Video Generation Models Are Inherent Lighting Estimators

Ziqi Cai[1,3], Shuchen Weng[2,1*], Kaiqi Liu[1], Zifeng Wang[1], Zhiquan Zhang[1], Minggui Teng[1], Han Jiang[3], and Boxin Shi[1*]

[1] Peking University [2] Beijing Academy of Artificial Intelligence
[3] OpenBayes Information Technology Co., Ltd.
czq@stu.pku.edu.cn, {shuchenweng, shiboxin}@pku.edu.cn

Abstract. Recovering dynamic environment maps from a single in-the-wild video is crucial for photorealistic rendering, yet remains a challenge. Recent video generation models can produce photorealistic scenes with complex lighting, possessing an inherent understanding of lighting. In this paper, we introduce V-LITE (Video generation models are inherent lighting estimators), a framework that unlocks this internal knowledge by reframing lighting estimation as a guided video inpainting task. Inspired by visual effects (VFX) industry practices, we insert a synthetic chrome ball into the scene to compel the model to generate physically plausible reflections from the surrounding spatio-temporal context. To bridge the gap from LDR-native models to the HDR domain, we design an HDR-aware VAE and employ an efficient LoRA-based fine-tuning strategy. We then construct a mixed dataset comprising high-fidelity HDR images to provide realistic HDR priors, and in-the-wild HDR videos to provide dynamic spatio-temporal context. Extensive experiments demonstrate that V-LITE produces temporally coherent HDR environment maps, revealing that modern video diffusion models are not merely synthesizers but also powerful, inherently capable estimators of physical scene lighting.

Keywords: Video Diffusion Models · Dynamic Lighting Estimation · High Dynamic Range

1 Introduction

As the primary factor to determine the visual appearance of in-the-wild scenes, an accurate lighting representation is the foundational cornerstone for photorealistic augmented reality [12], virtual object insertion [25], and scene relighting [5, 30]. However, recovering the complete environment map from a single

[1] Ziqi Cai, Shuchen Weng, Zifeng Wang, Zhiquan Zhang, Minggui Teng, and Boxin Shi are with State Key Laboratory of Multimedia Information Processing and National Engineering Research Center of Visual Technology, School of Computer Science, Peking University. Kaiqi Liu is with School of Software and Microelectronics, Peking University.

[*] Corresponding authors.

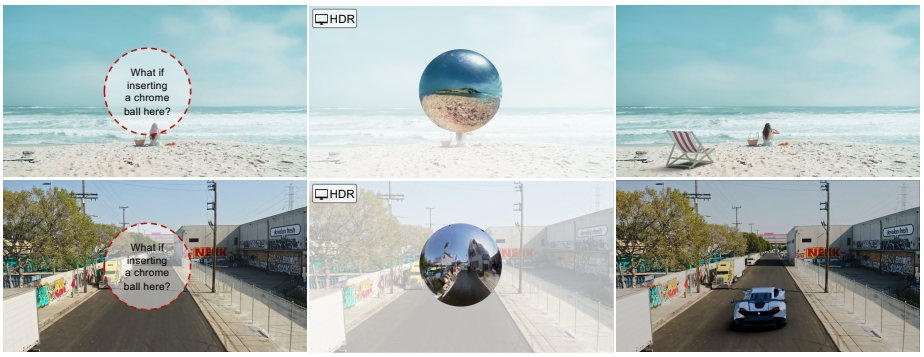

Fig. 1: An illustration of our V-LITE framework. (**a**) Inspired by on-set VFX practices, we reformulate dynamic lighting estimation as the task of inserting a virtual light probe into any in-the-wild LDR video. (**b**) We implement this concept as a video inpainting task, compelling the video diffusion model to leverage its inherent priors and render a physically plausible chrome ball that captures the dynamic illumination of the scene. (**c**) The resulting HDR environment map enables downstream applications, such as inserting virtual objects with realistic and temporally consistent lighting and shadows.

in-the-wild video remains a significant challenge, due to the inherently partial illumination observations and the complex interactions among camera motion, object occlusions, and diverse material properties.

Classical physics-based methods [24] typically solve a simplified problem, assuming videos have static lighting, Lambertian surfaces, or known geometry. Meanwhile, multi-view approaches [33] solve the problem by introducing external guidance (*e.g.*, images from additional perspectives). However, both paradigms fundamentally deviate from the intended task of processing an in-the-wild video. In contrast, recent work [26] feeds videos as conditions to generate low dynamic range (LDR) results for high dynamic range (HDR) fusion. While this approach presents a practical solution, it treats the model as a simple translator, ignoring its *inherent* generative priors for illumination and thus potentially producing implausible results, such as exposure-fusion artifacts, color bias, and temporally unstable reflections.

Recent video generation models [7,14,19,36,44] have demonstrated an impressive capability to produce photorealistic scenes with complex light and shadow interactions, which highlights their implicit understanding of illumination and suggests an inherent capability for lighting estimation. This motivates us to explore these capabilities to directly generate HDR environment maps. However, extending existing video generation frameworks for lighting estimation introduces two key challenges: *(i)* adapting the model to handle HDR processing, and *(ii)* enabling HDR representations within the latent diffusion model.

In this paper, we introduce V-LITE (**V**ideo generation models are inherent **LI**gh**T**ing **E**stimators), a unified HDR-integrated framework that estimates en-

vironment maps directly from a single in-the-wild video. We are inspired by the foundational practice in the visual effects industry, where a light probe (typically a chrome ball) is inserted into a scene to capture the complete surrounding illumination [28, 29]. This enables rendering virtual objects with photorealistic lighting consistency. Therefore, we reformulate the lighting estimation problem as a video inpainting task, requiring the model to fill in a synthetic chrome ball, which is then unwrapped to produce the final environment map. This approach forces the video generation model to consider both spatial context and temporal consistency when plausibly inserting the ball, thereby *inherently* leveraging its generative priors. As shown in Fig. 1, our method robustly generates high-quality environment maps even in extreme cases (*e.g.*, fast camera motion).

We build our framework upon a diffusion-based network for LDR video generation [36]. To process HDR videos with original LDR Variational AutoEncoder (VAE), we design and integrate learnable tonemap adapters, ensuring the fine-grained illumination cues (*e.g.*, color temperature variations across time) are faithfully preserved in the latent space. To further align the pretrained latent space with these HDR latents, we introduce a LoRA-based fine-tuning scheme [20]. This effectively enhances the lighting understanding while preserving its original generative priors for spatial and temporal context. To facilitate the model training and evaluation, we also construct V-LITESet, a dataset including over 8K HDR video pairs with dynamic temporal lighting and 800 static HDR images, totaling over 648K frames for effective co-training and evaluation.

Our contributions can be summarized as follows:

- We reformulate video lighting estimation as a light probe inpainting task, producing a dynamic environment map from a single in-the-wild video.
- We design an HDR-aware VAE to preserve lighting cues and a LoRA-based fine-tuning scheme to align the latent space with HDR representations.
- We construct a hybrid dataset comprising 8K videos with dynamic temporal lighting and 800 images with diverse luminance distributions, facilitating robust model training and comprehensive evaluation.

2 Related Work

2.1 Lighting Estimation from Images

Estimating scene illumination from a single image has been studied extensively in both vision and graphics. Early deep methods directly regress HDR indoor environment maps from limited-FoV LDR photos, showing that plausible light probes can be inferred without explicit geometry or material supervision [15]. Subsequent works decompose the task into geometry/visibility completion and LDR-to-HDR mapping to better preserve high-frequency cues and material-dependent effects [32, 45]. To improve spatial fidelity, fast indoor lighting estimation is explored [16], while HDR panorama generation enables lighting editing and robust HDR mapping [6, 34, 37]. Neural inverse rendering provides complementary,

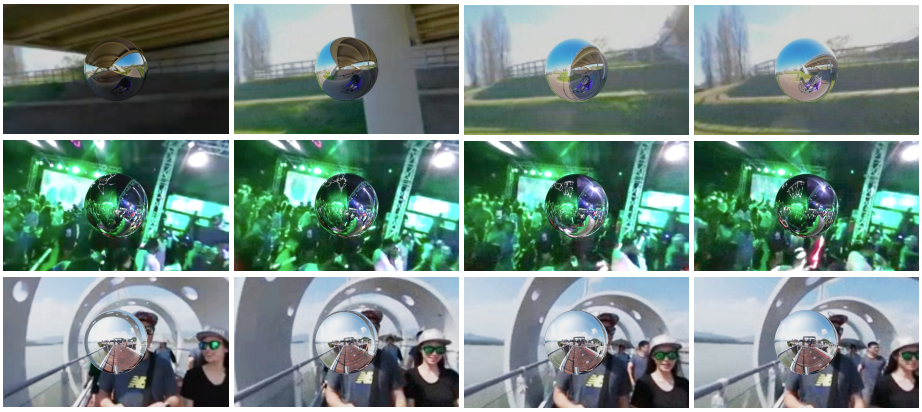

Fig. 2: Samples from the V-LITESet dataset. Each row corresponds to a video, and the images in the row represent different timestamps sampled from the video.

physics-grounded priors by jointly reasoning about geometry, BRDFs, and illumination from a single image [31]. For outdoor scenes, single-image methods often predict parametric sun/sky lighting [18]. Despite these advances, single-image approaches typically lack *temporal coupling* and rarely provide *native HDR* training/inference under unconstrained conditions.

2.2 Lighting Estimation from Videos

Moving from images to videos introduces the challenge of maintaining temporal consistency while still modeling dynamic illumination. Early progress leverages multi-view or stereo cues to infer lighting volumes with 3D coherence [33]. Closer to our setting, Li *et al.* [24] represent video illumination as a spatiotemporally consistent spherical-Gaussian lighting volume and train recurrent priors to suppress flicker and enforce cross-frame consistency for indoor HDR lighting. Recent work further studies spatiotemporally consistent indoor lighting estimation directly from in-the-wild videos, emphasizing smoothness of predicted lighting fields under changing viewpoints and local light variations [35]. In parallel, sequence-based HDR lighting reconstruction methods encode spatially-varying illumination as Gaussian splats from image sequences [4], but typically require controlled capture settings and do not target generic casually captured videos. Methods that aim at lightweight 3D-coherent representations can also improve spatial consistency (*e.g.*, LightOctree), but they are primarily designed for single-image inputs and do not directly couple predictions across time [39]. However, robust estimation for *in-the-wild* videos that is simultaneously *temporally consistent* and supports *native HDR* remains under-explored, which we address by leveraging video diffusion priors and task-specific adaptation.

2.3 Generative Diffusion Priors

Generative diffusion priors have been widely used in various computer vision tasks. Due to their strong capability to model the complex distribution of natural images, they are naturally explored to handle low-level vision tasks (*e.g.*, dehaze [23], colorization [8], and super-resolution [43]). This further motivates researchers to extend their application to classic vision tasks (*e.g.*, object detection [10], segmentation [9], and pose estimation [13]) and even 3D vision for reconstruction [38] and point cloud completion [21]. Recently, researchers have also demonstrated that video diffusion models have zero-shot reasoning capabilities for the visual world, implicitly encoding complex physics, geometry, and material optics [41]. Based on these previous works, we reformulate the video lighting estimation problem and explore the approach to leverage the generative priors of video diffusion models to capture the dynamic illumination.

3 V-LITESet Dataset

Existing large-scale video generation datasets [2, 40] primarily focus on LDR video content, making them ill-suited for developing HDR models. In contrast, existing HDR video datasets either focus on indoor scenarios [15] or provide limited in-the-wild samples [17]. To overcome these limitations, we introduce V-LITESet, a high-quality dataset specifically designed for training and evaluating HDR video lighting estimation.

Dynamic spatiotemporal subset. We build the dynamic subset of V-LITESet upon the PanoVid dataset [42], which includes diverse in-the-wild LDR panoramic videos. To lift these LDR videos to HDR, we employ a rigorous offline curation pipeline. Rather than assuming a single camera response function, we process the source panoramas using three distinct inverse tone-mapping curves (linear, $\gamma = 2.2$, and filmic) to span a diverse set of candidate HDR distributions. To filter out severe photometric losses, we map these candidates back to the LDR domain, and query a VLM [1] to evaluate the physical plausibility of the frames (*e.g.*, unnatural color banding, severe overexposure, and irreversible detail loss). Approximately 57% of the raw sequences are discarded due to low confidence scores. The remaining videos provide robust dynamic spatiotemporal context.

Static photometric subset. Since photometric approximations for tonemapped HDR videos may introduce certain artifacts, we additionally integrate 800 high-quality HDR panoramic images from PolyHaven [17] to form the static subset of V-LITESet. We adapt these images into static video sequences to match the input format of our video diffusion backbone. These real-world samples serve as absolute physical anchors during training, enabling the model to effectively learn the intrinsic luminance distributions and HDR contrasts of light sources (*e.g.*, the accurate intensity ratio between the sun and the sky).

Summary. The final V-LITESet comprises over 8K video pairs at 480×832 resolution with dynamic temporal lighting, alongside 800 HDR static sequences with diverse luminance distributions. This amounts to over 648K frames, split into

567K training and 81K evaluation samples. This provides a robust and comprehensive foundation for advancing research in dynamic HDR lighting estimation from a single video.

4 Methodology

In this section, we introduce the details of our V-LITE framework. We first review the flow-matching video generation and editing models that form our foundation (Sec. 4.1). Next, we propose the end-to-end VAE for HDR video processing to adapt the pretrained video backbone into the HDR domain (Sec. 4.2). Finally, we present our strategy to reformulate lighting estimation as a video inpainting task, leveraging its inherent generative priors to accurately capture the scene's dynamic lighting (Sec. 4.3). The overall pipeline is shown in Fig. 3.

4.1 Preliminaries

Video generation models Video generation models have demonstrated strong capabilities in producing diverse and high-quality dynamic content. As a representative practice of the video backbone, Wan 2.1 [36] adopts a diffusion Transformer architecture, modeling complex spatio-temporal dynamics through a latent-space training scheme based on the flow matching framework [27].

Given a target video x_1, the pretrained VAE encoder $\mathcal{E}$ first maps it into a latent code $z_1 = \mathcal{E}(x_1)$. During training, the model randomly samples a noise latent $z_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and creates an intermediate latent code via linear interpolation for a timestep $t \in [0, 1]$: $z_t = t \cdot z_1 + (1 - t) \cdot z_0$. The model is trained to predict the ground-truth velocity field associated with the interpolation path. Formally, the training objective is:

$$\mathcal{L} = \mathbb{E}_{z_0, z_1, c_{\text{txt}}, t}\big[\|v_\theta(z_t, c_{\text{txt}}, t) - (z_1 - z_0)\|^2\big], \tag{1}$$

where c_{txt} denotes the text condition and θ represents the model parameters. $v_\theta(\cdot)$ is the diffusion Transformer to predict the velocity field at latent code z_t.

Video editing models Video editing aims to modify the video content based on the text description indicating the editing direction and the semantics of the original video. To selectively transform targeted regions while maintaining temporal and spatial coherence in unedited areas, a recent method [22] proposes to incorporate text, mask, and other visual conditions into a unified conditioning representation. This conditioning representation is injected into the video backbone via an additional adapter. The training objective is thus formulated as:

$$\mathcal{L} = \mathbb{E}_{z_0, z_1, \mathcal{V}, t}\big[\|v_\theta(z_t, \mathcal{V}, t) - (z_1 - z_0)\|^2\big], \tag{2}$$

where $\mathcal{V} = (T, V, M)$ denotes the conditioning input, with T representing the text description, V denoting the visual conditions, and M providing the corresponding masks.

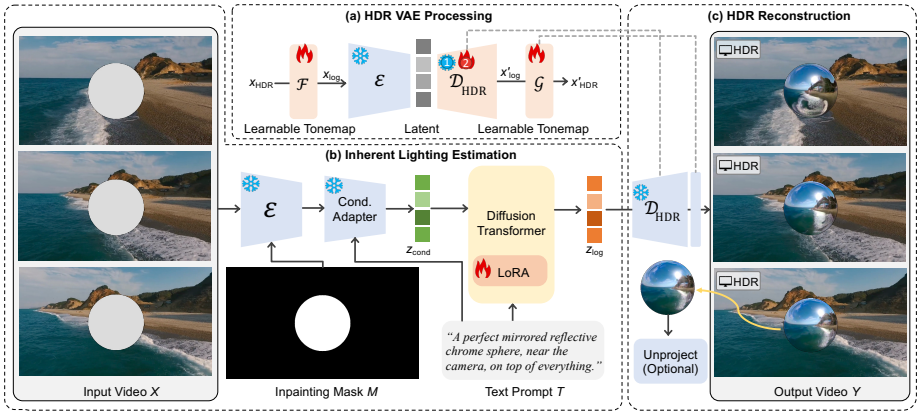

Fig. 3: Overview of the V-LITE pipeline. **(a)** HDR VAE processing. To operate in log-domain space, an input HDR video x_{HDR} is mapped by a learnable tone mapping network $\mathcal{F}$ to $x_{\log}$, encoded by $\mathcal{E}$, and decoded by $\mathcal{D}_{\text{HDR}}$ into $x'_{\log}$. An inverse network $\mathcal{G}$ then reconstructs the HDR video x'_{HDR}. We pretrain $\mathcal{F}$ and $\mathcal{G}$ before jointly optimizing them with $\mathcal{D}_{\text{HDR}}$. **(b)** Inherent lighting estimation. For an LDR video X masked by M (grey central region), $\mathcal{E}$ extracts latents that a condition adapter encodes into control signals z_{cond}. Conditioned on z_{cond} and a text prompt T, a LoRA-adapted flow-matching diffusion Transformer inpaints the masked region, outputting the full latent sequence $z_{\log}$. **(c)** HDR reconstruction. The HDR decoder $\mathcal{D}_{\text{HDR}}$ maps $z_{\log}$ to the final video Y featuring physically coherent reflections. Optionally, these chrome balls can be unprojected to estimate the scene's dynamic HDR environment map sequence, providing a powerful tool for downstream tasks like virtual object insertion.

4.2 HDR VAE Processing

The success of leveraging inherent generative priors for lighting estimation depends critically on operating in the HDR domain. However, since existing video generation backbones [36, 44] are designed for LDR content, it is essential to adapt their architectures to accurately support end-to-end HDR processing.

Existing methods [11,26,29] typically address this gap by producing multiple LDR variants (*e.g.*, exposure-bracketed frames) that are subsequently fused to approximate an HDR result. However, such approaches require multiple inference passes or an additional fusion stage, leading to increased computational cost and potential error accumulation across the reconstruction pipeline. In contrast, we propose to integrate the HDR domain conversion directly into the VAE, bridging the distribution gap within the latent space. As presented in Fig. 3 (a), this design inherently makes the LDR-trained video backbone compatible with HDR data.

Tonemap adapters. We first introduce a pair of learnable tonemap adapters and integrate them into the VAE [36] for end-to-end HDR video processing. Specifically, given a linear HDR video x_{HDR}, our tonemap adapter converts it to the log domain. It then applies a single 3D convolutional layer $\mathcal{F}$, a learnable

scale $\mathbf{s}$ and bias $\mathbf{b}$ to calculate the log-domain representation:

$$x_{\log} = \mathbf{s} \odot \big[\log(x_{\text{HDR}} + \epsilon) + \mathcal{F}\big(\log(x_{\text{HDR}} + \epsilon)\big)\big] + \mathbf{b}. \tag{3}$$

This adaptive mapping enables the encoder to dynamically stretch and compress signal ranges. The log domain representation $x_{\log}$ is fed into the LDR VAE encoder $\mathcal{E}$ to generate the compressed latent code $z = \mathcal{E}(x_{\log})$. After that, the HDR VAE decoder $\mathcal{D}_{\text{HDR}}$ produces the log-domain representation $x'_{\log} = \mathcal{D}_{\text{HDR}}(z)$. The inverse tonemap adapter then maps this representation back to the linear HDR domain:

$$x'_{\text{HDR}} = \exp\left[\frac{x'_{\log} - \mathbf{b}}{\mathbf{s}} + \mathcal{G}\left(\frac{x'_{\log} - \mathbf{b}}{\mathbf{s}}\right)\right], \tag{4}$$

where $\mathcal{G}$ is a single 3D convolutional layer and x'_{HDR} is the reconstructed HDR video. During the training process, these lightweight adapters learn the dynamic illumination and extreme ranges within a stable log-domain latent space. Notably, our HDR VAE decoder has the exact same architecture as the pretrained LDR VAE from the video backbone [36].

Training scheme. The VAE training is formulated as a reconstruction task, requiring the model to minimize the error between the input video x_{HDR} and the reconstructed video x'_{HDR}. To adapt our VAE with tonemap adapters for stable HDR processing, we employ a two-stage finetuning strategy. In the first stage, we freeze all VAE parameters and train the two tonemap adapters to establish a stable log-domain representation. In the second stage, we unfreeze the VAE decoder and jointly finetune it with the adapters. After finetuning, the combined VAE and tonemap adapters operate directly in the HDR domain, preserving visual fidelity across extreme luminance ranges typical of HDR data.

4.3 Inherent Lighting Estimation

Previous work [26] employs video generation models as a video translator for lighting estimation. In contrast, observing that recent video generation models [36] demonstrate an implicit ability to model lighting, producing temporally and spatially light-coherent results, we are motivated to directly exploit their internal lighting modeling capability rather than formulating the task as a translation problem. As illustrated in Fig. 3 (b), we reformulate the task of video lighting estimation as an inpainting problem.

Since the central region of the scene is typically irrelevant for global illumination, we leverage a central mask M to define the inpainting region for the synthesis of a physically plausible chrome ball. Given an in-the-wild LDR video x_{LDR}, we construct the masked input as $x_{\text{input}} = x_{\text{LDR}} \odot (1-M) + C \odot M$, where C is a constant gray value. We then decouple this input into the environment region $x_e = x_{\text{input}} \odot M$ and the scene content $x_s = x_{\text{input}} \odot (1 - M)$ as two conditions for the generation model. This setup allows the model to simultaneously reconstruct the HDR environment map from the surrounding regions and enhance the dynamic range of the original scene.

Diffusion process. During training and inference, both x_s and x_e are represented in the latent space as z_s and z_e using the pretrained LDR VAE encoder $\mathcal{E}$. At each inference step t, we utilize a condition adapter [22] to inject these latent codes into the video backbone to predict the velocity field:

$$\hat{v} = f_\theta\big(\hat{z}_t, f_c([z_s, z_e, c_{\text{txt}}]), c_{\text{txt}}\big), \tag{5}$$

where f_θ denotes the diffusion Transformer of the video backbone, f_c denotes the condition adapter for video inpainting, and c_{txt} represents the text embedding driving the model to fill the chrome ball. During inference, we iteratively solve the ordinary differential equation $\mathrm{d}z_t/\mathrm{d}t = \hat{v}$ from $t = 1$ to $t = 0$ using a numerical solver to synthesize the chrome ball. In particular, the prompt we provide is *"A perfect mirrored reflective chrome sphere, near the camera, on top of everything"*, which explicitly specifies the appearance and placement of the chrome ball.

After the diffusion process, our pretrained VAE decoder $\mathcal{D}_{\text{HDR}}$ is employed to reconstruct HDR video frames $\hat{x}_{\text{HDR}}$. To extract the final HDR environment map $\mathbf{E}$, we perform a post-processing step, where the chrome ball is isolated from the reconstructed HDR frames using the corresponding mask and re-projected into the environment map:

$$\mathbf{E} = f_{\mathcal{R}}(\hat{x}_{\text{HDR}} \odot M), \tag{6}$$

where $f_{\mathcal{R}}$ denotes the equirectangular transformation. This procedure produces a high-fidelity and temporally coherent sequence of HDR environment maps that faithfully capture the dynamic lighting of the scene.

Optimization process. We train only the core video diffusion Transformer using LoRA-based fine-tuning [20] to align its generative priors with the HDR latent distribution. During training, we randomly mix samples from the dynamic spatiotemporal and static photometric subsets at a $10:1$ ratio. To enable this joint training, we replicate the HDR images within the photometric subset to form static video sequences. Consequently, this mixed-data training strategy ensures that the model preserves its inherent temporal consistency priors through the dynamic videos, while utilizing the static sequences as physical anchors to learn accurate luminance distributions (*e.g.*, precise sun intensities).

To simultaneously reconstruct the HDR environment map and enhance the dynamic range of the original scene context, we define our objective as follows:

$$\mathcal{L} = \mathbb{E}_{z,\epsilon,t}\big[\|M' \odot (v_\theta(z_t, \mathcal{C}, t) - v)\|^2\big], \tag{7}$$

where $\mathcal{C} = [z_{\text{cond}}, c_{\text{txt}}]$ denotes the full conditioning input, $z_{\text{cond}} = \{z_e, z_s\}$, and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ represents a latent noise sample. Here, $v = z - \epsilon$ denotes the target velocity, v_θ is the predicted velocity field, and $M' = 1 + (\alpha - 1)M$ is a modulation mask with a weighting factor $\alpha = 5$ to explicitly emphasize the lighting estimation within the central inpainting region.

5 Experiments

In this section, we conduct a series of experiments to validate the effectiveness of our proposed method, V-LITE. We first introduce the experimental setup, including our implementation details, the dataset used, and the evaluation metrics. We then present comprehensive quantitative and qualitative comparisons against state-of-the-art methods. Finally, we perform detailed ablation studies to analyze the contribution of each key component in our framework.

We compare V-LITE against DiffusionLight [29], DiffusionLightTurbo [11] and StyleLight [37]. Since most existing methods do not support dynamic lighting from a single video, we adapt state-of-the-art single-image lighting estimation methods to the video domain by applying them in a per-frame manner.

5.1 Implementation Details

Our framework is built upon the pretrained Wan 2.1-1.3B video model [36]. Our custom-trained HDR-aware VAE is specifically designed to operate in the linear Rec. 2020 color space, which provides a wide-gamut, scene-linear representation aligned with HDR standards, ensuring physically accurate color reproduction. To effectively model the high dynamic range, the VAE's latent codes are compressed into a log space before being processed by the core diffusion model. For HDR-aware VAE fine-tuning, we adopt a two-stage training scheme with 10k steps in the first stage and 5k steps in the second. We optimize the model using the standard VAE objective, which includes the reconstruction loss and KL divergence loss, identical to the losses used in the original VAE training. For diffusion Transformer fine-tuning, we employ LoRA [20] with a rank of $r = 32$ and apply it to the attention layers of the diffusion Transformer. The model is trained on our V-LITESet for 100K steps with a global batch size of 64 on 8 NVIDIA H100 GPUs. We use the AdamW optimizer with a learning rate of 1×10^{-5}. During training and inference, the input videos containing the masked sphere are processed at a resolution of 480×832. After the model inpaints the sphere, its reflection is unwrapped to produce the final HDR environment map at a resolution of 256×512, which is commonly used in prior work [11,26,29,37].

5.2 Quantitative Evaluation

Evaluation protocol. We conduct quantitative evaluations on two public benchmarks following the evaluation protocol of DiffusionLight [29]. *(i) Editable Indoor* assesses illumination quality through scene rendering and reports MSE, SI-MSE, Angular Error in Radians (AER), and Lighting Stability (LS, the standard deviation of SI-MSE). *(ii) EnvMapNet* estimates the dominant lighting direction from the environment maps and reports both the Angular Error in Degrees (AED) and Angle Stability (AS, the standard deviation of AED).

Baselines. We compare our method against DiffusionLight [29], DiffusionLightTurbo [11], and StyleLight [37]. Note that DiffusionLight is extremely time-consuming (*i.e.*, over one day for a single video), so we report comparisons on a

Table 1: Quantitative Comparison. We highlight the best score in boldface. †Comparison on a subset of 10 videos.

	Editable Indoor				EnvMapNet		Efficiency
Method	MSE ↓	SI-MSE ↓	AER ↓	LS ↓	AED ↓	AS ↓	Seconds ↓
DiffusionLight† [29]	0.10	0.05	**4.58**	**0.03**	**40.07**	16.76	145,800
Ours†	**0.09**	**0.03**	4.68	**0.03**	42.53	**15.96**	**80**
DiffusionLightTurbo [11]	0.11	0.05	4.90	0.04	**37.74**	16.68	713
StyleLight [37]	0.13	0.07	6.01	0.05	44.22	17.83	2,002
Ours	**0.10**	**0.03**	**4.77**	**0.03**	41.63	**16.30**	**80**
LDR Baseline	0.11	0.04	5.45	0.03	44.71	15.87	80
Frozen backbone	3.00	0.07	12.95	0.04	51.71	19.45	80
Video-only	0.10	0.03	4.82	0.02	41.96	15.81	80

subset of 10 videos. The quantitative results are provided in Tab. 1. Compared to DiffusionLight, our model achieves competitive lighting accuracy and superior temporal stability while requiring only a fraction of the computation time. Compared to DiffusionLight-Turbo and StyleLight, our model attains the best overall performance.

5.3 Qualitative Evaluation

We provide a qualitative comparison with state-of-the-art image relighting methods in Fig. 4. The baselines including DiffusionLight [29], DiffusionLight-Turbo [11], and StyleLight [37], are designed for single-image manipulation. To adapt them for video evaluation, we apply each method independently to each frame of source video sequences. As illustrated, this per-frame approach leads to significant temporal artifacts. The baseline methods exhibit noticeable flickering, where the intensity and color of highlights and shadows change erratically between consecutive frames. This is an expected outcome, as these models lack any inherent mechanism to enforce temporal consistency.

Generalization to in-the-wild videos. As shown in Fig. 5, V-LITE generalizes to diverse in-the-wild videos under both day and night conditions, producing spatially accurate and temporally coherent lighting estimates even in complex real-world footage.

5.4 Ablation Study

To validate our key design choices, we conduct a set of ablation experiments that isolate the contributions of the HDR-aware VAE, the LoRA-based fine-tuning procedure, and our mixed-data training strategy. We evaluate three ablated variants against our full model and report the quantitative results in Tab. 1.

LDR baseline. We first evaluate the original pretrained model of Jiang *et al.* [22] without both our HDR VAE and the fine-tuning stage. Although this LDR baseline can inpaint chrome balls in videos, it cannot generate HDR content. To enable a comparison with HDR-capable models, we apply the inverse

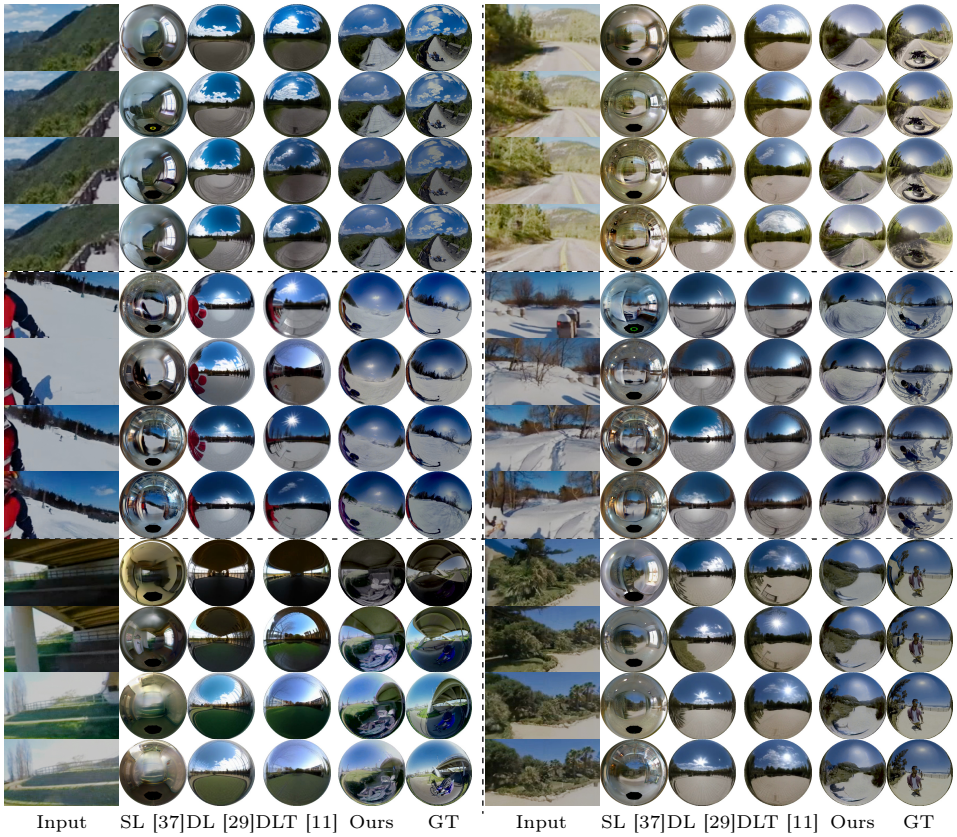

Fig. 4: Qualitative comparison with baseline lighting estimation models. We compare our method against three methods: StyleLight [37] (SL), DiffusionLight [29] (DL), and DiffusionLightTurbo [11] (DLT). For each block, the first row shows input video frames, followed by the predicted lighting conditions from each method, and ground truth (GT) in the final row. Our method produces lighting estimates that closely match the ground truth in terms of intensity and direction, while maintaining temporal consistency.

tone-mapping derived from the HDR ground truth to lift its LDR predictions into the HDR domain. The results support our claim that while video diffusion models inherently encode lighting information, they lack the representational precision to express it in the HDR space.

Frozen backbone. We incorporate our pretrained HDR VAE directly into the baseline model without any task-specific fine-tuning. This configuration introduces a domain mismatch between the original LDR training distribution and the new HDR latent space, significantly degrading the generative performance.

Video-only. To evaluate the impact of our mixed-data training strategy, we fine-tune the model equipped with the HDR VAE exclusively on the video dataset. While this variant aligns the generative priors to the video inpainting task, rely-

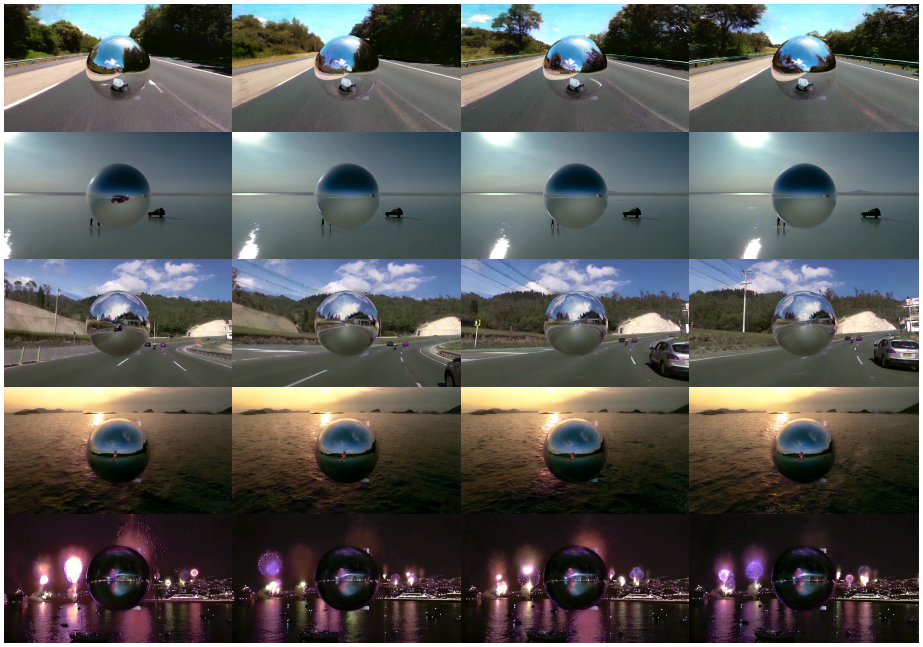

Fig. 5: Qualitative results on in-the-wild videos. For visualization purposes, all frames are tone-mapped from their original HDR format.

ing solely on tonemapped HDR video data limits the model's ability to predict accurate physical HDR values.

Instead, our full model incorporates both the HDR VAE and the LoRA-based fine-tuning, utilizing a mixed-data training strategy. It outperforms all ablation variants, demonstrating that all proposed components are indispensable.

5.5 Virtual Object Insertion

We demonstrate the effectiveness of our estimated HDR environment maps by inserting virtual objects into real-world videos. We first process the input LDR video to solve for the 3D camera motion and establish a ground plane using built-in motion tracker from Blender [3]. The dynamic HDR environment map sequence generated by V-LITE is then used as the exclusive lighting source for a virtual object. This object is rendered from the solved camera's perspective and composited back onto the original video. The results in Fig. 6 show that the virtual objects are seamlessly integrated, with lighting and shadows that are both physically plausible and temporally consistent.

To rigorously validate perceptual realism and lighting consistency, we conducted a user study with 10 participants across 20 diverse videos. For perceptual realism, 64% of responses rated our insertions as "Perfect", and 28% as "Accept-

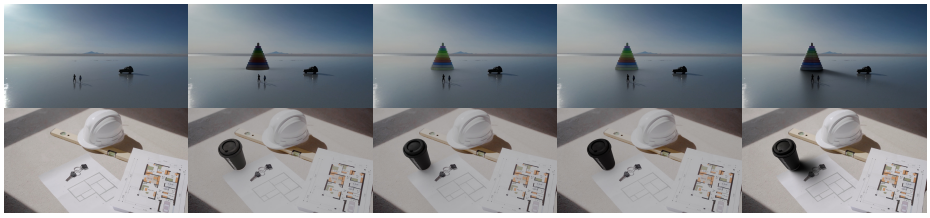

Fig. 6: Qualitative comparison of virtual object insertion. We compare our method against StyleLight [37], DiffusionLight [29], and DiffusionLightTurbo [11]. To ensure a fair comparison, we use the dynamic HDR environment map sequence estimated by each method to illuminate a virtual object within a fixed 3D scene. The first column displays the input video frame, followed by the rendered results from each competing method and our own. Our estimated lighting produces a more physically plausible result, enabling seamless integration of the virtual object. Note the superior quality of the specular highlights and the accurate rendering of both hard and soft cast shadows.

able". For lighting consistency, 66.5% rated the results as "Perfect", with less than 2% of total responses considered failures.

5.6 Discussion

Tonemapped HDR videos and mixed-data training. Capturing in-the-wild HDR videos requires specialized multi-exposure cameras, which fundamentally limits dataset scale and scene diversity. To learn complex dynamic lighting, our V-LITESet utilizes computationally tonemapped HDR videos. While these are not photometrically perfect HDRs, our rigorous VLM-based filtering explicitly discards 57% of heavily clipped data to ensure reasonable intensity ratios. By mixing these dynamic tonemapped videos with static real-world HDR images, V-LITE effectively integrates absolute photometric priors to overcome the physical scarcity of real-world HDR video data.

Ill-posed estimation and visual trade-offs. Estimating an HDR environment from a single LDR observation is an inherently ill-posed problem, as the original tone-mapping process is non-invertible. Because V-LITE leverages the generative priors of diffusion models to naturally resolve this ambiguity rather than performing a strict pixel-to-pixel translation, the output may exhibit slight background visual shifts (*e.g.*, lower contrast on certain bright surfaces). These are expected physical trade-offs for accurately extracting HDR reflections and dominant light directions. As evidenced by our high-fidelity virtual object insertions, V-LITE robustly estimates practical dynamic lighting.

Zero-shot generalization and spatial variance. We further examine the influence of the probe's position and size, observing that the inpainted reflections dynamically respond to its spatial positioning. As depicted in Fig. 7, rows 1–3 illustrate off-center placements from top-left to top-right, while row 4 explores different mask coverage ratios (30%, 50%, and 70%). All preserve structural

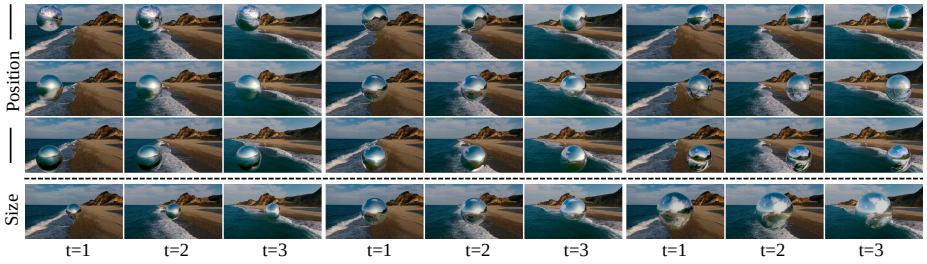

Fig. 7: Visualization of zero-shot generalization by varying probe position and size.

Fig. 8: Visualization of a failure case. When encountering extreme, out-of-distribution inputs, the model may occasionally produce mismatched environment maps.

integrity and realistic illumination, implying that V-LITE inherently captures underlying scene geometry.

6 Conclusion

We present V-LITE, a unified HDR-aware framework that produces dynamic HDR environment maps from in-the-wild videos. We also introduce V-LITESet, a hybrid dataset of tonemapped HDR videos and real-world HDR images curated to advance video-based lighting estimation. Our formulation natively treats lighting recovery as a latent video-inpainting problem, letting the model fully exploit the generative priors of video diffusion models. Experiments show that V-LITE delivers high-fidelity, temporally coherent environment maps in challenging dynamic scenes.

Limitations and failure cases. While V-LITE demonstrates strong generalization, its data-driven nature makes it sensitive to out-of-distribution inputs. As illustrated in Fig. 8, videos with extreme illumination conditions can occasionally lead to mismatched environment maps. Additionally, since our underlying video backbones [36, 44] struggle with long-duration generation, V-LITE's ability to estimate temporally stable lighting for extended sequences remains constrained.

Acknowledgment

This work is supported by National Natural Science Foundation of China (Grant No. 62136001) and Beijing Major Science and Technology Project (Grant No. Z251100008125009). PKU-affiliated authors thank openbayes.com for providing computing resources.

References

1. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-VL technical report. arXiv preprint arXiv:2502.13923 (2025)
2. Bain, M., Nagrani, A., Varol, G., Zisserman, A.: Frozen in time: A joint video and image encoder for end-to-end retrieval. In: International Conference on Computer Vision (2021)
3. Blender Online Community: Blender. `https://www.blender.org`, accessed: 2026-03-01
4. Bolduc, C., Hold-Geoffroy, Y., Shu, Z., Lalonde, J.F.: GaSLight: Gaussian splats for spatially-varying lighting in hdr. In: International Conference on Computer Vision (2025)
5. Cai, Z., Jiang, K., Chen, S.Y., Lai, Y.K., Fu, H., Shi, B., Gao, L.: Real-time 3D-aware portrait video relighting. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
6. Cai, Z., Weng, S., Xia, Y., Shi, B.: PhyS-EdiT: Physics-aware semantic image editing with text description. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
7. Cai, Z., Yang, T., Chang, Z., Li, S., Jiang, H., Weng, S., Shi, B.: Lighting-grounded video generation with renderer-based agent reasoning. arXiv preprint arXiv:2604.07966 (2026)
8. Chang, Z., Weng, S., Zhang, P., Li, Y., Li, S., Shi, B.: L-CAD: Language-based colorization with any-level descriptions using diffusion priors. In: Advances in Neural Information Processing Systems (2023)
9. Chen, K., Ramanan, D., Khurana, T.: Using diffusion priors for video amodal segmentation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2025)
10. Chen, S., Sun, P., Song, Y., Luo, P.: DiffusionDet: Diffusion model for object detection. In: International Conference on Computer Vision (2023)
11. Chinchuthakun, W., Phongthawee, P., Raj, A., Jampani, V., Khungurn, P., Suwajanakorn, S.: DiffusionLight-Turbo: Accelerated light probes for free via single-pass chrome ball inpainting. IEEE Transactions on Pattern Analysis and Machine Intelligence (2026)
12. Debevec, P.: Rendering synthetic objects into real scenes: Bridging traditional and image-based graphics with global illumination and high dynamic range photography. In: ACM SIGGRAPH Conference Papers (2008)
13. Feng, R., Gao, Y., Tse, T.H.E., Ma, X., Chang, H.J.: DiffPose: Spatiotemporal diffusion model for video-based human pose estimation. In: International Conference on Computer Vision (2023)
14. Gao, Y., Guo, H., Hoang, T., Huang, W., Jiang, L., Kong, F., Li, H., Li, J., Li, L., Li, X., et al.: Seedance 1.0: Exploring the boundaries of video generation models. arXiv preprint arXiv:2506.09113 (2025)
15. Gardner, M.A., Sunkavalli, K., Yumer, E., Shen, X., Gambaretto, E., Gagné, C., Lalonde, J.F.: Learning to predict indoor illumination from a single image. ACM Transactions on Graphics (2017)
16. Garon, M., Sunkavalli, K., Hadap, S., Carr, N., Lalonde, J.F.: Fast spatially-varying indoor lighting estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

17. Haven, P.: Poly Haven: The public 3d asset library. `https://polyhaven.com/`, accessed: 2026-03-01
18. Hold-Geoffroy, Y., Sunkavalli, K., Hadap, S., Gambaretto, E., Lalonde, J.F.: Deep outdoor illumination estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2017)
19. Hong, W., Ding, M., Zheng, W., Liu, X., Tang, J.: CogVideo: Large-scale pretraining for text-to-video generation via transformers. In: International Conference on Learning Representations (2023)
20. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Chen, W.: LoRA: Low-rank adaptation of large language models. In: International Conference on Learning Representations (2022)
21. Huang, T., Yan, Z., Zhao, Y., Lee, G.H.: ComPC: Completing a 3D point cloud with 2D diffusion priors. In: International Conference on Learning Representations (2025)
22. Jiang, Z., Han, Z., Mao, C., Zhang, J., Pan, Y., Liu, Y.: VACE: All-in-one video creation and editing. In: International Conference on Computer Vision (2025)
23. Lan, Y., Cui, Z., Liu, C., Peng, J., Wang, N., Luo, X., Liu, D.: Exploiting diffusion prior for real-world image dehazing with unpaired training. In: Association for the Advancement of Artificial Intelligence (2025)
24. Li, Z., Yu, L., Okunev, M., Chandraker, M., Dong, Z.: Spatiotemporally consistent HDR indoor lighting estimation. ACM Transactions on Graphics (2023)
25. Liang, R., Gojcic, Z., Nimier-David, M., Acuna, D., Vijaykumar, N., Fidler, S., Wang, Z.: Photorealistic object insertion with diffusion-guided inverse rendering. In: European Conference on Computer Vision (2024)
26. Liang, R., He, K., Gojcic, Z., Gilitschenski, I., Fidler, S., Vijaykumar, N., Wang, Z.: LuxDiT: Lighting estimation with video diffusion transformer. In: Advances in Neural Information Processing Systems (2025)
27. Lipman, Y., Chen, R.T., Ben-Hamu, H., Nickel, M., Le, M.: Flow matching for generative modeling. In: International Conference on Learning Representations (2023)
28. Murmann, L., Gharbi, M., Aittala, M., Durand, F.: A dataset of multi-illumination images in the wild. In: International Conference on Computer Vision (2019)
29. Phongthawee, P., Chinchuthakun, W., Sinsunthithet, N., Raj, A., Jampani, V., Khungurn, P., Suwajanakorn, S.: DiffusionLight: Light probes for free by painting a chrome ball. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
30. Rao, P., Fox, G., Meka, A., BR, M., Zhan, F., Weyrich, T., Bickel, B., Pfister, H., Matusik, W., Elgharib, M., et al.: Lite2Relight: 3D-aware single image portrait relighting. In: ACM SIGGRAPH Conference Papers (2024)
31. Sengupta, S., Gu, J., Kim, K., Liu, G., Jacobs, D.W., Kautz, J.: Neural inverse rendering of an indoor scene from a single image. In: International Conference on Computer Vision (2019)
32. Song, S., Funkhouser, T.: Neural illumination: Lighting prediction for indoor environments. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)
33. Srinivasan, P.P., Mildenhall, B., Tancik, M., Barron, J.T., Tucker, R., Snavely, N.: Lighthouse: Predicting lighting volumes for spatially-coherent illumination. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2020)
34. Tang, J., Zhong, H., Weng, S., Shi, B.: LuminAIRe: Illumination-aware conditional image repainting for lighting-realistic generation. In: Advances in Neural Information Processing Systems (2023)

35. Tong, M., Wu, R., Zheng, C.: Spatiotemporally consistent indoor lighting estimation with diffusion priors. arXiv preprint arXiv:2508.08384 (2025)
36. Wan, T., Wang, A., Ai, B., Wen, B., Mao, C., Xie, C.W., Chen, D., Yu, F., Zhao, H., Yang, J., et al.: Wan: Open and advanced large-scale video generative models. arXiv preprint arXiv:2503.20314 (2025)
37. Wang, G., Yang, Y., Loy, C.C., Liu, Z.: StyleLight: HDR panorama generation for lighting estimation and editing. In: European Conference on Computer Vision (2022)
38. Wang, Q., Zhao, Y., Ma, J., Li, J.: How to use diffusion priors under sparse views? In: Advances in Neural Information Processing Systems (2024)
39. Wang, X., Xiao, S., Liang, X.: LightOctree: Lightweight 3D spatially-coherent indoor lighting estimation. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
40. Wang, Y., He, Y., Li, Y., Li, K., Yu, J., Ma, X., Li, X., Chen, G., Chen, X., Wang, Y., et al.: InternVid: A large-scale video-text dataset for multimodal understanding and generation. In: International Conference on Learning Representations (2024)
41. Wiedemer, T., Li, Y., Vicol, P., Gu, S.S., Matarese, N., Swersky, K., Kim, B., Jaini, P., Geirhos, R.: Video models are zero-shot learners and reasoners. arXiv preprint arXiv:2509.20328 (2025)
42. Xia, Y., Weng, S., Yang, S., Liu, J., Zhu, C., Teng, M., Jia, Z., Jiang, H., Shi, B.: PanoWan: Lifting diffusion video generation models to 360° with latitude/longitude-aware mechanisms. In: Advances in Neural Information Processing Systems (2025)
43. Yang, T., Wu, R., Ren, P., Xie, X., Zhang, L.: Pixel-aware stable diffusion for realistic image super-resolution and personalized stylization. In: European Conference on Computer Vision (2024)
44. Yang, Z., Teng, J., Zheng, W., Ding, M., Huang, S., Xu, J., Yang, Y., Hong, W., Zhang, X., Feng, G., et al.: CogVideoX: Text-to-video diffusion models with an expert transformer. In: International Conference on Learning Representations (2025)
45. Zhan, F., Zhang, C., Yu, Y., Chang, Y., Lu, S., Ma, F., Xie, X.: EMLight: Lighting estimation via spherical distribution approximation. In: Association for the Advancement of Artificial Intelligence (2021)