Versatile Vision Foundation Model for Image and Video Colorization

Vukasin Bozic
ETH Zürich
Zürich, Switzerland
vbozic@student.ethz.ch

Abdelaziz Djelouah
Disney Research Studios
Zürich, Switzerland
aziz.djelouah@disneyresearch.com

Yang Zhang
Disney Research Studios
Zürich, Switzerland
yang.zhang@disneyresearch.com

Radu Timofte
University of Wurzburg
Würzburg, Germany
radu.timofte@uni-wuerzburg.de

Markus Gross
ETH Zürich
Zürich, Switzerland
grossm@inf.ethz.ch

Christopher Schroers
Disney Research Studios
Zürich, Switzerland
christopher.schroers@disneyresearch.com

Figure 1: We propose a single model to handle the various aspects of the colorization problem: multiple colorization results, user guidance through hints or prompts, and temporal stability. Existing methods are not able to address all these aspects at once while achieving competitive results on all of them.

ABSTRACT

Image and video colorization are among the most common problems in image restoration. This is an ill-posed problem and a wide variety of methods have been proposed, ranging from more traditional computer vision strategies to most recent development with transformer-based or generative neural network models. In this work we show how a latent diffusion model, pre-trained on text-to-image synthesis, can be finetuned for image colorization and provide a flexible solution for a wide variety of scenarios: high quality direct colorization with diverse results, user guided colorization through colors hints, text prompts or reference image and finally video colorization. Some works already investigated using diffusion models for colorization, however the proposed solutions are often more complex and require training a side model guiding the denoising process (*à la* ControlNet). Not only is this approach increasing the number of parameters and compute time, it also results in sub optimal colorization as we show. Our evaluation demonstrates that our model is the only approach that offers a wide flexibility while either matching or outperforming existing methods specialized in each sub-task, by proposing a group of universal, architecture-agnostic mechanisms which could be applied to any pre-trained diffusion model.

CCS CONCEPTS

• **Computing methodologies** → **Computational photography**.

KEYWORDS

Colorization, Image and Video Colorization, Image Restoration

ACM Reference Format:
Vukasin Bozic, Abdelaziz Djelouah, Yang Zhang, Radu Timofte, Markus Gross, and Christopher Schroers. 2024. Versatile Vision Foundation Model for Image and Video Colorization. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Conference Papers '24 (SIGGRAPH Conference Papers '24), July 27–August 01, 2024, Denver, CO, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3641519.3657509

SIGGRAPH Conference Papers '24, July 27–August 01, 2024, Denver, CO, USA

ACM ISBN 979-8-4007-0525-0/24/07
https://doi.org/10.1145/3641519.3657509

1 INTRODUCTION

Modification of the colors, whether it is for artistic purposes [Coen et al. 2001] or for the restoration of heritage footage [ame 2017], is a very common task in video production. As a result, there is a large body of works addressing several aspects of the problem with some notable ones [Antic 2020; Iizuka and Simo-Serra 2019] in the recent years. It is interesting to note that the problem is far from

being solved and remains an important topic in modern workshops on image and video restoration [Kang et al. 2023a].

Given the complexity of the task and the particularly ill-posed nature of the problem, earlier colorization works have focused on the propagation of color hints provided by the user. This can be either in the context of colorizing a single image [Levin et al. 2004], or temporal propagation [Welsh et al. 2002]. Among the most recent advances here we can cite *UniColor* [Huang et al. 2022] that uses a combination of a Chroma-VQGAN model and a transformer for diverse results. For unconditional colorization, a recently published work [Kang et al. 2023b] achieves impressive results using transformers, with one notable limitation of producing only one colorization variant per image. Additionally, user guidance or multi-frame output are not considered. To tackle diverse colorization, generative models have been used, with some of the most recent works relying on diffusion models [Liu et al. 2023b; Saharia et al. 2022]. In particular leveraging models that have been trained for text-to-image synthesis on large datasets. However these works often consider complex changes, with either a parallel model that is trained to condition the denoising [Liu et al. 2023a,b; Zhang et al. 2023] or a more involved modification of denoising model [Chang et al. 2023].

Despite more and more impressive results, there is still no generic solution that would be competitive on the various sub-tasks associated with colorization. In addition to this, text-to-image diffusion models that have been trained on extremely large-scale datasets embed a high-level visual understanding, which we think is important to address a problem such as colorization. In this work we show how a latent diffusion model (namely stable diffusion [Rombach et al. 2022]) can be finetuned after minor modifications to address the colorization problem. Additionally, we show how to take into account different modalities such as text, color hints from user or previously colorized frames both during training and inference time. On top of that, by leveraging various consistency-improving techniques, the model could be utilized for video colorization without any additional training, bringing the high quality colorization to the video domain. We note that such a strategy for using a vision foundation model have been explored concurrently to our work on related tasks such as optical flow [Saxena et al. 2023a] or depth [Ke et al. 2023] estimation.

A thorough evaluation of the trained model is provided using the most common datasets and evaluation metrics, including a user study. All evaluations clearly demonstrate the versatility of our model and its top performance on a wide range colorization subtasks, rivaling other more specialized models focused on single aspects of the problem. Ultimately we present a unified solution for the general colorization problem.

Our contributions can be summarized as follow:

- Leveraging vision foundation models for the colorization problem.
- A single versatile model capable of achieving state-of-the-art results in the various image colorization sub-tasks.
- A detailed evaluation including a user study which demonstrates the superiority of our model.

2 RELATED WORK

In general state-of-the-art methods and algorithms suffer from one or more of the following problems:

- Single colorization output or limited diversity.
- Sub-optimal visual performance in general.
- High specialization towards a particular sub-task, with little or no room for generalization.
- Complex architecture design around latent diffusion models.

Our model aims at addressing all of these issues, obtaining high quality and diverse colorization results, while accepting various forms of guidance. Next we review the different types of colorization works, focusing on most recent and best performing ones.

2.1 Automatic colorization

We can cite [Iizuka et al. 2016; Zhang et al. 2016] among the initial works adapting classical CNN architectures, essentially treating the problem as a classification task. Others use patch matching like strategy [Cheng et al. 2015], object detection [Su et al. 2020] or instance segmentation [Zhao et al. 2020]. Some [Larsson et al. 2016; Xia et al. 2022] also predict color distribution.

Transformer models [Vaswani et al. 2017] are exploited for the task of image colorization [Kumar et al. 2021], with Weng *et al.* and Ji *et al.* [2022; 2022a] further specializing the standard architecture. Transformers are largely used in the recent top performing model DDColor [Kang et al. 2023b]. Still, all of these methods provide single or very limited number of colorization results, which, given the nature of the problem, is not always desirable.

For diverse colorization results, methods based on generative models such as VAE [Deshpande et al. 2017] or GANs [Wu et al. 2021] were proposed. Vitoria *et al.* [2020] additionally utilize classification module for colorization guidance. Different methods [Huang et al. 2022] use GAN both as an auxiliary module to generate images queries from the latent space of a pretrained GAN [Vitoria et al. 2020], or condition the GAN network directly [Kim et al. 2022], efficiently sampling the colorized images. Finally, a range of specialized methods with custom procedures have shown a lot of promise [Antic 2020], combining the classical and adversarial training methods.

2.2 User-guided colorization

If we consider the end use case for colorization problem, it is very likely to be part of an interactive image (or video) editing tool. The nature of the problem is such that users would very likely want to guide the colorization. To that end, a handful of guiding modalities were introduced: color hints, reference images and text.

Color hint based colorization is the most prominent one, with some early works formulating the color propagation as an optimization problem [Levin et al. 2004]. More recently neural network based models also explored user guidance [Endo et al. 2016; Zhang et al. 2017]. In their work, Xiao *et al.* [2019] discriminate between the local and global inputs, enabling finer control over the colorization process. Finally a range of transformer [Yun et al. 2023] or GAN [Dou et al. 2021] based methods where also proposed for leveraging user inputs.

Reference-based colorization aims at transferring the color from a reference image to the grayscale one. Earlier works [Welsh et al.

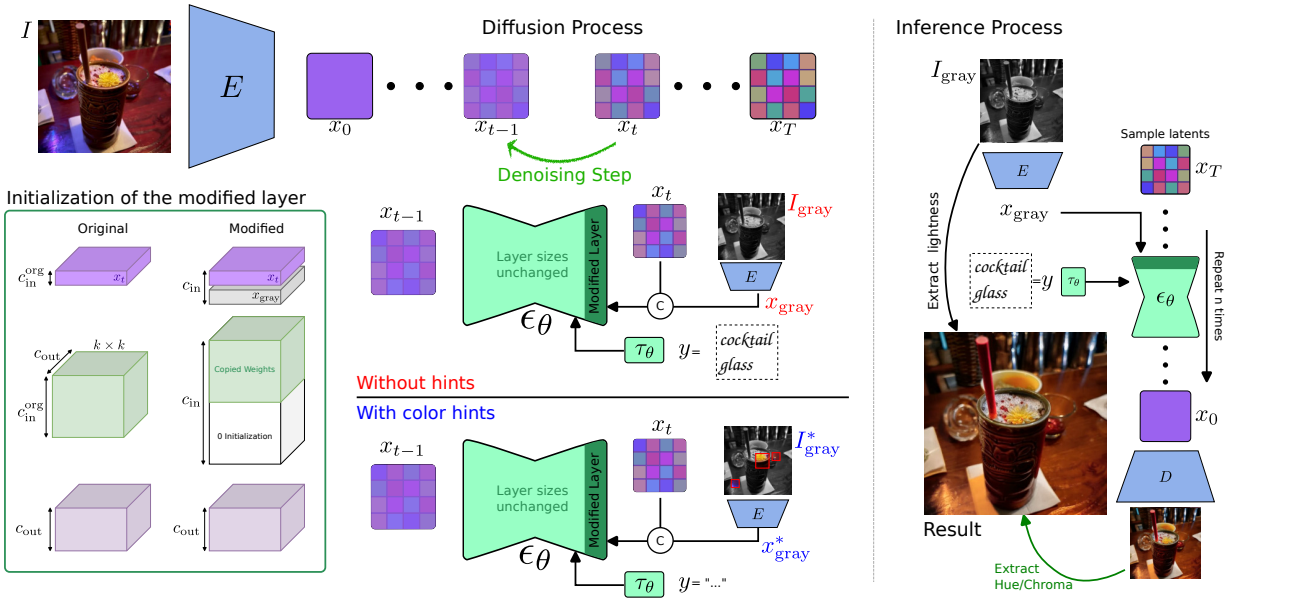

Figure 2: Overview of the proposed model. We re-purpose a pre-trained latent diffusion model. During training, We randomly sample an image that is encoded with E. The denoising UNet is trained to revert diffusion steps. In this case the model additionally takes as input the encoding of the grayscale image with potentially some areas with colors. We change the first convolution to adjust for this change with a specific weight initialization. The right side shows the inference process with a grayscale (no hints). The decoded image after denoising may contain artifacts due to the limited capacity of the decoder. This is not an issue for colorization as only Hue/Chroma values need to be copied and image details are maintained by using the original lightness of the grayscale image.

2002] rely on classical processing and matching between pixels, while more recent ones [He et al. 2018] propose a matching network to find the best reference image. Xue *et al.* and Bai *et al.* [2022; 2020] propagate the color in a coarse-to fine manner. More recently transformers [Wang et al. 2023; Yin et al. 2021] enable the long-context attention operation and matching between reference and target images.

Text-based colorization emerged most recently, following the success of language processing in other research areas [Li et al. 2022; Radford et al. 2021]. Manjunatha *et al.* [2018] utilize LSTM combined with extracted visual features. Chen *et al.* [2018]. introduce the color via object segmentation. Further improvements were made using attention-based modules [Chang et al. 2022; Weng et al. 2022b].

Finally, in multi-modal conditioning several of the aforementioned guidance strategies are used. Usually the problem is tackled through the reduction of the different multi-modalities into a single one, by applying different pre-processing modules [Huang et al. 2022; Yan et al. 2023].

2.3 Video Colorization

Video colorization poses the additional challenge of temporal consistency across the frames needs. Condition-based image colorization methods can be utilized, usually along with optical flow [Jampani et al. 2017] or instance tracking [Akimoto et al. 2020]. However this is less reliable and prone to accumulated correspondence errors. Different deep learning modules [Shi et al. 2023; Yang et al. 2022; Zhang et al. 2019] have been introduced to tackle the problem of consistency. Similarly to the single image case, generative models [Zhao et al. 2022] enable more diverse colorization results. We refer to Kang et al. [2023a] for an in-depth discussion around video colorization methods.

2.4 Diffusion models for colorization

With the emergence of diffusion models [Ho et al. 2020; Rombach et al. 2022], a plethora of different applications in the image and video domain have been explored. In particular, conditional image generation [Zhang et al. 2023], image editing [Brooks et al. 2023; Hertz et al. 2022; Kawar et al. 2023; Mokady et al. 2023], monocular depth estimation [Saxena et al. 2023b] etc. Saharia *et al.* [2022] propose an image space diffusion model that can be used for various image enhancement problems, including colorization. This model shares the same limitation as other image space diffusion models in terms of compute or lack of publicly available pre-trained model on large datasets. Reusing pre-trained latent diffusion model is the most commonly adopted path. A first option is training a ControlNet [Zhang et al. 2023] model for gray-scale image conditioning. Another is to design a colorization specific auxiliary module [Liu et al. 2023b]. This line of work was extended to video colorization [Liu et al. 2023a], where the auxiliary attention modules for temporal consistency were implemented and trained. Some approaches explored the finetuning of pre-trained diffusion models for this specific task [Lin et al. 2023] and furthermore introducing additional conditioning modalities, such as hints [Carrillo et al. 2023]. Lastly, Change *et al.* [2023], propose cross-modality modules for increased local, object level control. Although they share the

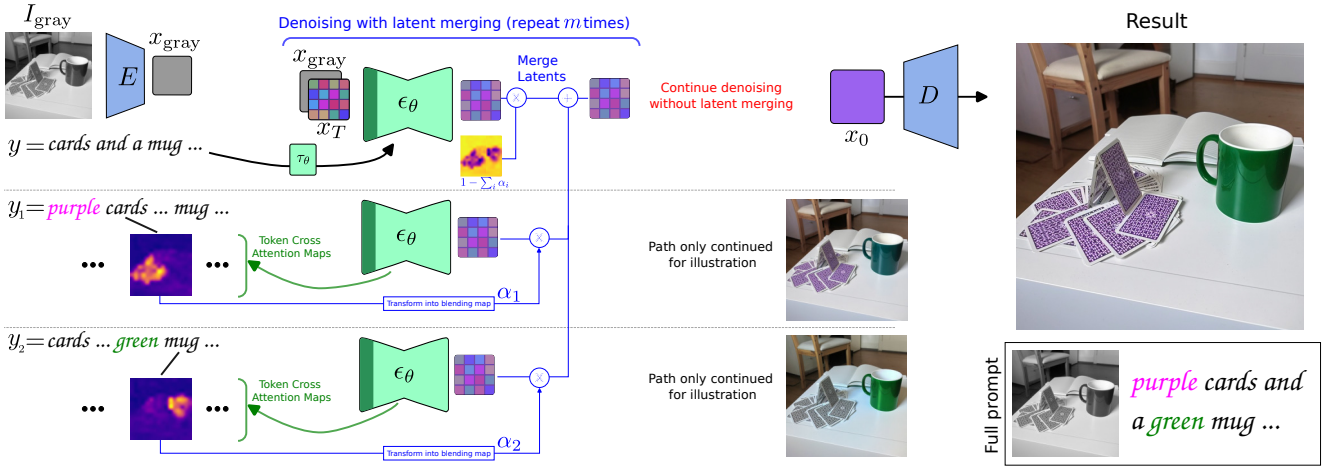

Figure 3: Overview of the color guidance through text. To colorize the grayscale image using the full prompt, we create parallel denoising paths: first a main path with a prompt y where all color information is removed; then, a path for each color from the full prompt (in this case y_1 and y_2). After each denoising steps the latents from the parallel paths are merged using the attention maps associated with each object token. This is repeated m times until color information is integrated in the main path (Here we continue the parallel paths for illustration).

common objective of re-purposing a pre-trained model, all these works propose complex and involved modifications. This in opposition to our proposal: while the changes are simpler they still demonstrate a large versatility in terms of control and great colorization performance validated both with quantitative evaluation and a user study.

3 METHOD

Our objective is to learn the conditional distribution of color images $p(I|I_{\text{gray}})$ given the gray scale image I_{gray}. To use a latent diffusion model [Rombach et al. 2022], images are encoded using a variational auto-encoder (VAE). Applied to our use case, the images I and I_{gray} are respectively encoded via the encoder E into their latent space representation x and x_{gray}, and $p(x|x_{\text{gray}})$ is the conditional distribution the latent diffusion model learns. In the forward process, noise is added according to the current time step $t \in \{0, \ldots, T\}$

$$x_t = \sqrt{\gamma} x_0 + \sqrt{1-\gamma}\epsilon \tag{1}$$

with $\epsilon \sim \mathcal{N}(0, I)$ and the parametrization of γ controls the variance schedule of the process. To model the reverse process we train a denoising UNet ϵ_θ with parameters θ, that minimizes the following loss

$$\mathcal{L} = \mathop{\mathbb{E}}_{x,y,t,\epsilon} ||\epsilon - \epsilon_\theta(x_t, \tau_\theta(y), t, x_{\text{gray}})||_2^2 \tag{2}$$

where $x = E(I)$ and I are sampled from an image dataset with corresponding text prompts y, while t is uniformly sampled in a set of diffusion steps. τ_θ is the part of the model that transforms text prompts.

3.1 Adapting a Foundation Model for Colorization

Text-to-image synthesis models trained on billions of images learn a powerful representation of visual content. To repurpose such a model for colorization we need to address two issues: *first*, the model needs to be adapted to use gray scale image conditioning, and *second*, we need to overcome the stable diffusion decoder limitations in reconstructing image details [Betker et al. 2023].

We address the first point by modifying the neural network architecture, as illustrated in Figure 2. The denoising UNet is augmented to incorporate the latent representation, x_{gray}, of the grayscale image as an additional input. However, a crucial aspect is the initialization of the weights for the kernel of the first convolution layer. Employing any of the default strategies [He et al. 2015] for this initialization doesn't allow the model to converge. Instead, we need to repurpose the pre-trained weights. Specifically, the first convolutional layer utilizes a $k \times k \times c_{\text{in}} \times c_{\text{out}}$ kernel. This differs from the original pre-trained model that uses a $k \times k \times c_{\text{in}}^{\text{org}} \times c_{\text{out}}$ kernel. The details in Figure 2 demonstrate how values from the original kernels are copied into the new one, while the remaining weights are initialized to 0.

The second issue: lack of quality in the reconstruction of the VAE, can be effectively circumvented in colorization case by employing a different color space. Since lightness channel of the image encapsulates all the texture details, we can retain original grayscale image lightness while incorporating the Chroma/Hue information from the model prediction. In our case, we utilize the CIELab color space.

3.2 Versatile User Guided Colorization

Despite its simplicity, the proposed method demonstrates adaptability accross a diverse range of scenarios.

Diverse Colorization Results. The right side of Figure 2 illustrates the process of colorizing a gray scale image. Firstly, latent values x_T are sampled, then a series of n denoising steps are conducted with the available conditioning information, in this case x_{gray} and y. After denoising, the image is decoded using the decoder D from the pre-trained VAE. We exclusively retain the hue/chroma from the decoded image and keep the lightness from the original grayscale image. During the denoising stage, we employ the classifier-free

guidance [Ho and Salimans 2022] and DDIM sampling [Song et al. 2020] similar to standard Diffusion Model approaches.

$$\bar{\epsilon} = (1+w)\epsilon_\theta(x_t, t, y, x_{\text{gray}}) - w\epsilon_\theta(x_t, t, x_{\text{gray}}) \quad (3)$$

Since we use a generative model for colorization, we can output diverse colorization results by sampling different latent x_T. As illustrated in Figure 4, this ability to produce multiple interpretations is crucial for ill-posed problems like colorization: multiple plausible colorizations can exist for a given grayscale image.

Color Hints. Since the model utilizes the pre-trained encoder, incorporating color hints at the image level is straightforward: simply providing the desired colors hints directly within the grayscale conditioning image. We denote I^*_{gray} (see Figure 2) a grayscale image that contains colors hints and its corresponding encoding x^*_{gray}. During training, the conditioning image I^*_{gray} can contain a variable number of colored patches. In this case, we avoid providing text (both during training and testing) and set y to an empty string to avoid interference with the provided color hints.

During inference, a simple user interface permits any number of color hints (color prompts) on the image. We also employ the classifier-free guidance, incorporating the hints

$$\bar{\epsilon} = (1+w)\epsilon_\theta(x_t, t, x^*_{\text{gray}}) - w\epsilon_\theta(x_t, t, x_{\text{gray}}) \quad (4)$$

Figure 4 demonstrates that even within this color hints guidance, a number of variations are possible.

Reference Image Colorization. Propagating colors to a grayscale image from a reference image, can also be done as long as hints can be propagated in a meaningful way. To achieve this, we leverage recent advances in semantic matching [Luo et al. 2023; Tang et al. 2023] and semantic segmentation [Kirillov et al. 2023]. Given a reference image I_{ref}, semantic segmentation are extracted from each image. Then, within each matching semantic region, keypoint matches are computed against the gray scale image I_{gray}. Colors from the reference image are copied to the corresponding positions in the grayscale image, which is then provided as conditioning to the denoising model (See Figure 4.c). This approach builds upon existing color hints conditioning, inspired by [Huang et al. 2022], and aims at automating the color transfer between semantically similar objects, diverging from the standard definition of reference image conditioning in image colorization.

3.3 Attention for Text Edits and Multi-frame Colorization

During the denoising process, it is possible to extend the possibilities of the model by manipulating the spatial attention maps. We demonstrate this for text prompts guidance for object colorization and cross-frame color propagation for video colorization.

Object colorization through text. During the fine tuning of the model for colorization, text prompts are still provided. This means it is possible to guide the colorization through text input. To offer a finer, instance-level control over this process and avoid color leakage, we use the cross-attention maps associated with each object token. More specifically, we create a list of "color" words (i.e $\{red, green, \ldots\}$) which can be identified in the text prompts during inference.

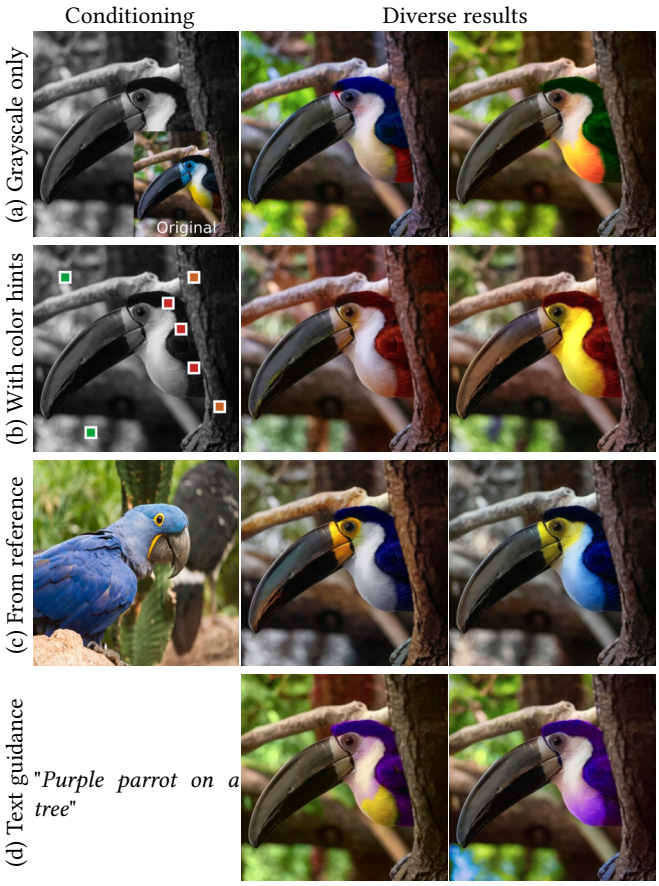

Figure 4: Diverse colorization results under various conditioning. (a) From a single gray scale image, several valid colorization can be sampled. (b) When color hints are provided (note the red and green pixels), the hints are respected while still offering variability in the colorization. (c) A reference image can be used for colorization hints: Note how the colors around the eyes and feathers match. (d) Finally, a text prompt can be used.

We use the example illustrated in Figure 3, to present the method without loss in generality. In the text input we match objects and associated colors. Given the two objects to colorize with specific colors, we create 3 parallel denoising processes with text prompts y, y_1 and y_2. The main denoising process has the prompt y where specific color prompts have been removed (to avoid color leakage). Then a text prompt is created for each object to colorize: y_1 and y_2 in this case. We note that we obtain the attention maps in a fashion similar to [Hertz et al. 2022] with the key difference that we do note require the compute expensive DDIM inversion [Song et al. 2020] or Null-Text Optimization [Mokady et al. 2023]. After each denoising step, the attention maps for prompted words are extracted and reformed into the latent masks, which is used to merge the latents into the main denoising process with prompt y. This denoising process with merging is repeated m times, after which the color information is integrated, and the parallel branches are not needed. Although m can be freely adjusted, we notice that

Table 1: Image colorization evaluation, COCO validation set. Bold, underlined and double-underlined formatting respectively indicate first, second and third performing method.

		FID↓	SSIM↑	LPIPS↓	PSNR↑	CF↑	ΔCF↓	CIE2K↓
1 sample (default)	DeOldify [Antic 2020]	10.79	**0.940**	**0.177**	**22.37**	20.4	18.11	**9.51**
	BigColor [2022]	9.02	0.910	0.215	20.33	33.92	17.85	11.45
	CT2 [2022a]	8.55	0.909	0.233	20.67	31.11	14.79	10.65
	DISCO [2022]	8.19	0.912	0.223	20.83	29.58	15.05	10.28
	UniColor [2022]	8.01	0.919	0.209	20.95	17.60	34.43	10.72
	ControlNet [2023]	8.47	0.902	0.241	19.64	**43.35**	15.81	12.41
	PBDiffusion [2023b]	8.31	0.918	0.221	20.24	36.21	14.03	11.83
	DDColor Large [2023b]	**7.35**	0.935	0.184	22.36	32.12	**11.22**	9.88
	VVFM (Ours)	7.92	0.930	0.201	20.90	34.43	13.01	10.25
Best of 5 samples	**VVFM** (Ours)	**7.29**	**0.942**	**0.162**	**23.11**	32.19	**9.44**	**8.53**
	PBDiffusion [2023b]	8.02	0.929	0.191	21.99	33.41	11.73	9.70
	UniColor [2022]	7.93	0.923	0.200	21.85	31.62	13.52	10.25
	ControlNet [2023]	8.18	0.914	0.214	21.21	**39.53**	12.75	10.24
	BigColor [2022]	8.68	0.915	0.223	20.44	32.13	15.56	10.95

setting it in range between $50-80\%$ of the the total number of steps consistently produces good results.

Cross-frame Color propagation for video colorization. Similarly to text, we leverage self attention maps to guide the colorization by paying attention to the appropriate regions, i.e., the other frames. We adopt a strategy similar to [Khachatryan et al. 2023] where self-attention operation in the model is repurposed to cross-frame attention. Given a key-frame k and a subsequent frame i the attention becomes for a given head

$$\text{CF-Attn}(Q^i, K^k, V^k) = \text{Softmax}\left(\frac{Q^i (K^k)^T}{\sqrt{c}}\right) V^k \tag{5}$$

where Q, K and V respectively denote queries, keys and values, following the notation introduced by [Rombach et al. 2022]. Considering Hue/Chrome as the predicted quantities, helps reducing the temporal stability issues associated with latent diffusion models. We can create longer temporally stable video, which means the key-frame k the model uses in cross-frame attention cannot remain the same, so after the pre-defined period, a new keyframe is chosen and its self-attention maps are used as the conditioning for the following frames. The experimental section evaluates various temporal propagation strategies and key-frame intervals. FInally, the described approach not only demands no further fine-tuning or training, but it also introduces no additional computations compared to the sampling of the same number of independent images, efficiently extending the existing model to video colorization.

4 RESULTS

We evaluate various aspects of the proposed model. This includes demonstrating the flexibility in terms of colorization, addressing several use cases including with and without user guidance and exploring multi-frame colorization. We provide extensive comparisons against state-of-the-art methods in the different settings.

4.1 Datasets and Implementation Details

Datasets. Similar to existing methods, for evaluation we use the COCO [Lin et al. 2014] validation set. We also use the validation set from the NTIRE video colorization challenge [Kang et al. 2023a].

Ground truth captions were already provided in the COCO dataset. For NTIRE dataset, we caption images using BLIP [Li et al. 2022]. Finally, to successfully train the model for the colorization task, it was essential to filter out grayscale images from the training sets. Those images deteriorate the quality of the colorization results, especially on historical black-and-white images. To achieve this, we again utilize image captioning to detect and remove grayscale or near-grayscale images from the training sets.

Implementation Details. The proposed solution should be compatible with any latent space diffusion model. However one of our key insight is to leverage a vision foundation model, and StableDiffusion 2.1 [Rombach et al. 2022; StabilityAI 2022] is the one we select as starting point. We train the model for 10k and 40k iterations for NTIRE and COCO respectively, using the batch size in range $[4, 8]$ with learning rate $1e-4$ and Adam optimizer. By introducing the text prompt dropout probability of 0.4 we increase the robustness of the model to the missing text prompts and stimulate the model to infer the colors from the grayscale directly or from the color hints, if provided. During training, image hints are randomly sampled as a number of patches of various sizes from the ground truth image which were left colorized on the conditioning grayscale image.

How to evaluate colorization? Similar to state-of-the-art we report common image metrics for evaluation (PSNR, SSIM and LPIPS [Zhang et al. 2018]). These are however not very suitable for the evaluation of colors, and it is common to include more representative metrics such Frechet inception distance (FID) [Heusel et al. 2017], CF [Hasler and Suesstrunk 2003] and CIEDE2000 [Alessi et al. 2014]. Following [Kang et al. 2023b], we also compute the difference, per-sample, in colorfulness (ΔCF) with respect to the ground truth. The objective is to take into account methods that create vivid colors which are unrealistic.

One issue with methods generating multiple outputs is the selection of a result for evaluation. Depending on the initial seed, some results can be worse than others and relying on a single sample is not sufficient to entirely judge the quality of these models. To better reflect this, whenever a model can produce multiple results, we additionally provide an evaluation row where 5 image results are randomly generated and the best sample is used for the evaluation. To determine this best sample, we compute several metrics: PSNR, MSE, SSIM, MS-SSIM, UQI [Wang and Bovik 2002], Relative Dimensionless Global Error (ERGAS) and Visual Information Fidelity (VIF) [Sheikh and Bovik 2006]. Each time a sample is the best performing according to one of the metrics, it receives one vote. The sample with most votes is used for the evaluation.

4.2 Grayscale Image Colorization

In comparisons to other models we limit ourselves to the most recent and top-performing methods. Qualitative comparison of the colorization results is presented in Figure 7.

The quantitative evaluation on COCO validation set is presented in Table 1. The rows are split in 2 groups. In the first group a single output is evaluated. For the deterministic methods this is their only output. For generative methods, we randomly sample the seed value once and use that 1 sample for evaluation. In the second group,

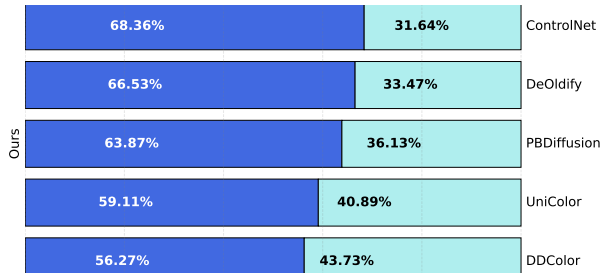

Figure 5: User study results: pairwise comparison between ours and other methods. (ELO in supplementary material).

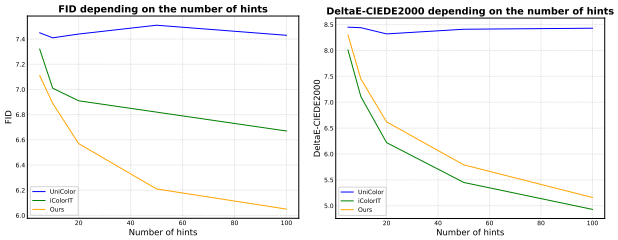

Figure 6: Hint colorization. Evaluation on the COCO dataset using FID and CIEDE2000.

generative methods are sampled 5 times, and the best is used for evaluation.

For the colorization task, it is common to consider that FID and ΔCF to be the most representative of the quality of the colorization results [Kang et al. 2023b]. While being extremely competitive with a single random sample (DDColor Large [Kang et al. 2023b] performs best in this case), our method largely outperforms all the existing methods when considering multiple samples (see Table 1). Similar conclusions are made with ImageNet. Note that we limited the comparisons in this case to the top performing models on COCO dataset.

For a task that is as subjective as colorization, it is important to consider the user evaluation, which we designed to compare against best-performing models. Similar to the numerical evaluation, we take into account the capacity of generative models, by creating multiple samples for each grayscale image. In a first filtering process, 3 users are asked to blindly select the best colorization for each grayscale image. After this filtering stage is done, we have a single color image for each method. The results of the survey are presented in Figure 5 where we clearly see a preference of the users for our method. Additional details about the survey can be found in supplementary material, including ELO scores [1978].

One aspect we also demontrate is the ability of our model to colorize image with various skin tones. A thorough analysis needs to be done before any real life applications, however since lightness contains a lot of information about the expected skin tone it seems reasonable to expect good results as illustrated in Figure 11.

Our proposed architecture adjustment and training strategy can be applied on any pre-trained diffusion model. To illustrate this we have implemented the proposed change on a different version of stable diffusion (SD1.4) and on two different architectures: SDXL [Podell et al. 2023] and Pixart-α [Chen et al. 2023]. Results in

Table 2: Evaluation using NTIRE Video Colorization validation set. We have two sets of comparisons, without and without using BistNet as a way to propagate color accross key-frames. Bold and underlined fonts respectively indicate first and second method.

	FID ↓	CDC ↓
Baseline (DeOldify)	47.15	0.00348
TCVC [2021]	46.5	**0.00309**
VCGAN [2022]	44.9	0.00337
DDColor [2023b]	34.88	0.06600
VVFM (without Cross Attention)	**26.8**	0.01246
VVFM	33.0	0.00343
Baseline (DeOldify) + BistNet	48.53	0.00342
DDColor + BistNet	36.29	0.00306
VVFM (without Cross Attention) + BistNet	**32.0**	0.00351
VVFM + BistNet	37.7	**0.00301**

Figure 13 show that all the different models are able to successfully colorize grayscale images.

4.3 User Guided Colorization

Color Hints. Our model accepts the color hints by users, which provides a convenient way to quickly colorize images in a desirable color. To evaluate this aspect we use the COCO dataset, where we keep a number of color hints provided from 5 to 100. Here we compared to 2 other state-of-the-art methods: UniColor [Huang et al. 2022] and iColorIT [Yun et al. 2023]. Evaluation is presented in Figures 6 using FID and CIEDE2000 with respect to ground truth images. We show qualitative results in Figure 8, where our method which produces realistic colors even in regions without hints.

Reference Image Colorization. Our reference colorization pipeline is inspired by UniColor [Huang et al. 2022] which propagate hints from the reference image to the grayscale. Besides Figures 4 and 10 that illustrate this, additional examples are provided in the supplementary material.

Text Guided Colorization. Text guided colorization pipeline is set against two works from state-of-the-art: DiffColor [2023] and UniColor [2022]. The first propose a complex model built around latent diffusion, and uses context text embedding optimization. The second makes use of CLIP to design a text-to-hints model and then transforming the text to local color hints. In our case by simply considering the appropriate attention map during the denoising process, we able to achieve vivid and realistic colorization that match the text prompt as illustrated in Figure 9.

4.4 Multi-frame Colorization

For qualitative comparison we used the model trained on COCO image dataset and showcased its performance on several clips (see Figure 14), comparing against the state-of-the-art and demonstrating the benefits of cross-frame attention. Overall both TCVC [2021] and VCGAN [2022] result in significant colorization artifacts (color bleeding, etc). Best performing single frame method DDColor [2023b], demonstrates interesting colorization results but without extensive temporal stability. This problem can be mitigated by using BiSTNet [Yang et al. 2022] to temporally smoothen the colorization

over the sequence. Using our model without cross-frame attention (VVFM-without cross-attention) produces colorful but temporally inconsistent results, which is expected. Adding the cross frame attention, maintains temporally stable results.

Numerical evaluation on NTIRE video colorization dataset use two official metrics, namely FID and *Color Distribution Consistency Index* (CDC) [Liu et al. 2021]. For this quantitative evaluation (and only here) we finetune the model on the training set of the NTIRE challenge. We then run the evaluations on the validation dataset. It is possible to use BiSTNet exemplar-based colorization for better temporal consistency, and we evaluation both DDCcolor and our model with and without this additional processing. The results are presented in Table 2. It is interesting to note that BiSTNet consistently improves the temporal stability measure at the cost of reduced colorfulness. Additionally, we do the ablation study on the keyframe period of the cross-attention mechanism, varying it from 5 to 50 frames. The results are presented in supplementary material.

5 DISCUSSION

In this work we have presented how to use a vision foundation model to address several aspects of the colorization problem. For single shot grayscale image colorization the model achieves competitive results against most recent works. Additionally multiple samples can be created, in which case the model achieves state-of-the-art performance. More importantly the model is versatile and can be used on a wide variety of scenarios: hint based colorization, text guidance and video. By leveraging the cross attention maps, we demonstrate that we can achieve good results on text guidance and temporal coherency with a simpler approach than existing works. Pushing for further improvement would require to overcome some of the limitations associated with the latent diffusion model: such as the limited capacity of the autoencoder that can lead to imperfect colorization for small objects. Additionally a finer control over the text guided colorization or handling strong motion would require more work, and these are active areas of research in the community.

REFERENCES

2017. America in color. https://www.smithsonianchannel.com/shows/america-in-color. Accessed: 2024-01-20.

Naofumi Akimoto, Akio Hayakawa, Andrew Shin, and Takuya Narihira. 2020. Reference-based video colorization with spatiotemporal correspondence. *arXiv preprint arXiv:2011.12528* (2020).

PJ Alessi, M Brill, J Campos Acosta, E Carter, R Connelly, J Decarreau, R Harold, R Hirschler, B Jordan, C Kim, et al. 2014. Colorimetry-part 6: CIEDE2000-colour-difference formula. *ISO/CIE* (2014), 11664–6.

Jason Antic. 2020. DeOldify. https://github.com/jantic/DeOldify. Accessed: 2024-01-20.

Yunpeng Bai, Chao Dong, Zenghao Chai, Andong Wang, Zhengzhuo Xu, and Chun Yuan. 2022. Semantic-sparse colorization network for deep exemplar-based colorization. In *European Conference on Computer Vision*. Springer, 505–521.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf* 2 (2023), 3.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 18392–18402.

Hernan Carrillo, Michaël Clément, Aurélie Bugeau, and Edgar Simo-Serra. 2023. Diffusart: Enhancing Line Art Colorization with Conditional Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3485–3489.

Zheng Chang, Shuchen Weng, Yu Li, Si Li, and Boxin Shi. 2022. L-CoDer: Language-based colorization with color-object decoupling transformer. In *European Conference on Computer Vision*. Springer, 360–375.

Zheng Chang, Shuchen Weng, Peixuan Zhang, Yu Li, Si Li, and Boxin Shi. 2023. L-CAD: Language-based Colorization with Any-level Descriptions using Diffusion Priors. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Jianbo Chen, Yelong Shen, Jianfeng Gao, Jingjing Liu, and Xiaodong Liu. 2018. Language-based image editing with recurrent attentive models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 8721–8729.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. 2023. PixArt-α: Fast Training of Diffusion Transformer for Photorealistic Text-to-Image Synthesis. arXiv:2310.00426 [cs.CV]

Zezhou Cheng, Qingxiong Yang, and Bin Sheng. 2015. Deep colorization. In *Proceedings of the IEEE international conference on computer vision*. 415–423.

Ethan Coen, Joel Coen, and Roger Deakins. 2001. Painting with Pixels. https://www.imdb.com/title/tt28520463/. Accessed: 2024-01-20.

Aditya Deshpande, Jiajun Lu, Mao-Chuang Yeh, Min Jin Chong, and David Forsyth. 2017. Learning diverse image colorization. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 6837–6845.

Zhi Dou, Ning Wang, Baopu Li, Zhihui Wang, Haojie Li, and Bin Liu. 2021. Dual color space guided sketch colorization. *IEEE Transactions on Image Processing* 30 (2021), 7292–7304.

Arpad E Elo and Sam Sloan. 1978. The rating of chessplayers: Past and present. *(No Title)* (1978).

Yuki Endo, Satoshi Iizuka, Yoshihiro Kanamori, and Jun Mitani. 2016. Deepprop: Extracting deep features from a single image for edit propagation. In *Computer Graphics Forum*, Vol. 35. Wiley Online Library, 189–201.

David Hasler and Sabine E Suesstrunk. 2003. Measuring colorfulness in natural images. In *Human vision and electronic imaging VIII*, Vol. 5007. SPIE, 87–95.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2015. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision*. 1026–1034.

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), 1–16.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-or. 2022. Prompt-to-Prompt Image Editing with Cross-Attention Control. In *The Eleventh International Conference on Learning Representations*.

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. *Advances in neural information processing systems* 30 (2017).

Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

Jonathan Ho and Tim Salimans. 2022. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598* (2022).

Zhitong Huang, Nanxuan Zhao, and Jing Liao. 2022. Unicolor: A unified framework for multi-modal colorization with transformer. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 1–16.

Satoshi Iizuka and Edgar Simo-Serra. 2019. Deepremaster: temporal source-reference attention networks for comprehensive video enhancement. *ACM Transactions on Graphics (TOG)* 38, 6 (2019), 1–13.

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 (ToG)* 35, 4 (2016), 1–11.

Varun Jampani, Raghudeep Gadde, and Peter V Gehler. 2017. Video propagation networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 451–461.

Xiaozhong Ji, Boyuan Jiang, Donghao Luo, Guangpin Tao, Wenqing Chu, Zhifeng Xie, Chengjie Wang, and Ying Tai. 2022. ColorFormer: Image colorization via color memory assisted hybrid-attention transformer. In *European Conference on Computer Vision*. Springer, 20–36.

Xiaoyang Kang, Xianhui Lin, Kai Zhang, Zheng Hui, Wangmeng Xiang, Jun-Yan He, Xiaoming Li, Peiran Ren, Xuansong Xie, Radu Timofte, et al. 2023a. NTIRE 2023 video colorization challenge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1570–1581.

Xiaoyang Kang, Tao Yang, Wenqi Ouyang, Peiran Ren, Lingzhi Li, and Xuansong Xie. 2023b. DDColor: Towards Photo-Realistic Image Colorization via Dual Decoders. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 328–338.

Bahjat Kawar, Shiran Zada, Oran Lang, Omer Tov, Huiwen Chang, Tali Dekel, Inbar Mosseri, and Michal Irani. 2023. Imagic: Text-based real image editing with diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6007–6017.

Bingxin Ke, Anton Obukhov, Shengyu Huang, Nando Metzger, Rodrigo Caye Daudt, and Konrad Schindler. 2023. Repurposing Diffusion-Based Image Generators for Monocular Depth Estimation. *arXiv preprint arXiv:2312.02145* (2023).

Levon Khachatryan, Andranik Movsisyan, Vahram Tadevosyan, Roberto Henschel, Zhangyang Wang, Shant Navasardyan, and Humphrey Shi. 2023. Text2video-zero: Text-to-image diffusion models are zero-shot video generators. *arXiv preprint*

arXiv:2303.13439 (2023).

Geonung Kim, Kyoungkook Kang, Seongtae Kim, Hwayoon Lee, Sehoon Kim, Jonghyun Kim, Seung-Hwan Baek, and Sunghyun Cho. 2022. Bigcolor: colorization using a generative color prior for natural images. In *European Conference on Computer Vision*. Springer, 350–366.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643* (2023).

Manoj Kumar, Dirk Weissenborn, and Nal Kalchbrenner. 2021. Colorization transformer. *arXiv preprint arXiv:2102.04432* (2021).

Gustav Larsson, Michael Maire, and Gregory Shakhnarovich. 2016. Learning representations for automatic colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14*. Springer, 577–593.

Anat Levin, Dani Lischinski, and Yair Weiss. 2004. Colorization using optimization. In *ACM SIGGRAPH 2004 Papers*. 689–694.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.

Jianxin Lin, Peng Xiao, Yijun Wang, Rongju Zhang, and Xiangxiang Zeng. 2023. DiffColor: Toward High Fidelity Text-Guided Image Colorization with Diffusion Models. *arXiv preprint arXiv:2308.01655* (2023).

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: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

Hanyuan Liu, Minshan Xie, Jinbo Xing, Chengze Li, and Tien-Tsin Wong. 2023a. Video Colorization with Pre-trained Text-to-Image Diffusion Models. *arXiv preprint arXiv:2306.01732* (2023).

Hanyuan Liu, Jinbo Xing, Minshan Xie, Chengze Li, and Tien-Tsin Wong. 2023b. Improved Diffusion-based Image Colorization via Piggybacked Models. *arXiv preprint arXiv:2304.11105* (2023).

Yihao Liu, Hengyuan Zhao, Kelvin C. K. Chan, Xintao Wang, Chen Change Loy, Yu Qiao, and Chao Dong. 2021. Temporally Consistent Video Colorization with Deep Feature Propagation and Self-regularization Learning. arXiv:2110.04562 [cs.CV]

Grace Luo, Lisa Dunlap, Dong Huk Park, Aleksander Holynski, and Trevor Darrell. 2023. Diffusion Hyperfeatures: Searching Through Time and Space for Semantic Correspondence. *arXiv preprint arXiv:2305.14334* (2023).

Varun Manjunatha, Mohit Iyyer, Jordan Boyd-Graber, and Larry Davis. 2018. Learning to color from language. *arXiv preprint arXiv:1804.06026* (2018).

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2023. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6038–6047.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.

Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. 2022. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*. 1–10.

Saurabh Saxena, Charles Herrmann, Junhwa Hur, Abhishek Kar, Mohammad Norouzi, Deqing Sun, and David J Fleet. 2023a. The Surprising Effectiveness of Diffusion Models for Optical Flow and Monocular Depth Estimation. *arXiv preprint arXiv:2306.01923* (2023).

Saurabh Saxena, Abhishek Kar, Mohammad Norouzi, and David J Fleet. 2023b. Monocular depth estimation using diffusion models. *arXiv preprint arXiv:2302.14816* (2023).

Hamid R Sheikh and Alan C Bovik. 2006. Image information and visual quality. *IEEE Transactions on image processing* 15, 2 (2006), 430–444.

Min Shi, Jia-Qi Zhang, Shu-Yu Chen, Lin Gao, Yu-Kun Lai, and Fang-Lue Zhang. 2023. Reference-based deep line art video colorization. *IEEE Transactions on Visualization & Computer Graphics* 29, 06 (2023), 2965–2979.

Jiaming Song, Chenlin Meng, and Stefano Ermon. 2020. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502* (2020).

StabilityAI. 2022. *Stable Diffusion v2.1*. https://stability.ai/news/stablediffusion2-1-release7-dec-2022

Jheng-Wei Su, Hung-Kuo Chu, and Jia-Bin Huang. 2020. Instance-aware image colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7968–7977.

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

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

Patricia Vitoria, Lara Raad, and Coloma Ballester. 2020. Chromagan: Adversarial picture colorization with semantic class distribution. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2445–2454.

Hanzhang Wang, Deming Zhai, Xianming Liu, Junjun Jiang, and Wen Gao. 2023. Unsupervised deep exemplar colorization via pyramid dual non-local attention. *IEEE Transactions on Image Processing* (2023).

Zhou Wang and Alan C Bovik. 2002. A universal image quality index. *IEEE signal processing letters* 9, 3 (2002), 81–84.

Tomihisa Welsh, Michael Ashikhmin, and Klaus Mueller. 2002. Transferring color to greyscale images. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*. 277–280.

Shuchen Weng, Jimeng Sun, Yu Li, Si Li, and Boxin Shi. 2022a. CT 2: Colorization transformer via color tokens. In *European Conference on Computer Vision*. Springer, 1–16.

Shuchen Weng, Hao Wu, Zheng Chang, Jiajun Tang, Si Li, and Boxin Shi. 2022b. L-code: Language-based colorization using color-object decoupled conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 2677–2684.

Yanze Wu, Xintao Wang, Yu Li, Honglun Zhang, Xun Zhao, and Ying Shan. 2021. Towards vivid and diverse image colorization with generative color prior. In *Proceedings of the IEEE/CVF international conference on computer vision*. 14377–14386.

Menghan Xia, Wenbo Hu, Tien-Tsin Wong, and Jue Wang. 2022. Disentangled Image Colorization via Global Anchors. *ACM Transactions on Graphics (TOG)* 41, 6 (2022), 204:1–204:13.

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)*. IEEE, 1887–1891.

Zhongyou Xu, Tingting Wang, Faming Fang, Yun Sheng, and Guixu Zhang. 2020. Stylization-based architecture for fast deep exemplar colorization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9363–9372.

Dingkun Yan, Ryogo Ito, Ryo Moriai, and Suguru Saito. 2023. Two-Step Training: Adjustable Sketch Colourization via Reference Image and Text Tag. In *Computer Graphics Forum*. Wiley Online Library.

Yixin Yang, Zhongzheng Peng, Xiaoyu Du, Zhulin Tao, Jinhui Tang, and Jinshan Pan. 2022. BiSTNet: Semantic Image Prior Guided Bidirectional Temporal Feature Fusion for Deep Exemplar-based Video Colorization. *arXiv preprint arXiv:2212.02268* (2022).

Wang Yin, Peng Lu, Zhaoran Zhao, and Xujun Peng. 2021. Yes," Attention Is All You Need", for Exemplar based Colorization. In *Proceedings of the 29th ACM international conference on multimedia*. 2243–2251.

Jooyeol Yun, Sanghyeon Lee, Minho Park, and Jaegul Choo. 2023. iColoriT: Towards Propagating Local Hints to the Right Region in Interactive Colorization by Leveraging Vision Transformer. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 1787–1796.

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/CVF conference on computer vision and pattern recognition*. 8052–8061.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

Richard Zhang, Phillip Isola, and Alexei A Efros. 2016. Colorful image colorization. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part III 14*. Springer, 649–666.

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. *arXiv preprint arXiv:1705.02999* (2017).

Jiaojiao Zhao, Jungong Han, Ling Shao, and Cees GM Snoek. 2020. Pixelated semantic colorization. *International Journal of Computer Vision* 128 (2020), 818–834.

Yuzhi Zhao, Lai-Man Po, Wing Yin Yu, Yasar Abbas Ur Rehman, Mengyang Liu, Yujia Zhang, and Weifeng Ou. 2022. VCGAN: video colorization with hybrid generative adversarial network. *IEEE Transactions on Multimedia* (2022).

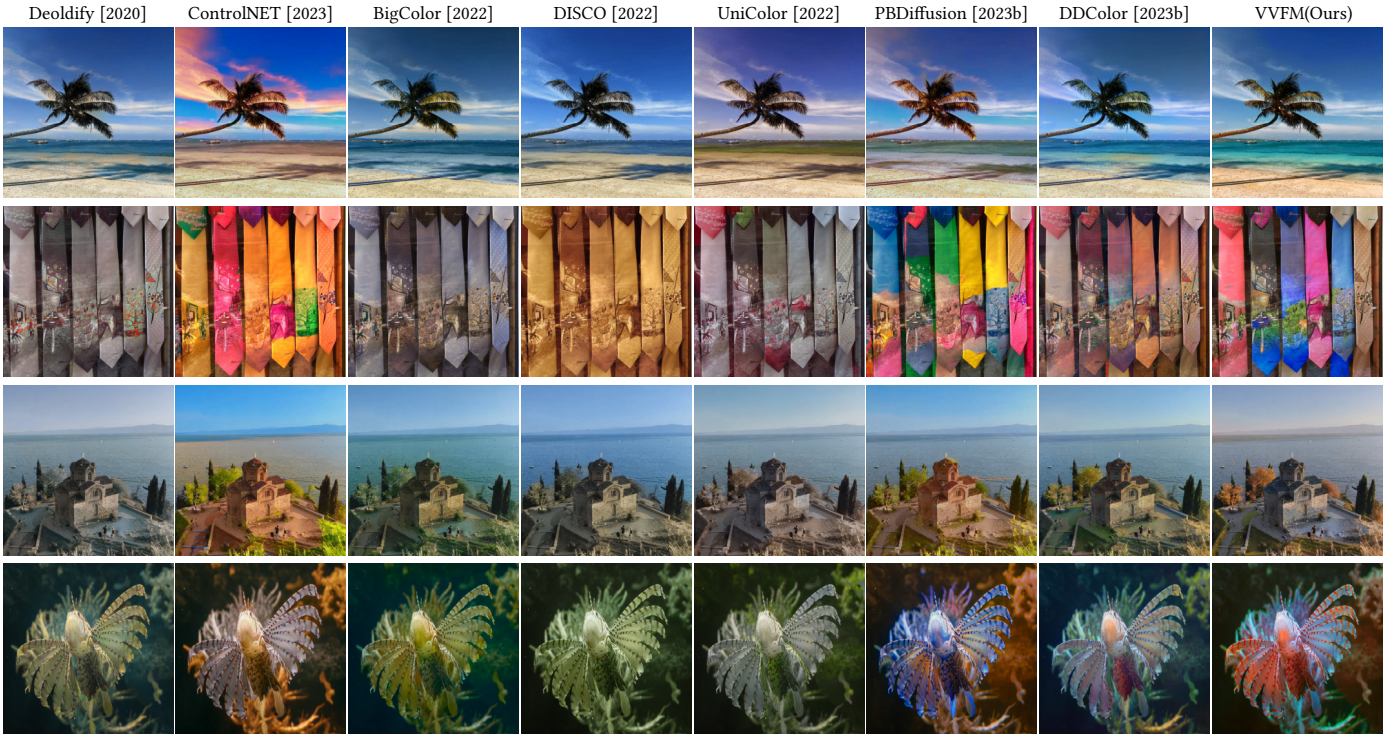

Figure 7: Qualitative comparisons of image colorization methods.

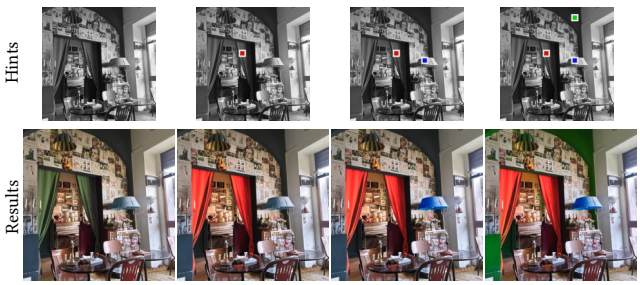

Figure 8: Qualitative results for hint based colorization.

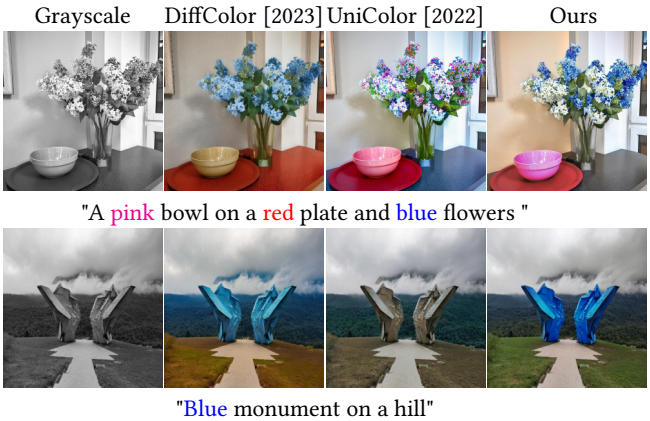

Figure 9: Comparisons for text guided colorization.

Figure 10: Examples of reference colorization.

Figure 11: Example of image colorization of different skin tones.

Figure 12: Examples of historical image colorization.

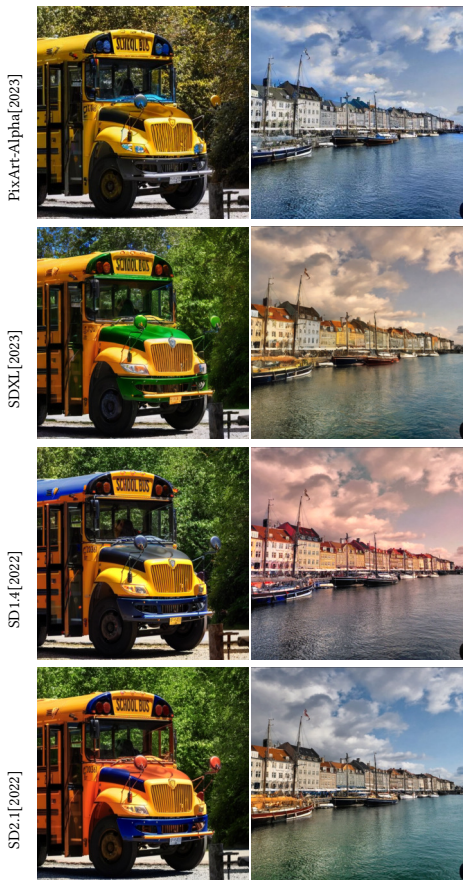

Figure 13: Examples of image colorization for different pre-trained model backbones.

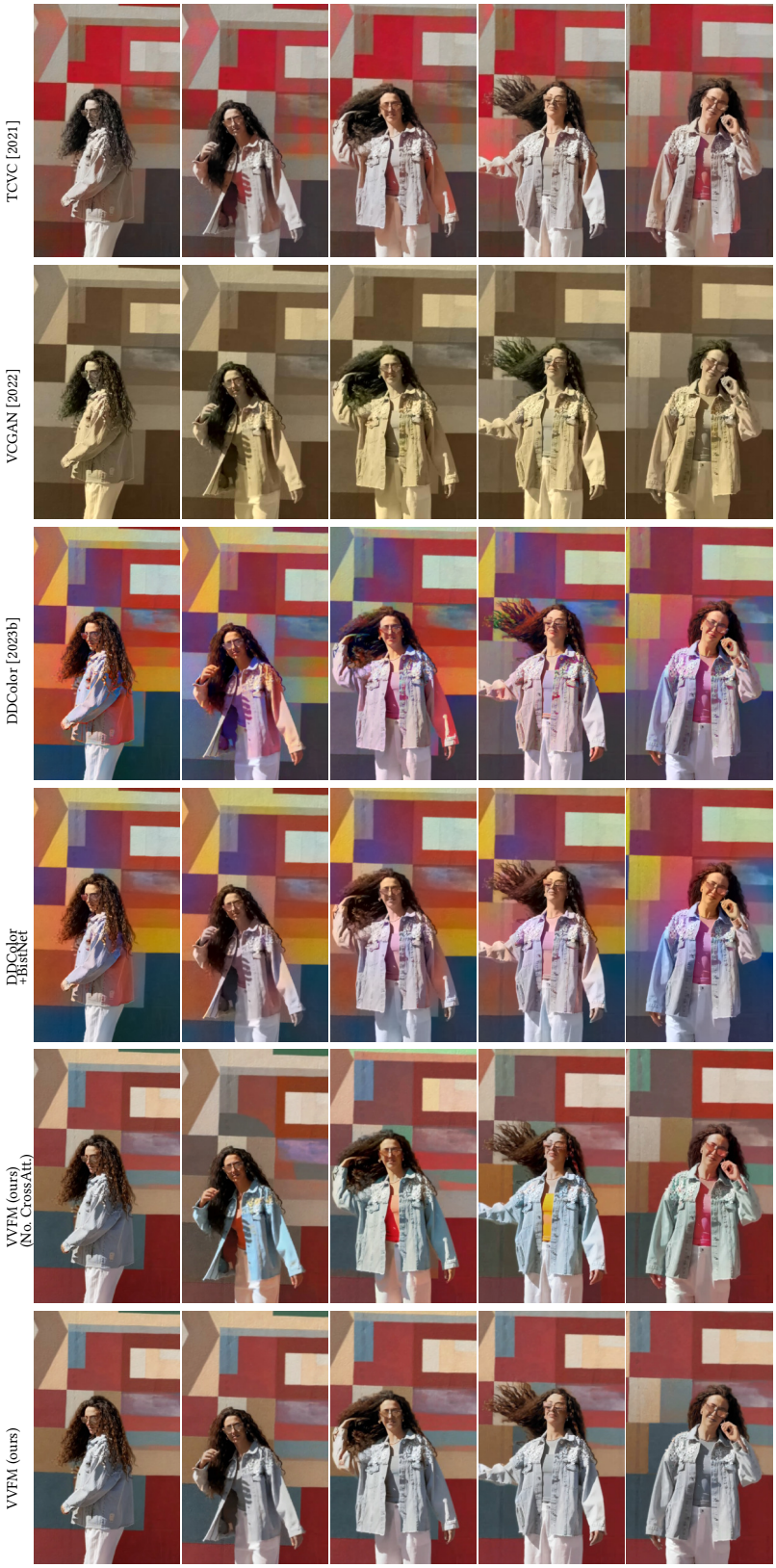

Figure 14: Qualitative comparisons of video colorization methods.