Linear Image Generation by Synthesizing Exposure Brackets

Supplementary Material

A. Additional Ablation Studies

This section presents additional ablation studies of our method, investigating the influence of positional encoding, the number of exposure brackets, key architectural components, and alternative exposure modulation choices on generation quality and multi-exposure consistency. We provide both quantitative results and qualitative analyses to further validate the effectiveness and necessity of our proposed modules.

Positional Encoding. As shown in Table 1, positional encoding plays a crucial role in separating frame brightness. Using 2D RoPE alone makes tokens at the same spatial location across different exposure brackets difficult to distinguish, leading to noticeable checkerboard-like artifacts. Introducing a Layer Embedding (LE) significantly improves performance, indicating that explicit bracket identity helps resolve token ambiguity. However, when 3D RoPE is used, adding LE offers no clear improvement. This suggests that 3D RoPE already encodes bracket identity effectively through positional embedding. Therefore, we adopt pure 3D RoPE as our default design.

Method	AS ↑	NIQE ↓	LS ↑
2D RoPE	4.099	4.332	4.10
2D RoPE + LE	**5.792**	3.853	6.21
3D RoPE + LE	5.695	3.721	21.07
3D RoPE (Ours)	5.700	**3.658**	**23.06**

Table 1. Ablation study on positional encoding methods. LE represents Layer Embedding. 3D RoPE achieves the best performance across all metrics. LS represents the luminance scale which measures the ratio between the brightest and darkest images.

Number of Exposure Brackets. To determine the optimal number of exposure brackets, we test how different bracket number influences image generation quality and the luminance scale (LS) between the brightest and darkest frames. For all configurations, we set the exposure value (EV) interval between each adjacent frame to 2. Specifically, for 2 brackets we use EVs of $[-2, 0]$; for 3 brackets: $[-2, 0, 2]$; for 4 brackets: $[-4, -2, 0, 2]$; and for 5 brackets: $[-6, -4, -2, 0, 2]$. As shown in Table 2, using 4 exposure brackets achieves the best overall performance. Although 5 brackets can slightly improve luminance scale, increasing the number of generated brackets in fact leads to a degradation in image quality, as shown in Fig. 1.

Model Architecture Ablation. We evaluate different model architecture design choices including the impact of LoRA [4] module and exposure modulation components.

# Brackets	AS ↑	NIQE ↓	LS ↑	CLIP Sim.↑
2	**5.820**	3.864	4.76	25.68
3	5.543	4.000	7.00	25.62
4 (Ours)	5.700	**3.658**	23.06	**26.02**
5	5.258	4.294	**25.76**	23.92

Table 2. Quantitative comparisons on the number of exposure brackets. 4 brackets achieve the best trade-off between image quality and dynamic range.

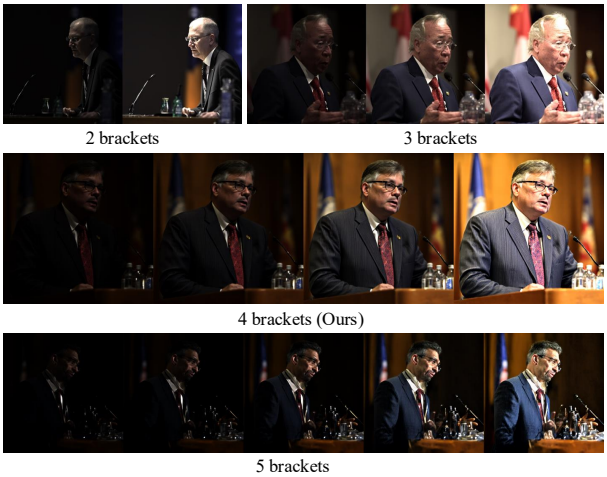


Figure 1. Visual comparison of models trained with different numbers of exposure brackets. Increasing the number of brackets expands the dynamic range but may degrade image quality. Training with more than five brackets tends to introduce noticeable artifacts; therefore, we adopt four brackets as our baseline, offering a balanced trade-off between dynamic range and visual fidelity.

Table 3 shows that both LoRA and our exposure modulation module are essential for optimal performance. As shown in Fig. 2, removing LoRA results in severe distorted global structure; removing the exposure modulation module causes poor contrast across brackets. Since the MM-DiT [2] branch at the early stage of Flux [6] is mainly responsible for modeling the overall structure of the image, introducing exposure modulation module at this stage destabilizes training and leads to structural distortions. Therefore, we only apply our exposure modulation module to the single-DiT component, which enables the model to maintain consistent global structure across all brackets, while allowing fine-grained control of luminance and detail alignment for each bracket.

Exposure Modulation Methods Comparison. To enable effective exposure modulation, we systematically explore three major approaches: (1) considering exposure value as conditions like text prompt or time. (2) injecting explicit

Method	AS ↑	NIQE ↓	LS ↑	CLIP Sim.↑
w/o LoRA	3.183	3.899	52.58	17.88
w/o Modulation	**6.245**	4.143	2.13	25.54
Modulation on MM-DiT	4.937	3.987	**86.61**	24.15
Ours	5.700	**3.658**	23.06	**26.02**

Table 3. Quantitative comparisons on model architecture ablation. Our exposure modulation module, when applied specifically to the single-DiT architecture, is crucial for stable multi-exposure generation and coherent exposure transitions across brackets.

(a) w/o LoRA

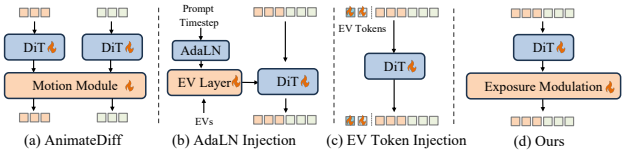
(b) w/o Modulation

(c) Modulation also on MM-DiT

(d) Ours

Figure 2. Visual comparison of generation results under different ablation settings. Without LoRA (a), the model becomes unstable during training due to the large distribution shifts across exposure levels in the sequence, leading to degraded and inconsistent outputs. Without modulation (b), the model produces visually plausible images but lacks meaningful exposure control. Applying modulation only to the MM-DiT branch (c) enhances exposure variation but introduces structural distortions. Our full method (d) achieves stable training, preserves structural fidelity, and produces coherent and physically plausible exposure transitions.

exposure tokens to the sequence, and (3) introducing additional dedicated network modules for modulation as shown in Fig. 3. To ensure stable training and robust adaptation across different methods, we consistently add a LoRA module with rank 64 on the main DiT backbone in all experiments. We provide a comprehensive evaluation of these exposure modulation strategies in Table 4 and Fig. 5. Specifically, the AnimateDiff-style approach (a) encodes exposure information implicitly by concatenating exposure brackets along the batch dimension, relying on the motion module



(a) AnimateDiff (b) AdaLN Injection (c) EV Token Injection (d) Ours

Figure 3. Comparison of different EV injection strategies for exposure modulation. In (a) AnimateDiff [3], exposure information is implicitly encoded by concatenating brackets along the batch dimension, followed by sequence alignment using the motion module, but without explicit exposure conditioning. In (b) AdaLN [10] Injection, EV guidance is introduced by appending a modulation layer after each LayerNorm in DiT, which adjusts the scaling and shifting parameters (α, β, γ) in an exposure-aware manner. In (c) EV Token Injection, learnable EV tokens are inserted into the token sequence, and an attention mask constrains each bracket to attend only to its corresponding EV token. In (d) Ours, we propose an exposure modulation module that directly adjusts feature activations based on EVs, enabling stable training and fine-grained, spatially coherent exposure control across brackets.

Method	AS ↑	NIQE ↓	LS ↑	CLIP Sim.↑
AnimateDiff	4.521	6.201	2.34	23.94
AdaLN Injection	4.481	5.098	14.86	21.15
EV Token	5.372	4.291	18.04	24.23
Ours	**5.700**	**3.658**	**23.06**	**26.02**

Table 4. Detailed comparison of EV injection methods. Our exposure modulation self-attention mechanism demonstrates superior performance in maintaining exposure consistency while preserving image quality.

for sequence alignment. The AdaLN-based method (b) inserts a zero-initialized modulation layer after each LayerNorm [1] in the DiT blocks. This EV modulation layer takes the scalar exposure value (EV) as input, encodes it via Fourier features followed by a lightweight MLP, and produces bracket-specific FiLM [11] parameters $(\Delta\gamma, \Delta\beta)$ as well as gate offsets (Δgate). These parameters are then applied to the normalized hidden features and gating signals in a residual manner, enabling exposure-dependent adjustment. All EV-dependent offsets are zero-initialized and magnitude-constrained, ensuring that the model initially behaves identically to the original DiT and gradually learns stable exposure-aware modulation during training. The EV Token Injection approach (c) introduces learnable exposure tokens and leveraging attention masks, but can destabilize image structure (see Fig. 3(c) and Table 4). In contrast, our proposed exposure modulation self-attention module (d) directly modulates features according to exposure values, producing stable, spatially consistent, and physically plausible exposure transitions, consistently achieving the best results in both exposure alignment and image quality.

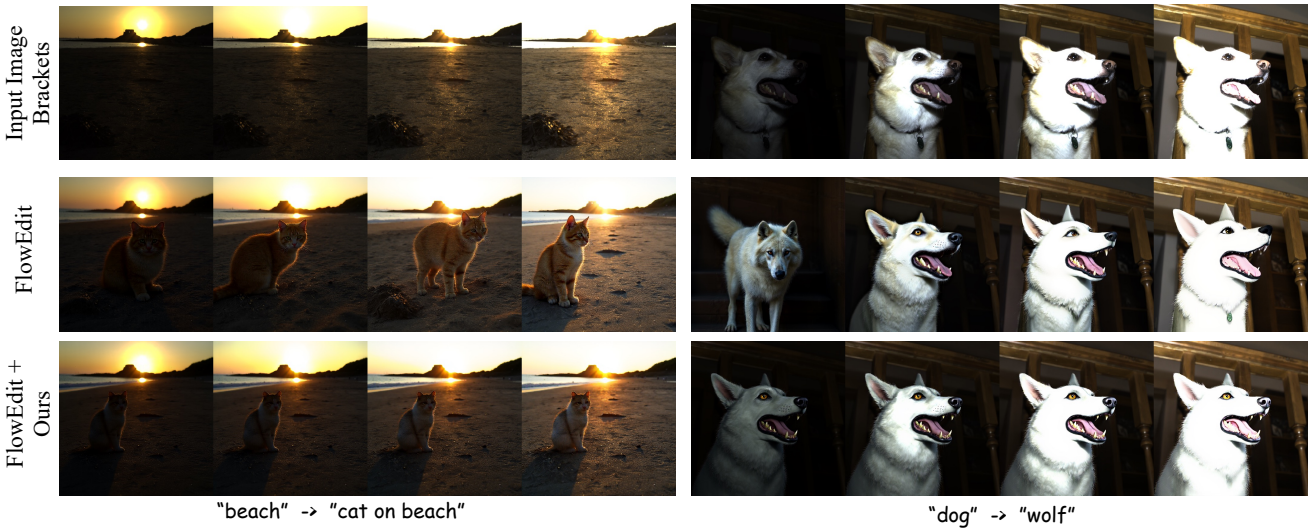
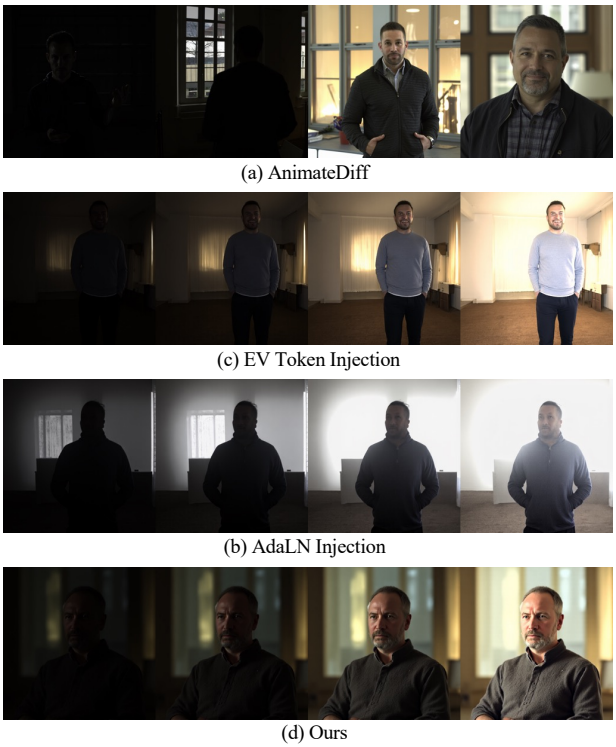


Figure 4. Visualization results of our method with FlowEdit [5]. By integrating FlowEdit, our method enables intuitive and consistent editing across different exposure brackets without finetuning. Without our method, FlowEdit struggles to achieve consistent edits for the various exposure brackets, making it unsuitable for linear image editing.



(a) AnimateDiff

(c) EV Token Injection

(b) AdaLN Injection

(d) Ours

Figure 5. Visual comparison of various exposure modulation methods. Unlike other approaches, our method produces consistently aligned content across different exposure levels while preserving high image quality.

Figure 6. Linear image inpainting results. By combining our exposure-aware generation model with RePaint, missing regions across exposure brackets can be faithfully reconstructed. The inpainted content is well-aligned with the surrounding structures and maintains consistent exposure relationships across both shadow and highlight areas.

B. Post-Editing Flexibility

A key motivation for linear image generation is enabling downstream post-editing with greater flexibility than display-referred or HDR images. To quantify this, we compare post-editing capabilities against HDR generation methods SingleHDR [7] and LEDiff [12] in display space. All methods are tone-mapped to display space using a global Reinhard operator with $\gamma = 2.2$. Since SingleHDR and LEDiff follow an image-to-HDR paradigm (requiring an SDR input), we use our EV0 bracket output as their input to ensure aligned comparison. We then apply brightness adjustments (± 2 EV) and white balance shifts (± 2000 K) to each method's output, reporting FID against unedited ground-truth images.

As shown in Table 5, our method outperforms prior HDR generation methods on all post-editing metrics, demonstrating the advantage of generating scene-referred linear images for downstream editing. "Ours (SDR)" denotes

Figure 7. ControlNet-based linear image generation results with Canny edge guidance. Our method enables the synthesis of diverse exposure brackets with varied content while preserving consistency with the given edge condition, demonstrating strong controllability and flexibility in content generation.

Figure 8. Reference-based HDR exposure brackets generation from an SDR image. Starting from a display-referred SDR image (left), we apply a gamma correction to approximate its linear domain representation and synthesize a set of pseudo-linear reference exposure brackets (middle-top). Although these brackets have limited dynamic range, they serve as structural and exposure-aware guidance. Using an SDEdit [9] noise–denoise process, we progressively inject noise into the reference brackets and perform conditional denoising to generate HDR exposure brackets with enriched dynamic range (middle-bottom). By adjusting the noise strength during SDEdit, we can control the trade-off between fidelity to the reference exposure structure and the generation of plausible high-dynamic-range content (right). This enables reference-guided HDR bracket generation while preserving semantic consistency.

our EV0 frame rendered directly without HDR fusion; its degraded performance after large edits confirms that full bracket generation and fusion is essential for robust post-editing.

Method	AS ↑	FID ↓	FID +2EV ↓	FID -2EV ↓	FID +2000K ↓	FID -2000K ↓
SingleHDR [7]	5.781	31.57	29.17	32.84	31.82	30.78
LEDiff [12]	5.806	29.94	28.67	28.81	30.34	28.98
Ours (SDR)	5.700	32.26	31.14	38.10	32.15	32.31
Ours	**5.819**	**27.87**	**26.40**	**28.19**	**27.96**	**28.28**

Table 5. Post-editing comparison with HDR generation methods in display space (after Reinhard tone-mapping with $\gamma = 2.2$). FID is reported before and after ± 2 EV brightness and ± 2000 K white balance adjustments. Our method consistently achieves lower FID after editing, demonstrating the advantages of linear-space generation for post-processing.

C. Downstream Applications

Linear Image Inpainting. In addition to exposure bracket generation, our model can be directly applied to linear image inpainting. By leveraging the exposure-aware representation, the model is able to restore masked regions in each bracket while preserving the relative exposure relationships across brackets. As shown in Fig. 6, combining our model with RePaint [8] enables structurally coherent and exposure-consistent inpainting across a wide dynamic range. The reconstructed regions seamlessly blend with the original content under both underexposed and overexposed conditions, demonstrating the applicability of our approach to linear and HDR inpainting tasks.

Linear Image Editing. Our method can also be used to edit linear images using a training-free flow-matching image editing method. We use FlowEdit [5] to edit linear images by first decomposing them into multiple exposure brackets, applying aligned edits on each bracket using our pretrained model, and then fusing them back with our multiple exposure brackets fusion strategy. This approach ensures edits are consistent and radiometrically correct across exposures. As shown in Fig. 4, by integrating FlowEdit, our method enables intuitive and consistent editing across different exposure brackets without finetuning. Without our method, FlowEdit struggles to achieve consistent edits across the various exposure brackets, making it unsuitable for linear image editing.

ControlNet-based Linear Image Generation. We demonstrate the compatibility of our method with ControlNet [13] guidance for conditional linear image generation. Fig. 7 shows results with different control conditions. By controlling the exposure brackets, we can ultimately synthesize a linear image that adheres to the given condition. Since Flux's ControlNet is primarily designed for 1024×1024 resolution, occasional misalignment may still occur at lower resolutions.

HDR Rendering. After generating a set of exposure brackets with our method, we can synthesize a linear image by merging these brackets in linear space. To produce a display-ready HDR image, we apply gamma correction to the merged linear result, effectively mapping it into the appropriate non-linear space for visualization or HDR export. This workflow allows for the preservation of high dynamic range and accurate scene representation in the final HDR output. We have attached a website containing HDR images rendered from our generated linear images.

Reference-based HDR Image Generation. As illustrated in Fig. 8, our approach also enables reference-based HDR image generation from standard SDR images. We simulate pseudo-linear exposure brackets from an input SDR image, then use an SDEdit-style strategy: noise is injected into these brackets, and our model leverages its generative prior to reconstruct the noisy latents back into the linear domain. The final outputs can then be merged and mapped to produce an HDR image, enriching the dynamic range while preserving semantic consistency with the original SDR reference.

D. Additional Qualitative Results

Fig. 9 demonstrates our method's capability to generate linear images across various artistic styles while maintaining proper exposure relationships across brackets.

E. Limitations

While our method achieves promising results for linear image generation, several limitations remain: Since Flux itself cannot guarantee high-quality generation for high-resolution images (e.g., 2K), increasing the number of exposure brackets further affects image quality. When the frame number becomes too large, it becomes difficult to maintain stable generation even at 1024 resolution, leading to inconsistent exposure relationships and potential artifacts across brackets. Besides, our training dataset has inherent limitations in terms of aesthetic scores, which are not particularly high, and consists entirely of realistic photographic scenes. Consequently, when generating non-realistic objects such as robots, mechanical structures, or highly stylized content, the model's performance will be affected.

References

[1] Jimmy Lei Ba, Jamie Ryan Kiros, and Geoffrey E Hinton. Layer normalization. *arXiv preprint arXiv:1607.06450*, 2016. 2

[2] 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. 1

[3] Yuwei Guo, Ceyuan Yang, Anyi Rao, Zhengyang Liang, Yaohui Wang, Yu Qiao, Maneesh Agrawala, Dahua Lin, and Bo Dai. Animatediff: Animate your personalized text-to-image diffusion models without specific tuning. In *ICLR*, 2024. 2

[4] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. *ICLR*, 2022. 1

[5] Vladimir Kulikov, Matan Kleiner, Inbar Huberman-Spiegelglas, and Tomer Michaeli. Flowedit: Inversion-free text-based editing using pre-trained flow models. In *ICCV*, 2025. 3, 4

[6] Black Forest Labs. Flux. https://github.com/black-forest-labs/flux, 2024. 1

[7] Yu-Lun Liu, Wei-Sheng Lai, Yu-Sheng Chen, Yi-Lung Kao, Ming-Hsuan Yang, Yung-Yu Chuang, and Jia-Bin Huang. Single-image HDR reconstruction by learning to reverse the camera pipeline. In *CVPR*, 2020. 3, 4

[8] Andreas Lugmayr, Martin Danelljan, Andres Romero, Fisher Yu, Radu Timofte, and Luc Van Gool. Repaint: Inpainting using denoising diffusion probabilistic models. In *CVPR*, 2022. 4

[9] Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. Sdedit: Guided image synthesis and editing with stochastic differential equations. In *ICLR*, 2022. 4

[10] William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, 2023. 2

[11] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *AAAI*, 2018. 2

[12] Chao Wang, Zhihao Xia, Thomas Leimkuhler, Karol Myszkowski, and Xuaner Zhang. LEDiff: Latent exposure diffusion for HDR generation. In *CVPR*, 2025. 3, 4

[13] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *ICCV*, 2023. 5

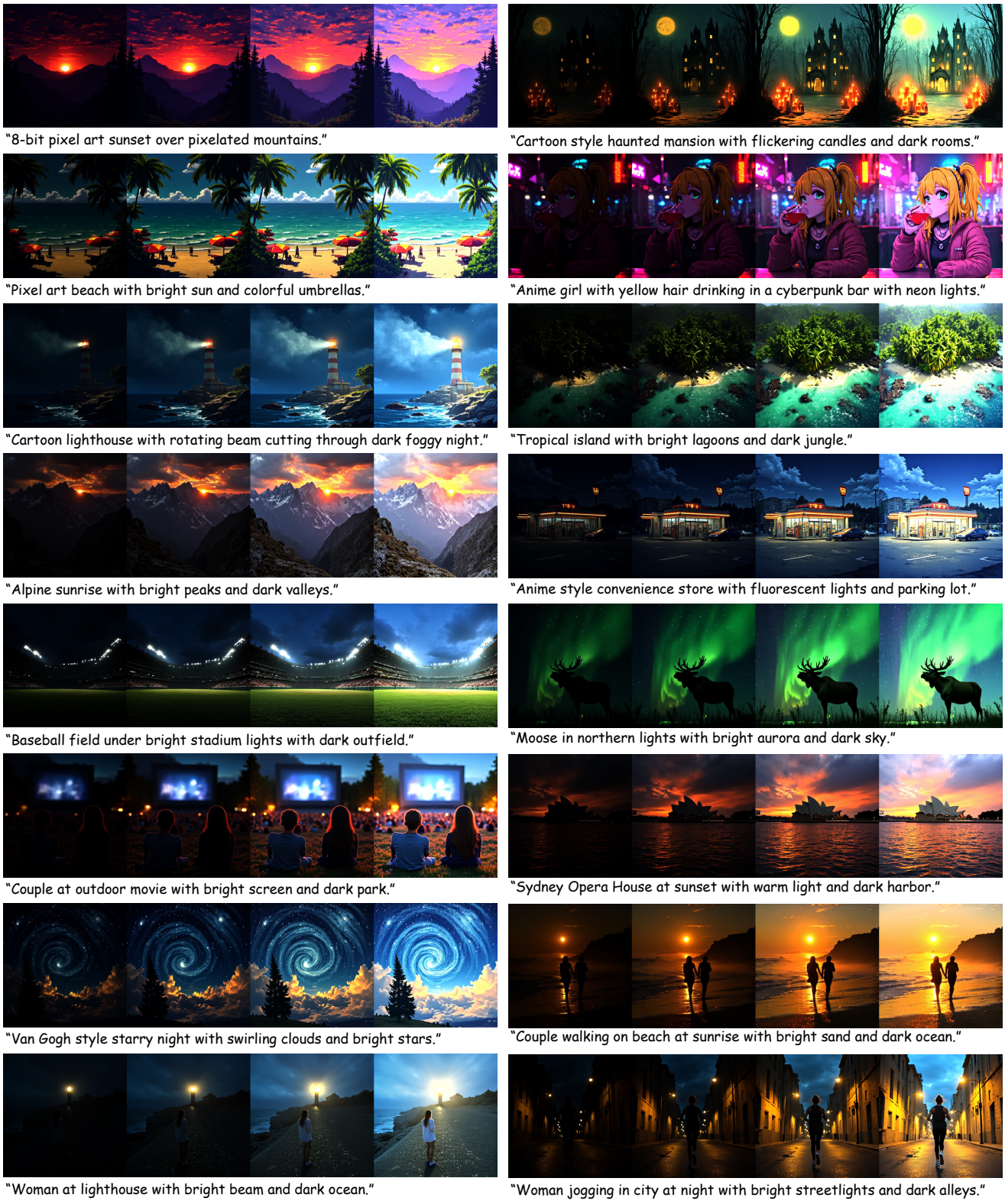


Figure 9. Diverse style generation results. Our method can generate exposure brackets with various artistic styles including cartoon, photorealistic, painterly, and cinematic looks while preserving proper exposure relationships across brackets.