance, it sometimes struggles to determine the correct class of the input due to *categorical-ambiguity*, especially when two different objects look very similar shape-wise (Fig. 12) in their *abstract* and *deformed* sketch forms (*e.g.*, apple *vs*. pear, guitar *vs*. violin). In future, we aim to extend our method with the flexibility to include additional class labels. The sketch+label *composed-conditioning* [68] might mitigate the categorical-ambiguity of confusing classes.

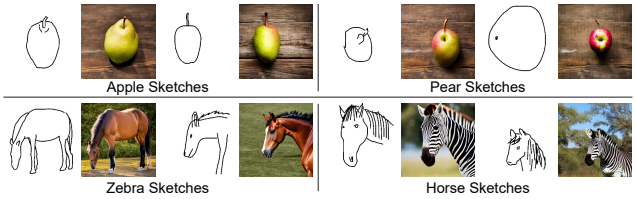

Figure 12. Failure cases where sketches from certain classes (*e.g.*, zebra) produce images from other similar-looking classes (*e.g.*, horse) or vice-versa. Please note that we *do not* use text prompts.

7. Conclusion

Our work takes a significant step towards democratising sketch control in diffusion models. We exposed the limitations of current approaches, showcasing the deceptive promise of sketch-based generative AI. By introducing an abstraction-aware framework, featuring a sketch adapter, adaptive time-step sampling, and discriminative guidance, we empower amateur sketches to yield precise, high-fidelity images without the need for textual prompts during inference. We welcome the community to scrutinise our results. Please refer to the demo video for a detailed real-time comparison with state-of-the-arts.

References

[1] Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended Diffusion for Text-driven Editing of Natural Images. In *CVPR*, 2022. 5

[2] Dmitry Baranchuk, Ivan Rubachev, Andrey Voynov, Valentin Khrulkov, and Artem Babenko. Label-Efficient Semantic Segmentation with Diffusion Models. In *ICLR*, 2021. 2

[3] Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Aneeshan Sain, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. More Photos are All You Need: Semi-Supervised Learning for Fine-Grained Sketch Based Image Retrieval. In *CVPR*, 2021. 2

[4] 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

[5] Ayan Kumar Bhunia, Subhadeep Koley, Abdullah Faiz Ur Rahman Khilji, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketching Without Worrying: Noise-Tolerant Sketch-Based Image Retrieval. In *CVPR*, 2022. 4, 7

[6] Ayan Kumar Bhunia, Subhadeep Koley, Amandeep Kumar, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch2Saliency: Learning to Detect Salient Objects from Human Drawings. In *CVPR*, 2023. 2, 3

[7] John Canny. A Computational Approach to Edge Detection. *IEEE TPAMI*, 1986. 4, 5, 7

[8] Z. Cao, G. Hidalgo Martinez, T. Simon, S. Wei, and Y. A. Sheikh. OpenPose: Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *IEEE TPAMI*, 2019. 5

[9] Caroline Chan, Fredo Durand, and Phillip Isola. Informative Drawings: Learning to generate line drawings that convey geometry and semantics. In *CVPR*, 2022. 5

[10] Dar-Yen Chen, Subhadeep Koley, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, Ayan Kumar Bhunia, and Yi-Zhe Song. DemoCaricature: Democratising Caricature Generation with a Rough Sketch. In *CVPR*, 2024. 2

[11] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Viswanatha Reddy Gajjala, Aneeshan Sain, Tao Xiang, and Yi-Zhe Song. Partially Does It: Towards Scene-Level FG-SBIR With Partial Input. In *CVPR*, 2022. 2

[12] Pinaki Nath Chowdhury, Tuanfeng Wang, Duygu Ceylan, Yi-Zhe Song, and Yulia Gryaditskaya. Garment ideation: Iterative view-aware sketch-based garment modeling. In *3DV*, 2022. 2

[13] Pinaki Nath Chowdhury, Ayan Kumar Bhunia, Aneeshan Sain, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. SceneTrilogy: On Human Scene-Sketch and its Complementarity with Photo and Text. In *CVPR*, 2023. 1, 5

[14] 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, 3

[15] Bram de Wilde, Anindo Saha, Richard PG ten Broek, and Henkjan Huisman. Medical diffusion on a budget: textual inversion for medical image generation. *arXiv preprint arXiv:2303.13430*, 2023. 2

[16] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *NeurIPS*, 2021. 1, 2, 3, 5, 6

[17] Mathias Eitz, James Hays, and Marc Alexa. How do humans sketch objects? *ACM TOG*, 2012. 3, 7

[18] Songwei Ge, Taesung Park, Jun-Yan Zhu, and Jia-Bin Huang. Expressive Text-to-Image Generation with Rich Text. In *ICCV*, 2023. 5

[19] Arnab Ghosh, Richard Zhang, Puneet K Dokania, Oliver Wang, Alexei A Efros, Philip HS Torr, and Eli Shechtman. Interactive Sketch & Fill: Multiclass Sketch-to-Image Translation. In *CVPR*, 2019. 2

[20] David Ha and Douglas Eck. A Neural Representation of Sketch Drawings. In *ICLR*, 2017. 7

[21] Cusuh Ham, Gemma Canet Tarres, Tu Bui, James Hays, Zhe Lin, and John Collomosse. Cogs: Controllable generation and search from sketch and style. In *ECCV*, 2022. 2

[22] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-Prompt Image Editing with Cross Attention Control. In *ICLR*, 2022. 2

[23] Aaron Hertzmann. Why Do Line Drawings Work? A Realism Hypothesis. *Perception*, 2020. 3, 5

[24] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. *arXiv preprint arXiv:2207.12598*, 2022. 1, 2

[25] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *NeurIPS*, 2020. 1, 2, 3

[26] Ziqi Huang, Kelvin CK Chan, Yuming Jiang, and Ziwei Liu. Collaborative Diffusion for Multi-Modal Face Generation and Editing. In *CVPR*, 2023. 5

[27] Ziqi Huang, Tianxing Wu, Yuming Jiang, Kelvin CK Chan, and Ziwei Liu. ReVersion: Diffusion-Based Relation Inversion from Images. *arXiv preprint arXiv:2303.13495*, 2023. 5

[28] Zhengyu Huang, Haoran Xie, Tsukasa Fukusato, and Kazunori Miyata. AniFaceDrawing: Anime Portrait Exploration during Your Sketching. *arXiv preprint arXiv:2306.07476*, 2023. 2

[29] Quan Huynh-Thu, Marie-Neige Garcia, Filippo Speranza, Philip Corriveau, and Alexander Raake. Study of Rating Scales for Subjective Quality Assessment of High-Definition Video. *IEEE TBC*, 2010. 6

[30] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-Image Translation with Conditional Adversarial Networks. In *CVPR*, 2017. 6

[31] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *CVPR*, 2019. 6

[32] Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. Imagic: Text-Based Real Image Editing with Diffusion Models. In *CVPR*, 2023. 2

[33] Muhammad Uzair Khattak, Hanoona Rasheed, Muhammad Maaz, Salman Khan, and Fahad Shahbaz Khan. MaPLe: Multi-modal Prompt Learning. In *CVPR*, 2023. 6

[34] Kwanyoung Kim and Jong Chul Ye. Noise2Score: Tweedie's Approach to Self-Supervised Image Denoising without Clean Images. In *NeurIPS*, 2021. 5

[35] Kazuma Kobayashi, Lin Gu, Ryuichiro Hataya, Takaaki Mizuno, Mototaka Miyake, Hirokazu Watanabe, Masamichi Takahashi, Yasuyuki Takamizawa, Yukihiro Yoshida, Satoshi Nakamura, et al. Sketch-based Medical Image Retrieval. *arXiv preprint arXiv:2303.03633*, 2023. 2

[36] 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. 2, 4, 6

[37] 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. 2

[38] 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

[39] 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. 2

[40] Gihyun Kwon and Jong Chul Ye. Diffusion-based Image Translation using Disentangled Style and Content Representation. In *ICLR*, 2023. 5

[41] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Frechèt Inception Distance. In *ICLR*, 2023. 6

[42] Lambda Labs. Stable Diffusion Image Variations, 2022. 4

[43] Alexander C Li, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak. Your Diffusion Model is Secretly a Zero-Shot Classifier. *arXiv preprint arXiv:2303.16203*, 2023. 2

[44] Changjian Li, Hao Pan, Adrien Bousseau, and Niloy J Mitra. Free2CAD: Parsing freehand drawings into CAD commands. *ACM TOG*, 2022. 2

[45] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *ICML*, 2022. 5, 6

[46] Minchen Li, Alla Sheffer, Eitan Grinspun, and Nicholas Vining. Foldsketch: Enriching garments with physically reproducible folds. *ACM TOG*, 2018. 2

[47] Runtao Liu, Qian Yu, and Stella X Yu. Unsupervised Sketch-to-Photo Synthesis. In *ECCV*, 2020. 2

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

[49] Yongyi Lu, Shangzhe Wu, Yu-Wing Tai, and Chi-Keung Tang. Image Generation from Sketch Constraint Using Contextual GAN. In *ECCV*, 2018. 2

[50] Ling Luo, Yulia Gryaditskaya, Tao Xiang, and Yi-Zhe Song. Structure-Aware 3D VR Sketch to 3D Shape Retrieval. In *3DV*, 2022. 2

[51] Ling Luo, Pinaki Nath Chowdhury, Tao Xiang, Yi-Zhe Song, and Yulia Gryaditskaya. 3D VR Sketch Guided 3D Shape Prototyping and Exploration. In *ICCV*, 2023. 2

[52] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided Image Synthesis and Editing with Stochastic Differential Equations. In *ICLR*, 2021. 2, 3

[53] Qiang Meng, Shichao Zhao, Zhida Huang, and Feng Zhou. MagFace: A Universal Representation for Face Recognition and Quality Assessment. In *CVPR*, 2021. 5, 6

[54] Aryan Mikaeili, Or Perel, Daniel Cohen-Or, and Ali Mahdavi-Amiri. SKED: Sketch-guided Text-based 3D Editing. In *CVPR*, 2023. 2

[55] 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. 1, 2, 3, 4, 5, 6, 7

[56] 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, 4, 5, 6

[57] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-Shot Text-to-Image Generation. In *ICML*, 2021. 2

[58] 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

[59] René Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler, and Vladlen Koltun. Towards robust monocular depth estimation: Mixing datasets for zero-shot cross-dataset transfer. *IEEE TPAMI*, 2022. 5

[60] Elad Richardson, Yuval Alaluf, Or Patashnik, Yotam Nitzan, Yaniv Azar, Stav Shapiro, and Daniel Cohen-Or. Encoding in Style: a StyleGAN Encoder for Image-to-Image Translation. In *CVPR*, 2021. 2

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1, 3, 4, 5, 6, 7, 8

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

[63] 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

[64] 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. 1, 2, 3

[65] Aneeshan Sain, Ayan Kumar Bhunia, Yongxin Yang, Tao Xiang, and Yi-Zhe Song. StyleMeUp: Towards Style-Agnostic Sketch-Based Image Retrieval. In *CVPR*, 2021. 2, 3, 5

[66] Aneeshan Sain, Ayan Kumar Bhunia, Pinaki Nath Chowdhury, Subhadeep Koley, Tao Xiang, and Yi-Zhe Song. CLIP for All Things Zero-Shot Sketch-Based Image Retrieval, Fine-Grained or Not. In *CVPR*, 2023. 2, 5, 6

[67] Aneeshan Sain, Ayan Kumar Bhunia, Subhadeep Koley, Pinaki Nath Chowdhury, Soumitri Chattopadhyay, Tao Xiang, and Yi-Zhe Song. Exploiting Unlabelled Photos for Stronger Fine-Grained SBIR. In *CVPR*, 2023. 2, 5

[68] Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. Pic2Word: Mapping Pictures to Words for Zero-shot Composed Image Retrieval. In *CVPR*, 2023. 8

[69] Patsorn Sangkloy, Nathan Burnell, Cusuh Ham, and James Hays. The sketchy database: learning to retrieve badly drawn bunnies. *ACM TOG*, 2016. 5, 6, 7

[70] Patsorn Sangkloy, Wittawat Jitkrittum, Diyi Yang, and James Hays. A Sketch Is Worth a Thousand Words: Image Retrieval with Text and Sketch. In *ECCV*, 2022. 1, 5

[71] Idan Schwartz, Vésteinn Snæbjarnarson, Hila Chefer, Serge Belongie, Lior Wolf, and Sagie Benaim. Discriminative Class Tokens for Text-to-Image Diffusion Models. In *ICCV*, 2023. 2, 3

[72] Jiaming Shen, Kun Hu, Wei Bao, Chang Wen Chen, and Zhiyong Wang. Bridging the Gap: Fine-to-Coarse Sketch Interpolation Network for High-Quality Animation Sketch Inbetweening. *arXiv preprint arXiv:2308.13273*, 2023. 2

[73] Harrison Jesse Smith, Qingyuan Zheng, Yifei Li, Somya Jain, and Jessica K Hodgins. A Method for Animating Children's Drawings of the Human Figure. *ACM TOG*, 2023. 2

[74] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep Unsupervised Learning using Nonequilibrium Thermodynamics. In *ICML*, 2015. 2

[75] Jifei Song, Yi-Zhe Song, Tony Xiang, and Timothy M Hospedales. Fine-Grained Image Retrieval: the Text/Sketch Input Dilemma. In *BMVC*, 2017. 5

[76] Zhuo Su, Wenzhe Liu, Zitong Yu, Dewen Hu, Qing Liao, Qi Tian, Matti Pietikäinen, and Li Liu. Pixel Difference Networks for Efficient Edge Detection. In *ICCV*, 2021. 4

[77] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *CVPR*, 2016. 6

[78] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion. *arXiv preprint arXiv:2306.03881*, 2023. 2

[79] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *CVPR*, 2023. 2, 3

[80] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In *NeurIPS*, 2017. 4

[81] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-Guided Text-to-Image Diffusion Models. In *ACM SIGGRAPH*, 2023. 1, 2, 3, 4, 6

[82] 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. 2, 3, 6, 7

[83] Saining Xie and Zhuowen Tu. Holistically-Nested Edge Detection. In *ICCV*, 2015. 4

[84] Jiarui Xu, Sifei Liu, Arash Vahdat, Wonmin Byeon, Xiaolong Wang, and Shalini De Mello. Open-Vocabulary Panoptic Segmentation with Text-to-Image Diffusion Models. In *CVPR*, 2023. 2

[85] 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 *CVPR*, 2023. 4, 5

[86] Lan Yang, Kaiyue Pang, Honggang Zhang, and Yi-Zhe Song. SketchAA: Abstract Representation for Abstract Sketches. In *ICCV*, 2021. 4

[87] Lan Yang, Kaiyue Pang, Honggang Zhang, and Yi-Zhe Song. Finding Badly Drawn Bunnies. In *CVPR*, 2022. 2, 5

[88] Emilie Yu, Rahul Arora, J Andreas Baerentzen, Karan Singh, and Adrien Bousseau. Piecewise-smooth surface fitting onto unstructured 3D sketches. *ACM TOG*, 2022. 2

[89] Qian Yu, Feng Liu, Yi-Zhe Song, Tao Xiang, Timothy M Hospedales, and Chen-Change Loy. Sketch Me That Shoe. In *CVPR*, 2016. 3

[90] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *ICCV*, 2023. 1, 2, 3, 4, 6, 7

[91] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*, 2017. 6