SketchDeco: Training-Free Latent Composition for Precise Sketch Colourisation

Supplementary Material

A. Supplementary Material

This supplementary material complements the main paper, "SketchDeco: Training-Free Latent Composition for Precise Sketch Colourisation," by providing additional details: ablation study of textual prompts (Sec. B), the unusual effectiveness of K-D Tree (Sec. C), ablation on local sketch colourisation (Sec. D), additional discussion on design choices (Sec. E), clarification on contributions (Sec. F), limitation and future study (Sec. G), and full quantitative results on in-the-wild sketch dataset (Sec. H).

B. Ablation Study of Textual Prompts

We conduct an additional ablation study to investigate the contribution of each textual information on our global sketch colourisation process in the following cases: *(i)* Baseline, where the colourised results are generated without any textual guidance (i.e., only input sketches are utilised as conditions); *(ii)* only class semantics (i.e., "`[class]`"), derived from BLIP-2[6] prediction, are applied for textual guidance; *(iii)* class semantics and positive prompts (i.e., "`[class]`, *hyper-realistic, quality, photography style*") are used; *(iv)* negative prompts are combined; *(v)* finally, four main textual information, including class semantics, colour names, positive prompts (i.e., "`[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 (i.e., "*drawing look, sketch look, line art style, cartoon look, unnatural colour, unnatural texture, unrealistic look, low-quality*") are utilised. The experimental outcomes demonstrate that each component of textual guidance plays a pivotal role in generating faithful result. Notably, the presence of colour names and class semantics significantly enhances the visual outputs, as depicted in Fig. S1.

Fig. S1. **Ablation study of different textual prompt settings.** In our global sketch colourisation pipeline, the textual prompts t_p are composed of four main textual information, including class semantics, colour names, positive prompts (i.e., "`[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 (i.e., "*drawing look, sketch look, line art style, cartoon look, unnatural colour, unnatural texture, unrealistic look, low-quality*")

C. The Unusual Effectiveness of K-D Tree

[i] Compatability with professional designer workflow: In our envisioned user interface (UI), we prioritise convenience by allowing designers to specify region masks and their desired colour palettes using Photoshop. While these palettes are typically provided as colour codes (*e.g.*, `#FFD700` or `rgb(255,215,0)`), our diffusion models require text-based input to specify colours (*i.e.*, "`gold`"). To bridge this gap between user-friendly inputs and the requirements of Stable Diffusion [13], we propose a straightforward yet highly effective solution – a binary search tree (specifically, a K-D Tree with $K = 3$) that maps RGB values $\mathcal{P}_{rgb} = \{(r, g, b) \mid r, g, b \in [0, 255]\}$ to a W3C standard database of 147 CSS3 colour names.

[ii] Choice of standard CSS3 colour database: Remarkably, we found that using these standard CSS3 colour names as input to the pre-trained text encoder CLIP [9] effectively conditions Stable Diffusion to accurately reproduce the specified colour in the generated images, which are depicted in Fig. S2. This unexpected effectiveness may stem from the fact that both CLIP and Stable Diffusion were trained on large-scale internet data, where Cascading Style Sheets (CSS) are widely used as a language for describing the rendering of HTML and XML documents. By leveraging the inherent compatibility of these models with web standards, we eliminate the need for training a custom Hexcode encoder, keeping SketchDeco entirely training-free.

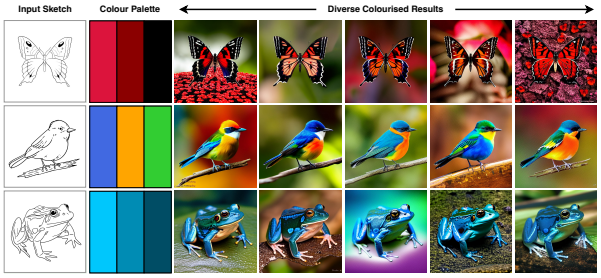

Fig. S2. **Diverse colourised outputs.** Our training-free global sketch colourisation technique leverages the rich generative priors from a large-scale text-to-image (T2I) diffusion model to produce diverse and realistic outputs. Colour names are derived from the K-D Tree algorithm with specific colours $[c_s^1]$, $[c_s^2]$, ..., $[c_s^n]$

[iii] Benefits over LLMs: When compared to LLMs, which are computationally intensive for tasks such as colour naming, a KD-tree provides a more efficient solution. LLMs are prone to generating non-existent colours, which can lead to inaccuracies in diffusion models. In contrast, a KD-tree enables quick and efficient lookups with minimal overhead. By aligning a KD-tree with standard CSS3 colours, we ensure accuracy, scalability, and easy future expansions.

D. Ablation on Local Sketch Colourisation

[i] Benefits of staged generation: Our method employs a divide-and-conquer strategy to facilitate a complex task (see Fig.2). The global stage generates initial outputs adhering to predefined sketch and colour palettes. These outputs are subsequently combined in the local stage, guided by region masks, which carefully refine the results to ensure seamless colour transitions and maintain the integrity of the sketch and colours. Moreover, this design also offers user flexibility, allowing for verification and potential regeneration of global colourised results, as depicted in Fig.S2.

[ii] Choice of local colour composition The composition of global colourised results serves as an essential bridge between the global and local stages. This process ensures the preservation of sketch and colour integrity for $\mathcal{I}^*$. As the global outcomes are consistently derived from the same sketch, their integration is seamless. We further examine the efficacy of our approach by contrasting it with the user-guided image colourisation technique, IDeepColor [14], which utilises convolutional neural network (CNN) to propagate colour hints. Our experimental setup involved providing a grayscale image along with 100 random hints to IDeepColor for recolourisation. Fig. S3 illustrates that IDeepColor not only fails short to retain the original colours but also introduces visual artefacts and colour bleeding.

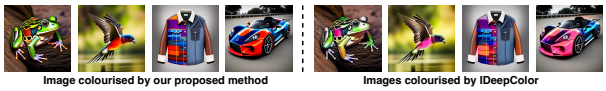

Fig. S3. **Impact of local colour composition.** Our method can maintain colour accuracy in the local stage due to the implementation of the local colour composition step. In contrast, IDeepColour [14] falls short and also introduces colour artefacts, e.g., grayish.

[iii] Does inversion of composited image $\mathcal{I}^*$ and local refinement steps help?: In this ablation study, we examine the effectiveness of our local refinement process initiated by the ODE inversion of $\mathcal{I}^*$. We compare this to SDEdit [8], an inversion-free image editing method, using only Gaussian noise addition to facilitate editing process. We fixed the guidance scale at 6.5, varying only the intensity of the noise to assess the impact on the local nuances. As shown in Fig. S4, SDEdit generally smooths images but struggles with creative colour blending and detail preservation, leading to distorted textures (e.g., in a cow at $s = 0.5$) and unrealistic features (e.g., in a bird at $s = 0.5$). In contrast, our method surpasses SDEdit by integrating ODE inversion of $\mathcal{I}^*$ together with $\mathbf{T}_{except}$ and injection of self-attention to preserve the integrity of sketch and colour. Then, we enhance smooth colour transitions through strategic noise incorporation and textual guidance. Additionally, we also introduce the parameter τ to effectively balance these elements during the final sampling process, facilitating an optimal trade-off between fidelity and aesthetic refinement.

Fig. S4. **Ablation study on local refinement steps.** Evaluation of our inversion-based local refinement technique compared to the inversion-free SDEdit[8] method. SDEdit tends to oversmooth images, compromising creative colour blending and texture fidelity, resulting in distorted textures (e.g., cow at $s = 0.5$) and unrealistic features (e.g., bird at $s = 0.5$). In contrast, our method retains the original colours and sketches from the composite image while facilitating harmonious colour transitions across different regions.

[iv] Can naive ControlNet match our fine-grained colour control?: We additionally investigate whether a standard pretrained Scribble ControlNet [13], originally developed for text-guided sketch-to-image generation, can approximate the performance of our training-free, region-controlled colourisation framework. To construct a fair comparison, captions were generated from our method's outputs $\mathcal{I}^{\mathcal{L}}$ using multimodal LLMs, and these captions were subsequently used as text prompts for ControlNet. Even after multiple attempts and manual selection of the top three result candidates, the reproduced images remained unable to match the spatial fidelity and chromatic precision achieved by our method. This outcome indicates that text-guided conditioning is inherently limited for the task of precise local sketch colourisation, thereby reinforcing the strength of our approach in delivering fine-grained, spatially aligned colour control without any additional training (see Fig. S5).

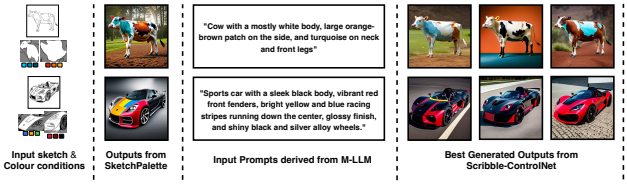

Fig. S5. **Comparison with a naive ControlNet baseline.** We compare our method, which leverages ControlNet in global stage, against a traditional text-guided ControlNet model. The naïve ControlNet fails to reproduce our results, demonstrating its inability to enforce fine-grained, spatially localised colour constraints.

E. Additional Discussion on Design Choices

In our study, we assess the impact of several design choices within the contexts of AFHQ-cat[4] and AFHQ-dog[4]. Firstly, within Visual Question Answering (VQA) tasks, substituting BLIP2 with GIT [11] results in FID/LPIPS scores of $(65.71/0.693)$ and $(129.19/0.742)$, respectively. Comparatively, the adoption of llama-3-vision-alpha [7] yields $(57.23/0.673)$ and $(61.22/0.683)$, while llava-1.5-7b [7] delivers $(55.97/0.669)$ and $(62.96/0.686)$. Secondly, in the domain of local sketch colourisation through inversion schemes, replacing DPM-Solver++ with DDIM

results in FID/LPIPS scores of $(35.09/0.716)$, with DPM-Solver achieving $(33.74/0.714)$. These findings highlight the nuanced influence of specific algorithmic changes on model performance across different tasks.

F. Clarification on Contributions

Image/sketch colourisation is a long-standing [3, 5, 12] open-problem in Computer Vision. While prior works used GANs [12], the latest works [5] use fine-tuned diffusion models. We take the next step by introducing *(i)* training-free paradigm for sketch colourisation that *outperforms fine-tuned SOTA* (see Fig. S6), *(ii)* existing diffusion-based colourisation only allow global colour consistency, whereas SketchDeco enable precise user-directed local colourisation, *(iii)* intuitively blend several modules for a fast pipeline that is compatible with consumer grade GPUs. While the use of modules like BLIP-2, ControlNet, SD-1.5 have become ubiquitous in computer vision (due to their reliable performance) our method intuitively integrates them to introduce several "firsts" that advances image/sketch colourisation literature. Particularly, *(i)* We use coarse-to-fine multi-staged generation with diffusion models to combine global and local colourisation, *(ii)* Prior colourisation methods using SD-1.5 lacks local colour consistency. We adapt SD-1.5 to ensure both global and local colour consistency defined by region masks and colour palette. *(iii)* A latent-space composition technique combining diffusion inversion and guided sampling to ensure both local colour fidelity and global harmony. *(iv)* A custom attention mechanism that adaptively balances sketch structure preservation with controlled colour diffusion.

G. Limitation and Future Study

In our work, we acknowledge several limitations and suggest avenues for future research. Firstly, when processing small selected regions for local colourisation, the initial blending ability derived from global colourisation pipeline may lead to some losses in colour fidelity, necessitating exploration of novel techniques to address this challenge. This could involve the development of adaptive blending strategies or localised colour correction mechanisms to better preserve colour accuracy in intricate or fine-textured areas. Additionally, while our approach relies on matching colour hexcodes with name entries in a standard CSS3 colour database[1], there is potential to enhance the database to accommodate a wider variety of colours. However, caution must be exercised to ensure compatibility with the pre-trained Stable Diffusion[10] model, as ambiguous or non-standard colour names may introduce uncertainty and inconsistency in the colourisation process.

H. Qualitative Results on Random Sketches

We perform qualitative comparison against two state-of-the-art (SOTA) diffusion-based approaches: DiffBlender [5] and DiSS[2]. DiffBlender aims to enhance the functionality of Text-to-Image diffusion models by integrating inputs from multiple modalities. These inputs are processed based on their conditional types, which include image/sketches, spatial tokens (*i.e.*, bounding boxes), and non-spatial tokens (*i.e.*, colour palettes). For fair comparison, we use a combination of sketch and colour palette modalities, equivalent to our global sketch colourisation. As for DiSS, the primary objective is to generate realistic images based on sketch and stroke conditions. DiSS also proposes a region-sensitive stroke-to-image method using partially coloured strokes as input, to synthesise diverse content in the non-coloured regions. For a fair comparison with our approach, we use a single colour instead of a colour palette. To the best of our knowledge, there is no prior work addressing the use of masks and colour palettes for *region-aware colourisation*. Importantly, unlike our method, existing SOTA methods [2, 5] are *not training-free* and require expensive fine-tuning. As a result, from Fig. S6, our training-free colourisation approach outperforms fine-tuned state-of-the-art (SOTA) models in various aspects including colour vividness, colour harmonisation, fidelity to the original sketch, and overall realistic look. Additionally, Figs. S7 to S9 represents qualitative results on 15 random sketches obtained from `www.freepik.com` using search keywords: "*black-and-white* `[class]` *sketch*". Notably, the seamless colour transition effects before and after undergoing our local refinement processes in the local stage are highlighted using **`blue bounding boxes`**.

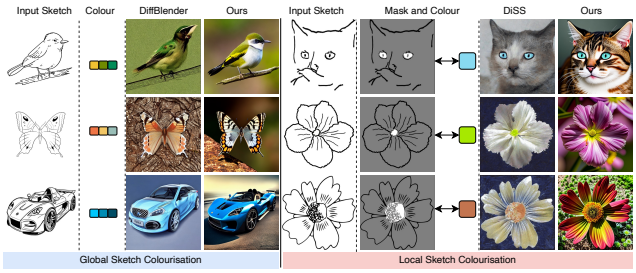

Fig. S6. **Qualitative assessment against fine-tuned SOTA methods using randomly sourced internet sketches.** We demonstrate that our training-free approach surpasses contemporary state-of-the-art methods (i.e., DiffBlender [5] and DiSS [2]) that rely on extensive fine-tuning, particularly in areas such as sketch fidelity, the application of creative colour schemes, and enhanced realism.

References

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

[2] 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. 3

[3] Zezhou Cheng, Qingxiong Yang, and Bin Sheng. Deep Colorization. In *ICCV*, 2015. 3

[4] Yunjey Choi, Youngjung Uh, Jaejun Yoo, and Jung-Woo Ha. StarGAN v2: Diverse Image Synthesis for Multiple Domains. In *CVPR*, 2020. 2

[5] 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] 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. 1

[7] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual Instruction Tuning. *arXiv preprint arXiv:2304.08485*, 2023. 2

[8] 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*, 2022. 2

[9] 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. 1

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

[11] Jianfeng Wang, Zhengyuan Yang, Xiaowei Hu, Linjie Li, Kevin Lin, Zhe Gan, Zicheng Liu, Ce Liu, and Lijuan Wang. GIT: A Generative Image-to-text Transformer for Vision and Language. *arXiv preprint arXiv:2205.14100*, 2022. 2

[12] 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. 3

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

[14] 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

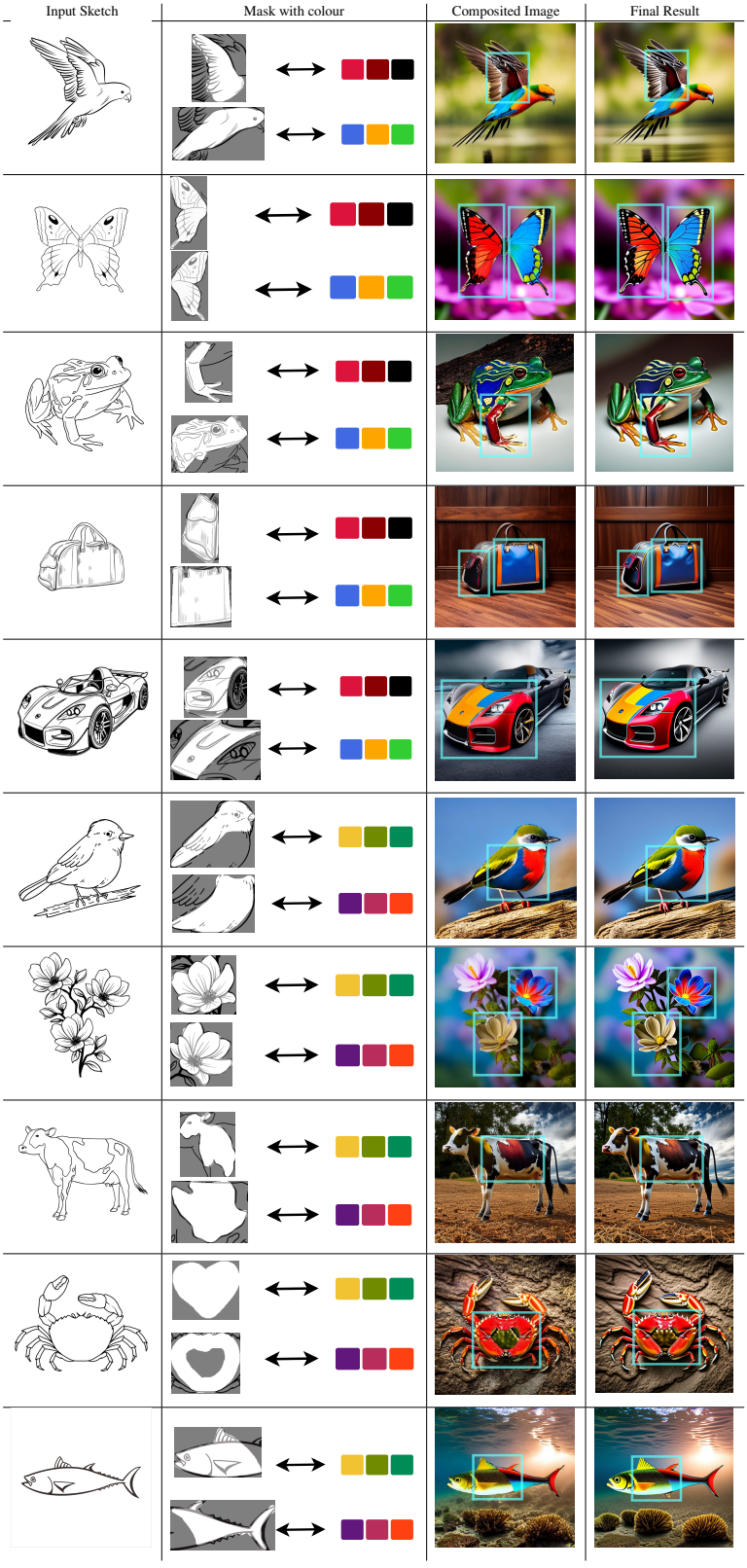

Fig. S7. **Qualitative evaluation with *in-the-wild* sketches.** The local sketch colourisation pipeline showcased notable adaptability, allowing users to incorporate three conditional inputs: a variety of sketches, hand-drawn masks, and preferred colour palettes. Consequently, our approach yields precise outcomes tailored to user-defined specifications. It is non-trivial to mention that these sketch images are randomly sourced from `www.freepik.com` with search keywords: "`black-and-white [class] sketch`".

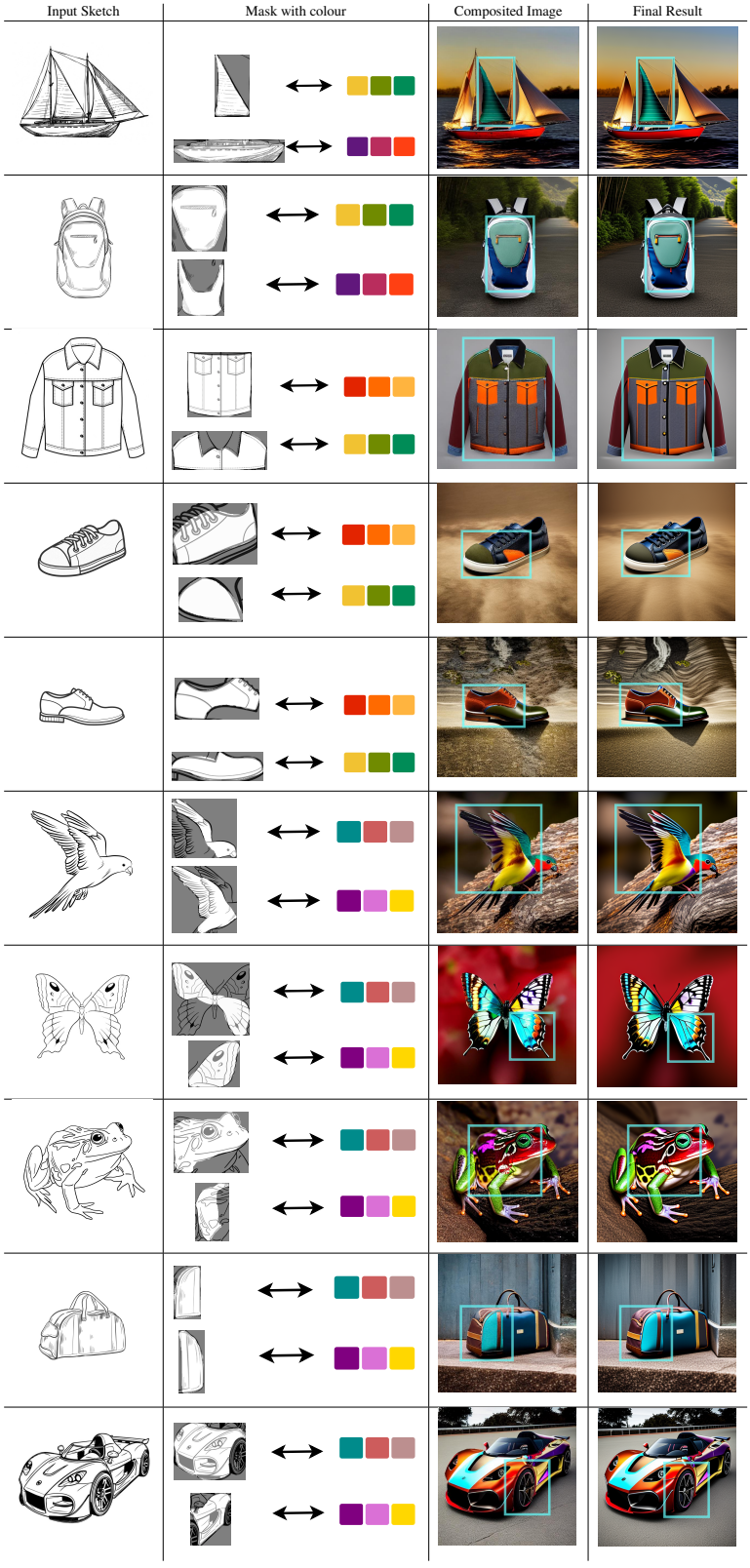

Fig. S8. **Qualitative evaluation with *in-the-wild* sketches.** The local sketch colourisation pipeline showcased notable adaptability, allowing users to incorporate three conditional inputs: a variety of sketches, hand-drawn masks, and preferred colour palettes. Consequently, our approach yields precise outcomes tailored to user-defined specifications. It is non-trivial to mention that these sketch images are randomly sourced from `www.freepik.com` with search keywords: "`black-and-white [class] sketch`".

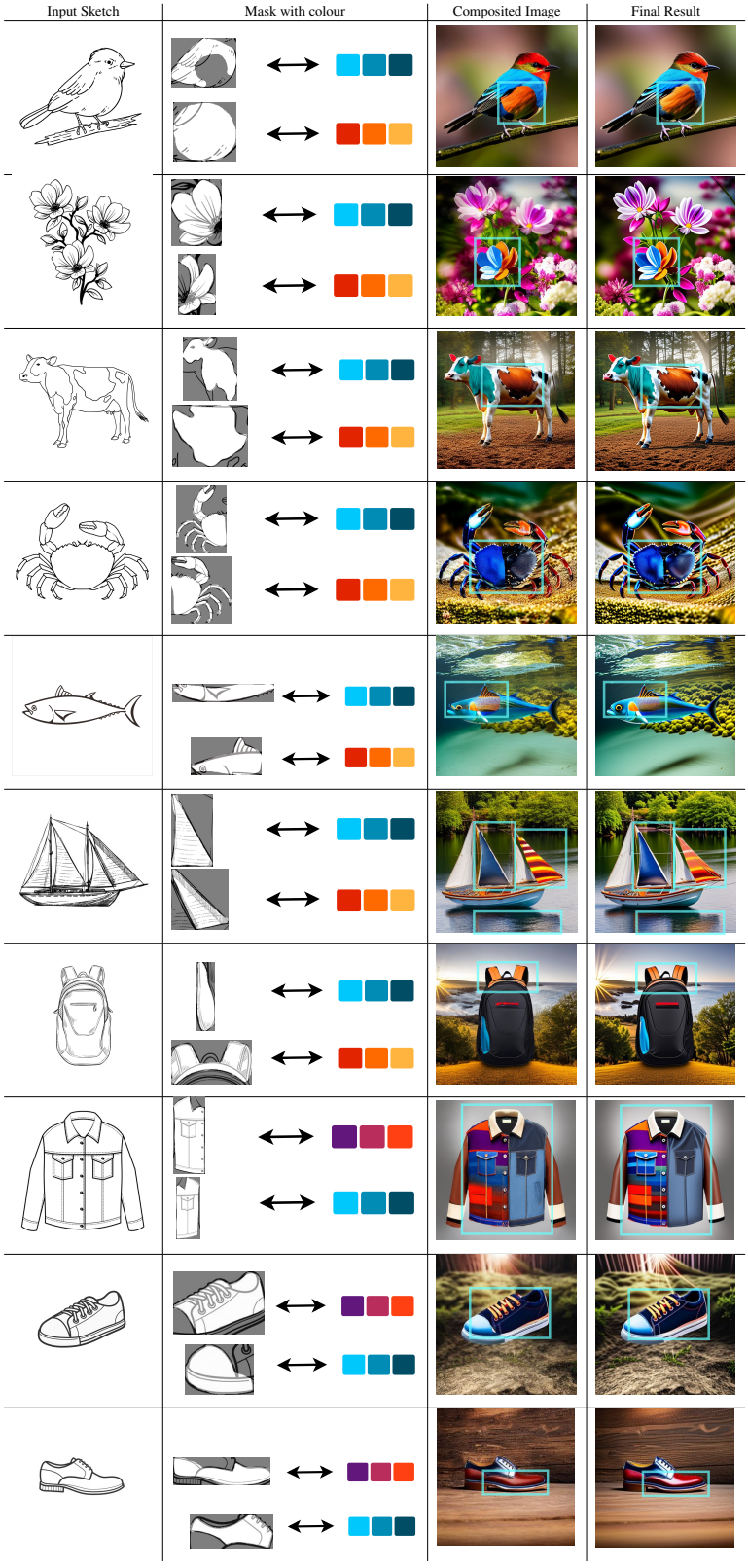

Fig. S9. **Qualitative evaluation with *in-the-wild* sketches.** The local sketch colourisation pipeline showcased notable adaptability, allowing users to incorporate three conditional inputs: a variety of sketches, hand-drawn masks, and preferred colour palettes. Consequently, our approach yields precise outcomes tailored to user-defined specifications. It is non-trivial to mention that these sketch images are randomly sourced from `www.freepik.com` with search keywords: "`black-and-white [class] sketch`".