LumiSculpt: Enabling Consistent Portrait Lighting in Video Generation

Yuxin Zhang[1*], Dandan Zheng[2], Biao Gong[2†], Shiwen Wang[1],
Jingdong Chen[2], Ming Yang[2], Weiming Dong[1✉], Changsheng Xu[1]
[1]MAIS, Institute of Automation, Chinese Academy of Sciecnes [2]Ant Group
{zhangyuxin2020, weiming.dong, changsheng.xu}@ia.ac.cn
{yuandan.zdd,gongbiao.gb,jingdongchen.cjd,m.yang}@antgroup.com

Figure 1: *LumiSculpt* allows user-specified control over the intensity, position, and trajectories of an assumed light source, with textual conditions as input. Being trained once, *LumiSculpt* is capable of generating diverse results at inference time.

Abstract

Lighting plays a pivotal role in ensuring the naturalness and aesthetic quality of video generation. However, the impact of lighting is deeply coupled with other factors of videos, *e.g.*, objects and scenes. Thus, it remains challenging to disentangle and model coherent lighting conditions independently, limiting the flexibility to control lighting in video generation. In this paper, inspired by the established controllable T2I models, we propose *LumiSculpt* , which enables precise and consistent lighting control in T2V generation models. *LumiSculpt* equips the video generation with new interactive capabilities, allowing the input of reference image sequences with customized lighting conditions. Furthermore, the core learnable plug-and-play module of *LumiSculpt* facilitates direct control over the intensity, position and trajectory of an assumed light source in video diffusion models. To effectively train *LumiSculpt* and address the issue of insufficient lighting data, we construct *LumiHuman* , a new lightweight and flexible dataset for portrait lighting of images and videos. Experimental results demonstrate that *LumiSculpt* achieves precise and high-quality lighting control in video generation. The analysis demonstrates the flexibility of *LumiHuman* .

CCS Concepts

• **Computing methodologies** → **Computer vision**.

Keywords

Light Control, Video generation, Text-to-video generation, Diffusion models

1 Introduction

If a video tells a story, then lighting is the voice that shapes its tone and mood. Lighting is essential for video generation, which is one of the defining factors for the overall aesthetic quality of the generated video, and is also used to convey emotions, highlight character traits, and guide the audience's attention. Mainstream video generation methods currently employ latent diffusion models (LDMs) to achieve video generation through multi-step denoising in a latent space. Research on controllable image and video generation based on LDMs supports our studies of consistent lighting control. Several methods [4, 8, 11, 12, 31] have been developed to achieve relatively accurate text-controlled video generation, as well as video editing [2, 3], customization [36], and controlling [9, 32, 35]. These works have improved the controllability, aesthetics, and usability of video generation. However, due to the deep coupling between lighting and the other non-stationary factors of videos, it is challenging to model coherent lighting conditions independently, resulting in a lack of handy approaches to controlling lighting in videos.

The challenge of customizing lighting lies in three aspects: the lack of training data, the representation of lighting, and the mechanism of injecting lighting conditions without influencing other attributes. Specifically, although there are relighting datasets based on light stages [5], the data format of light stages is not readily applicable in video generation scenarios. Therefore, a new dataset that is adaptable to text-controlled content generation is needed. Obtaining the projection of lighting on the camera's imaging plane requires knowledge of the lighting and the surface texture of the

∗ Work done during internship at Ant Group. † Project lead.
✉ Corresponding author.

illuminated objects [17, 23, 29], which cannot be achieved in an end-to-end video generation scenario. Thus, a simple lighting representation that is only related to lighting parameters is important. Finally, similar to most control tasks, lighting control faces the problem of the deep decoupling of lighting from other factors, such as semantics and color.

In this paper, we propose *LumiHuman* to address the problem of limited training data. *LumiHuman* is a portrait lighting dataset that can constitut more than 220K videos of humans with known lighting parameters. This is a lightweight and flexible dataset that is not limited to specific lighting movements but is presented in freely combinable frames, laying the foundation for a more diverse range of lighting paths and combinations. We then use virtual engine rendering with known lighting parameters to obtain projections of different directional lighting on planes as a lighting representation. To achieve video lighting control, we propose a lighting control method, *LumiSculpt* , which learns an accurate plug-and-play lighting module capable of controlling the direction and movement of lighting in video generation. To solve the problem of lighting feature injection, we introduce a light control module that takes the lighting projection as input and integrates lighting control injected into the generative model layer by layer. Furthermore, to better decouple lighting from appearance, we design a decoupling loss based on a dual-branch structure, preserving diverse generative capabilities.

We implemented *LumiSculpt* on Open-Sora-Plan [19] to enable precise lighting control. *LumiSculpt* focuses specifically on lighting control in text-to-video generation, rather than addressing image-to-image or video-to-video relighting tasks. It goes beyond managing fixed types and directions of lighting based on text prompts. Instead, it enables precise control over the position and moving trajectory of the light source within a single video, offering great flexibility and customization. We conducted comprehensive quantitative and qualitative evaluations. The experimental results show that *LumiSculpt* has achieved state-of-the-art performance in the control of text-to-video lighting in precision and diversity, as shown in Figure 1. In summary, our main contributions are as follows:

- We introduce a portrait lighting dataset *LumiHuman* . *LumiHuman* is a lighting video dataset comprising over 220K different videos (*i.e.,* 2.3M images). *LumiHuman* includes over 30K lighting positions, and over 3K light source trajectories for each individual. *LumiHuman* paves the way for lighting control in both image and video generation.
- We introduce *LumiSculpt* , which enables control of the position and moving trajectories of light sources in video generation. We propose a lighting representation method, a lighting injection approach, and a lighting decoupling loss for the text-to-video generation scenario, which enables diverse content generation with limited data.
- Extensive experiments prove that *LumiSculpt* has achieved state-of-the-art performance. *LumiSculpt* can enable consistent portrait lighting in diverse scenarios. The analysis demonstrate the comprehensiveness and flexibility of *LumiHuman* , which provides a generalizable lighting prior.

2 Related Works

2.1 Relighting

In recent years, deep learning techniques have made significant progress in portrait relighting [17, 23, 24, 26, 27, 30, 34, 38, 42], which often rely on paired data captured by light stage systems [5] for supervised learning. Typically, these methods require the use of high dynamic range (HDR) environmental maps as input. This process involves estimating intermediate surface properties, including normal vectors, albedo, diffuse reflectance, and specular reflection characteristics. However, the reliance on HDR environmental maps limits the practical application of these techniques in video generation scenarios. Besides, researchers also explore portrait relighting techniques that do not depend on light stage data [13, 14, 33].

Recently, diffusion-based models have shed light on new approaches to relighting. Ren et al. [29] propose a three-stage lighting-aware diffusion model called Relightful Harmonization, which aims to provide complex lighting coordination for foreground portraits with any background image. Zeng et al. [39] propose a three-stage portrait relighting method using a fine diffusion model called DiLightNet, which calculates radiance cues to re-synthesize and refine the foreground object by combining the rough shape of the foreground object inferred from the preliminary image. Xing et al. [37] propose a natural image relighting method called Retinex-Diffusion, which treats the diffusion model as a black-box image renderer and decomposes its energy function to be consistent with the image formation model. However, there still lack of methods for lighting control in text-to-video generation. The most related works are diffusion-based relighting methods. LightIt [18] is an image-guided method for image relighting conditioning on shading estimation and normal maps. IC-Light [41] is an image relighting method to generate harmonized background with the user input foreground. Light-A-Video [46] proposes a training-free video relighting method, which leverages on the image relighting model to achieve video relighting controlled by text. It is worth noting that the above methods are image-to-image or video-to-video methods that can leverage the input prior, while our method aims at lighting control in text-to-video generation, which is more challenging. Moreover, our method is able to control the exact position and trajectory of the light source within a single video, instead of setting a fixed light type using text prompts.

2.2 Text-to-video synthesis and controlling

Recently, several researches, such as [4, 8, 11, 12, 31], have adopted diffusion models to create highly realistic video content, utilizing text as conditions in guiding the generation process. These studies focus on ensuring alignment between textual descriptions and the final video output. Addressing the issue of difficulty in precisely describing specific visual attributes through text conditions, some studies have attempted to achieve finer video control by fine-tuning models or introducing additional control parameters. Tune-A-Video [36] propose a fine-tuning framework that allows users to customize specific videos. VideoComposer [32] use explicit control signals to guide the temporal dynamics of the video. Gong et al. [7] introduce TaleCrafter to handle interactions among multiple characters, featuring layout and structural editing capabilities. He et al. [10] propose a retrieval-based deep guidance method

Figure 2: Diverse indivisuals in *LumiHuman* .

Table 1: Comparison of Openillumination [21],DPR [45] and *LumiHuman*

Dataset	Synthesis	Light Positions	Light Movement	Number of Images	Subject	Resolutions
DPR	2D	7	None	138K	-	1024×1024
Openillumination	Light Stage	142	None	108K	64 objects	3000×4096
LumiHuman	3D	**35,937**	**>3K**	**2.3M**	**65 individuals**	1024×1024

that can integrate existing video clips into a coherent narrative video by customizing the appearance of characters. These studies mainly focus on the appearance of objects and scenes. Several methods [9, 35, 43, 44] learn and control motion through customized diffusion models. These attempts have made substantial progress in controlling video in specific aspects. Nevertheless, further dedicated efforts are required on precise control of lighting in videos.

3 LumiHuman

Our goal is to achieve unified control over video lighting—a challenging task with significant implications. The primary difficulties can be categorized into three areas: *(1) Dataset Scarcity*: There is a notable lack of lighting-specific datasets, particularly for videos. Few annotated examples explicitly capture lighting variations, and even fewer provide well-defined lighting information. *(2) Complexity of Lighting Attributes*: Lighting involves multiple factors, including light source type, direction of illumination, and the material properties of objects. Accurately representing lighting effects within the camera's field of view becomes especially important. *(3) Attribute Decoupling*: Similar to other control tasks, lighting control requires effective decoupling of specific attributes. A major technical challenge lies in isolating lighting information from object appearance in training data, preventing the model from overfitting.

We introduce a portrait lighting dataset, referred to as *LumiHuman* . *LumiHuman* is a continuous lighting video dataset comprising over *220K* different videos (*i.e., 2.3 million* images). The resolution of each video is 1024 × 1024. *LumiHuman* is created using Unreal Engine [6] for lighting simulation, allowing for the production of data with known lighting information. As shown in Figure 2, *LumiHuman* includes 65 diverse human subjects, 30K lighting positions, and over 3K lighting trajectories for each people. As shown in Table. 1, compared to other lighting datasets Openillumination [21] and Deep Portrait Relighting (DPR) dataset [45] (generated from face image dataset Celeb-A [22]), LumiHuman outperforms in light positions, light movements and number of images.

Details. As shown in Figure 3(a), *LumiHuman* can be combined to generate various types of character lighting. Its 30K lighting positions enable the creation of light and shadow effects across *all areas* of the human face. In Figure 3(b), we present the brightness distribution map for different facial regions. Each ridge in the ridge plot represents a specific facial area, where the horizontal axis indicates brightness and the vertical axis denotes the number of samples corresponding to each brightness level. *LumiHuman* comprehensively covers all facial areas and distributes samples across a wide range of brightness levels. As illustrated in Figure 3(c), we present continuous video frames composed of samples from *LumiHuman* . These samples can be flexibly arranged to form diverse lighting trajectories based on user specifications—such as horizontal, vertical, diagonal, arc-shaped, or multi-light-source superpositions.

Lighting Representation. A straightforward approach to representing lighting effects is to embed lighting parameters as additional input to the model. However, this method requires a large amount of annotated data to establish a reliable mapping between lighting vectors and the two-dimensional image plane. To better align with the model's latent space, we propose projecting lighting information into a blank canvas, as illustrated in Figure 5(b). For each lighting position, this is visualized as an image in which brighter regions indicate stronger illumination, and darker regions indicate weaker lighting. This representation enables more effective integration of lighting information into the video generation model.

Data Collection. The *LumiHuman* collection consists of five key stages: *(1) Lighting Design*: As illustrated in Figure 4(a), we constructed a lighting position matrix—a three-dimensional grid measuring 160 cm × 160 cm × 160 cm. Points within the grid are uniformly spaced at 5 cm intervals and serve as lighting positions. A point light source moves across these grid points to illuminate the subject from diverse angles. *(2) Lighting Trajectory Design*: Within this 3D grid, we defined horizontal, vertical, and diagonal trajectories composed of grid points to simulate a variety of lighting change effects. *(3) Character Construction*: To generate high-quality portrait lighting data, we employed the MetaHuman dataset [25],

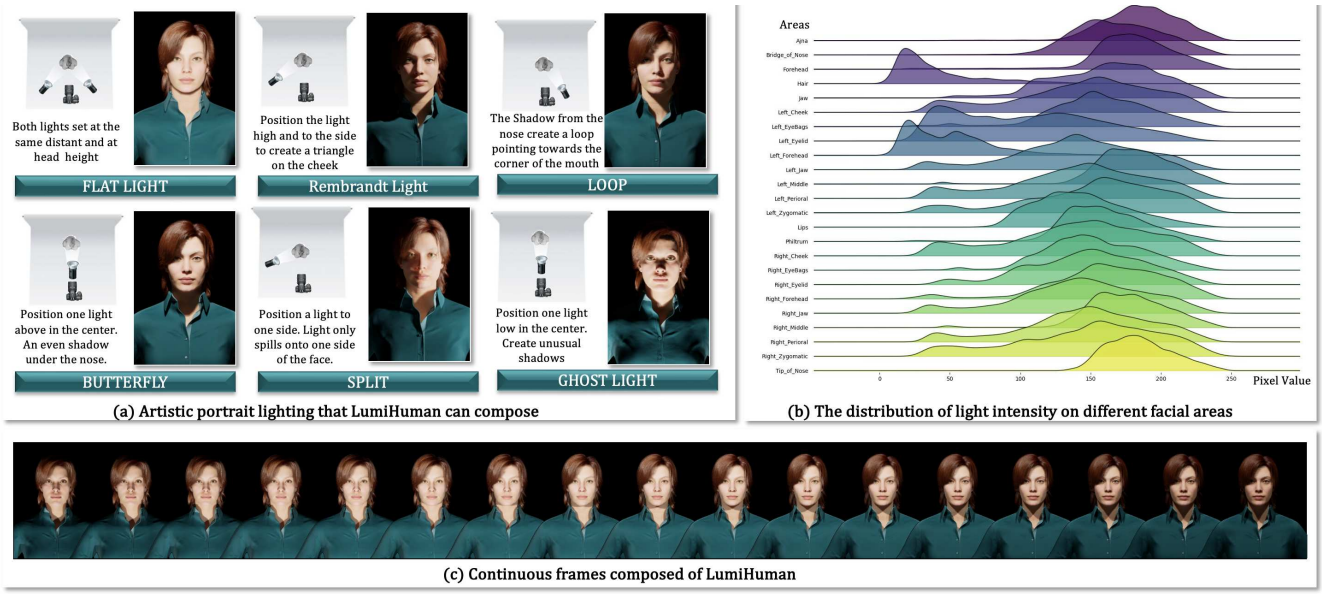

Figure 3: (a) *LumiHuman* offers a variety of basic elements that can be combined to form various types of portrait lighting, widely applicable to a range of tasks related to character lighting. (b) shows the distribution of light intensity on different facial areas of the characters; *LumiHuman* 's lighting matrix can cover all areas of the face and produce a significant range of light and shadow variations. (c) shows an example of creating a continuous lighting video using *LumiHuman* .

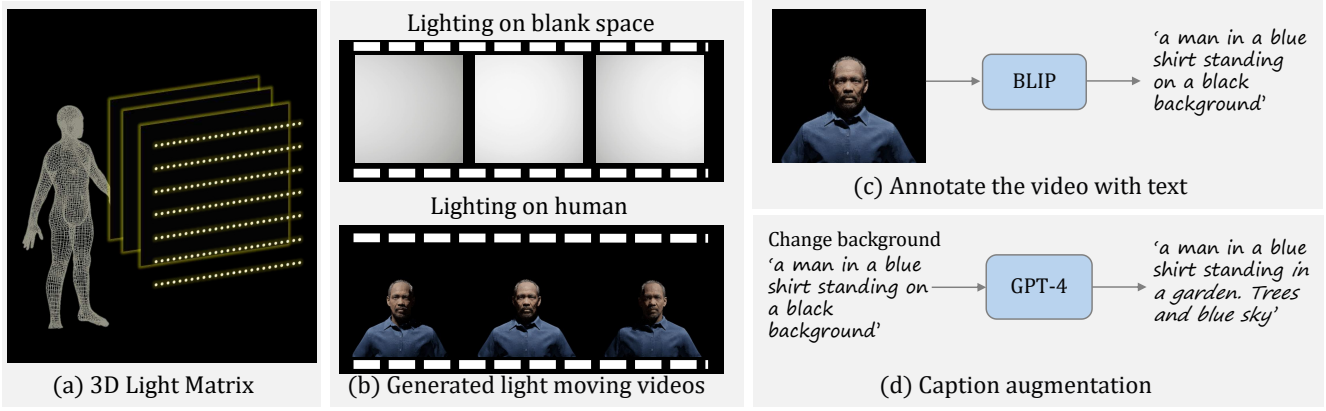

Figure 4: The collection process of *LumiHuman* includes: (a) designing a 3D point light source matrix of $33 \times 33 \times 33$ lighting points, (b) rendering single-frame images and generating portrait lighting videos with various path lighting and lighting reference videos, (c) annotating with a BLIP model, and (d) producing enhanced background captions using a large language model.

which provides 3D models of 65 diverse individuals. This variation enhances the richness of visual lighting effects across different characters. *(4) Flexibility and Storage Optimization*: To support a wide range of lighting path variations while managing storage demands—particularly due to duplicate frames from identical lighting positions—we provide a flexible, lightweight image-video dataset. The dataset includes images rendered from different lighting positions within the 3D grid, with each character associated with $33 \times 33 \times 33$ unique samples. Videos can be generated using predefined lighting trajectories, or users can define new paths to simulate additional lighting effects, as demonstrated in Figure 4(b). *(5) Text Annotation and Augmentation*: For automated video captioning, we utilized BLIP [20]. To enrich the contextual diversity of the dataset, we further applied GPT-4 [1] for caption augmentation, generating a wide range of descriptive narratives, as illustrated in Figures 4(c) and (d).

4 LumiSculpt

4.1 Integrating Lighting into Video Generators

LumiHuman enables lighting to be represented in pixel space and parameterized as an input to standard visual models. We extract lighting features using a Variational Autoencoder (VAE), shared with the generative model, and feed them into a dedicated lighting

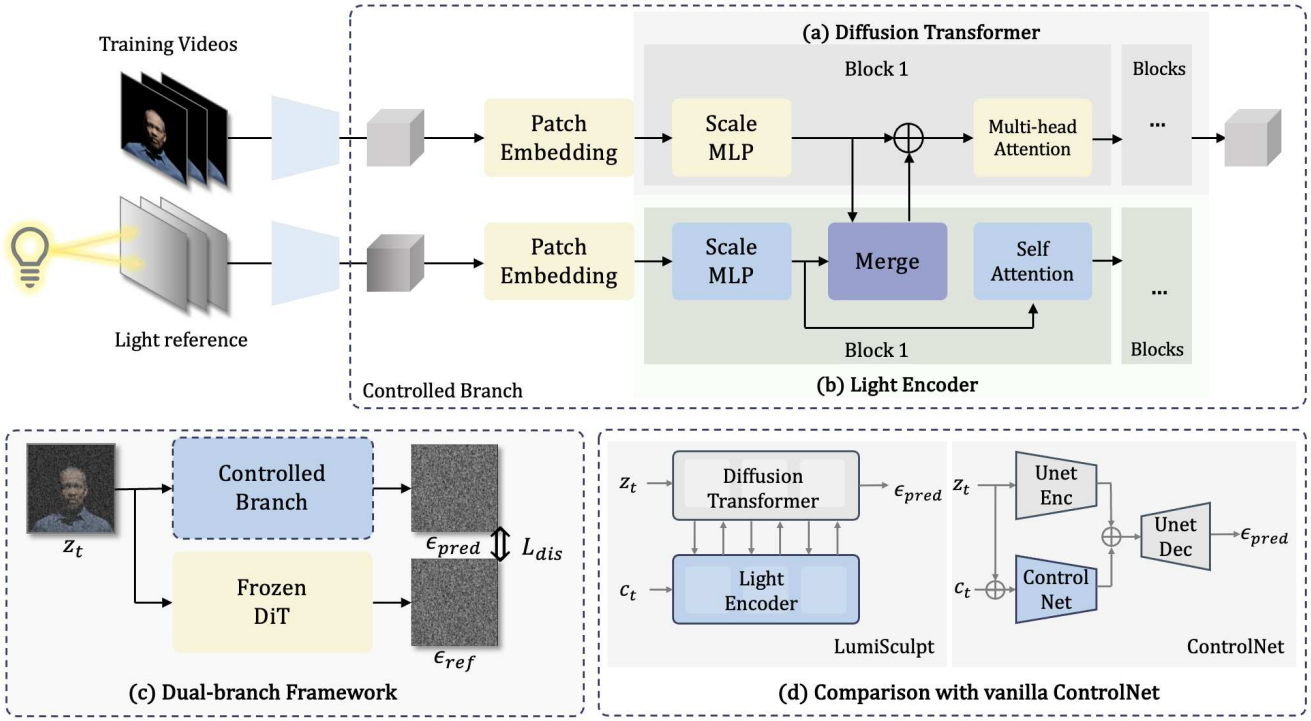

Figure 5: The pipeline of *LumiSculpt* consists of the generation backbone, i.e. the controlled branch, which includes (a) a diffusion transformer (DiT), a pre-trained video denoising network, and (b) a light encoder, a trainable external transformer network. The light encoder takes light reference latents as input and processes them through various blocks to produce a light condition sequence. This sequence is integrated into the generation backbone using several merge modules within each block. During training, we propose a (c) dual-branch framework including a controlled branch and a frozen branch, which provide regularization for diverse appearances. The frozen branch is a DiT with frozen parameters, sharing weights with (a). Both branches predict noise, resulting in ϵ_{pred} and ϵ_{reg}, which are used to compute the disentanglement loss $\mathcal{L}_{dis}$. (c) and (d) show that *LumiSculpt* differs form ControlNet [40] in terms of model structure, condition injection, training manners and objectives.

encoder. As illustrated in Figure 5(b), the lighting encoder adopts a transformer architecture composed of self-attention layers, enabling it to compute global attention scores across video frames. This design effectively captures the spatial and temporal dynamics of lighting throughout a video clip.

The lighting encoder takes the lighting features as input and passes them through transformer blocks that match the number of layers in the backbone model. These blocks output a sequence of latent representations with dimensions aligned to those of the backbone model, enabling seamless feature fusion. Our objective is to integrate these latents into the DiT architecture of the T2V model. Specifically, the latent video features, denoted z_t, are combined with the lighting features c_t via element-wise addition. The resulting features are then passed through a linear layer to produce the output for the next layer, with a guidance scale hyperparameter set to 0.5.

As shown in Figure 5(d), *LumiSculpt* employs a 3D self-attention mechanism in the lighting encoder and utilizes multi-stage weighting as the conditional injection mechanism. In comparison, ControlNet uses a U-Net encoder to extract features and injects conditions via additive latent fusion.

4.2 Lighting Learning

We adopt a data-driven approach to learn complex lighting effects, specifically capturing the impact of light projection from various positions on the human face. This approach is implemented using *LumiHuman* . However, a key challenge arises when learning lighting directly from data rendered in Unreal Engine: the leakage of appearance information. This issue stems from the dataset's consistent backgrounds and layouts, which, while beneficial for stable training, also increase the risk of the model overfitting to specific appearances.

This highlights a central challenge in lighting control: how can lighting be disentangled from other visual attributes? We address this by proposing a lighting-appearance disentanglement method, which incorporates a dual-branch architecture alongside a novel disentanglement loss function. This framework enables the model to isolate lighting information while mitigating the influence of appearance-related features.

Dual-Branch Framework. A straightforward approach to obtaining diverse appearance data is to use additional video datasets as regular samples, thereby exposing the model to a wider range

of appearances. However, acquiring large-scale, diverse data with known lighting conditions remains challenging. To address this, we generate regular samples using generative models.

As illustrated in Figure 5(c), we propose a dual-branch framework consisting of a training branch and a frozen reference branch. The frozen branch utilizes a pre-trained foundational denoising model to serve as an appearance reference. During training, both branches receive the same textual condition and noisy latent input, producing predicted noise values ϵ_t and ϵ_t^{reg}, respectively. The diverse appearance priors from the pre-trained model are captured in ϵ_t^{reg}, enabling the generation of regular samples in a cost-efficient manner.

Table 2: Quantitative experimental results and ablation study results.The best results are marked as bold and the seconds one are marked by underline.

Method	Consistency		Lighting Accuracy		Quality
	CLIP↑	LPIPS↓	Direction↓	Brightness↑	CLIP↑
Open-Sora	0.9845	1.3503	0.4542	0.8229	0.3182
IC-Light	0.9703	2.5329	0.5264	0.8632	0.3145
ControlNet	0.8081	5.9324	0.5500	0.8032	<u>0.3440</u>
Ours(full model)	<u>0.9951</u>	1.1312	0.3500	<u>0.8779</u>	**0.3597**
Ours(w/o caption aug)	0.9948	<u>1.1211</u>	<u>0.2992</u>	**0.9269**	0.3416
Ours(w/o $\mathcal{L}_{dis}$)	**0.9957**	**1.1033**	**0.1945**	0.8363	0.2909

Loss Functions. We train the model to capture the overall distribution of the dataset using a denoising loss $\mathcal{L}_{\text{denoise}}$, which captures both lighting and appearance information. However, since appearance variation is not required, we propose a disentanglement loss, $\mathcal{L}_{\text{dis}}$, to evaluate and suppress appearance information. $\mathcal{L}_{\text{dis}}$ satisfy two criteria: (1) it should effectively quantify the consistency of appearance between the latent representations of two videos, and (2) it should be invariant to planar spatial information, as dependence on pixel location could hinder the learning of lighting distributions. Inspired by neural style transfer [16], which similarly aims to preserve style while discarding structural information, we adopt Adaptive Instance Normalization (AdaIN) [15] to compare the feature statistics—mean and variance—of the latents from two branches. This effectively captures the distribution of appearance features. By aligning the appearance distribution of the training branch with that of a frozen reference branch, we retain the diverse generative capacity of the original model while filtering out redundant appearance signals. Under this framework, the generative model is optimized to match the training video via $\mathcal{L}_{\text{denoise}}$, while $\mathcal{L}_{\text{dis}}$ enforces appearance disentanglement. Together, they enable precise lighting control while preserving generalization. The overall training objective is defined as:

$$\begin{aligned} \mathcal{L}_{\text{dis}} &= \left\| \sigma(z_0^{\text{pred}}) - \sigma(z_0^{\text{reg}}) \right\|_2 + \left\| \mu(z_0^{\text{pred}}) - \mu(z_0^{\text{reg}}) \right\|_2, \\ \mathcal{L}_{\text{denoise}} &= \mathbb{E}_{z_{1:N},\epsilon,c_t,t} \left[\left\| \hat{\epsilon}(z_{1:N}^{\text{pred}}, c_t, t) - \epsilon \right\|^2 \right], \\ \mathcal{L}_{\text{total}} &= \mathcal{L}_{\text{denoise}} + \beta \mathcal{L}_{\text{dis}}, \end{aligned} \tag{1}$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ denote the standard deviation and mean, respectively. $z_0^{\text{reg}} = z_t^{\text{reg}} - \epsilon_{\text{reg}}$ and $z_0^{\text{pred}} = z_t^{\text{pred}} - \epsilon_{\text{pred}}$ represent the predicted denoised latents of the frozen and training branches at timestep t. N is the total number of denoising steps, c_t is the textual condition, and β is a balancing coefficient set to 3.0.

5 Experiments

5.1 Experimental Setup

Methods for comparison . We compare our approach with state-of-the-art text-to-video generation methods Open-Sora [19], image relighting method IC-light [41], and image control method ControlNet [40].

Metrics . We employ a variety of quantitative and qualitative metrics to assess the lighting accuracy, inter-frame coherence, and visual-text similarity of generated videos.

Evaluation dataset . We use 500 different light paths and captions not present in the training dataset as conditions to guide the comparative methods in generating evaluation videos.

Implementation details. In all video generation experiments, we use Open-Sora v1.2.0 [19] with the default network architecture. We set a learning rate of 1×10^{-4}. The input video resolution is $640 \times 480 \times 29$. The training process for each motion requires approximately $800 \sim 1500$ iterations using eight NVIDIA A100. The number of inference steps is set to $T = 50$ and the guidance scale is set to $w = 7.5$.

5.2 Quantitative Evaluations

Table 2 presents five quantitative metrics used for evaluation: (1) *Frame-wise CLIP image similarity* assesses semantic-level video coherence by measuring the similarity of frame-wise CLIP [28] image embeddings. A higher value indicates stronger inter-frame similarity and better semantic stability in the generated video. (2) *Frame-wise LPIPS consistency* evaluates feature-level coherence by calculating frame-wise LPIPS. A lower value reflects smaller feature discrepancies, indicating higher inter-frame consistency. (3) *Lighting direction RMSE* is computed for each frame and assess the consistency of the generated video's lighting direction with the reference. A smaller RMSE indicates better alignment with the target lighting direction. (4) *Brightness consistency* involves segmenting each video frame into patches and computing the average brightness for each patch to construct a brightness distribution. This distribution is used to measure the consistency between the generated video and the reference, independent of absolute brightness values. (5) *CLIP text-image similarity* measures the alignment between the generated video frames' CLIP image embeddings and the text embedding of the caption. A higher similarity indicates better generation quality. As shown in the results, *LumiSculpt* outperforms Open-Sora-Plan [19] and IC-Light [41] by maintaining strong inter-frame consistency and text-image alignment, while also achieving precise lighting control.

5.3 Qualitative Evaluations

As shown in Figure 6, due to the absence of video illumination control methods, we compare our approach with the image illumination control method IC-Light based on diffusion models, and the video generation method Open-Sora. We consider two light intensity levels, strong and soft, as well as horizontal and vertical lighting movement directions. Since IC-Light is designed for relighting existing images, we use portraits generated by our method

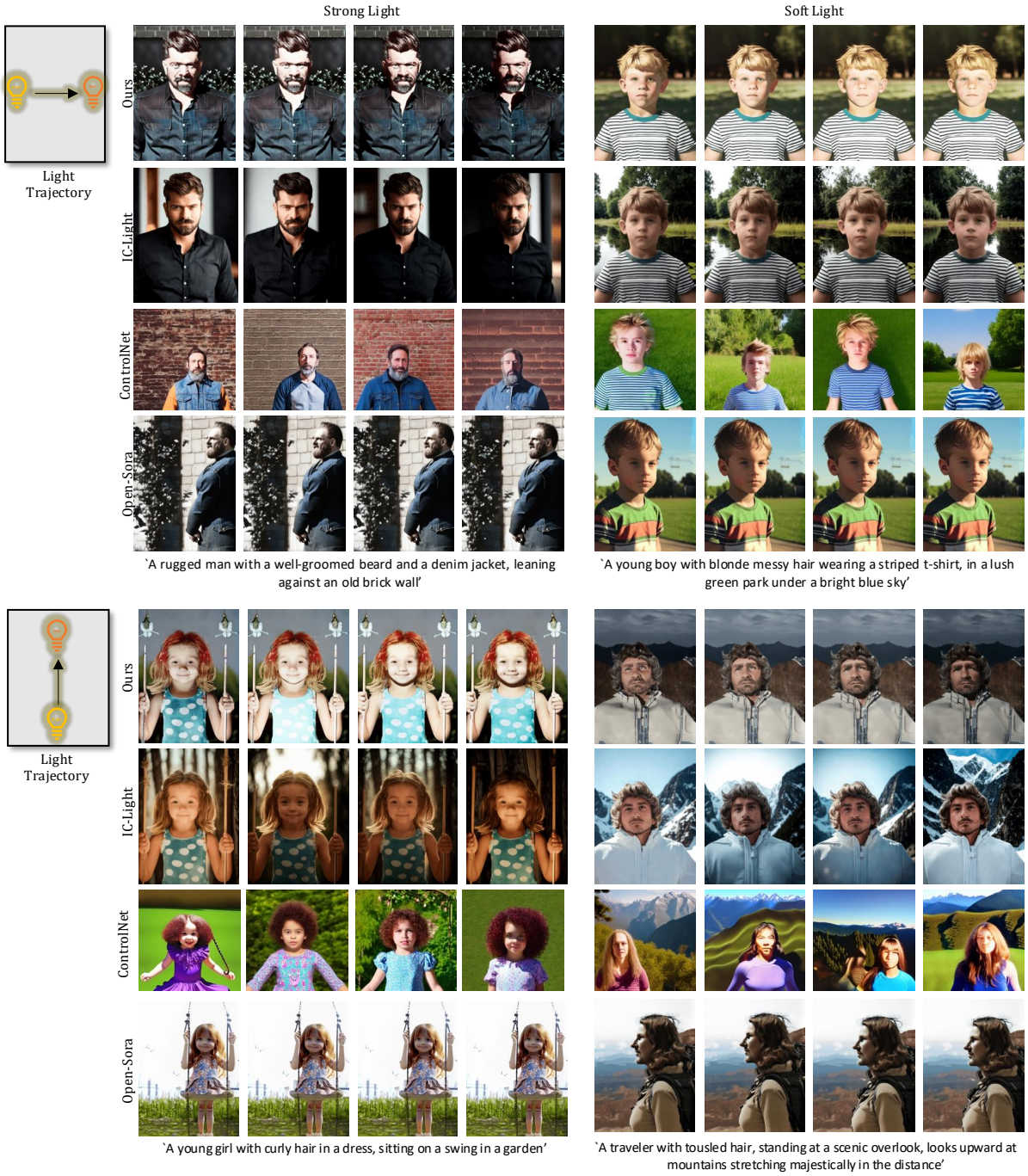

Figure 6: Comparison results with state-of-the-art methods IC-Light [41], ControlNet [40] and Open-Sora [19]. The classic horizontal and vertical directions for light movement and two brightness levels are tested to achieve a comprehensive qualitative evaluation.

as foreground guidance. IC-Light is capable of producing single-frame images with accurate lighting directions, but due to a lack of inter-frame awareness, the coherence of the output video is poor, with noticeable flickering in the background. Open-Sora can generate coherent and aesthetically pleasing videos, but struggles to control lighting direction via textual conditions, resulting in relatively unchanged lighting throughout the video. Our method not only ensures video coherence and visual quality but also achieves precise control over lighting trajectory and intensity. Video results are provided in the supplementary material.

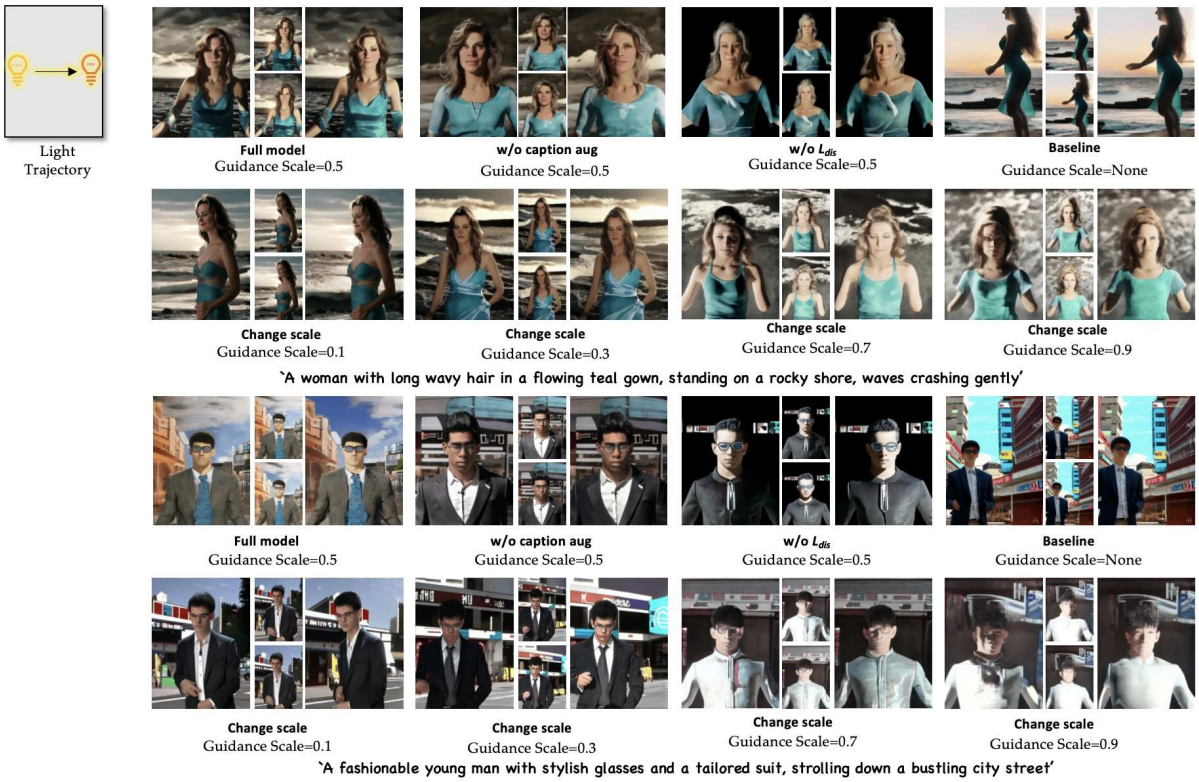

Figure 7: Ablation study results. We conducted ablations on key components related to model training, including caption augmentation and $\mathcal{L}_{dis}$, both of which contribute to the diversity of generated content. We presented results with varying values of the important parameter *guidance scale*, which affects the accuracy of lighting and the diversity of appearance, related to model inference.

Table 3: Experimental results of hyper-parameter *guidance scale*. The best results are marked as bold and the seconds one are marked by underline

Scale	Consistency↑	Accuracy↓	Quality↑
scale=0.1	0.9943	0.4239	0.3163
scale=0.3	**0.9964**	0.3825	**0.3814**
scale=0.5	<u>0.9951</u>	0.3500	<u>0.3597</u>
scale=0.7	0.9939	**0.2484**	0.3499
scale=0.9	0.9902	<u>0.2922</u>	0.2941

5.4 Ablation Study

As shown in the $1^{st} \sim 3^{rd}$ and last rows of Figure 7, we present the results of ablating different modules of *LumiSculpt* . Removing caption augmentation from *LumiHuman* leads to a lack of diverse textual guidance, causing the model during training to rely solely on text conditions that exactly match the dataset, thus improving appearance fitting. As shown in the second row, the generated results exhibit consistent pose and layout. Without the dual-branch structure and decoupling loss, as shown in the third row, the generated appearances tend to overfit the training data, making it challenging to produce diverse backgrounds. As illustrated in the first row, the complete *LumiSculpt* successfully balances diverse appearances with accurate lighting. As shown in the $4^{th} \sim 7^{th}$ rows of Figure 7 and Table 3, we present both qualitative and quantitative analysis results of varying the hyper-parameter *guidance scale*. During inference, the standard *LumiSculpt* sets the guidance scale to 0.5. Increasing the guidance scale intensifies the strength of lighting guidance, enhancing the accuracy of lighting direction and brightness, but an excessively large guidance scale can undermine the model's ability to generate diverse outputs. In contrast, decreasing the guidance scale reduces the strength of lighting guidance, leading to a decrease in lighting accuracy.

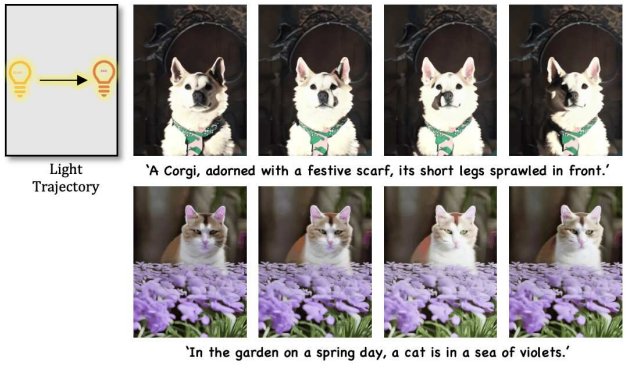

Figure 8: The results of *LumiSculpt* on animals.

Generalization. As shown in the Figure 8, *LumiSculpt* provides lighting priors on animals, which demonstrates the generalization ability to real-world cases.

6 Conclusion

In this paper, we address the challenge of lighting control in text-to-video generation. To address data scarcity, lighting representation difficulties, and lighting injection complexities, we introduce a flexible portrait lighting dataset, *LumiHuman* , along with a plug-and-play lighting guidance method, *LumiSculpt* . Our proposed dual-branch structure and associated loss function for decoupling are not only effective for lighting control but also have the potential to be generalized across a variety of generative tasks. As video generation techniques continue to evolve, enabling precise lighting control becomes an increasingly important research direction, driven by significant aesthetic demands from both professionals and the general public. We believe that *LumiHuman* and *LumiSculpt* will serve as valuable resources and methodologies for future explorations in this domain.

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[2] Duygu Ceylan, Chun-Hao P Huang, and Niloy J Mitra. 2023. Pix2Video: Video editing using image diffusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23206–23217.

[3] Wenhao Chai, Xun Guo, Gaoang Wang, and Yan Lu. 2023. StableVideo: Text-driven consistency-aware diffusion video editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 23040–23050.

[4] Haoxin Chen, Menghan Xia, Yingqing He, Yong Zhang, Xiaodong Cun, Shaoshu Yang, Jinbo Xing, Yaofang Liu, Qifeng Chen, Xintao Wang, Chao Weng, and Ying Shan. 2023. VideoCrafter1: Open Diffusion Models for High-Quality Video Generation. *arXiv preprint arXiv:2310.19512* (2023).

[5] Paul Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. 2000. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques (SIGGRAPH '00)*. ACM Press/Addison-Wesley Publishing Co., USA, 145–156.

[6] Epic Games. 2024. https://www.unrealengine.com/en-US

[7] Yuan Gong, Youxin Pang, Xiaodong Cun, Menghan Xia, Yingqing He, Haoxin Chen, Longyue Wang, Yong Zhang, Xintao Wang, Ying Shan, and Yujiu Yang. 2023. TaleCrafter: Interactive Story Visualization with Multiple Characters. In *ACM SIGGRAPH Asia Conference Proceedings*. 101:1–101:10.

[8] Yuwei Guo, Ceyuan Yang, Anyi Rao, Yaohui Wang, Yu Qiao, Dahua Lin, and Bo Dai. 2024. AnimateDiff: Animate Your Personalized Text-to-Image Diffusion Models without Specific Tuning. In *International Conference on Learning Representations (ICLR)*.

[9] Hao He, Yinghao Xu, Yuwei Guo, Gordon Wetzstein, Bo Dai, Hongsheng Li, and Ceyuan Yang. 2024. CameraCtrl: Enabling camera control for text-to-video generation. *arXiv preprint arXiv:2404.02101* (2024).

[10] Yingqing He, Menghan Xia, Haoxin Chen, Xiaodong Cun, Yuan Gong, Jinbo Xing, Yong Zhang, Xintao Wang, Chao Weng, Ying Shan, et al. 2023. Animate-A-Story: Storytelling with Retrieval-Augmented Video Generation. *arXiv preprint arXiv:2307.06940* (2023).

[11] Jonathan Ho, William Chan, Chitwan Saharia, Jay Whang, Ruiqi Gao, Alexey Gritsenko, Diederik P Kingma, Ben Poole, Mohammad Norouzi, David J Fleet, et al. 2022. Imagen Video: High definition video generation with diffusion models. *arXiv preprint arXiv:2210.02303* (2022).

[12] Jonathan Ho, Tim Salimans, Alexey Gritsenko, William Chan, Mohammad Norouzi, and David J Fleet. 2022. Video Diffusion Models. In *International Conference on Learning Representations (ICLR)*.

[13] Andrew Hou, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. 2022. Face relighting with geometrically consistent shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4217–4226.

[14] Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. 2021. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 14719–14728.

[15] Xun Huang and Serge Belongie. 2017. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*. 1501–1510.

[16] Yongcheng Jing, Yezhou Yang, Zunlei Feng, Jingwen Ye, Yizhou Yu, and Mingli Song. 2020. Neural Style Transfer: A Review. *IEEE Transactions on Visualization and Computer Graphics* 26, 11 (nov 2020), 3365–3385.

[17] Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. 2024. SwitchLight: Co-design of Physics-driven Architecture and Pre-training Framework for Human Portrait Relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 25096–25106.

[18] Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Niebner, and Yannick Hold-Geoffroy. 2024. LightIt: Illumination Modeling and Control for Diffusion Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE Computer Society, Los Alamitos, CA, USA, 9359–9369. doi:10.1109/CVPR52733.2024.00894

[19] PKU-Yuan Lab and Tuzhan AI etc. 2024. *Open-Sora-Plan*. doi:10.5281/zenodo.10948109

[20] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*. PMLR, 12888–12900.

[21] Isabella Liu, Linghao Chen, Ziyang Fu, Liwen Wu, Haian Jin, Zhong Li, Chin Ming Ryan Wong, Yi Xu, Ravi Ramamoorthi, Zexiang Xu, and Hao Su. 2024. OpenIllumination: A Multi-Illumination Dataset for Inverse Rendering Evaluation on Real Objects. arXiv:2309.07921 [cs.CV]

[22] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

[23] Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M Patel. 2024. Holo-Relighting: Controllable Volumetric Portrait Relighting from a Single Image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4263–4273.

[24] Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, and Vishal M Patel. 2023. LightPainter: interactive portrait relighting with freehand scribble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 195–205.

[25] MetaHuman. 2023. https://www.unrealengine.com/en-US/metahuman

[26] Thomas Nestmeyer, Jean-François Lalonde, Iain Matthews, and Andreas Lehrmann. 2020. Learning physics-guided face relighting under directional light. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5124–5133.

[27] Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. 2021. Total Relighting: learning to relight portraits for background replacement. *ACM Transactions on Graphics* 40, 4 (2021), 43:–43:21.

[28] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. 8748–8763.

[29] Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. 2024. Relightful Harmonization: Lighting-aware Portrait Background Replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6452–6462.

[30] Tiancheng Sun, Jonathan T Barron, Yun-Ta Tsai, Zexiang Xu, Xueming Yu, Graham Fyffe, Christoph Rhemann, Jay Busch, Paul Debevec, and Ravi Ramamoorthi. 2019. Single image portrait relighting. *ACM Transactions on Graphics* 38, 4 (2019), 79:1–79:12.

[31] Jiuniu Wang, Hangjie Yuan, Dayou Chen, Yingya Zhang, Xiang Wang, and Shiwei Zhang. 2023. ModelScope Text-to-Video Technical Report. *arXiv preprint arXiv:2308.06571* (2023).

[32] Xiang Wang, Hangjie Yuan, Shiwei Zhang, Dayou Chen, Jiuniu Wang, Yingya Zhang, Yujun Shen, Deli Zhao, and Jingren Zhou. 2023. VideoComposer: Compositional Video Synthesis with Motion Controllability. In *Advances in Neural Information Processing Systems (NeurIPS)*.

[33] Yifan Wang, Aleksander Holynski, Xiuming Zhang, and Xuaner Zhang. 2023. SunStage: Portrait reconstruction and relighting using the sun as a light stage. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20792–20802.

[34] Zhibo Wang, Xin Yu, Ming Lu, Quan Wang, Chen Qian, and Feng Xu. 2020. Single image portrait relighting via explicit multiple reflectance channel modeling. *ACM Transactions on Graphics* 39, 6 (2020), 220:1–220:13.

[35] Yujie Wei, Shiwei Zhang, Zhiwu Qing, Hangjie Yuan, Zhiheng Liu, Yu Liu, Yingya Zhang, Jingren Zhou, and Hongming Shan. 2024. DreamVideo: Composing Your Dream Videos with Customized Subject and Motion. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 6537–6549.

[36] Jay Zhangjie Wu, Yixiao Ge, Xintao Wang, Stan Weixian Lei, Yuchao Gu, Yufei Shi, Wynne Hsu, Ying Shan, Xiaohu Qie, and Mike Zheng Shou. 2023. Tune-a-Video: One-shot tuning of image diffusion models for text-to-video generation. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 7623–7633.

[37] Xiaoyan Xing, Vincent Tao Hu, Jan Hendrik Metzen, Konrad Groh, Sezer Karaoglu, and Theo Gevers. 2024. Retinex-Diffusion: On Controlling Illumination Conditions in Diffusion Models via Retinex Theory. *arXiv preprint arXiv:2407.20785* (2024).

[38] Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. 2022. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics* 41, 6 (2022), 23:1–231:21.

[39] Chong Zeng, Yue Dong, Pieter Peers, Youkang Kong, Hongzhi Wu, and Xin Tong. 2024. Dilightnet: Fine-grained lighting control for diffusion-based image generation. In *ACM SIGGRAPH 2024 Conference Papers*. 1–12.

[40] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2023. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 3836–3847.

[41] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. 2024. IC-Light GitHub Page.

[42] Xiuming Zhang, Sean Fanello, Yun-Ta Tsai, Tiancheng Sun, Tianfan Xue, Rohit Pandey, Sergio Orts-Escolano, Philip Davidson, Christoph Rhemann, Paul Debevec, et al. 2021. Neural light transport for relighting and view synthesis. *ACM Transactions on Graphics* 40, 1 (2021), 9:1–9:17.

[43] Yuxin Zhang, Fan Tang, Nisha Huang, Haibin Huang, Chongyang Ma, Weiming Dong, and Changsheng Xu. 2023. MotionCrafter: One-shot motion customization of diffusion models. *arXiv preprint arXiv:2312.05288* (2023).

[44] Rui Zhao, Yuchao Gu, Jay Zhangjie Wu, David Junhao Zhang, Jiawei Liu, Weijia Wu, Jussi Keppo, and Mike Zheng Shou. 2023. MotionDirector: Motion Customization of Text-to-Video Diffusion Models. *arXiv preprint arXiv:2310.08465* (2023).

[45] Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. 2019. Deep Single Portrait Image Relighting. In *International Conference on Computer Vision (ICCV)*.

[46] Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, et al. 2025. Light-A-Video: Training-free Video Relighting via Progressive Light Fusion. *arXiv preprint arXiv:2502.08590* (2025).