Event-Driven Lighting for Immersive Attention Guidance with Video Diffusion Model

Anonymous Author(s)
Affiliation
Address
email

Abstract

In immersive VR environments, users may miss crucial story events that occur outside their field of view. To mitigate this problem, we propose a method and framework for enhancing users' attention in virtual environments by autonomously identifying significant events and adjusting environment lighting via latent video diffusion models (Video-LDMs) to achieve controllable video relighting accordingly. Our method accepts video frames and an English text script as input, detects events in the script using large language models (LLMs), then extracts the event arguments with corresponding G-buffer frames. Given the localized event and its mask, our system makes it more obvious to the user by automatically adjusting lighting by leveraging pre-trained diffusion model. Through a user study, we validate that our approach improves users' sense of presence within and memory of virtual environments. Our proposed framework is versatile, and can be applied to scenarios like creating cinematic videos and guiding users in interactive games.

1 Introduction

In immersive virtual environments (VEs), the user is typically free to move the camera in any direction they please. This can lead to a common problem in VR experiences, wherein the user may turn to face away from an object or event of interest that is crucial to their virtual experience. That is, the important events of the virtual experience occur outside of the user's field of view and they may become confused due to missing critical information. To overcome this problem, researchers have developed different attention-guidance techniques [25]. Though their implementations differ, all attention-guidance techniques introduce some stimulus (usually visual) that attracts the user's gaze towards the virtual area of interest. When choosing an attention-guiding stimulus, it is important to manage the tradeoff between the stimulus' power (ability to reliably influence the user's attention) and how immersive it is.

Recent advancements in video generative models, such as Sora [2] and Stable Video Diffusion (SVD) [1], have captured significant attention in the field. A key aspect of their performance is the implicit encoding of illumination-related appearance priors, particularly in Latent Diffusion Models (LDMs) [24], which enable the generation of complex scenes under various lighting effects. In this study, we hypothesize that LDMs inherently learn and reproduce intricate lighting effects without the need for traditional computer graphics-based lighting algorithms. Our objective is to bring the lighting-related priors embedded within these models through a control mechanism so that we can extend this technique to event-driven lighting (EDL), wherein the model dynamically enhances specific regions of interest based on the alignment of target objects or movements within the field of view and the narrative context.

The proposed pipeline in our study is designed to generate photorealistic relighting effects on VR input video sources. For this study, we simplify the problem by focusing on a single light source

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

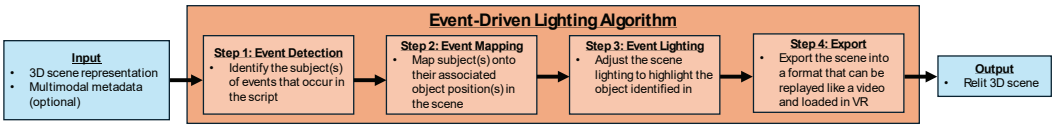


Figure 2: There are four steps in our *Event-Driven Lighting* pipeline: (1) identifying events; (2) locating corresponding objects; (3) relighting the scene with a suitable lighting configuration; (4) exporting the rendering into a vidoe or multimodal VR environment.

mask as input combined with various consistent geometry layers that are lighting-insensitive across frames.

In this work, we introduce an *event-driven lighting* framework for attention-guidance in immersive virtual environments. Our system automatically detects events of interest, locates their positions in the VE relative to the user, and introduces diegetic lighting stimuli that draw the user's gaze towards the events of interest. We demonstrate that our event-driven lighting framework is agnostic to the underlying 3D representation, by realizing implementations of an immersive narrated tour application in both Video-LDMs-based and traditional game-engine-based VEs. Results of a user study evaluation show that our method maintains user immersion while improving their spatial memory when quizzed about the location of scene contents afterwards. Our main contributions are:

- A general framework for guiding user attention in VR via event-driven lighting.
- Automatic detection of narrative events and their respective arguments using large language models, then extract *event mask* within 2D image space.
- A video latent diffusion relighting model using rgb, depth, normal, albedo, and *event mask* frames.
- Formal user study to evaluate the proposed framework's effectiveness.

To our knowledge, our work is the first to integrate language-based event detection, semantic segmentation, and diffusion-based video relighting in a unified pipeline, enabling controllable, temporally coherent lighting changes in video, grounded in viewer-relevant semantic events. This bridges the gap between generative video diffusion and scene-aware, event-responsive editing.

2 Background

2.1 Stable Video Diffusion

In this section, we first introduce the latent diffusion model, SVD [1], which serves as the foundation of our approach, in Sec.2.1. Subsequently, we describe our EDL-relighting SVD pipeline in Sec.3.3, with an overview provided in Fig. 3.

SVD is a generative model designed to produce temporally coherent video frames based on an initial prompt or video conditioning. It extends the LDM to the video domain by introducing a temporal dimension and optimizing over both spatial and temporal features to maintain consistency across frames. The LDM comprises two main components: an autoencoder and a diffusion model. The autoencoder is responsible for compressing and reconstructing images, utilizing an encoder $\mathcal{E}(\cdot)$ and a decoder $\mathcal{D}(\cdot)$. Specifically, the encoder projects an image x into a lower-dimensional latent space z, followed by the decoder reconstructing the original image from this latent representation. The reconstruction process yields an image $\tilde{x} = \mathcal{D}(z)$, approximating the original image x.

In the forward process, a video sample $\mathcal{X} = \{I_0, I_1, \ldots, I_{N-1}\}$ composed of N frames, is first encoded in the latent space $\mathbf{z} = \mathcal{E}(\mathbf{x})$, then the intermediate noisy video at time step t is created as $\mathbf{z}_t = \alpha_t \mathbf{z} + \sigma_t \epsilon$, where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ is Gaussian noise, and α_t and σ_t define a fixed noise schedule. Given that the data distribution $\mathbf{z}_0 \sim q(\mathbf{z}_0)$ is progressively corrupted by Gaussian noise over T steps, this process follows a variance schedule denoted by $\beta_1, \ldots, \beta_T$:

$$q(\mathbf{z}_t \mid \mathbf{z}_{t-1}) = \mathcal{N}\left(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t \mathbf{I}\right), \quad t = 1, \ldots, T \tag{1}$$

The denoising U-Net, $\epsilon_\theta\left(\mathbf{z}_t; t\right)$, receives this noisy video latent $\mathbf{z}_t$ and the conditioning $\mathbf{c}$ computed from the input image, i.e., the first frame I_0 in the video, and is trained to predict this added noise using a loss function:

$$\mathcal{L}(\theta) = \mathbb{E}_{t\sim U(1,T),\epsilon_t\sim\mathcal{N}(0,1)}\left[\left\|\epsilon_t - \epsilon_\theta\left(\mathbf{z}_t; t, \mathbf{c}\right)\right\|^2\right], \tag{2}$$

where x_t is the noisy sample of x_0 at timestep t, the target vector ϵ_t here is $\epsilon_t = \alpha_t\epsilon - \sigma_t\mathbf{z}_t$, referred to as v-prediction.

Once the denoising network is trained, starting from pure noise $\mathbf{z}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, the sampling process iteratively denoises the noisy latent by predicting the noise in the input and then applying an update step to remove a portion of the estimated noise from the noisy latent

$$\mathbf{z}_{t-1} = \text{update}(\mathbf{z}_t, \epsilon_\theta(\mathbf{z}_t; t, \mathbf{c}), t) \tag{3}$$

until we get clean latent $\mathbf{z}_0$, followed by decoding $\mathcal{D}(z_0)$ to get the generated video. The exact implementation of the update $(\cdot, \cdot)$ function depends on the specifics of the sampling method; we incorporate the original SVD's Euler Discrete Scheduler [13].

3 Method

In this section, we present a framework for event-driven video relighting through two stages: (1) language-guided event detection using Segment-Anything for semantic mask generation; (2) conditioning these masks within our SVD Latent Diffusion Model with fine-tuning a U-Net backbone. Unlike approaches requiring 3D bounding boxes and geometry reconstruction, our method operates on 2D video with paired lighting sequences. The result is a system enabling photorealistic, element-specific lighting control with superior efficiency.

3.1 Event Detection with large Language Instructed Segmentation Assistant

Our method leverages the large Language Instructed Segmentation Assistant (LISA) [16] to detect and extract events in video sequences. LISA extends the capabilities of the Segment-Anything Model (SAM) [15] by incorporating language understanding to identify semantically meaningful regions within frames.

Given input video $V = \{I_0, I_1, ..., I_{T-1}\}$ consisting of T frames and a language prompt L describing the event of interest, we obtain a sequence of binary masks $M = \{m_0, m_0, ..., m_{T-1}\}$ where:

$$m_i = \text{LISA}(I_i, L) \tag{4}$$

By leveraging the precise segmentation masks from LISA, our approach can selectively relight specific events or objects within the video while maintaining the appearance of the surrounding environment. This enables fine-grained control over the relighting process without requiring explicit 3D scene understanding or reconstruction.

3.2 Lighting Design

Using the three-point lighting principle, we opt for an extremely simple lighting design recipe that can apply to a wide range of scenarios, simply placing a spotlight as key directly above the subject bounding box. In a more general setting, the positioning of lights would also depend upon the camera viewing angle toward the subject, as this determines whether a light source is perceived as a backlight or keylight to the viewer.

We acknowledge that lighting design can be subjective and that more sophisticated recipes may be similarly effective in guiding user attention or fulfilling other objectives, such as appearing natural or motivated by the context of the story. An exhaustive comparison of such alternatives is a promising direction for future work.

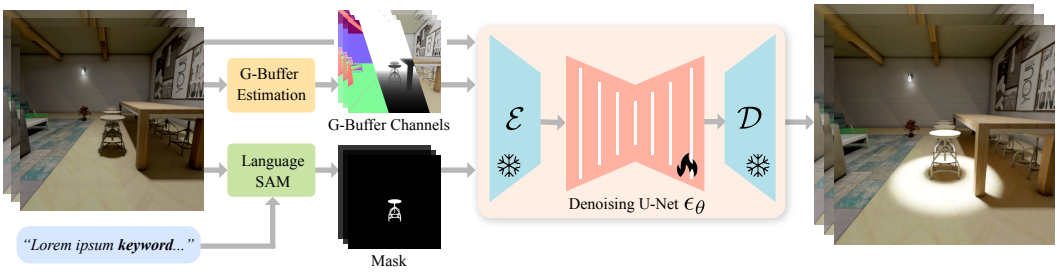


Figure 3: Our relighting system architecture. In the relighting module, we fine-tune the Diffusion 3D U-Net by feeding in encoded latents of the original video concatenated with encoded latents of its corresponding G-buffer channels.

3.3 Relighting Network Architecture

Our problem can be formalized as follows. Given a video sequence $\mathbf{I}$, the objective is to generate a relighted video $\mathbf{I}^{'}$ controlled by a sequence of event masks $\mathbf{m}$. To achieve this, we adapt the image-to-video SVD [1] framework introduced by DiffusionRenderer [18]. This perspective allows us to leverage the powerful image synthesis capabilities of diffusion models for the specific task of lighting manipulation. Following this paradigm, we fine-tune only the denoising U-Net ϵ_θ of the Stable Video Diffusion model while keeping its VAE encoder-decoder pair $\{\mathcal{E}(\cdot), \mathcal{D}(\cdot)\}$ frozen.

The model accepts multiple input channels which serve as helpful priors: original video frames $\mathbf{I}$, depth $\mathbf{d}$, normal $\mathbf{n}$, and albedo maps $\mathbf{a}$, and a *Relighting Object Mask* $\mathbf{m}$. Using these informative maps as input to the model enables precise control over the relighting process while preserving scene structure. To accommodate these 20 additional G-Buffer channels, we expand the first convolutional layer of the diffusion U-Net architecture. We use the VAE encoder $\mathcal{E}$ to separately encode each G-buffer from $\{\mathbf{I}, \mathbf{d}, \mathbf{n}, \mathbf{a}, \mathbf{m}\}$ into the latent space and concatenate them to produce the pixel-aligned scene attribute latent map $\mathbf{g} = \{\mathcal{E}(\mathbf{I}), \mathcal{E}(\mathbf{d}), \mathcal{E}(\mathbf{n}), \mathcal{E}(\mathbf{a}), \mathcal{E}(\mathbf{m})\} \in \mathbb{R}^{F\times h\times w\times 20}$.

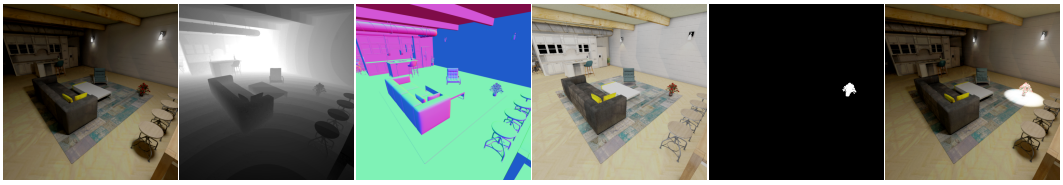
As a result, the diffusion U-Net ϵ_θ takes the noisy latent $\mathbf{z}_t$ and G-buffer latent $\mathbf{g}$ as pixel-wise input. At each U-Net level k, the 3D U-Net architecture operates as a cross-attention mechanism that *queries* the CLIP-embedded input video frames. This multi-level cross-attention design enables the model to incorporate visual semantic information at various scales and resolutions, with each U-Net layer generating its own set of *keys* and *values* from the CLIP embeddings for effective feature aggregation and integration into the generative process. Through the multi-level self-attention and cross-attention layers, the diffusion model is able to learn to shade G-buffers with lighting. During inference, the diffusion target can be computed as $\hat{I}_0 = D(\hat{z}_0)$, producing realistic relit images via iterative denoising.

4 Experiments

Evaluating video relighting methods in real-world scenarios is inherently challenging due to the lack of ground truth videos captured with controlled lighting variations. Therefore, we conduct our evaluation using a diverse synthetic dataset that incorporates varied camera viewpoints and relighting target object configurations. This dataset allows for extensive assessment of our proposed EDL-relighting SVD method. The specific details of our experimental setup are provided in Section 4.1. Further information regarding our training objective and implementation can be found in Section 4.3 and Section 4.4, respectively. We refer to more model implementation details in the Supplement.

4.1 Experiment Setup

Task Definition. We evaluate our method on the task of event-driven video relighting. Given an input video sequence, its corresponding G-buffers (specifically, color frames $\mathbf{I}$, depth $\mathbf{d}$, surface normals $\mathbf{n}$, albedo $\mathbf{a}$, and an event masks $\mathbf{m}$ providing lighting guidance, our method outputs a relit



(a) Input (b) Depth (c) Normal (d) Albedo (e) Object mask (f) Relit

Figure 4: Our synthetic relighting video dataset is rendered in six modalities.

version of the video sequence $\hat{\mathbf{I}}^{'}$. Performance is evaluated by comparing the generated relit frames against the ground truth relit frames $\mathbf{I}^{'}$ provided in our synthetic dataset.

Baselines. As there are currently no existing video diffusion models specifically addressing event-driven relighting with comparable inputs, we compare our approach against ScribbleLight [3], a state-of-the-art diffusion model for single-image relighting based on masking guidance, representing the closest publicly available method for controllable diffusion-based relighting, despite its focus on static images. More comprehensive details regarding the quantitative evaluation metrics and full comparison results are provided in the supplemental material.

4.2 Dataset

Synthetic data curation. To train our event-driven relighting models, we require high-quality video data with comprehensive ground-truth information for materials, geometry, and lighting conditions. Each training sample in our dataset consists of paired frames containing RGB video, depth maps, normal maps, albedo masks, and corresponding relit video frames under various lighting conditions. These multi-modal buffers are typically unavailable in real-world datasets, presenting a significant challenge for training.

We construct a synthetic dataset of 3D indoor scenes rendered from Sketchfab models available under Creative Commons licenses using a custom BlenderProc pipeline [4]. We import the associated model file for each model and apply structural normalization, including bounding box-based centering and grounding. We infer a suitable room type from metadata to scale the scene to realistic proportions. A semantic filter excludes background and ceiling elements, allowing us to sample object-centric camera trajectories with consistent framing. For each valid scene, a dynamic camera animation of 14 frames is generated, and rendered at a resolution of 512×512 pixels. We construct a synthetic dataset, named *Sketchfab-Synth*, based on three indoor scenes. For each scene, we generated 60 distinct video sequences by varying combinations of target object and camera pose, resulting in a total of 180 videos. Each sequence is 14 frames and includes comprehensive per-frame ground truth modalities: color, depth, surface normals, albedo, segmentation masks, and corresponding relit versions under various lighting conditions. This dataset was partitioned into training (80%, 144 videos) and testing (20%, 36 videos) sets, denoted as *Sketchfab-Synth-Train* and *Sketchfab-Synth-Test*, respectively.

4.3 Training Objective

We train our rendering model on a combination of a synthetic video dataset and real-world auto-labeled data, using paired G-buffer, lighting, and RGB videos. During training, for an RGB video $\mathbf{I}$, the target latent variable is defined as $\mathbf{z}_0 := \mathcal{E}(\mathbf{I})$. Noise is added to $\mathbf{z}_0$ to produce noisy image latent $\mathbf{z}_t$. The training objective is:

$$\mathcal{L}(\theta) = \| \epsilon_t - \epsilon_\theta \left(\mathbf{z}_t; \mathbf{g}, \mathbf{c}_{\text{msk}}, t \right) \|_2^2 \tag{5}$$

4.4 Implementation Details

We implement our model using PyTorch and utilize SVD [1] as our backbone, following the original SVD fine-tuning setup [1]. During training, we apply the EDM noise scheduler with 25 diffusion steps. Training our method takes 8,000 iterations using a batch size of 1.

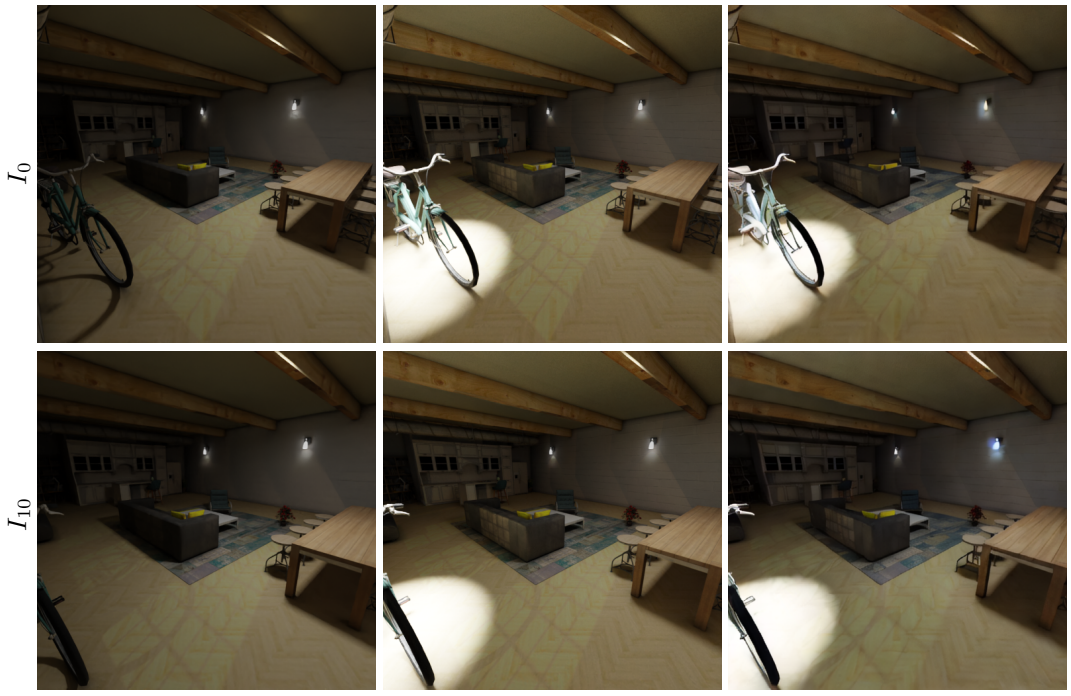


Figure 5: Visual qualitative comparison between default lighting and EDL in a video diffusion environment. We query the *bike* entity of the *big room* scenario in the narration. From left to right: original input video frames, ground truth relight frames, our relight inference. From top to bottom, showing its capability to process both single images and dynamic video sequences featuring changing camera viewpoints.

To accelerate training and facilitate initial spatial feature learning, we employed a two-stage training strategy. First, the network was trained at a resolution of 128x128 with a batch size of four. The weights from this lower-resolution model were then used to fine-tune the network at the target 512x512 resolution. To accommodate training on a single GPU, we utilized gradient accumulation with 16 steps and trained with mixed-precision (FP16). Input images were preprocessed by resizing and center cropping to 512×512. Optimization was performed using the Adam optimizer with an initial learning rate of 10^{-4} and a cosine annealing with restarts learning rate scheduler. To consider, a 0.1 dropout is applied independently to each condition channel to reduce reliance on individual conditions and potentially enhance robustness. Training our method to convergence takes approximately 2 days on four Nvidia RTX A6000 GPU cards.

5 Evaluation

In this section, we detail the design and analysis of our approach's evaluation for immersive attention guidance. A Virtual Reality (VR) application was developed, featuring a virtual environment with an accompanying narrator describing the scene's narrative. Our system dynamically applied lighting effects to objects of interest in synchronization with the narrator's description.

5.1 User Study

Experiment Design and Setup. To assess the efficacy of our event-driven lighting (EDL) approach, we conducted a between-subject study involving two groups: one experiencing the scene with our EDL approach, and the other viewing the original scene without any lighting effects (denoted as *baseline*). Participants' experiences were evaluated using the Simulator Sickness Questionnaire (SSQ) [14] and Immersive Presence Questionnaire (IPQ) [26]. Additionally, we investigated the impact of our approach on users' ability to remember the virtual scene's content by designing two tasks: an

information recall task requiring participants to identify objects present in the scene from a list of items, and a spatial memory task requiring participants to circle the placement of objects mentioned by the narrator on a map of the virtual scene.

Figure 6: Our EDL-relighting SVD is rendered as a 360° equirectangular video for the VR user study, with and without Event-Driven Lighting.

Two scene types were examined: a *static scene* generated using our language-SAM-guided SVD-based EDL implementation (exported to a 360° video to ensure that participants were able to view the scene at a high frame rate) and a *dynamic scene* featuring animated objects implemented in the Unity game engine. For each scene, the narrator delivered a 90-second script describing nine objects distributed strategically around the participant (i.e., not all objects were placed within the user's field of view upon starting the experiment).

Participants. Twenty-two participants (15 males, 7 females) were recruited from university students (age range 19-33, $\mu = 27$). Participants self-reported their gaming experience as one of four categories: 11 non-gamers, 8 casual gamers, 2 core gamers, and 1 hardcore gamer. Four participants had no prior VR experience. Participants were evenly divided into two groups for the between-subject study.

Procedure. Participants, equipped with a Meta Quest Pro headset connected to a computer, were seated on a swivel chair for the duration of the study. Participants were immersed in a 360° scene. A training session familiarized them with viewing a 360°, requiring identification of virtual objects. Subsequently, participants completed two trials, viewing either the *static scene* or the *dynamic scene*, with the order of presentation counterbalanced.

During each trial, participants were instructed to follow the narrator's story, attempting to locate referenced objects. Post-trial, participants completed SSQ and IPQ questionnaires. After viewing all scenes, participants underwent information recall and scene understanding tasks for both scenes. We presented these tasks after participants completed both trials in order to prevent participants from learning the objective of the experiment and having unnatural gaze behavior during the second trial. Post-experiment, a semi-structured interview was conducted to elicit comments and feedback from participants. The entirety of the experimental process, from the introduction to the last interview, was completed within 30 minutes. Participants were compensated with a $10 Amazon gift card upon completion of the experiment.

5.2 Quantitative Study

While diffusion models benefit significantly from large-scale data for real-world generalization, our synthetic dataset is limited in size. Consequently, our evaluation primarily validates the event-

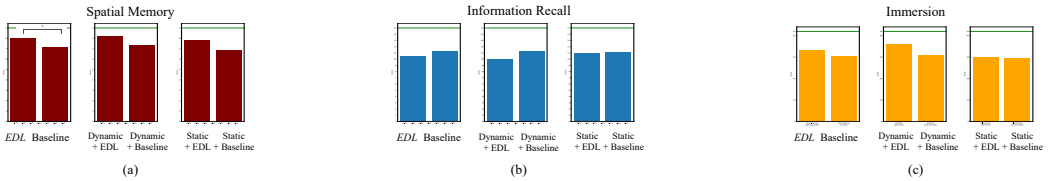


Figure 7: Experiment results for (a) the spatial memory task, (b) information recall task scores, and (c) immersion scores indicate a statistically significant difference ($p < .05$) between *EDL* and the baseline in the scene understanding task. This observation implies that our approach enhances users' ability to remember the locations of objects within the space.

driven relighting concept through a user study assessing perceptual outcomes. We further compare our approach qualitatively against the state-of-the-art masking-based relighting diffusion model, ScribbleLight [3]. Representative results are shown in Fig. 5; comprehensive quantitative evaluation and additional comparisons are provided in the **supplemental material**.

5.3 User Study Results

To assess the normality of our data, we initially conducted a Shapiro-Wilk test, revealing a departure from normal distribution ($p < .001$). Subsequently, a Mann-Whitney U Test was employed for data analysis. The results are illustrated in Figure 7.

Spatial Memory For the scene understanding task, participants labeled objects mentioned during the trial on a floor map based on their memory. The results revealed that *EDL* achieved a higher score ($\mu = 8, \sigma = 1.15$) than the baseline ($\mu = 7.14, \sigma = 1.32$), indicating a significant difference ($Z = -2.205, \mathbf{p < 0.05}$). This suggests that EDL effectively supports users in spatial memory tasks.

Information Recall In the task, participants categorized 13 objects based on whether they were i) within the scene and mentioned by the narrator, ii) in the scene but not mentioned, iii) not in the scene. Both EDL ($\mu = 10.45, \sigma = 1.71$) and baseline ($\mu = 11.23, \sigma = 1.66$) conditions yielded similar scores, with no significant difference noted ($Z = -1.346, p = .178$).

Immersion Participants reported a higher perceived immersion, the sense of "being there", with EDL ($\mu = 33.10, \sigma = 11.45$) compared to the baseline ($\mu = 30.32, \sigma = 8.71$) according to the IPQ scores. The observed difference is not statistically significant ($Z = -0.693, p = 0.488$).

Motion Sickness To assess the occurrence of motion sickness during the trials, participants completed SSQ questionnaires before and after each trial. The results of a Wilcoxon Signed Ranks Test showed no significant difference between pre- and post-SSQ scores for both the first trial ($Z = -.656, p = .512$) and the second trial ($Z = -.211, p = .833$). These findings suggest that participants overall experience was not adversely affected by motion sickness.

6 Related Work

Attention Guidance in Virtual Environments. The problem of directing user attention has long been explored in the context of virtual reality (VR), with applications spanning narrative media, gaming, and immersive training environments. Traditional techniques such as automatic viewpoint redirection [22, 19] aim to rotate users toward points of interest, improving recall and focus. While effective, these techniques may induce discomfort or disrupt user agency, motivating more subtle alternatives.

Visual manipulation techniques—e.g., spatial blurring [28, 11] and differential rendering—guide gaze while preserving realism. Later work introduced more immersive methods, such as SwiVRChair [9], which physically rotates users toward targets, and particle-based cues like HiveFive [17] to simulate natural attention shifts. However, such systems often require hardware coupling and are not easily transferable to non-interactive video content. Our approach reinterprets these ideas for passive video settings: instead of explicitly guiding gaze, we induce soft attention through dynamic, context-aware relighting derived from high-level semantic events.

Event Detection via Large Language Models. Event extraction traditionally relies on predefined schemas and supervised models [7, 21], with representations defined over textual entities and roles [20]. Transformer-based models like BERT [5] have demonstrated strong performance on both event detection and argument extraction, often framed as classification [29] or question answering [6]. More recently, zero-shot prompting of large language models (LLMs) such as GPT-3 has enabled more flexible and open-domain event detection [8], reducing reliance on task-specific training data.

Unlike most prior work, which focuses on purely textual modalities, our system operates across modalities: we use an LLM to detect semantically salient events from either captions or prompts and then associate these events with spatial regions via SAM-2 object segmentation [23]. This multi-stage approach enables contextual relighting based on high-level intent, bridging vision-language interfaces with physically grounded scene manipulation.

Video Editing and Generation with Diffusion Models. Diffusion-based generative models have recently been extended to video domains [2, 27]. Techniques like ControlVideo [32] and AnimateDiff [10] improve temporal coherence via shared noise maps or latent warping, enabling consistent motion and appearance across frames. These models excel at stylization and animation synthesis but lack mechanisms for localized relighting or scene-aware editing. Moreover, they operate primarily in the generative setting, where video content is synthesized from scratch.

In contrast, our work focuses on relighting existing video content. We build on stable video diffusion backbones but introduce spatial-temporal control via lighting prompts and semantic attention maps, allowing fine-grained, frame-consistent illumination adjustments without retraining on entire video sequences.

Diffusion-Based Relighting and Illumination Control. Single-image relighting via diffusion models has emerged as a powerful tool for photorealistic manipulation. DiLightNet [30] employs depth-aware radiance priors for fine-grained edits, while IC-Light [31] harmonizes light transport across composite objects. Neural Gaffer [12] achieves category-agnostic relighting using environment maps, and ScribbleLight [3] supports localized lighting edits via sparse scribbles. However, all of these methods focus on static images and do not account for temporal consistency, making them unsuitable for video.

Furthermore, their reliance on explicit masks or depth priors can be fragile in dynamic settings. Our method eliminates this dependency by grounding lighting in semantically meaningful events and adapting the target object over time via masking control within a diffusion framework. We demonstrate that such event-driven relighting can replicate viewer-guidance effects while maintaining temporal coherence and realism across frames.

7 Conclusion

We proposed an automatic method for driving lighting control in immersive environments based on narrative events detected through language-guided segmentation. Our approach leverages SVD to enable video relighting using semantic masks from language input, demonstrating controllable relighting without complex 3D scene representations. Human evaluation revealed our EDL technique improved users' spatial memory, with semi-structured interviews showing preference for our approach, characterized as more "cinematic" and immersive. Participants showed statistically significant spatial memory improvement, noting the EDL approach aided narrative understanding and attention direction.

Our approach successfully extends SVD to generate relit video sequences. However, limitations exist, mainly our reliance on synthetic scenes for training data, where incorporating real-world datasets could improve generalization.

Future work. Future directions include improving language-model event detection reliability, studying more complex lighting schemes, and extending our method to account for real-world videos or spatially varying materials. The automatic event-driven lighting control promises benefits for immersive applications where narrative-driven lighting enhances engagement and spatial understanding. This capability holds significant promise for enhancing viewer immersion across various media and applications, including film, television, and interactive virtual reality.

References

[1] Andreas Blattmann, Tim Dockhorn, Sumith Kulal, Daniel Mendelevitch, Maciej Kilian, Dominik Lorenz, Yam Levi, Zion English, Vikram Voleti, Adam Letts, et al. Stable video diffusion: Scaling latent video diffusion models to large datasets. *arXiv preprint arXiv:2311.15127*, 2023.

[2] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. Video generation models as world simulators. 2024.

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

[4] Maximilian Denninger, Dominik Winkelbauer, Martin Sundermeyer, Wout Boerdijk, Markus Wendelin Knauer, Klaus H Strobl, Matthias Humt, and Rudolph Triebel. Blenderproc2: A procedural pipeline for photorealistic rendering. *Journal of Open Source Software*, 8(82):4901, 2023.

[5] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[6] Xinya Du and Claire Cardie. Event extraction by answering (almost) natural questions. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 671–683, Online, November 2020. Association for Computational Linguistics.

[7] Shaoyang Duan, Ruifang He, and Wenli Zhao. Exploiting document level information to improve event detection via recurrent neural networks. In Greg Kondrak and Taro Watanabe, editors, *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 352–361, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[8] Jun Gao, Huan Zhao, Changlong Yu, and Ruifeng Xu. Exploring the feasibility of chatgpt for event extraction. *arXiv preprint arXiv:2303.03836*, 2023.

[9] Jan Gugenheimer, Dennis Wolf, Gabriel Haas, Sebastian Krebs, and Enrico Rukzio. Swivrchair: A motorized swivel chair to nudge users' orientation for 360 degree storytelling in virtual reality. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pages 1996–2000, 2016.

[10] 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. *arXiv preprint arXiv:2307.04725*, 2023.

[11] Hajime Hata, Hideki Koike, and Yoichi Sato. Visual guidance with unnoticed blur effect. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 28–35, 2016.

[12] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. *arXiv preprint arXiv:2406.07520*, 2024.

[13] Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in neural information processing systems*, 35:26565–26577, 2022.

[14] Robert S Kennedy, Norman E Lane, Kevin S Berbaum, and Michael G Lilienthal. Simulator sickness questionnaire: An enhanced method for quantifying simulator sickness. *The international journal of aviation psychology*, 3(3):203–220, 1993.

[15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 4015–4026, 2023.

[16] Xueyan Lai, Rui Ji, Changhuai Lang, Yahong Song, Xiuzhe Du, and Qi Tian. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 14*–*, 2023.

[17] Daniel Lange, Tim Claudius Stratmann, Uwe Gruenefeld, and Susanne Boll. Hivefive: Immersion preserving attention guidance in virtual reality. In *Proceedings of the 2020 CHI conference on human factors in computing systems*, pages 1–13, 2020.

[18] Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Zhi-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, et al. Diffusionrenderer: Neural inverse and forward rendering with video diffusion models. *arXiv preprint arXiv:2501.18590*, 2025.

[19] Yen-Chen Lin, Yung-Ju Chang, Hou-Ning Hu, Hsien-Tzu Cheng, Chi-Wen Huang, and Min Sun. Tell me where to look: Investigating ways for assisting focus in 360 video. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, pages 2535–2545, 2017.

[20] Kang Liu, Yubo Chen, Jian Liu, Xinyu Zuo, and Jun Zhao. Extracting events and their relations from texts: A survey on recent research progress and challenges. *AI Open*, 1:22–39, 2020.

[21] Yaojie Lu, Hongyu Lin, Jin Xu, Xianpei Han, Jialong Tang, Annan Li, Le Sun, Meng Liao, and Shaoyi Chen. Text2Event: Controllable sequence-to-structure generation for end-to-end event extraction. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2795–2806, Online, August 2021. Association for Computational Linguistics.

[22] Lasse T Nielsen, Matias B Møller, Sune D Hartmeyer, Troels CM Ljung, Niels C Nilsson, Rolf Nordahl, and Stefania Serafin. Missing the point: an exploration of how to guide users' attention during cinematic virtual reality. In *Proceedings of the 22nd ACM conference on virtual reality software and technology*, pages 229–232, 2016.

[23] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

[24] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[25] Sylvia Rothe, Daniel Buschek, and Heinrich Hußmann. Guidance in cinematic virtual reality-taxonomy, research status and challenges. *Multimodal Technologies and Interaction*, 3(1):19, 2019.

[26] Thomas Schubert, Frank Friedmann, and Holger Regenbrecht. The experience of presence: Factor analytic insights. *Presence: Teleoperators & Virtual Environments*, 10(3):266–281, 2001.

[27] Uriel Singer, Adam Polyak, Thomas Hayes, Xi Yin, Jie An, Songyang Zhang, Qiyuan Hu, Harry Yang, Oron Ashual, Oran Gafni, et al. Make-a-video: Text-to-video generation without text-video data. *arXiv preprint arXiv:2209.14792*, 2022.

[28] Wayne S Smith and Yoav Tadmor. Nonblurred regions show priority for gaze direction over spatial blur. *The Quarterly Journal of Experimental Psychology*, 66(5):927–945, 2013.

[29] Sen Yang, Dawei Feng, Linbo Qiao, Zhigang Kan, and Dongsheng Li. Exploring pre-trained language models for event extraction and generation. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 5284–5294, 2019.

[30] 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*, pages 1–12, 2024.

[31] 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.

[32] Yabo Zhang, Yuxiang Wei, Dongsheng Jiang, Xiaopeng Zhang, Wangmeng Zuo, and Qi Tian. Controlvideo: Training-free controllable text-to-video generation. *arXiv preprint arXiv:2305.13077*, 2023.

A Technical Appendices and Supplementary Material

More visual and quantitative results please check our supplemental material.

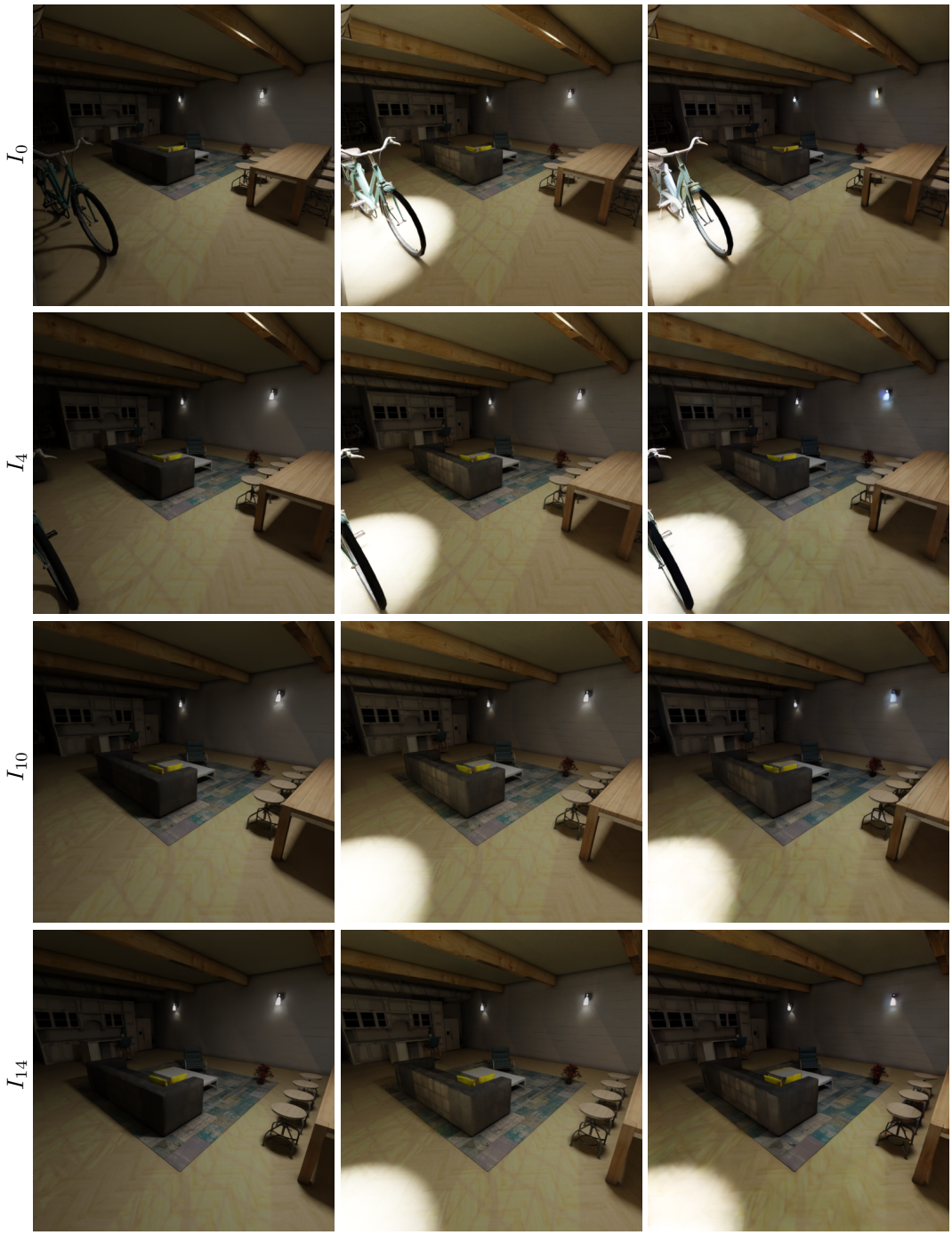


Figure 8: Relight object *bike* in scene *big room*. (View model in Sketchfab: Big Room 3D Model).

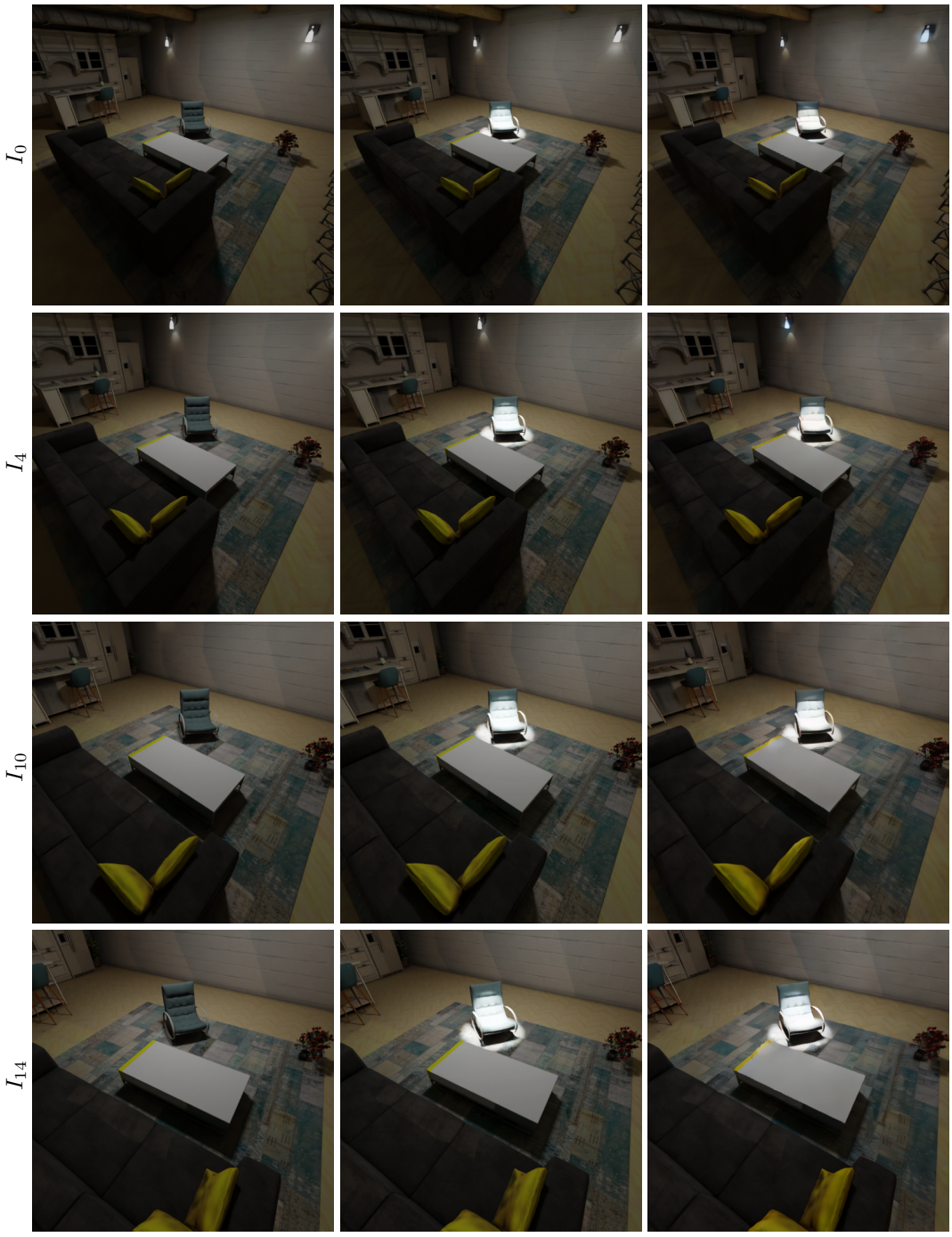


Figure 9: Relight object *lounge chair* in scene *big room*

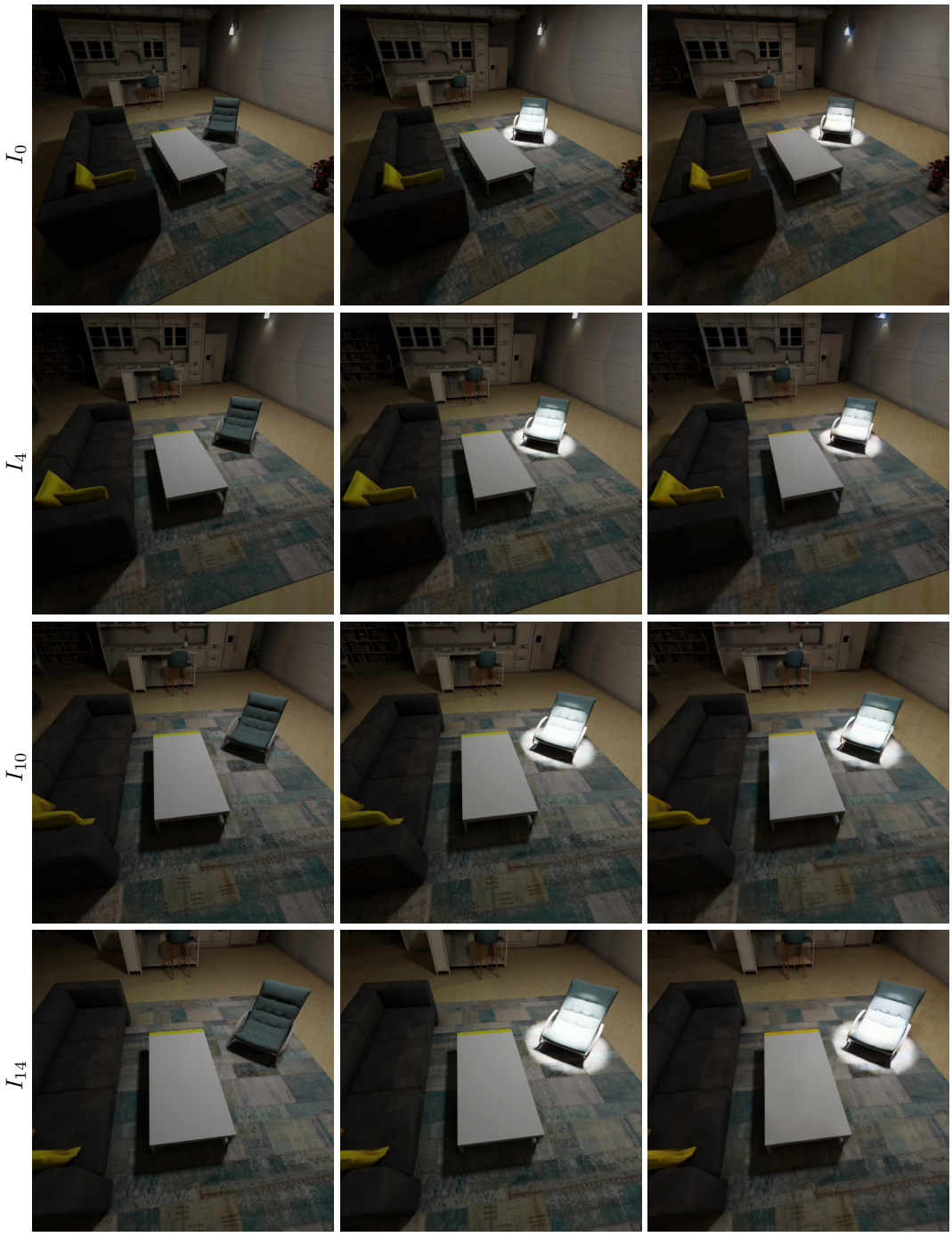


Figure 10: Relight object *lounge chair* in scene *big room*

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]