

Training-Free Light-Guided Text-to-Image Diffusion Model via Initial Noise Manipulation

Ryugo Morita^{1,2}, Stanislav Frolov¹, Brian Bernhard Moser¹, Ko Watanabe¹, Riku Takahashi^{1,2}, Andreas Dengel¹

¹RPTU Kaiserslautern-Landau & DFKI GmbH, Kaiserslautern, Germany

²Faculty of Science and Engineering, Hosei University, Tokyo, Japan

Abstract—Diffusion models have demonstrated high-quality performance in conditional text-to-image generation, particularly with structural cues such as edges, layouts, and depth. However, lighting conditions have received limited attention and remain difficult to control within the generative process. Existing methods handle lighting through a two-stage pipeline that relights images after generation, which is inefficient. Moreover, they rely on fine-tuning with large datasets and heavy computation, limiting their adaptability to new models and tasks. To address this, we propose a novel Training-Free Light-Guided Text-to-Image Diffusion Model via Initial Noise Manipulation (LGTM), which manipulates the initial latent noise of the diffusion process to guide image generation with text prompts and user-specified light directions. Through a channel-wise analysis of the latent space, we find that selectively manipulating latent channels enables fine-grained lighting control without fine-tuning or modifying the pre-trained model. Extensive experiments show that our method surpasses prompt-based baselines in lighting consistency, while preserving image quality and text alignment. This approach introduces new possibilities for dynamic, user-guided light control. Furthermore, it integrates seamlessly with models like ControlNet, demonstrating adaptability across diverse scenarios.

Index Terms—Light-Guided Text-to-Image, Generative Model, Diffusion Model

I. INTRODUCTION

Diffusion models have revolutionized image synthesis by enabling high-fidelity text-to-image generation [1]–[3], with wide-ranging applications such as art, design, and education [4]–[6]. To better reflect user preferences, recent studies have explored conditional generation using structural cues such as edges, segmentation maps, and layouts [7]–[12]. These methods focus on object features, with limited attention to lighting, an essential factor for realism and mood.

Recent works [13], [14] address lighting control via two-stage workflows that first generate an image and then apply a separate relighting module to modify its illumination. However, such pipelines are inefficient and typically depend on illumination-annotated datasets and heavy fine-tuning. IC-Light [13] is trained on approximately 10 million images using 8×H100 (80GB) GPUs over 100 hours, while DelightNET [14] constructs a synthetic dataset of 25K objects, each rendered under 4 viewpoints and 12 lighting conditions, and is trained with 8×V100 GPUs for 30 hours. In the fast-evolving landscape of generative models, such resource-heavy pipelines are increasingly impractical.

On the other hand, prompt engineering [15] provides a lightweight, training-free way to influence generation, but it



Fig. 1. Existing prompt-engineering methods fail to generate differences in generated images with or without light-specific prompts, resulting in outputs that overlook specified lighting conditions. Our proposed LGTM effectively guides lighting during image generation, ensuring outputs align with text prompts and desired lighting directions without fine-tuning.

remains unreliable for controlling illumination. As shown in Fig. 1, Stable Diffusion [3] fails to achieve consistent illumination even when users explicitly specify a light direction via text prompt. This highlights the need for a more direct and seamless method to incorporate user-defined lighting conditions into the generation process.

To address this challenge, we propose a novel Training-Free Light-Guided Text-to-Image Diffusion Model (LGTM) that manipulates the initial noise to steer illumination throughout the diffusion process. We first conduct a channel-wise sensitivity analysis of the VAE latent noise in Latent Diffusion Models (LDMs), and find that *channel 1* is strongly correlated with global brightness and perceived light direction. Guided by this analysis, LGTM selectively manipulates *channel 1* to enable intuitive and fine-grained lighting control without fine-tuning.

Extensive experiments demonstrate that LGTM achieves more accurate and coherent lighting aligned with user-specified directions than prompt-based methods in Stable Diffusion, while preserving visual quality and text-image alignment. In addition, by modifying only the initial noise, our method can be seamlessly applied to conditional modules such as ControlNet [7], enabling simultaneous control over structural cues (e.g., edges) and illumination, and demonstrating strong adaptability to diverse generation scenarios and user constraints. Our contributions are as follows:

- We define light-guided text-to-image generation as a

novel task and propose a Training-Free Light-Guided Text-to-Image Diffusion Model (LGTM) to address this.

- We are the first to explore light control via latent-space manipulation by leveraging the disentangled structure of the VAE latent channels in LDM, identifying *channel 1* as a key factor for encoding lighting information.
- LGTM achieves effective illumination control by modifying only the initial latent noise, without altering the model architecture or parameters, making it compatible with extended frameworks like ControlNet and adaptable to a wide range of image/video generation scenarios.

II. RELATED WORKS

A. Conditional Text-to-Image Generation

Diffusion models have significantly advanced text-to-image generation [3], [16]. Conditional text-to-image methods incorporate additional modalities, such as edges [7], [8], segmentation [9], [10], and layouts [11], [17], to better align generated images with user preferences. As image generation models rapidly evolve, training-based approaches suffer from limited adaptability due to the need for extensive retraining. This has motivated growing interest in training-free conditional generation methods. Recent training-free approaches mainly manipulate attention mechanisms to control structural properties, such as object layouts [12], [18], inter-object relationships [19], [20], and panoramic compositions [21], [22]. However, lighting remains unexplored in training-free settings. Unlike structural attributes, illumination requires costly annotations or curated datasets with diverse lighting conditions.

B. Initial Noise of Diffusion Model

Diffusion models begin with Gaussian noise and denoise it to synthesize images aligned with a target distribution. The initial noise plays a pivotal role in this process, as they strongly influence both visual quality and semantic alignment [23]–[25]. Recent studies optimize the initial noise with human preference signals or attention-based metrics to improve image fidelity and text-image alignment without requiring additional model training [26], [27]. Furthermore, selecting optimal noise seeds or modifying localized regions within the noise affects object placement and overall image quality [28], [29]. Beyond optimization, direct manipulation of the initial noise is leveraged to exert fine-grained control over image layouts, enabling layout-aware and layered image generation [30]–[33]. Inspired by these findings, we explore illumination control beyond structural control via initial noise manipulation.

C. Diffusion Models for Light Control

Existing lighting control methods primarily focus on relighting existing images rather than incorporating lighting guidance directly into the text-to-image generation process. Relightful Harmonization [34] adjusts illumination through post-processing, while other approaches such as DelightNET [14], FlashTex [35], and LightIt [36] rely on 3D rendering or synthetic datasets. IC-Light [13] further introduces a physically motivated light transport mechanism during training.

Despite their effectiveness, these methods rely on a two-stage relighting paradigm that modifies illumination after image generation. In contrast, we define light-guided text-to-image generation as a new task, where lighting conditions are specified at generation time. We address this task with a training-free approach that directly integrates lighting control into the diffusion process, without a separate relighting stage.

III. METHODS

As illustrated in Fig. 2, LGTM takes a text prompt p and a user-specified light direction l as inputs, and generates an output image I whose illumination follows l while remaining consistent with p . To achieve this, we first analyze channel-wise sensitivities of the initial latent noise in the VAE latent space, and then guide illumination by manipulating the initial noise according to a light mask derived from l .

A. Preliminaries

Our method extends Latent Diffusion Models (LDM) [3], operating in the latent space of a VAE encoders $\mathcal{E}$ and decoders $\mathcal{D}$ [37]. The encoder $\mathcal{E}$ encodes an image $x \in \mathbb{R}^{H \times W \times 3}$ in RGB space into a latent representation $z = \mathcal{E}(x)$, where $z \in \mathbb{R}^{H/8 \times W/8 \times 4}$ and the decoder $\mathcal{D}$ is trained to reconstruct x as $\hat{x} = \mathcal{D}(z)$, which is approximately identical to x .

Stable Diffusion employs a Denoising Diffusion Probabilistic Model (DDPM) [2] operating in the latent space of LDM. It trains a U-Net model, ϵ_θ , to predict noise added to an initial latent, denoted as z_t , which is the latent $z = \mathcal{E}(x)$ with noise added at timestep $t \in T$. Given a condition y (i.e., text prompt), the objective of a text-to-image LDM is

$$\mathcal{L}_{\text{LDM}} := \mathbb{E}_{\mathcal{E}(x), y, \epsilon \sim \mathcal{N}(0,1), t} [\|\epsilon - \epsilon_\theta(z_t, t, \tau_\theta(y))\|^2], \quad (1)$$

where both ϵ_θ and τ_θ are jointly optimized. However, Stable Diffusion adopts a frozen CLIP text encoder instead of a trainable text encoder τ_θ .

B. Analysis: Channel-wise Effects of Initial Noise

We analyze how channel-wise perturbations of the initial latent noise influence illumination-related attributes in the generated images. Specifically, we conduct a controlled channel-wise perturbation on the initial latent noise $z_T \in \mathbb{R}^{H/8 \times W/8 \times 4}$. For each channel $c \in \{1, 2, 3, 4\}$, we apply a constant scaling while keeping all other channels unchanged:

$$\hat{z}_T^{(c)} = \alpha \odot z_T^{(c)}, \quad (2)$$

where α is a constant scaling factor. We generate images using the same text prompt and random seed, varying only the perturbed channel, in order to isolate its effect.

Fig. 3 shows the qualitative results of this analysis. We observe that perturbing *channel 1* consistently induces global changes in brightness and alters the apparent illumination direction across the scene. In contrast, perturbations applied to channels 2–4 primarily affect color tone and background hue, with minimal influence on lighting or shadow structure.

These observations indicate that *channel 1* is strongly correlated with illumination-related factors in the latent space. This

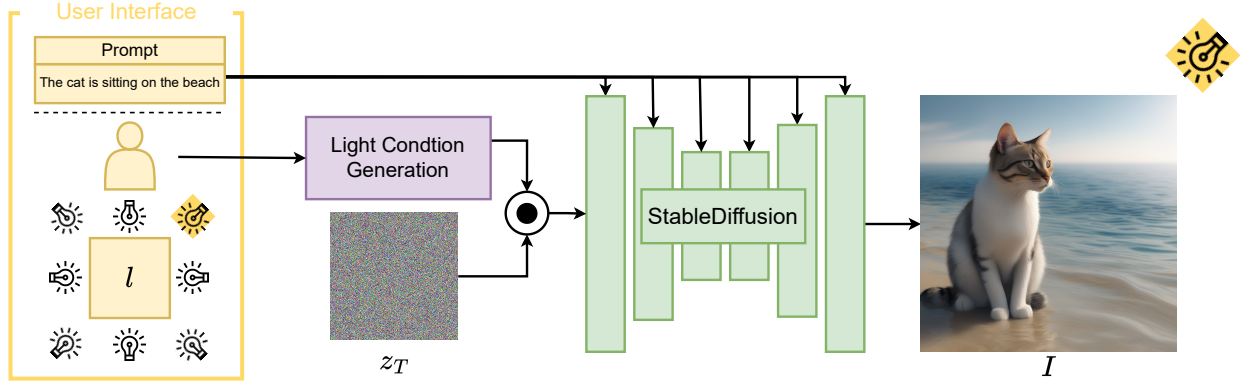


Fig. 2. Overview of our proposed LGTM. The user inputs a prompt p and a light condition l . The Light Conditional Generation module generates the light direction mask m_l according to l for manipulating the initial noise in Stable Diffusion. Then, the vanilla Stable Diffusion model integrates these inputs, dynamically adjusting the latent space—particularly *channel 1*—to reflect user-defined lighting conditions. Finally, it outputs the final image I .

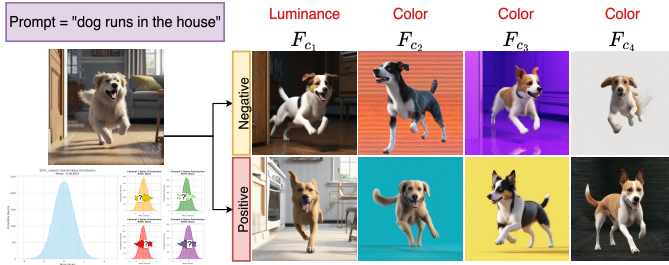


Fig. 3. **Channel-wise sensitivity analysis via scaling the initial latent noise.** We scale a single latent channel $z_T^{(c)}$ ($c \in \{1, 2, 3, 4\}$) by a constant factor α , while keeping the prompt, random seed, and the other channels fixed. Scaling *channel 1* consistently induces global brightness changes and alters the perceived illumination direction, whereas scaling channels 2–4 mainly affects chromatic attributes with limited impact on lighting.

channel-wise sensitivity analysis motivates our design choice to guide lighting by selectively manipulating *channel 1* of the initial latent noise, as described in the following sections.

C. Light Conditional Generation (LCG)

Generating lighting conditions requires modeling complex interactions between light, shadows, and reflections. To simplify this, we introduce Light Conditional Generation (LCG), where users specify light direction via a graphical interface by selecting a point or line indicating the light source. The system generates a light mask m_l that defines the light’s origin and spread, creating smooth and natural lighting effects.

In our approach, we employ a linear gradient to transition from white to black based on the distance d of each pixel (i, j) from the user-defined light source. Specifically, we use:

$$m_l(i, j) = \max\left(0, 1 - \frac{d(i, j)}{r}\right), \quad (3)$$

where r is a user-defined radius that controls the extent of the light’s influence. Within this radius, the mask value linearly decreases from 1 (white) at the light source to 0 (black) at $d(i, j) = r$. Pixels beyond this radius remain at 0, effectively modeling regions with negligible light.

This linear formulation offers a straightforward way to control the range and smoothness of the light gradation, making it more intuitive for users to specify how far the lighting should extend in the image. The resulting mask m_l is applied to guide the lighting during image generation.

D. Latent Space Light Guidance (LSLG)

Building on the generated light mask m_l , we propose a Latent Space Light Guidance (LSLG) technique to guide lighting in Stable Diffusion’s latent space. The mask is applied to *channel 1* of the initial noise z_T , modulating light intensity according to the user’s input. The transformation is defined as:

$$\hat{z}_T^1 = z_T^1 \odot (1 + m_l) \quad (4)$$

where z_T^1 represents the initial latent noise at timestep T for channel 1, and m_l scales light intensity across the scene.

Our experiments reveal that *channel 1* encodes illumination-related information, making it the ideal target for light manipulation. Adjusting this channel allows for intuitive and precise control of lighting conditions directly in the generation process without additional training.

IV. EXPERIMENTS

A. Experimental Setup

We conduct experiments using the Stable Diffusion XL (SDXL) [38] to generate images at a resolution of 1024×1024 . For inference, we utilize the DDIMSampler with a guidance scale of 7.5 and 50 time steps. Since no prior work has addressed light-guided text-to-image generation, we use vanilla SDXL with prompt engineering as a baseline. Light-specific prompts such as “light coming from the *light_direction*” are added to evaluate its ability to guide lighting.

B. Dataset and Metrics

We use the Dog and Cat dataset [39], containing 2,000 images evenly split between cats and dogs. Captions were generated using BLIP [40] to focus the comparison on light control rather than object generation accuracy. Limiting object categories enabled controlled experiments on light conditions.



Fig. 4. **Qualitative Results.** The existing model fails to control lighting conditions, often generating images with random or inconsistent lighting. In contrast, our approach effectively incorporates user-specified light direction and intensity, producing more natural and coherent lighting effects in the generated images.

TABLE I
QUANTITATIVE RESULTS COMPARING OURS WITH THE BASELINE.

Method	FID ↓	NIMA ↑	CLIP-I ↑	CLIP-T ↑	Left ↑	Right ↑
Cat						
SDXL	69.57	5.44	0.671	0.317	52.8%	52.9%
Ours	79.13	5.66	0.663	0.320	79.0%	77.3%
Dog						
SDXL	71.15	5.67	0.618	0.309	51.9%	52.3%
Ours	81.08	5.68	0.610	0.312	77.3%	76.7%

To assess visual realism and aesthetics, we employ Fréchet Inception Distance (FID) [41] and Neural Image Assessment (NIMA) [42]. Text-image alignment is assessed using CLIP-I and CLIP-T [43]. To evaluate light control, we propose light accuracy. First, we use YOLOv8 [44] to detect the object and expand their bounding boxes by 1.25x to include surrounding areas. Within these regions, we apply a shadow detection model [45] to analyze shadow directions. This metric assesses whether object shadows align correctly with the specified light direction, leveraging the principle that shadows extend opposite the light source. For example, the shadow should extend to the right if the light is from the left. This metric thus provides a direct quantitative assessment of how effectively our model positions shadows according to the intended lighting.

C. Qualitative Results

Fig. 4 shows qualitative comparisons between the baseline and our method. While the baseline model generates high-quality images aligned with text prompts, it fails to account for light-specific directives, often placing light sources at random. In contrast, our model incorporates both text prompts and light direction, accurately control the specified light source. These results demonstrate our method’s capability to understand and reflect light-shadow relationships.

D. Quantitative Results

Table I summarizes the scores for visual quality, text alignment, and light control accuracy. In terms of perceptual quality

and text alignment, the two methods remain comparable, as reflected by similar NIMA and CLIP-based scores. However, FID increases with our method, which is consistent with its known sensitivity to global appearance changes such as brightness and illumination. Since LGTM steers lighting, the generated images deviate from the dataset’s marginal lighting distribution, even when perceptual quality is preserved. Thus, the higher FID reflects a controllability–distribution trade-off rather than a degradation in visual quality.

The critical difference lies in light control accuracy. While the baseline model produces near-random shadow orientations (approximately 52% for both left and right lighting), our method achieves substantially higher accuracy. For example, under left-side lighting, LGTM correctly aligns shadows in 79.0% (Cat) and 77.3% (Dog) of the generated images, compared to 52.8% and 51.9% for the baseline. Similarly, under right-side lighting, LGTM attains 77.3% (Cat) and 76.7% (Dog) accuracy, whereas the baseline remains close to chance level (52.9% and 52.3%). These results demonstrate that LGTM provides reliable and consistent lateral lighting control without additional training, while maintaining high visual quality and text–image alignment.

V. APPLICATION

We demonstrate the LGTM’s flexibility by integrating ControlNet [7], enabling simultaneous control over text prompts, edges, and lighting. LGTM only manipulates the initial latent

TABLE II
QUANTITATIVE RESULTS WITH THE CONTROLNET.

Method	FID ↓	NIMA ↑	CLIP-I ↑	CLIP-T ↑	Left ↑	Right ↑
Cat						
ControlNet	71.56	5.46	0.664	0.296	51.2%	51.8%
Ours	83.63	5.66	0.668	0.315	77.3%	76.2%
Dog						
ControlNet	76.63	5.56	0.606	0.283	51.6%	52.1%
Ours	83.54	5.62	0.613	0.307	77.7%	73.2%

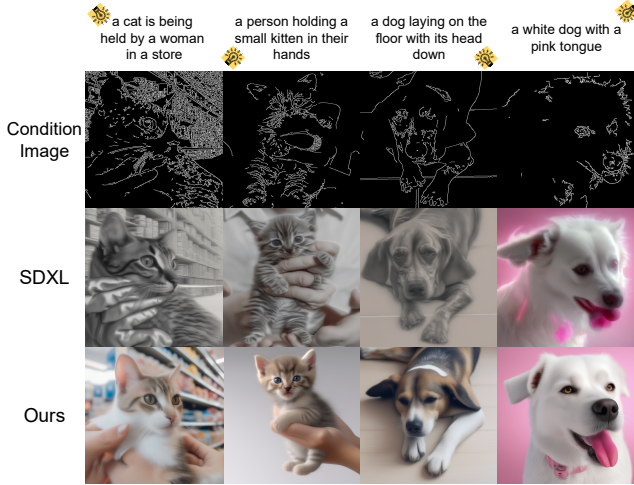


Fig. 5. **Qualitative Results in Application.** Existing models integrated with ControlNet [7] control edges but fail to handle lighting, often producing inconsistent results. Our approach combines user-specified light direction and intensity with edge control, generating images with natural lighting and precise structure, demonstrating versatility in handling multiple controls.

noise, it can be plugged into a latent-based generation model straightforwardly and used jointly with structural conditions.

A. Qualitative Comparison with Standard ControlNet

Fig. 5 illustrates a qualitative comparison between standard ControlNet and our extended method. While ControlNet successfully generates images conditioned on text prompts and canny edges, it fails to account for specified lighting directions. In contrast, our method incorporates all conditional information, including edge and light guidance, to generate images with more coherent lighting and shadow placement.

B. Quantitative Comparison with Standard ControlNet

As shown in Table II, our model surpasses the standard ControlNet in almost visual quality and text alignment metrics. Light control’s accuracy highlights ControlNet’s shortcomings in generating images aligned with light direction. In contrast, our method effectively integrates lighting conditions, ensuring accurate shadow placement and consistency.

VI. LIMITATIONS AND FUTURE WORK

As shown in Fig. 6, the generated subjects tend to align their orientation with the direction of the light source. This tendency

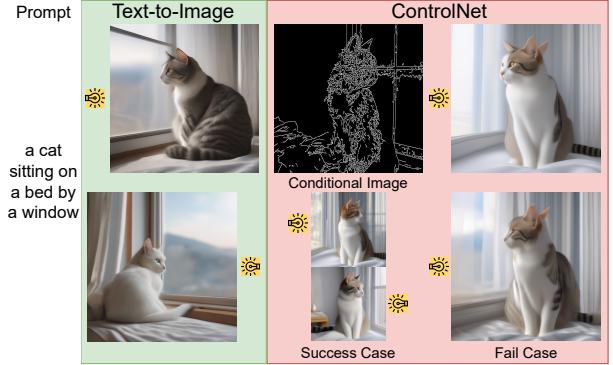


Fig. 6. **Illustration of Light Alignment Behavior.** LGTM aligns generated subjects with the specified light direction, even overriding explicit constraints from ControlNet (e.g., canny-edge-based facial orientation). This can enhance realism but may also result in unintended subject orientation.

persists even under ControlNet [7] constraints, indicating that the diffusion model prioritizes consistency with lighting cues over geometric orientation. This limitation becomes apparent in scenarios that require independent control of lighting and subject pose. As this work represents an initial step toward light-guided text-to-image generation, a deeper investigation into the interaction between lighting conditions and generative biases remains an important direction for future research.

VII. CONCLUSION

This work is the first to explicitly explore the relationship between the VAE latent space in Stable Diffusion and light control, identifying *channel 1* as key to intuitive and precise light manipulation. Our method provides a user-friendly interface that integrates text prompts and light conditions, enabling seamless control of lighting during image generation. Furthermore, its adaptability allows integration with models like ControlNet, broadening its potential applications. Extensive result show that LGTM effectively aligns generated images with both textual descriptions and user-specified light directions without additional training. This advancement highlights the practicality and versatility of our approach to dynamic, user-guided image generation.

ACKNOWLEDGEMENTS

This work was supported by the BMBF Project Albatross (Grant 01IW24002). All compute was done thanks to the Pegasus cluster at DFKI.

REFERENCES

- [1] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020.
- [2] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 33, 2020.
- [3] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022.
- [4] Jaerin Lee, Daniel Sungho Jung, Kanggeon Lee, and Kyoung Mu Lee. Semanticdraw: towards real-time interactive content creation from image diffusion models. In *CVPR*, pages 13021–13030, 2025.
- [5] Zhendong Wang, Jianmin Bao, Shuyang Gu, Dong Chen, Wengang Zhou, and Houqiang Li. Designdiffusion: High-quality text-to-design image generation with diffusion models. In *CVPR*, pages 20906–20915, 2025.
- [6] Ryugo Morita, Ko Watanabe, Jinjia Zhou, Andreas Dengel, and Shoya Ishimaru. Genareading: Augmenting human cognition with interactive digital textbooks using large language models and image generation models. In *Proceedings of the Augmented Humans International Conference 2025*, pages 289–301, 2025.
- [7] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023.
- [8] Subhadeep Koley, Ayan Kumar Bhunia, Deeptanshu Sekhri, Aneeshan Sain, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. It's all about your sketch: Democratising sketch control in diffusion models. In *CVPR*, 2024.
- [9] Yogesh Balaji, Seungjun Nah, Xun Huang, Arash Vahdat, Jiaming Song, Qingsheng Zhang, Karsten Kreis, Miika Aittala, Timo Aila, Samuli Laine, et al. ediff-i: Text-to-image diffusion models with an ensemble of expert denoisers. *arXiv preprint arXiv:2211.01324*, 2022.
- [10] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *ECCV*. Springer, 2022.
- [11] Guangcong Zheng, Xianpan Zhou, Xuewei Li, Zhongang Qi, Ying Shan, and Xi Li. Layoutdiffusion: Controllable diffusion model for layout-to-image generation. In *CVPR*, 2023.
- [12] Jiafeng Mao and Xueting Wang. Training-free location-aware text-to-image synthesis. In *ICIP*. IEEE, 2023.
- [13] 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.
- [14] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*, 2024.
- [15] Vivian Liu and Lydia B Chilton. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI conference on human factors in computing systems*, pages 1–23, 2022.
- [16] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *NeurIPS*, 35, 2022.
- [17] Zhengyuan Yang, Jianfeng Wang, Zhe Gan, Linjie Li, Kevin Lin, Chenfei Wu, Nan Duan, Zicheng Liu, Ce Liu, Michael Zeng, et al. Reco: Region-controlled text-to-image generation. In *CVPR*, 2023.
- [18] Sicheng Mo, Fangzhou Mu, Kuan Heng Lin, Yanli Liu, Bochen Guan, Yin Li, and Bolei Zhou. Freecontrol: Training-free spatial control of any text-to-image diffusion model with any condition. In *CVPR*, 2024.
- [19] Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM transactions on Graphics (TOG)*, 42(4):1–10, 2023.
- [20] Yanyu Li, Pencheng Wan, Liang Han, Yaowei Wang, Liqiang Nie, and Min Zhang. Countdiffusion: Text-to-image synthesis with training-free counting-guidance diffusion. *arXiv preprint arXiv:2505.04347*, 2025.
- [21] Omer Bar-Tal, Lior Yariv, Yaron Lipman, and Tali Dekel. Multidiffusion: fusing diffusion paths for controlled image generation. In *ICML*, pages 1737–1752, 2023.
- [22] Stanislav Frolov, Brian B Moser, and Andreas Dengel. Spotdiffusion: A fast approach for seamless panorama generation over time. In *WACV*, pages 2073–2081. IEEE, 2025.
- [23] Song Yan, Min Li, Bi Xinliang, Jian Yang, Yusen Zhang, Guanye Xiong, Yunwei Lan, Tao Zhang, Wei Zhai, and Zheng-Jun Zha. Beyond randomness: Understand the order of the noise in diffusion. *arXiv preprint arXiv:2511.07756*, 2025.
- [24] Shuangqi Li, Hieu Le, Jingyi Xu, and Mathieu Salzmann. All seeds are not equal: Enhancing compositional text-to-image generation with reliable random seeds. *arXiv preprint arXiv:2411.18810*, 2024.
- [25] Zipeng Qi, Lichen Bai, Haoyi Xiong, and Zeke Xie. Not all noises are created equally: Diffusion noise selection and optimization. *arXiv preprint arXiv:2407.14041*, 2024.
- [26] Aravindan Kamatchi Sundaram, Ujjayan Pal, Abhimanyu Chauhan, Aishwarya Agarwal, and Srikrishna Karanam. Cocono: Attention contrast-and-complete for initial noise optimization in text-to-image synthesis. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 9862–9870, 2025.
- [27] Luca Eyring, Shyamgopal Karthik, Karsten Roth, Alexey Dosovitskiy, and Zeynep Akata. Reno: Enhancing one-step text-to-image models through reward-based noise optimization. *NeurIPS*, 37:125487–125519, 2024.
- [28] Luca Eyring, Shyamgopal Karthik, Alexey Dosovitskiy, Nataniel Ruiz, and Zeynep Akata. Noise hypernetworks: Amortizing test-time compute in diffusion models. *arXiv preprint arXiv:2508.09968*, 2025.
- [29] Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Boqing Gong, Cho-Jui Hsieh, and Minhao Cheng. The crystal ball hypothesis in diffusion models: Anticipating object positions from initial noise. *arXiv preprint arXiv:2406.01970*, 2024.
- [30] Xueting Wang Jiafeng Mao and Kiyoharu Aizawa. The lottery ticket hypothesis in denoising: Towards semantic-driven initialization. *ECCV*, 2024.
- [31] Ryugo Morita, Sho Kuno, Ryunosuke Tanaka, Rongzhi Li, Hoang Dai Dinh, and Issey Sukeeda. Sawna: Space-aware text to image generation. In *Proceedings of the Special Interest Group on Computer Graphics and Interactive Techniques Conference Posters*, pages 1–2, 2025.
- [32] Ryugo Morita, Stanislav Frolov, Brian Bernhard Moser, Takahiro Shirakawa, Ko Watanabe, Andreas Dengel, and Jinjia Zhou. Tkg-dm: Training-free chroma key content generation diffusion model. *arXiv preprint arXiv:2411.15580*, 2024.
- [33] Daichi Nagai, Ryugo Morita, Shunsuke Kitada, and Hitoshi Iyatomi. Tuae: Training-free noise transplant and cultivation diffusion model. *arXiv preprint arXiv:2511.02580*, 2025.
- [34] 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*, 2024.
- [35] Kangle Deng, Timothy Omerick, Alexander Weiss, Deva Ramanan, Jun-Yan Zhu, Tinghui Zhou, and Maneesh Agrawala. Flashtex: Fast relightable mesh texturing with lightcontrolnet. In *ECCV*. Springer, 2025.
- [36] Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Nießner, and Yannick Hold-Geoffroy. Lightit: Illumination modeling and control for diffusion models. In *CVPR*, 2024.
- [37] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.
- [38] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.
- [39] Jeremy Elson, John R Douceur, Jon Howell, and Jared Saul. Asirra: a captcha that exploits interest-aligned manual image categorization. *CCS*, 7(366-374):15, 2007.
- [40] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*. PMLR, 2022.
- [41] 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. *NeurIPS*, 30, 2017.
- [42] Hossein Talebi and Peyman Milanfar. Nima: Neural image assessment. *IEEE TIP*, 27(8), 2018.
- [43] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*, 2021.
- [44] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *ADICS*. IEEE, 2024.
- [45] Runmin Cong, Yuchen Guan, Jinpeng Chen, Wei Zhang, Yao Zhao, and Sam Kwong. Sddnet: Style-guided dual-layer disentanglement network for shadow detection. In *ACM MM*, 2023.