Content-Aware Image Color Editing with Auxiliary Color Restoration Tasks

Yixuan Ren[1]*, Jing Shi[2], Zhifei Zhang[2], Yifei Fan[2], Zhe Lin[2], Bo He[1], Abhinav Shrivastava[1]

[1]University of Maryland, College Park [2]Adobe Research

{yxren,bohe,abhinav}@cs.umd.edu, {jingshi,zzhang,yifan,zlin}@adobe.com

Abstract

Diversified image color editing is typically modeled as a multimodal image-to-image translation (MMI2IT) problem with an impact on multiple applications such as photo enhancement and retouching. Although previous GAN-based algorithms successfully generate diverse edits with clear control, we observe two issues remaining: Firstly, they tend to apply the same color style to all kinds of input images when sampling with the same style latent, regardless of the input content and scenes. Secondly, they usually edit the color style globally in an image and fail to keep each semantic region and instance in harmonic colors individually. We attribute these issues to the strong independence between the style latent and the condition image in most current MMI2IT methods.

To edit the raw image into a more harmonic direction with awareness of its global content and local semantics, we introduce auxiliary color restoration tasks by reducing the input color information and training jointly. We also increase the model's capacity and enrich the noise's locality with diffusion models. Furthermore, we propose a new set of metrics to measure the content-awareness of MMI2IT models, that is, how the generated style is adaptive to the input image's content. Our model is also extensible to several downstream applications including exemplar-based color editing and language-guided color editing, without imposing extra demands on the already trained model.

1. Introduction

We investigate the task of diversified image color editing. Given an input image, the goal is to generate multiple edited images in different photographic color styles while keeping their original content, structure and texture. Image color editing is essential for various downstream applications such as image enhancement [10, 26] and retouching [3, 13]. The problem is typically formulated as a multimodal image-to-image translation (MMI2IT) task, where the model is trained to edit a raw image in diverse color styles given different randomly sampled noise. Early work has difficulty extracting the multimodal capability. BicycleGAN [30] and DivCo [14] are limited to the simple network design. SpaceEdit [23] leverages the latent space of the StyleGAN2 [9] and achieves excellent multimodal generation capability with clear control.

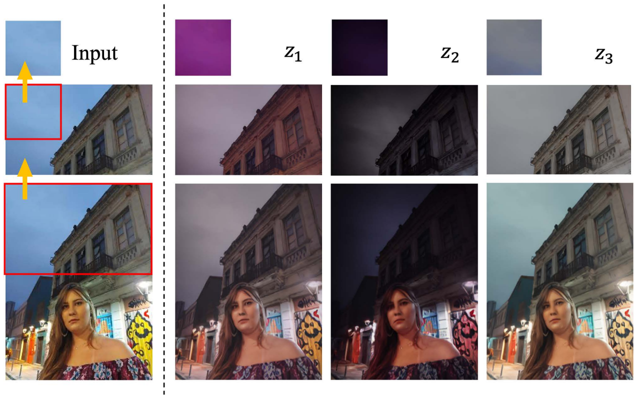

Figure 1. Our model generates content-aware color editing results. Given different patches cropped out of the same image as the input, our model applies adaptive color styles to different semantics to produce harmonic and aesthetic output. For example, sky usually has the most artistic diversity, while buildings and human portraits have their preferred color tone ranges. When a randomly sampled latent $z \sim \mathcal{N}(0, 1)$ edits a sky image into very extraordinary colors (1st row), it yet adapts less colorful but brighter tones to buildings, and least colorful but warmer tones to humans (2nd and 3rd rows). In the meantime the major editing direction of each latent z is retained the same for smooth controllability.

In practice, non-expert users are more used to the raw photos being tuned and polished automatically and smartly in batch. Therefore we expect the model to be able to perform semantic-adaptive color editing mainly based on the input image. Although SpaceEdit can achieve diversified and controllable color editing results, we find its understanding of the input semantics still insufficient for advanced cases.

*Part of this work was done when Yixuan was an intern at Adobe Research.

When randomly sampling from noise as the style latent, we observe that the same or very similar color styles are applied to all kinds of input images, no matter if they match well. However, images of natural scenes and human portraits can prefer different color tones, such as the examples in Fig. 1, where the sky patch can be extremely artistic, but humans are suitable for milder color tone to keep the face clear. Besides, we also find SpaceEdit usually only performs global editing, turning the entire frame toward one direction. A uniform color tone is forced to be applied to every region and instance without considering the semantics, resulting in unappealing output and even artifacts. For example, while polishing the color of the sky to more blue or red in an artistic way, people's clothing may also be dyed, especially when the original color is shallow. Figs. 3, 4 and 5 provide some such samples.

We suspect reasons causing the aforementioned observations are low content awareness in both data and model aspects. Different types of raw images may share commonly welcomed color editing styles as well as holding their own preferences dependent on comprehensive and complicated circumstances. Thus the ground truth editing styles in the data lack an evident correlation with the input images that are easy to capture. Besides, many samples have relatively subtle color editing and thus increase the risk of a trivial shortcut that learns an almost identity transform plus some random perturbations.

From the perspective of models, although SpaceEdit has leveraged the upgraded structure of CoModGAN [29] connecting the input image's feature to the style sampling network as well to conditionally co-modulate, this path might not be fully utilized when the styles don't strongly correlated with the input images in the training data. The vanilla GAN objective also contributes to this issue as an indirect constraint, especially when the target domain is not far from the source domain. SpaceEdit reports that a conditional discriminator feeding on the input image contrastively is critical to the performance compared to an unconditional one. Furthermore, the spatial alignment for the style sampling and modulation process of StyleGAN-based methods [22] is yet insufficient for all detailed cases. This limits the model to perform fine-grained regional editing.

Diffusion models [2,5] provide a bigger latent space with a closer spatial alignment of the same shape of the condition image. For I2I tasks, the latent noise is concatenated with the input image and thus sets up a dense correlation between the input content and editing style spatially. Besides, the direct L1/L2 loss is a more strict constraint to enforce the learned mapping not to be trivial. Therefore, we employ diffusion models to improve the content-aware co-modulation and spatial alignment between the editing styles and input content.

We further introduce multiple auxiliary color restoration tasks that enforce the model to learn and utilize more information on the input content implicitly. The tasks include colorization, which predicts an image's chrominance given its luminance, and its complementary task to infer the luminance from the chrominance. By reducing the input information in color channels the auxiliary tasks require the model a deeper understanding of the input semantics to produce reasonable output, forcing the model to take the input image's features into consideration when determining the output style, instead of mapping the latent noise to a dedicated color style and simply applying it on any raw image. The colorization task aims for a proper color tone and its complementary tackles brightness to avoid under- or over-exposure outcomes.

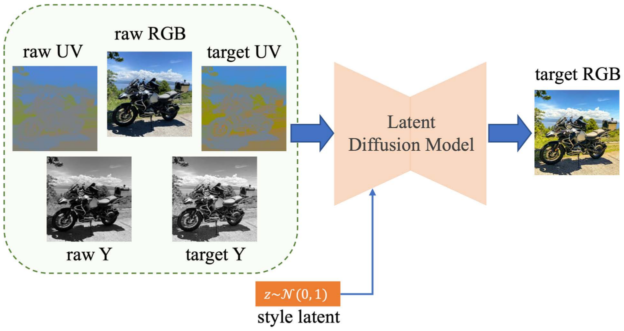

Figure 2. Our model is built upon Latent Diffusion Models . There are 5 tasks being trained jointly: the main task has the input of raw RGB, and the auxiliary color restoration tasks are fed with extra input conditions decomposed from YUV color space, for enhancing content-adaptive luminance and chrominance editing respectively. In every training iteration one of the input types is randomly chosen as the input.

In summary, our contributions are:

- We introduced a diffusion-based model for multimodal diversified image color editing, and enhanced its content awareness via auxiliary color restoration tasks with joint training strategy.
- We analyzed the non-adaptiveness issues in existing MMI2IT algorithms and proposed a new framework of metrics to measure content awareness quantitatively.
- We conducted massive experiments to validate that our model learns to always apply reasonable color styles adaptive to the input semantics. We further extended its functionality for various downstream applications.

2. Related Work

Multimodal Image Editing. Multimodal image editing task aims to edit an input image with multiple diverse styles given random noise latent. BicycleGAN [30] first performs multimodal image-to-image translation trained on paired data. [7, 11, 12, 16] extended it to unpaired training data.

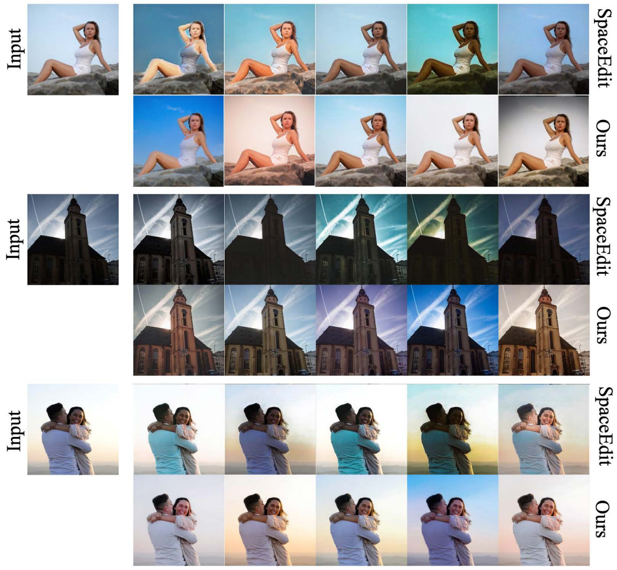

Figure 3. Qualitative comparisons between our model and SpaceEdit on randomly sampled color editing styles. Although SpaceEdit sometimes generates fancier and more extraordinary colors, it risks a lot on the essential quality and harmony, and results in heavy artifacts, such as dark humans and buildings, gapped edges around foreground objects, dyed clothes, etc. Our model instead understands the input semantics and processes them spatially, assigning adaptive color styles to different background and foreground content with also good diversity.

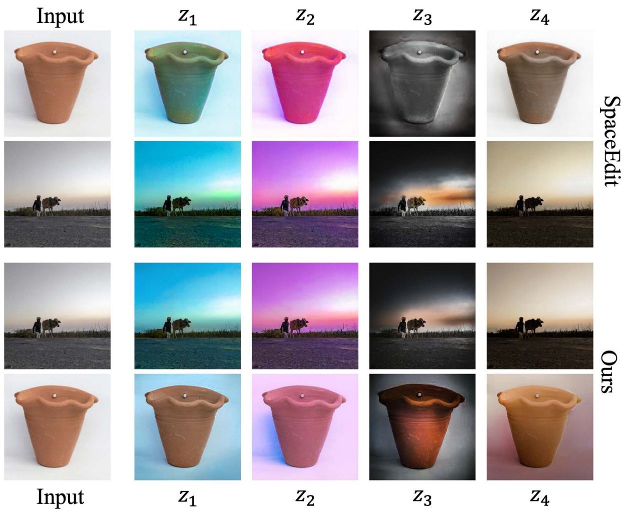

Figure 4. Content-awareness comparisons given the same style latent z. This is an extreme case containing two distinct scenarios: sky and landscape tend to have striking artistic color editing, while for close objects and indoor scenes most editings are only in a small range. Both SpaceEdit and our model are able to perform distinctive and colorful editings. However, SpaceEdit applies them to other input images without adaptation, and thus dyes and twists the clay pot in an unnatural and unpleasing way. By contrast our model manages to retain the reasonable scene and object and edits them via adding various lighting and shadowing in corresponding colors. This ensures that for any randomly sampled style latent $z \sim \mathcal{N}(0, 1)$, our model can always produce proper color editing, diversified as well as adaptive to the input content. This is thus friendly and reliable for common users without any guidance instruction or post-selection required.

StyleGAN [8] injects the latent code into multiple layers of the generator via normalization and modulation. It has been demonstrated to have well-disentangled latent space and thus has been applied to many downstream image editing tasks. CoModGAN [29] further addresses the importance of the condition image modulating the editing style and embedding both conditional and stochastic style representations via co-modulation. SpaceEdit [23] leverages this structure for multimodal image color editing pretraining. However, we notice that there still exist artifacts in their results, especially when the condition image is less relied on than the style vector when determining the editing style.

Diffusion Models. The diffusion model is a type of score-based generative models that iteratively denoising the step to map a Gaussian noise to the sample within the empirical data distribution. [5, 24] provided the foundational work for subsequent diffusion models. Denoising Diffusion Implicit Models (DDIMs) [25] achieves deterministic generation by omitting the noise sampled for each middle timesteps. Diffusion with Classifier-Free Guidance (CFG) [6] lifts the need for an external classifier which in addition guides a pretrained diffusion model to generate images following certain criteria. Latent Diffusion Models (LDMs) [18] that transformed the diffusion model into a high-performance generator for general conditional inputs by introducing a cross-attention layer. It significantly reduces the computational cost while almost preserving the generative quality, and thus plenties of work and pipelines derived from it for various applications, such as Stable Diffusion (SD) for large-scale text-to-image generation and many other image-to-image editing tasks. Diffusion models have been demonstrated to be indeed powerful in image synthesis and editing tasks. Therefore, we leverage LDMs for diversified image color editing tasks.

3. Image Color Editing with Diffusion Models

We formulate the diversified image color editing task as a multimodal image-to-image translation problem: we learn a mapping $x = f(z, y)$ where x is the target edited image, y is the raw input image as the condition and z is random noise as the style latent.

Fig. 2 shows the main architecture of our proposed system. Diffusion models work as iteratively mapping noise from standard Gaussian distribution into the empirical data

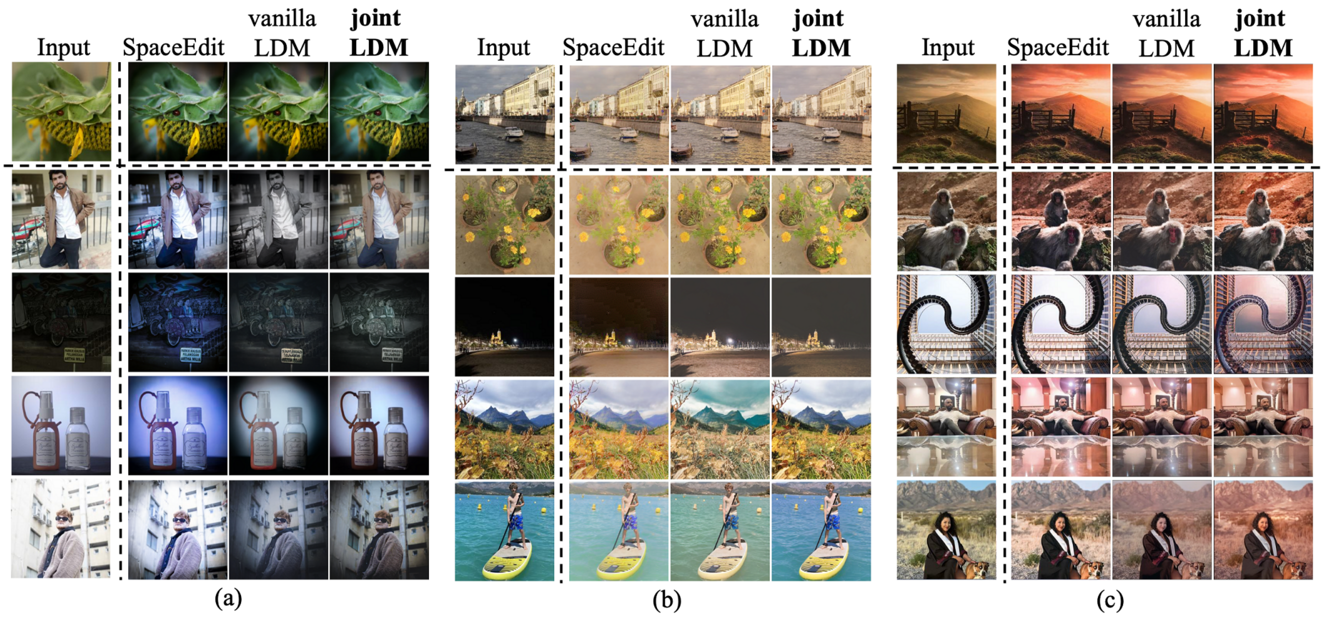

Figure 5. Qualitative comparisons based on exemplar-based color editing. In subfigure (a), the reference style is to tune toward colder color tone, and to add a vignetting effect. SpaceEdit simply adjusts all images into blue color even for human faces. Vanilla LDM fails to apply a consistent vignetting region around the margins and sometimes turns the main objects too dark or greyscale. Our jointly trained model (**4th column**) is able to produce the most accurate and proper effect as in the reference row. In subfigure (b), our model adapts brightness to the diverse input content, preventing overexposure on both day and night scenes, as well as preserves the original color hue faithfully. In subfigure (c), our model produces the most vibrant red color to the background uniformly, without touching the background and foreground texture. Note that on the sofa and table case (4th row), our model is able to recognize the reflection on the shiny tabletop surface, and dims the luminance a bit compared to the directly lit ceiling and walls. More results are shown in the appendix.

distribution via a denoising Gaussian process. To achieve image-to-image translation, we adopt conditional diffusion model [18, 20], where the conditional input image is concatenated with the noise at every denoising step. We employ latent diffusion models (LDMs) [21] for computational efficiency.

A LDM works in two stages. In the first stage, an encoder $\mathcal{E}$ and a decoder $\mathcal{D}$ are trained to bi-directionally convert images between the pixel space and the latent space. In the second stage, we train a denoising UNet [19] ϵ_θ with parameters θ, and it only needs to predict the low-dimensional latent representation of the target image. At each timestep t, we concatenate the input condition together with the noise z_t following [18, 20], and the training objective is

$$L_{LDM} := \mathbb{E}_{\mathcal{E}(y),x,\epsilon\sim\mathcal{N}(0,1),t} \left\|\epsilon - \epsilon_\theta((z_t \oplus \mathcal{E}(x)), t)\right\|_n , \tag{1}$$

where $\oplus$ denotes concatenation operation, and n can be either 1 or 2 indicating L_1 or L_2 loss. The appendix provides more details on the model mechanism.

4. Auxiliary Color Restoration Tasks

The lack of understanding of the semantics when the model generates non-adaptive color editing styles inspires us to enhance its content awareness. Our goal is that the same z will adapt to different image content and regions to get a realistically and aesthetically edited image. We propose two auxiliary tasks to achieve our goal: colorization (chrominance completion) and luminance completion.

4.1. Joint Training of Colorization Task

The colorization task [20] maps a grayscale image to a colorful image, and we observed that the assigned color is highly dependent on the input content and semantics in real-world data. For example, the palettes of living creatures and artificial objects diverge a lot. Mammals and fish also have different skin colors due to their species. Clothing colors vary among summer shorts and winter coats because of their different materials and their usual surroundings. These instances require the model to understand the semantics of the input image thoroughly to generate proper colors. Hence we propose to train our color editing task with the colorization task jointly.

We leverage YUV [15] color space, which has been widely applied to colorful video and image compression and transmission, to decompose the chrominance (UV) from luminance (Y), that is, the grayscale of the color image. The most straightforward option is to use the grayscale of the target image to predict its colorful format. Moreover, we can also input the grayscale of the raw image and still predict its edited colorful image. This becomes harder because the model additionally needs to learn to transfer the luminance from the raw image to the target image before colorizing it. The full editing path can also be decomposed as colorizing on raw images first and then transferring the luminance changes toward target images.

In every training step, raw RGB or target Y or raw Y images are randomly selected and input, and the reconstruction loss is calculated for all output. The output target has to always be the same otherwise the joint training will be unstable and inconstant.

4.2. Chrominance to Luminance Completion

We find that our model with above colorization tasks sometimes outputs under- or over-exposure images, which indicates that it is still less capable to produce a proper luminance range. To address this issue, we further design an auxiliary task of chrominance to luminance completion. Following the YUV color space, we set the Y channel of the image to 127.5 out of 255 and map the color space back to RGB space. The degraded image only remains chrominance information and loses the luminance signal. Then the model takes it in and is expected to recover the complete RGB target with the original luminance. With these tasks jointly trained, our model is promoted to comprehend the chrominance space and estimates the appropriate luminance and thus benefiting our color editing task with improved luminance quality. Similar to the colorization tasks, we use both the target UV images and the raw UV images as the input, and they are randomly chosen to compose a batch. Reconstruction losses are calculated between all output and target RGB images.

5. Downstream Applications

Our model is primarily designed and trained for the case that given randomly sampled noise, it is able to always generate aesthetic and proper style of color editing with the awareness of the input semantics, which simplify users' operations and complexity and serves as an automated pipeline for common users needless of any instructions in practice. In addition to this, our model can also be applied to more downstream applications including exemplar-based and language-guided color editing, as well as still provide diversified and content-adaptive results.

Table 1. Quantitative results of our models and variants. *Ours Baseline* refers to the vanilla LDM version and *Ours Joint* refers to upgrades with auxiliary tasks and joint training as checked for each item. Our models outperform the state-of-the-art methods on both generative fidelity and diversity. The auxiliary tasks are denoted by their input data format, for example, x_Y refers to colorization on target images and y_{UV} the chrominance to luminance completion on raw images.

Approach	Auxiliary Tasks				FID↓	LPIPS↑	$\sigma_{\times 10^{-2}}$ ↑
	x_Y	y_Y	x_{UV}	y_{UV}			
SpaceEdit					7.752	0.159	1.289
SpaceEdit Joint	✓	✓	✓	✓	11.715	0.163	1.082
Ours Baseline L1					7.103	0.096	0.518
Ours Joint L1	✓				7.075	0.101	0.610
Ours Joint L1	✓	✓			7.055	0.125	0.847
Ours Joint L1	✓	✓	✓		7.043	0.147	1.180
Ours Joint L1 [main]	✓	✓	✓	✓	6.988	0.153	1.205
Ours Baseline L2					7.644	0.189	1.621
Ours Joint L2	✓	✓	✓	✓	7.815	**0.220**	**2.052**
Ours L1 Perturb Aug					**6.538**	0.069	0.353

Table 2. The quantitative results of the proposed Content-Awareness Metrics (CAMs). Our main model with joint auxiliary tasks outperforms other competitors on these metrics. More details, including the self-validations for the metrics design, are in the appendix.

Approach	CAM-1↑	CAM-2		CAM-3	
		F-stat↑	p-val	F-stat↑	p-val
SpaceEdit (60k data)	0.2467	15.093	0.0181	2.4619	0.0243
SpaceEdit	0.2816	16.257	0.0122	2.9985	0.0253
Ours LDM baseline	0.3571	21.553	0.0148	4.9446	0.0204
Ours Joint L1 [main]	**0.5767**	**28.343**	0.0142	**7.1684**	0.0278

Table 3. User study results.

Approach	5-star Ratings↑	Binary Choice↑
SpaceEdit	3.523	226
Ours Joint L1 [main]	**4.062**	**361**

5.1. Latent Space Inversion

An invertible diffusion probabilistic model encoder (DPM-Encoder) is proposed by [27] for denoising diffusion probabilistic models (DDPMs). It maps generated images back into their full latent space including all timesteps. We apply this idea to our latent conditional diffusion models with a latent encoder $\mathcal{E}$ and additional input conditions. For our models that translate raw images y to edited images x, the complete latent noise is $z_t = \epsilon_t \oplus \mathcal{E}(y)$ at timestep t, and the full noise space $\{\epsilon_t\}_{t=T,\ldots,1}$ can be inverted by

$$\begin{aligned} &x_1, \ldots, x_{T-1}, x_T \sim q(x_{1:T}|x_0), \\ &\epsilon_t = (x_{t-1} - \mu_\theta(x_t, \mathcal{E}(y), t))/\sigma_t \ominus \mathcal{E}(y), \end{aligned} \tag{2}$$

where $\ominus$ denotes splitting as inverse concatenation.

5.2. Exemplar-Based Color Editing

Given a pair of raw and edited images, our model can invert them to acquire the editing style in the latent diffusion model's noise space following the above inversion method, and then apply it to another new input image to transfer the color style editing. The results can be viewed in Figs. 4 and 5, where both the style preservation and the content-awareness of our model during the transfer is highlighted.

5.3. Language-Guided Color Editing

Without being trained with text captions describing the specific color editing style, our model can also perform open-vocabulary language-guided image color editing by leveraging the CLIP-guided diffusion inference process.

[2] introduced classifier-guidance for diffusion model inference, and [1] extended it by replacing the conventional classifier with a pretrained CLIP [17] model to apply language guidance on the sampling steps of diffusion models. To demonstrate the full capacity of our model, we follow the pipeline to perform CLIP-guided image color editing. In particular, with a pretrained CLIP model consisting of an image encoder E_{img} and a text encoder E_{txt}, and the input text prompt c for instruction, the new μ for each denoising step will be updated with the gradient of the CLIP loss w.r.t. the predicted image of that step:

$$\begin{gathered}\hat{\mu_\theta}((y_t, \mathcal{E}(x), t)|c) = \mu_\theta(y_t, \mathcal{E}(x), t) + s \cdot \sigma_t \nabla_{y_t} L_{CLIP}, \\ L_{CLIP} = E_{img}(\mathcal{D}(y_t)) \cdot E_{txt}(c),\end{gathered} \tag{3}$$

where s is the guidance scale.

6. Experiments

6.1. Datasets and Implementation Details

Datasets. We follow [23] to use the Adobe Discover dataset collected from the Adobe Discover website, which is all contributed by Lightroom users uploading their images and edits. The edits mainly concentrate on color and tone retouching, without changing the original content or geometry. We only use one edit per image to form one-to-one pairs, and we use all data from the community without any cherrypicking for only higher-quality samples. Compared to [23] which uses 60k samples, more data are collected to reach 500k in total. A larger amount of data represents a richer and more precise real-world distribution, as well as demands a higher capacity of the model to handle. We split the full dataset into 400k/50k/50k pairs for train/val/test sets respectively.

Experiment Settings. All our and compared experiments are trained with images of resolution 256 × 256 following [23]. We use a pretrained VQGAN model for $\mathcal{E}$ and $\mathcal{D}$ in the LDM with the downsampling factor $f = 4$. The UNet in the LDM is trained from scratch.

Metrics. We evaluate all methods with three aspects of computational metrics: generative fidelity, diversity, and our proposed content-awareness.

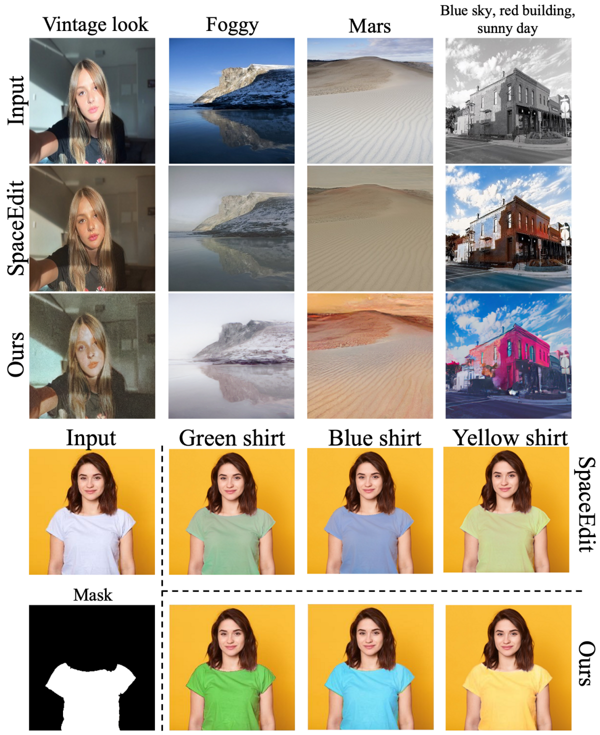

Figure 6. Language-guided editing results with or without masks. Our edited colors are more realistic and adaptive to the specific environments in terms of both luminance and chrominance benefited from our auxiliary tasks. Our vintage portrait maintains the original cold color tone as well as slightly lowers the clarity with film grain effect to simulate analog photography. Our foggy mountain also preserves the original brightness with saturation faded. For the Mars desert and red building, our model produces more vibrant color. In the masked task, our clothing has more accurate and vivid colors considering its material and the surrounding lighting.

FID [4]: it measures the gap between the distributions of the generated and ground truth data using features calculated by Inception Network. It is the lower the better. This number reflects the generative fidelity of the editing result.

LPIPS [28]: We generate 16 different edited images for each input image with randomly sampled noise, then we compute the mean pair-wise LPIPS distance among the output images. It measures the diversity of generated styles based on the same set of input images.

Variance σ: Similar to LPIPS, given one input image we generate 16 edited images, and then compute the pair-wise pixel L1 distance among the output images. The variance of these L1 distances is used as this metric. This also measures the generative diversity.

Furthermore, to quantitatively measure and compare the content-awareness property across multimodal generative models, we propose a new framework of metrics based on the correlations between the input image's content and their

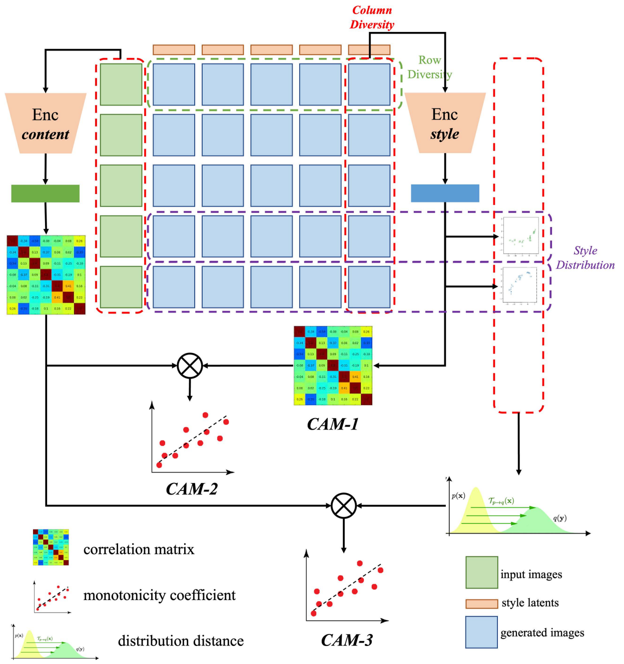

Figure 7. Illustration of our proposed framework of metrics for content-awareness in MMI2IT tasks. Conventional diversity is calculated over each row, measuring the output various given different latent noise for the same image. Our proposed metrics are calculated over each column to reflect how the same latent noise have adaptive editing style for different input content. Detailed definitions and formulas of each metric variant are illustrated in the appendix.

output editing styles. We name them Content-Awareness Metrics (CAMs). There are three variants as shown in Fig. 7, *i.e. CAM-1*, *CAM-2* and *CAM-3*. It follows the inference settings for calculating the conventional diversity above, to generate a grid of output images corresponding to a set of input images and a set of style latents. Then we calculate the diversity of all images generated from the same style latent and different input images, *i.e.* along each column instead of each row, as *CAM-1*. *CAM-2* and *CAM-3* further calculate the correlations between the input images' content and the output images' styles, or their distributions, given the same style latent or the same set of style latents. These reflect how well the model produces content-aware editing styles in MMI2IT tasks. Please refer to the appendix for more details.

Following the settings in [18], since FID is dependent on the amount of samples, all metrics are calculated and averaged on 16 random sets each containing 5000 random samples from the test split. Note that this number might vary among other previous work so we run and test their models with our uniform settings.

6.2. Diversified Image Color Editing

The quantitative results of our methods and the variants are listed in Tab. 1. We mainly compared with SpaceEdit, the previous state-of-the-art approach. We conduct our full model with both L1 and L2 losses. Our L2 LDM without auxiliary tasks has already surpassed in both fidelity and diversity metrics than SpaceEdit. For the L1 LDM, our fully jointly trained model shows further boosted fidelity and comparable diversity. Tab. 2 lists the results of our proposed CAMs. It also shows that our main model has the leading positions on all three variants of the content-awareness measurement. And our ablation models also have intermediate values above previous GAN-based methods. More experiments and results are provided in the appendix.

Fig. 3 shows the qualitative comparison. Although SpaceEdit has a higher diversity score than some of our ablation variants, it is traded by the generative quality and sacrifices aesthetics. Our results are more realistic with detailed shapes and textures faithfully preserved after appealing editing, as well as comparable and reasonable diversity.

6.3. Content Awareness Analysis

Our model displays clear patterns of content awareness in color editing. Using the techniques of paired latent noise space inversion and exemplar-based color editing in Secs. 5.1 and 5.2, we conduct content-awareness comparisons in Figs. 4 and 5.

In Fig. 4, we first select reference editing images for the sky and landscape input, and acquire identical color styles by exemplar-based color transfer (2nd and 3rd rows) to each model. Then the corresponding style latent z is applied to a new input image, a close shot of a pottery pot in a simple indoor background for each model (1st and 4th rows). SpaceEdit insists on applying the identical global color tone to all raw images regardless of their content, and our model adapts them according to the major and minor objects while still preserving the intended editing color style as faithfully as possible.

In each subfigure of Fig. 5, the first row contains an initial input image and three output images of SpaceEdit, the vanilla LDM, and our joint model. In this row, all the reference output are managed to be visually identical via paired latent noise space inversions. Then the same color style latent z of each model is respectively applied to new input images and generates corresponding edited images as paired color style transfer. Our joint model is shown to have the advantage of the most consistent color style following the style latent as well as the most adaptive to various input images respectively for aesthetics over others.

6.4. User Studies

We conduct two user studies to validate that our model yields superior images in terms of both generative qual-

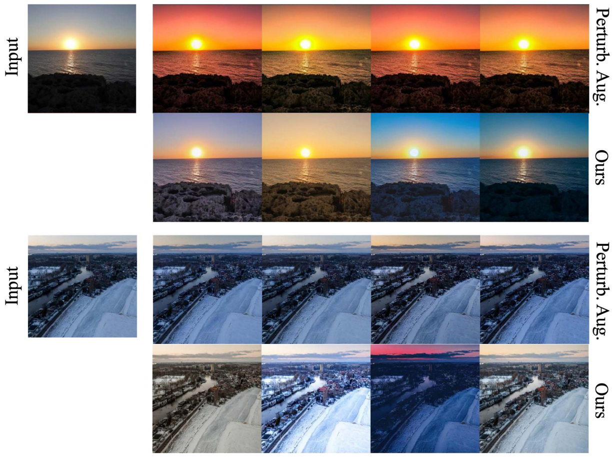

Figure 8. Mode collapse when training with color perturbation augmentations on the ground truth data. The model easily overfits on a single data point and significantly sacrifices diversity.

ity and semantic-adaptiveness to previous methods. In the studies, each user is shown one raw input image, and two sets of output images edited by our model and SpaceEdit respectively. Each set is sampled with 16 different random noise as an automated recommendation. In the first study, users are instructed to rate the overall quality of the two sets of edited images as a whole from the aspects of aesthetics, realism and diversity out of 5 stars. In the second study, users are asked to simply choose one from the two sets as they think better than the other following the same criteria. We released 300 different questionnaires of randomly selected raw images and their editing results on Amazon MTurk and each questionnaire is performed by 2 users. There are 577 and 587 valid responses for the two studies. The results are listed in Tab. 3. Our model surpasses SpaceEdit with an average rating of 4.062 over 3.523 and 361 choices in total over 226.

6.5. Ablation Studies

Auxiliary Tasks. Our models are dominatingly boosted by the auxiliary tasks. As listed in Tab. 1, our baseline model built upon vanilla LDM is progressively enhanced with each auxiliary task incorporated. According to the results, the colorization and chrominance to luminance completion tasks on the target image domain contribute the most to the lift of the performance. We observed from the experiments that with the colorization task, the chrominance in generated images becomes more content-adaptive. And with the chrominance to luminance completion task, the brightness of output is more dependent on the input image than the latent noise. The other two tasks also help by decomposing the whole edit path into multiple segments to tackle individually. Notably, the chrominance to luminance completion task substantially increases the generative diversity.

Trivial data augmentation by color perturbations. We also investigate if traditional data augmentations on chrominance and luminance can seemingly reach similar performance as competitive to our auxiliary tasks. We applied four types of image color perturbations on the input images: hue, saturation, brightness, and contrast, covering our auxiliary tasks' manipulation of images. The quantitative results are listed in Tab. 1. It shows that this data augmentation scheme can significantly elevate the fidelity, *i.e.* generating much closer results to the ground truth distribution; however, the diversity is meanwhile largely lost and it in fact results in mode collapse. Fig. 8 visualizes the comparison. It demonstrates that trivial color perturbations are too simple for the model to tackle and form shortcuts.

6.6. Language-Guided Color Editing

Our results and comparisons on open-vocabulary language-guided color editing are displayed in Figs. 6. We performed two downstream tasks, with and without a region mask respectively. In both tasks, our model generates more accurate and vibrant colors according to the text prompts. Our method benefits from the joint auxiliary tasks of colorization and luminance completion to adapt the most accurate and attractive color tone and brightness. More results are shown in the appendix.

7. Conclusion and Future Work

We explored the diversified image color editing task with diffusion models and auxiliary color restoration tasks. The diffusion model exhibits outstanding color editing ability and enables to match better with the target image distribution than GAN-based methods. Joint training with the colorization and the chrominance to luminance completion tasks boosts the deeper semantic understanding and content awareness for the model. Also, the latent of our model represents a unique color editing style that can be applied to a different image with adaptive adjustment based on the new image semantics. We believe that the content-awareness is an important property for many MMI2IT models and tasks to measure and has practical meaning to real-world applications and user interactions.

Our future work is to extend our color editing framework to support broader color-related tasks, including image harmonization, composition and *etc*. Besides, extended auxiliary color tasks, especially those whose data are easier to expand to a larger scale, such as colorization, may further boost the main tasks with limited and expensive data. We hope that our work serves as a versatile joint training framework that can not only promote the color editing field but also facilitate various color-related tasks.

References

[1] Katherine Crowson. Clip guided diffusion hq 256x256.ipynb. https://colab.research.google.com/drive/1lQJXS55mRyN7TWDomTNo8tbeUErO7rMJ, 2021. 6

[2] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2, 6

[3] Jingwen He, Yihao Liu, Yu Qiao, and Chao Dong. Conditional sequential modulation for efficient global image retouching. In *European Conference on Computer Vision*, pages 679–695. Springer, 2020. 1

[4] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017. 6

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2, 3

[6] Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. 3

[7] Xun Huang, Ming-Yu Liu, Serge Belongie, and Jan Kautz. Multimodal unsupervised image-to-image translation. In *Proceedings of the European conference on computer vision (ECCV)*, pages 172–189, 2018. 2

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3

[9] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 1

[10] Han-Ul Kim, Young Jun Koh, and Chang-Su Kim. Pienet: Personalized image enhancement network. In *European Conference on Computer Vision*, pages 374–390. Springer, 2020. 1

[11] Hsin-Ying Lee, Hung-Yu Tseng, Jia-Bin Huang, Maneesh Singh, and Ming-Hsuan Yang. Diverse image-to-image translation via disentangled representations. In *Proceedings of the European conference on computer vision (ECCV)*, pages 35–51, 2018. 2

[12] Hsin-Ying Lee, Hung-Yu Tseng, Qi Mao, Jia-Bin Huang, Yu-Ding Lu, Maneesh Singh, and Ming-Hsuan Yang. Drit++: Diverse image-to-image translation via disentangled representations. *International Journal of Computer Vision*, 128(10):2402–2417, 2020. 2

[13] Jie Liang, Hui Zeng, Miaomiao Cui, Xuansong Xie, and Lei Zhang. Ppr10k: A large-scale portrait photo retouching dataset with human-region mask and group-level consistency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 653–661, 2021. 1

[14] Rui Liu, Yixiao Ge, Ching Lam Choi, Xiaogang Wang, and Hongsheng Li. Divco: Diverse conditional image synthesis via contrastive generative adversarial network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16377–16386, 2021. 1

[15] Joe Maller. Rgb and yuv color. *FXScript Reference*, 2003. 4

[16] Sanghyeon Na, Seungjoo Yoo, and Jaegul Choo. Miso: Mutual information loss with stochastic style representations for multimodal image-to-image translation. *arXiv preprint arXiv:1902.03938*, 2019. 2

[17] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 6

[18] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3, 4, 7

[19] Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical image computing and computer-assisted intervention*, pages 234–241. Springer, 2015. 4

[20] Chitwan Saharia, William Chan, Huiwen Chang, Chris Lee, Jonathan Ho, Tim Salimans, David Fleet, and Mohammad Norouzi. Palette: Image-to-image diffusion models. In *ACM SIGGRAPH 2022 Conference Proceedings*, pages 1–10, 2022. 4

[21] Chitwan Saharia, Jonathan Ho, William Chan, Tim Salimans, David J Fleet, and Mohammad Norouzi. Image super-resolution via iterative refinement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. 4

[22] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1532–1540, 2021. 2

[23] Jing Shi, Ning Xu, Haitian Zheng, Alex Smith, Jiebo Luo, and Chenliang Xu. Spaceedit: learning a unified editing space for open-domain image editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022. 1, 3, 6

[24] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 3

[25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 3

[26] Yuda Song, Hui Qian, and Xin Du. Starenhancer: Learning real-time and style-aware image enhancement. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4126–4135, 2021. 1

[27] Chen Henry Wu and Fernando De la Torre. Unifying diffusion models' latent space, with applications to cyclediffusion and guidance. *arXiv preprint arXiv:2210.05559*, 2022. 5

[28] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018. 6

[29] Shengyu Zhao, Jonathan Cui, Yilun Sheng, Yue Dong, Xiao Liang, Eric I Chang, and Yan Xu. Large scale image completion via co-modulated generative adversarial networks. *arXiv preprint arXiv:2103.10428*, 2021. 2, 3

[30] Jun-Yan Zhu, Richard Zhang, Deepak Pathak, Trevor Darrell, Alexei A Efros, Oliver Wang, and Eli Shechtman. Toward multimodal image-to-image translation. *Advances in neural information processing systems*, 30, 2017. 1, 2