This CVPR paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

SketchDeco: Training-Free Latent Composition for Precise Sketch Colourisation

Chaitat Utintu Yi-Zhe Song
SketchX, CVSSP, University of Surrey.
{chaitat.u, y.song}@surrey.ac.uk

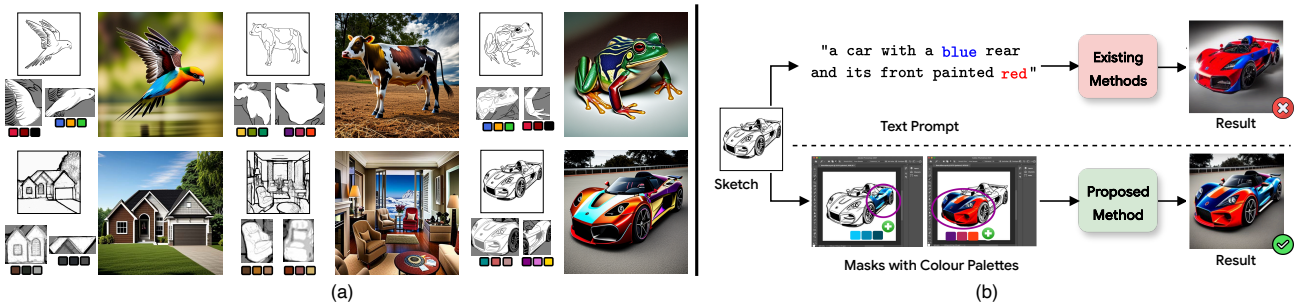


Figure 1. (a) Set of results generated by the proposed method. (b) Prior works on user-guided sketch colouring [79] that use text descriptions often lack precision. Hence, we focus on precision and convenience by requiring designers to specify the region mask (*e.g.*, use Photoshop) and its desired colour palettes. Our training-free method finds the optimal way to colour specified regions with chosen colours.

Abstract

We introduce `SketchDeco`*, a training-free approach to sketch colourisation that bridges the gap between professional design needs and intuitive, region-based control. Our method empowers artists to use simple masks and colour palettes for precise spatial and chromatic specification, avoiding both the tediousness of manual assignment and the ambiguity of text-based prompts. We reformulate this task as a novel, training-free composition problem. Our core technical contribution is a guided latent-space blending process: we first leverage diffusion inversion to precisely "paint" user-defined colours into specified regions, and then use a custom self-attention mechanism to harmoniously blend these local edits with a globally consistent base image. This ensures both local colour fidelity and global harmony without requiring any model fine-tuning. Our system produces high-quality results in 15–20 inference steps on consumer GPUs, making professional-quality, controllable colourisation accessible.*

1. Introduction

Transforming a line-art sketch into a fully coloured image is a fundamental task in creative workflows, from animation storyboards and product design to concept art. This process requires not just the artistic application of colour, but also precise spatial control to define regions and maintain coherence. While large-scale diffusion models [52, 54] have revolutionized image synthesis, achieving this fine-grained, region-specific *colour* control remains a major challenge. Text prompts are semantically powerful but spatially ambiguous. And while structural controls like sketches can define form [79], they do not solve the core problem of how to grant artists direct command over *where* and *how* colour is applied without compromising the model's rich generative priors.

The challenge is a trade-off between control and quality. Traditional methods, from manual assignment [56, 78] to reference-based transfer [35, 76], offer precision but are far too labour-intensive for modern workflows. Conversely, text-guided approaches [11, 77] are fast but imprecise, often resulting in semantic errors and "colour bleeding" that require manual correction.

This paper argues that the solution is not more training, but a novel, training-free *composition* framework. We introduce `SketchDeco`, a system that reformulates region-controlled colourisation as a latent-space blending problem. We empower artists with simple, practical inputs – a sketch, region masks, and colour palettes – and integrate them *without* fine-tuning.

Our core intuition is to separate global consistency from local control. Instead of trying to force a single model to do both at once, our system first generates a globally consistent, fully-coloured base image from the sketch. Then, it *composites* the user's specific colour choices (from palettes) into their specified regions (from masks) at the latent level. This approach allows us to blend the local edits harmoniously with the global base image, rather than fighting against the model's priors.

More specifically, to achieve this composition, we employ a *guided inversion and blending* process. We leverage diffusion inversion [48, 60] with DPM-Solver++ [43] to find an optimal noise vector and precisely "paint" the user's colours into the target latent regions. To ensure this edit is both precise and stable, we use guided sampling with classifier-free guidance (CFG) [25] and an exceptional prompt [44]. Finally, a custom self-attention mechanism, modulated by a scaling factor τ, harmoniously blends these local edits, preserving the original sketch structure while ensuring global coherence.

Our unified, training-free approach provides the following contributions:

- A novel region-guided framework that composes masks and palettes for precise, flexible colour control.
- A latent-space composition technique combining diffusion inversion and guided sampling to ensure both local colour fidelity and global harmony.
- A custom attention mechanism that adaptively balances sketch structure preservation with controlled colour diffusion.
- An efficient, training-free method that produces high-quality results in 15-20 steps on consumer GPUs.

2. Related Works

Sketch-to-Image Generation: Sketch-to-image generation [32] diverges from traditional image-to-image translation [14, 28, 34, 50, 84, 85] to generate realistic images from input sketches, bridging the gap between hand-drawn sketches and real images. SketchyGAN [12] pioneers this task by employing GANs [21] alongside edge-preserving image augmentations, while ContextualGAN [45] utilises conditional GANs to learn joint image-sketch representations. Given the difficulty in acquiring paired sketch-image datasets, some researchers have explored unsupervised [40] and self-supervised [39, 55] learning paradigms to tackle this challenge. Subsequent studies have enhanced results by fine-tuning pretrained StyleGAN models [69] or leveraging VQGAN to minimise discrepancies between ground truth image-sketch embeddings [22]. Despite these advancements, existing GANs-based methods for sketch-to-image generation often suffer from issues of unstable training and inferior quality, leading to a shift towards diffusion models [26, 38, 49, 66, 68, 79]. In this paper, we utilise pretrained text-to-image (T2I) diffusion models and address sketch-to-image generation problem as a downstream task.

Diffusion Models for Vision-based Tasks: Diffusion models [26, 59, 60], a recent class of generative models, outperform GANs [21] and VAEs [31] in both output quality and training stability by learning to reverse a Markov process through a denoising autoencoder. To improve efficiency, Rombach *et al*. [52] introduced Latent Diffusion Models (LDMs), later popularised as Stable Diffusion (SD), which perform the diffusion process in a latent space rather than in pixel space, significantly reducing computational costs while enabling conditional control via cross-attention [65]. Text-to-image (T2I) diffusion models employ textual inputs to iteratively guide denoising, achieving state-of-the-art results in applications such as image synthesis [13], style transfer [27], content manipulation [74], inpainting [46], and compositional generation [44]. Beyond text-based conditioning, additional modalities, such as sketches, enhance control and fidelity; however, naive methods necessitate full model retraining [13, 27, 68], leading to substantial computational overhead. To mitigate this, techniques like ControlNet [79], GLIGEN [38], and T2I-Adapter [49] fine-tune hypernetworks for domain adaptation. More recent approaches introduce conditioning mechanisms that circumvent the need for retraining or fine-tuning [44, 47, 58], refining the UNet's [53] denoising process by leveraging cross-attention maps [65]. In this work, we adopt a training-free methodology to address the realistic sketch colourisation problem, ensuring efficiency while preserving fidelity and control in the generated outputs.

Image and Sketch Colourisation: Deep learning (DL)-based image colourisation has advanced well beyond traditional methods [19, 20], with approaches commonly categorised as reference-based [10, 35, 41, 73, 75], text-based [4, 9, 11, 70, 77], or user-guided [4, 8, 27, 36, 80]. Reference-based methods transfer colour via pixel or semantic correspondences [10, 41], often using GANs [73] or pretrained networks [75], but remain prone to semantic misalignment. Text-based methods employ language cues through attention [11], palettes [4], or large-scale models [77], yet suffer from ambiguity and limited control. User-guided methods spread user-specified hints [4, 7], optimise colour propagation [36], or combine global [8, 27] and local inputs [6, 80], though spatial coherence and intuitive control remain challenging. Sketch colourisation [42, 71, 72, 76, 86, 87] extends this field, sharing similar principles but facing distinct difficulties due to sparse structure and lack of texture. As such, it remains a relatively new and challenging area with considerable potential for further exploration. In this study, we introduce a user-guided, text-free framework that enhances creative expression by leveraging a sketch, region-controlled masks, and corresponding colour palettes, thereby enabling precise, coherent, and spatially controllable sketch colourisation.

3. Proposed Method

Overview: We introduce a novel training-free paradigm for sketch colourisation, as illustrated in Fig. 2. Our method runs efficiently on a consumer-grade GPU and leverages the rich prior knowledge embedded in large-scale diffusion models [52]. Given an input sketch $\mathcal{S} \in \mathbb{R}^{512\times512\times3}$, a set of colour palettes $\{\mathcal{P}_{H1}, \mathcal{P}_{H2}, ..., \mathcal{P}_{Hn}\}$, where $\mathcal{P}_{\mathcal{H}}$

comprises hexadecimal colour codes, and corresponding region masks $\{\mathcal{M}^{(1)}, \mathcal{M}^{(2)}, \ldots, \mathcal{M}^{(n)}\}$, with each $\mathcal{M} \in \mathbb{R}^{512\times512\times1}$, we adopt a divide-and-conquer strategy consisting of two sequential stages: *global sketch colourisation* (Sec. 3.2) and *local sketch colourisation* (Sec. 3.3). The global stage generates multiple plausible colourised outputs $\mathcal{I}^G$ that preserve both the structure of $\mathcal{S}$ and the specified colour palettes $\mathcal{P}_\mathcal{H}$. The local stage refines these outputs using region masks $\mathcal{M}$, ensuring seamless composition and smooth colour transitions. The final result, $\mathcal{I}^L$, preserves spatial coherence while satisfying user-defined constraints.

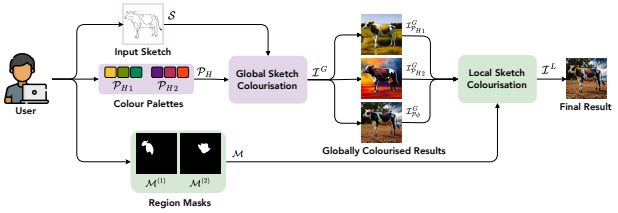


Figure 2. **Framework Overview.** Given an input sketch and region-specific colour palettes with corresponding masks, our method employs a divide-and-conquer strategy consisting of two sequential stages: **Global** and **Local Sketch Colourisations**.

3.1. Background

Pixel Space Diffusion Models: Diffusion Probabilistic Models (DPMs) [17, 26, 60] approximate a data distribution $p(x)$ by iteratively denoising a Gaussian-distributed variable. The denoising function is trained to reverse a fixed Markov Chain process of length T. Diffusion modeling consists of two stochastic processes: $forward$ and $backward$ diffusion [60]. In training, $forward$ diffusion gradually corrupts an image $x_0 \in \mathbb{R}^{H\times W\times3}$ by adding Gaussian noise, producing x_t as $x_t = \sqrt{\bar{\alpha}_t}x_0 + (\sqrt{1-\bar{\alpha}_t})\epsilon$, where, α_t controls noise levels from $\alpha_0 = 1$ to $\alpha_T \simeq 0$, and t is sampled from $\{1, ..., T\}$ [60]. The $backward$ process employs denoising autoencoders $\epsilon_\theta(\cdot)$ to predict clean images by minimizing:

$$\mathcal{L}_{\text{DM}} = \mathbb{E}_{x,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(x_t, t)\|_2^2\right]. \quad (1)$$

During inference, the trained network $\epsilon_\theta(\cdot)$ iteratively refines a Gaussian noise sample x_T over T steps to reconstruct x_0, recovering the data distribution [60].

Latent Diffusion Models: Latent Diffusion Models (LDMs, e.g., Stable Diffusion [52]) learn data distributions in a compact latent space rather than directly in pixel space. An autoencoder with encoder $\mathcal{E}(\cdot)$ and decoder $\mathcal{D}(\cdot)$ (UNet backbone [53]) performs diffusion efficiently [52]. Given an image $x_0 \in \mathbb{R}^{H\times W\times3}$, the encoder downsamples it by $f = H/h = W/w$ into $z_0 \in \mathbb{R}^{h\times w\times d}$ as $z_0 = \mathcal{E}(x_0)$, and the decoder reconstructs $\tilde{x_0} = \mathcal{D}(z_0) = \mathcal{D}(\mathcal{E}(x_0))$. Conditional generation is enabled by adding cross-attention [65] so that the denoising network $\epsilon_\theta(z_t, t, c)$ learns $p(z|c)$, where c is the condition. The training objective is:

$$\mathcal{L}_{\text{LDM}} = \mathbb{E}_{\mathcal{E}(x),c,\epsilon\sim\mathcal{N}(0,1),t}\left[\|\epsilon - \epsilon_\theta(z_t, t, c)\|_2^2\right] \quad (2)$$

LDMs support various conditions (*e.g.*, text [52], sketch [79]) using the same objective. For text conditioning, a prompt p is embedded via $\mathbf{T}(\cdot)$ and encoded by the pretrained CLIP text encoder $\mathcal{E}_t(\cdot)$ [51] to obtain $t_p = \mathcal{E}_t(\mathbf{T}(p))$. The text encoding is injected into UNet layers through cross-attention [65], defined as $\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)\cdot V$, with

$$Q = W_Q \cdot z_t,\ K = W_K \cdot t_p,\ V = W_V \cdot t_p \quad (3)$$

where W_Q, W_K, W_V are learnable projection matrices; W_K, W_V linearly project $t_p \in \mathbb{R}^{77\times768}$ to Keys and Values, and W_Q projects noisy latents to Query maps [52].

3.2. Global Sketch Colourisation

We present a novel user-guided global sketch colourisation framework (see Fig. 3 a) operating on input sketch $\mathcal{S} \in \mathbb{R}^{512\times512\times3}$ and colour palette $\mathcal{P}_\mathcal{H} = \{h \mid h \text{ is a colour hexcode}\}$. The model leverages generative priors from pretrained Stable Diffusion to achieve two goals: *(a)* generate faithful images $\mathcal{I}^G$ preserving sketch identity and structure (*i.e.* $id(\mathcal{I}^G) \approx id(\mathcal{S})$), and *(b)* produce vivid colours aligned with $\mathcal{P}_\mathcal{H}$ (*i.e.* $colour(\mathcal{I}^G) \approx colour(\mathcal{P}_\mathcal{H})$). The main challenges are: *(i)* the prompt-free setup hinders sketch-to-image generation without the sketch's class label in textual conditioning [79], and *(ii)* colour hexcodes must be mapped to standard colour names.

To address these, we use BLIP-2 [37] in a VQA setup to infer sketch semantics, a K-Dimensional Tree (K-D Tree) [5] to match hexcodes with colour names, and a pretrained Scribble ControlNet [79] for structure-preserving sketch-to-image generation. Building on [16], our pipeline directly processes the input sketch, integrates BLIP-2 class cues, and employs K-D Tree colour retrieval for enabling training-free, user-guided sketch colourisation.

Sketch Semantic Prediction: We utilise BLIP-2 [37], a vision–language pre-training framework that effectively aligns image and text modalities via its Q-former module [37]. The Q-former connects CLIP's image encoder with a large language model (LLM) through learnable query embeddings that attend to relevant visual regions. We adopt BLIP-2's Visual Question Answering (VQA) setup to infer class semantics from input sketches by prompting (*e.g.*, "*What is the object in this sketch?*"). This approach provides strong zero-shot generalisation across diverse sketch types, outperforming previous works [30, 76, 78] that relied solely on task-specific sketch encoders.

Colour Name Searching: We employ the CSS3 colour database [3] and a K-Dimensional Tree (K-D Tree) [5] to efficiently map colour hexcodes to their nearest named counterparts in RGB space. A K-D Tree with $K = 3$ is constructed from the 147 CSS3 reference colours (see Fig. 4), ensuring that any colour in $\mathbb{R}^3$ can be associated with a valid name. Every colour hexcodes $\mathcal{P}_\mathcal{H} = \{h_i\}_{i=1}^n$ are converted to RGB values, and the tree retrieves the closest colour names $\mathcal{P}_{name} = \{c_s^i\}_{i=1}^n$. The retrieved names are integrated into textual prompts that incorporate class semantics and descriptive cues. The final guidance t_p is formatted as:

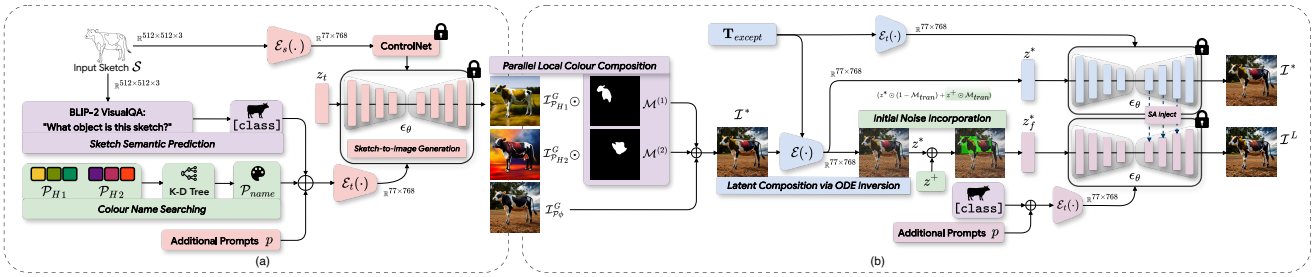


Figure 3. **Global and Local Sketch Colourisation Stages.** (a) In **global stage** 3.2, given a sketch $\mathcal{S}$ and colour palettes $\{\mathcal{P}_{\mathcal{H}}\}_{i=1}^{n}$, BLIP-2 [37] infers sketch class semantics, a K-D Tree [5] maps palette hexcodes to colour names, and Scribble ControlNet [79] generates globally colourised results $\{\mathcal{I}^G_{\mathcal{P}_{\mathcal{H}i}}\}_{i=1}^{n}$ and an auxiliary image $\mathcal{I}^G_{\mathcal{P}_\phi}$, preserving $id(\mathcal{S})$ and $colour(\mathcal{P}_{\mathcal{H}})$, while enabling interactive refinement for adjusting unsatisfying textures further. (b) In **local stage** 3.3, pre-colourised regions from $\{\mathcal{I}^G_{\mathcal{P}_{\mathcal{H}i}}\}_{i=1}^{n}$ are composed with background from $\mathcal{I}^G_{\mathcal{P}_\phi}$ to form composited image $\mathcal{I}^*$, which is then inverted to noisy latents z^* and refined via guided-sampling [43]. Noise incorporation enhances boundary smoothness, while SA injection, guided by τ, produces the final result $\mathcal{I}^L$ with smooth blending and structural fidelity.

"`[class]`, *hyper-realistic, quality, photography style, using only colours in colour palette of* `[`c_s^1`]`, - `[`c_s^2`]` - ... `[`c_s^n`]`" and negative prompts "*drawing look, sketch look, ...*". See supplement for more details about the K-D Tree algorithm (e.g., design choices, and advantages over LLMs).

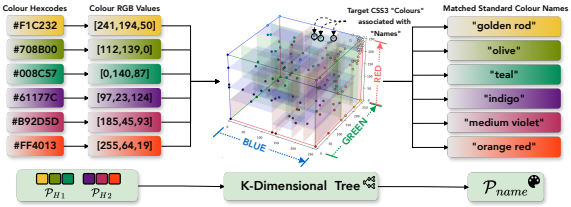


Figure 4. 3D search space constructed using K-D Tree algorithm.

Sketch-to-Image Generation: To generate a realistic image $\mathcal{I}^G$ based on input sketch $\mathcal{S}$, we employ the generative prior derived from a robust pretrained T2I diffusion model. Unlike textual conditioning t_p, sketches embody spatial information [57]. Thus, we harness the capabilities of the pretrained Scribble ControlNet [79] to attain fidelity and realism in the sketch-to-image generation. Given the sketch latents $z_s = \mathcal{E}_s(\mathcal{S})$; $z_s \in \mathbb{R}^{h\times w\times d}$, the noise estimator $\epsilon_\theta(z_t, t, t_p, z_s)$ needs to incorporate sketch condition $p(z|t_p, z_s)$. In this step, we generate $n+1$ images: the globally colourised results $\{\mathcal{I}^G_{\mathcal{P}_{\mathcal{H}i}}\}_{i=1}^{n}$, derived from the given colour palettes $\{\mathcal{P}_{\mathcal{H}i}\}_{i=1}^{n}$, and an auxiliary image $\mathcal{I}^G_{\mathcal{P}_\phi}$, which has no colour description (see Fig. 3). This extra image serves as the background for composition steps.

User-interactive Refinement: Users can apply local colour composition in the latent space z using user-defined masks $\{\mathcal{M}^{(i)}\}_{i=1}^{n}$. However, since interactions in latent space are not directly perceptible to the human eye, this approach offers limited controllability and interpretability for creative workflows, where human feedback is essential [16]. To address this, our pipeline renders global previews (*i.e.*, $\{\mathcal{I}^G_{\mathcal{P}_{\mathcal{H}i}}\}_{i=1}^{n}$ and $\mathcal{I}^G_{\mathcal{P}_\phi}$) in pixel space, allowing users to inspect, regenerate, and compare variants with different random seeds (see Supp. video). Rather than explicitly compositing latent features, our method initiates composition in pixel space, then applies ODE inversion to blend the results back into latent space during the local stage (Fig. 3 b), implicitly reproducing the same compositional effect. This hybrid approach retains the flexibility of latent editing while enabling direct, interpretable interaction ideal for creative domains, like interior design, where rapid exploration of colour-texture combinations is critical (see Fig.5).



Figure 5. User-interactive refinement process for interior design.

3.3. Local Sketch Colourisation

In local sketch colourisation (see Fig. 3 b), we are, to the best of our knowledge, the first to reframe this task as a combination of *image composition* [44] and *reconstruction* [43]. For text-guided approach [79], textual prompts often struggle to provide precise colour control (see Fig. 1 b), so we instead introduce spatial colour information directly into the noisy latent space rather than relying solely on the cross-attention mechanism [65]. In our approach, pre-colourised regions derived from $\mathcal{I}^G_{\mathcal{P}_H}$ are composed with $\mathcal{I}^G_{\mathcal{P}_\phi}$. Since all global results $\mathcal{I}^G$ originate from the same sketch, this composition step inherently preserves sketch faithfulness. The composited image is then inverted back into the latent space, *allowing the colour information to be directly embedded in the noisy latents*, and finally performs latent composition to refine boundaries and ensure coherent visual blending, all without any model retraining or fine-tuning.

Therefore, the final colourised image $\mathcal{I}^L$ aims to satisfy three objectives: (a) preserve the identity of the input sketch (*i.e.*, $id(\mathcal{I}^L) \approx id(\mathcal{S})$), (b) maintain local colour consistency within each mask region $\mathcal{M}$ such that the resulting colours align with the provided palette $\mathcal{P}_{\mathcal{H}}$ (*i.e.*, $colour(\mathcal{I}^L \odot \mathcal{M}) \approx colour(\mathcal{P}_{\mathcal{H}})$), and (c) ensure smooth and coherent transitions between locally coloured regions.

Parallel Local Colour Composition: Since $\mathcal{I}^G$ inherits both $id(\mathcal{S})$ and $colour(\mathcal{P}_\mathcal{H})$, we first apply each mask $\mathcal{M}^{(i)}$ to the corresponding $\mathcal{I}^G_{\mathcal{P}_{\mathcal{H}i}}$ to isolate its region of interest: $\mathcal{I}^G_{\mathcal{M}^{(i)}} = \mathcal{I}^G_{\mathcal{P}_{\mathcal{H}i}} \odot \mathcal{M}^{(i)}, i = 1, \ldots, n$, where $\odot$ denotes the Hadamard product. The composited image $\mathcal{I}^*$ is then obtained as: $\mathcal{I}^* = \mathcal{I}^G_{\mathcal{P}_\phi} \odot (1 - \mathcal{M}_{tot}) + \sum_{i=1}^n \mathcal{I}^G_{\mathcal{M}^{(i)}}$, with $\mathcal{M}_{tot} = \max_{i=1}^n \mathcal{M}^{(i)}$. This design enables *parallel processing of multiple masks*, where the number of regions only affects the number of intermediate globally colourised results $\mathcal{I}^G_{\mathcal{P}_H}$, without impacting the quality of $\mathcal{I}^*$ and $\mathcal{I}^L$.

Latent Composition via ODE Inversion: The composited image $\mathcal{I}^*$ satisfies our second objective – local colour consistency – but lacks smooth transition between local coloured regions (third objective). To achieve smooth transitions between local colourised regions, a crucial step involves transforming the composited image $\mathcal{I}^*$ into a latent representation z^*. The sampling process within diffusion probabilistic models (DPM) can be executed by solving ordinary differential equations (ODE) from T to 0 [61]:

$$\frac{dx_t}{dt} = f(t)x_t + \frac{g^2(t)}{2\sigma_t}\epsilon_\theta(x_t, t), \quad \text{where,} \quad x_T \sim \mathcal{N}(0, \tilde{\sigma}^2 I), \quad (4)$$

To perform the inversion process, we integrate the diffusion ODE from (0) to (T) using DPM-Solver++[43], chosen for its efficiency in generating high-quality samples within 15–20 steps, compared with the 100–250 steps typically required by DDIM[60]. We adopt DPM-Solver++ for its efficiency and strong trajectory alignment, which better preserves colour fidelity during inversion. Nevertheless, for text-to-image diffusion models, several works[23, 48, 64, 67] report substantial reconstruction errors when applying inversion in the text-conditioned setting $\epsilon_\theta(x_t, t, t_p)$. This issue primarily arises from the inherent instability of classifier-free guidance (CFG) mechanisms:

$$\hat{\epsilon}_\theta(x_t, t, t_p, \varnothing) = s \cdot \epsilon_\theta(x_t, t, t_p) + (1 - s) \cdot \epsilon_\theta(x_t, t, \varnothing) \quad (5)$$

where, $t_p = \mathcal{E}_t(\mathbf{T}(p))$ and $\varnothing = \mathcal{E}_t(\text{“ ”})$ denotes text encodings of normal and null prompt (empty string) embeddings, and s is the guidance scale. To address this, we replace null prompt with exceptional prompt embedding $\mathbf{T}_{except}$[44], which sets all token numbers to a common value and removes excessive positional embeddings and special tokens like `[startoftext]`, `[endoftext]`, and `[pad]`. This leads to more accurate inversion $\epsilon_\theta(x_t, t, \mathcal{E}_t(\mathbf{T}_{except}))$, as any information included in the prompt causes backward ODE trajectories to deviate from forward trajectories.

Initial Gaussian Noise Incorporation: After inversion process $\mathcal{I}^* \rightarrow z^*$, we incorporate the additional Gaussian noise z^+ to enhance transition smoothness by allowing the rich generative priors [52] to effectively in-paint the transition area between regions inside (*i.e.*, $\mathcal{M}_{tot}$) and outside (*i.e.*, $1 - \mathcal{M}_{tot}$) the user masks. Initially, we determine the spatial extent of each regional mask $\mathcal{M}$ by computing its coverage rectangular area $\mathcal{M}_{rect}$. Subsequently, we identify the transitional area $\mathcal{M}_{tran}$ by employing the summation of XOR operations $\oplus$ between each rectangular area and its corresponding user-defined mask as $\mathcal{M}_{tran} = \sum_{i=1}^n (\mathcal{M}^i_{rect} \oplus \mathcal{M}^i)$. Therefore, the initial noise incorporation for final diffusion ODE solving becomes: $z^*_f = z^* \odot (1 - \mathcal{M}_{tran}) + z^+ \odot \mathcal{M}_{tran}$ (see Fig.6).

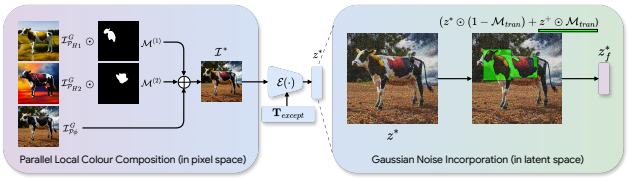


Figure 6. Initial Gaussian noise incorporation in latent space.

Self-attention Injection for Global Consistency: In the final stage of latent composition, the initial noise z^*_f, derived from the composited image $\mathcal{I}^*$, serves as the starting point for solving the diffusion ODE from T to 0 [61]. Throughout this stage, sketch class semantics along with additional prompt p (*e.g.*, “*hyper-realistic* `[class]` *in photography style.*”) is instrumental in producing smooth colour transitions across boundary regions. Nonetheless, relying solely on the initial noise may not sufficiently preserve the desired colour consistency (*i.e.*, $colour(\mathcal{I}^L \odot \mathcal{M}) \not\approx colour(\mathcal{P}_\mathcal{H})$) or maintain sketch identity (*i.e.*, $id(\mathcal{I}^L) \not\approx id(\mathcal{S})$). To address this, we inject the self-attention map $\mathcal{A}^*_{l,t}$ of $\mathcal{I}^*$ to guide the denoising steps. This self-attention map is derived using self-attention modules from the pretrained Stable Diffusion model [52], incorporating exceptional prompt embedding $\mathbf{T}_{except}$ for enhanced contextual understanding. Particularly, the self-attention module at layer l composes of three projection matrices as $W_{Q_l}, W_{K_l}, W_{V_l}$ and the encoded features of composited image at timestep t and layer l is $z^*_{l,t} \in \mathbb{R}^{h \times w \times d}$, we compute the self-attention map $\mathcal{A}^*_{l,t}$ by $\mathcal{A}^*_{l,t} = \text{softmax}\left(\frac{Q_{l,t}K^T_{l,t}}{\sqrt{d}}\right) \cdot V_{l,t}$ with:

$$Q_{l,t} = W_{Q_l} \cdot z^*_{l,t}, \; K_{l,t} = W_{K_l} \cdot z^*_{l,t}, \; V_{l,t} = W_{V_l} \cdot z^*_{l,t}. \quad (6)$$

However, achieving a balanced trade-off between textual and spatial guidance is essential. Therefore, we introduce a scaling factor τ to regulate this trade-off. During the initial phase $t \in [T, T \cdot (1 - \tau)]$ we inject the self-attention map $\mathcal{A}^*_{l,t}$ to maintain *global fidelity*; whereas, in the subsequent phase $t \in [T \cdot (1 - \tau), 0]$ the text encoding t_p from the text prompt p facilitates *smooth transitions* and *colour-blending* throughout local regions of the final colourised result $\mathcal{I}^L$.

4. Experiments

Dataset: We evaluate our method on four diverse datasets chosen to cover distinct domains: AFHQ-cat and AFHQ-dog [15] for animal face images, Place365-small [83] for complex indoor and outdoor scenes, Danbooru2023[1] for anime-style content, and PascalVOC2012 [18] for multi-object natural scenes. Specifically, we use ∼5k images each from AFHQ-cat/dog training sets, 900 test images from five randomly selected indoor and outdoor categories

[1] https://huggingface.co/datasets/nyanko7/danbooru2023

in Place365-small, the first 10k samples of Danbooru2023, and ∼12k images from PascalVOC2012. To obtain high-quality sketches, we combine the edge detection algorithms PiDiNet [62] and LineartDetector [79]. Experiments are divided into *global* and *local* sketch colourisation. For global stage, we use K-Means clustering with K=4 to extract dominant colours (choice of K detailed in Sec. 4.3). For local stage, two random masks are applied—sufficient as mask count has no impact on results (see Sec. 3.3)—and dominant colours are computed within masked regions using the same K=4. We further test our method in-the-wild on 15 randomly sourced sketches from `www.freepik.com` using the query "*black-and-white* `[class]` *sketch*".

Implementation Details: We use the official Stable Diffusion v1.5 implementation [52] with Huggingface versions of Scribble ControlNet [79] and BLIP-2's Q-former [37] for the global stage, following default settings and p described in Sec. 3.2. Class semantics are derived from BLIP-2's VQA features, and colour names are retrieved via a K-D Tree [5] search over the RGB space restricted to the CSS3 colour database [3]. The local stage employs official Stable Diffusion v1.5 [52], DPM-Solver++ [43], TF-ICON [44], and Prompt-to-Prompt [23] implementations with CFG = 2.5, $\tau = 0.4$, and p described in Sec. 3.3. All experiments are conducted on a single Nvidia RTX 4090 Super GPU.

Evaluation Metrics: We evaluate performance using five metrics. *(i) Fréchet Inception Distance (FID)* [24] measures the distributional difference between real and generated samples using InceptionV3 activations [63], where lower scores indicate better fidelity. *(ii) Learned Perceptual Image Patch Similarity (LPIPS)* [82] computes the weighted $l2$ distance between deep features from an ImageNet-pretrained AlexNet [33], with lower values implying closer perceptual alignment. *(iii) Peak Signal-to-Noise Ratio (PSNR)* compares the maximum signal power to noise distortion, where higher values reflect superior reconstruction quality. *(iv) SSIM* measures luminance, contrast, and structural similarity, with scores closer to 1 denoting higher consistency. *(v) Dynamic Closest Colour Warping (DCCW)* [29] measures palette similarity by estimating the perceptual distance between predicted and reference colour palettes, providing accurate reflection of colour consistency.

Competitors: We evaluate our framework against SOTA methods for global and local sketch colourisation: (i) For *global* colourisation, **DiffBlender** [30] enhances text-to-image diffusion with sketch and palette inputs, while **T2I-Adapter** [49] serves as a ControlNet-based baseline [79] using identical sketches and textual prompts. **IDeepColor** [81] is adapted by colourising grayscale T2I-Adapter outputs with 100 sampled hints from ground-truths. **Anime-Painter** [1] and **CounterfeitXL** [2], tuned for anime styles, are tested on Danbooru2023 to assess domain generalisation. (ii) For *local* colourisation, we compare with **ColorizeDiffusion** [76], **ColorFlow** [86], **MangaNinja** [42], and **Cobra** [87]. All models receive identical reference patches; baselines use them directly, while our method extracts dominant colours and processes them with our two-stage pipeline (Sec. 3) to ensure fairness and consistency.

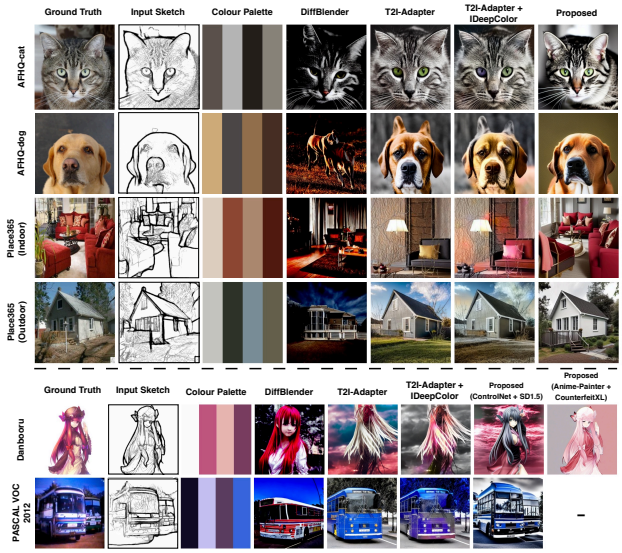


Figure 7. **Qualitative results on diverse datasets (*global*).** Our method outperforms SOTA techniques, showing better sketch fidelity, colour vividness, realism, and overall colourisation quality.

4.1. Qualitative Evaluation

Local Colourisation: Compared with reference-based methods, our approach yields more realistic results with vivid colours and stronger fidelity to input sketch and reference local colours (see Fig. 8). While MangaNinja [42] transfers local colours effectively, it lacks global consistency. On Danbooru2023 dataset, Cobra [87] provides convincing texture and shading but produces less vibrant backgrounds. Thus, the results demonstrate the robustness of our *training-free* framework, achieving competitive or superior local colourisation even against domain-specific methods.

Global Colourisation: Fig. 7 demonstrates the effectiveness of our global colourisation framework in preserving sketch structure and colour consistency across all diverse domains, from simple animal sketches (AFHQ [15]) to complex multi-object scenes (PascalVOC2012 [18]). Our method achieves higher sketch fidelity than competing approaches. While IDeepColor [81] attains accurate colour placement using colour hints from ground-truth images, it still exhibits colour artefacts such as colour bleeding and desaturation, suggesting that image colourisation techniques are suboptimal for improving sketch colourisation pipeline. In contrast, our training-free design *integrates seamlessly with domain-specific backbones* (*i.e.*, Anime-Painter [1] and CounterfeitXL [2]), showing strong adaptability across domains, as evidenced on the Danbooru2023 results.

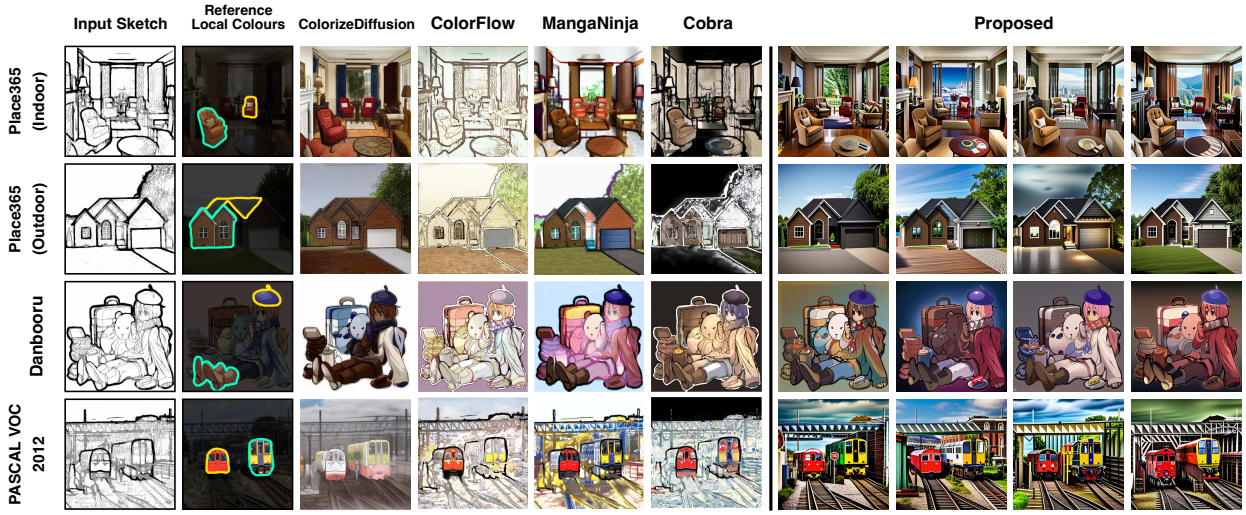


Figure 8. **Qualitative evaluation on diverse datasets. (*local*)** Our method outperforms SOTA techniques, achieving more accurate local colour propagation, and enhanced sketch fidelity. By effectively integrating local reference cues through the proposed two-stage pipeline (see Sec. 3), it produces coherent and realistic colourisation with improved consistency and visual quality across diverse styles and datasets.

Table 1. Benchmarks on diverse datasets (*local colourisation*).

Methods	Place365 (Indoor)					Place365 (Outdoor)					Danbooru2023					PascalVOC2012				
	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓
ColorizeDiffusion [76]	151.52	0.645	11.18	0.310	15.30	125.936	0.554	12.04	0.386	9.73	229.44	0.455	9.56	0.485	9.65	110.80	0.517	**11.42**	0.257	24.37
ColorFlow [86]	354.07	0.643	5.97	0.225	17.05	204.45	0.603	8.13	0.361	16.05	94.87	**0.313**	**10.02**	**0.541**	10.23	367.69	0.491	8.14	0.295	14.98
MangaNinja [42]	134.57	0.548	9.57	0.331	15.19	127.70	0.568	11.24	0.403	9.46	304.17	0.507	9.57	0.485	19.61	289.21	0.492	9.83	0.302	10.61
Cobra [87]	221.38	0.603	8.01	0.186	14.96	193.58	0.732	4.18	0.107	9.05	138.48	0.373	6.27	0.402	13.57	382.70	0.554	5.72	0.164	13.96
Proposed	**123.87**	**0.527**	**11.65**	**0.352**	**11.85**	**112.48**	**0.489**	**12.51**	**0.408**	**7.65**	**134.54**	0.450	7.31	0.416	**9.23**	**95.64**	**0.481**	8.97	**0.298**	**8.89**

Table 2. Benchmarks on diverse datasets (*global colourisation*).

Methods	AFHQ-cat					AFHQ-dog				
	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓
DiffBlender [30]	86.82	0.811	6.94	0.032	11.60	145.50	0.858	6.81	0.033	11.74
T2I-Adapter [49]	68.95	0.706	7.80	0.134	10.44	107.12	0.746	8.12	0.134	11.66
T2I-Adapter + IDeepColor [81]	68.41	0.673	7.89	0.133	**6.93**	116.95	0.699	8.27	0.135	**7.66**
Proposed	**50.31**	**0.671**	**8.65**	**0.187**	10.30	**89.70**	**0.687**	**8.78**	**0.214**	11.44

Methods	Place365 (Indoor)					Place365 (Outdoor)				
	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓
DiffBlender [30]	133.12	0.720	6.59	0.045	10.24	144.37	0.724	7.00	0.050	13.72
T2I-Adapter [49]	121.48	0.643	8.80	0.154	10.81	176.12	0.740	9.22	0.100	13.16
T2I-Adapter + IDeepColor [81]	121.57	0.614	8.86	0.153	**7.28**	176.40	0.690	9.42	0.119	**8.61**
Proposed	**113.99**	**0.553**	**9.03**	**0.199**	10.69	**129.11**	**0.605**	**9.46**	**0.166**	12.11

Methods	Danbooru2023					PascalVOC2012				
	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓	FID↓	LPIPS↓	PSNR ↑	SSIM ↑	DCCW ↓
DiffBlender [30]	232.01	0.779	4.06	0.068	18.09	82.15	0.776	6.43	0.071	14.16
T2I-Adapter [49]	232.70	0.735	6.57	0.148	17.11	154.30	0.697	8.61	0.155	11.77
T2I-Adapter + IDeepColor [81]	233.36	0.709	6.77	0.152	**13.15**	160.62	0.670	8.74	0.156	**8.17**
Proposed (ControlNet [79] + SD1.5 [52])	149.66	0.657	6.90	0.237	15.71	**75.27**	**0.558**	**8.97**	**0.163**	11.12
Proposed (Anime-Painter [1] + CounterfeitXL [2])	**138.74**	**0.483**	**8.82**	**0.438**	15.10	-	-	-	-	-

4.2. Quantitative Evaluation

Local Colourisation: As shown in Table 1, our method shows superior performance for indoor and outdoor scenes (see Fig. 8), demonstrating stronger generalisation to real-world settings. On Danbooru2023, other methods achieve better LPIPS, PSNR, and SSIM, reflecting their capabilities in the line-art domain, yet our lower FID and higher DCCW indicate greater realism and local colour fidelity. For PascalVOC2012, while methods like ColorizeDiffusion [76] score well on LPIPS, PSNR, and SSIM, likely due to its resemblance to line-art scenes, their outputs appear less coherent with visually unappealing backgrounds (see Fig. 8). Overall, our approach achieves the best FID and DCCW scores across all datasets, confirming its *strong global consistency* while facilitating *precise local colour transfer*.

Global Colourisation: As shown in Table 2, our method consistently outperforms competing approaches across scenarios, including animal faces and indoor/outdoor settings. Although IDeepColor [81] attains superior DCCW performance due to its direct use of ground-truth colour hints, it degrades perceptual quality with noticeable colour artefacts (see Fig. 7). On Danbooru2023, the benefits of our *training-free* design are evident, offering strong domain adaptability and seamless integration of stronger pretrained backbones, such as transitioning from SD1.5 to SDXL (*i.e.*, CounterfeitXL [2]). Using SDXL further boosts PSNR and SSIM scores, confirming improved pixel-level fidelity for $\mathcal{I}^{G}_{\mathcal{P}_{\mathcal{H}}}$.

4.3. Ablation Study

Impact of colour palette size: We evaluate the influence of palette size $\mathcal{P}_H$ on the globally colourised result $\mathcal{I}^{G}_{\mathcal{P}_H}$ using LPIPS [82] for perceptual quality and DCCW [29] for colour fidelity. As shown in Table 3, larger palettes improve colour faithfulness by capturing greater chromatic diversity, but excessive sizes introduce textual ambiguity in the T2I diffusion model, slightly reducing structural coherence, as indicated by higher LPIPS when n increases from 4 to 5.

Table 3. Ablation on different colour palette sizes.

Methods	AFHQ-cat		Place365 (Indoor)		Danbooru2023		PascalVOC2012	
	LPIPS↓	DCCW↓	LPIPS↓	DCCW↓	LPIPS↓	DCCW↓	LPIPS↓	DCCW↓
w/ 1 colour	0.693	14.23	0.562	15.72	0.694	32.16	0.673	15.06
w/ 2 colours	0.674	12.12	0.559	12.32	0.676	19.10	0.676	12.36
w/ 3 colours	0.686	10.61	0.557	11.21	0.671	16.66	0.665	11.20
w/ 4 colours	**0.671**	10.30	**0.553**	10.69	**0.657**	15.71	**0.558**	11.12
w/ 5 colours	0.682	**9.21**	0.560	**10.22**	0.668	**15.55**	0.662	**10.58**



Figure 9. **Qualitative evaluation on *in-the-wild* sketches.** We randomly selected sketches from `www.freepik.com` using the keywords "*black-and-white* `[class]` *sketch*." Our sketch colourisation technique allows users to specify colour palettes and corresponding region masks (using Photoshop), with generative priors embedded in diffusion model determining the best way to apply the chosen palettes.

Comparison of Attention Map Injections: Fig. 10 compares results from different attention map injection strategies. The baseline solves the diffusion ODE from T to 0 using DPM-Solver++ [43] without injections (Eq. 4). *(i)* Increasing injected maps does not improve quality, especially for local colourisation, where colour and sketch fidelity are critical. *(ii)* Injecting multiple maps adds unnecessary computational cost, as each requires solving another diffusion ODE. Hence, we inject only the self-attention map of the composited image, fixing the diffusion process count at two. This design achieves better detail preservation, while global consistency is maintained through textual guidance.

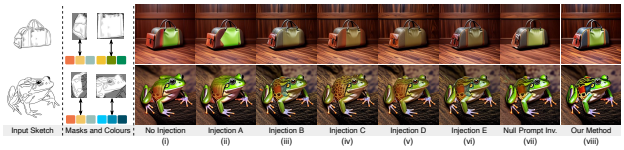


Figure 10. **Contributions of Attention Maps.** ***(i)* No injection** (without attention map injection), ***(ii-vi)* Injection A-E** (different combinations of attention map injection: **A** = $SA(\mathcal{I}^*)$+$CA(\mathcal{I}^*, \mathcal{I}^G_{\mathcal{M}})$, **B** = $SA(\mathcal{I}^*)$+$SA(\mathcal{I}^G_{\mathcal{M}})$, **C** = $SA(\mathcal{I}^*)$+$CA(\mathcal{I}^*, \mathcal{I}^G_{\mathcal{M}})$+$SA(\mathcal{I}^G_{\mathcal{M}})$, **D** = $SA(\mathcal{I}^G_{\mathcal{M}})$+$CA(\mathcal{I}^G_{\mathcal{M}}, \mathcal{I}^G_{\mathcal{P}_\phi})$+$SA(\mathcal{I}^G_{\mathcal{P}_\phi})$, **E** = $SA(\mathcal{I}^G_{\mathcal{M}})$+$CA(\mathcal{I}^G_{\mathcal{M}}, \mathcal{I}^G_{\mathcal{P}_\phi})$, (*SA*: self-attention and *CA*: cross-attention)), ***(vii)* Null inversion** (change from exceptional prompt to null prompt inversion [48]), ***(viii)* Our method** (default settings as in Sec. 3.3).

Effect of Exceptional Prompt $\mathbf{T}_{except}$: To assess $\mathbf{T}_{except}$, we replace it with a null prompt (*i.e.*, the unconditional term in classifier-free guidance (CFG), Eq. 5) during inversion. As shown in Fig. 10, this results in reduced colour richness and content fidelity. The degradation arises from special tokens in the null prompt (*e.g.*, `[startoftext]`, `[endoftext]`, ...) that cause embedding shifts and distort spatial coherence. $\mathbf{T}_{except}$, with these tokens removed, yields a more stable and accurate inversion, maintaining alignment between forward and backward ODE trajectories.

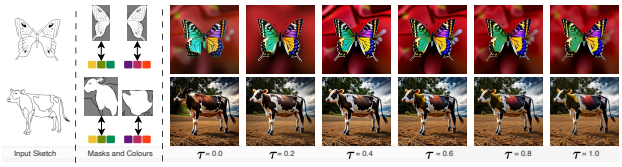


Figure 11. **Trade-off between harmonisation and faithfulness.** A low $\tau = 0.0$ gives better harmonisation of colours but lacks faithfulness (*e.g.*, missing regions in the wings of 'butterfly') whereas a high $\tau = 1.0$ gives highly faithful image generation but lacks harmonisation of colours (*e.g.*, the red patch on "cow").

Benefit of Scaling Factor τ: Fig. 11 shows the effect of varying scaling factor τ. Lower τ values yield smoother colour transitions guided mainly by the prompt (*e.g.*, `[class]`, *hyper-realistic*), but at $\tau = 0$ the output loses alignment with the sketch and colours. Increasing τ strengthens structural and colour fidelity via the self-attention injection (Sec. 3.3), yet excessive values reduce blending and make results resemble $\mathcal{I}^*$. Optimal coherence occurs for $\tau \in [0.3, 0.6]$; we set $\tau = 0.4$ in all experiments.

5. Conclusion

In conclusion, this paper introduces a novel approach to sketch colourisation that expertly balances precision and user convenience. Our method leverages easily created Photoshop region masks and user-defined colour palettes to offer intuitive control. We utilise staged generation, latent-space composition via ODE inversion and guided sampling, and a custom self-attention mechanism with an innovative scaling factor. Our approach addresses existing limitations, such as *limited user creativity* from restrictive textual prompts, inaccurate local colour assignment, and unpleasant results. Our design, which requires no training or fine-tuning, enhances zero-shot capabilities and integrates smoothly with more advanced foundational methods. Finally, our model is fast (takes less than 20 steps), training-free, and works on consumer-grade Nvidia RTX 4090 GPU.

Acknowledgements

We acknowledge the support of UK Research and Innovation (UKRI) through the UKRI AI: Centre for Doctoral Training in AI for Digital Media Inclusion under Grant No.: EP/Y030915/1. For the purpose of open access, the author has applied a Creative Commons Attribution (CC BY) license to any Author Accepted Manuscript version arising.

References

[1] Anime Painter. https://huggingface.co/xinsir/anime-painter. Accessed: 2024-10-06. 6, 7

[2] CounterfeitXL. https://huggingface.co/gsdf/CounterfeitXL. Accessed: 2024-10-06. 6, 7

[3] CSS Color Module Level 3. https://www.w3.org/TR/css-color-3/. Accessed: 2024-03-03. 3, 6

[4] Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David K. Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. Coloring with Words: Guiding Image Colorization Through Text-based Palette Generation. In *ECCV*, 2018. 2

[5] Jon Louis Bentley. Multidimensional binary search trees used for associative searching. *Communications of the ACM*, 1975. 3, 4, 6

[6] 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

[7] 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

[8] Huiwen Chang, Ohad Fried, Yiming Liu, Stephen DiVerdi, and Adam Finkelstein. Palette-based Photo Recoloring. In *SIGGRAPH*, 2015. 2

[9] Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. L-CoDer: Language-Based Colorization with Color-Object Decoupling Transformer. In *ECCV*, 2022. 2

[10] Guillaume Charpiat, Matthias Hofmann, and Bernhard Schölkopf. Automatic Image Colorization Via Multimodal Predictions. In *ECCV*, 2008. 2

[11] Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. Language-Based Image Editing with Recurrent Attentive Models. In *CVPR*, 2018. 1, 2

[12] Wengling Chen and James Hays. SketchyGAN: Towards Diverse and Realistic Sketch to Image Synthesis. In *CVPR*, 2018. 2

[13] Shin-I Cheng, Yu-Jie Chen, Wei-Chen Chiu, Hung-Yu Tseng, and Hsin-Ying Lee. Adaptively-Realistic Image Generation from Stroke and Sketch with Diffusion Model. In *WACV*, 2023. 2

[14] Yen-Chi Cheng, Hsin-Ying Lee, Min Sun, and Ming-Hsuan Yang. Controllable Image Synthesis via SegVAE. In *ECCV*, 2020. 2

[15] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *CVPR*, 2020. 5, 6

[16] Xiaoyan Cong, Yue Wu, Qifeng Chen, and Chenyang Lei. Automatic Controllable Colorization via Imagination. In *CVPR*, 2024. 3, 4

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

[18] Mark Everingham, Luc Gool, Christopher K. Williams, John Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vision*, 2010. 5, 6

[19] Beck Fabian, Dachsbacher Carsten, and Sadlo Filip. A Fast and Efficient Semi-guided Algorithm for Flat Coloring Line-arts. In *VMV*, 2018. 2

[20] Chie Furusawa, Kazuyuki Hiroshiba, Keisuke Ogaki, and Yuri Odagiri. Comicolorization: Semi-Automatic Manga Colorization. In *SIGGRAPH*, 2017. 2

[21] 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

[22] 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

[23] 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*, 2023. 5, 6

[24] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *NeurIPS*, 2017. 6

[25] Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021. 2

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

[27] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and Controllable Image Synthesis with Composable Conditions. In *ICML*, 2023. 2

[28] Junho Kim, Minjae Kim, Hyeonwoo Kang, and Kwang Hee Lee. U-GAT-IT: Unsupervised Generative Attentional Networks with Adaptive Layer-Instance Normalization for Image-to-Image Translation. In *ICLR*, 2020. 2

[29] Suzi Kim and Sunghee Choi. Dynamic Closest Color Warping to Sort and Compare Palettes. In *ACM TOG*, 2021. 6, 7

[30] Sungnyun Kim, Junsoo Lee, Kibeom Hong, Daesik Kim, and Namhyuk Ahn. DiffBlender: Scalable and Composable Multimodal Text-to-Image Diffusion Models. *arXiv preprint arXiv:2305.15194*, 2023. 3, 6, 7

[31] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[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. 2

[33] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, 2012. 6

[34] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Kumar Singh, and Ming-Hsuan Yang. Diverse Image-to-Image Translation via Disentangled Representations. In *ECCV*, 2018. 2

[35] Junsoo Lee, Eungyeup Kim, Yunsung Lee, Dongjun Kim, Jaehyuk Chang, and Jaegul Choo. Reference-Based Sketch Image Colorization using Augmented-Self Reference and Dense Semantic Correspondence. In *CVPR*, 2020. 1, 2

[36] Anat Levin, Dani Lischinski, and Yair Weiss. Colorization Using Optimization. In *SIGGRAPH*, 2004. 2

[37] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. *arXiv preprint arXiv:2301.12597*, 2023. 3, 4, 6

[38] Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-Set Grounded Text-to-Image Generation. In *CVPR*, 2023. 2

[39] Bingchen Liu, Yizhe Zhu, Kunpeng Song, and Ahmed Elgammal. Self-Supervised Sketch-to-Image Synthesis. In *AAAI*, 2021. 2

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

[41] Xiaopei Liu, Liang Wan, Yingge Qu, Tien-Tsin Wong, Stephen Lin, Chi-Sing Leung, and Pheng-Ann Heng. Intrinsic colorization. In *SIGGRAPH*, 2008. 2

[42] Zhiheng Liu, Ka Leong Cheng, Xi Chen, Jie Xiao, Hao Ouyang, Kai Zhu, Yu Liu, Yujun Shen, Qifeng Chen, and Ping Luo. MangaNinja: Line Art Colorization with Precise Reference Following. In *CVPR*, 2025. 2, 6, 7

[43] Cheng Lu, Yuhao Zhou, Fan Bao, Jianfei Chen, Chongxuan Li, and Jun Zhu. DPM-Solver++: Fast Solver for Guided Sampling of Diffusion Probabilistic Models. *arXiv preprint arXiv:2211.01095*, 2023. 2, 4, 5, 6, 8

[44] Shilin Lu, Yanzhu Liu, and Adams Wai-Kin Kong. TF-ICON: Diffusion-Based Training-Free Cross-Domain Image Composition. In *ICCV*, 2023. 2, 4, 5, 6

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

[46] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. RePaint: Inpainting using Denoising Diffusion Probabilistic Models. In *CVPR*, 2022. 2

[47] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. FreeControl: Training-Free Spatial Control of Any Text-to-Image Diffusion Model with Any Condition. *arXiv preprint arXiv:2312.07536*, 2023. 2

[48] Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text Inversion for Editing Real Images using Guided Diffusion Models. *arXiv preprint arXiv:2211.09794*, 2022. 2, 5, 8

[49] Chong Mou, Xintao Wang, Liangbin Xie, Yanze Wu, 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, 6, 7

[50] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic Image Synthesis with Spatially-Adaptive Normalization. In *CVPR*, 2019. 2

[51] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*, 2021. 3

[52] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *CVPR*, 2022. 1, 2, 3, 5, 6, 7

[53] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation. In *MICCAI*, 2015. 2, 3

[54] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L. Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic Text-to-Image Diffusion Models with Deep Language Understanding. In *NeurIPS*, 2022. 1

[55] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. Sketch3T: Test-time Training for Zero-Shot SBIR. In *CVPR*, 2022. 2

[56] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling Deep Image Synthesis with Sketch and Color. In *CVPR*, 2017. 1

[57] 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. 4

[58] Chenyang Si, Ziqi Huang, Yuming Jiang, and Ziwei Liu. FreeU: Free Lunch in Diffusion U-Net. *arXiv preprint arXiv:2309.11497*, 2023. 2

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

[60] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. In *ICLR*, 2021. 2, 3, 5

[61] Yang Song, Jascha Sohl-Dickstein, Diederik P. Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-Based Generative Modeling through Stochastic Differential Equations. In *ICLR*, 2021. 5

[62] 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. 6

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

[64] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. *arXiv preprint arXiv:2211.12572*, 2022. 5

[65] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention Is All You Need. In *NeurIPS*, 2017. 2, 3, 4

[66] Andrey Voynov, Kfir Aberman, and Daniel Cohen-Or. Sketch-Guided Text-to-Image Diffusion Models. In *SIGGRAPH*, 2023. 2

[67] Bram Wallace, Akash Gokul, and Nikhil Naik. EDICT: Exact Diffusion Inversion via Coupled Transformations. In *CVPR*, 2023. 5

[68] Qiang Wang, Di Kong, Fengyin Lin, and Yonggang Qi. DiffSketching: Sketch Control Image Synthesis with Diffusion Models. In *BMVC*, 2022. 2

[69] Sheng-Yu Wang, David Bau, and Jun-Yan Zhu. Sketch Your Own GAN. In *ICCV*, 2021. 2

[70] Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. L-CoDe:Language-Based Colorization Using Color-Object Decoupled Conditions. In *AAAI*, 2022. 2

[71] Shukai Wu, Xiao Yan, Weiming Liu, Shuchang Xu, and Sanyuan Zhang. Self-driven dual-path learning for reference-based line art colorization under limited data. In *IEEE TCSVT*. IEEE, 2023. 2

[72] Shukai Wu, Yuhang Yang, Shuchang Xu, Weiming Liu, Xiao Yan, and Sanyuan Zhang. FlexIcon: Flexible Icon Colorization via Guided Images and Palettes. In *ACM MM*, pages 8662–8673, 2023. 2

[73] Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. Towards Vivid and Diverse Image Colorization with Generative Color Prior. In *ICCV*, 2021. 2

[74] Sihan Xu, Ziqiao Ma, Yidong Huang, Honglak Lee, and Joyce Chai. CycleNet: Rethinking Cycle Consistency in Text-Guided Diffusion for Image Manipulation. In *NeurIPS*, 2023. 2

[75] Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. Stylization-Based Architecture for Fast Deep Exemplar Colorization. In *CVPR*, 2020. 2

[76] Dingkun Yan, Liang Yuan, Erwin Wu, Yuma Nishioka, Issei Fujishiro, and Suguru Saito. ColorizeDiffusion: Adjustable Sketch Colorization with Reference Image and Text. In *WACV*, 2025. 1, 2, 3, 6, 7

[77] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing Colors: Image Colorization with Text Guided Diffusion. In *SIGGRAPH*, 2023. 1, 2

[78] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. Two-stage Sketch Colourization. *SIGGRAPH*, 2018. 1, 3

[79] 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

[80] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. Real-Time User-Guided Image Colorization with Learned Deep Priors. In *SIGGRAPH*, 2017. 2

[81] Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S. Lin, Tianhe Yu, and Alexei A. Efros. Real-Time User-Guided Image Colorization with Learned Deep Priors. In *SIGGRAPH*, 2017. 6, 7

[82] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *CVPR*, 2018. 6, 7

[83] Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2017. 5

[84] Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks. In *ICCV*, 2017. 2

[85] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A. Efros, Oliver Wang, and Eli Shechtman. Toward Multimodal Image-to-Image Translation. In *NeurIPS*, 2017. 2

[86] Junhao Zhuang, Xuan Ju, Zhaoyang Zhang, Yong Liu, Shiyi Zhang, Chun Yuan, and Ying Shan. ColorFlow: Retrieval-Augmented Image Sequence Colorization. *arXiv preprint arXiv:2412.11815*, 2025. 2, 6, 7

[87] Junhao Zhuang, Lingen Li, Xuan Ju, Zhaoyang Zhang, Chun Yuan, and Ying Shan. Cobra: Efficient Line Art COlorization with BRoAder References. In *SIGGRAPH*, 2025. 2, 6, 7