Bringing AI-Generated Images Into the Light With HDR

Originally published at https://www.unite.ai/bringing-ai-generated-images-into-the-light-with-hdr/

AI images and video may be impressive, but they're not 'professional'-standard – an issue that a new research project seeks to address.

In the professional audio-visual community, one of the most frequent objections to the encroachment of AI that crops up is the current lack of professional standards of image and video reproduction. Not the least of these is the ability to work with High Dynamic Range (HDR) images and video.

HDR images are the modern equivalent of a 19th/20th-century photographic practice called bracketing, where the same picture is taken multiple times with increasing amounts of light being allowed to reach the film emulsion:

Above, a short bracketed sequence. Inset below, the high dynamic range that can be extrapolated from these photos into a single image. Source

In traditional photography, this resulted in multiple pictures which could, with some expertise and effort, be composed into a single print that benefited from all the different levels of detail available across the range of the exposures. But it was not a trivial or easy process.

These days an 'auto-bracketed' image sequence can either produce multiple images or be combined into a single HDR image – effectively a multiplicity of exposures in one image, which HDR-capable image-editing applications such as Photoshop can iterate through, and allow the photographer to orchestrate into one, ideal output image.

If you're wondering why you should care, or how this kind of thing impacts your own photography, the illustration for this article is intended to demonstrate this in a familiar way:

Above, on the left we see a typical example of an sRGB (i.e., non-HDR) image. Just brightening (shown on the right) it does *not* show the monster in the closet, because that detail got thrown away when the photographer, and the automated processes of the camera, decided what to prioritize in the photo:

Below is an indication (left) of how 'washed out' the foreground would have to be *at the time of exposure* to register the closet-monster in a non-HDR photo, and (right) how the monster is plunged into darkness when the exposure is made suitable for the well-lit foreground subjects instead:

Below, we see the kind of detail that can be 'rescued' from an HDR image or image-sequence. In this case, the monster was 'hiding' in the very lowest visual registers of the HDR sequence, in a level where the rest of the content would have been 'blown out' into near-white (above, left). By specifying that a wide range of brightness levels should be expressed, selectively, in the same image, these dissonant elements can be composed into one rational picture:

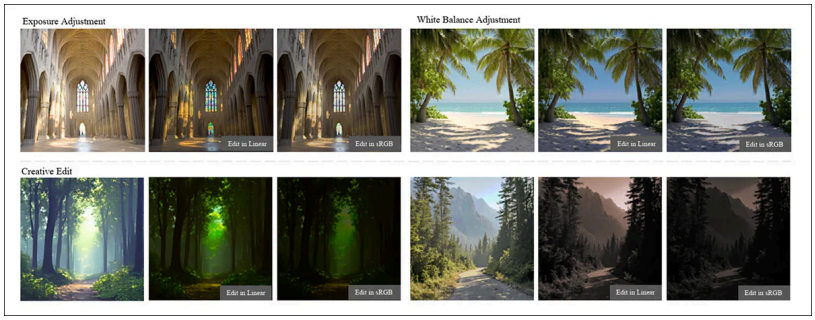
A non-HDR image is known as a *display-referred image*, and a high-gamut HDR image is known as a *scene-referred image*.

HDR video is also a thing, and this kind of tonal versatility and ductility really gives film-makers some latitude to rescue, grade, and interpret footage in a number of creative and consistent ways; unsurprising, then, that creatives are reluctant to work with the 'flattened out' sRGB output typical of most generative AI frameworks.

HDR in AI

Naturally, the research scene is interested in bringing AI-generated frameworks into the HDR era. However, it's not a trivial task, both because of the fundamental architecture of diffusion-based generative systems, and because good HDR data takes up a great deal of disk space, making for unwieldy collections; consequently, datasets fit for the task are scarce.

Nonetheless, a collaboration between a university at Singapore and Adobe ADBE 5.73% Research is offering a method of producing HDR image sequences, in a methodology that can theoretically be applied to video as well as still images:



From the project site for the new work, examples of 'bracketed' text-to-image output. Source

The new system generates several aligned versions of the same image at different brightness levels and learns how bright the scene really was, then combines these into a single result that keeps detail in both shadows and highlights, allowing later edits to exposure or color to behave more like adjustments to a real camera capture, rather than fragile tweaks to a fully-processed image.

The system leverages a diversity of different models for the task, including variants of Qwen and Flux:

Examples from the new paper, showing how the system can generate multiple exposure versions of the same scene while keeping the underlying structure fixed. the model produces consistent images across very dark to very bright settings, whether the prompt describes moonlight, sunlight, sunset, or even a small object li composition remaining stable as only the lighting changes. The method can vary brightness in a controlled, camera-like way, rather than drifting or inventing new c Source

The authors state:

> *'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.'*

The new work is titled *Linear Image Generation by Synthesizing Exposure Brackets*, and comes from four authors across S-Lab at Nanyang Technological University, Adobe NextCam, and Adobe Research. Besides the aforementioned project page and YouTube video accompanying the release, there is also a (currently sparse) GitHub repo, and the promise of a dataset release.

Though the authors supply many examples of output from the system at the associated project page, viewers will need an HDR-capable monitor to really distinguish the characteristics of the HDR output presented. Nonetheless, please find the researchers' YouTube overview embedded at the end of this article – but be aware that the differences between shown examples may not be clear on a non-HDR monitor.

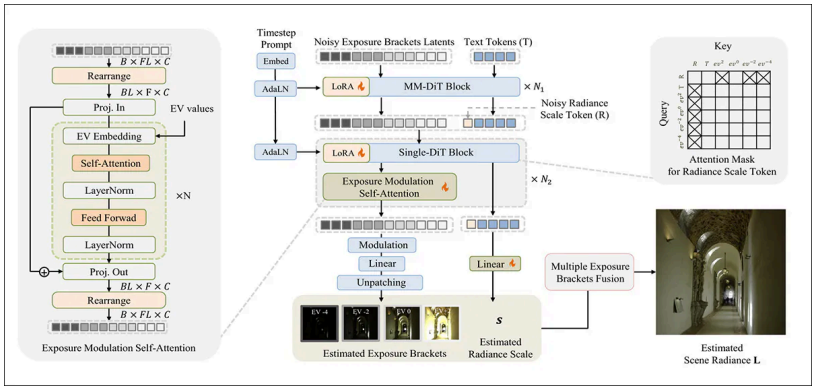
Method and Data

The authors emphasize the extent to which data-gathering is a challenge in this particular pursuit:

> *'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.'*

The researchers made creative use of the few options to hand, leveraging the RAISE dataset as actual training data, and the MIT-Adobe FiveK dataset as evaluation data*.

To build usable HDR training data, the researchers ran the RAW camera files through a standardized pipeline to strip away camera-specific quirks, converting the images into a consistent, scene-referred linear format:



Schema for the authors' workflow: the system starts with noise representing four exposure levels of the same scene, along with a text prompt and a brightness tok through stacked transformer blocks that keep the different exposures aligned, while adjusting for lighting. The system then predicts both the set of exposure image brightness scale, and subsequently decodes and combines them into a single scene-referred image, retaining detail in both shadows and highlights.

This entailed reconstructing full RGB from sensor data, applying color correction, normalizing white balance, and briefly moving into a perceptual color space for denoising before returning to a clean linear signal. The actual light in the scene was then recovered using the camera's exposure settings, so that each pixel would reflect real brightness rather than a display-ready approximation.

Because such values can vary widely, the data was then stabilized by scaling each image based on its own brightness distribution, using mid-range and highlight statistics to avoid both washed-out images and blown highlights, finally obtaining a normalized linear image that preserved the true range of light in the scene, while remaining stable enough for training.

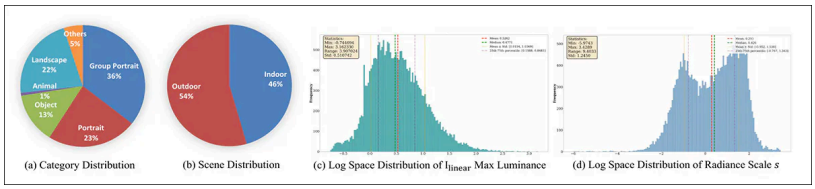
Text labels for the images were then created with the Qwen2.5-VL 7B model, with prompts crafted to match the characteristics of the Flux model that would be used at generation-time.

Each image was split into exposure 'slices' and passed through a shared VAE encoder, converting all exposures into a common latent space designed to capture the full brightness range. The latents were then refined from noise, and decoded back into images, allowing consistent reconstruction across dark and bright regions, without collapsing them into a single, 'flattened' exposure.

LoRA finetuning was used to adapt the pretrained Flux backbone to linear-image data with minimal extra parameters, helping the Single-Diffusion Transformers (single-DiT) model remain stable, even as brightness varied across exposure brackets.

Exposure Modulation Self-Attention (central column in schema illustration above) was introduced to jointly process all brackets, allowing luminance to be adjusted per exposure while keeping structure and fine detail aligned.

3D Rotary Positional Embedding (3D-R[o]PE) was used to encode both spatial position and exposure identity, so that the model could distinguish which bracket each token belonged to, while preserving spatial consistency, enabling clean separation of brightness variation from scene content.



An overview of the dataset used in the study, showing how images are distributed across content types and indoor versus outdoor scenes, alongside the spread o processed data. The histograms plot luminance and the radiance scale in log space, illustrating how widely real-world brightness can vary, with higher radiance va brighter scenes and highlighting the strong dynamic range the model is trained to handle.

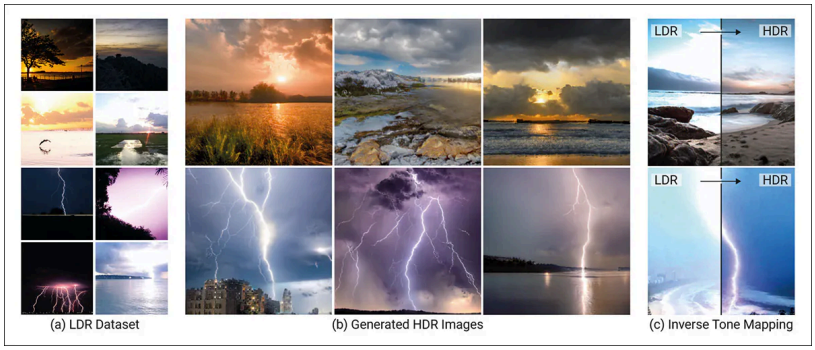
3D-RoPE split out *where* a feature was and 'which exposure it comes from' into separate signals, so that brightness variation could be adjusted independently, without corrupting spatial detail.

Tests

The researchers used Flux-dev as the generative framework, with training occurring on four NVIDIA A100 GPUs, each with 80GB of VRAM. The batch size was set at 4 (per GPU), over 10,000 iterations.

LoRA fine-tuning used a rank of 64. The AdamW optimizer was used at a learning rate of 2×10^{2} (for the exposure modulation aspect).

The authors note that while there are two prior works which are similar in scope, neither was an obvious contender for a testing phase. The Max Planck-led 2022 outing GlowGAN is limited to generating specific image categories, while 2025's Bracket Diffusion (again, led by the Max Planck Institute) can only generate an HDR image at 256x256px, and takes several minutes to do so.



From the original GlowGAN paper, typical low dynamic range (LDR) images lose detail in shadows and highlights, while the model learns to produce high dynamic detail across brightness levels and enable recovery of saturated regions through inverse tone mapping. Source

Therefore, in the absence of direct baselines for linear-image generation, the authors compared their method with adapted versions of strong existing models, rather than purpose-built alternatives.

One set of experiments (*'T2I Fine-Tuning'*) fine-tuned the text-to-image diffusion model Flux using LoRA, training it to generate linear images directly, and evaluating how a state-of-the-art T2I model adapted to this domain.

A second comparison (*'T2V fine-tuning'*) used the text-to-video model Wan 2.1, whose VAE compresses multiple frames into a shared latent; in this setup, four exposure brackets were encoded into a single latent representation, and then decoded back, testing whether a video-style pipeline could model exposure variation.

The third set of experiments (*'T2I Model Inflation'*) compared against CameraCtrl and Generative Photography, which both extend image diffusion models via temporal modules, to produce multi-frame outputs. These were also finetuned on the same data, for a consistent comparison.

Metrics used were Fréchet Inception Distance (FID); Aesthetic Score (AS); Naturalness Image Quality Evaluator (NIQUE); CLIP Sim score; and Luminance Similarity (LS):

Type	Model	FID ↓	AS ↑	NIQE ↓	CLIP Sim.↑	LS ↑
T2I Finetuning	Flux [20]	32.12	4.712	5.304	25.90	/
T2V Finetuning	Wan 2.1 [34]	/	4.537	5.412	24.79	1.12
T2I Model Inflation	CameraCtrl [14] (w/ F)	37.25	5.230	4.131	**26.89**	8.97
	Gen. Photography [42] (w/ F)	40.17	4.619	4.514	23.71	7.11
	Gen. Photography [42] (w/o F)	43.83	3.909	4.870	20.51	5.56
	Ours	**28.29**	**5.700**	**3.658**	26.02	**23.06**

A comparison of the authors' method against several adapted baselines for generating linear, scene-referred images. Text-to-image (Flux) and text-to-video (Wan 2.1) models are fine-tuned with LoRA to test how well existing generative systems handle this setting, while CameraCtrl and Generative Photography extend diffusion models with temporal components. Some scores are missing, because certain models cannot reliably produce consistent exposure brackets, which are required to recover full dynamic range. Across the reported metrics, the new method achieves the strongest overall results, particularly on measures tied to image quality and accurate brightness reconstruction.

Regarding these results, the authors state:

> *'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× 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.'*

A comparison with LoRA-adapted Flux and Wan 2.1, illustrating how each method handles exposure changes across the same scenes. Competing approaches tend to lose detail in very dark or very bright regions, while the proposed method maintains consistent structure and recovers usable detail across the full range of exposures. Please refer to the source paper and appellant project site for better-quality results examples.

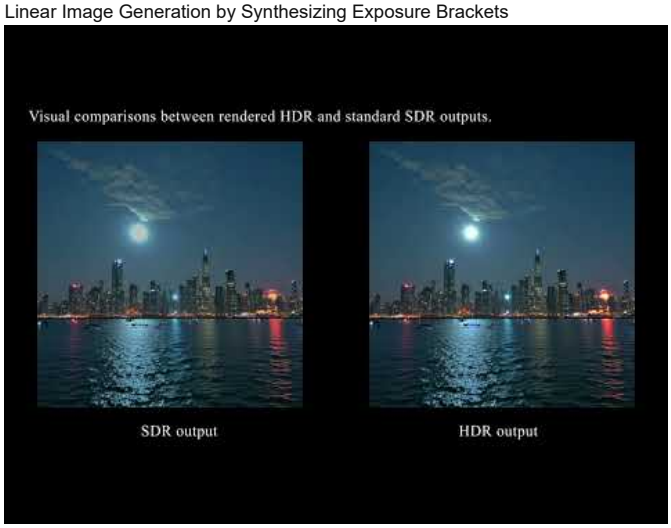
Please refer to the source paper's extended experiments and supplementary materials section for further tests.

Conclusion

For media professionals, such as those working in film and TV production, the same output that has captured the imagination (and, increasingly, the ire) of the world has left them nonplussed, since all nearly of their pipelines rely in some way on HDR captures.

Therefore this is a timely project, representing a facility that one would hope would become an optional standard across new frameworks – albeit that it is certain to at least double rendering times; clearly, also, latency will need to be seriously addressed if HDR AI content is not to be relegated to the 'in post' rather than in-camera category.

Linear Image Generation by Synthesizing Exposure Brackets



* *Normally we would show examples, but since the reader may not have an HDR-capable monitor, we omit these in this case.*

First published Sunday, April 26, 2026

Writer on machine learning, domain specialist in human image synthesis. Former head of research content at Metaphysic.ai, until its dissolution into DNEG's Brahma.ai.
Portfolio site: martinanderson.ai
Contact: martin@martinanderson.ai

Discover More