DreamLight: Towards Harmonious and Consistent Image Relighting

Yong Liu[1]*, Wenpeng Xiao[2]*, Qianqian Wang[2] , Junlin Chen[2] , Shiyin Wang[2] , Yitong Wang[2] , Xinglong Wu[2] , Yansong Tang[1]†
[1]Tsinghua Shenzhen International Graduate School, Tsinghua University [2]ByteDance Inc.
liuyong23@mails.tsinghua.edu.cn, tang.yansong@sz.tsinghua.edu.cn

Abstract

We introduce a model named DreamLight for universal image relighting in this work, which can seamlessly composite subjects into a new background while maintaining aesthetic uniformity in terms of lighting and color tone. The background can be specified by natural images (image-based relighting) or generated from unlimited text prompts (text-based relighting). Existing studies primarily focus on image-based relighting, while with scant exploration into text-based scenarios. Some works employ intricate disentanglement pipeline designs relying on environment maps to provide relevant information, which grapples with the expensive data cost required for intrinsic decomposition and light source. Other methods take this task as an image translation problem and perform pixel-level transformation with autoencoder architecture. While these methods have achieved decent harmonization effects, they struggle to generate realistic and natural light interaction effects between the foreground and background. To alleviate these challenges, we reorganize the input data into a unified format and leverage the semantic prior provided by the pretrained diffusion model to facilitate the generation of natural results. Moreover, we propose a Position-Guided Light Adapter (PGLA) that condenses light information from different directions in the background into designed light query embeddings, and modulates the foreground with direction-biased masked attention. In addition, we present a post-processing module named Spectral Foreground Fixer (SFF) to adaptively reorganize different frequency components of subject and relighted background, which helps enhance the consistency of foreground appearance. Extensive comparisons and user study demonstrate that our DreamLight achieves remarkable relighting performance.

1 Introduction

With the rapid development of generative models in recent years [1, 2, 3, 4], image composition has received increasing attention owing to its capacity for controlled generation [5, 6, 7]. However, since the implanted foreground and the new background originate from different sources, this discrepancy can easily lead to an unrealistic perception of the composite image. In this paper, we focus on a challenging task named universal image relighting, which aims to seamlessly composite a subject into a new background while maintaining realism and aesthetic uniformity in terms of lighting and color tone. The background can either be specified by provided natural images (image-based relighting), or be created from unlimited text prompts (text-based relighting). This task ensures that users and digital elements coexist naturally within any environment and has wide application in virtual reality and intelligent editing for the film and advertising industries [8, 9, 10, 11, 12, 13].

*Equal Contribution
†Corresponding author

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

 https://doi.org/10.52202/085713-1211

Figure 1: Our DreamLight is a unified image relighting model capable of performing relighting guided by arbitrary text or natural background images without any other prior such as HDR maps. The top left of each set of figures is the original picture of the foreground, the bottom left is the given condition, and the right is the generated result.

Early researches about relighting tend to focus on the image condition, *e.g.*, image harmonization and portrait relighting, while with scant exploration into text-based scenarios. Portrait relighting methods [14, 15, 16, 17] mainly take the idea of physics-guided design that explicitly models the image intrinsics and formation physics. They decompose the input images into several components and leverage a phased pipeline to incrementally learn image components such as surface normals and albedo maps. Some harmonization methods [18, 19] also take similar disentanglement idea. Despite the promising results achieved by these approaches, obtaining training data with pairs of high-quality relighting images and their corresponding intrinsic attributes from light stage is expensive and difficult. Most harmonization methods [20, 21, 22, 23] take this task as an image-to-image translation paradigm and propose to perform pixel-level transformation based on autoencoder architecture. However, they lack object semantic guidance and tend to overlook distinct variations in illumination. Recently, some works [24, 8, 25] exploit the powerful semantic modeling ability of generative diffusion models [1] to enhance relighting effect, but their light source relies on the given environment maps and such maps are not always feasible to acquire in real world. IC-Light [26] is proposed to address both image-based and text-based relighting for natural images without introducing other signal guidance. It imposes light alignment training strategy to enhance the light consistence. However, IC-Light takes two separate models for different conditions and does not tailor the structure. Simply concatenating foreground, background, and noise limits the model's understanding of foreground-background interactions, leading to severe color bleeding and foreground distortions in some scenarios.

To address the above challenges, we present a model named DreamLight that can perform both image-based and text-based relighting. In DreamLight, to generate natural interaction effects of light and color tone between the foreground and the background, we propose a Position-Guided Light Adapter (PGLA), which condenses light information from different directions in the background into several groups of light query embeddings and selectively modulates the foreground area with direction-biased masked attention. In addition, we present an effective Spectral Foreground Fixer (SFF) as a post-processing module to enhance the consistency of foreground appearance and avoid subject distortion. Based on the wavelet transform, SFF is trained to learn dynamic calibration coefficients for high-frequency textures of input foreground and relighted low-frequency light. By adaptively reorganize these information, SFF can output stunningly consistent foreground. Besides, to facilitate the training of our model, we develop different kinds of data generation processes, *e.g.*, 3D rendering and training relighting lora, to produce diverse training samples.

We have performed extensive quantitative and qualitative comparisons. Experiment results demonstrate that our DreamLight exhibits superior generalization and performance on the universal relighting of natural images. Related ablations also prove the rationality and effectiveness of the proposed designs. Furthermore, we observe that thanks to the unified learning of text and image conditions in a single model, our DreamLight can generate results with the guidance of both conditions.

https://doi.org/10.52202/085713-1211

Our contributions can be summarized as follows:

- We propose a model named DreamLight for universal image relighting, which can seamlessly composite a subject into a new background with either image or text conditions. We also develop high-quality data generation process to benefit the training of our model.
- We introduce a Position-Guided Light Adapter (PGLA), which is designed to enable foreground elements at different positions to interact with background light from various directions in a tendentious manner for generating more natural lighting effects.
- We propose a Spectral Foreground Fixer (SFF) module to adaptively reorganize different frequency components of the input and relighted subjects, which helps enhance the consistency of foreground textures.

2 Related Work

Image-based Relighting: There are two principal sets of related image-based relighting methods. The first is portrait relighting approaches [15, 14, 16, 27, 28, 29, 30, 8, 17, 31]. They mainly take the idea of physics-guided model design that typically involve the intermediate prediction of surface normals, albedo, and a set of diffuse and specular maps with ground truth supervision. To achieve that, some methods rely on the paired training data acquired with the light stage system [32] and a target HDR environment map as the external light source. However, the dependence on light stage data and HDR maps incurs substantial data collection cost and significantly limits their implementation in real-world situations, where obtaining HDR maps may not always be feasible. Some methods [33, 31] employ multi-stage frameworks. As a result, the accuracy and performance of these systems hinge on the precision of each individual stage. This makes the entire process complex and susceptible to errors that could propagate throughout these intermediate steps. The second is image harmonization methods that aim to match the color statistics of the foreground object with those of the background for natural composition [18, 20, 22, 23, 34, 21, 19, 35, 36, 37, 38]. These methods tend to take this task as an end-to-end image-to-image translation paradigm, where the network is trained to predict a harmonized image from the input composite. Some works [39, 40] collect pixel-aligned paired data by color transfer or altering foreground color in real images with pre-designed or learned augmentations. While these methods have achieved decent harmonization effects, they struggle to generate realistic and natural light interaction effects between the foreground and the background. Recently, some works [24, 8, 25] exploit generative models [1] to enhance relighting effect, but most of them still rely on the given environment maps for lighting guidance, which limits the potential application scenarios. IC-Light [26] proposes to perform pure natural image relighting by imposing light alignment training strategy and achieves excellent performance for scenarios with strong lighting. However, its simple network design limits the model's performance and is prone to generating severe color bleeding. To alleviate above issues, we propose DreamLight designed for natural images without any additional prior source. A position-guided light adapter is introduced to boost the reasonable interaction between subjects and backgrounds, thus contributing to generating natural and harmonious lighting effect.

Text-based Relighting: Currently, there is relatively less research focused on text-based relighting. Although Lasagna [41] takes text as input, the text serves to indicate the light direction rather than to describe the target background. The pioneer IC-Light [26] utilizes two separate models to handle image-based and text-based relighting. Actually, some text-guided inpainting methods [42, 43] can achieve similar effect. However, they tend to preserve the original lighting and color of the subject, which may result in unnatural results. Conversely, our DreamLight achieves powerful relighting performance of natural images using a single model for both image-based and text-based situations.

3 Method

3.1 Overview

Figure 2 illustrates the pipeline of our DreamLight. It takes a triplet of foreground image, background image, and text prompt as input. For image-based relighting, the text prompt is set to "blend these two images". For text-based relighting, the background image is designed as an all-black image. The **green** and **brown** lines show the specific processes of image-based and text-based relighting, respectively. In detail, we first utilize a pretrained segmentation model [26, 44] to extract the subject

 https://doi.org/10.52202/085713-1211

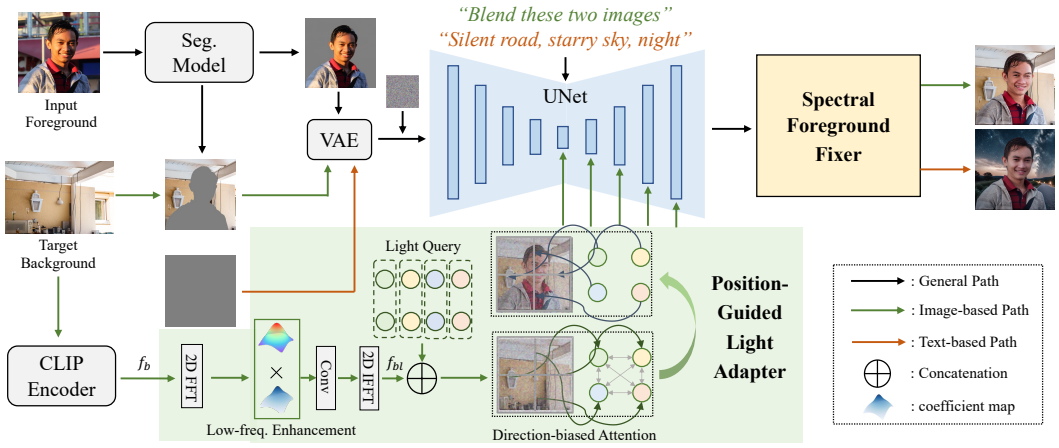

Figure 2: Pipeline of our DreamLight. It takes the foreground image, background image, and text prompt as input. The masked foreground and background images are encoded by pretrained VAE and concatenated with the random noise to serve as the input of UNet. The Position-Guided Light Adapter is proposed to selectively inject background light from different directions into the foreground at different locations for more natural relighting results. The Spectral Foreground Fixer is utilized to enhance the consistency of the subject.

region and remove the original background from the input image. Following [45, 26], the masked subject and background are encoded by VAE and then concatenated with the random noise to serve as the input for diffusion UNet. For image-based relighting, a Position-Guided Light Adapter (PGLA) is proposed to selectively inject background light information from different directions into the foreground with direction-biased masked attention. Finally, a Spectral Foreground Fixer (SFF) is utilized to enhance the consistency of the foreground.

3.2 Position-Guided Light Adapter

Although simply concatenating the background with random noise can convey some information about environment lighting, it imposes pixel-aligned light prior and overlooks the natural interaction between the subject and light from various directions in the background, thus resulting in undesirable relighting results in some scenes. To alleviate this problem, we propose the Position-Guided Light Adapter (PGLA), which enhances the foreground's response to light sources from different directions in background while reducing potential unreasonable light alignment. This is achieved by additional encoding and organization of the background light information.

Specifically, the overall pipeline of our PGLA is based on the process of IP-Adapter [46]. We first utilize a CLIP image encoder to encode the target background image into a feature map $f_b \in \mathbb{R}^{H \times W \times C}$, where C indicate the embedding dimension. H, W denote the spatial size of feature map. To mitigate the potential noise effects of background textures on light information extraction, we perform a low-frequency enhancement operation on the encoded background features, which helps emphasize the high-level lighting and color tones information. The process can be formulated as:

$$\begin{aligned} g &= Gaussian(H, W, \sigma), \\ f_{bl} &= FFT(f_b) * g, \\ f_{bl} &= IFFT(ReLU(Conv(f_{bl}))) + f_b, \end{aligned} \tag{1}$$

where FFT and $IFFT$ are Fourier transform and Fourier inverse transform. g denotes the filtering coefficient map. σ is cutoff frequency and $*$ means element-wise product. $Conv$ and $ReLU$ operations are utilized to update the features in the spectral domain for efficient global interaction.

Then, we restructure the background features and encode light information from different directions into predefined light queries f_Q, which are randomly initialized learnable embeddings. Considering that the condition is vanilla 2D natural background images, we assume that the light sources can be split into four basic directions: left, right, top, and down. Thus, we leverage four sets of query embeddings to selectively extract light information from background features.

https://doi.org/10.52202/085713-1211

To achieve that, we propose a direction-biased masked attention. As shown in Figure 3, we use cross attention mechanism to condense the information of background into the light query. In detail, we initialize four sets of light query and design them to separately perceive the background light sources of the four directions, *i.e.*, left, right, top, and down. Taking the query f_Q^{left} corresponding to the left light as an example, we generate a coefficient map that decays from left to right with the same size as the background feature, as shown by the "left decay map" in Figure 3. It is then flattened and multiplied with the initial attention weight of corresponding regions to modulate the cross attention, making the query f_Q^{left} focus more on the information on the left side of the background feature. Similarly for the light query responsible for the other directions. Besides, we concatenate the light query and background features as Key and Value to ensure the interaction between different group of queries, which contributes to the overall harmonization.

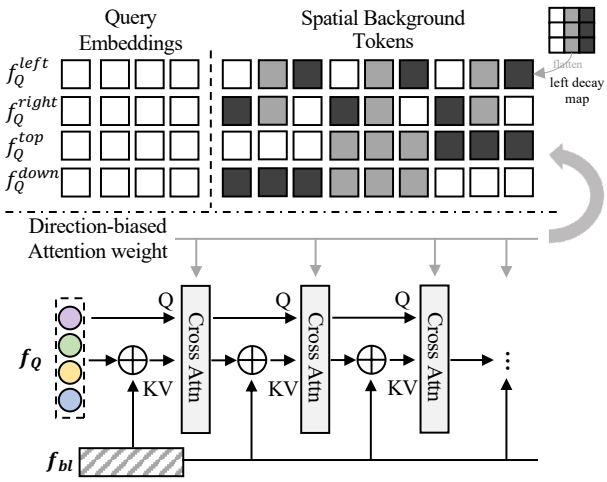

Figure 3: Process of direction-biased masked attention. The upper figure is the design of the attention mask in cross attention. Here the darker color indicates the tendency toward 0. To allow different groups of query to perceive different directions of light, we generate coefficient maps with different attenuation directions and multiply them to the original attention weight. For ease of understanding, we note the transformation of the left decay map in the figure. Besides, the attention weight between different queries has not been modified for overall light information harmonization.

Having these condensed light queries, the next step is to inject their information into the foreground area of the latent features in UNet with similar masked attention. That is, we adjust the attention weight of condition cross attention so that different regions of the foreground object have a stronger response to nearby light sources. In addition to the exchange of Q and KV, we additionally add a mask to the background area to ensure that the background is not changed. Furthermore, as shown in Figure 2, we only inject these light prior in the middle and up block of the UNet. This is due to the fact that feature changes in down blocks may have an impact on the overall semantics [47]. We only need to make changes to the object's light based on its overall representation, which helps to avoid the potential distortion problems caused by extra information injection.

With such designs, the combined subject elements can perform selective interaction with the background to produce more natural relighting results. Related experiments in Section 4.3 also justify the rationality of our designs.

3.3 Spectral Foreground Fixer

Diffusion-based methods tend to face the problem of foreground distortion, especially in small and detailed areas such as face and text. On the one hand, the latent space of large-scale pre-trained models tends to embellish the foreground, which can lead to inconsistency and ID variation with the input. On the other hand, the encoding process of VAE may cause information loss in small regions with high information density, leading to difficulties in maintaining the texture of the subject. Therefore, we propose a Spectral Foreground Fixer (SFF) to address this challenge. This module is based on the assumption that the high-frequency components of an image correspond to the pixels varying drastically, such as object boundaries and textures, while the low-frequency components correspond to the general semantic information such as the color and light.

As shown in Figure 4, we utilize Wavelet Transform to extract the high-frequency and low-frequency components of the input foreground image and the initial predicted results. It can be seen that the high-frequency part maintains the details and textures, while low-frequency part indicate rough colors and tones. Then we combine the high-frequency part of input foreground image with the low-frequency component of the initial relighting result and feed them into a Modulator, which takes

https://doi.org/10.52202/085713-1211

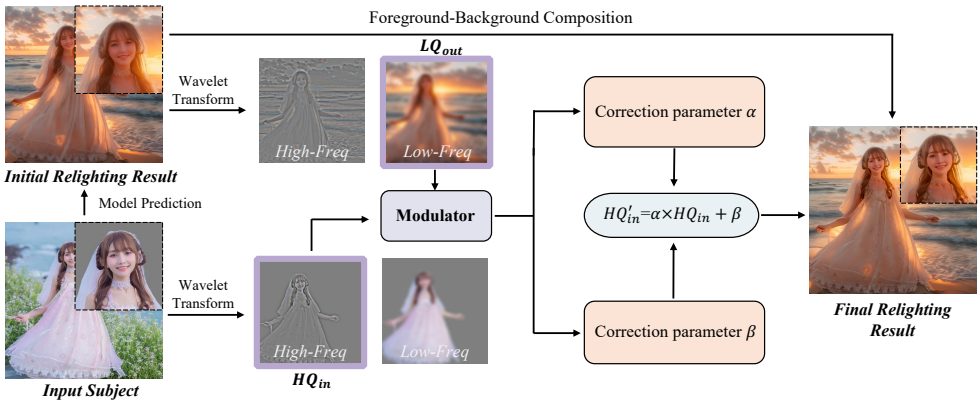

Figure 4: Process of Spectral Foreground Fixer. The modulator is utilized to predict a set of coefficients for modulating the textures of foreground with the light of background.

image-to-image generation paradigm and is trained to predict a set of coefficients to reorganize these information. The reorganization process can be formulated as:

$$\begin{aligned} \alpha, \beta &= \mathcal{M}(HQ_{in}, LQ_{out}), \\ HQ'_{in} &= HQ_{in} * \alpha + \beta, \\ I'_{out} &= HQ'_{in} + LQ_{out}, \end{aligned} \tag{2}$$

where HQ and LQ denote the high-frequency and low-frequency part extracted by Wavelet Transform, respectively. $\mathcal{M}$ is the modulator. $\alpha \in \mathbb{R}^{H\times W\times 3}$ and $\beta \in \mathbb{R}^{H\times W\times 3}$ are the predicted modulation coefficients. α is utilized to control the influence degree of foreground texture and β contains the balance information for harmonizing the foreground. Compared to directly reorganize the foreground texture and background semantics, such design helps to output more consistent and natural results, avoiding artifacts from forced combination. Finally, we replace the foreground region of initial relighting images by the predicted I'_{out} according to the foreground mask.

The modulator M is individually trained in a self-supervised manner. Specifically, we perform random color transformation on arbitrary natural images to obtain pseudo pairs of relighting data, where the transformed image is input and original image serves as target. Then we extract the high-frequency component of the transformed image and the low-frequency component of the original image as inputs to the modulator for modulation coefficient prediction. The modulator is expected to adaptively combine the high-frequency and low-frequency parts from different source and avoid potential color noise or artifacts. We utilize MSE and perceptual loss [48] to supervise the learning of the modulator. Furthermore, to promote training stability and improve coordination, the supervision is applied on both the predicted high-frequency part HQ'_{in} and entire output image I'_{out}.

3.4 Data Generation

We design data generation pipeline to facilitate the training of our model. Our data has three sources. Firstly, we construct pairs by training a relighting ominicontrol [49] lora in a bootstrapping manner, *i.e.*, training – incorporating results into training set – continue training. The initial set consists of 100 image pairs collected from time-lapse photography videos and self-photographed photos. After each training, the model is used to relight vanilla images. High quality pairs are selected and incorporated into the training set for continue training. We will open source this relighting lora to benefit community. Secondly, we utilize available 3D assets [50] to render a number of consistent images with lighting of different color and directions. We construct an automatically rendering pipeline on 3D Arnold Renderer, and generate various lighting effects with random light sources and HDR images for corresponding pairs. Finally, to enhance data diversity, we also process vanilla images with IC-Light [26] and filter out high-quality synthetic data pairs with aesthetic score [51]. The prompts are generated through a two-step process: GPT-4 [52] initially brainstorms over 200 fundamental scenarios, which are then tailored by LLaVA [53] according to the main subjects present in the images. Totally, the quantities of the three types of data are about 600k, 150k, and 300k, respectively. Please see the supplementary material for detailed analysis about the training data.

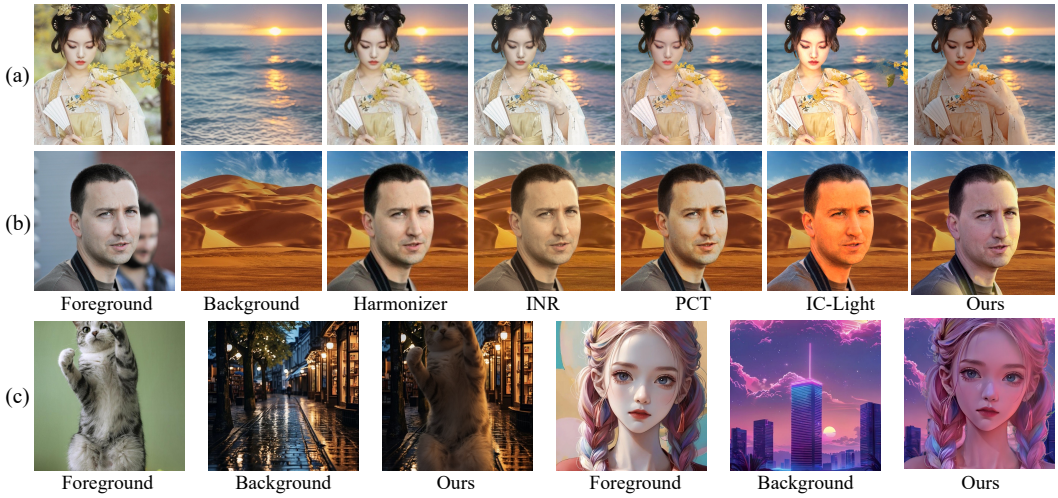

Figure 5: Image-based relighting results. (a) and (b) demonstrate the comparisons with popular available image-based relighting methods, *i.e.*, Harmonizer [38], INR [21], PCT [20], and IC-Light [26]. (c) shows the generalizability of our DreamLight to foregrounds of different categories and various styles of images. Please see the supplementary material for more qualitative results.

4 Experiment

4.1 Implementation Details

Our model is implemented in PyTorch [54] using 8 $\times$ A100 at 512 $\times$ 512 resolutions. The main model and fixer model are trained separately. The main model is trained end-to-end with the batch size of 512. The learning rate is set to 5e-5. We leverage StableDiffusion-v1.5 [1] as the base generative model and CLIP-H [55] as the encoder for position-guided light adapter. Following [26], we takes RMBG-1.4 as the segmentation model for extracting the region of subject. The spectral foreground fixer is finetuned on the VAE model of StableDiffusion-v1.5. The cutoff frequency of spectral filter is set to 5 in default. The number of light query is set to 4 for each direction. The evaluation benchmark contains 600 high-quality image pairs rendered by Arnold Renderer from real objects. For image-based relighting, we take the popular metrics of standard Peak Signal-to-Noise Ratio (PSNR), Structural Similarity (SSIM), Learned Perceptual Image Patch Similarity (LPIPS) [56], and image similarity score calculated by CLIP [55] (CLIP-IS) to verify the effectiveness of our method. For text-based relighting, we utilize image-text matching CLIP score, aesthetic score [51], and Image Reward (IR) [57] score to assess the plausibility of the generated results. IR score is calculated by text-to-image human preference evaluation models trained on large-scale datasets of human preference choices. Aesthetic score is a linear model trained on image quality rating pairs of real images. Please refer to the supplementary material for more details and illustrations.

4.2 Main Results

Qualitative Comparison: In Figure 5 and Figure 6, we present relighting results of our method and the comparison with existing methods. Our DreamLight achieves excellent performance of both image-based and text-based relighting in a single model. More relighting results can be found in the supplementary materials. From (a) and (b) in Figure 5 we can see that our method not only harmonizes the lighting of the foreground and background, but also more effectively models the interaction between light sources and objects within images. In addition to human and natural backgrounds, we illustrate the generalization ability of our model for foregrounds of different categories and other stylistic images in (c). In Figure 6 we compare our method with IC-Light [26] and existing state-of-the-art prompt-guided inpainting methods since existing studies have paid limited attention to relighting based on given prompts. Although inpainting-based methods are capable of generating backgrounds that are in accordance with text prompts, they exhibit a propensity towards maintaining the color and lighting of the foreground unchanged, leading to an incongruity with the background. IC-Light, in contrast, grapples with issues of excessive color variations in

https://doi.org/10.52202/085713-1211

Figure 6: Results of text-based relighting. PowerPaint [43] and BrushNet [42] are powerful inpainting methods that can perform prompt-guided inpainting. To better illustrate, we demarcate the foreground, outpainting methods, and relighting methods with dashed lines.

Method	PSNR↑	SSIM↑	LPIPS ↓	CLIP-IS ↑
INR [21]	16.79	0.694	0.231	0.872
PCT [20]	18.45	0.714	0.206	0.895
IH [19]	20.66	0.762	0.177	0.896
IC-Light [26]	20.27	0.771	0.181	0.889
Ours	**22.15**	**0.783**	**0.158**	**0.908**

Table 1: Quantitative results comparison about image-based relighting.

Method	CLIP Score↑	Aesthetic Score↑	IR Score↑
PP-V1 [43]	0.605	5.58	1.87
PP-V2 [43]	0.627	5.97	2.50
BrushNet [42]	0.613	5.73	1.96
IC-Light [26]	0.629	6.04	2.25
Ours	**0.644**	**6.32**	**3.47**

Table 2: Quantitative results about text-based relighting. PP denotes PowerPoint.

the generated images as well as distortions of the foreground. Our approach can produce relighting results with more natural light. Additionally, the bottom row illustrates the capacity of our method to maintain consistency for the subject. We can observe that all other methods generate results with significant facial distortion, whereas ours ensures excellent subject consistency.

Quantitative Comparison: Table 1 and Table 2 display the evaluation metrics of different methods about image-based and text-based relighting, respectively. Table 1 indicates that our approach can yield more consistent and harmonious outcomes when provided with a target background image. Results in Table 2 show that our DreamLight exhibits advantages in terms of vision-text compatibility, aesthetic appeal, and rationality. Results of user study are reported in the supplementary materials.

4.3 Case Study

Position-Guided Light Adapter: In Figure 7 we conduct visualization comparison of the ablations about the PGLA. As depicted in Figure 7, the model struggles to learn the light interaction between the foreground and background without any prior imposition (W/o adapter). When applying vanilla IP-Adapter [46], which performs arbitrary interaction between subject and background, information from diverse directions in the background interferes with each other, thereby affecting the final relighting performance. Through the utilization of direction-biased masked attention, the model selectively transmits background lighting information, enabling the foreground to acquire lighting that is harmonious with the background (the last two columns). The design of low-frequency enhancement further discards irrelevant information exist in the background textures, thus contributing to robust training

Method	PSNR↑	SSIM↑	LPIPS↓	CLIP-IS↑
w/o adapter	18.58	0.732	0.221	0.865
vanilla IP.A.	20.23	0.752	0.184	0.891
w/o Filter	21.81	0.778	0.162	0.903
PGLA (Ours)	**22.15**	**0.783**	**0.158**	**0.908**

Table 3: Results of different light adapter designs. “IP.A." means IP-Adapter [46]. Filter is the spectral filter used to enhance the low-frequency component of background.

https://doi.org/10.52202/085713-1211

Figure 7: Visualization comparisons with different light adapter design strategies.

Figure 8: Visualization comparisons with different foreground fixer strategies. For the ease of observation, we accentuate the changes in the facial region on the right side of each relighting result. *The green box above showcases the magnified facial region of the relighted results.* The red box below presents the magnified facial region of the input foreground image. Best viewed zoom-in.

of the model and promoting generation of more natural and consistent results. Quantitative results in Table 3 also demonstrate the rationality of our designs.

Spectral Foreground Fixer: Figure 8 shows the qualitative analysis about the proposed SFF. The introduction of ControlNet [58] to provide subject information fails to alleviate the issue of foreground distortion, as distortions often occur on small areas and such design struggles to avoid the problem of encoding information loss. Our SFF achieves robust foreground preservation effects, as also substantiated by the quantitative results in Table 4. As mentioned above, SFF primarily works on critical small regions. While the refinement of these regions is important for visual quality, it has limited impact on global metrics. Thus, to better test the refinement effect, in Table 4 we crop small face regions for metric calculation.

Method	PSNR↑	SSIM↑	CLIP-IS↑	AS↑
W/o fixer	19.31	0.652	0.846	5.31
ControlNet	19.69	0.674	0.868	5.46
SFF (Ours)	**25.53**	**0.831**	**0.946**	**6.84**

Table 4: Different foreground fix strategies.

Handle Both Conditions: We observe that thanks to the unified learning of text and image conditions in a single model, our DreamLight can generate results with the guidance of both conditions. As shown in Figure 9, our method can maintain the structure and elements of the given background while making additional adjustments based on text prompts. Note that this ability is emergent and our model does not undergo training for such situation.

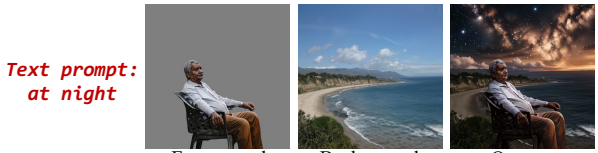

Figure 9: Visualization results of our DreamLight conditioned on both text prompt and background image.

5 Conclusion

In this paper, we present DreamLight that can perform both image-based and text-based relighting in a single model. In addition to extending application scenarios, our model emerges with the ability to handle both conditions simultaneously. By performing tendentious interaction between foreground and background in Position-Guided Light Adapter (PGLA), our model can achieve a natural and harmonious relighting effect. In addition, the Spectral Foreground Fixer (SFF) greatly enhance the consistency of the subject by adaptively leveraging information of different frequency bands. To train the model, we also develop a high-quality data generation pipeline. Experiment results have demonstrated that our DreamLight achieves excellent relighting performance and superior generalization ability for natural images. We hope this work could inspire more related researches and potential practical applications.

https://doi.org/10.52202/085713-1211

Acknowledgements

This work was supported by Guangdong Natural Science Funds for Distinguished Young Scholar (No. 2025B1515020012).

References

[1] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022.

[2] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4195–4205, 2023.

[3] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, pages 6840–6851, 2020.

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

[5] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization. In *CVPR*, pages 6593–6602, 2024.

[6] Lianghua Huang, Di Chen, Yu Liu, Yujun Shen, Deli Zhao, and Jingren Zhou. Composer: Creative and controllable image synthesis with composable conditions. *arXiv preprint arXiv:2302.09778*, 2023.

[7] Binxin Yang, Shuyang Gu, Bo Zhang, Ting Zhang, Xuejin Chen, Xiaoyan Sun, Dong Chen, and Fang Wen. Paint by example: Exemplar-based image editing with diffusion models. In *CVPR*, pages 18381–18391, 2023.

[8] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *CVPR*, pages 6452–6462, 2024.

[9] Anand Bhattad, James Soole, and DA Forsyth. Stylitgan: Image-based relighting via latent control. In *CVPR*, pages 4231–4240, 2024.

[10] Yong Liu, Ran Yu, Fei Yin, Xinyuan Zhao, Wei Zhao, Weihao Xia, Jiahao Wang, Yitong Wang, Yansong Tang, and Yujiu Yang. Learning high-quality dynamic memory for video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

[11] Zhuoyan Luo, Yicheng Xiao, Yong Liu, Shuyan Li, Yitong Wang, Yansong Tang, Xiu Li, and Yujiu Yang. Soc: Semantic-assisted object cluster for referring video object segmentation. *Advances in Neural Information Processing Systems*, 36:26425–26437, 2023.

[12] Kunyang Han, Yong Liu, Jun Hao Liew, Henghui Ding, Jiajun Liu, Yitong Wang, Yansong Tang, Yujiu Yang, Jiashi Feng, Yao Zhao, et al. Global knowledge calibration for fast open-vocabulary segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 797–807, 2023.

[13] Yong Liu, Sule Bai, Guanbin Li, Yitong Wang, and Yansong Tang. Open-vocabulary segmentation with semantic-assisted calibration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3491–3500, 2024.

[14] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. Switchlight: Co-design of physics-driven architecture and pre-training framework for human portrait relighting. In *CVPR*, pages 25096–25106, 2024.

[15] Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, and Vishal M Patel. Lightpainter: interactive portrait relighting with freehand scribble. In *CVPR*, pages 195–205, 2023.

[16] Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image. In *CVPR*, pages 4263–4273, 2024.

[17] Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, pages 43–1, 2021.

[18] Zonghui Guo, Haiyong Zheng, Yufeng Jiang, Zhaorui Gu, and Bing Zheng. Intrinsic image harmonization. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 16367–16376, 2021.

[19] Chris Careaga, S Mahdi H Miangoleh, and Yağız Aksoy. Intrinsic harmonization for illumination-aware image compositing. In *SIGGRAPH Asia*, pages 1–10, 2023.

[20] Julian Jorge Andrade Guerreiro, Mitsuru Nakazawa, and Björn Stenger. Pct-net: Full resolution image harmonization using pixel-wise color transformations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5917–5926, 2023.

[21] Jianqi Chen, Yilan Zhang, Zhengxia Zou, Keyan Chen, and Zhenwei Shi. Dense pixel-to-pixel harmonization via continuous image representation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023.

[22] Ben Xue, Shenghui Ran, Quan Chen, Rongfei Jia, Binqiang Zhao, and Xing Tang. Dccf: Deep comprehensible color filter learning framework for high-resolution image harmonization. In *ECCV*, pages 300–316, 2022.

[23] Wenyan Cong, Xinhao Tao, Li Niu, Jing Liang, Xuesong Gao, Qihao Sun, and Liqing Zhang. High-resolution image harmonization via collaborative dual transformations. In *CVPR*, pages 18470–18479, 2022.

[24] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. *arXiv preprint arXiv:2406.07520*, 2024.

[25] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *SIGGRAPH*, pages 1–12, 2024.

[26] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *ICLR*, 2025.

[27] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. Learning physics-guided face relighting under directional light. In *CVPR*, pages 5124–5133, 2020.

[28] Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli: Diffusion face relighting. In *ICCV*, pages 22646–22657, 2023.

[29] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. Single image portrait relighting. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

[30] Longwen Zhang, Qixuan Zhang, Minye Wu, Jingyi Yu, and Lan Xu. Neural video portrait relighting in real-time via consistency modeling. In *ICCV*, pages 802–812, 2021.

[31] Yazhou Xing, Yu Li, Xintao Wang, Ye Zhu, and Qifeng Chen. Composite photograph harmonization with complete background cues. In *ACM MM*, pages 2296–2304, 2022.

[32] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *SIGGRAPH*, pages 145–156, 2000.

[33] Anand Bhattad and David A Forsyth. Cut-and-paste object insertion by enabling deep image prior for reshading. In *3DV*, pages 332–341, 2022.

 https://doi.org/10.52202/085713-1211

[34] Ke Wang, Michaël Gharbi, He Zhang, Zhihao Xia, and Eli Shechtman. Semi-supervised parametric real-world image harmonization. In *CVPR*, pages 5927–5936, 2023.

[35] Zonghui Guo, Dongsheng Guo, Haiyong Zheng, Zhaorui Gu, Bing Zheng, and Junyu Dong. Image harmonization with transformer. In *ICCV*, pages 14870–14879, 2021.

[36] Yifan Jiang, He Zhang, Jianming Zhang, Yilin Wang, Zhe Lin, Kalyan Sunkavalli, Simon Chen, Sohrab Amirghodsi, Sarah Kong, and Zhangyang Wang. Ssh: A self-supervised framework for image harmonization. In *ICCV*, pages 4832–4841, 2021.

[37] Linfeng Tan, Jiangtong Li, Li Niu, and Liqing Zhang. Deep image harmonization in dual color spaces. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 2159–2167, 2023.

[38] Zhanghan Ke, Chunyi Sun, Lei Zhu, Ke Xu, and Rynson WH Lau. Harmonizer: Learning to perform white-box image and video harmonization. In *ECCV*, pages 690–706, 2022.

[39] Wenyan Cong, Jianfu Zhang, Li Niu, Liu Liu, Zhixin Ling, Weiyuan Li, and Liqing Zhang. Dovenet: Deep image harmonization via domain verification. In *CVPR*, pages 8394–8403, 2020.

[40] Yi-Hsuan Tsai, Xiaohui Shen, Zhe Lin, Kalyan Sunkavalli, Xin Lu, and Ming-Hsuan Yang. Deep image harmonization. In *CVPR*, pages 3789–3797, 2017.

[41] Dina Bashkirova, Arijit Ray, Rupayan Mallick, Sarah Adel Bargal, Jianming Zhang, Ranjay Krishna, and Kate Saenko. Lasagna: Layered score distillation for disentangled object relighting. *arXiv preprint arXiv:2312.00833*, 2023.

[42] Xuan Ju, Xian Liu, Xintao Wang, Yuxuan Bian, Ying Shan, and Qiang Xu. Brushnet: A plug-and-play image inpainting model with decomposed dual-branch diffusion. *arXiv preprint arXiv:2403.06976*, 2024.

[43] Junhao Zhuang, Yanhong Zeng, Wenran Liu, Chun Yuan, and Kai Chen. A task is worth one word: Learning with task prompts for high-quality versatile image inpainting. *arXiv preprint arXiv:2312.03594*, 2023.

[44] Yong Liu, Cairong Zhang, Yitong Wang, Jiahao Wang, Yujiu Yang, and Yansong Tang. Universal segmentation at arbitrary granularity with language instruction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3459–3469, 2024.

[45] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *CVPR*, pages 18392–18402, 2023.

[46] Hu Ye, Jun Zhang, Sibo Liu, Xiao Han, and Wei Yang. Ip-adapter: Text compatible image prompt adapter for text-to-image diffusion models. *arXiv preprint arXiv:2308.06721*, 2023.

[47] Mingdeng Cao, Xintao Wang, Zhongang Qi, Ying Shan, Xiaohu Qie, and Yinqiang Zheng. Masactrl: Tuning-free mutual self-attention control for consistent image synthesis and editing. In *ICCV*, pages 22560–22570, 2023.

[48] Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *ECCV*, pages 694–711, 2016.

[49] Zhenxiong Tan, Songhua Liu, Xingyi Yang, Qiaochu Xue, and Xinchao Wang. Ominicontrol: Minimal and universal control for diffusion transformer. *arXiv preprint arXiv:2411.15098*, 2024.

[50] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, pages 13142–13153, 2023.

[51] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *NeurIPS*, pages 25278–25294, 2022.

[52] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[53] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *NeurIPS*, 2024.

[54] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *NeurIPS*, 2019.

[55] 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 *ICML*, pages 8748–8763, 2021.

[56] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pages 586–595, 2018.

[57] Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. Imagereward: Learning and evaluating human preferences for text-to-image generation. *NeurIPS*, 2024.

[58] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, pages 3836–3847, 2023.

 https://doi.org/10.52202/085713-1211

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We clearly provide our motivation and contribution in these parts and the claims accurately reflect them.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the potential limitations in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: We aim to address a practical and applicable task named unified image relighting rather than proposing theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided our implementation details in the main paper and supplementary materials. We will open source our code and models.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

https://doi.org/10.52202/085713-1211

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: Codes and models need to be reviewed by our organization before they are made available. We will make them public as soon as possible.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have claimed the corresponding information in the main paper and supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: We have demonstrated the generation results of different seeds in the supplementary materials.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.

- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have claimed the compute resources.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have followed the NeurIPS code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We have discussed the potential broader impacts in the supplementary materials.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.

https://doi.org/10.52202/085713-1211

- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: Our work is under the CC-BY 4.0 license.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.

- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our work does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.

 https://doi.org/10.52202/085713-1211

- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: We only use LLMs to benefit writing.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.