Event-Driven Lighting for Attention Guidance with Video Diffusion Model
Supplemental Material

A. Implementation Details

As detailed in the main manuscript, we adopt a two-stage training protocol. Both the 128 × 128 and 512 × 512 models are trained on our *Sketchfab-Synth-Train* dataset, requiring 6000 and 2000 epochs, respectively. We further evaluate model performance on the *Sketchfab-Synth-Test* dataset at 100-epoch intervals.

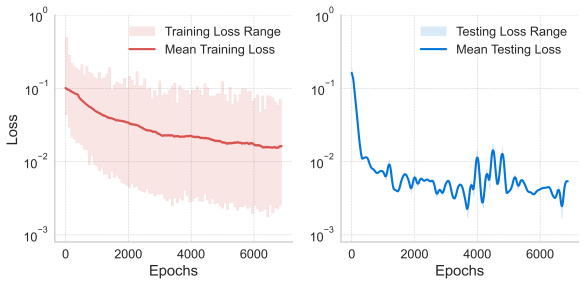

Figure 1. Our 128x128 pretrained model loss.

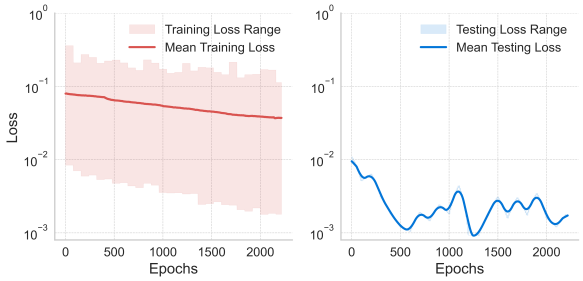

Figure 2. Our 512x512 fine-tune model loss.

Room	Sample Indices
Room 1	1, 4, 5, 15, 28, 38[3], 39, 42, 46, 50, 53, 54[4]
Room 2	1, 3[5], 5, 20, 29, 38[6], 43, 45, 46, 47, 51, 57
Room 3	1, 3, 4, 7[7], 21[8], 26, 40, 46, 48, 49, 50, 60

Table 1. Sample indices chosen for each room in our *Sketchfab-Synth-Test*. Highlight inference was displayed at Sec. E.

We randomly select a representative subset of frames from each room's sixty rendering samples to form our testing dataset (see Table 1).

B. Evaluation

Baselines Since no existing method supports mask-guided relighting for indoor video, we adapt the single-image diffusion model ScribbleLight, and IC-Light [2, 4] as our baseline without retraining on our *Sketchfab-Synth-Test* dataset. For the ScribbleLight, we convert each ground-truth mask into a scribble map (relit region = 1, background = 0.5) per the original specification, and supply the normals, diffuse maps, and these scribbles to ScribbleLight to produce relit frames on a per-frame basis. For IC-Light, we extract the foreground from the input image by BRIA v1.4 [1], then apply the same scribble map as the background relighting condition. Both methods are inferred at 512 × 512 resolution with the aid of a text prompt generated by BLIP-2 [3]. Qualitative results are shown in Section E.

Metrics We assess relighting quality using four standard measures: root-mean-square error (RMSE), peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and Learned Perceptual Image Patch Similarity (LPIPS) [5]. RMSE and LPIPS (lower is better) quantify pixel-level and perceptual differences, respectively, while PSNR and SSIM (higher is better) capture overall fidelity and structural consistency. For each metric, we report both the average and the best scores over all test samples.

	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
ScrribleLight	0.3000(0.172)	10.612(15.281)	0.323(0.487)	0.584(0.483)
IC-Light v1	0.367(0.259)	8.781(11.723)	0.202(0.370)	0.769(0.615)
Ours (128x128)	0.061(0.031)	24.528(30.118)	0.823(0.916)	0.120(0.046)
Ours (512x512)	**0.042(0.019)**	**28.423(34.322)**	**0.904(0.948)**	**0.070(0.034)**

Table 2. Quantitative comparison of relighting accuracy between our EDL-Diffusion, ScribbleLight [2], and IC-Light v1 [4]. We compute the mean(best) errors with respect to a target relit image.

We evaluate our method against baseline methods on our *Sketchfab-Synth-Test* (see Table 2). Even at low resolution, our approach outperforms the baseline across all metrics (RMSE, PSNR, SSIM, LPIPS). At full resolution, these gains become even more pronounced, demonstrating that our diffusion-based pipeline consistently surpasses ScribbleLight and IC-Light v1.

	RMSE ↓	PSNR ↑	SSIM ↑	LPIPS ↓
ScrribleLight	0.032(0.000)	Inf(Inf)	0.990(1.000)	0.016(0.000)
IC-Light v1	0.043(0.000)	Inf(Inf)	0.990(1.000)	0.017(0.000)
Ours (128x128)	0.015(0.000)	Inf(Inf)	0.995(1.000)	0.016(0.000)
Ours (512x512)	**0.012(0.000)**	**Inf(Inf)**	**0.997(1.000)**	**0.004(0.000)**

Table 3. Quantitative comparison of relighting accuracy between our EDL-Diffusion, ScribbleLight [2], and IC-Light v1 [4] based on cropped masking area.

We also evaluate quantitative analysis on the cropped masking area only to see the performance on the highlighted attention area. As shown in Table 3, our 512×512 EDL-Diffusion-Relight yields better performance on masked regions, outperforming ScribbleLight and IC-Light v1. Even the 128×128 variant surpasses both,

proving superior local relighting accuracy.

C. Webpage Materials

In our supplementary webpage, the topmost section features a demonstration video illustrating how EDL-Diffusion-Relight transforms an input sequence. The Demo panel provides additional side-by-side examples: users can select various training and test cases to inspect our video-diffusion relighting results and compare them directly against ScribbleLight under identical inputs. For full details and interactive exploration, please visit our website: https://quantum-whisper.github.io.

D. Video

We provide a sample of relight videos with our supplemental materials.

E. Qualitative Visualization

Our qualitative evaluation examines sequential time frames I_i, which are characterized by varied relit objects and camera motion within a consistent room setting. Unlike prior methods such as ScribbleLight, which employ diffuse, normal, and scribble inputs, or IC-Light, which relights foregrounds based on background-conditioned pseudo-lighting, our approach leverages the original RGB frame alongside G-buffers. This video-to-video input strategy is critical for ensuring that our generated output frames accurately preserve the inherent illumination levels and ambient room conditions of the input video.

In the qualitative examples from our test set, ScribbleLight maintains overall room ambience by aligning diffuse layers but fails to generalize its relighting to our diverse indoor scenes. IC-Light improves illumination within EDL-defined regions yet its VAE redraw introduces temporal inconsistency and alters the original video style. In contrast, our event-driven-lighting video diffusion model preserves the input video's appearance, enforces frame-to-frame coherence, and confines relighting strictly to the EDL-masked areas.

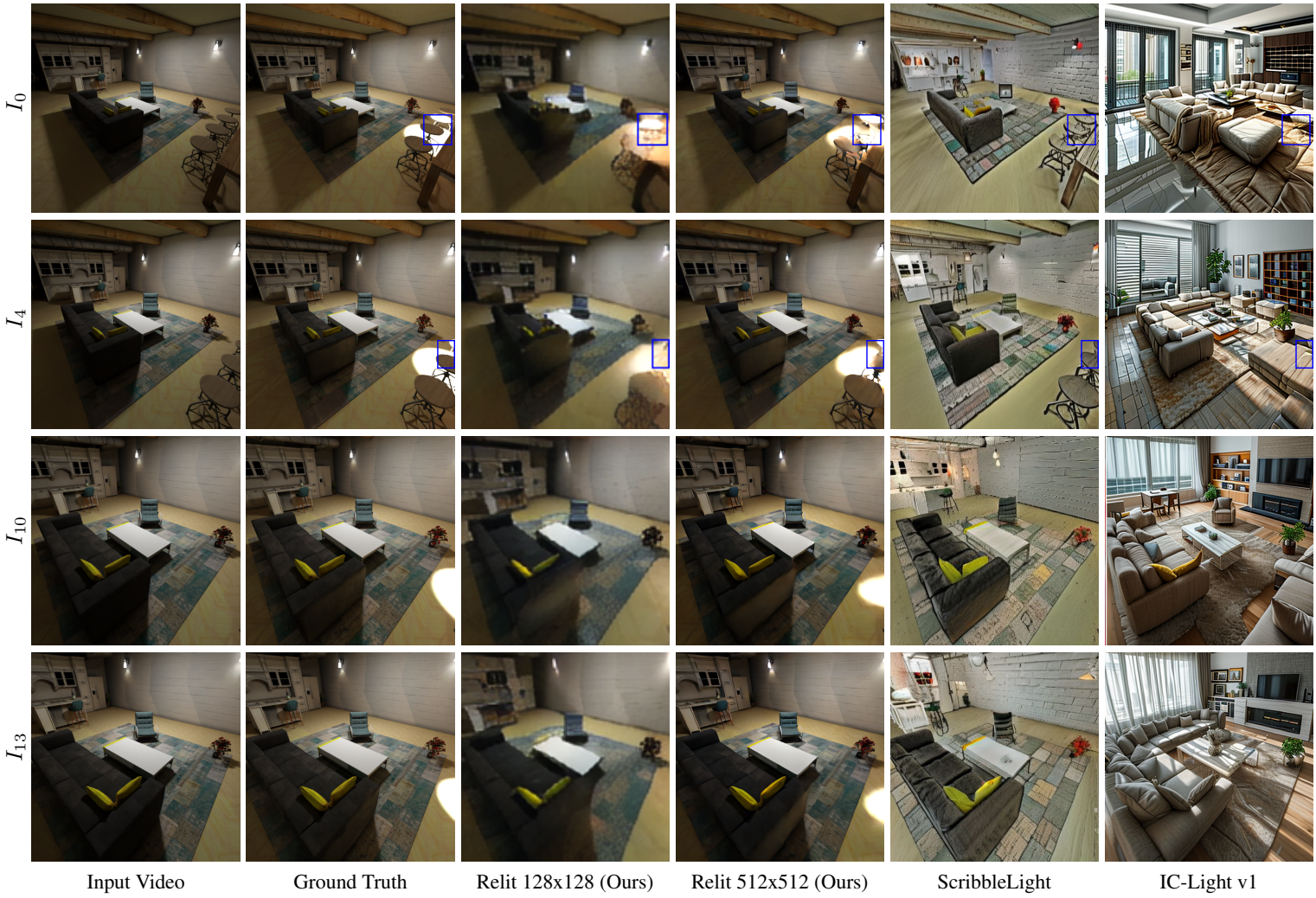

Figure 3. Sample 38 of room 1: relight object *bar chair 3* in scene *room 1*. (View model in Sketchfab: Big Room 3D Model). The relight object area was highlight by a blue bounding box for better visualization. Our method leverages the original RGB frame and accompanying G-buffers to faithfully preserve the input video's intrinsic illumination and ambient room conditions. In contrast to prior work (e.g., ScribbleLight and IC-Light), it maintains global scene lighting while selectively enhancing the region within the blue bounding box to guide viewer attention.

Figure 4. Sample 54 of room 1: relight object *plant 1* in scene *room 1*. (View model in Sketchfab: Big Room 3D Model). The relighting target region—specified via masking or scribble—is delineated with a blue bounding box to enhance visual clarity. Our method leverages the original RGB frame and accompanying G-buffers to faithfully preserve the input video's intrinsic illumination and ambient room conditions. In contrast to prior work (e.g., ScribbleLight and IC-Light), it maintains global scene lighting while selectively enhancing the region within the blue bounding box to guide viewer attention.

Figure 5. Sample 3 of room 2: relight object *plant 1* in scene *room 2*. (View model in Sketchfab: Cozy bathroom design). The relighting target region—specified via masking or scribble—is delineated with a blue bounding box to enhance visual clarity. Our method leverages the original RGB frame and accompanying G-buffers to faithfully preserve the input video's intrinsic illumination and ambient room conditions. In contrast to prior work (e.g., ScribbleLight and IC-Light), it maintains global scene lighting while selectively enhancing the region within the blue bounding box to guide viewer attention.

Figure 6. Sample 38 of room 2: relight object *right faucet handle 1* in scene *room 2*. (View model in Sketchfab: Cozy bathroom design). The relighting target region—specified via masking or scribble—is delineated with a blue bounding box to enhance visual clarity. Our method leverages the original RGB frame and accompanying G-buffers to faithfully preserve the input video's intrinsic illumination and ambient room conditions. In contrast to prior work (e.g., ScribbleLight and IC-Light), it maintains global scene lighting while selectively enhancing the region within the blue bounding box to guide viewer attention.

Figure 7. Sample 7 of room 3: relight object *book* in scene *room 3*. (View model in Sketchfab: Kidman Room). The relighting target region—specified via masking or scribble—is delineated with a blue bounding box to enhance visual clarity. Our method leverages the original RGB frame and accompanying G-buffers to faithfully preserve the input video's intrinsic illumination and ambient room conditions. In contrast to prior work (e.g., ScribbleLight and IC-Light), it maintains global scene lighting while selectively enhancing the region within the blue bounding box to guide viewer attention.

Figure 8. Sample 20 of room 3: relight object *couch* in scene *room 3*. (View model in Sketchfab: Kidman Room). The relighting target region—specified via masking or scribble—is delineated with a blue bounding box to enhance visual clarity. Our method leverages the original RGB frame and accompanying G-buffers to faithfully preserve the input video's intrinsic illumination and ambient room conditions. In contrast to prior work (e.g., ScribbleLight and IC-Light), it maintains global scene lighting while selectively enhancing the region within the blue bounding box to guide viewer attention.

References

[1] BRIA. Bria background removal v1.4 model card, 2025. [Online; accessed 2025-05-23].

[2] Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, and Roni Sengupta. Scribblelight: Single image indoor relighting with scribbles, 2024.

[3] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[4] 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 *The Thirteenth International Conference on Learning Representations*, 2025.

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