Linear Image Generation by Synthesizing Exposure Brackets

Yuekun Dai[1] Zhoutong Zhang[2] Shangchen Zhou[1] Nanxuan Zhao[3]
[1]S-Lab, Nanyang Technological University [2]Adobe NextCam [3]Adobe Research
{ydai005, s200094}@ntu.edu.sg, {zhoutongz, nanxuanz}@adobe.com
https://ykdai.github.io/projects/raw_gen

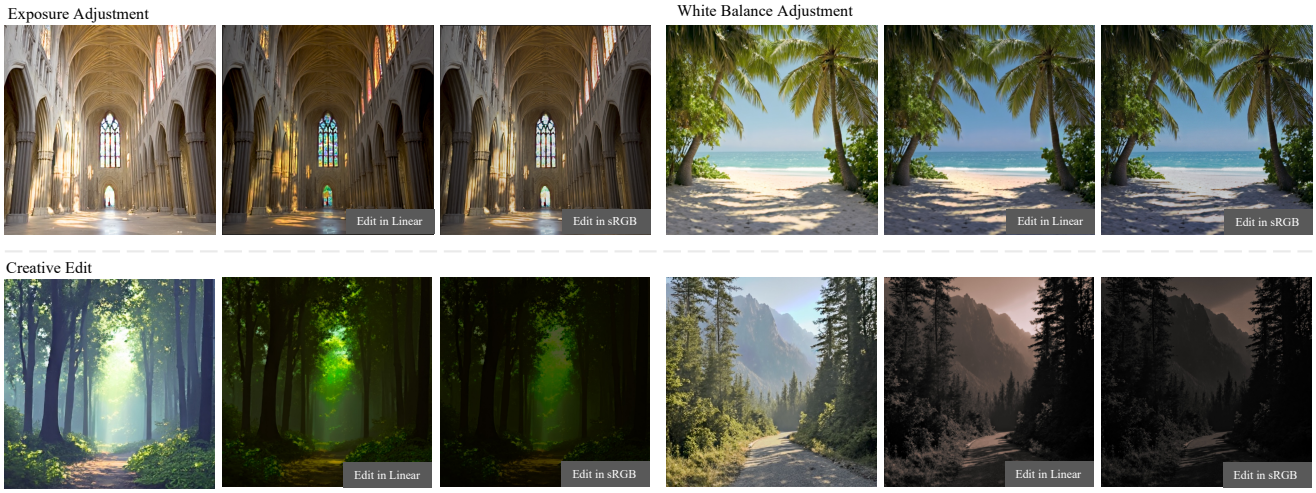


Figure 1. Linear image provides more room for the user to edit compared to sRGB images. Here we show the difference between linear and sRGB images for post editing. All four images are generated from our model, and the same edits are applied in both linear space and sRGB to showcase the difference. Linear images provide more details in highlights, are more accurate for white balance adjustments and support more dramatic creative edits.

Abstract

The life of a photo begins with photons striking the sensor, whose signals are passed through a sophisticated image signal processing (ISP) pipeline to produce a display-referred image. However, such images are no longer faithful to the incident light, being compressed in dynamic range and stylized by subjective preferences. In contrast, RAW images record direct sensor signals before non-linear tone mapping. After camera response curve correction and demosaicing, they can be converted into linear images, which are scene-referred representations that directly reflect true irradiance and are invariant to sensor-specific factors. Since image sensors have better dynamic range and bit depth, linear images contain richer information than display-referred ones, leaving users more room for editing during post-processing. Despite this advantage, current generative models mainly synthesize display-referred images, which inherently limits downstream editing. In this paper, we address the task of ***text-to-linear-image generation****: synthesizing a high-quality, scene-referred linear image that preserves full dynamic range, conditioned on a text prompt, for professional post-processing. Generating linear images is challenging, as pre-trained VAEs in latent diffusion models struggle to simultaneously preserve extreme highlights and shadows due to the higher dynamic range and bit depth. To this end, we represent a linear image as a sequence of exposure brackets, each capturing a specific portion of the dynamic range, and propose a DiT-based flow-matching architecture for text-conditioned exposure bracket generation. We further demonstrate downstream applications including text-guided linear image editing and structure-conditioned generation via ControlNet.*

1. Introduction

Most of the images we encounter every day are stylized renditions of the real world: they are the results of complex imaging pipelines where every pixel encodes display colors rather than how bright the underlying scene truly is. These images are typically referred to as display-referred images. In contrast, linear images are scene-referred representations

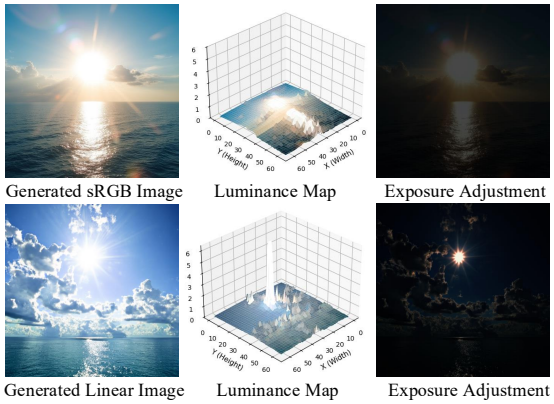


Figure 2. Linear images are scene-referred representations that directly reflect the irradiance at the imaging plane without the nonlinear tone mapping or compression applied in sRGB or HDR images. The high bit depth and dynamic range of the linear image help prevent color banding as marked in the red boxes and enables flexible exposure adjustments without losing highlight details.

that record the scene irradiance at each pixel location on the imaging plane, prior to any nonlinear tone mapping or compression. A linear image can be derived from a RAW capture by correcting the camera response curve, performing demosaicing, and converting to a camera-independent color space. Through these operations, linear images become faithful and sensor-agnostic descriptions of physical radiance, providing a robust foundation for post-processing, relighting as shown in Fig. 1, as well as benefiting some computational photography applications [4, 14, 15, 29].

Modern image generative models [5, 10, 22] are trained almost exclusively on display-referred datasets like LAION-5B [23]. Consequently, they are capable of producing aesthetically pleasing yet tone-mapped results whose brightness and contrast are constrained by display dynamic range. When generating scenes containing both bright highlights and deep shadows, these models tend to produce flattened tone-mapped images with limited post-editing flexibility. By contrast, linear images preserve the full dynamic range of the scene, offering photographers and downstream systems significantly greater room for exposure, white-balance, and tone adjustment without introducing highlight clipping as shown in Fig. 2. Despite the advantages of linear images, research on their generative modeling remains limited. Ideally, a user should be able to generate a linear image directly from a text description, obtaining a scene-referred result ready for professional post-processing without any lossy conversion. At the same time, AI-based image editing tools are widely used, but they are designed for display-referred images and cannot directly process linear images. As a result, users must convert linear images to sRGB before editing, discarding dynamic range information that cannot be recovered after the fact. This gap highlights the need for *text-to-linear-image* generative models and downstream editing tools that natively support linear image formats.

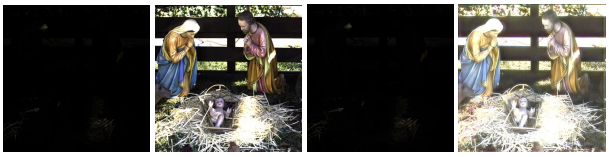


Figure 3. Visual comparison of brightened VAE reconstructed image and exposure-adjusted ground truth. The VAE fails to preserve details in very dark region in linear images, causing severe information loss after the encode-decode process. We use the VAE's float32 precision to mitigate quantization artifacts in this case.

Building a generative model for linear images is inherently challenging. First, data scarcity poses a major bottleneck: scene-referred images typically remain private to photographers, while only their tone-mapped versions are publicly shared. This makes it impractical to train large-scale diffusion models purely from linear data. Second, the high dynamic range and bit depth of linear images exceed the representational capacity of the pre-trained variational autoencoder (VAE) used in latent diffusion models. VAEs are trained to reconstruct display-referred images within limited value ranges; when applied to linear data, they struggle to simultaneously preserve details in both highlights and shadows, as shown in Fig. 3. The result resembles an image captured by a sensor with insufficient dynamic range, where regions of extreme brightness or darkness are clipped or compressed. This analogy motivates our solution: inspired by exposure bracketing in photography, we represent a linear image as a sequence of exposure-specific sub-images, each capturing a distinct portion of the overall dynamic range. By reconstructing these exposure brackets instead of a single high-dynamic-range frame, we circumvent the limitations of existing VAEs and enable reliable synthesis of high-bit-depth content. Specifically, we develop a flow-matching-based generative framework that synthesizes multiple exposure brackets and fuses them into a unified linear image. The framework builds upon Flux [16], enhanced with exposure modulation self-attention and LoRA. We further introduce a radiance scale-token denoising mechanism for joint radiance scale-image generation, enabling explicit prediction of the scene radiance scale. Our proposed exposure modulation self-attention also ensures consistency across exposure brackets, allowing the model to generate well-aligned exposure brackets and scene radiance scale jointly.

Extensive experiments demonstrate that our method produces realistic, physically accurate linear images with rich dynamic range, and enables training-free linear image editing as well as ControlNet-guided conditional generation.

Our main contributions can be summarized as follows: (1) A text-conditioned flow-matching framework for high-bit-depth linear image generation, employing multi-exposure synthesis to expand dynamic range beyond VAE limits. (2) A radiance-scale token denoising mechanism for joint prediction of radiance scale and image content, improving scene radiance reconstruction. (3) Integration of exposure modulation self-attention and LoRA fine-tuning within a DiT backbone, achieving efficient and high-fidelity synthesis. (4) Comprehensive evaluations and applications demonstrating realistic text-to-linear-image generation and downstream tasks including linear image editing.

2. Related Work

HDR Image Reconstruction and Generation. Similar to our method, HDR image reconstruction and generation also aim to recover high dynamic range content. Most existing HDR generation methods [6, 9, 11, 27, 28] adopt a two-stage pipeline, where an LDR image is first synthesized and then converted to HDR via an inverse tone-mapping (ITM) module. Eilertsen *et al*. [9] introduce deep learning to the ITM task by employing convolutional neural networks (CNNs) to enhance highlight details in HDR reconstruction. Following this paradigm, Text2Light [6] and StyleLight [28] focus on HDR panorama generation by first estimating the LDR panorama and subsequently training an ITM network for HDR estimation. LEDiff [27] adopts a fused latent diffusion framework to progressively infer shadow and highlight regions, thereby achieving a more accurate and continuous HDR formation process. Guan *et al*. [11] further propose a two-stage diffusion-based pipeline that first synthesizes standard dynamic range (SDR) images using a pretrained diffusion model and then estimates a gain map via adapted stable diffusion. In contrast to the aforementioned two-stage approaches, several recent methods [1, 26] focus on direct HDR generation without relying on intermediate LDR representations. GlowGAN [26] leverages Gaussian exposure modeling to capture HDR-LDR relationships, enabling unsupervised HDR synthesis and improved reconstruction of over-exposed areas using pretrained generative priors. Bracket Diffusion [1] leverages Denoising Diffusion Probabilistic Models (DDPMs) [13] to directly simulate multi-exposure bracketed imaging for HDR synthesis through test-time optimization. Among these, only GlowGAN and Bracket Diffusion are capable of direct HDR generation. However, GlowGAN is restricted to specific image categories, while Bracket Diffusion takes several minutes to generate a single 256×256 HDR image. Since HDR images are typically produced through tone-mapping operations, their generation priors are difficult to leverage for downstream linear image tasks, and these methods do not support HDR-related editing, limiting their applicability in content creation workflows.

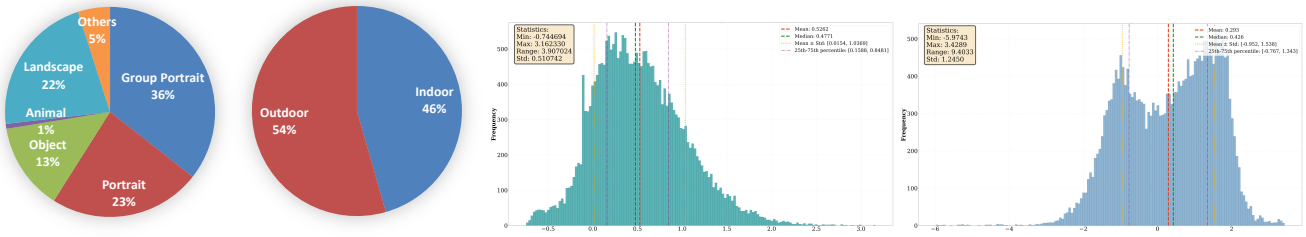
RAW/Linear Image Reconstruction. Although no prior work has directly addressed RAW/Linear image generation, numerous studies have focused on sRGB-to-RAW reconstruction. The same RAW image can be rendered into different sRGB images through various in-camera Image Signal Processing (ISP) pipelines. Thus, sRGB-to-RAW is an ill-posed problem without any RAW-image-prior, most RAW reconstruction methods assume the RAW to sRGB mapping follows a certain pattern to avoid this ill-posedness. UPI [2] assumes that the ISP applies global tone mapping and introduces a method to synthesize realistic RAW data by inverting the standard camera ISP pipeline, enabling effective training of denoising models on unpaired data. CycleISP [34], InvISP [31] and ReverseISP [7] propose paired learning-based sRGB-to-RAW and RAW-to-sRGB networks, which assume the generated sRGB images are produced by a specific invertible ISP pipeline. RAW-Diffusion [21] uses DDPM [13] for sRGB-to-RAW image generation, also following the global tone mapping ISP. Other methods take additional metadata [17, 18, 20, 30, 32] or reference image [19] for sRGB to RAW image reconstruction. The requirements of the metadata and specific ISP can all be attributed to the lack of a RAW image generation prior. Therefore, in this paper, we aim to address this problem through a text-to-linear-image generation framework to establish this kind of prior.

RAW Image Dataset. Unlike sRGB images, RAW data are less publicly available due to their large storage requirements and the need for specialized camera capture settings. The Adobe FiveK dataset [3] is one of the most widely used benchmarks for RAW image processing, containing 5,000 RAW-sRGB image pairs retouched by professional photographers. Another commonly used dataset is RAISE [8], which includes 8,156 high-resolution RAW photographs captured under diverse lighting conditions. In this work, we collect additional RAW images and use them together with RAISE [8] as the training data, while adopting Adobe FiveK [3] as the evaluation set for our experiments.

3. Data Collection and Processing

3.1. Data Preprocessing

Acquiring a large number of linear images is extremely challenging in practice. Moreover, most public HDR datasets are either panoramic (thus focusing almost exclusively on large-scale scene content) or do not provide true linear images, making them unsuitable for our purposes. Therefore, we primarily use RAW image datasets as the basis for training. As discussed in the previous section, we use RAISE [8] and our own collected RAW images as the training set, and adopt the MIT-Adobe FiveK [3] as our test set. To ensure high visual quality, we filter out images with aesthetic scores below 4.5, ultimately retaining 25k images



(a) Category Distribution (b) Scene Distribution (c) Log Space Distribution of I_{linear} Max Luminance (d) Log Space Distribution of Radiance Scale s

Figure 4. Overview of our dataset statistics. We show the overall composition of our dataset in terms of content categories and scene types, along with the distributions (in $\log_{10}$ space) of the radiance scale s and linear image's luminance values. Larger radiance scale means the image is brighter in the physical world. These distributions highlight the large dynamic-range variation of the linear image.

in our training set.

We develop a preprocessing pipeline to convert camera-specific RAW data into camera-independent, scene-referred linear images. Our pipeline consists of four main stages: (1) demosaicing to convert Bayer pattern data to full RGB; (2) applying color correction matrices (CCM) and lookup tables (LUT) to transform sensor signals into camera-independent RGB space; (3) white balance adjustment to standardize color temperature to 5000K; and (4) converting RGB to CIE-XYZ for denoising, then back to linear RGB as linear sensor signals $\mathbf{I}_s$.

Given the processed sensor signal $\mathbf{I}_s$, we recover scene radiance via the standard exposure inversion:

$$\mathbf{L} = \mathbf{I}_s \cdot \frac{F^2}{t \cdot ISO} \cdot 2^{-EV}, \tag{1}$$

where t, ISO, and F denote exposure time, sensor gain, and aperture F-number, and EV is the exposure-compensation setting in stops.

To stabilize training and avoid extremely wide range of radiance values, we normalize $\mathbf{L}$ using a radiance scale s to obtain the normalized linear image $\mathbf{I}_{\text{linear}}$. Instead of enforcing a fixed median value, we compute s from two robust percentile statistics of the radiance distribution of $\mathbf{L}$. Specifically, we estimate a mid-level radiance and a highlight radiance as:

$$m = \text{Percentile}_{0.5}(\mathbf{L}), \qquad h = \text{Percentile}_{0.9}(\mathbf{L}), \tag{2}$$

and derive two candidate radiance scales:

$$s_{\text{med}} = \frac{m}{0.18}, \qquad s_{\text{hi}} = \frac{h}{0.8}. \tag{3}$$

To address cases where large dark backgrounds might cause subject regions to become abnormally bright after normalization, we incorporate the highlight-based radiance scale s_{hi} to help constrain subject exposure. We use the maximum of the two candidate radiance scales as:

$$s = \max(s_{\text{med}},\ s_{\text{hi}}). \tag{4}$$

Finally, the normalized linear image used for training is obtained as:

$$\mathbf{I}_{\text{linear}} = \frac{\mathbf{L}}{s}. \tag{5}$$

This percentile-based normalization effectively controls overall scene radiance while preventing highlight saturation, producing linear images which are suitable for training. The distribution of radiance scale s and $\mathbf{I}_{\text{linear}}$ is shown in Fig. 4, together with the category and scene compositions of our dataset.

3.2. Image Captioning

To create text labels, we use the Qwen2.5-VL 7B [24] with the instruction "Describe the content and details of the image in a direct, concise way, without introductory phrases." to generate captions for the EV0 (base exposure) images in our dataset. The instruction style is chosen to match the caption distribution of the base Flux model's pretraining data. The generated captions serve as the text conditioning signal during training, enabling text-to-linear-image generation.

4. Proposed Method

4.1. Problem Formulation

Our method aims to jointly predict a radiance scale s and its corresponding scene-referred linear image $\mathbf{I}_{\text{linear}} \in \mathbb{R}^{H\times W\times 3}$, both conditioned on input text prompts. Due to the high bit depth and extensive dynamic range of linear images, it is difficult to directly reconstruct linear images using a VAE. To address this challenge, we decompose the linear image into a sequence of exposure brackets, each capturing the scene at a different exposure level. Given a list of exposure values $\text{EV} = [ev_1, ev_2, ..., ev_K]$, each bracket image $\mathbf{I}_k$ is constructed from the normalized linear image:

$$\mathbf{I}_k = \text{clip}(\mathbf{I}_{\text{linear}} \cdot 2^{ev_k},\ 0, 1), \tag{6}$$

where $\text{clip}(\cdot,\ 0, 1)$ restricts the values to the range $[0, 1]$. In this work, we adopt $\text{EV} = [-4, -2, 0, 2]$ as our set of exposure values, resulting in $K = 4$ bracket images derived from each linear image. To enable independent encoding

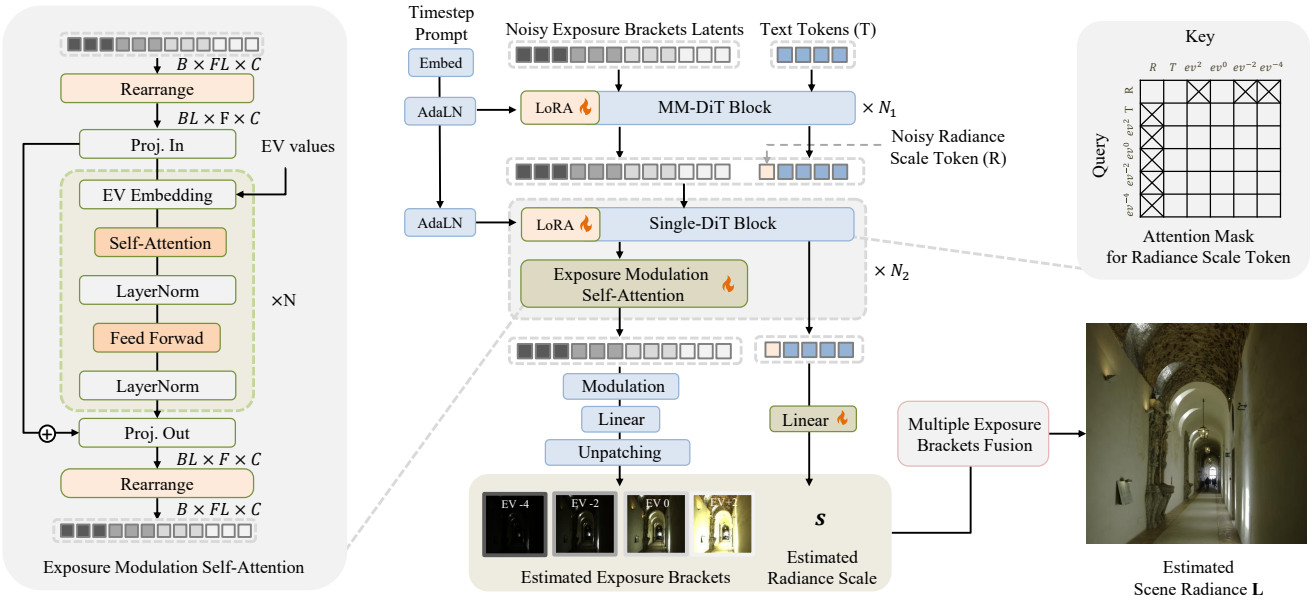


Figure 5. Overview of the proposed framework. The model takes as input the concatenated noise of $K{=}4$ exposure brackets, text prompt, and a noisy radiance-scale token R. These are processed through N_1 MM-DiT blocks (with LoRA), followed by N_2 Single-DiT blocks (with LoRA) that incorporate our Exposure Modulation Self-Attention for exposure-aware feature modulation. The radiance-scale token attends only to text tokens and the EV0 bracket (as shown in the attention mask), and is projected via a linear layer to produce the estimated radiance scale s. The denoised bracket tokens are decoded into K exposure brackets, which are then fused by our Multiple Exposure Brackets Fusion module to yield the final scene-referred linear image $\mathbf{L}$.

through the shared VAE encoder without modification, the K exposure bracket images $\mathbf{I}_k$ are concatenated along the batch dimension and passed through a shared VAE encoder. This produces K latent sequences $\mathbf{z}_k \in \mathbb{R}^{L\times C}$, where L is the sequence length and C is the embedding dimension. These latent sequences are then concatenated along the sequence dimension to form a combined latent representation $\mathbf{z}_{\text{all}} \in \mathbb{R}^{KL\times C}$, which aggregates information from all exposure brackets.

In our approach, we begin by initializing two sets of Gaussian noise: one latent tensor $\mathbf{z}_t \in \mathbb{R}^{KL\times C}$ for the exposure brackets and another noise token of shape $(1\times C)$ for the radiance scale. Through flow-matching denoising, these noisy representations are transformed into the corresponding exposure bracket tokens and a radiance scale token. The exposure bracket tokens are subsequently decoded by the VAE decoder to generate multi-exposure bracket images, while the radiance scale token is projected back to a scalar via a linear layer for radiance scale estimation. This procedure enables the joint generation of both exposure brackets and the global radiance scale.

4.2. Architecture Design

LoRA Finetuning. LoRA fine-tuning adapts the pre-trained Flux backbone to the linear image domain with minimal parameter overhead. It prevents the Single-DiT backbone from destabilizing under the rapidly changing exposure distributions across different bracket frames.

Exposure Modulation Self-Attention. As shown in Fig. 5(b), we introduce Exposure Modulation Self-Attention, which jointly attends across all exposure brackets. This mechanism allows the model to flexibly adjust luminance for each bracket while preserving structural alignment and detail consistency across exposures.

3D-RoPE for Sequence Disentanglement. We use 3D Rotary Positional Embedding (3D-RoPE) to encode both spatial position and bracket identity into each token, helping the model distinguish and correctly process tokens from different exposure brackets. In the standard Flux formulation, each image token is assigned a 2D positional coordinate (i, j) corresponding to its spatial location in the image grid. To extend this to multi-bracket generation, we augment the positional encoding to a 3D tuple (index, i, j), where i and j retain their original spatial meaning, and index is set to the bracket index $k \in \{0, 1, \ldots, K{-}1\}$ of the exposure frame to which the token belongs. This design allows the model to unambiguously identify which exposure bracket each token belongs to while preserving intra-bracket spatial structure, enabling effective disentanglement of luminance levels across the jointly attended sequence.

4.3. Radiance Scale Tokenization and Prediction

For quantization and prediction, we use $s_l = \log_{10}(s)$, the log-transformed radiance scale. We discretize the ground-truth log-radiance s_l into 20 uniform bins over the interval $[-6, 4]$. Each s_l value is mapped to a one-hot discrete

Figure 6. Comparison of our proposed method with LoRA finetuning on Flux and Wan [25]. Unlike other baseline methods, our approach effectively preserves shadow details even in over-exposed situations.

code $\mathbf{s}_d \in \{0,1\}^{20}$ indicating its quantized bin. This discrete code is then embedded into the diffusion token space through a shared linear projection:

$$\mathbf{t}_s = W\,\mathbf{s}_d, \qquad W \in \mathbb{R}^{20\times d}, \tag{7}$$

which is then jointly updated with image tokens during self-attention. At inference time, the updated radiance token $\hat{\mathbf{t}}_s$ is projected back to the bin space:

$$\hat{\mathbf{s}}_d = \mathrm{softmax}(\hat{\mathbf{t}}_s W^\top), \tag{8}$$

and the predicted log-radiance $\hat{s}_l$ is obtained as the expectation over the 20 bin centers μ_i:

$$\hat{s}_l = \sum_{i=1}^{20} \hat{\mathbf{s}}_d[i]\;\mu_i. \tag{9}$$

If recovering the predicted radiance scale s is needed, it can be computed by $s = 10^{\hat{s}_l}$.

To ensure robust radiance scale prediction, we carefully design an attention mask for the radiance scale token, as illustrated in Fig. 5. Specifically, the radiance scale token is only allowed to attend to the text tokens and the tokens corresponding to the EV0 (base exposure) bracket. This allows the model to derive a global scene radiance primarily from the reference exposure, mitigating the risk of being influenced by the over- or under-exposed bracket images. Conversely, the other image and text tokens do not attend to the radiance scale token, aiming to maintain the fidelity of generated images. Additionally, the radiance scale token does not undergo 3D-RoPE positional embedding, as it does not represent a spatial location or exposure index.

4.4. Training Objective

Our model is trained under a flow-matching formulation. Let $\mathbf{z}_t$ denote the latent at continuous time $t \in [0,1]$, and let $\phi_t(\mathbf{z}_0, \mathbf{z}_1)$ be the ground-truth probability flow that maps clean latents $\mathbf{z}_0$ to noise $\mathbf{z}_1$. Flow matching constructs a velocity field $u_t(\mathbf{z}_t)$ satisfying the probability flow ODE

$$\frac{d\mathbf{z}_t}{dt} = u_t(\mathbf{z}_t), \tag{10}$$

and the model learns to approximate this field via

$$\mathcal{L}_{\text{img}} = \mathbb{E}_{t,\mathbf{z}_0,\mathbf{z}_1}\left[\|u_t(\mathbf{z}_t) - u_\theta(\mathbf{z}_t, \mathbf{c}, t)\|_2^2\right]. \tag{11}$$

The radiance-scale token is supervised in the same velocity space: if $\mathbf{t}_{s,t}$ and $\hat{\mathbf{t}}_{s,t}$ denote the ground-truth and predicted radiance tokens following the probability flow, then

$$\mathcal{L}_{\text{rad}} = \left\|u_t(\mathbf{t}_{s,t}) - u_\theta(\hat{\mathbf{t}}_{s,t})\right\|_2^2, \tag{12}$$

where u_t denotes analytically defined token-space velocity derived from the quantized radiance interpolation schedule.

To ensure physically consistent exposure ordering among the generated exposure brackets, we further impose a bracket-consistency loss computed in pixel space. After denoising, we obtain the predicted clean latents $\hat{\mathbf{z}}_0$, decode them with the VAE, reshape them into (B, F, C, H, W), and enforce that all frames align with the reference exposure (EV0). Let $\hat{\mathbf{I}}_k$ denote the predicted frame at exposure ev_k and let k_0 be the index where $ev_k = 0$, so that $\hat{\mathbf{I}}_{k_0}$ is the predicted EV=0 (base exposure) bracket serving as the reference. We compute

$$\mathcal{L}_{\text{bracket}} = \sum_k \left\| \frac{\hat{\mathbf{I}}_k}{2^{ev_k}} - \hat{\mathbf{I}}_{k_0} \right\|_1, \tag{13}$$

which enforces the expected multiplicative radiance ratios between bracketed frames without relying on masking or thresholding. The total loss is defined as:

$$\mathcal{L} = \mathcal{L}_{\text{img}} + \lambda_{\text{rad}}\mathcal{L}_{\text{rad}} + \lambda_{\text{bracket}}\mathcal{L}_{\text{bracket}}, \tag{14}$$

where λ_{rad} and λ_{bracket} are set to 1.0 and 0.5 respectively.

4.5. Multiple Exposure Brackets Fusion

At inference time, we synthesize K exposure-bracketed images $\hat{\mathbf{B}} = \{\hat{\mathbf{I}}_1, \ldots, \hat{\mathbf{I}}_K\}$ (from darkest to brightest) and subsequently fuse them to obtain the final linear image $\hat{\mathbf{I}}_{\text{linear}}$. For each pair of consecutive exposure images $\hat{\mathbf{I}}_k$ and $\hat{\mathbf{I}}_{k+1}$, we compute the mean intensity of each channel in non-saturated regions for each frame. We then compute a per-channel ratio vector $r_k \in \mathbb{R}^3$ by dividing the mean value of each channel in $\hat{\mathbf{I}}_{k+1}$ by the corresponding mean in $\hat{\mathbf{I}}_k$, where the mean is taken over valid (non-saturated) pixels. This three-channel ratio ensures accurate alignment of brightness transitions across the exposure levels.

The fusion process proceeds in a hierarchical fashion, starting from the brightest bracket and iteratively incorporating information from darker brackets. At each step k ($k = K-1, ..., 1$), a soft mask $\mathbf{M}_k \in [0,1]^{H\times W\times 3}$ is constructed to delineate spatial regions where the current bracket provides unsaturated, high-fidelity information, with smoothing to avoid blending artifacts. The fusion at each stage is then performed as a weighted combination:

$$\hat{\mathbf{I}}_{\text{fused}} \leftarrow \hat{\mathbf{I}}_{\text{fused}} \cdot (1 - \mathbf{M}_k) + \left(\hat{\mathbf{I}}_k \cdot r_k\right) \cdot \mathbf{M}_k, \qquad (15)$$

where $\hat{\mathbf{I}}_k \cdot r_k$ denotes per-channel multiplication to align the luminance of $\hat{\mathbf{I}}_k$ with the fused result. The fused image is initialized as $\hat{\mathbf{I}}_K$ and updated recursively. This method enables the reconstruction of a high-dynamic-range linear image from multiple exposure-biased generations, effectively recovering highlight and shadow details while ensuring radiometric consistency and smooth transitions.

5. Experiments

5.1. Training Details

In our experiment, we use Flux-dev as our framework and conduct training on 4 NVIDIA A100 80GB GPUs. The batch size is 4 per GPU, and we train for 10,000 iterations in total. For LoRA finetuning, we set the rank to 64 and α to 128. In our Exposure Modulation Self-Attention module, we use 8 attention heads, an inner dimension of 512, and set the exposure modulation attention head dimension to 64. We adopt the AdamW optimizer, setting the learning rate for the exposure modulation module to 2×10^{-5} and the learning rate for LoRA parameters to 1×10^{-4}. All training is performed using bf16 precision to improve computational efficiency and reduce memory consumption.

5.2. Comparison with Previous Methods

Existing works have not addressed the task of direct RAW or linear image generation, as discussed in Section 2. While some recent methods focus on HDR image generation, such as GlowGAN [26], their applicability is limited: GlowGAN can only generate HDR images for specific image categories and does not produce true linear images suitable for direct linear image workflows. Moreover, most prior HDR generation approaches operate in display-referred or tone-mapped spaces rather than predicting scene-referred linear content. Due to the lack of direct competitors for linear image generation, we mainly compare our method with strong baselines by adapting existing architectures. Specifically, we conduct the following three types of comparisons:

1. **T2I Finetuning:** We finetune the state-of-the-art text-to-image diffusion model Flux [16] using LoRA on our dataset, training the model to generate linear images directly, rescaled to the $[0, 1]$ range. This allows us to evaluate how a state-of-the-art T2I model adapts to the linear image domain.
2. **T2V Finetuning:** We also test using text-to-video model Wan 2.1 [25], whose VAE is capable of compressing $4T + 1$ frames into T latent representations. In our setting, we select the EV=0 frame as the reference and compress the 4 exposure brackets using Wan's VAE encoder into a single latent map. We then finetune a LoRA module to generate such latent maps, which are subsequently decoded back into exposure brackets. The same dataset as used for our baseline experiments is adopted for this finetuning procedure.
3. **T2I Model Inflation:** We further compare against CameraCtrl [12] and Generative Photography [33], which inflate image diffusion architectures with temporal modules to generate multi-frame sequences. Since these methods are not designed for linear image generation, we fine-tune them on our training set for a fair comparison (denoted "w/ F").

Table 1 provides a comprehensive quantitative comparison, and Fig. 6 and Fig. 8 show qualitative results. FID and LS are computed on linear images, while AS, NIQE, and CLIP Sim. are evaluated on the EV0 frame. Due to the wide distribution of linear images, directly finetuning T2I Model on linear data makes it difficult to balance shadow and highlight details. T2I Model Inflation methods suffer from both limited dynamic range and significant image quality degradation even after fine-tuning. For T2V Finetuning, Wan 2.1's $4\times$ temporal downsampling entangles the 4 exposure brackets into a single latent representation, causing a severe distribution mismatch that cannot be resolved through fine-tuning alone. By directly modeling scene-referred properties using exposure brackets, our method achieves superior visual quality and dynamic range across all baselines.

5.3. Radiance Scale Estimation

Accurate radiance scale estimation requires integrating both semantic context from text tokens and spatial luminance cues from image tokens. Our token denoising strategy achieves this naturally: the radiance scale token participates in the joint self-attention over all tokens during denoising, allowing it to simultaneously attend to text descriptions and

Figure 7. More visualization results generated by our method. Our approach supports the generation of images with various styles.

Type	Model	FID ↓	AS ↑	NIQE ↓	CLIP Sim.↑	LS ↑
T2I Finetuning	Flux [16]	32.12	4.712	5.304	25.90	/
T2V Finetuning	Wan 2.1 [25]	/	4.537	5.412	24.79	1.12
T2I Model Inflation	CameraCtrl [12] (w/ F)	37.25	5.230	4.131	**26.89**	8.97
	Gen. Photography [33] (w/ F)	40.17	4.619	4.514	23.71	7.11
	Gen. Photography [33] (w/o F)	43.83	3.909	4.870	20.51	5.56
	Ours	**28.29**	**5.700**	**3.658**	26.02	**23.06**

Table 1. Comprehensive quantitative comparison with all baselines. T2I/T2V Finetuning directly adapt existing generative models to linear image generation via LoRA. T2I Model Inflation methods extend image diffusion architectures with temporal modules and are fine-tuned on our training set (w/ F). "/" indicates the metric cannot be meaningfully computed: Wan 2.1 cannot produce consistent exposure brackets required for HDR fusion (LS), and T2I Finetuning lacks the temporal capacity to span the full dynamic range (LS). "F" denotes fine-tuning on our training set. "Gen. Ph." denotes Generative Photography [33].

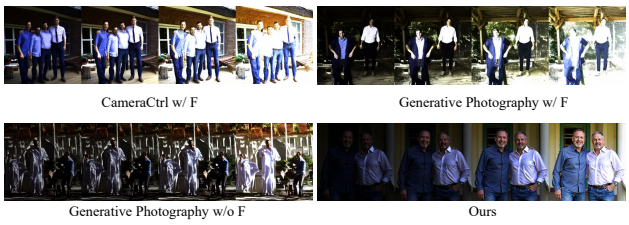


Figure 8. Visual comparison with CameraCtrl and Generative Photography. Our method produces better-quality exposure brackets with more consistent luminance transitions.

image content. To validate this design, we compare against a global pooling baseline, where pooled output tokens are passed through an MLP for radiance scale prediction. We test three variants: using only text tokens, only image tokens, or their concatenation. As shown in Table 2, the token denoising method better leverages both text and image information simultaneously, leading to higher estimation accuracy than all global pooling alternatives.

	Text-MLP	Image-MLP	Merge-MLP	Ours
MAE ↓	0.782	1.213	0.792	**0.737**

Table 2. Quantitative comparison of radiance scale estimation methods using MAE.

5.4. More Analysis

As shown in Fig. 7, our method can generate exposure brackets with diverse visual styles. More visualization results, ControlNet-based generation, linear image inpainting, and text-guided linear image editing are provided in the supplementary materials. Additional ablation studies covering positional encoding, number of exposure brackets, model architecture design, and EV injection strategy are also provided in the supplementary materials.

6. Conclusion

In this paper, we present a novel generative framework for linear image synthesis with high dynamic range. Our method adopts a flow-matching paradigm with multi-exposure bracket prediction to overcome the limitations of VAE models in modeling scene radiance, enabling improved dynamic range and radiance fidelity in the generated images. Building upon a DiT backbone, we introduce exposure modulation self-attention and radiance-scale token denoising, which ensure accurate generation across diverse exposure levels and enable explicit prediction of scene radiance scale. Our approach achieves efficient adaptation on limited linear data while maintaining image quality. Additionally, the framework enables downstream applications including linear image editing and ControlNet-based conditional generation, bridging the gap between generation model and professional photography workflows.

References

[1] Mojtaba Bemana, Thomas Leimkühler, Karol Myszkowski, Hans-Peter Seidel, and Tobias Ritschel. Bracket Diffusion: Hdr image generation by consistent ldr denoising. In *Computer Graphics Forum*, 2025. 3

[2] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *CVPR*, 2019. 3

[3] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *CVPR*, 2011. 3

[4] Haoming Cai, Tsung-Wei Huang, Shiv Gehlot, Brandon Y Feng, Sachin Shah, Guan-Ming Su, and Christopher Metzler. Parametric shadow control for portrait generation in text-to-image diffusion models. In *ICCV*, 2025. 2

[5] Junsong Chen, Chongjian Ge, Enze Xie, Yue Wu, Lewei Yao, Xiaozhe Ren, Zhongdao Wang, Ping Luo, Huchuan Lu, and Zhenguo Li. Pixart-σ: Weak-to-strong training of diffusion transformer for 4k text-to-image generation. In *ECCV*. Springer, 2024. 2

[6] Zhaoxi Chen, Guangcong Wang, and Ziwei Liu. Text2Light: Zero-shot text-driven hdr panorama generation. *ACM TOG*, 41(6):1–16, 2022. 3

[7] Marcos V Conde, Radu Timofte, Yibin Huang, Jingyang Peng, Chang Chen, Cheng Li, Eduardo Pérez-Pellitero, Fenglong Song, Furui Bai, Shuai Liu, et al. Reversed image signal processing and raw reconstruction. aim 2022 challenge report. In *ECCVW*, 2022. 3

[8] Duc-Tien Dang-Nguyen, Cecilia Pasquini, Valentina Conotter, and Giulia Boato. RAISE: A raw images dataset for digital image forensics. In *ACM Multimedia Systems*, 2015. 3

[9] Gabriel Eilertsen, Joel Kronander, Gyorgy Denes, Rafał K Mantiuk, and Jonas Unger. Hdr image reconstruction from a single exposure using deep cnns. *ACM TOG*, 36(6):1–15, 2017. 3

[10] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024. 2

[11] Yuanshen Guan, Ruikang Xu, Yinuo Liao, Mingde Yao, Lizhi Wang, and Zhiwei Xiong. HDR image generation via gain map decomposed diffusion. In *ICCV*, 2025. 3

[12] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. CameraCtrl: Enabling camera control for video diffusion models. In *ICLR*, 2025. 7, 8

[13] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *NeurIPS*, 2020. 3

[14] Haofeng Huang, Wenhan Yang, Yueyu Hu, Jiaying Liu, and Ling-Yu Duan. Towards low light enhancement with raw images. *IEEE TIP*, 31:1391–1405, 2022. 2

[15] Eric Kee, Adam Pikielny, Kevin Blackburn-Matzen, and Marc Levoy. Removing reflections from raw photos. In *CVPR*, 2025. 2

[16] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 2, 7, 8

[17] Seonghyeon Nam, Abhijith Punnappurath, Marcus A Brubaker, and Michael S Brown. Learning srgb-to-raw-rgb de-rendering with content-aware metadata. In *CVPR*, 2022. 3

[18] Rang MH Nguyen and Michael S Brown. RAW image reconstruction using a self-contained sRGB-JPEG image with only 64 kb overhead. In *CVPR*, 2016. 3

[19] Junji Otsuka, Masakazu Yoshimura, and Takeshi Ohashi. Self-supervised reversed image signal processing via reference-guided dynamic parameter selection. *CoRR*, 2023. 3

[20] Abhijith Punnappurath and Michael S Brown. Spatially aware metadata for raw reconstruction. In *WACV*, 2021. 3

[21] Christoph Reinders, Radu Berdan, Beril Besbinar, Junji Otsuka, and Daisuke Iso. RAW-Diffusion: RGB-Guided Diffusion Models for High-Fidelity RAW Image Generation. In *WACV*, 2025. 3

[22] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, 2022. 2

[23] 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*, 2022. 2

[24] Qwen Team. Qwen2.5-vl, 2025. 4

[25] Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025. 6, 7, 8

[26] Chao Wang, Ana Serrano, Xingang Pan, Bin Chen, Karol Myszkowski, Hans-Peter Seidel, Christian Theobalt, and Thomas Leimkühler. GlowGAN: Unsupervised learning of hdr images from ldr images in the wild. In *ICCV*, 2023. 3, 7

[27] Chao Wang, Zhihao Xia, Thomas Leimkuhler, Karol Myszkowski, and Xuaner Zhang. LEDiff: Latent exposure diffusion for HDR generation. In *CVPR*, 2025. 3

[28] Guangcong Wang, Yinuo Yang, Chen Change Loy, and Ziwei Liu. StyleLight: Hdr panorama generation for lighting estimation and editing. In *ECCV*, 2022. 3

[29] Tianfu Wang, Mingyang Xie, Haoming Cai, Sachin Shah, and Christopher A Metzler. Flash-split: 2d reflection removal with flash cues and latent diffusion separation. In *CVPR*, 2025. 2

[30] Yufei Wang, Yi Yu, Wenhan Yang, Lanqing Guo, Lap-Pui Chau, Alex C Kot, and Bihan Wen. Raw image reconstruction with learned compact metadata. In *CVPR*, 2023. 3

[31] Yazhou Xing, Zian Qian, and Qifeng Chen. Invertible image signal processing. In *CVPR*, 2021. 3

[32] Lu Yuan and Jian Sun. High quality image reconstruction from raw and jpeg image pair. In *ICCV*, 2011. 3

[33] Yu Yuan, Xijun Wang, Yichen Sheng, Prateek Chennuri, Xingguang Zhang, and Stanley Chan. Generative photography: Scene-consistent camera control for realistic text-to-image synthesis. In *CVPR*, 2025. 7, 8

[34] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, Ming-Hsuan Yang, and Ling Shao. CycleISP: Real image restoration via improved data synthesis. In *CVPR*, 2020. 3