

Rebuttal for “*LumiSculpt*”

ICLR 2025,

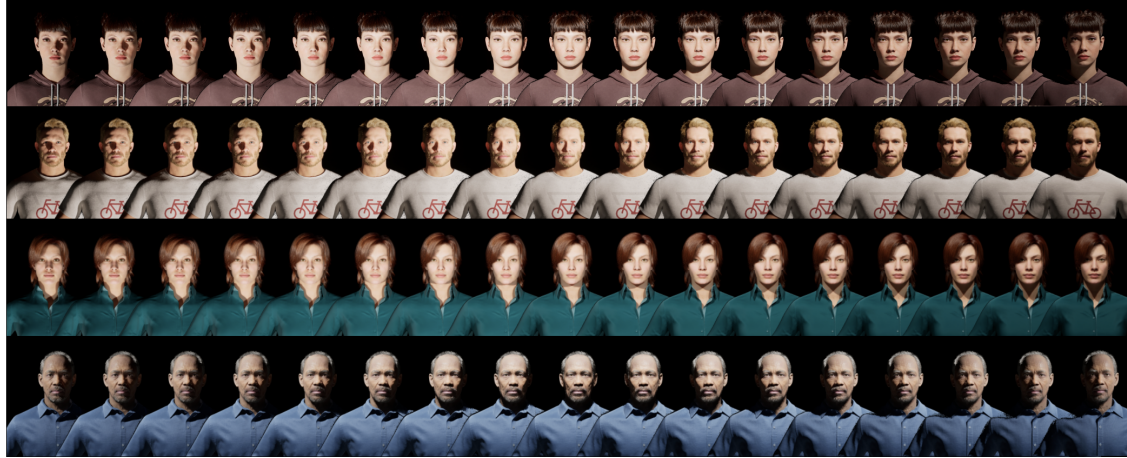
Manuscript ID: 4419

[Common Concern 1] Statistics and Diversity of *LumiHuman* Dataset

We introduce the *LumiHuman* dataset, a continuous lighting video dataset comprising over **220K** different videos (*i.e.*, **2.3 million** images). The resolution of each video is 1024×1024 . As shown in Fig. 1, *LumiHuman* includes 65 diverse human subjects, 30K lighting positions, and over 3K lighting trajectories for each people.



(a) Divers individuals in LumiHuman



(b) Frames with various lighting trajectories in LumiHuman

Fig. 1. Samples in *LumiHuman*.

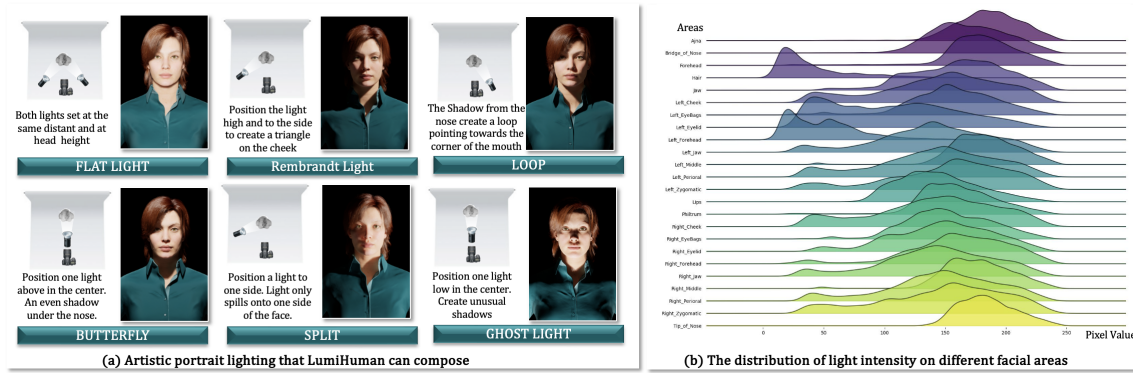


Fig. 2. Illustration of the light sources and camera, and the ridge plot of illuminated areas (*i.e.*, face patches.)

Table I: Comparison of other lighting-related datasets.

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

Our *LumiHuman* of 65 human identities is sufficient for training *LumiSculpt*, which is supported by extensive qualitative and quantitative experiments. The scalability of synthetic data lies in the ability to construct diverse light trajectories, leveraging varied lighting data to facilitate the model’s learning of illumination harmonization.

[Common Concern 2] Similar to ControlNet

LumiSculpt is distinct from ControlNet in terms of its task, motivation, module design, training objective, training data, backbone, and generated results. A detailed explanation for each point is provided below:

- **Task:** *LumiSculpt* is a specialized lighting control method designed for DiT based T2V models. ControlNet is a control method that focuses on image geometry (pose, depth map, canny, etc.) for U-Net based T2I models.
- **Motivation:** *LumiSculpt*’s motivation focuses on elements in videos that affect realism and aesthetics, i.e., lighting, and proposes a method to achieve coherent video generation with controllable lighting. ControlNet’s motivation stems from the randomness in T2I diffusion models, hence it introduces a method for generating images with controllable geometry.
- **Module Design:** As shown in Fig. 3(d), *LumiSculpt* employs self-attention mechanisms as the lighting encoder and uses linear layers and latent weighting as condition injection mechanisms. ControlNet uses the U-Net Encoder to extract features and injects conditions by adding latents. These atomic components are commonly used and necessary for feature extraction and condition injection, which are not limited to a specific method.
- **Training Objective:** *LumiSculpt* tackles the core challenge of the entanglement of lighting and appearance. As shown in Fig. 3(c), *LumiSculpt* employs a dual-branch structure and an appearance-lighting disentanglement loss. ControlNet is trained with the diffusion noise prediction loss.
- **Training Data:** *LumiSculpt* utilizes video data with coherent inter-frame lighting changes, whereas ControlNet is based on independent images.
- **Backbone:** *LumiSculpt* is build upon DiT-based Open-Sora-Plan (Lab & etc., 2024), and ControlNet is designed for U-Net structured Stable Diffusion (Rombach et al., 2022).
- **Generated Results:** *LumiSculpt* generates coherent videos while ControlNet generates images.

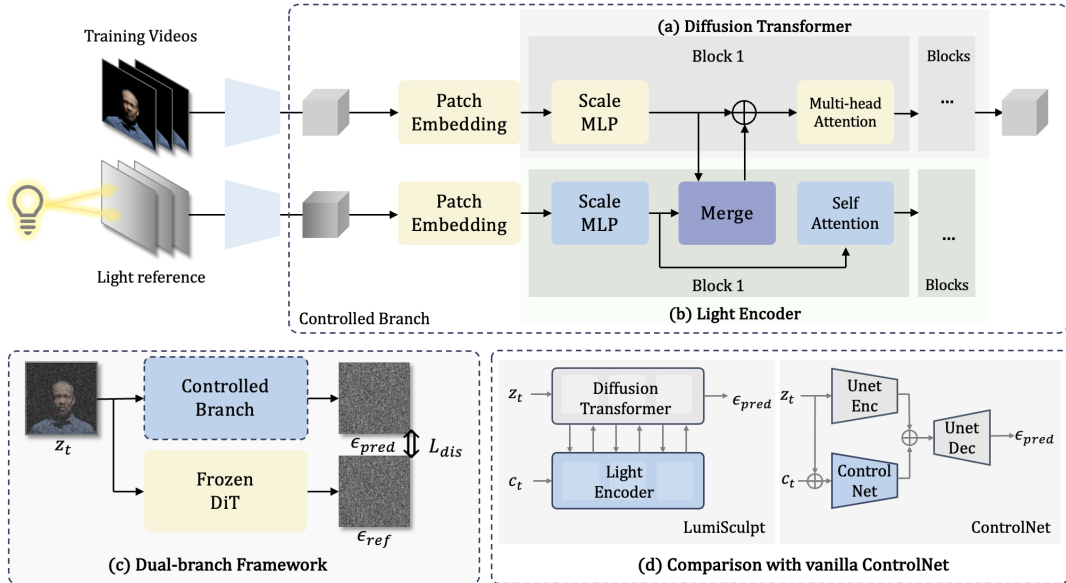


Fig. 3. Differences between *LumiSculpt* and ControlNet.

We implement ControlNet to video lighting control by training with paired frames in *LumiHuman* and generating image sequence as video. The comparison results are shown in Fig. 4 and Tab. II. ControlNet struggles to achieve lighting control, generating images with random lighting. This validates the effectiveness of our model structure and training methodology.

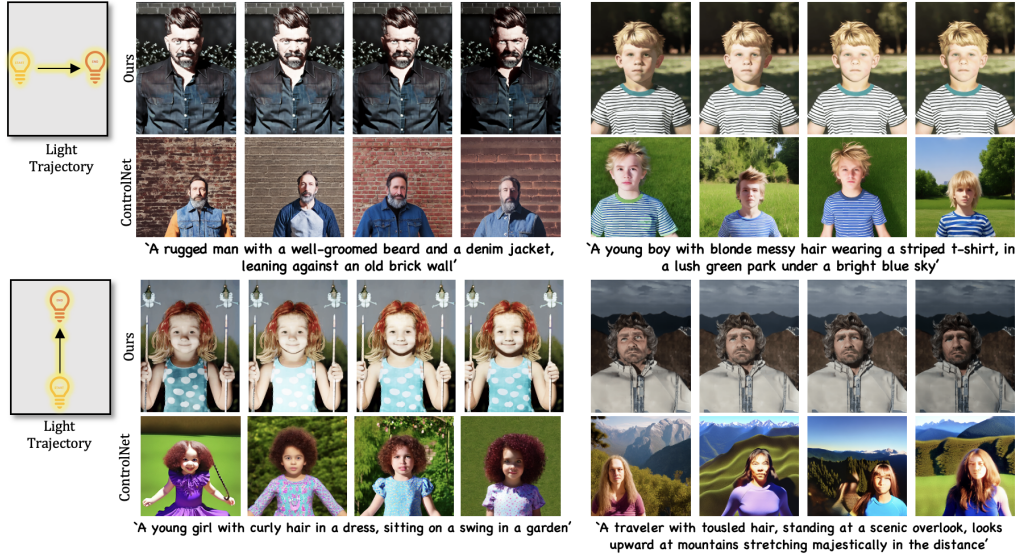


Fig. 4. Comparison results with state-of-the-art methods ControlNet (Zhang et al., 2023).

Table II: Quantitative experimental results and ablation study results. The best results are marked as **bold**.

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	0.3440
Ours	0.9951	1.1312	0.3500	0.8779	0.3597

Referee: #1 rLXS

Comment #1

LumiHuman is synthetic, which may limit the model’s performance in real-world cases. I wonder if there can be a thorough evaluation of real-world cases. There are only 65 individuals in the dataset, which may limit the model to generalize to new portraits.

Response: Thanks for your suggestion. As shown in the Fig. 5, *LumiSculpt* supports the generation of videos featuring diverse backgrounds, environments, and characters and also provides lighting priors on **non-human objects**. This demonstrates the generalization ability to real-world cases.



Fig. 5. More results with *LumiSculpt*.

Synthetic data does not compromise the model’s generalization. During training, *LumiSculpt* employ various strategies to mitigate overfitting, ensuring that the light control module primarily learns the patterns of light variation rather than the appearance of the characters. To evaluate with real-world case, we employ the commonly used FID (Seitzer, 2020) score to assess the photo-realism of both *LumiSculpt* and Open-Sora (Lab & etc., 2024) within the FFHQ (Karras et al., 2019) dataset. As shown in Table III, *LumiSculpt* achieves a better FID score, demonstrating its ability to generate realistic videos.

Table III: FID of *LumiSculpt* and Open-Sora using the FFHQ (Karras et al., 2019) dataset

Method	Open-Sora	LumiSculpt
FID ↓	35.7	33.0

The “65 individuals” is also not the limiting factor for model training. *LumiSculpt* learns lighting variation patterns and achieves generalization through diverse light trajectories constructed from synthetic data, rather than relying on human appearances.

Comment #2

The generated videos are not informative enough. The motion dynamics are not enough. I wonder if there are results where the portrait and background can move more vividly

Response: Thanks for the comment. Yes, with motion descriptions, *LumiSculpt* exhibits motion dynamics where the portrait and background can move more vividly. As shown in Fig. 6, we have marked the regions with significant motion changes. Actually, generating portrait and background with vivid dynamics is challenging for T2V models, and it is even harder to control both lighting and motion dynamics. As illustrated in Fig. 7, applying image-based lighting control methods (since there is no suitable video-based model available) cannot achieve inter-frame consistency. Therefore, *LumiSculpt* provides a novel solution for controllable video generation, particularly focused on lighting.

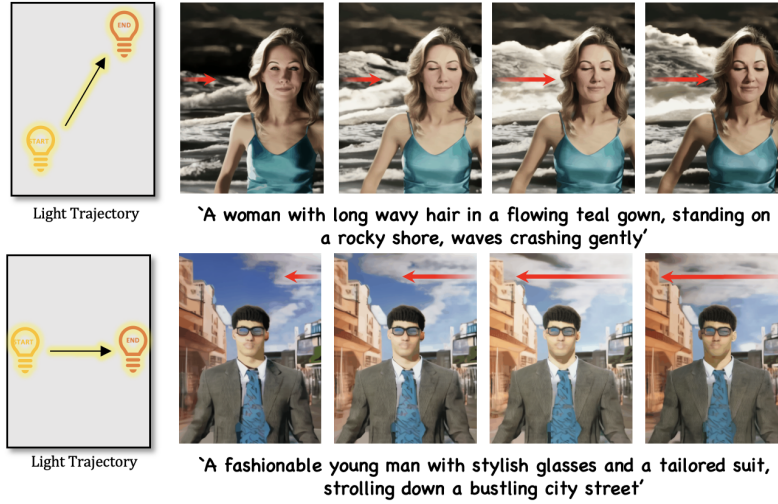


Fig. 6. Dynamic videos generated by *LumiSculpt*.



Fig. 7. Comparison results with state-of-the-art methods IC-Light (Zhang et al., 2024).

Referee: #2 MY7D

Comment #1

This algorithm seems more suitable for image generation, as I did not observe any specific design tailored for video tasks. Video generation is merely an extension of the algorithm’s application.

Response: Thanks for the comment. **Firstly**, *LumiSculpt* incorporated 3D attention specifically designed for temporal modeling in videos. All light injection modules in this work are built upon the backbone of the video diffusion generation model, ensuring consistent temporal modeling of light dynamics without compromising the model’s original generative capabilities. **Secondly**, lighting control in image generation primarily focuses on harmonizing lighting between the background and the subject. When directly applied the image based method to video generation, it may result in severe temporal inconsistencies, as each frame may exhibit different visual content. In contrast, our approach demonstrates smooth and stable lighting across video frames, reflecting the effectiveness of our current design, which including conditional extraction and injection methods, for video generation.

Comment #2

In the comparisons, the authors use images generated by the network as the foreground. Does this imply that, limited by the synthetic data used during training, the algorithm may not generalize well to real-world scenes? I also noticed unnatural foreground (human) generation results in the video demo.

Response: Thanks. Synthetic data does not compromise the model’s generalization. During training, *LumiSculpt* also employ various strategies to mitigate overfitting, ensuring that our light control module primarily learns the patterns of light variation rather than the appearance or content of the characters. We employ the commonly used FID (Seitzer, 2020) score to assess the realism of the generated results for both *LumiSculpt* and Open-Sora (Lab & etc., 2024) within the FFHQ (Karras et al., 2019) dataset. As shown in Table IV, *LumiSculpt* achieves a better FID score, demonstrating its ability to generate realistic videos.

Table IV: FID of *LumiSculpt* and Open-Sora using the FFHQ (Karras et al., 2019) dataset

Method	Open-Sora	LumiSculpt
FID ↓	35.7	33.0

LumiSculpt is a T2V method, aiming at generating lighting controllable videos by texts. Thus, the ability to generate both foreground and background with text is an advantage of *LumiSculpt*. IC-Light’s goal is re-lighting, which involves harmonizing lighting between foreground and background images. Thus, IC-Light’s foreground is generated by *LumiSculpt* because it needs a foreground image.

Comment #3

Can this dataset be open-sourced to ensure reproducibility for future work?

Response: Yes, it certainly will be open-sourced upon acceptance.

Comment #4

I find the caption augmentation section somewhat unclear. Is it simply replacing captions, or does it involve corresponding changes in the image background as well?

Response: It is replacing captions. During training, the augmented captions serve as textual conditions input into the dual-branch models. These captions can guide the frozen branch to produce latents for the same character against different backgrounds, which act as regularization samples providing stronger appearance constraints for the \mathcal{L}_{dis} . This drives the Controlled Branch to generate richer backgrounds instead of only black backgrounds. As shown in the first and second rows of Fig. 8, the inclusion of augmented captions enhances the model’s ability to generate diverse backgrounds and layouts.



Fig. 8. Ablation results of augmented captions.

Referee: #3 wtpx

Comment #1

The synthetic renderings could follow the usual light stage setup with full coverage, not just frontal lighting.

Response: We sincerely appreciate your valuable suggestions regarding lighting settings. *LumiHuman* only include light sources in front of the characters, because in an environment with point light sources, the light behind the characters would be blocked by the human body, resulting in a black image, or it appears as a near-white light spot, making it difficult to see the object. These phenomena exist in both generated data and real-world light-stage data (Liu et al., 2024).

Our current light matrix is capable of creating rich light and shadow effects. *LumiHuman* provides over 30K lighting positions and over 3K lighting trajectories for each individual. These lighting positions can create light and shadow effects in **all areas** of the human face. As shown in Fig. 9, we present the brightness distribution map of different regions of the human face. Each ridge in the ridge plot represents a different facial area, with the horizontal axis indicating brightness and the vertical axis indicating the number of samples at the corresponding brightness. *LumiHuman* covers all areas of the face and distributes samples across a wide range of brightness levels.

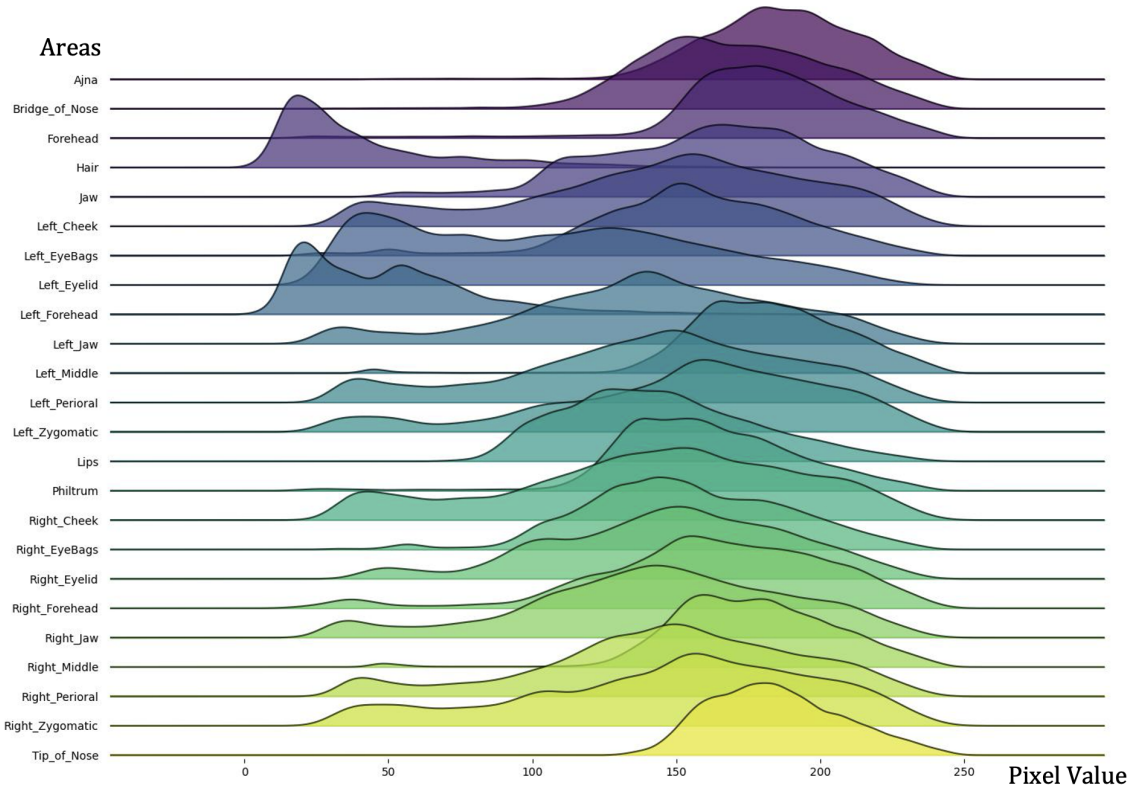


Fig. 9. The distribution of light intensity on different facial areas of the characters. *LumiHuman*'s lighting matrix can cover all areas of the face and produce a significant range of light and shadow variations.

Comment #2

Furthermore, it is not clear whether 65 identities can provide enough diversity. I believe that the main advantage of using a synthetic dataset is that it can be scaled.

Response: Thanks for the comment. Our *LumiHuman* of 65 human identities can provide sufficient diversity to train *LumiSculpt*, which is supported by extensive qualitative and quantitative experiments. The scalability of synthetic data lies in the ability to construct diverse light trajectories, leveraging varied lighting data to facilitate the model’s learning of illumination harmonization. As shown in Tab. V, compared to other lighting datasets Openillumination (Liu et al., 2024) and Deep Portrait Relighting (DPR) dataset (Zhou et al., 2019)(generated from face image dataset Celeb-A (Liu et al., 2015)), *LumiHuman* outperforms in light positions, light movements and number of images.

Table V: Comparison of other lighting-related datasets.

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

Comment #3

It would also be crucial to show that available real-world light-stage datasets cannot provide enough supervision to achieve such quality for lighting control.

Response: Thanks for the valuable suggestion. **Firstly**, available public light-stage datasets, e.g., Openillumination (Liu et al., 2024), do not contain human subject data, and its One-Light-At-a-Time (OLAT) data comprises only 142 lighting positions, which is hard to achieve smooth changes in lighting. Relying solely on publicly light-stage datasets is insufficient for T2V model training. **Secondly**, real-world light-stage datasets rely on HDR maps that have a significant domain gap with T2I and T2V scenarios. **In summary**, as shown in Figure 10, *LumiHuman* provides a coordinated, large spatial range of light sources, enabling users to freely combine the types of lighting they require.

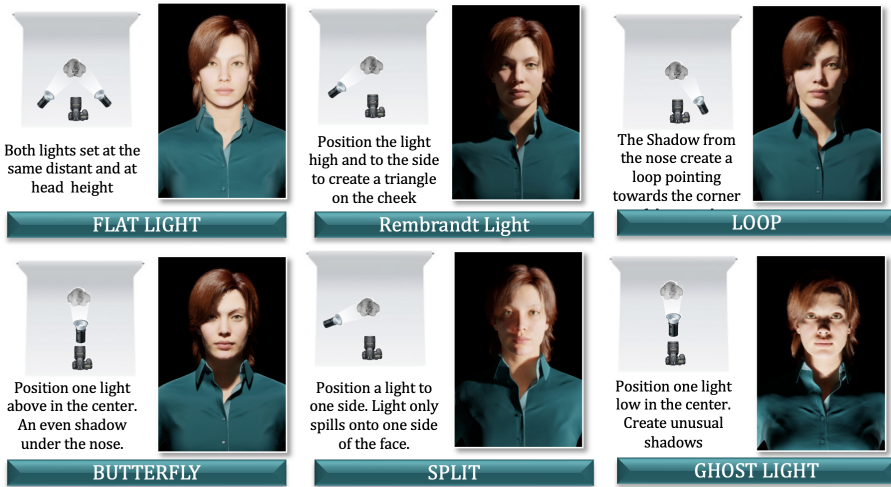


Fig. 10. *LumiHuman* offers a variety of basic elements that can be combined to form various types of portrait lighting, widely applicable to a range of tasks related to character lighting.

Comment #4

It would be important to highlight the key difference to ControlNet.

Response: Thanks. *LumiSculpt* is distinct from ControlNet in terms of its task, motivation, module design, model backbone, generated results, training objective and training data.

- **Task:** *LumiSculpt* is a specialized lighting control method designed for DiT based T2V models. ControlNet is a control method that focuses on image geometry (pose, depth map, canny, etc.) for U-Net based T2I models.
- **Motivation:** *LumiSculpt*'s motivation focuses on elements in videos that affect realism and aesthetics, i.e., lighting, and proposes a method to achieve coherent video generation with controllable lighting. ControlNet's motivation stems from the randomness in T2I diffusion models, hence it introduces a method for generating images with controllable geometry.
- **Module Design:** As shown in Fig. 11(d), *LumiSculpt* employs self-attention mechanisms as the lighting encoder and uses linear layers and latent weighting as condition injection mechanisms. ControlNet uses the U-Net Encoder to extract features and injects conditions by adding latents. These atomic components are commonly used and necessary for feature extraction and condition injection, which are not limited to a specific method.
- **Training Objective:** *LumiSculpt* tackles the core challenge of the entanglement of lighting and appearance. As shown in Fig. 11(c), *LumiSculpt* employs a dual-branch structure and an appearance-lighting disentanglement loss. ControlNet is trained with the diffusion noise prediction loss.
- **Training Data:** *LumiSculpt* utilizes video data with coherent inter-frame lighting changes, whereas ControlNet is based on independent images.
- **Backbone:** *LumiSculpt* is build upon DiT-based Open-Sora-Plan (Lab & etc., 2024), and ControlNet is designed for U-Net structured Stable Diffusion (Rombach et al., 2022).
- **Generated Results:** *LumiSculpt* generates coherent videos while ControlNet generates images.

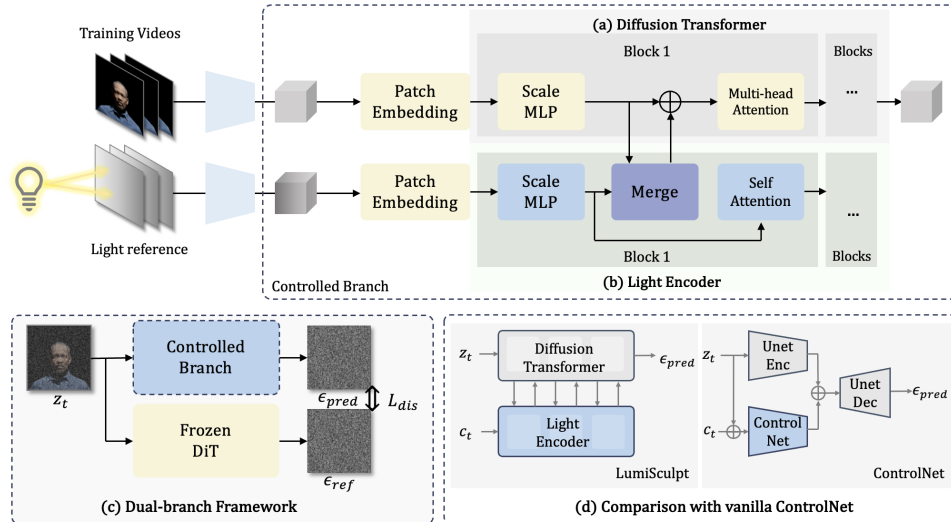


Fig. 11. Differences between *LumiSculpt* and ControlNet.

Comment #5

Now, it seems that the key difference is the dual-branch predictions, although the effectiveness of this idea is questionable based on the ablation. Furthermore, the proposed disentanglement loss is not well-motivated. The key assumption is that the latents reflect the appearance. However, the latents contain geometric, material, and also lighting features, thus not being disentangled.

Response: The dual-branch framework is proposed to address the core challenge of the entanglement of illumination and appearance. The proposed disentanglement loss is designed with the motivation for forcing the appearance distribution follow the backbone model, thus achieve disentanglement of appearance and lighting. **Specifically**, the \mathcal{L}_{dis} calculates the mean and variance of each channel of the latent features, i.e. distributional differences between two latents without considering geometric features. This method of appearance disentanglement has been proven effective in a series of style transfer tasks (Huang & Belongie, 2017; Johnson et al., 2016). As shown in Fig. 12, without \mathcal{L}_{dis} , the background would overfit to black.



Fig. 12. Ablation results of \mathcal{L}_{dis} .

Comment #6

It would be great to show the diversity of the generated samples - more samples with the same conditioning.

Response: Thanks for the suggestion. As shown in Fig. 13, we present more results with the same prompt.



Fig. 13. LumiSculpt results with same prompt.

Comment #7

Additional baseline comparisons would be important. Although the method uses the T2V models for light editing, the resulting videos are static, making it fair to compare against T2I models. Such comparisons could also give interesting insights about the lighting priors of T2I and T2V models.

Response: Thanks. The only appropriate open-source light control **T2I** methods is IC-Light. Existing relighting methods, such as Relightful Harmonization (Ren et al., 2024), target on **harmonizing** the lighting of a given foreground image and a background image. Our method achieves controllable lighting for T2V generation, where both characters and backgrounds are specified by text prompts. Therefore, relighting methods are not applicable to our task.

Comment #8

The key contribution is not clear. Based on the title and abstract it is LumiSculpt, based on the intro (L.087 - Additionally...) it is the dataset LumiHuman.

Response: Thanks. We will revise the manuscript to avoid confusion. Both the dataset and methods are integral contributions of our work, which are **equally important**. Since we introduce a new task, it requires collecting suitable training data from scratch. The proposed *LumiHuman* dataset consists of videos showcasing varied and controllable lighting changes. Additionally, our model, *LumiSculpt*, is specifically designed for this task. The core contribution of *LumiSculpt* is achieving temporally stable light control through a DiT based generative model. **In conclusion**, the allocation of contributions in this work is similar to previous works like IC-Light (Zhang et al., 2024) and Relightful Harmonization (Ren et al., 2024), where the dataset and the method are equally significant.

Comment #9

Recent T2I lighting control methods, such as LightIt could be discussed.

Response: Thanks for for introducing LightIt (Kocsis et al., 2024). We will cite this work and highlight the differences between LightIt and our approach. Specifically, LightIt is an image-guided (I2I) method for image relighting which requires additional estimated shading and normals. Our method, in contrast, is text-guided (T2V) and requires only text and target lighting conditions to achieve video lighting control. These differences provide valuable insights for our method design.

Comment #10

It might be better to narrow the title, reflecting that the domain is human portraits.

Response: Thanks. *LumiSculpt* is **not restricted** to humans, we have experimented with some animal cases and also achieved stable lighting control effects, as shown in Fig. 14. It shows that *LumiSculpt* enables the model to learn about lighting priors and extend this knowledge to non-human objects.



Fig. 14. *LumiSculpt* results with non-human objects.

Comment #11

What is the reason that the generated samples have a very similar geometry and appearance as IC Light, but highly different to Open-Sora, although the proposed method uses Open-Sora.

Response: This issue arises from our experimental settings. The foreground image fed to IC-Light is generated by *LumiSculpt*, as IC-Light is a relighting method that focuses on generating backgrounds and the overall lighting harmony. In contrast, Open-Sora results are generated from random noise. It is worth noting that *LumiSculpt* is a fully functional and comprehensive T2V generative model designed to create controllable videos with lighting effects beyond relighting.

Comment #12

Could you please give a bit more details, how exactly are the augmented captions used? If I understand it correctly, the goal with those is to give additional noise to the model to avoid overfitting.

Response: The goal of the augmented captions is to provide regularization samples to the model to avoid overfitting. The regularization samples are latents of the same character against different backgrounds. Specifically, during training, the augmented captions serve as textual conditions into the dual-branch models. These captions can guide the frozen branch to produce latents for the same character against different backgrounds, which act as regularization samples providing strong appearance constraints for the \mathcal{L}_{dis} . This drives the Controlled Branch to generate richer backgrounds instead of only black backgrounds.

Comment #13

The results look oversaturated, what can be the reason for that?

Response: We are unsure which specific case the reviewer refers to regarding oversaturated. While some color deviations might occur due to the VAE and the pretrained backbone, overall, we think the results align well with standard aesthetic expectations. We employ the commonly used FID (Seitzer, 2020) score to assess the realism of the generated results for both *LumiSculpt* and Open-Sora (Lab & etc., 2024) within the FFHQ (Karras et al., 2019) dataset. As shown in Table VI, the FID score of *LumiSculpt* is better, demonstrating its ability to generate realistic videos.

Table VI: FID of *LumiSculpt* and Open-Sora using the FFHQ (Karras et al., 2019) dataset.

Method	Open-Sora	<i>LumiSculpt</i>
FID ↓	35.7	33.0

Referee: #4 LUeN

Comment #1

How diverse the MetaHuman dataset is since it only contains 65 individuals.

Response:

The diversity of *LumiHuman* mainly lies in the variety of light trajectories rather than the individuals, leveraging varied lighting data to facilitate the model’s learning of illumination rather than human appearance. Specifically, as shown in Tab. VII, compared to other lighting datasets Openillumination (Liu et al., 2024) and Deep Portrait Relighting (DPR) dataset (Zhou et al., 2019) (generated from face image dataset Celeb-A (Liu et al., 2015)), *LumiHuman* outperforms in light positions, light movements and number of images, which demonstrates the diversity of *LumiHuman*. Moreover, our *LumiHuman* of 65 human identities is sufficient for training *LumiSculpt*, which is supported by extensive qualitative and quantitative experiments. Fig. 15 shows real samples of human individuals in *LumiHuman*.

Table VII: Comparison of other lighting-related datasets.

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



Fig. 15. Real samples in *LumiHuman*.

Comment #2

How accurate your caption could describe the lighting since lighting caption is a very unique task that current LLM model is not doing well. From the results, I didn't see any caption related to lighting.

Response: The caption only provides a supplementary semantic condition, such as background, character details, *etc.*, and the precision of light control is guided by the input lighting reference video. Each frame in *LumiHuman* is paired with a lighting reference, allowing the descriptions of the lighting to be added to the captions, without relying on a Large Language Model (LLM). As commented by the reviewer, determining lighting remains a challenge for LLMs, and even for humans, since lighting itself is inherently difficult to describe in language. In contrast, the lighting reference video captures accurate lighting conditions, which serves as input and is easily interpreted by diffusion models.

Comment #3

Since the model is trained on synthetic rendered images, the results are far from photo-realistic and most of the results from the teaser images are 'fake' portrait with unrealistic facial texture.

Response: Thanks. Synthetic data does not compromise the model's generalization. During training, *LumiSculpt* also employ various strategies to mitigate overfitting, ensuring that our light control module primarily learns the patterns of light variation rather than the appearance or content of the characters. We employ the commonly used FID (Seitzer, 2020) score to assess the realism of the generated results for both *LumiSculpt* and Open-Sora (Lab & etc., 2024) within the FFHQ (Karras et al., 2019) dataset. As shown in Table VIII, the FID score of *LumiSculpt* is better, demonstrating its ability to generate realistic videos.

Table VIII: FID of *LumiSculpt* and Open-Sora using the FFHQ (Karras et al., 2019) dataset

Method	Open-Sora	LumiSculpt
FID ↓	35.7	33.0

Comment #4

It is not clear how authors control the lighting intensity.

When constructing *LumiHuman*, the light source distance varies in $50cm \sim 210cm$, which can create a noticeable effect of light intensity transitioning on the character's face. During **inference**, light intensity can be freely controlled using a user-specified lighting reference video. The light intensity of lighting reference video is changed by the distance between the light source and the illuminated subject. During model **training**, *LumiSculpt* can learn the mapping between the reference lighting intensity and the visual effects on the character's face from paired training data.

Comment #5

IC-light has much better photo-realistic results compared with your methods. And what's the advantage of authors method ?

Response: We kindly invite the reviewer to revisit our comparison results in Fig. 16 and the supplemented video. IC-Light fails to achieve stable lighting control in videos, as it is an image-based relighting method. It results in inconsistent lighting across frames, with significant variations in both the subject and background across frames. It is worth noting that *LumiSculpt* is a fully functional and comprehensive T2V generative model designed to create controllable videos with lighting effects. While IC-Light does requires a portrait as the foreground input. Regarding photo-realistic, we shown the FID results in Response 3. Our method shows

fairly photo-realistic results.



Fig. 16. Comparison results with state-of-the-art methods IC-Light (Zhang et al., 2024).

Comment #6

It seems that authors only show white/black lighting but not color lighting which ICnet could do.

Response: At present, no T2V generation methods are capable of controlling lighting, which is our primary objective. Modifying the color of the light is beyond the scope of our current work, which we plan to explore in future work.

Comment #7

Regarding model, I don't see any difference between yours and controlnet besides it is a video version.

Response: Thanks. The model of *LumiSculpt* is distinct from ControlNet in terms of its module design, backbone and training objective.

- **Module Design:** As shown in Fig. 17(d), *LumiSculpt* employs self-attention mechanisms as the lighting encoder and uses linear layers and latent weighting as condition injection mechanisms. ControlNet uses the U-Net Encoder to extract features and injects conditions by adding latents. These atomic components are commonly used and necessary for feature extraction and condition injection, which are not limited to a specific method.
- **Backbone:** *LumiSculpt* is build upon DiT-based Open-Sora-Plan (Lab & etc., 2024), and ControlNet is designed for U-Net structured Stable Diffusion (Rombach et al., 2022).
- **Training Objective:** *LumiSculpt* tackles the core challenge of the entanglement of lighting and appearance. As shown in Fig. 17(c), *LumiSculpt* employs a dual-branch structure and an appearance-lighting disentanglement loss. ControlNet is trained with the diffusion noise prediction loss.

We implement ControlNet to video lighting control by training with frames in *LumiHuman* and generating image sequence as video. The comparison results are shown in Fig. 18 and Tab. IX. ControlNet struggles to achieve lighting control, generating images with random lighting. This validates the effectiveness of our model structure and training methodology.

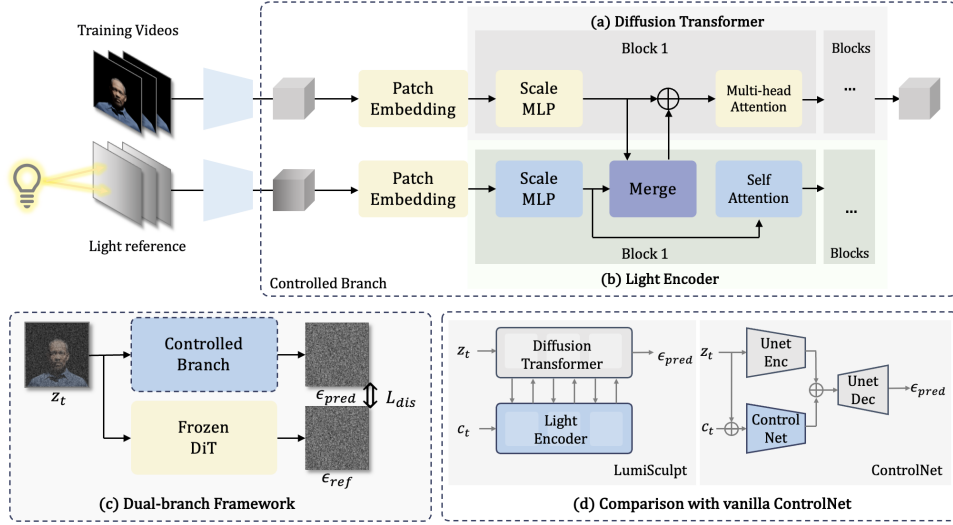


Fig. 17. Differences between *LumiSculpt* and ControlNet.

Table IX: Quantitative experimental results and ablation study results. The best results are marked as **bold**.

Method	Consistency		Lighting Accuracy		Quality
	CLIP \uparrow	LPIPS \downarrow	Direction \downarrow	Brightness \uparrow	CLIP \uparrow
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	0.3440
Ours	0.9951	1.1312	0.3500	0.8779	0.3597



Fig. 18. Comparison results with state-of-the-art methods ControlNet (Zhang et al., 2023).

References

- Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pp. 1501–1510, 2017.
- Justin Johnson, Alexandre Alahi, and Li Fei-Fei. Perceptual losses for real-time style transfer and super-resolution. In *European Conference on Computer Vision (ECCV)*, pp. 694–711. Springer, 2016.
- Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pp. 4401–4410. Computer Vision Foundation / IEEE, 2019. . URL http://openaccess.thecvf.com/content_CVPR_2019/html/Karras_A_Style-Based_Generator_Architecture_for_Generative_Adversarial_Networks_CVPR_2019_paper.html.
- Peter Kocsis, Julien Philip, Kalyan Sunkavalli, Matthias Niebner, and Yannick Hold-Geoffroy. LightIt: Illumination Modeling and Control for Diffusion Models . In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9359–9369, Los Alamitos, CA, USA, June 2024. IEEE Computer Society. . URL <https://doi.ieeecomputersociety.org/10.1109/CVPR52733.2024.00894>.
- PKU-Yuan Lab and Tuzhan AI etc. Open-sora-plan, April 2024. URL <https://doi.org/10.5281/zenodo.10948109>.
- 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. Openillumination: A multi-illumination dataset for inverse rendering evaluation on real objects, 2024.
- Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.
- Mengwei Ren, Wei Xiong, Jae Shin Yoon, Zhixin Shu, Jianming Zhang, HyunJoon Jung, Guido Gerig, and He Zhang. Relightful harmonization: Lighting-aware portrait background replacement. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6452–6462, 2024.
- Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.
- Maximilian Seitzer. pytorch-fid: FID Score for PyTorch. <https://github.com/mseitzer/pytorch-fid>, August 2020. Version 0.3.0.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3836–3847, 2023.
- Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Ic-light github page, 2024.
- Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep single portrait image relighting. In *International Conference on Computer Vision (ICCV)*, 2019.