UniColor: A Unified Framework for Multi-Modal Colorization with Transformer

ZHITONG HUANG*, City University of Hong Kong, China
NANXUAN ZHAO*, University of Bath, United Kingdom
JING LIAO†, City University of Hong Kong, China

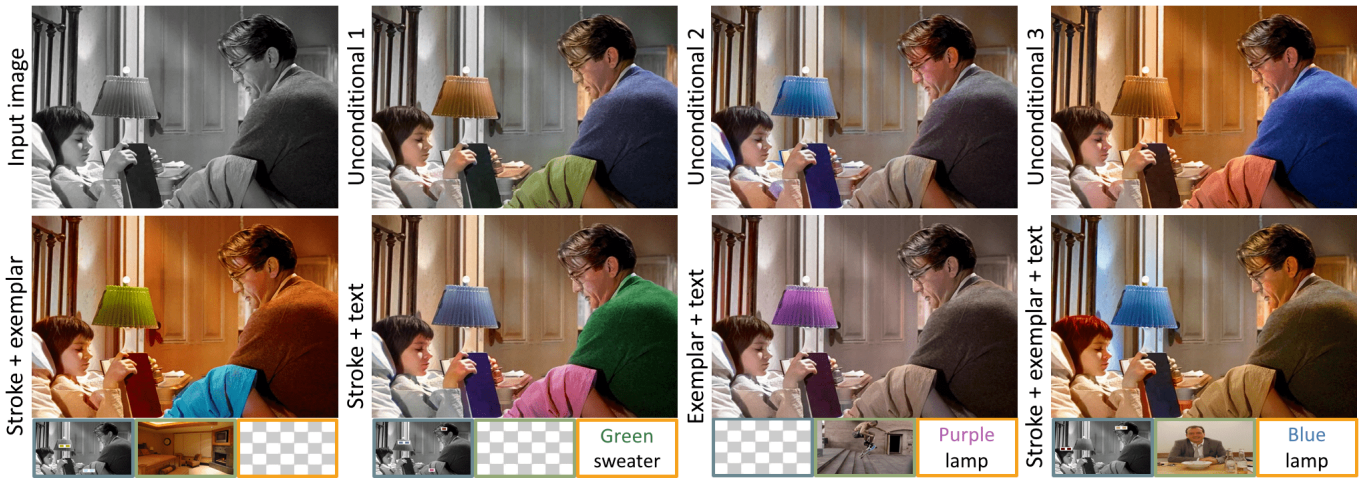

Fig. 1. Given an input grayscale image, our unified framework UniColor is able to: (a) produce diverse colorization results unconditionally (1^{st} row), and (b) colorize the image from a hybrid of controls (2^{nd} row), including stroke (in blue frame), exemplar (in green frame), and text (in orange frame). Input image: Gregory Peck & Mary Badham in film *To Kill a Mockingbird*, 1962. Reference images are from MSCOCO.

We propose the first unified framework *UniColor* to support colorization in multiple modalities, including both unconditional and conditional ones, such as stroke, exemplar, text, and even a mix of them. Rather than learning a separate model for each type of condition, we introduce a two-stage colorization framework for incorporating various conditions into a single model. In the first stage, multi-modal conditions are converted into a common representation of hint points. Particularly, we propose a novel CLIP-based method to convert the text to hint points. In the second stage, we propose a Transformer-based network composed of *Chroma-VQGAN* and *Hybrid-Transformer* to generate diverse and high-quality colorization results conditioned on hint points. Both qualitative and quantitative comparisons demonstrate that our method outperforms state-of-the-art methods in every control modality and further enables multi-modal colorization that was not feasible before. Moreover, we design an interactive interface showing the effectiveness of our unified framework in practical usage, including automatic colorization, hybrid-control colorization, local recolorization, and iterative color editing. Our code and models are available at *https://luckyhzt.github.io/unicolor*.

CCS Concepts: • **Computing methodologies** → **Image manipulation**; *Graphics systems and interfaces*; *Computational photography*; *Neural networks*.

Additional Key Words and Phrases: colorization, multi-modal controls, color editing, Transformer

*Both authors contributed equally to this research.
†Corresponding author.

Authors' addresses: Zhitong Huang, luckyhzt@gmail.com, City University of Hong Kong, Hong Kong SAR, China; Nanxuan Zhao, nanxuanzhao@gmail.com, University of Bath, Bath, United Kingdom; Jing Liao, jingliao@cityu.edu.hk, City University of Hong Kong, Hong Kong SAR, China.

1 INTRODUCTION

Colorization, a task of adding colors to grayscale images, has been actively studied recently [Jin et al. 2021; Kumar et al. 2021; Vitoria et al. 2020; Wu et al. 2021b]. As an ill-posed problem, colorization often has a one-to-many mapping, where many results could be semantically meaningful and visually pleasing for a single input grayscale image. For example, the lamp and sweater in Fig. 1 could have many diverse colors in the colorized images. Previous unconditional colorization methods [Iizuka et al. 2016; Su et al. 2020; Vitoria et al. 2020; Zhang et al. 2016] can only generate a deterministic result for a single grayscale image, failed to maintain the diverse and expressive nature. Some other methods [Deshpande et al. 2017; Kumar et al. 2021; Saharia et al. 2021] can produce diverse colorization results, but fail to customize the colors based on user control. To alleviate these problems, various conditional methods have been proposed, which can be categorized into: stroke-based methods [Endo et al. 2016; Xiao et al. 2019; Zhang et al. 2017], exemplar-based methods [He et al. 2018; Xu et al. 2020; Zhang et al. 2019], and text-based methods [Manjunatha et al. 2018]. The generated colors depend on how these conditions are designed.

Although conditional methods allow customized results, they are still limited to a single modality, decreasing the flexibility of the general usage. In practice, a combination of different interaction manners is often required to achieve a satisfactory colorization result. For example, in the last result of Fig. 1, we use different modalities to control different objects, *i.e.*, exemplar for large object

(sweater) and background (door), text for smaller object of the lamp, and stroke for details of the hair. Thus, a framework that enables multi-modal conditions for colorization is a natural choice and in demand for practical usage. In this work, we aim to develop the first unified colorization framework, producing diverse results under multi-modal controls, including both unconditional and conditional ones (*e.g.,* stroke, exemplar, and text).

However, creating such a unified colorization framework is not an easy task with two unique challenges. 1). *Multi-modal controls*: existing works often design and train the model supporting only a single type of user interaction and cannot generalize to the other ones directly. Given the complete different distributions on various conditions (*e.g.,* stroke, exemplar, and text), how to encode different modalities and take control of results with freely integrated conditions is critical to the framework design. 2). *Diversity and quality*: being a stochastic task, the framework needs to output diverse results with high quality, which are both semantically meaningful and visually pleasing.

To this end, we introduce *UniColor*, a novel unified framework for colorization with multi-modal interactions. Our framework can colorize a grayscale image from scratch or based on conditions either in a single type or multi-modal ones. We mainly consider modalities, including stroke, exemplar, and text, which are the common interactive ways for the colorization task. To unify different modalities, we adopt a two-stage framework by taking hint points as an intermediate representation. A hint point is a point with conditional colors, where the minimum size is a pixel. This is because hint points can be naturally decomposed or extracted from different modalities. For the stroke-based condition, the hint points can be sampled along with strokes. For the exemplar-based condition, a colorful exemplar is warped to the input grayscale image based on semantic matching, and then hint points with high matching confidence can be selected. For the text-based condition, we introduce a new method based on CLIP embedding [Radford et al. 2021] for assigning hint points on the objects corresponding to the input text. We thus convert all modalities into a unified representation (*i.e.,* hint points) in the first stage. Then the model can concentrate on learning colorization in the second stage.

To ensure the diversity and quality of generated results, in the second stage, we take advantage of Transformer architecture [Vaswani et al. 2017], which has shown great success on various generation tasks [Esser et al. 2021; Ramesh et al. 2021; Wu et al. 2021a; Zhang et al. 2021]. Given a grayscale image, together with hint points, we design a Transformer-based network for diverse colorization. More specifically, we first introduce *Chroma-VQGAN*, a subnetwork composed of two separated gray and color encoders and one joint decoder, to disentangle chroma representation from the gray one. The chroma representation is discretized through a learned codebook, while gray representation remains continuous features for keeping input details. We then propose a *Hybrid-Transformer* for predicting chrominance values. Different from previous works [Esser et al. 2021; Wu et al. 2021a; Zhang et al. 2021] taking all the input and conditions as discrete tokens, our Hybrid-Transformer is created for mix-type inputs, including hint points in pure colors, continuous gray representation, and quantized chroma representation, which avoids quantization loss on the gray input and hint points. We follow a BERT-style [Devlin et al. 2019] training scheme for incorporating color hints in any position.

The extensive experiments on both unconditional and various conditional colorization tasks demonstrate the effectiveness of our method for generating diverse results with high-quality. We also design a UI tool for interactive colorization under our unified framework. Please see more details in the supplementary video demo. In summary, we make the following contributions:

- We propose the first unified framework (UniColor) for interactive colorization, allowing multi-modal conditions in hybrid-mode.
- We present a method for unifying multi-modal conditions, including stroke, exemplar, and text, by taking hint points as an intermediate representation. Especially, we introduce a novel CLIP-based method for text-to-hint-point conversion.
- We propose a colorization network composed of Chroma-VQGAN and Hybrid-Transformer for generating diverse and high-quality colorization results both conditionally and unconditionally.
- We design an interface, showing the effectiveness of our unified framework in practical usage.

2 RELATED WORKS

2.1 Unconditional Colorization

Deep neural network has been proven to be successful in image colorization, where the model is learned from hundreds of thousands of gray-color image pairs. An earlier work [Zhang et al. 2016] treats colorization as a classification task, and proposes an end-to-end network for learning the mapping from the grayscale images to the distribution of quantized chrominance values. Global semantic information is further added to guide the colorization [Iizuka et al. 2016], where the model jointly learns to colorize and predict class labels.With the development of generative adversarial network (GAN) [Goodfellow et al. 2014], it has been used in colorization tasks [Isola et al. 2017; Kiani et al. 2020; Vitoria et al. 2020]. To enhance the quality of colorization on images with multiple objects, instance-aware colorization is introduced [Su et al. 2020] by extracting object-level features.

Diverse colorization. As colorization is an ill-posed problem with multiple possible solutions, it is essential to generate diverse colorization results instead of a deterministic one. Cao et al. [2017] initialize the color channels with random noise and train the generator to produce diverse colors from the noise with adversarial loss. Deshpande et al. [2017] model the VAE-encoded color embeddings as a mixture of Gaussian distributions conditioned on the gray image, and sample diverse colors from the Gaussian distributions. A Gaussian conditional random field layer is further applied to the output color distribution [Messaoud et al. 2018], to enhance global structural consistency and enable controllability in form of sparse color points. Wu et al. [2021b] propose a framework for diverse colorization results, with guidance from the diverse color priors generated by a pretrained GAN. Coltran [Kumar et al. 2021] models the probability of color distribution with Transformer and samples diverse colors autoregressively. Recently, Palette [Saharia et al. 2021]

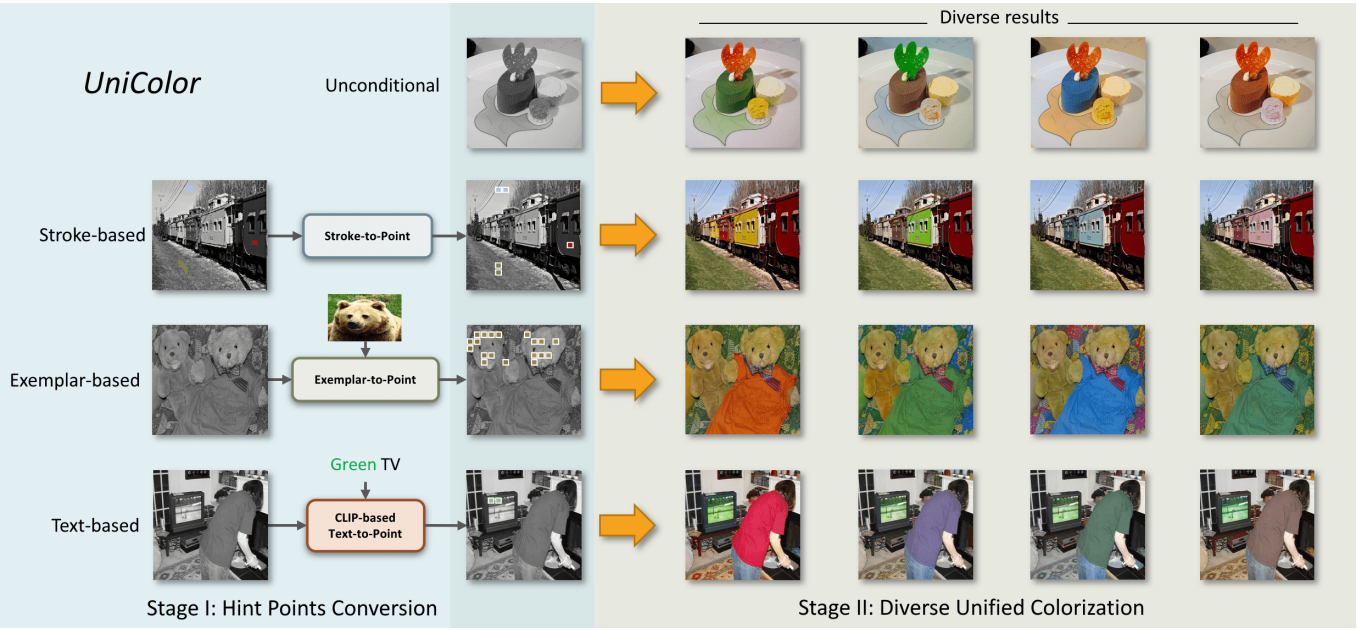

Fig. 2. Our unified colorization pipeline. The pipeline consists of two stages. In the first stage, all different conditions are unified as hint points. In the second stage, diverse results are generated automatically either from scratch or based on the condition of hint points. Input images: the 1^{st} row is from ImageNet and all others are from MSCOCO.

trains a diffusion model [Ho et al. 2020a] to denoise the color images from the Gaussian noises, conditioned on the gray images, for diverse colorization.

2.2 Conditional Colorization

To allow user controls, some conditional colorization methods were proposed. According to the control modality, they can be categorized into: stroke-based colorization, exemplar-based colorization, and text-based colorization. However, there is no existing method unifying multi-modal controls.

Stroke-based colorization. Stroke-based colorization requires users to draw color strokes as the condition. Earlier works [Levin et al. 2004; Luan et al. 2007] rely on optimization methods for propagating the user input strokes to the whole image, based on intensity and spatial position. The later work [Endo et al. 2016] learns the similarity mapping explicitly through a neural network by generating a probability map. Recently, rather than guiding with explicit similarity metrics, several methods [Xiao et al. 2019; Zhang et al. 2017] propose to conduct end-to-end training for direct propagation.

Exemplar-based colorization. Exemplar-based colorization aims to colorize a grayscale image based on a user input reference image. The colorization pipeline usually consists of two steps. The first step warps the reference image to the grayscale input, based on either the deep similarity metrics [He et al. 2018; Zhang et al. 2019] or transfer color from the reference image through a learned network [Xu et al. 2020]. The second step is to generate the final colorization based on the warped reference. The performance of exemplar-based colorization largely relies on the fully warped or transferred reference, and it fails to generate feasible results in the regions with incorrect correspondences.

Text-based colorization. Text-based colorization generates results according to the user-input text description, which often contains an object with a color word. Manjunatha et al. [2018] encodes the input text by an LSTM and fuses it with the visual feature of the input grayscale image through feature-wise affine transformations. Chen et al. [2018] introduces a recurrent attentive model to fuse the text feature with the image feature, but limits images to specific classes, such as flowers and geometry shapes. Bahng et al. [2018] generate color palette from text to control the global color distribution. Different from stroke-based and exemplar-based methods, text-based colorization is still in an early stage and needs further exploration.

2.3 Transformer in Image Generation and colorization

Transformer architecture [Vaswani et al. 2017] is originally introduced for natural language processing, which models long-range relations between the input tokens through attention [Bahdanau et al. 2015; Kim et al. 2017]. It is then extended to the image domain by treating each image pixel as a visual word [Chen et al. 2020]. However, this direct usage cannot deal with high-resolution images as the computational cost increases quadratically with the number of tokens increasing. Later variants are introduced to improve the efficiency including: 1) restricting attention to local fields [Parmar et al. 2018; Weissenborn et al. 2020]; 2) replacing the full attention with successive partial attentions [Child et al. 2019; Ho et al. 2020b]; and 3) reducing the length of input tokens [Dosovitskiy et al. 2021; Esser et al. 2021]. A two-stage framework for high-resolution image

synthesis [Esser et al. 2021] is proposed by using a convolutional neural network (CNN) based VQGAN to encode and tokenize the input image. This method not only reduces the length of input tokens but also incorporates the advantages of CNN (*e.g.,* local interactions and inductive bias).

Colorization with Transformers. Transformer architecture is first introduced to image colorization in Coltran [Kumar et al. 2021]. The colorization process is divided into three coarse-to-fine modules including a conditional autoregressive (AR) transformer, a color upsampler, and a spatial upsampler. The work regards each image pixel as a token and uses axial attention [Ho et al. 2020b] to enable image colorization in a resolution of 256×256, but hard to be applied to higher resolution. The encoded feature of the grayscale image is used to modulate the layers of the transformer, which generates diverse colorized images under multinomial sampling. However, this method does not accept any kinds of controls and only supports unconditional colorization. Additionally, with pure Transformer architecture, the method cannot take advantage of the local interactions and inductive bias from CNN.

3 UNICOLOR

We aim to design a unified colorization framework producing diverse results unconditionally or conditioned on a mixed set of multi-modal inputs, including stroke, exemplar, and text. To implement such a framework, we mainly face two challenges.

First, conditions under different modalities have various representations, which require different network architecture. For example, an exemplar image is usually warped by a correspondence network [Zhang et al. 2019] while the text needs to be encoded with the LSTM [Manjunatha et al. 2018]. Therefore, multi-modal conditions are difficult to be incorporated into a single network. To deal with multi-modal conditions, we introduce hint points as an intermediate unified representation for all three types of conditions (*i.e.,* stroke, exemplar, and text). As shown in Fig. 2 (stage 1), all modalities are converted into the form of hint points, which are then used to guide and control the colorization process. This design also enhances the generalization of our framework since this representation can be easily extended to process new modalities.

Second, as a one-to-many mapping problem, the framework should generate colorization in diverse results, which are both aesthetically pleasing and semantically valid. To achieve this goal, we propose a network combining both VQGAN and Transformer architectures, which have shown promising results on general image generation and editing tasks [Esser et al. 2021; Wu et al. 2021a; Zhang et al. 2021]. There are two benefits: 1) by quantizing the image into a codebook through VQGAN, the method can be converted into a classification formulation for sampling different results based on probabilities, and thus increases the diversity; 2) by building correlations among tokens in Transformer, the model learns global consistency for ensuring the quality of results. While naive usage of VQGAN and Transformer generates serious artifacts, we introduce novel Chroma-VQGAN and Hybrid-Transformer for our unique unified colorization problem.

An overview of our framework is shown in Fig. 2, which consists of two stages. The first stage called *Hint Points Conversion* aims to convert multi-modal conditions (*i.e.,* stroke s, reference image I_r, and text t) into a uniform hint point representation h_c. The second stage, called *Diverse Unified Colorization*, aims to generate diverse high-quality colorization results from the input grayscale image $I_g \in \mathbb{R}^{H\times W}$, conditioned on a mixed set of conditions $\mathbb{P}(\{s, I_r, t\})$ ($\mathbb{P}$ is the power set):

$$\{\hat{I}_c^i\} \sim P(\hat{I}_c | I_g, \mathbb{H}(\mathbb{P}(\{s, I_r, t\}))) \tag{1}$$

where $\mathbb{H}$ is the Hint Points Conversion and $\hat{I}_c^i$ is the i-th image, sampled from the multinomial probability distribution P, in the set of diverse colorization results $\{\hat{I}_c^i\}$.

3.1 Hint Points Conversion

Hint points, a set of points assigned with target colors, as shown in Fig. 2, is an accurate and flexible way of representing the color condition. Different modalities commonly used in the colorization task (*i.e.,* stroke [Xiao et al. 2019; Zhang et al. 2017], exemplar [He et al. 2018; Xu et al. 2020; Zhang et al. 2019], and text [Manjunatha et al. 2018]) can be converted to hint points by different modules easily and then be used in a mixed way for practical application. For fast processing, we divide an image into a grid of cells with size $d \times d$. So a hint point corresponds to one cell in a single color.

Stroke to Hint Points Conversion. Given user-drawn strokes on a grayscale image, we traverse the cells along the stroke, and regard the cell as a hint point if the number of colored pixels within a cell surpasses a threshold (*e.g.,* $0.75d$), then we assign the color of this hint point as the stroke's color. After repeating this process for all the strokes, the grayscale image with hint points will be sent into the second stage as shown in Fig. 2.

Exemplar to Hint Points Conversion. The exemplar used in colorization is often an image sharing similar semantic content with the target one. To convert the exemplar image into hint points, we are inspired by the conventional way of warping the exemplar image to the grayscale image with semantic matching [He et al. 2018; Zhang et al. 2019]. But different from previous works warping all the pixels equally, we only keep the $d\times d$ cells with high matching confidences as hint points. This is to avoid the injected noises caused by mismatches and allow more diverse sampling on final results. The confidences are measured from the output correlation matrix of the correspondence network in [Zhang et al. 2019], which is based on pretrained VGG19 [Simonyan and Zisserman 2015]. More specifically, we keep a hint point if the average similarity of its corresponding cell in the warped image is larger than a threshold (*i.e.,* we set to 0.23 empirically), and assign the color to the hint point as the mean color of the corresponding cell.

Text to Hint Points Conversion. Text-based image generation and manipulation tasks have been actively studied recently as the development of natural language processing techniques. It is extended to colorization tasks but is still under-explored without a mature and conventional method. In view of this, we propose a novel method based on the Contrastive Language-Image Pre-training (CLIP) embedding [Radford et al. 2021], which has shown to be a powerful text-image representation. Note that compared with previous referring object segmentation methods [Luo et al. 2020; Wang et al. 2022], our CLIP-based method is training-free without the need for

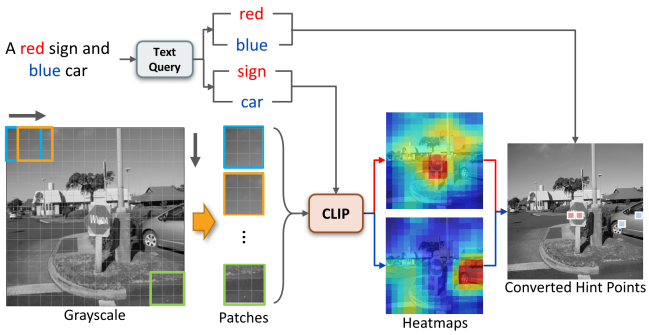

Fig. 3. Text-to-hint-points conversion. For each object concept, we calculate a correspondence map by measuring the similarity between cells and the textual concept through CLIP embedding, and the cells with top-2 correspondence values are selected as hint points. Input image: from MSCOCO.

ground-truth annotations (*e.g.*, RefCOCO [Kazemzadeh et al. 2014]) and can naturally deal with objects in open vocabulary. Given a text containing the described objects with colors, we first extract the object and color concepts based on the text parsing. We then divide the grayscale image into grid with a cell size of d, and slide an $n \times n$ window (*i.e.*, we set $n = 3$) with stride 1 for extracting patches across the image. At each location, we calculate the correspondence value between the patch of $n \times n$ cells and each of the object concepts by cosine similarity between the features from the CLIP embedding, as shown in Fig. 3, and this correspondence value contributes to all the $n \times n$ cells within a patch. For each of the object concepts, after averaging all values within a cell, we obtain a correspondence map, and regard the top-2 cells as hint points. The specified color of the corresponding object in the given text is further assigned to the hint points, according to a color table.

3.2 Diverse Unified Colorization

Given color hints and a grayscale image, the second stage of UniColor aims to propagate the hint colors for generating diverse colorization results. We thus propose a Transformer-based architecture. As shown in Fig. 4, it contains two sub-networks, one is the Chroma-VQGAN for learning a disentangled and quantized chroma representation from the continuous gray one, and the other is the Hybrid-Transformer for learning colorization based on unified condition.

3.2.1 Chroma-VQGAN. Before introducing our Chroma-VQGAN, we first revisit the typical VQGAN [Esser et al. 2021]. VQGAN encodes color images and quantizes them into a discrete codebook. Each image can be represented by a spatial collection of codebook entries. VQGAN learns the codebook through a reconstruction task by decoding such spatial collection with a decoder. Since detailed structure information is lost during the quantization process, the reconstruction result can contain serious distortions and artifacts (will be validated in Sec. 5.3.1). However, this is not acceptable in the colorization task, as the structure should be aligned well with the grayscale input. Therefore, to adapt VQGAN to our task, we create a variant called Chroma-VQGAN for encoding gray and color features separately to preserve the detailed structure with continuous gray but quantized chrominance representations.

Given a color image I_c with its gray version I_g as inputs, Chroma-VQGAN takes two actions to mitigate the distortion artifacts and improve the reconstruction quality. First, we introduce a side branch, *i.e.*, gray encoder shown in Fig. 4, for extracting features from the gray input. Second, to avoid the information loss during quantization, we directly fuse the features of gray input without quantizing for sending it into the decoder to reconstruct the input color image. More formally, we first obtain color features f_c from the input color image I_c through the color encoder, while obtaining gray features f_g from the input gray image I_g through the gray encoder. To incorporate with the hint points, both features are down-sampled by the encoders with a factor of d, which is same as the cell size of the hint points. We then tokenize the color features f_c into x_c through a learnable codebook $\mathcal{Z} = \{z_k\}_{k=0}^{N-1}$, but remain the gray features as the continuous one. The tokenized $x_c \in \mathbb{R}^{H/d \times W/d}$ contains the indices of the entries from the learned codebook [Esser et al. 2021] and the index of i^{th} row and j^{th} column is obtained by:

$$x_c^{ij} = T(f_c^{ij}) = \underset{k \in [0,N-1]}{\arg\min} \|f_c^{ij} - z_k\|, \tag{2}$$

where $T(.)$ is a tokenization operation.

Before combing with the continuous gray features f_g, the color indices x_c are detokenized by T^{-1} for restoring the color features as:

$$\hat{f}_c = T^{-1}(x_c) = \{z_k \in \mathcal{Z}, k = x_c\}. \tag{3}$$

The detokenized features $\hat{f}_c$ are then concatenated with the gray features f_g along the channel dimension. After decoding through a decoder, we obtain the reconstructed color image I_{rec}. The Chroma-VQGAN is trained in an adversarial way by learning accurate and perceptually rich features with the help of a discriminator, following the strategy in the previous work [Esser et al. 2021].

Unlike the previous work [Esser et al. 2021] training separate VQGANs for quantizing both the condition and the image, our Chroma-VQGAN uses a single VQGAN with two encoder branches and the gray features are kept unquantized, as shown in Fig. 4. This design enables high-quality reconstructions, where the structure and content are well preserved from the input images. Another benefit of our Chroma-VQGAN is that it can disentangle the chrominance features in f_c from the gray one f_g, because of the little information loss from gray image encoding (will be validated in Sec. 5.3.1). This allows the Hybrid-Transformer introduced next to focus on the chrominance prediction, achieving better final results.

3.2.2 BERT-Style Hybrid-Transformer. In this subsection, we introduce how we generate diverse colorization results from the unified hint points condition and grayscale input. We take a BERT-style scheme [Devlin et al. 2019] for training our Transformer. During the training, the model needs to fulfill a color completion task. That is, we randomly mask out a portion of input color tokens, and ask the model to restore these tokens based on gray image, hint points, and the unmasked color tokens. Different from the traditional Transformer only relying on discrete color tokens as the inputs, our model has a hybrid input.

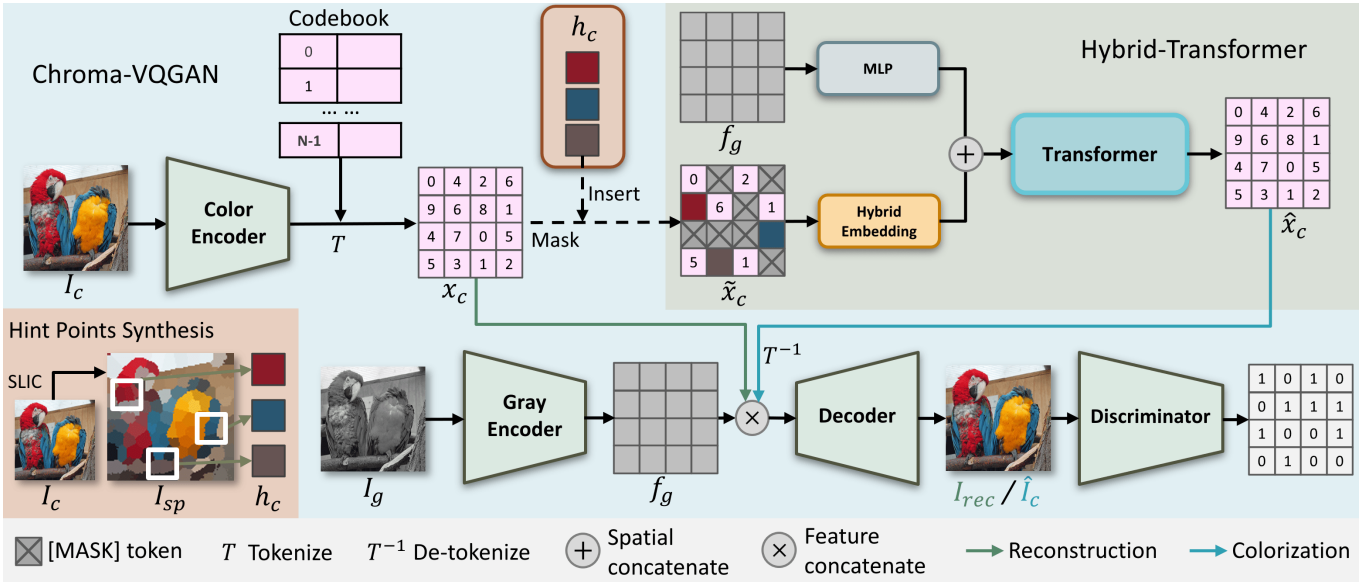

Fig. 4. Architecture of diverse unified colorization network. The network consists of two sub-nets: (a) a Chroma-VQGAN to disentangle and quantize chroma representation from the continuous gray one, and (b) a Hybrid-Transformer to generate diverse colorization results from unified conditions and continuous gray features. Input image: from ImageNet.

The hybrid inputs to our Transformer are extracted continuous gray features f_g, masked color tokens $\tilde{x}_c$, and hint points h_c, where $\tilde{x}_c = x_c^M \cup x_c^{\overline{M}}$, and M indicates the indices of the tokens masked out. For each masked token, we replace the original codebook index with a special token $[MASK]$. Since our model should also support unconditional colorization without hint points during the inference time, we first introduce a simplified formulation, and then discuss how the hint points are synthesized with an adapted formulation. Conditioned on the unmasked color tokens $x_c^{\overline{M}}$, and the gray features f_g, our Hybrid-Transformer is trained to learn the likelihood of the indices at the masked positions for restoring x_c:

$$P(x_c^M | x_c^{\overline{M}}, f_g) = \prod_{i \in M} P(x_c^i | x_c^{\overline{M}}, f_g). \tag{4}$$

Then the Transformer is trained to minimize the softmax cross-entropy loss between the output probabilities and the ground-truth color indices.

Hint Points Injection. Rather than quantizing hint points as tokens to the Transformer, we directly use the continuous color values to keep the accurate color of the hint points. To better fuse the hint points h_c with the color tokens x_c, we create a hybrid embedding space by mapping color tokens through an embedding layer for generating embedding color features e_{x_c}. We learn a mapping function from the color of hint points h_c to the feature space f_h, sharing the same channel dimension d_e as the embedded features e_{x_c}:

$$f_h = W_h h_c + p_h, \tag{5}$$

where $W_h \in \mathbb{R}^{3 \times d_e}$ is the learnable weights and p_h is a learnable positional embedding to distinguish the positions of hint points from color tokens. Note that we only insert the hint points at the masked positions M during the training time. Given the mapped features of the hint points f_h as another conditional input, the conditional probability of the Hybrid-Transformer in Eqn. (4) can be rewritten as:

$$P(x_c^M | x_c^{\overline{M}}, f_g, f_h) = \prod_{i \in M} P(x_c^i | x_c^{\overline{M}}, f_g, f_h). \tag{6}$$

Hint Points Synthesis. Collecting a large scale annotated data covering all kinds of multi-modal conditions, and converting them into unified hint points are impractical, we thus create a novel method for synthesizing hint points during the training process. The key idea is to randomly sample grid cells from the original color image, and extract the hint point from each of these cells. For each selected cell, the color of the hint point should be sampled from the $d \times d$ image cell. A simple way is to take the mean color of the whole cell as the hint color, but we find this naive solution is problematic, as a single cell may cover multiple color regions and boundaries. As shown in Fig. 4, the cells of I_{sp} in white frames may cover multiple colors, *e.g.*, orange and blue, and the mean color cannot represent either one.

To avoid such vague and inaccurate color assignment, we propose to use the dominant color of each cell. We first compute the superpixel segmentation from the color image I_c through Simple Linear Iterative Clustering (SLIC) Algorithm [Achanta et al. 2012]. Then we get the superpixel image I_{sp} by representing each superpixel with the mean color of the segment. During training, we randomly select a grid cell from the superpixel images, and take the dominant color value within that cell as the color of the hint point. To let the Hybrid-Transformer also deal with unconditional colorization in the inference time, there is a chance (*i.e.*, 30%) that no hint point is added at the training time.

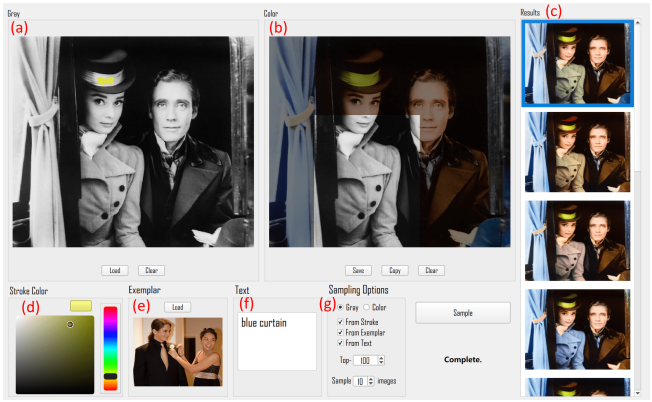

Fig. 5. Our interface of interactive multi-modal image colorization. (a) displays the input grayscale image; (c) shows all the diverse colorization results; (b) is the user-selected colorized image from (c), which can be further re-colorized and iteratively edited; (d) is the color picker of the stroke that can be drawn in (a) or (b) for stroke-based colorization; (e) is the input reference image for exemplar-based colorization; (f) is the input text description for text-based colorization; (g) is the panel for input and modality options. Input image: Audrey Hepburn and Mel Ferrer, while filming *War and Peace*, 1955. Reference image: from MSCOCO.

3.2.3 Inference. During the inference time, the model starts with the gray features f_g, hint points h_c and color tokens $\tilde{x}_c$ all filled with $[MASK]$ tokens. Then the predicted color tokens $\hat{x}_c$ are sampled autoregressively (AR) in raster scan order with Hybrid-Transformer:

$$p(\hat{x}_c|f_g, f_h) = \prod_i p(\hat{x}_c^i|\hat{x}_c^{<i}, f_g, f_h), \tag{7}$$

where the probability of the current token is conditioned on all previously sampled tokens, the gray features, and the hint points. For each generated color token, we apply multinomial sampling within the top-k (*i.e.,* k is set to 100) indices under the predicted probability distribution. We keep the hint point tokens fixed during the sampling process. After sampling all the tokens within $\hat{x}_c$, we detokenize the tokens to continuous features before concatenating them with the gray features along the channel dimension. Finally, the concatenated features are fed into the decoder of our Chroma-VQGAN to obtain the colorization image result $\hat{I}_c$.

3.3 Interactive Interface

Driven by our proposed UniColor framework, we design an interactive tool for multi-modal colorization. Fig. 5 shows one screenshot of the user interface of our tool. Our tool supports various types of image colorization, including unconditional, stroke-based, exemplar-based, text-based, and hybrid colorization. In the hybrid colorization mode, the user can choose to add conditions in a combined way. Our tool has four main components: 1) a canvas for showing the grayscale input image and drawing strokes (a), 2) a panel showing all the diverse colorization results (c), 3) a canvas for re-colorizing the color image (b), and 4) an interface for inputting various types of conditions (d,e,f) and a panel for input and modality selection (g).

Unconditional Colorization. For unconditional colorization, the user just needs to input a gray image in (a) and uncheck all the conditions in (g) to let the system colorize the image automatically. The diverse results will be displayed in (c).

Multi-modal Colorization. Our system allows the user to select one modality for colorization from the panel (g) and then specify the conditions in different forms. For stroke-based colorization, the user can select colors in (d) and draw color strokes onto the grayscale image in (a); for exemplar-based colorization, the user can import the reference color image in (e); for the text-based colorization, the user can type in the text description of objects and colors in the textbox (f). Different types of conditions will be converted into hint points to guide the colorization process.

Hybrid Controls. Our system also supports hybrid controls. The user can input more than one type of condition and check multiple modalities in (g), while the system will mix all the hint points generated from the selected conditions for colorization. To avoid conflict, we define the default priority of hybrid controls as: stroke, text, and exemplar, *e.g.,* if stroke and exemplar conditions generate hint points on the same location, the points generated from the exemplar will be ignored, and only the points from the stroke will be considered.

Re-colorizing & Iterative Editing. After colorization, the system will show the diverse colorization results in (c), and the user may select one to be displayed in (b). If the user wants to edit the selected result further, he or she could select an image subregion to re-colorize. Multi-modal conditions can also be applied to the subregion to reflect the user's intents. And the user may select and re-colorize subregions iteratively to interactively refine the result. Additionally, if a color image is imported, the system will also show the color version of the image in (b) so that the user could directly re-colorize an original color image in the same manner.

4 IMPLEMENTATION DETAILS

4.1 Network Architectures

In this subsection, we introduce the detailed network architectures of both Chroma-VQGAN and Hybrid-Transformer.

Chroma-VQGAN. We mainly follow the implementations of the previous work [Esser et al. 2021]. For the color encoder, we increase the number of channels from 256 to 512. For the quantization module, we set the size of the codebook as $N = 4096$. For the gray encoder, we use the same structure as the color encoder, except changing the input channel to 1. The input color images and grayscale images are down-sampled by a factor of $d = 16$. After concatenating the gray and color features, the number of input channels to the decoder is 1024.

Hybrid-Transformer. We use a similar structure as iGPT [Chen et al. 2020] with learnable positional encoding. The Transformer consists of 24 multi-head self-attention (MHSA) [Vaswani et al. 2017] layers with 16 heads. The color tokens are embedded into features of channel dimension $d_e = 512$. The gray features and color of hint points are passed through two separate linear layers, respectively, both with output dimensions of 512. After that, the gray features $f_g \in \mathbb{R}^{16\times16\times512}$ and the embedded color features $e_{x_c} \in \mathbb{R}^{16\times16\times512}$ are flattened and concatenated in spatial dimension to form the input of shape 512×512 (*i.e.,* $512 = 2 \times 16^2$). We compute the attentions among all tokens without adding any mask. The final

dimension of the predicted probability is the same as the length of the codebook (*i.e.,* 4096).

4.2 Training Strategy

Training Dataset. We train both the Chroma-VQGAN and Hybrid-Transformer on the ImageNet ILSVRC2012 [Russakovsky et al. 2015] dataset, with around 12M training images. During training, the images are first resized to 294×294 and then randomly cropped to 256×256 followed by random horizontal flipping.

Chroma-VQGAN. The Chroma-VQGAN is trained on 4 Nvidia V100 GPUs with a total batch size of 32 for 260,000 steps (around 60 hours). The learning rate is set to 5.12×10^{-5} (base learning rate $1.6 \times 10^{-6} \times$ batch size 32) throughout the whole training stage with no warm-up or decay. For stability, we begin to update the parameters of the discriminator after 6000 steps.

Hybrid-Transformer. The hybrid-Transformer is trained in BERT-style [Devlin et al. 2019], where 16 to 256 randomly selected input color tokens are masked and replaced with a learnable [*MASK*] token. We also apply an additional probability of 5% to mask all the 256 color tokens to ensure that the model can predict the colors from scratch. Among the masked tokens, we randomly select 1 to 16 positions to insert hint points, which are generated from the superpixel images as described in Sec. 3.2.2. To ensure the transformer is capable of predicting color tokens without hint points, we only insert the hint points with a probability of 70%. All the numbers and the positions of the masking and the hint points are sampled from the uniform distribution. The hybrid-Transformer is trained on 4 Nvidia V100 GPUs with an accumulated batch size of $2 \times 4 \times 16 = 128$ for 142,000 steps (around 46 hours). The learning rate is set to 2.05×10^{-4} (base learning rate $1.6 \times 10^{-6} \times$ batch size 128) and decayed by 0.1 at 10^{th} epoch.

4.3 Inference & Interaction Speed

Our model takes an average of 4.6 seconds to colorize a 256×256 image with a single Nvidia V100 GPU, which is the common speed of an autoregressive Transformer. For text-based colorization, it will take an additional 1.8 seconds to convert the text to hint points. For the interactive system, it normally takes the user 10-15 seconds to input every single modality (*i.e.,* stroke, exemplar, and text).

5 EXPERIMENTS

In this section, we first compare our unified framework with previous works in all four types of colorization (unconditional, stroke-based, exemplar-based, and text-based) respectively (Sec. 5.1). We then conduct the user study and show the results in Sec. 5.2. Lastly, we perform ablation studies in Sec. 5.3.

Test Data. We select testing images from two datasets: ImageNet ILSVRC2012 [Russakovsky et al. 2015] and MSCOCO 2017 [Lin et al. 2014]. For ImageNet, we randomly select 5,000 images from the 50,000 images in the validation set, where 5 images are drawn from each class. For MSCOCO, we use all the 5,000 images from the validation set with ground-truth text caption and segmentation.

Table 1. Comparison with unconditional colorization methods on both ImageNet and MSCOCO datasets. The models are trained on ImageNet if not specified (* Trained on MSCOCO). Metrics calculated on original images are taken as references.

	ImageNet		MSCOCO	
	FID↓	Colorful↑	FID↓	Colorful↑
Original	—	38.00	—	37.46
CIC [2016]	21.31	34.25	32.62	34.36
User-guided (auto) [2017]	12.53	26.16	19.18	27.04
Deoldify [2018]	9.59	21.39	12.29	22.84
InstColor* [2020]	12.74	26.00	12.72	27.26
ChromaGAN [2020]	16.27	26.92	25.50	27.08
GenPrior [2021b]	9.57	35.29	—	—
Coltran [2021]	12.31	36.59	14.20	36.31
Ours	**9.46**	**39.01**	**11.16**	**39.11**

5.1 Comparisons on Multi-modal Colorization

As a unified multi-modal colorization framework, we need to examine the performance on different condition modalities, and we first compare with previous state-of-the-art (SOTA) methods on: unconditional, stroke-based, exemplar-based, and text-based colorization, respectively. We show both quantitative and qualitative results for each of the conditions.

Evaluation Metrics. To measure the overall quality and fidelity of the generated images, we calculate **FID** [Heusel et al. 2017] between the generated and ground-truth images. We also calculate **Colorfulness** [Hasler and Suesstrunk 2003] to measure how colorful and vivid are the colorized images. For stroke-based colorization, we further calculate **LPIPS** [Zhang et al. 2018] to measure the perceptual similarity between the colorized and original images, because the generated images are supposed to be close to the original one, from which the hint points are sampled. For exemplar-based colorization, we introduce **Contextual Loss** [Mechrez et al. 2018] with pixel-wise L2-loss to measure the similarity between the non-aligned images, *i.e.,* the colorized images and reference images. For text-based colorization, we use **CLIP score** [Radford et al. 2021] to measure the relevance between the colorized images and the text prompts. All the metrics are calculated with images of size 256×256, except 96×96 for Contextual Loss.

Unconditional Colorization. For unconditional colorization, we compare with three types of SOTA methods that can colorize grayscale images without user hints: a) CNN-based methods: CIC [Zhang et al. 2016], User-guided (auto) [Zhang et al. 2017], InstColor [Su et al. 2020], Deoldify (software) [Deoldify 2018], ChromaGAN [Vitoria et al. 2020] and GenPrior [Wu et al. 2021b]; b) Transformer-based method: Coltran [Kumar et al. 2021]; and c) Diffusion-based model: Palette [Saharia et al. 2021].

Quantitatively, as shown in Tab. 1, our method outperforms CNN-based and Transformer-based methods on both datasets in terms of FID and colorfulness. Generally, Transformer-based methods

Fig. 6. Comparison with unconditional colorization methods: InstColor [Su et al. 2020], ChromaGAN [Vitoria et al. 2020], and Coltran [Kumar et al. 2021]. Our model can generate diverse results for each of the input grayscale images. Input images: the 1^{st} and 4^{th} rows are from ImageNet; others are from MSCOCO.

(Coltran and ours) with multinomial sampling produce more vivid and colorful colors than CNN-based methods, except that CIC encourages rare colors in the loss function (also observed in [Wu et al. 2021b; Zhang et al. 2019]) and GenPrior uses auto-generated reference images. We show the qualitative comparison in Fig. 6. [1] Our method can generate diverse and vivid colors, with multinomial sampling from Hybrid-Transformer (*e.g.,* diverse colors of the bus in 2^{nd} row and the pillows in 3^{rd} row). Thanks to the global attention module, different from CNN-based methods, our Transformer-based framework generates consistent color across distant pixels sharing the same semantics (*e.g.,* the flowers in 1^{st} row and the sofa in 3^{rd} row). Compared with Coltran, the existence of CNN-based Chroma-VQGAN makes our model more sensitive to local contours (*e.g.,* the dog and sofa in 3^{rd} row) and details (*e.g.,* the books in 4^{th} row).

As the code of Palette [Saharia et al. 2021] is not available, we only compute the FID score following the evaluation protocol used in both Coltran and Palette on ImageNet validation set. We obtain the FID scores as: 19.37 (Coltran), 15.78 (Palette), and 16.80 (Ours). Under this protocol, our unconditional method outperforms Coltran and is comparable to Palette. For qualitative comparison, we demonstrate the results on the images shown in the original Palette paper. As can be seen in Fig. 7, both Palette and our method can produce diverse

[1]Because of the limited space, we only choose two representatives from CNN-based methods for qualitative comparison and leave the full results in the supplementary document.

Fig. 7. Comparison with diffusion-based model Palette [Saharia et al. 2021]. Our method generates diverse results comparable to Palette. Input images: from ImageNet.

colorization results with high fidelity and vivid colors. Besides, our method can further support multi-modal control.

Stroke-based Colorization. For stroke-based colorization, we compare with User-guided [Zhang et al. 2017], a recent SOTA work. To be fair, instead of generating the hint points from the superpixel images as specified in Sec. 3.2.2, we adopt the same method as the

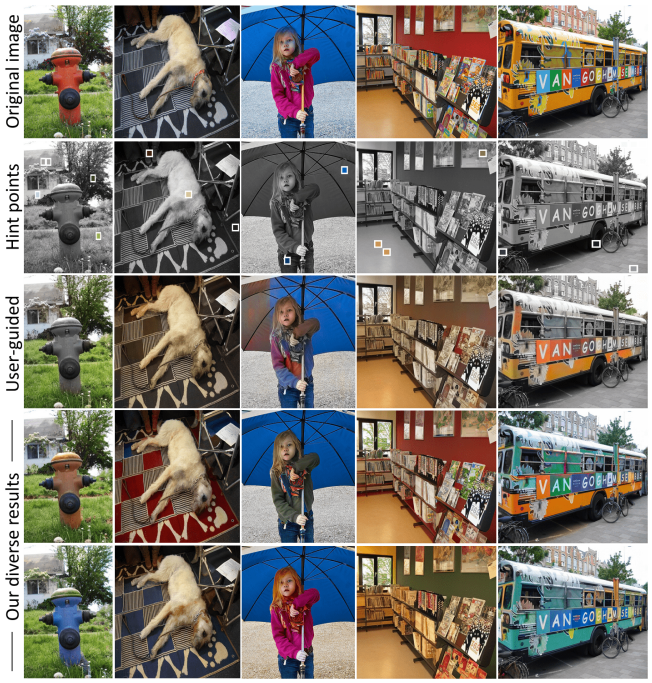

Fig. 8. Comparison with the stroke-based method (User-guided [Zhang et al. 2017]). Input images: the 2^{nd} and 4^{th} columns are from ImageNet; others are from MSCOCO.

User-guided one [Zhang et al. 2017] to assign the mean color of a cell to the hint point. To avoid selecting hint points on the intersections of two objects, we first perform clustering and segmentation on the images with mean-shift and then randomly select 2 to 16 cells from the large segments as the hint points.

We show the quantitative results in Tab. 2. To verify whether our model follows the input condition, we also compare with our unconditional variant. As can be seen, the lower FID and LPIPS indicate that our model propagates the hint points derived from strokes properly. Compared with the previous stroke-based method [Zhang et al. 2017], ours achieves better FID and colorfulness, and comparable LPIPS. As shown in Fig. 8, our method propagates the stroke color (hint points) into the whole object smoothly and consistently (*e.g.,* the blue umbrella in 3^{rd} column), while User-guided fails to spread the hint colors to the whole object. For the region without specified hint points, our method can also generate diverse and vivid colors (*e.g.,* the books and wall in 4^{th} column and the bus in 5^{th} column).

Exemplar-based Colorization. For exemplar-based colorization, we compare with Deep Exp. [He et al. 2018] and Exp. Video [Zhang et al. 2019]. We only test on ImageNet because similar reference images could hardly be obtained for images in MSCOCO. We use the retrieval method based on gray images in [He et al. 2018] to obtain the reference images from the ImageNet training set, which are closest to the input grayscale images.

Table 2. Comparison with the stroke-based colorization method - User-guided [Zhang et al. 2017] on both ImageNet and MSCOCO dataset. The unconditional variant (*i.e.,* Uncond.) helps on verifying the effectiveness of input conditions.

	ImageNet			MSCOCO		
	FID↓	LPIPS↓	Colorful↑	FID↓	LPIPS↓	Colorful↑
User-guided	9.76	0.1144	32.48	14.68	**0.1166**	31.44
Ours (Uncond.)	9.46	0.1945	**39.01**	11.16	0.1909	**39.11**
Ours (Stk.)	**7.04**	**0.1119**	36.16	**8.89**	0.1189	35.71

Table 3. Comparison with the exemplar-based methods on ImageNet.

	FID↓	Contextual↓	Colorful↑
Deep Exp. [2018]	10.79	**1.75**	29.64
Exp. Video [2019]	10.70	1.90	25.92
Ours (Uncond.)	9.46	2.65	**39.01**
Ours (Exp.)	**7.39**	1.82	38.80

Table 4. Quantitative results of text-based colorization on MSCOCO.

	CLIP similarity↑	FID↓
Original	24.05	—
Ours (Uncond.)	23.55	**11.16**
Ours (Text)	**24.50**	11.29

As shown in Tab. 3, our exemplar-based method obtains lower FID and contextual loss than our unconditional variant, which demonstrates the effectiveness of the reference images. Compared with other exemplar-based methods, ours achieves the best FID and colorfulness scores. In the aspect of contextual loss, ours performs a bit worse than Deep Exp. [He et al. 2018], due to the fact that our method selectively inherits the colors from the warped image with high confidence, as stated in Sec. 3.1. This mechanism makes our method more robust when the warped image is unreliable. As shown in 3^{rd} column of Fig. 9, the color of the girl's arm is not misguided by the wrongly warped blue. When the warped image of the sugar store in 2^{nd} column and the bed in 5^{th} column are noisy, other methods generate results of low colorfulness, while our method still generates images with vivid and diverse colors.

Text-based Colorization. For text-based colorization, we only test MSCOCO with ground-truth text caption and segmentation. Because most of the text captions include no color descriptions, we automatically insert a color word before each object word. We first cluster the RGB values of all the pixels within the ground-truth segmentation of each object, and then assign the color with the highest occurrence to that object. Then we colorize the images from the new captions with color descriptions. Since the implementation and full results of the previous text-based method Learn-Color-Lang [Manjunatha et al. 2018] are not publicly available, we only perform

Fig. 9. Comparison with exemplar-based methods (Deep Exp. [He et al. 2018] and Exp. Video [Zhang et al. 2019]). The reference images are shown in the right-bottom corner of each input grayscale image. Input images: the reference image in the 5^{th} column is from MSCOCO, and the other images are from ImageNet.

Fig. 10. Comparison with the text-based colorization method Learn-Color-Lang [Manjunatha et al. 2018]. We show the comparison in the first group, and our diverse results in the second group. Input images: 1^{st} group is from paper [Manjunatha et al. 2018], 2^{nd} group is from MSCOCO.

visual comparison on the images shown in their original paper in Fig. 10.

Thus, for quantitative results, we only compare with our unconditional variant to verify whether the model follows the input text, and the score of ground-truth images is shown for reference. As shown in Tab. 4, our text-based method gets a higher CLIP similarity score and similar FID compared with the unconditional method. This indicates that our text-based method can correctly locate the objects and colorize them with accurate colors. Because of the ambiguities of color words, though our method generates different colors from the ground truths, they are still aligned well with the input color words. This may be the reason for the higher CLIP similarity score compared to the ground truths. Compared with Learn-Color-Lang (Fig. 10), our method produces more accurate color (*e.g.,* the gray horse in 1^{st} row), with less artifacts (*e.g.,* the blue sofa in 2^{nd} row). With CLIP-based hint points conversion, our method can respond to tiny objects (*e.g.,* the red sign in 4^{th} row) and open-vocabulary objects which may not appear in MSCOCO (*e.g.,* the purple hat in 3^{rd} row). This reflects the feasibility and flexibility of our method.

5.2 User Study

To further validate the effectiveness of our method, we conduct a user study and discuss the results in this subsection. As text-based colorization has no suitable previous work for comparison, we mainly compare three different design scenarios: unconditional, stroke-based, and exemplar-based colorization through the user study. For unconditional colorization, we select the two representative works used in qualitative comparison from CNN-based methods: InstColor [2020] and ChromaGAN [2020], and Coltran [2021] for Transformer-based method.

For each modality, we randomly collected 30 design cases based on different input grayscale images. For unconditional and stroke-based ones, we selected 15 images per dataset from ImageNet and MSCOCO, while for exemplar-based methods, we selected all from the MSCOCO dataset. For each design case, given the input grayscale image, hint points for stroke-based colorization, and a reference image for exemplar-based colorization, we showed the results from different methods in random order. The participants were asked to select the best one based on two metrics: realistic and consistent with input condition (if it has). Each participant was asked for 8 formal questions per scenario with an additional validation question, deriving 25 questions in total. The validation question is very simple by putting the only ground truth image among grayscale ones for testing the faithfulness of each participant.

We finally collected 390 questionnaires, where 301 of them are valid with passing the validation question. Among the 301 participants, 163 participants are below 20 years old, 70 range from 20 to 40 years of age, and 68 are above 40 years old. We show the result of the user study in Fig. 11, and we find that our method outperforms other methods in all unconditional, stroke-based, and

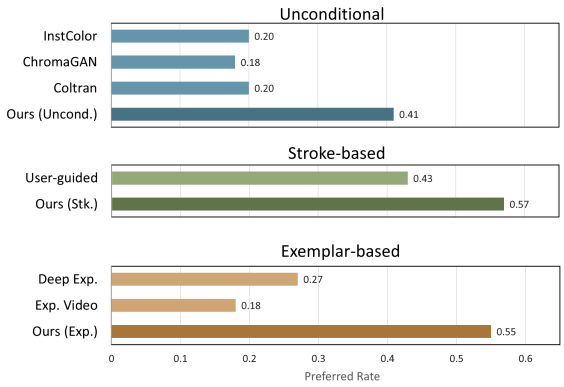

Fig. 11. Results of the user study on unconditional, stroke-based, and exemplar-based colorization tasks.

exemplar-based colorization with a preferred rate of 41%, 57%, and 55%, respectively.

5.3 Ablation Study

5.3.1 The effect of Chroma-VQGAN. To verify whether our Chroma-VQGAN can better reconstruct the images with unquantized gray features, we train a vanilla VQGAN and Quant-VQGAN, which has the same architecture as Chroma-VQGAN, but the gray features are quantized into tokens same as the color tokens. To compare the performance, we reconstruct the 5,000 images from ImageNet and compute the FID, PSNR, LPIPS, and SSIM [Wang et al. 2004]. As shown in Tab. 5 and Fig. 12, our Chroma-VQGAN reconstructs the color image with less distortions and obtains better metrics, compared with both the vanilla VQGAN and Quant-VQGAN. Our method is even better than vanilla VQGAN with a 4 times smaller downsampling rate and 16 times larger number of tokens (leads to 256 times more computational cost and 4096 times more inference time). The experiment indicates that the additional unquantized gray features preserve the structural details during the decoding of the color tokens, which enables better reconstruction with much less computational cost.

To examine whether our Chroma-VQGAN learns disentangled chrominance representation in color tokens, we combine different pure color images with the input grayscale image. A good disentangled chrominance representation only controls the color without influencing the underlying structure. As shown in Fig. 13, our Chroma-VQGAN still preserves the structure details even though the input color image is changed, whereas the results of Quant-VQGAN are blurred. This implies that the color tokens in Chroma-VQGAN contain solely the chrominance information, whereas the color tokens in Quant-VQGAN still carry structure details. Therefore, by keeping gray features unquantized, the disentangled chrominance representation can help Hybrid-Transformer focus on color prediction without distracting from structural details.

5.3.2 The effect of Hybrid-Transformer. Our Hybrid-Transformer takes inputs in a hybrid format with quantized color tokens, continuous gray features, and color hint points. In this subsection, we first create a baseline by replacing continuous gray features with discrete tokens based on Quant-VQGAN (**Quant-gray**) and test under unconditional colorization (*i.e.,* without input color hints). We then create the other baseline (**Quant-hint**) by replacing the continuous color hints with color tokens. Specifically, we obtain the token index by sending the pure hint color image into the color encoder of Chroma-VQGAN. Note that the gray features remain continuous under hint points in discrete tokens.

We show the results in Tab. 6 and Fig. 14. By using continuous gray features, our model outperforms the baseline (Quant-gray) by a large margin in terms of FID and LPIPS and achieves comparable results on Colorfulness. As for the qualitative result (1^{st} *row* in Fig. 14), the predicted colors from Quant-gray are not aligned with the input content, leading to a mess of colors. By using continuous hint points, our model outperforms the baseline (Quant-hint) on all three metrics significantly. As shown in the 2^{nd} row, the Quant-hint method cannot correctly inherit the color of the hint points,

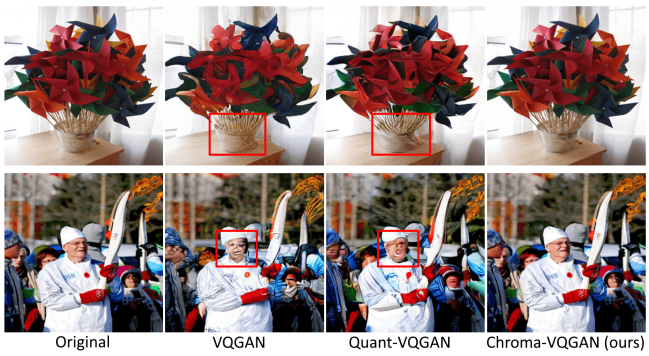

Fig. 12. Comparison results on image reconstruction for showing the effect of Chroma-VQGAN. Input original images: from ImageNet.

Table 5. Comparison results on image reconstruction. We compare our Chroma-VQGAN with the vanilla VQGAN [Esser et al. 2021] and a variant called Quant-VQGAN.

	Downsample rate	FID↓	PSNR↑	LPIPS↓	SSIM↑
VQGAN	4×	2.61	28.70	0.0433	0.8555
	16×	11.83	20.03	0.1691	0.5062
Quant-VQGAN	16×	11.78	20.67	0.1592	0.5316
Chroma-VQGAN (Ours)	16×	**1.68**	**29.73**	**0.0304**	**0.8770**

Table 6. The effect of Hybrid-Transformer. We compare with variants of replacing continuous gray features with discrete tokens (Quant-gray) and replacing continuous hint points with discrete tokens (Quant-hint).

	FID↓	LPIPS↓	Colorful↑
Quant-gray (Uncond.)	11.88	0.2035	**39.33**
Ours (Uncond.)	**9.46**	**0.1945**	39.01
Quant-hint (Stk.)	9.76	0.1445	24.45
Ours (Stk.)	**7.04**	**0.1119**	**36.16**

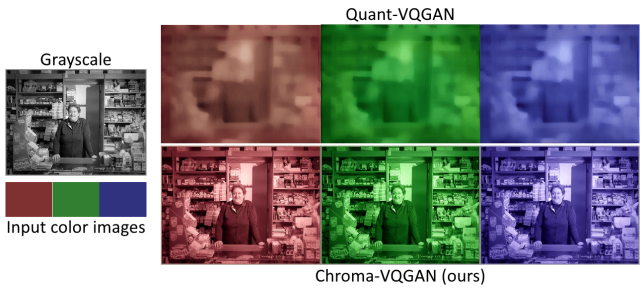

Fig. 13. Results of changing the input color images. Our Chroma-VQGAN generates colored images with high quality while results from Quant-VQGAN are blurred. Input image: from ImageNet.

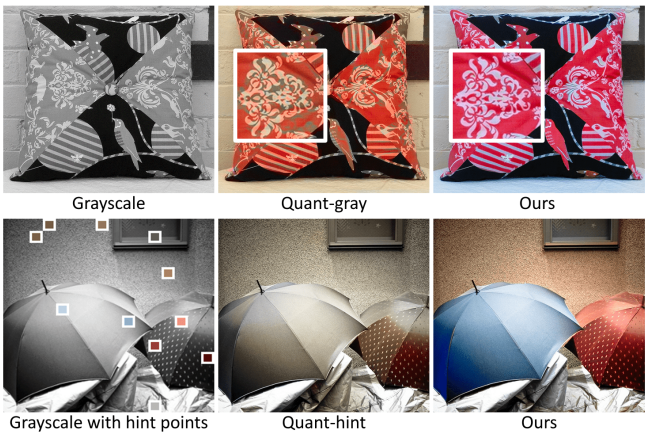

Fig. 14. The effect of Hybrid-Transformer. First row: comparison with Transformer trained using quantized gray features (Quant-gray). Second row: comparison with Transformer trained using quantized hint colors (Quant-hint). Input images: from ImageNet.

especially when the input hint points are in different colors. The results imply that our Hybrid-Transformer can receive more original and accurate conditional information from the unquantized colors of hint points.

6 APPLICATIONS

With the interactive system and our flexible framework, we show various practical applications in this section, including colorization with hybrid controls, recolorization, and iterative editing. Rather than only showing results on natural images, we also demonstrate more results on legacy old photos. To enable the model to deal with images with diverse contents, we retrain our model on MSCOCO dataset for showing results in this section.

6.1 Colorization with Hybrid Controls

Our model allows hybrid-modality for controlling the generated colorization results. We show the results in Fig. 15, and they follow the input hybrid conditions well. For example, in the last column of 4^{th} row, we use strokes to control the color of the lamp and hair, reference image to control the color of the curtain, and text for the color of the suit.

6.2 Recolorization

Except for colorization, our model also allows recolorization of existing color images under different controls. To recolorize the selected region, we simply mask the color tokens within the region and resample them with given conditions. We show the results in Fig. 16. Our model can adjust the colors of different objects, such as the jacket in 1^{st} example (from green to orange) and the motor in the 3^{rd} example (from green to red). Besides, instead of only recolorizing on single objects, our model is able to recolor larger scenes, such as the context of the orange bus in the 2^{nd} example.

6.3 Iterative Editing

When colorizing an image, a single pass is often not enough to fully convey the user's intention, and iterative editing becomes an essential way for adjusting and improving the result. Our model is also capable of doing so. We show several results in Fig. 17. For tiny objects (*e.g.,* the statue in the 1^{st} row and the pot in the 2^{nd} row), the users can use strokes or texts to edit the colors. For large objects (*e.g.,* the building in the 1^{st} row and the dress in the 3^{rd} row), the users can use reference images to control the colors. This also reflects the convenience and flexibility of our system, where users can select different controls to edit various types of objects and images.

7 CONCLUSIONS & LIMITATIONS

To conclude, we propose the first unified framework UniColor, which supports diverse colorization in different modalities, including both unconditional and conditional ones (*e.g.,* stroke, exemplar, and text). Different from existing works, which only support a specific type of user control and cannot generalize to other ones, our framework unifies all three types of controls into the form of hint points, which can be naturally extracted from stroke and exemplar conditions. To extract hint points from the text input, we propose a novel CLIP-based method to locate the objects described in the text and add corresponding colors to form the hint points. With the hint points, we propose a network, which consists of a Chroma-VQGAN and Hybrid-Transformer, for diverse colorization with high quality and colorfulness. Based on our unified framework, we design an interactive system to support all four types of image colorization (*i.e.,* unconditional, stroke-based, exemplar-based, and text-based). The system also enables the user to perform hybrid controls, recolorizing, and iterative editing.

Despite the superior performance of UniColor, we still encounter some limitations to be explored in the future. First, diversity is generally an advantage in image colorization, but, on the other hand, the stochastic sampling may seldom lead to unexpected colors, such as the green road and brown broccoli in Fig. 18 1^{st} row. This can be alleviated by limiting the range of stochastic sampling but sacrificing diversity to some degree. Or we may consider filtering the colorization results based on some semantic metric. For example, we may train a network to evaluate the quality of the produced colors. Moreover, control conflicts might occur when mixing conditions

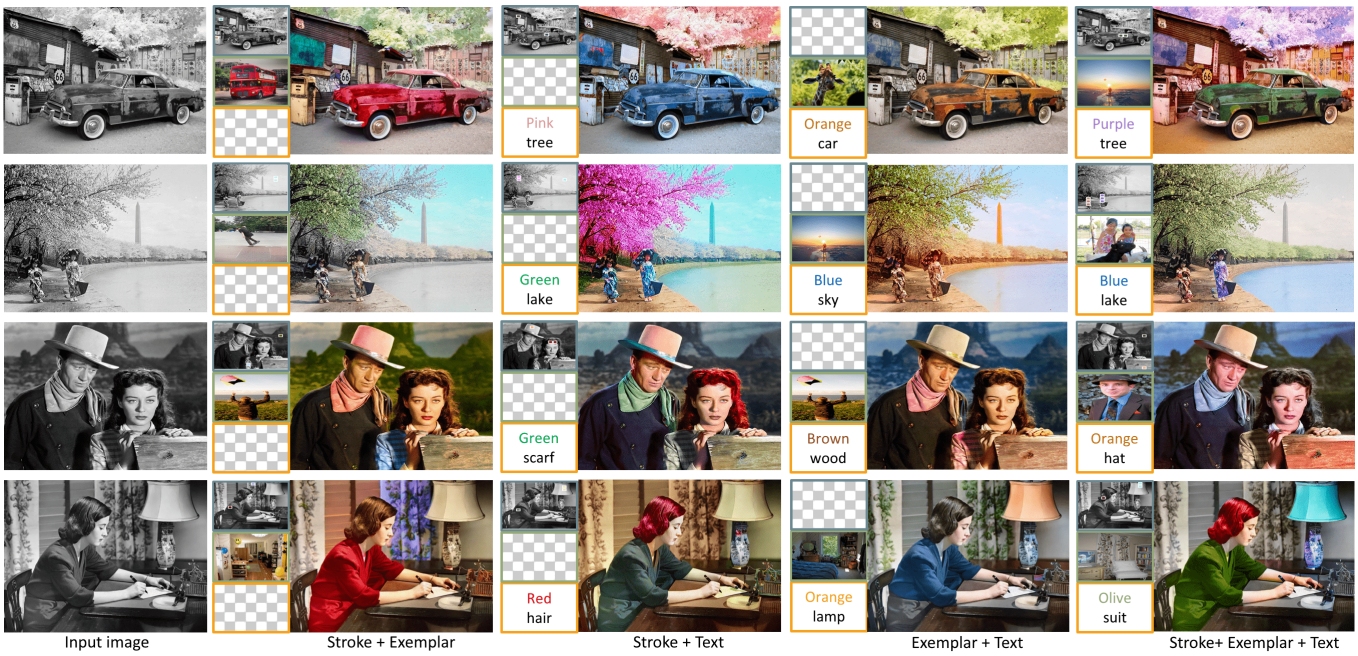

Fig. 15. Colorization on legacy old photos with hybrid controls. In each row, we first show the input grayscale photo, followed by four design cases under different hybrid conditions. We indicate stroke-based conditions with blue boundaries, exemplar-based conditions with green boundaries, and text-based conditions with orange boundaries. Input images (from top to bottom): 1) Oldtimer automobile Crom; 2) Sumi and Sada Tamura, 1925; 3) Gail Russell & John Wayne in *Angel and the Badman*, 1947; 4) Portrait of woman writing letter at desk, 1950. All reference images are from MSCOCO.

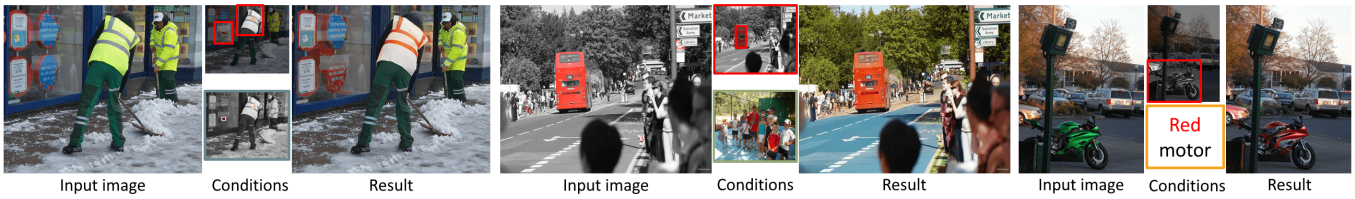

Fig. 16. Recolorization supported by our framework. For each design case, in the middle column, we show the selected region in the upper row, and the input condition in the lower row. Input images: all images are from MSCOCO, except the first input image is from ImageNet.

from different modalities. As shown in Fig. 18 2^{nd} row, the green hint points from the text condition, and the red stroke lay on the same clothes and thus generated a mixture of green and red. This problem can be avoided by more careful user interactions or designing an algorithm to auto-detect the conflicts based on image segmentation. For example, if the algorithm detects that the hint points from different controls lay on the same segment, it will automatically ignore some hint points according to the default priority or notify the user to make further decisions. It would be worthwhile to further improve user experiences in colorization tasks with our Unicolor framework.

8 ACKNOWLEDGMENTS

We thank the anonymous reviewers for helping us to improve this paper. We also thank the artists and photographers for approving us to use their photos. This work was supported by the Hong Kong Research Grants Council (RGC) GRF Scheme under Grant CityU 11216122.

REFERENCES

Radhakrishna Achanta, Appu Shaji, Kevin Smith, Aurelien Lucchi, Pascal Fua, and Sabine Süsstrunk. 2012. SLIC Superpixels Compared to State-of-the-Art Superpixel Methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 34, 11 (2012), 2274–2282. https://doi.org/10.1109/TPAMI.2012.120

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.0473

Hyojin Bahng, Seungjoo Yoo, Wonwoong Cho, David Keetae Park, Ziming Wu, Xiaojuan Ma, and Jaegul Choo. 2018. Coloring with Words: Guiding Image Colorization Through Text-Based Palette Generation. In *Computer Vision – ECCV 2018*, Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss (Eds.). Springer International Publishing, Cham, 443–459.

Yun Cao, Zhiming Zhou, Weinan Zhang, and Yong Yu. 2017. Unsupervised Diverse Colorization via Generative Adversarial Networks. In *Machine Learning and Knowledge Discovery in Databases*, Michelangelo Ceci, Jaakko Hollmén, Ljupčo Todorovski,

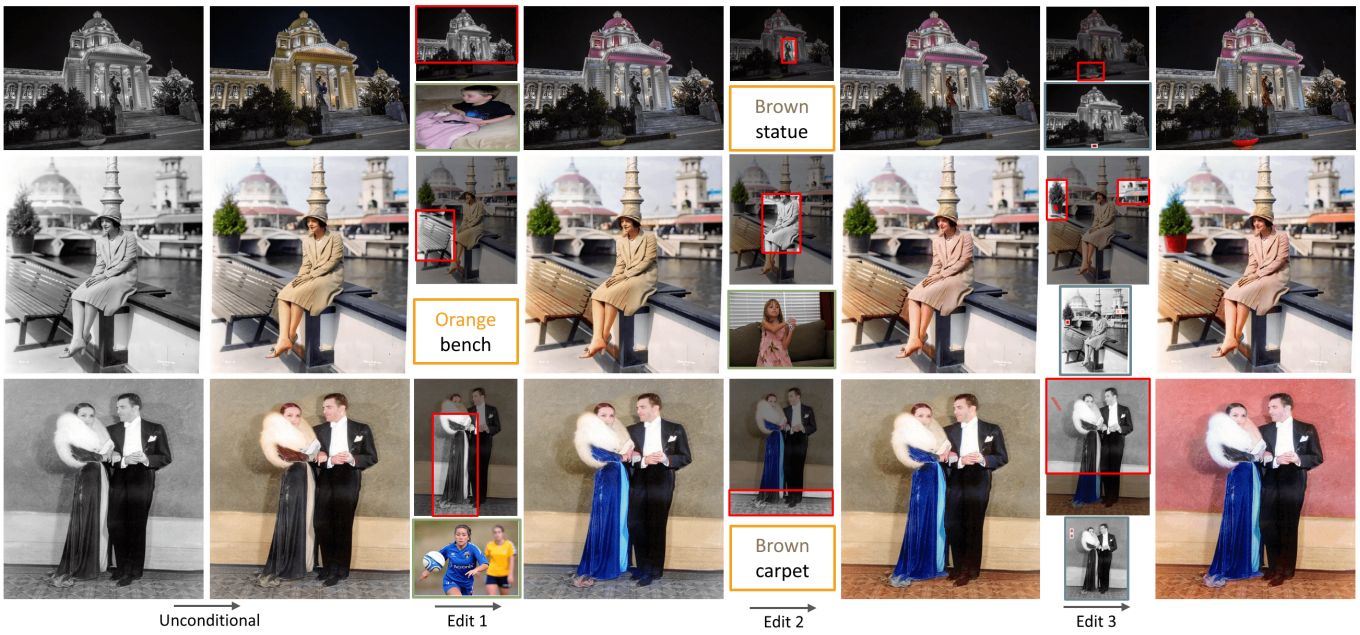

Fig. 17. Iterative editing on legacy old photos. Users can adjust the color based on various conditions iteratively. Input images (from top to bottom): 1) Night view of the serbian national assembly building; 2) Janet Gaynor, 1920s Coney Island; 3) Dolores del Río next to her husband Cedric Gibbons. Reference images: from MSCOCO.

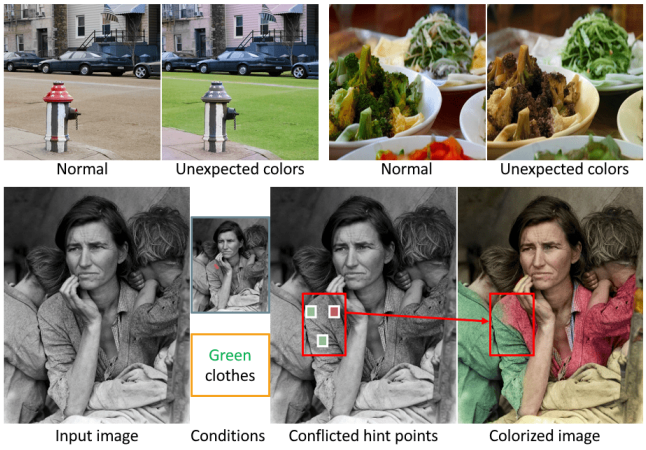

Fig. 18. Failure cases. 1^{st} row: unexpected colors, input images: from MSCOCO. 2^{nd} row: conflicts in hybrid controls, input image: Migrant Mother, 1936.

Celine Vens, and Sašo Džeroski (Eds.). Springer International Publishing, Cham, 151–166.

Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2018. Language-Based Image Editing with Recurrent Attentive Models. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), 8721–8729.

Mark Chen, Alec Radford, Rewon Child, Jeffrey Wu, Heewoo Jun, David Luan, and Ilya Sutskever. 2020. Generative Pretraining From Pixels. In *Proceedings of the 37th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 119)*, Hal Daumé III and Aarti Singh (Eds.). PMLR, 1691–1703. https://proceedings.mlr.press/v119/chen20s.html

Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. 2019. Generating Long Sequences with Sparse Transformers. *URL https://openai.com/blog/sparse-transformers* (2019).

Deoldify. 2018. Deoldify: A Deep Learning based Project for Colorizing and Restoring Old Images and Video. https://github.com/jantic/DeOldify.

Aditya Deshpande, Jiajun Lu, Mao Chuang Yeh, Min Jin Chong, and David Forsyth. 2017. Learning diverse image colorization. In *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017 (Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017)*. Institute of Electrical and Electronics Engineers Inc., United States, 2877–2885. https://doi.org/10.1109/CVPR.2017.307

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT 2019, Minneapolis, MN, USA, June, 2019, Volume 1*. 4171–4186. https://doi.org/10.18653/v1/n19-1423

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net. https://openreview.net/forum?id=YicbFdNTTy

Yuki Endo, Satoshi Iizuka, Yoshihiro Kanamori, and Jun Mitani. 2016. DeepProp: Extracting Deep Features from a Single Image for Edit Propagation. *Computer Graphics Forum* 35 (05 2016), 189–201. https://doi.org/10.1111/cgf.12822

Patrick Esser, Robin Rombach, and Bjorn Ommer. 2021. Taming Transformers for High-Resolution Image Synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 12873–12883.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K.Q. Weinberger (Eds.), Vol. 27. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2014/file/5ca3e9b122f61f8f06494c97b1afccf3-Paper.pdf

David Hasler and Sabine E. Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human Vision and Electronic Imaging VIII*, Bernice E. Rogowitz and Thrasyvoulos N. Pappas (Eds.), Vol. 5007. International Society for Optics and Photonics, SPIE, 87 – 95. https://doi.org/10.1117/12.477378

Mingming He, Dongdong Chen, Jing Liao, Pedro V Sander, and Lu Yuan. 2018. Deep exemplar-based colorization. *ACM Transactions on Graphics (TOG)* 37, 4 (2018), 47.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2017. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*,

Vol. 30.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020a. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 6840–6851. https://proceedings.neurips.cc/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf

Jonathan Ho, Nal Kalchbrenner, Dirk Weissenborn, and Tim Salimans. 2020b. Axial Attention in Multidimensional Transformers. https://openreview.net/forum?id=H1e5GJBtDr

Satoshi Iizuka, Edgar Simo-Serra, and Hiroshi Ishikawa. 2016. 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 (Proc. of SIGGRAPH 2016)* 35, 4 (2016).

Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A. Efros. 2017. Image-to-Image Translation with Conditional Adversarial Networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. 5967–5976. https://doi.org/10.1109/CVPR.2017.632

Xin Jin, Zhonglan Li, Ke Liu, Dongqing Zou, Xiaodong Li, Xingfan Zhu, Ziyin Zhou, Qi song Sun, and Qingyu Liu. 2021. Focusing on Persons: Colorizing Old Images Learning from Modern Historical Movies. *Proceedings of the 29th ACM International Conference on Multimedia* (2021).

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. 2014. ReferItGame: Referring to Objects in Photographs of Natural Scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, 787–798. https://doi.org/10.3115/v1/D14-1086

Leila Kiani, Masoudnia Saeed, and Hossein Nezamabadi-pour. 2020. Image Colorization Using Generative Adversarial Networks and Transfer Learning. *2020 International Conference on Machine Vision and Image Processing (MVIP)* (2020), 1–6.

Yoon Kim, Carl Denton, Luong Hoang, and Alexander M. Rush. 2017. Structured Attention Networks. *CoRR* abs/1702.00887 (2017). arXiv:1702.00887 http://arxiv.org/abs/1702.00887

Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. 2021. Colorization Transformer. In *International Conference on Learning Representations*. https://openreview.net/forum?id=5NA1PinlGFu

Anat Levin, Dani Lischinski, and Yair Weiss. 2004. Colorization Using Optimization *(SIGGRAPH '04)*. Association for Computing Machinery, New York, NY, USA, 689–694. https://doi.org/10.1145/1186562.1015780

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014*, David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars (Eds.). Springer International Publishing, Cham, 740–755.

Qing Luan, Fang Wen, Daniel Cohen-Or, Lin Liang, Ying-Qing Xu, and Heung-Yeung Shum. 2007. Natural Image Colorization. In *Rendering Techniques*, Jan Kautz and Sumanta Pattanaik (Eds.). The Eurographics Association. https://doi.org/10.2312/EGWR/EGSR07/309-320

Gen Luo, Yiyi Zhou, Xiaoshuai Sun, Liujuan Cao, Chenglin Wu, Cheng Deng, and Rongrong Ji. 2020. Multi-Task Collaborative Network for Joint Referring Expression Comprehension and Segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. 2018. Learning to Color from Language. In *North American Chapter of the Association for Computational Linguistics*.

Roey Mechrez, Itamar Talmi, and Lihi Zelnik-Manor. 2018. The Contextual Loss for Image Transformation with Non-Aligned Data. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Safa Messaoud, David A. Forsyth, and Alexander G. Schwing. 2018. Structural Consistency and Controllability for Diverse Colorization. In *ECCV (6)*. 603–619. https://doi.org/10.1007/978-3-030-01231-1_37

Niki Parmar, Ashish Vaswani, Jakob Uszkoreit, Lukasz Kaiser, Noam Shazeer, Alexander Ku, and Dustin Tran. 2018. Image Transformer. In *Proceedings of the 35th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 80)*, Jennifer Dy and Andreas Krause (Eds.). PMLR, 4055–4064. https://proceedings.mlr.press/v80/parmar18a.html

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. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *ICML*.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-Shot Text-to-Image Generation. In *Proceedings of the 38th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 139)*, Marina Meila and Tong Zhang (Eds.). PMLR, 8821–8831. https://proceedings.mlr.press/v139/ramesh21a.html

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. 2015. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)* 115, 3 (2015), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Chitwan Saharia, William Chan, Huiwen Chang, Chris A. Lee, Jonathan Ho, Tim Salimans, David J. Fleet, and Mohammad Norouzi. 2021. Palette: Image-to-Image Diffusion Models. https://doi.org/10.48550/ARXIV.2111.05826

Karen Simonyan and Andrew Zisserman. 2015. Very Deep Convolutional Networks for Large-Scale Image Recognition. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.). http://arxiv.org/abs/1409.1556

Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. 2020. Instance-aware Image Colorization. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

Patricia Vitoria, Lara Raad, and Coloma Ballester. 2020. ChromaGAN: Adversarial Picture Colorization with Semantic Class Distribution. In *The IEEE Winter Conference on Applications of Computer Vision*. 2445–2454.

Zhou Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. 2004. Image Quality Assessment: From Error Visibility to Structural Similarity. *Trans. Img. Proc.* 13, 4 (apr 2004), 600–612. https://doi.org/10.1109/TIP.2003.819861

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. CRIS: CLIP-Driven Referring Image Segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*.

Dirk Weissenborn, Oscar Täckström, and Jakob Uszkoreit. 2020. Scaling Autoregressive Video Models. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rJgsskrFwH

Chenfei Wu, Jian Liang, Lei Ji, Fan Yang, Yuejian Fang, Daxin Jiang, and Nan Duan. 2021a. NÜWA: Visual Synthesis Pre-training for Neural visUal World creAtion. https://doi.org/10.48550/ARXIV.2111.12417

Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. 2021b. Towards Vivid and Diverse Image Colorization with Generative Color Prior. In *International Conference on Computer Vision (ICCV)*.

Yi Xiao, Peiyao Zhou, Yan Zheng, and Chi-Sing Leung. 2019. Interactive Deep Colorization Using Simultaneous Global and Local Inputs. In *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 1887–1891. https://doi.org/10.1109/ICASSP.2019.8683686

Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. 2020. Stylization-Based Architecture for Fast Deep Exemplar Colorization. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 9360–9369. https://doi.org/10.1109/CVPR42600.2020.00938

Bo Zhang, Mingming He, Jing Liao, Pedro V Sander, Lu Yuan, Amine Bermak, and Dong Chen. 2019. Deep exemplar-based video colorization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8052–8061.

Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful Image Colorization. In *ECCV*.

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

Richard Zhang, Jun-Yan Zhu, Phillip Isola, Xinyang Geng, Angela S Lin, Tianhe Yu, and Alexei A Efros. 2017. Real-Time User-Guided Image Colorization with Learned Deep Priors. *ACM Transactions on Graphics (TOG)* 9, 4 (2017).

Zhu Zhang, Jianxin Ma, Chang Zhou, Rui Men, Zhikang Li, Ming Ding, Jie Tang, Jingren Zhou, and Hongxia Yang. 2021. UFC-BERT: Unifying Multi-Modal Controls for Conditional Image Synthesis. In *Advances in Neural Information Processing Systems*, A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan (Eds.). https://openreview.net/forum?id=iEEAPq3TUEZ