MagicColor: Multi-Instance Sketch Colorization

Yinhan Zhang[1†] Yue Ma[2†] Bingyuan Wang[1] Qifeng Chen[2] Zeyu Wang[1,2*]
[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology

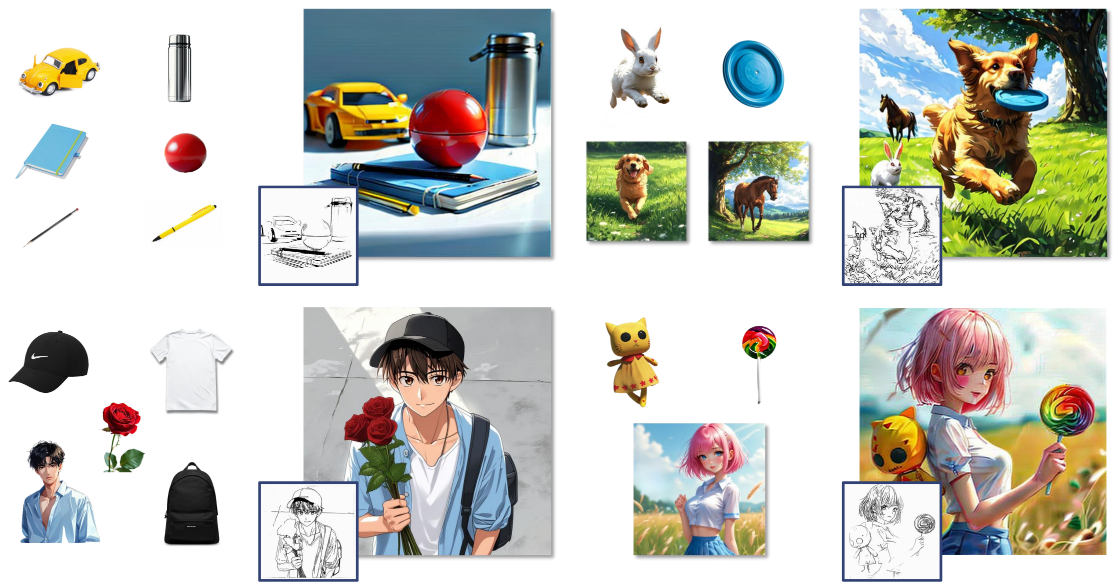

Figure 1. **Results of MagicColor.** Given a set of colored references, MagicColor can colorize a line art image while maintaining color consistency across multiple instances. Compared to traditional methods, our approach significantly improves coloring efficiency.

Abstract

We present MagicColor, a diffusion-based framework for multi-instance sketch colorization. The production of multi-instance 2D line art colorization adheres to an industry-standard workflow, which consists of three crucial stages: the design of line art characters, the coloring of individual objects, and the refinement process. The artists are required to repeat the process of coloring each instance one by one, which is inaccurate and inefficient. Meanwhile, current generative methods fail to solve this task due to the challenge of multi-instance pair data collection. To tackle these challenges, we incorporate three technical designs to ensure precise character detail transcription and achieve multi-instance sketch colorization in a single forward pass. Specifically, we first propose the self-play training strategy to address the lack of training data. Then we introduce an instance guider to feed the color of the instance. To achieve accurate color matching, we present fine-grained color matching with edge loss to enhance visual quality. Equipped with the proposed modules, MagicColor enables automatically transforming sketches into vividly-colored images with accurate consistency and multi-instance control. Experiments on our collected datasets show that our model outperforms existing methods regarding chromatic precision. Specifically, our model critically automates the colorization process with zero manual adjustments, so novice users can produce stylistically consistent artwork by providing reference instances and the original line art. Our code and additional details are available at https://yinhan-zhang.github.io/color.

† Equal contribution.
∗ Corresponding author.

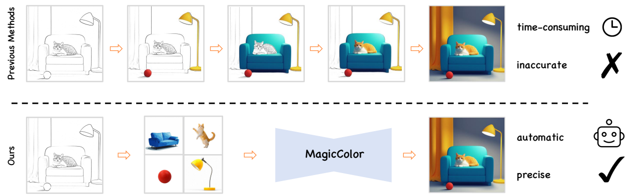

Figure 2. **Illustration of the workflow of multi-instance sketch colorization production.** Previous methods can only achieve multi-instance sketch colorization step by step, which is time-consuming and inaccurate. In contrast, our method can color a sketch while maintaining consistency, making multi-instance sketch colorization easier.

1. Introduction

The fast development of the digital cartoon industry has demonstrated widespread potential applications for generative artificial intelligence. Colorizing cartoon sketches, a crucial aspect within the broader realm of the cartoon industry, serves as an indispensable task. It not only heightens the visual allure but also enriches the narrative experience by effectively and vividly communicating emotions and actions. Using the automatic pipeline can streamline production workflows, accelerate content creation, reduce labor demands, and meet the growing demand for cartoons.

In the traditional process of cartoon sketch colorization, artists first analyze the line art to grasp the character and story context. They then choose suitable color palettes and manually apply colors layer by layer with attention to detail. However, traditional methods have two drawbacks. The first one is that they are time-consuming, as manual coloring requires great effort to match the original design. For instance, in an animation series, multiple identical objects in the same image need to be colored one by one, which is very inefficient. The other problem is inaccuracy since manual work is error-prone and different artists may have varied color interpretations, causing inconsistent results in large-scale projects. Our method slots into this workflow, supporting auto-colorization while preserving design fidelity and instance-level color consistency.

Cartoon sketch colorization has garnered considerable research interest due to its potential in digital art and animation production. Currently, user-guided methods [3, 5, 7, 23, 27, 58, 64, 95, 99] rely on explicit human inputs such as color points, scribbles, or textual descriptions to direct the coloring process. Reference-based approaches [1, 11, 20, 94, 101] employ fused attention mechanisms to transfer colors from exemplar images while preserving structural coherence. While recent advances [52, 57] have improved automated colorization, three critical limitations persist: (1) **Domain adaptation gap:** existing colorization pipelines rely heavily on reference image fidelity, as pronounced structural mismatches between line art and exemplars frequently induce erroneous chromatic mappings. This dependency imposes an impractical requirement for near-isomorphic geometry between inputs, which rarely holds in unconstrained animation workflows where stylized sketches often diverge significantly from real-world references. (2) **Instance-level control granularity:** existing methods demonstrate a lack of fine-grained control over instance-specific attributes during the color transfer process. Discrepancies in character pose, proportion, or viewpoint between reference instances and target sketches often lead to distorted textures and the loss of important details. As a result, crucial features from the reference image may be significantly diminished during the colorization process. (3) **Color consistency:** achieving instance-aware color consistency is essential. The coloring of regions with non-closed lines or multiple objects often results in color bleeding, which compromises the accuracy and stability of instance control. This instability undermines the overall harmony of the image, which detracts from the viewer's experience.

To overcome these challenges, we propose *MagicColor*, a novel framework that streamlines the colorization process. Our approach builds upon pre-trained diffusion model priors, capitalizing on their learned capacity to enforce visual consistency across generated outputs. The architectural cornerstones of our approach are articulated as follows. First, to tackle the misalignment between the reference instances and target line art, we incorporate an explicit reference mechanism that seamlessly infuses reference-derived color semantics and artistic styles into line art. In addition, we employ implicit latent control to ensure precise reference to each instance, significantly enhancing color accuracy and consistency. Second, to further improve the visual quality and color consistency of the output, we introduce edge loss and color matching. These methods compel the model to genuinely extract color information from the reference character design, thereby improving the accuracy of semantic correspondence and reducing the reliance on any color information that may inadvertently leak from the reference image. Third, our model adopts a two-stage, self-play training strategy, which addresses the challenges of limited multi-instance training data and subsequently incorporates additional reference images to refine the colorization capability. By facilitating colorization across multiple instances, our model achieves impressive color consistency with minimal human intervention. Our approach achieves state-of-the-art results across both quantitative metrics and qualitative evaluations, outperforming prior art in animated content creation. We aim for a critical step toward fully automated, high-efficiency animation pipelines with guaranteed stylistic coherence. This method can also be extended beyond anime to the broader digital art and media fields. Our contributions can be summarized as follows:

- We propose ***MagicColor***, the first multi-instance coloring method to support multiple instances integration for sketch colorization in a single forward pass.
- Technically, to solve the lack of multi-instance data, we design a two-stage, self-play training strategy. We also propose an instance guider and pixel-level color matching with edge loss to enhance the color correspondence.

2. Related Work

2.1. Line Art Colorization

Line art colorization techniques strive to decode the link between semantics and color by leveraging large-scale datasets [8, 47–51, 63, 91–93, 97]. Researchers utilize a series of semantic modules such as classification [22], semantic segmentation [17, 102], and instance-aware information [62, 69] to enrich color vibrancy. These methods generally perform well when the object's color has a strong semantic-based determinacy. However, when confronted with objects that exhibit a broad range of colors, they often produce lackluster results. In contrast, our proposed framework, equipped with an innovative imagination module, is engineered to transcend this constraint.

On the other hand, generative priors enshrined within pretrained Generative Adversarial Networks (GANs) [16, 72, 81] and Diffusion models have been cornerstones in pursuing photorealistic colorization. GANs are adept at both conditional and unconditional image synthesis. Specific conditioning factors, such as layout and semantic maps, are harnessed to fine-tune the image synthesis process [70, 71]. For example, the StyleGAN-family models have demonstrated remarkable prowess in generating high-resolution images without explicit conditioning [25, 26, 82]. Some models struggle to maintain the local spatial integrity of grayscale inputs but can use diverse color priors from pretrained models [13, 21, 62]. In our sketch colorization task, we aim to adapt these generative prior-based techniques. Our method can preserve sketch structures and semantics while using their color-generation abilities.

2.2. Visual Correspondence

In computer vision, visual correspondence aims to identify and match relevant features or points among different images. It is widely applied in tasks such as stereo vision and motion tracking. In the past, traditional methods [2, 45] relied on hand-designed features to establish corresponding relationships. Nowadays, deep learning methods [9, 12, 28, 31] obtain matching capabilities through supervised learning with labeled datasets. However, supervised learning faces significant scalability issues. Precise pixel-level annotations are not only time-consuming and labor-intensive but also costly. To address this, scholars have started to explore weakly-supervised or self-supervised visual correspondence models. For example, LightGlue [38] can match sparse local features across image pairs through an adaptive mechanism. CoTracker [24] adopts a semi-supervised training method by generating pseudo-labels using off-the-shelf models. DIFT [73] extracts features by diffusion models and can achieve pixel-level semantic point matching. Building upon the diffusion models' prior correspondence knowledge, our framework enables reference-based colorization via semantic-aligned color mapping between line art and reference instances, without any structural modifications.

2.3. Reference-Based Image Colorization

A great deal of research has been committed to colorizing photographs using reference-based priors [4, 15, 18, 74, 90]. Initially, efforts focused on the transfer of chromatic information to the corresponding regions through luminance and texture alignment, with various low-level feature-based correspondence techniques developed for more precise local color transfer [7, 32, 40, 52, 79]. However, these methods are vulnerable to complex appearance variations of the same object, as low-level features cannot capture semantic nuances [17, 98]. Line art colorization is notably different from natural image colorization [65, 83, 88, 95].

The Diffusion models have emerged as a powerful alternative. The Denoising Diffusion Probabilistic Model [46] and the Denoising Diffusion Implicit Model [67] paved the way for Latent Diffusion Models (LDMs) like Stable Diffusion [60], revolutionizing text-to-image generation. Building on LDMs, ControlNet [96] uses task-specific conditions and multi-modal inputs [41, 57, 85] to control pretrained diffusion models [14, 37, 76, 84, 86, 87, 103, 104]. Reference-based colorization [42, 43, 53, 55, 75, 80], guided by a user-provided reference image, has also become popular, with existing methods using strategies such as segmented graphs, active learning, and attention networks. AnimeDiffusion [6] and ColorizeDiffusion [89] introduce a diffusion-based reference-based framework for anime face colorization. Paint-by-Example [90] and ObjectStitch [68] leverage CLIP as their cross-modal encoder to extract instance-level visual-semantic embeddings, whereas AnyDoor [10] innovates by training on video sequences and adopting DINOv2 [54] for spatial-temporal feature learning. Despite these advancements, all frameworks focus on generic object categories, falling short of fine-grained part-level alignment required for intricate design tasks [77, 78].

3. Preliminaries

3.1. Latent Diffusion Model

As the core architecture of Stable Diffusion [60], Latent Diffusion Models (LDM) revolutionize text-to-image generation by executing diffusion-denoising processes in a compressed latent space rather than the raw pixel domain, enabling stable and efficient training. The pipeline begins with a Variational Autoencoder (VAE) projecting RGB images into low-dimensional latent codes, where semantic-guided diffusion sampling occurs under textual conditioning. Then, a UNet-based network incorporates self-attention and cross-attention mechanisms through UNet blocks to learn the reverse denoising process in the latent space. Cross-attention establishes bidirectional interactions between text embeddings and visual features, ensuring prompt semantics are continuously infused. The whole training objective of the UNet can be written as:

$$\mathcal{L}_{LDM} = \mathbb{E}_{t,\mathbf{z},\epsilon}\left[\left\|\epsilon - \epsilon_\theta\left(\sqrt{\alpha_t}\mathbf{z} + \sqrt{1-\alpha_t}\epsilon, c, t\right)\right\|^2\right], \tag{1}$$

where z notes the latent embedding of the training sample. ϵ_θ and ϵ represent predicted noise by the diffusion model and ground truth noise at corresponding timestep t, respectively. c is the condition embedding involved in the generation, and the coefficient α_t remains consistent with that employed in vanilla diffusion models.

3.2. Reference Condition Injection

In the current scenario, reference instances frequently suffer from noise, redundancy, and semantic conflicts, which significantly impede the model's ability to learn colors effectively. To address this issue, when presented with N user-provided reference instances along with the line art input, our model employs a dual-branch condition injection strategy. This strategy aims to achieve semantic similarity and structural alignment with the input. First, we align multiple instances in a layer and input them into the reference net. Then, we encode with CLIP and apply reference attention mechanisms [4, 18, 44] to inject color and semantic info into the UNet. Also, the line art is injected into the UNet to enhance line-structure features. This lets the model learn global reference info evenly and avoid interference. Second, for instance-aware image generation, we utilize instance images I_i embedding as latent control signals. Unlike other methods [44, 62], our paradigm supports the use of multiple instances, which enhances the model's generalizability. Zero-shot customization is tough, so we use pretrained vision models to extract the target object identity. Previous studies used CLIP for target embedding [34, 59]. We use DINOv2 as the feature encoder [54] to get discriminative spatial identity. DINOv2, trained with patch-level objectives under random masking, has highly expressive features. Its output includes a $26 \times 26 \times 1024$ spatial embedding s_i for patch-level features and a 1024 dimensional global embedding g_i.

4. Method

Problem Definition. We first formulate the task of multi-instance sketch colorization with per-instance chromatic control, enabling precise mapping from multiple reference objects to corresponding line drawing instances via semantic correspondence. We can represent the problem in a more concise way using the following formula:

$$I_{out} = (S, R, M), \tag{2}$$

where I_{out} is the output image to be generated. The condition C is composed of a line art image S, a set of reference instances $R = \{[I_1, \cdots, I_i] \mid i = 1, \cdots, N\}$, and a set of instance masks $M = \{[M_1, \cdots, M_i] \mid i = 1, \cdots, N\}$. Each I_i in R corresponds to an instance, and each M_i in M is a binary mask indicating the reference spatial location of an instance.

Architecture Design. Owing to the high-fidelity detail demands in anime sketch colorization, the core challenge resides in designing encoders capable of spatially precise reference image parsing to extract sub-pixel-level visual cues. Inspired by recent studies [19, 35, 44] which demonstrated the effectiveness of leveraging an additional UNet architecture for this purpose, we introduce a reference net following a similar design. Additionally, the sketch and instance guider are designed based on ControlNet [96], which provides an efficient condition for injecting line art and instance-level information into the generative process. Furthermore, we employ DINOv2 [54] to encode images and train a Feed-Forward Network layer to extract features from the image embeddings. These extracted features are then integrated into the latent control signals.

4.1. Self-Play Training Strategy

We then introduce a two-stage, self-play training approach to solve the problem of a lack of multi-instance training data and gradually boost the model's performance to build an advanced model that can give high-quality coloring results.

Single-Reference Colorization Training. In the first stage, training starts by activating the reference net, UNet, and the sketch guider. In both stages, we use color matching with edge loss for training. For each anime sequence, we perform random frame sampling without replacement: one frame is designated as the style reference exemplar, while the other serves as the raw input sketch for colorization. We use the whole reference image as a conditional input and extract the line art from the original image. Given a reference and line art image pair, the model maps color semantics from the reference image to the line art by minimizing a carefully designed edge loss function, aiming to make the model learn basic color-semantic relationships, forming the basis for later multi-instance refinement.

Multi-Instance Refinement. In the second stage, we train the reference net, UNet, sketch guider, and instance guider

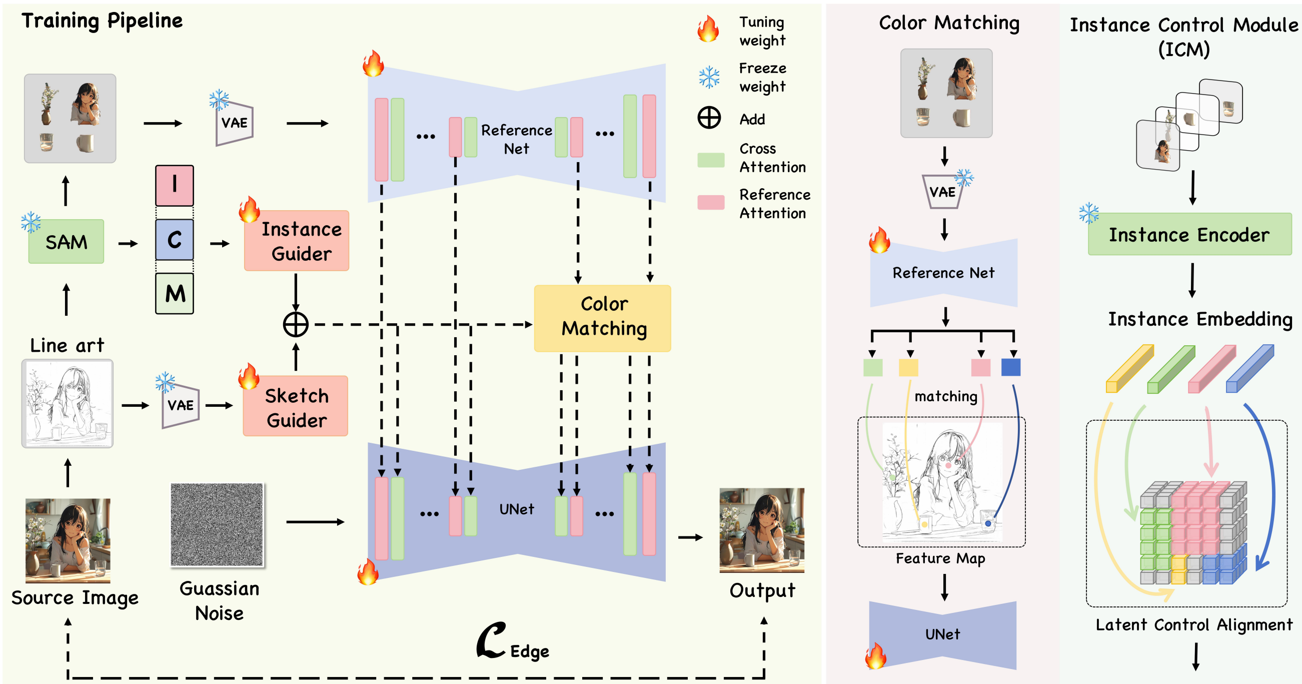

Figure 3. **Overview of the MagicColor pipeline.** We combine a dual-UNet framework with an instance control module (ICM). During training, we use multiple instances to set the overall color accurately. The color matching enables the model to better align the colors of the target image with those of the reference instances precisely. The edge loss helps the model pay more attention to the high-frequency areas and edges, resulting in a more accurate and vivid colorization for each instance.

to improve the model's ability to handle multiple instances and make local coloring more accurate. To precisely control each instance in the references, we use the Segment Anything Model (SAM) [29, 39, 100] to extract instances separately and perform operations like random fusion, scaling, shuffling, and adding noise to instances in a layer as reference net input to enhance the model's per-instance semantic perception for accurate colorization. Then, the model encodes all instances, aligning them in latent space, and starts coloring each instance from the sketch, considering their features and overall relationships. This process results in more accurate colorization, especially in areas with high color and semantic requirements.

4.2. Instance Control Module

Previous methods often resorted to directly imposing constraints within the reference attention mechanism for control purposes [4, 18, 19]. Nevertheless, when faced with the task of handling multiple instances as conditional input, this conventional approach proved to be extremely arduous. To uphold semantic correspondence and safeguard details, we align image embeddings in the latent space, thereby empowering more precise and effective handling of the intricate input scenarios presented by multiple instances. Given target regions (instance masks), we align the bounding rectangle of the mask with the input DINO feature maps using box coordinates. Subsequently, we use the dense grids of the mask within this rectangle to interpolate features from the input DINO features. This process is similar to ROI alignment, but the resolution of the ROI is flexible and follows the size of the mask. Finally, to further incorporate global information, we randomly drop 10% of the spatial embedding s_i and replace it with the DINO global embedding g_i. As our target mask may have a different silhouette from the shape of the input image, injecting global embedding enhances the model's generalizability and adaptability across diverse image conditions. The final ROI feature is then warped into the original region of the latent control signal:

$$l_c = \sum_{i=1}^{n} Drop\_out(Interpolate(s_i, M_i, B_i), g_i), \quad (3)$$

where B_i is the bounding box of the target mask M_i. To establish an instance-aware colorization framework, we use a latent control signal, denoted as l_c, which is a latent feature with a size of $l_c \in \mathbb{R}^{C\times H'\times W'}$. As shown in Figure 4, our model with an instance control module achieves precise instance-level control to produce varied colorization outcomes. Although the training data is derived from anime datasets, the references can incorporate real-life images, highlighting the model's adaptability across different domains and flexible utilization of diverse reference sources.

4.3. Structure-Content Supervision Enhancement

We present a comprehensive approach to enhance the performance of diffusion models in line art colorization. By introducing edge loss and color matching, we address crucial aspects of image quality, resulting in more visually appealing and perceptually accurate generated images.

Edge Loss. During the diffusion model's training process, each pixel in an image contributes equally to the supervision process. The diffusion model's original training objective is pixel-level mean squared error (MSE) loss. However, simultaneously, we aim to make the image content more consistent with human perception of image quality. To enhance the supervision over the high-frequency area and improve the generation quality, we propose the edge loss, which consists of the perceptual loss and the re-weighted edge loss. The formula for the perceptual loss is as follows:

$$\mathcal{L}_{\text{perceptual}} = \frac{1}{N}\sum_{i=1}^{N} |\phi_i(z) - \phi_i(\hat{z})|_2^2\,, \tag{4}$$

where $\hat{z}$ is the prediction latent embedding obtained by decoding ϵ_θ, N represents the number of feature layers used for calculation, and ϕ_i is the feature-extraction function corresponding to the i-th layer of the pre-trained neural network. λ is a hyper-parameter that balances $\mathcal{L}_{LDM}$ and the perceptual loss $\mathcal{L}_{perceptual}$. Moreover, in scenes characterized by complex hierarchical structures and overlapping objects, the importance of object edges may differ significantly from that of the plain background pixels. We calculate the overall instances edge map, which effectively mitigates interference from the background as we focus solely on the edges of individual objects. Then, we apply structure foreground enhancement from SyntheOcc [33] to the edge pixels. Since we perform edge detection in the latent space with the size of the input images, we regard the edge map as a loss weight map w to enhance structure supervision. Finally, our total loss can be written as:

$$\mathcal{L} = w \cdot \mathcal{L}_{LDM} + \lambda \cdot \mathcal{L}_{perceptual} \tag{5}$$

Color Matching. We employ pre-trained diffusion models to identify corresponding points within real-world images. Our model, which has a UNet at its core, processes noisy images by cleaning them and extracting features crucial for establishing correspondences. Two steps utilize a diffusion feature map to match pixels between two images. **(1) Semantic Matching.** This is accomplished by determining the nearest neighbors and computing the similarity using cosine distance. To obtain pixel correspondences, we begin by extracting dense features from both images and matching them. Let F_i be the feature map of an image i.

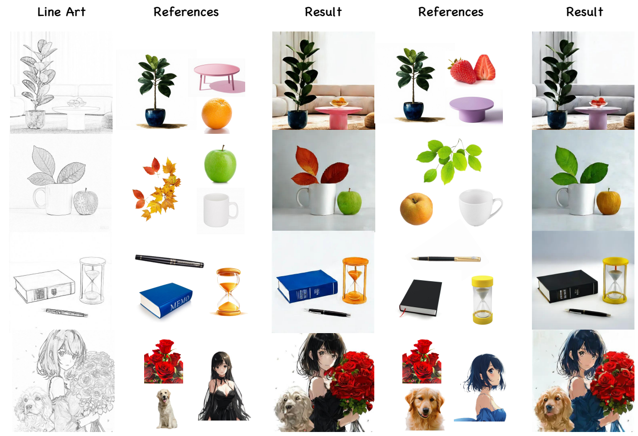

Figure 4. **Instance Control Ability.** With the same line art and diverse reference instances, our method achieves precise instance-level control for varied colorization and all without needing extra guidance.

For a pixel at position p, we obtain the feature vector $F(p)$ through bilinear interpolation. In terms of color matching, $C_{reference}$ denotes the color feature of the reference image, and C_{source} denotes the color feature of the source image. We obtain C_{source} and $C_{reference}$ by sampling features from the normalized source and target features. To quantify their similarity, we calculate the Euclidean distance (D) between $C_{reference}$ and the flattened C_{source}. Given two feature vectors $v_1 = C_{reference}$ and $v_2 = C_{source}$, the Euclidean distance is computed as follows:

$$D = \sqrt{\sum_{k=1}^{f} (v_{1k} - v_{2k})^2}, \tag{6}$$

where f represents the feature dimension. Subsequently, we find the nearest-neighbor indices. **(2) Feature injection**. Given a pixel p_1 in v_1, we find the corresponding pixel p_2 in v_2 as follows:

$$p_2 = \arg\min_{p} \; d\left(F_1\left(p_1\right), F_2(p)\right), \tag{7}$$

where d is the cosine distance. This approach transfers color-related information between the two images.

5. Experiments

5.1. Implementation Detail

Dataset. The dataset employed in this study comprises two key components: the anime video dataset and the image dataset. The data were preprocessed from Sakuga [56], ATD-12K [66], and manually collected from the internet. A total of 1,670 image pairs were selected from animations to form the test set, and the remaining data constituted the training set. We then extracted their instances as references.

Table 3. **Quantitative Comparisons Across Datasets**

TwoTwoTwopt¡ -Twopt¿ Cross Domain Dataset	FID ↓		PSNR ↑		SSIM ↑		LPIPS ↓	
	Animation	Hand-drawn	Animation	Hand-drawn	Animation	Hand-drawn	Animation	Hand-drawn
RSIC(256x)	128.971	300.636	16.93	8.872	0.548	0.416	0.508	0.566
SGA(256x)	124.73	306.881	19.00	7.253	0.621	0.414	0.454	0.533
AnimeDiffusion(256x)	154.78	305.021	11.26	10.207	0.416	0.426	0.557	0.549
ColorizeDiffusion(512x)	128.20	298.673	12.882	9.817	0.492	0.349	0.475	0.613
MangaNinja(512x)	43.16	295.223	14.28	10.249	0.543	0.361	0.362	0.598
Ours(512x)	**28.95**	**251.301**	**23.75**	**10.551**	**0.783**	**0.434**	**0.201**	**0.431**

Baseline design. The backbone architecture for both the UNet and reference net is derived from Stable Diffusion 1.5 [61]. The sketch guider and instance guider are initialized with pre-trained ControlNet weights. The instance encoder is initialized using DINOv2.

Hyper-parameters. During training, we resized the height and weight of the input image to 512 and kept the original aspect ratio. Our model undergoes 100,000 optimization iterations with a batch size of 1. The learning rate is set to 1×10^{-5}. The training phase takes around 7 days using 2 NVIDIA A800 80G GPUs. We perform a random horizon flip and random brightness adjustment for each reference image as data augmentation to simulate multi-view conditions.

5.2. Comparison

5.2.1. Qualitative Results

To achieve the same settings, we use a complete reference image as the reference condition for the comparative baseline methods, and our method employs instances extracted from the same reference image. We compare our approach with previous sketch colorization methods. For GAN-based approaches, we consider RSIC [30] and SGA [36]. RSIC and SGA utilize GANs for colorization, each with a distinct architecture and training strategy. We observe that GAN-based methods tend to produce color incoordination. Moreover, when there are numerous color block regions in the image, color bleeding issues are likely to occur, where the color of one color block contaminates the adjacent color blocks. Regarding diffusion-based models, AnimeDiffusion [6], ColorizeDiffusion [89], and MangaNinja [44] (without point guided) perform better in terms of color control. However, as shown in Figure 6, they still struggle to effectively transfer the colors from the reference instances to the line drawing sketch, especially for the colors of smaller details in the images. The root cause is that the models fail to fully learn the color correspondences between the original image and the reference instances.

5.2.2. Quantitative Results

To comprehensively evaluate the colorization ability of our model, we quantitatively compared our method with the state-of-the-art colorization method on our test set. We constructed an **Animation Dataset** (1,570 pairs) from two anime films (*Your Name* and *Spirited Away*) and a **Hand-Drawn Dataset** (100 pairs) collected from real-user

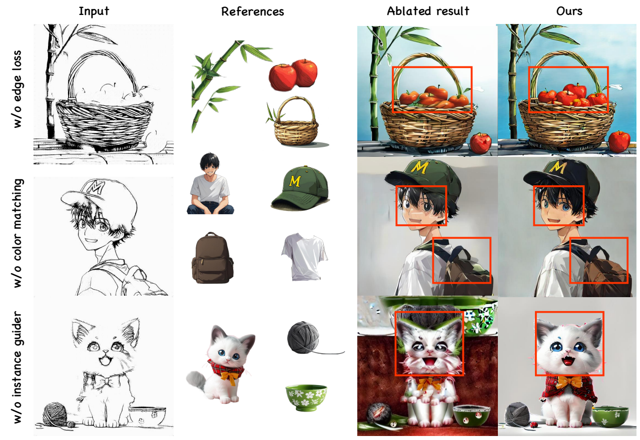

Figure 5. **Ablations on each component.** "w/o edge loss" indicates without the edge loss, "w/o color matching" indicates without the color matching, "w/o instance guider" indicates without the instance guider.

Table 4. **Quantitative results of ablation study.**

	FID ↓	PSNR ↑	SSIM ↑	LPIPS ↓
All	**62.531**	**22.587**	**0.806**	**0.203**
w/o Edge Loss	83.042	20.842	0.755	0.217
w/o Instance Guider	91.852	16.467	0.718	0.308
w/o Color Matching	88.573	20.832	0.749	0.226

sketches as a test set. The results are presented in Table 3. We did not consider background noise when evaluating the experimental results. Due to the limited resolution of most previous work, all measurements are performed at a resolution of 512 × 512. We report four metrics of different methods. **FID** gauges visual similarity, with lower values indicating better quality. **PSNR** measures distortion, and higher values imply less distortion. **SSIM** assesses luminance, contrast, and structure similarity, with values closer to 1 showing better structural preservation. **LPIPS** measures perceptual similarity, and lower values mean closer resemblance to the reference. By comparing our model with these GAN-based and diffusion-based methods using the established metrics, our approach demonstrates a significant advantage over previous methods.

5.3. Ablation Study

Effectiveness of Edge Loss. To evaluate edge loss, we replace its edge loss with the original diffusion loss. Our method fails to preserve structural edge information, resulting in mismatched colors on the apple's edges, as shown in the first row of Figure 5. When we eliminate the overall structure edge loss, the model struggles to maintain subtle edge features. Numerical evaluations in Table 4 confirm the importance of structure edge loss for preserving image in-

Figure 6. **Qualitative comparisons with existing methods.** Given a line drawing and multiple reference instances, our method demonstrates far more precise colorization and higher-quality results compared to other methods, effectively maintaining line-drawing structure and reference instances' color consistency.

tegrity.

Effectiveness of Color Matching. Removing the color-matching component diminishes our model's ability to retain color information and details from input references. For instance, in the second row of Figure 5, the colors of the bag and the character's facial details are inconsistent.

Effectiveness of Instance Guider. To illustrate the effectiveness of our instance guider, we remove it during the experiment. Visual results in Figure 5 show that without this module, our model struggles to transfer instance-level color information from references effectively. As can be seen in the third row, the lack of an instance guider hampers the learning of instance-level information, leading to significant color loss and noise in the final results when relying solely on the reference net.

5.4. Limitations and Discussion

Our method has many advantages and potential applications, but it also faces several limitations. We summarize them as follows: (1) **Flexible usage:** When stylizing the same sketch with diverse reference images, our method retains the character's identity. It adjusts details like lighting and background based on the reference styles. (2) **Semantic awareness:** For cartoon wallpapers or posters with multiple characters or objects, our approach uses multiple reference images to perform semantic-based colorization, ensuring each line art element gets colorized semantically. (3) **Multi-object and occlusions:** Single-subject line art can efficiently convert a sketch into a vivid, full-colored illustration, speeding up the production and enabling animators to test different color concepts quickly. However, in sketch images with many main objects or significant occlusions between them, the detailed inter-object colors may not be well maintained. (4) **More explorations:** The model can accurately color each element in multi-subject scenarios such as an anime battle scene or a forest scene while maintaining a harmonious overall color scheme.

6. Conclusion

This paper has presented *MagicColor*, a diffusion-based sketch colorization framework. Specifically, we introduce a multi-instance approach, a two-stage, self-play training strategy, and pixel-level color matching with edge loss. Our experiments demonstrated that *MagicColor* outperforms current methods in visual quality and style consistency, advancing the field of digital cartoon colorization. In future work, we plan to substantially augment the number of reference instances and enhance the model's capacity to maintain semantic and color consistency. We also plan to release the source code for better collaboration between creative practitioners and AI researchers.

7. Acknowledagment

This work was partially supported by the Guangdong Provincial Key Lab of Integrated Communication, Sensing and Computation for Ubiquitous Internet of Things #2023B1212010007. The Research Grant Council of the Hong Kong Special Administrative Region under grant number 16203122.

References

[1] Yunpeng Bai, Chao Dong, Zenghao Chai, Andong Wang, Zhengzhuo Xu, and Chun Yuan. Semantic-Sparse Colorization Network for Deep Exemplar-Based Colorization. In *European Conference on Computer Vision*, pages 505–521. Springer, 2022. 2

[2] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. SURF: Speeded Up Robust Features. In *Computer Vision–ECCV 2006: 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006. Proceedings, Part I 9*, pages 404–417. Springer, 2006. 3

[3] Tim Brooks, Aleksander Holynski, and Alexei A Efros. InstructPix2Pix: Learning to Follow Image Editing Instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 2

[4] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-Free Mutual Self-Attention Control for Consistent Image Synthesis and Editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22560–22570, 2023. 3, 4, 5

[5] Ruizhi Cao, Haoran Mo, and Chengying Gao. Line Art Colorization Based on Explicit Region Segmentation. In *Computer Graphics Forum*, pages 1–10, 2021. 2

[6] Yu Cao, Xiangqiao Meng, PY Mok, Xueting Liu, Tong-Yee Lee, and Ping Li. AnimeDiffusion: Anime Face Line Drawing Colorization via Diffusion Models. *arXiv preprint arXiv:2303.11137*, 2023. 3, 7

[7] Hernan Carrillo, Michaël Clément, Aurélie Bugeau, and Edgar Simo-Serra. Diffusart: Enhancing Line Art Colorization with Conditional Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3486–3490, 2023. 2, 3

[8] Evan Casey, Víctor Pérez, and Zhuoru Li. The animation transformer: Visual correspondence via segment matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11323–11332, 2021. 3

[9] Qihua Chen, Yue Ma, Hongfa Wang, Junkun Yuan, Wenzhe Zhao, Qi Tian, Hongmei Wang, Shaobo Min, Qifeng Chen, and Wei Liu. Follow-Your-Canvas: Higher-Resolution Video Outpainting with Extensive Content Generation. *arXiv preprint arXiv:2409.01055*, 2024. 3

[10] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-Shot Object-Level Image Customization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6593–6602, 2024. 3

[11] Alex Yong-Sang Chia, Shaojie Zhuo, Raj Kumar Gupta, Yu-Wing Tai, Siu-Yeung Cho, Ping Tan, and Stephen Lin. Semantic Colorization With Internet Images. *ACM Transactions on Graphics (TOG)*, 30(6):1–8, 2011. 2

[12] Seokju Cho, Sunghwan Hong, Sangryul Jeon, Yunsung Lee, Kwanghoon Sohn, and Seungryong Kim. CATs: Cost Aggregation Transformers for Visual Correspondence. *Advances in Neural Information Processing Systems*, 34: 9011–9023, 2021. 3

[13] Zhi Dou, Ning Wang, Baopu Li, Zhihui Wang, Haojie Li, and Bin Liu. Dual Color Space Guided Sketch Colorization. *IEEE Transactions on Image Processing*, 30:7292–7304, 2021. 3

[14] Kunyu Feng, Yue Ma, Bingyuan Wang, Chenyang Qi, Haozhe Chen, Qifeng Chen, and Zeyu Wang. Dit4Edit: Diffusion Transformer for Image Editing. *arXiv preprint arXiv:2411.03286*, 2024. 3

[15] Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel Cohen-Or. An Image is Worth One Word: Personalizing Text-To-Image Generation Using Textual Inversion. *arXiv preprint arXiv:2208.01618*, 2022. 3

[16] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative Adversarial Nets. *Advances in neural information processing systems*, 27, 2014. 3

[17] Sergio Guadarrama, Ryan Dahl, David Bieber, Mohammad Norouzi, Jonathon Shlens, and Kevin Murphy. PixColor: Pixel Recursive Colorization. *arXiv preprint arXiv:1705.07208*, 2017. 3

[18] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-To-Prompt Image Editing with Cross Attention Control. *arXiv preprint arXiv:2208.01626*, 2022. 3, 4, 5

[19] Li Hu. Animate Anyone: Consistent and Controllable Image-To-Video Synthesis for Character Animation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8153–8163, 2024. 4, 5

[20] Yifei Huang, Sheng Qiu, Changbo Wang, and Chenhui Li. Learning Representations for High-Dynamic-Range Image Color Transfer in a Self-Supervised Way. *IEEE Transactions on Multimedia*, 23:176–188, 2020. 2

[21] Zhitong Huang, Nanxuan Zhao, and Jing Liao. UniColor: A Unified Framework for Multi-Modal Colorization With Transformer. *ACM Transactions on Graphics (TOG)*, 41 (6):1–16, 2022. 3

[22] Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. Let There Be Color! Joint End-To-End Learning of Global and Local Image Priors for Automatic Image Colorization With Simultaneous Classification. *ACM Transactions on Graphics (TOG)*, 35(4):1–11, 2016. 3

[23] Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. Colorformer: Image Colorization via Color Memory Assisted Hybrid-Attention Transformer. In *European Conference on Computer Vision*, pages 20–36. Springer, 2022. 2

[24] Nikita Karaev, Iurii Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. COTRACKER3: Simpler and Better Point Tracking by Pseudo-Labelling Real Videos. *arXiv preprint arXiv:2410.11831*, 2024. 3

[25] Tero Karras. A Style-Based Generator Architecture for Generative Adversarial Networks. *arXiv preprint arXiv:1812.04948*, 2019. 3

[26] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of Stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 3

[27] Hyunsu Kim, Ho Young Jhoo, Eunhyeok Park, and Sungjoo Yoo. Tag2pix: Line Art Colorization Using Text Tag With Secat and Changing Loss. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9056–9065, 2019. 2

[28] Seungwook Kim, Juhong Min, and Minsu Cho. TransforMatcher: Match-To-Match Attention for Semantic Correspondence. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8697–8707, 2022. 3

[29] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment Anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026, 2023. 5

[30] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5801–5810, 2020. 7

[31] Jae Yong Lee, Joseph DeGol, Victor Fragoso, and Sudipta N Sinha. Patchmatch-Based Neighborhood Consensus for Semantic Correspondence. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13153–13163, 2021. 3

[32] Chenyang Lei and Qifeng Chen. Fully Automatic Video Colorization With Self-Regularization and Diversity. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3753–3761, 2019. 3

[33] Leheng Li, Weichao Qiu, Yingjie Cai, Xu Yan, Qing Lian, Bingbing Liu, and Ying-Cong Chen. SyntheOcc: Synthesize Geometric-Controlled Street View Images through 3D Semantic MPIs. *arXiv preprint arXiv:2410.00337*, 2024. 6

[34] Leheng Li, Weichao Qiu, Xu Yan, Jing He, Kaiqiang Zhou, Yingjie Cai, Qing Lian, Bingbing Liu, and Ying-Cong Chen. OmniBooth: Learning Latent Control for Image Synthesis with Multi-Modal Instruction. *arXiv preprint arXiv:2410.04932*, 2024. 4

[35] Xinghui Li, Qichao Sun, Pengze Zhang, Fulong Ye, Zhichao Liao, Wanquan Feng, Songtao Zhao, and Qian He. Anydressing: Customizable Multi-Garment Virtual Dressing via Latent Diffusion Models. *arXiv preprint arXiv:2412.04146*, 2024. 4

[36] Zekun Li, Zhengyang Geng, Zhao Kang, Wenyu Chen, and Yibo Yang. Eliminating Gradient Conflict in Reference-Based Line-Art Colorization. In *European Conference on Computer Vision*, pages 579–596. Springer, 2022. 7

[37] Jun Hao Liew, Hanshu Yan, Daquan Zhou, and Jiashi Feng. MagicMix: Semantic Mixing with Diffusion Models. *arXiv preprint arXiv:2210.16056*, 2022. 3

[38] Philipp Lindenberger, Paul-Edouard Sarlin, and Marc Pollefeys. Lightglue: Local Feature Matching at Light Speed. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 17627–17638, 2023. 3

[39] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding Dino: Marrying Dino With Grounded Pre-training for Open-Set Object Detection. In *European Conference on Computer Vision*, pages 38–55. Springer, 2025. 5

[40] Xueting Liu, Wenliang Wu, Chengze Li, Yifan Li, and Huisi Wu. Reference-Guided Structure-Aware Deep Sketch Colorization for Cartoons. *Computational Visual Media*, 8: 135–148, 2022. 3

[41] Yifan Liu, Zengchang Qin, Tao Wan, and Zhenbo Luo. Auto-Painter: Cartoon Image Generation from Sketch by Using Conditional Wasserstein Generative Adversarial Networks. *Neurocomputing*, 311:78–87, 2018. 3

[42] Zhiheng Liu, Ruili Feng, Kai Zhu, Yifei Zhang, Kecheng Zheng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones: Concept Neurons in Diffusion Models for Customized Generation. *arXiv preprint arXiv:2303.05125*, 2023. 3

[43] Zhiheng Liu, Yifei Zhang, Yujun Shen, Kecheng Zheng, Kai Zhu, Ruili Feng, Yu Liu, Deli Zhao, Jingren Zhou, and Yang Cao. Cones 2: Customizable Image Synthesis With Multiple Subjects. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, pages 57500–57519, 2023. 3

[44] 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. *arXiv preprint arXiv:2501.08332*, 2025. 4, 7

[45] David G Lowe. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 60:91–110, 2004. 3

[46] He Lyu, Ningyu Sha, Shuyang Qin, Ming Yan, Yuying Xie, and Rongrong Wang. Advances in Neural Information Processing Systems. *Advances in neural information processing systems*, 32, 2019. 3

[47] Yue Ma, Yali Wang, Yue Wu, Ziyu Lyu, Siran Chen, Xiu Li, and Yu Qiao. Visual Knowledge Graph for Human Action Reasoning in Videos. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4132–4141, 2022. 3

[48] Yue Ma, Xiaodong Cun, Yingqing He, Chenyang Qi, Xintao Wang, Ying Shan, Xiu Li, and Qifeng Chen. MagicStick: Controllable Video Editing via Control Handle Transformations. *arXiv preprint arXiv:2312.03047*, 2023.

[49] Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Xiu Li, and Qifeng Chen. Follow Your Pose: Pose-Guided Text-To-Video Generation Using Pose-Free Videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 4117–4125, 2024.

[50] Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, et al. Follow-Your-Click: Open-domain Regional Image Animation via Short Prompts. *arXiv preprint arXiv:2403.08268*, 2024.

[51] Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, et al. Follow-Your-Emoji: Fine-Controllable and Expressive Freestyle Portrait Animation. *arXiv preprint arXiv:2406.01900*, 2024. 3

[52] Akinobu Maejima, Hiroyuki Kubo, Takuya Funatomi, Tatsuo Yotsukura, Satoshi Nakamura, and Yasuhiro Mukaigawa. Graph Matching Based Anime Colorization With Multiple References. In *ACM SIGGRAPH 2019 Posters*, pages 1–2, 2019. 2, 3

[53] Yihao Meng, Hao Ouyang, Hanlin Wang, Qiuyu Wang, Wen Wang, Ka Leong Cheng, Zhiheng Liu, Yujun Shen, and Huamin Qu. AniDoc: Animation Creation Made Easier. *arXiv preprint arXiv:2412.14173*, 2024. 3

[54] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. DINOV2: Learning Robust Visual Features without Supervision. *arXiv preprint arXiv:2304.07193*, 2023. 3, 4

[55] Yulin Pan, Chaojie Mao, Zeyinzi Jiang, Zhen Han, and Jingfeng Zhang. Locate, Assign, Refine: Taming Customized Image Inpainting with Text-Subject Guidance. *arXiv preprint arXiv:2403.19534*, 2024. 3

[56] Zhenglin Pan, Yu Zhu, and Yuxuan Mu. Sakuga-42M Dataset: Scaling Up Cartoon Research. *arXiv preprint arXiv:2405.07425*, 2024. 6

[57] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-Supervised Dense Visual Alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13481, 2022. 2, 3

[58] Yingge Qu, Tien-Tsin Wong, and Pheng-Ann Heng. Manga Colorization. *ACM Transactions on Graphics (TOG)*, 25 (3):1214–1220, 2006. 2

[59] 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 *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 4

[60] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 3

[61] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis with Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, 2022. 7

[62] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 22500–22510, 2023. 3, 4

[63] Mehdi Safaee, Aryan Mikaeili, Or Patashnik, Daniel Cohen-Or, and Ali Mahdavi-Amiri. CLiC: Concept Learning in Context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6924–6933, 2024. 3

[64] Patsorn Sangkloy, Jingwan Lu, Chen Fang, Fisher Yu, and James Hays. Scribbler: Controlling Deep Image Synthesis with Sketch and Color. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5400–5409, 2017. 2

[65] Johannes L Schonberger and Jan-Michael Frahm. Structure-From-Motion Revisited. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4104–4113, 2016. 3

[66] Li Siyao, Shiyu Zhao, Weijiang Yu, Wenxiu Sun, Dimitris Metaxas, Chen Change Loy, and Ziwei Liu. Deep Animation Video Interpolation in the Wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6587–6595, 2021. 6

[67] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising Diffusion Implicit Models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[68] Yizhi Song, Zhifei Zhang, Zhe Lin, Scott Cohen, Brian Price, Jianming Zhang, Soo Ye Kim, and Daniel Aliaga. ObjectStitch: Object Compositing with Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18310–18319, 2023. 3

[69] Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. Instance-Aware Image Colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7968–7977, 2020. 3

[70] Wei Sun and Tianfu Wu. Learning Layout and Style Reconfigurable Gans for Controllable Image Synthesis. 3

[71] Wei Sun and Tianfu Wu. Image Synthesis From Reconfigurable Layout and Style. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 10531–10540, 2019. 3

[72] Vadim Sushko, Edgar Schönfeld, Dan Zhang, Juergen Gall, Bernt Schiele, and Anna Khoreva. You Only Need Adversarial Supervision for Semantic Image Synthesis. *arXiv preprint arXiv:2012.04781*, 2020. 3

[73] Luming Tang, Menglin Jia, Qianqian Wang, Cheng Perng Phoo, and Bharath Hariharan. Emergent Correspondence from Image Diffusion. *Advances in Neural Information Processing Systems*, 36:1363–1389, 2023. 3

[74] Luming Tang, Nataniel Ruiz, Qinghao Chu, Yuanzhen Li, Aleksander Holynski, David E Jacobs, Bharath Hariharan,

Yael Pritch, Neal Wadhwa, Kfir Aberman, et al. RealFill: Reference-Driven Generation for Authentic Image Completion. *ACM Transactions on Graphics (TOG)*, 43(4):1–12, 2024. 3

[75] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-Play Diffusion Features for Text-Driven Image-to-Image Translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1921–1930, 2023. 3

[76] Chaitat Utintu, Pinaki Nath Chowdhury, Aneeshan Sain, Subhadeep Koley, Ayan Kumar Bhunia, and Yi-Zhe Song. SketchDeco: Decorating B&W Sketches with Colour. *arXiv preprint arXiv:2405.18716*, 2024. 3

[77] Bingyuan Wang, Qifeng Chen, and Zeyu Wang. Diffusion-Based Visual Art Creation: A Survey and New Perspectives. *arXiv preprint arXiv:2408.12128*, 2024. 3

[78] Bingyuan Wang, Hengyu Meng, Rui Cao, Zeyu Cai, Lanjiong Li, Yue Ma, Qifeng Chen, and Zeyu Wang. MagicScroll: Enhancing Immersive Storytelling with Controllable Scroll Image Generation. In *2025 IEEE Conference Virtual Reality and 3D User Interfaces (VR)*. IEEE, 2025. 3

[79] Jiangshan Wang, Yue Ma, Jiayi Guo, Yicheng Xiao, Gao Huang, and Xiu Li. COVE: Unleashing the Diffusion Feature Correspondence for Consistent Video Editing. *arXiv preprint arXiv:2406.08850*, 2024. 3

[80] Jiangshan Wang, Junfu Pu, Zhongang Qi, Jiayi Guo, Yue Ma, Nisha Huang, Yuxin Chen, Xiu Li, and Ying Shan. Taming Rectified Flow for Inversion and Editing. *arXiv preprint arXiv:2411.04746*, 2024. 3

[81] Ting-Chun Wang, Ming-Yu Liu, Jun-Yan Zhu, Andrew Tao, Jan Kautz, and Bryan Catanzaro. High-Resolution Image Synthesis and Semantic Manipulation with Conditional Gans. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8798–8807, 2018. 3

[82] Xiaolong Wang, Allan Jabri, and Alexei A Efros. Learning Correspondence From the Cycle-Consistency of Time. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2566–2576, 2019. 3

[83] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image Quality Assessment: From Error Visibility to Structural Similarity. 3

[84] Shukai Wu, Yuhang Yang, Shuchang Xu, Weiming Liu, Xiao Yan, and Sanyuan Zhang. FlexIcon: Flexible Icon Colorization via Guided Images and Palettes. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 8662–8673, 2023. 3

[85] Shaoan Xie, Yang Zhao, Zhisheng Xiao, Kelvin CK Chan, Yandong Li, Yanwu Xu, Kun Zhang, and Tingbo Hou. DreamInpainter: Text-Guided Subject-Driven Image Inpainting With Diffusion Models. *arXiv preprint arXiv:2312.03771*, 2023. 3

[86] Zhongcong Xu, Jianfeng Zhang, Jun Hao Liew, Hanshu Yan, Jia-Wei Liu, Chenxu Zhang, Jiashi Feng, and Mike Zheng Shou. MagicAnimate: Temporally Consistent Human Image Animation Using Diffusion Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1481–1490, 2024. 3

[87] Jingyun Xue, Hongfa Wang, Qi Tian, Yue Ma, Andong Wang, Zhiyuan Zhao, Shaobo Min, Wenzhe Zhao, Kaihao Zhang, Heung-Yeung Shum, et al. Follow-Your-Pose v2: Multiple-Condition Guided Character Image Animation for Stable Pose Control. *arXiv preprint arXiv:2406.03035*, 2024. 3

[88] Dingkun Yan, Ryogo Ito, Ryo Moriai, and Suguru Saito. Two-Step Training: Adjustable Sketch Colourization via Reference Image and Text Tag. In *Computer Graphics Forum*, page e14791, 2023. 3

[89] Dingkun Yan, Liang Yuan, Erwin Wu, Yuma Nishioka, Issei Fujishiro, and Suguru Saito. Colorizediffusion: Adjustable Sketch Colorization With Reference Image and Text. *arXiv preprint arXiv:2401.01456*, 2024. 3, 7

[90] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by Example: Exemplar-Based Image Editing with Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18381–18391, 2023. 3

[91] Shenghai Yuan, Jinfa Huang, Xianyi He, Yunyuan Ge, Yujun Shi, Liuhan Chen, Jiebo Luo, and Li Yuan. Identity-Preserving Text-to-Video Generation by Frequency Decomposition. *arXiv preprint arXiv:2411.17440*, 2024. 3

[92] Shenghai Yuan, Jinfa Huang, Yujun Shi, Yongqi Xu, Ruijie Zhu, Bin Lin, Xinhua Cheng, Li Yuan, and Jiebo Luo. MagicTime: Time-Lapse Video Generation Models as Metamorphic Simulators. *arXiv preprint arXiv:2404.05014*, 2024.

[93] Nir Zabari, Aharon Azulay, Alexey Gorkor, Tavi Halperin, and Ohad Fried. Diffusing Colors: Image Colorization with Text Guided Diffusion. In *SIGGRAPH Asia 2023 Conference Papers*, pages 1–11, 2023. 3

[94] Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. Deep Exemplar-Based Video Colorization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8052–8061, 2019. 2

[95] Lvmin Zhang, Chengze Li, Tien-Tsin Wong, Yi Ji, and Chunping Liu. Two-stage Sketch Colorization. *ACM Transactions on Graphics (TOG)*, 37(6):1–14, 2018. 2, 3

[96] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding Conditional Control to Text-to-Image Diffusion Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3836–3847, 2023. 3, 4

[97] Qian Zhang, Bo Wang, Wei Wen, Hai Li, and Junhui Liu. Line Art Correlation Matching Feature Transfer Network for Automatic Animation Colorization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 3872–3881, 2021. 3

[98] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful Image Colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*, pages 649–666. Springer, 2016. 3

[99] Richard Zhang, Phillip Isola, and Alexei A Efros. Colorful Image Colorization. In *Computer Vision–ECCV 2016: 14th*

European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14, pages 649–666. Springer, 2016. 2

[100] Youcai Zhang, Xinyu Huang, Jinyu Ma, Zhaoyang Li, Zhaochuan Luo, Yanchun Xie, Yuzhuo Qin, Tong Luo, Yaqian Li, Shilong Liu, et al. Recognize Anything: A Strong Image Tagging Model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1724–1732, 2024. 5

[101] Hengyuan Zhao, Wenhao Wu, Yihao Liu, and Dongliang He. Color2Embed: Fast Exemplar-Based Image Colorization Using Color Embeddings. *arXiv preprint arXiv:2106.08017*, 2021. 2

[102] Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. Pixelated Semantic Colorization. *International Journal of Computer Vision*, 128:818–834, 2020. 3

[103] Chenyang Zhu, Kai Li, Yue Ma, Chunming He, and Xiu Li. MultiBooth: Towards Generating All Your Concepts in an Image from Text. *arXiv preprint arXiv:2404.14239*, 2024. 3

[104] Chenyang Zhu, Kai Li, Yue Ma, Longxiang Tang, Chengyu Fang, Chubin Chen, Qifeng Chen, and Xiu Li. InstantSwap: Fast Customized Concept Swapping across Sharp Shape Differences. *arXiv preprint arXiv:2412.01197*, 2024. 3