LUMIFORGE: INVERSE LIGHTING CURATION FOR PRECISE PORTRAIT VIDEO RELIGHTING

Anonymous authors
Paper under double-blind review

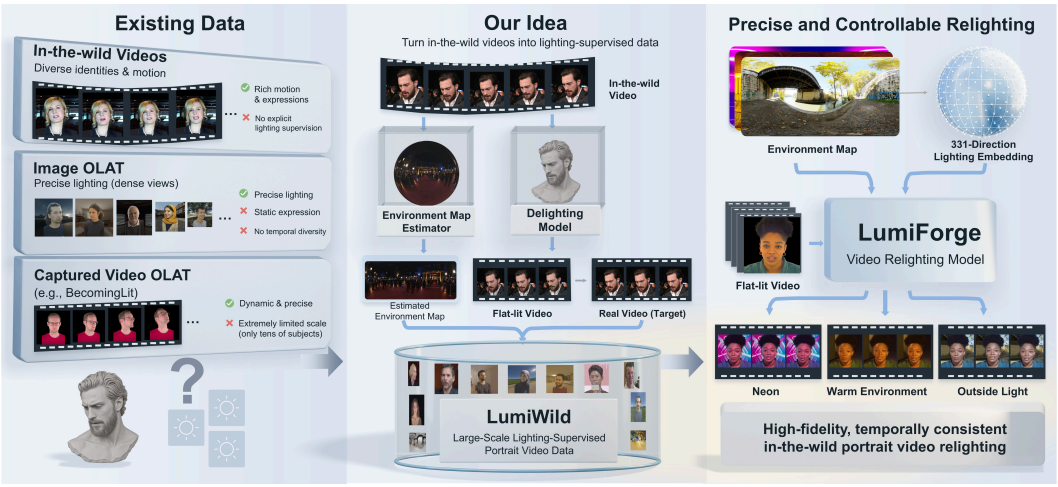

Figure 1: **LumiForge overview.** Recovered illumination turns in-the-wild videos into lighting supervision, complementing OLAT data for environment-controlled portrait video relighting.

ABSTRACT

Precise portrait video relighting requires a moving face to respond faithfully and consistently to a specified environment illumination, far beyond generic lighting stylization. The fundamental obstacle is supervision: OLAT provides accurate directional light responses, but extending such capture to diverse moving faces demands tightly synchronized high-speed cameras and extremely fast illumination control, making large-scale video supervision impractical. We present **LumiForge** with two simple principles: *recover rather than capture*, and *resolve rather than compress*. First, our **Inverse Lighting Curation (ILC)** recovers source illumination and canonical flat-lit appearance from in-the-wild videos, turning real performances into direct relighting supervision. Applying ILC to 15,191 VFHQ clips yields **LumiWild**, an explicitly environment-conditioned portrait video corpus with nearly one million frames. Second, our **ORBIT** conditioning preserves environment illumination as directional light tokens and jointly models subject identity, rather than collapsing lighting into a global appearance code. Together, scalable recovered supervision and resolved illumination enable high-fidelity, temporally consistent, and precisely controllable portrait video relighting, while our environment-map estimator achieves state-of-the-art performance on our held-out portrait-based illumination benchmark.

1 INTRODUCTION

Controllable portrait video relighting asks a moving face to respond to a specified environment: facial shading, highlights, and color should change with incident illumination while the performance remains intact. Learning this relationship faces a tension between *lighting accuracy, natural dynamics, and scale*. Image one-light-at-a-time (OLAT) capture measures directional responses at a fixed

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

facial state. Dynamic light stages add motion but require synchronized imaging and rapid lighting control. In-the-wild videos offer diverse performances at scale, with unknown lighting. These sources provide complementary pieces of the same learning problem.

Our starting point is to recover the missing lighting condition from real performances. **Inverse Lighting Curation (ILC)** constructs a triplet from a pseudo-flat-lit sequence, an estimated source environment, and the original video as target. Thus, real facial motion and appearance remain in the target while inverse models supply its conditioning. Applying ILC to VFHQ (Xie et al., 2022) produces **LumiWild**: 15,191 clips of 65 frames, or 987,415 frame occurrences. These are estimated lighting labels, whose uncertainty motivates complementary controlled OLAT supervision. Compared with the reference-appearance task on wild videos in Lux Post Facto (Mei et al., 2025), ILC makes the recovered environment an explicit condition of the forward relighting task.

Supervision must also reach the model without losing its directional structure. We introduce **ORBIT**, an OLAT-resolved identity–lighting conditioner that aligns 331 light-token slots with the basis used for image supervision and injects identity through a separate attention branch. Individual light slots also permit direct token-level edits, selectively suppressing a lighting component while retaining the remaining illumination. FaceOLAT (Rao et al., 2025) establishes the image relighting prior; BecomingLit (Schmidt et al., 2025) supplies studio motion and trains a delighting teacher; VFHQ-derived triplets expose the model to natural performances. LumiForge directly generates relit video, without reconstructing an OLAT video for every direction.

The remaining challenge is to learn from motion without losing the image prior's lighting response. Joint video adaptation can change that response even when image weights are frozen: altered hidden features change how lighting conditions are read. Our final *Temporal Residual Relighter* therefore retains the image computation and inserts explicit time-only modules. Its single-frame bypass preserves the image function, while motion supervision and novel-light teacher matching guide video adaptation. The resulting design separates an architectural retention property from the empirical quality of multi-frame predictions. Paired synthetic targets, changed-light video comparisons, and interventions evaluate this design. The residual model retains image benchmark scores and approaches joint adaptation's video reference fidelity with fewer trainable parameters. Our contributions are:

- **Environment-aware video supervision:** ILC and the VFHQ-derived LumiWild collection connect real performances to estimated source lighting.
- **Relighting-preserving video adaptation:** ORBIT conditions a frozen image path, with explicit temporal residuals and exact single-frame recovery under matched computation.
- **Selective lighting control:** ORBIT's directional slots expose individual lighting components for editing, supported by token interventions, paired-target measurements, and multi-method video comparisons.

2 Related Work

Portrait Relighting and Light-Transport Supervision. OLAT capture measures directional reflectance fields (Debevec et al., 2000), supporting learned light transport and shadow-aware relighting (Nestmeyer et al., 2020; Hou et al., 2021; R et al., 2021; Sun et al., 2021). Recent methods combine physical representations with stronger generative priors (Mei et al., 2023; Kim et al., 2024; Mei et al., 2024; Chaturvedi et al., 2025). Estimated supervision is also established: SfSNet (Sengupta et al., 2018) combines synthetic labels and real-image reconstruction, DPR (Zhou et al., 2019) constructs physically based training pairs, and Lumos (Yeh et al., 2022) adapts synthetic relighting supervision to real portraits. ILC applies this broader principle to video by retaining the real performance as target and estimating its source environment and flat-lit conditions.

Portrait Video Relighting. Subject-specific reconstruction (Wang et al., 2021; Cai et al., 2024; Lin et al., 2024) complements recent generative approaches (Wang et al., 2025; Guo et al., 2025; Mei et al., 2025). LCVD uses reconstructed-face shading hints, while Comprehensive Relighting uses a temporal-lighting objective. Most closely related, Lux Post Facto combines OLAT-derived HDR supervision with reference-based appearance copying on motion-rich wild videos. Our wild-video branch instead receives an estimated environment condition. BecomingLit (Schmidt et al.,

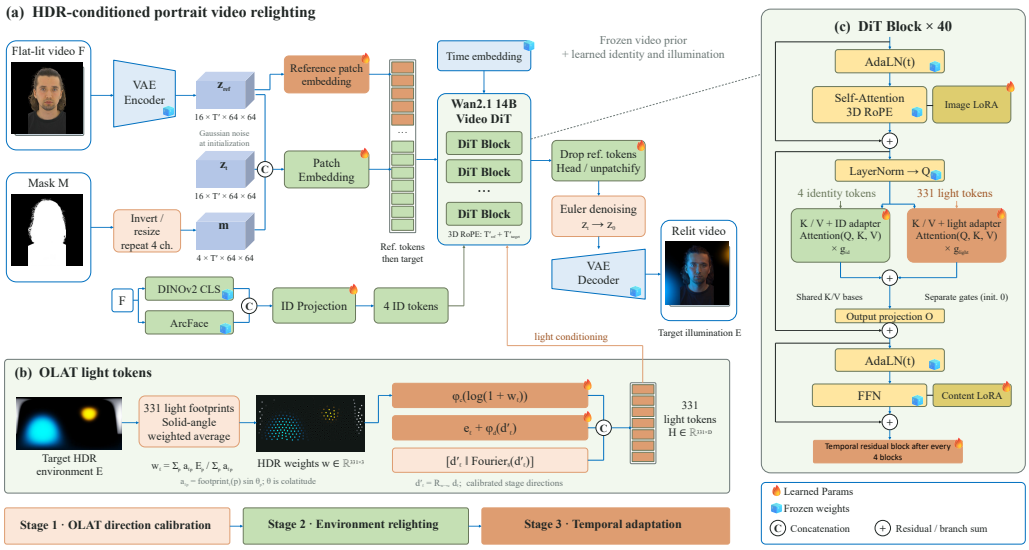

Figure 2: **LumiForge architecture.** Reference features and ORBIT tokens condition a frozen image relighter. Time-only residuals after blocks $4, 8, \ldots, 40$ provide temporal communication across 65 framewise latents; $T = 1$ bypasses these modules.

2025) captures dynamic lighting for subject-specific avatars; we use it for the joint-video predecessor and the delighting teacher. The final residual model preserves the image inference path and learns temporal corrections from VFHQ. Directional tokens and hybrid supervision motivate this design; our focus is their alignment with controlled image supervision and capability-preserving video adaptation.

3 Method

LumiForge connects three components: ILC supplies environment-conditioned real performances, ORBIT carries directional illumination and identity into the generator, and temporal residuals extend an image relighter to video. Figure 2 shows the final computation. The image and joint-video models serve as predecessors and experimental controls.

3.1 Task and Lighting Representation

Given flat-lit frames $\mathbf{F} = \{F_t\}_{t=1}^{T}$, foreground masks $\mathbf{M}$, and a target HDR environment E, we learn $\hat{\mathbf{Y}} = \mathcal{R}_\theta(\mathbf{F}, \mathbf{M}, E)$. The target edit preserves identity, expression, and motion. We assume distant illumination approximately constant over a short clip. For a fixed scene state and linear sensor response, light transport has the form

$$Y_t^{\text{lin}}(E) = \int_{\mathbb{S}^2} E(\omega)\mathcal{T}(F_t, \omega)d\omega \approx \sum_{i=1}^{K} \mathbf{w}_i(E) \odot \mathbf{B}_i(F_t), \tag{1}$$

where $\mathbf{B}_i$ is a directional RGB response, with basis normalization absorbed into it. OLAT captures measure these responses (Debevec et al., 2000). This physical relation motivates the targets and conditioning; the diffusion generator predicts the relit frame directly. It does not explicitly output K basis images, and its tone-mapped output is not constrained to obey linear superposition.

3.2 Inverse Lighting Curation

A wild portrait performance $\mathbf{V}$ observes the face under its source illumination, but lacks the conditions needed for forward relighting. ILC estimates them with an environment estimator $\mathcal{G}_{\text{env}}$ and a

video delighting model $\mathcal{G}_{\text{del}}$:

$$\hat{E}^{\text{src}} = \mathcal{G}_{\text{env}}(\mathbf{V}), \quad \hat{\mathbf{F}} = \mathcal{G}_{\text{del}}(\mathbf{V}), \quad \mathcal{C}_{\text{ILC}}(\mathbf{V}) = (\hat{\mathbf{F}}, \hat{E}^{\text{src}}, \mathbf{V}). \tag{2}$$

The resulting reconstruction task is $\mathcal{R}_\theta(\hat{\mathbf{F}}, \mathbf{M}, \hat{E}^{\text{src}}) \to \mathbf{V}$. The original performance remains the target. VFHQ yields 15,191 complete 65-frame LumiWild clips: 987,415 frame occurrences, which need not be unique source frames or independent examples.

Why inverse estimates are useful. The environment estimator combines visible scene context with portrait reference features. A separate Wan2.1 delighting model extends a frozen image prior with a BecomingLit-trained video adapter, without an HDRI input. Its prediction provides a canonical appearance approximation, rather than measured albedo. Illumination recovery is likewise ambiguous with reflectance, exposure, and camera processing. Residual input shading may permit copying under source-light reconstruction. We therefore combine controlled image supervision with real motion and evaluate changed-light behavior (Section 4). The estimator is evaluated independently through fixed OLAT rendering in Appendix A.

3.3 ORBIT: Directional Tokens and Selective Control

ORBIT (**O**LAT-**R**esolved **B**i-branch **I**dentity–lighting **T**okens) maintains $K = 331$ slots aligned with the directional regions of the image OLAT basis. Let a_{ip} be the solid-angle weight of panorama pixel p in region i. The light tokens are

$$\mathbf{w}_i = \frac{\sum_p a_{ip} E_p}{\sum_p a_{ip}}, \qquad \mathbf{h}_i = \Phi_{\text{light}}(\mathbf{w}_i, \mathbf{d}_i), \qquad \mathbf{H} = [\mathbf{h}_1, \ldots, \mathbf{h}_K] \in \mathbb{R}^{K \times D}. \tag{3}$$

The encoder combines log-compressed RGB radiance, camera-coordinate direction $\mathbf{d}_i$, Fourier direction features, and a learned slot embedding. Each coefficient is a regional radiance average, rather than integrated energy. A shared index links measured image responses, natural HDRIs, and video conditions, allowing edits to address a specific group of light slots. Pooling preserves between-region structure but discards subregion angular detail, so very narrow sources need not retain their sharp-shadow response.

Identity and local appearance are supplied for each reference frame using DINOv2 (Oquab et al., 2024) and ArcFace (Deng et al., 2022):

$$\mathbf{S}_t = \Phi_{\text{id}}(f_{\text{DINO}}(F_t), f_{\text{Arc}}(F_t)) \in \mathbb{R}^{4 \times D}, \qquad \mathbf{X}'_t = \mathbf{X}_t + g_{\text{light}} \mathcal{A}_{\text{light}}(\mathbf{X}_t, \mathbf{H}) + g_{\text{id}} \mathcal{A}_{\text{id}}(\mathbf{X}_t, \mathbf{S}_t). \tag{4}$$

Separate learned gates distinguish incident-light and subject conditioning without guaranteeing statistical disentanglement. Changing the environment recomputes light tokens while retaining reference frames, masks, and identity features.

Token-level lighting edits. The directional slots make illumination directly addressable. For a selected set $\mathcal{S}$, setting $\mathbf{h}_i = \mathbf{0}$ for $i \in \mathcal{S}$ suppresses encoded light components while retaining the other slots, reference frames, and identity features. Figure 3 shows complementary red/blue interventions and an inactive-region control. This edit acts after encoding; it differs from setting physical radiance coefficients to zero.

3.4 Preserving the Image Path during Temporal Adaptation

We instantiate LumiForge on Wan2.1-Fun-14B-Control (Team Wan et al., 2025), with hidden width $D = 5120$. The Image Relighter encodes and decodes each frame independently. The Joint Video Relighter instead compresses 65 frames into 17 latent times and trains rank-128 LoRA in Q/K/V/O of 40 joint space–time blocks. Although the image weights are frozen, the added projections alter the features querying light attention. Frozen parameters therefore do not by themselves preserve the image relighting function.

The final Temporal Residual Relighter restores the image computation: 65 independent frame latents, each frame's own reference and identity conditions, and the original reference/target token arrangement. Image embeddings, adapters, spatial blocks, and the output head remain frozen. We insert a trainable temporal module after blocks $4, 8, \ldots, 40$, without inheriting the joint-video LoRA. This assigns spatial relighting to the image path and cross-frame communication to explicit residual operators.

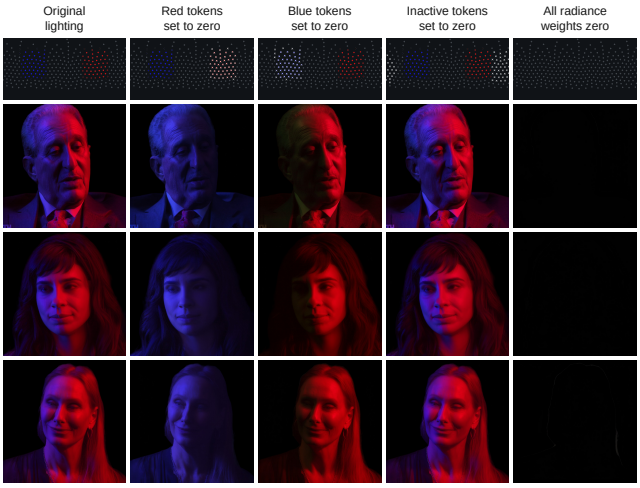

Figure 3: **Token-level lighting control.** Inputs and inference randomness are fixed. Maps mark the intervened slots: zeroing red or blue tokens selectively suppresses that colored contribution; inactive-token edits have a smaller effect. The last column zeros radiance coefficients instead. All-off suppression is imperfect. These interventions expose the representation's selective editing capability.

Time-only correction. For target-token features $X_{t,p} \in \mathbb{R}^{5120}$ at fixed patch p, let $Z_{t,p} = W_d \,\mathrm{LN}(X_{t,p})$, with affine-free normalization and $W_d \in \mathbb{R}^{256\times5120}$. A projection forms Q/K/V; four 64-dimensional heads attend over all T times at each patch, with temporal RoPE on Q/K. After concatenating heads, denote the attention result by A and its values by V. We update

$$U_{t,p} = A_{t,p} - V_{t,p}, \qquad \widetilde{U}_{t,p} = U_{t,p} - \frac{1}{T}\sum_{s=1}^{T} U_{s,p}, \qquad X'_{t,p} = X_{t,p} + W_u \widetilde{U}_{t,p}. \tag{5}$$

Here $W_u \in \mathbb{R}^{5120\times256}$. Subtracting the current value isolates the correction induced by temporal communication; centering removes a shared feature offset. This is a feature constraint, not conservation of final RGB or exposure. Only target tokens are directly updated at the insertion; later spatial blocks propagate the correction. Fixed patches need not track the same facial surface during motion, so the operator's benefit is empirical.

Single-frame recovery. The output projection starts at zero, making each new module an identity at initialization. An explicit bypass when $T = 1$ additionally guarantees, after training,

$$\boxed{f^{T=1}_{\text{residual}} = f^{T=1}_{\text{image}}} \tag{6}$$

under matched inputs, conditions, diffusion steps, precision, and computation. For $T > 1$, the modules intentionally alter features to use other frames; the bypass does not imply equality to independent image predictions. The ten modules add 28.18M trainable parameters, compared with 209.72M for joint-video LoRA. This is a reduction in adaptation capacity, not a measured runtime or memory saving.

3.5 Learning Lighting Responses and Temporal Corrections

Controlled image prior. FaceOLAT training first uses sparse white-light combinations to learn directional responses, then natural HDRIs to learn environment-conditioned appearance. We use the backbone's flow-matching objective for noisy target latent $\mathbf{Z}_\tau$, velocity target $\mathbf{v}^*_\tau$, and conditions $\mathbf{c} = (\mathbf{F}, \mathbf{M}, \mathbf{H}, \mathbf{S})$:

$$\mathcal{L}_{\text{FM}} = \mathbb{E}_{\tau,\mathbf{Z}_\tau}\left[\|\mathbf{v}_\theta(\mathbf{Z}_\tau, \tau; \mathbf{c}) - \mathbf{v}^*_\tau\|_2^2\right]. \tag{7}$$

The joint predecessor adapts on BecomingLit followed by VFHQ; a historical control adapts directly to VFHQ. Both use the studio-trained delighting teacher. These configurations establish the comparison for final residual adaptation, which starts from the image model rather than the joint weights.

Complementary residual-training branches. We sample 50% VFHQ reconstruction micro-batches, 25% FaceOLAT HDR/basis supervision, and 25% novel-HDR matching to the frozen image teacher. The supervised branches optimize $\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{FM}} + 0.1\mathcal{L}_{\Delta t}$, where $\mathcal{L}_{\Delta t}$ compares predicted

and reference clean-latent frame differences. It supervises the *change required by the target*, rather than rewarding a sequence simply for being static. FaceOLAT images repeat along time and provide lighting targets; VFHQ provides actual motion.

The teacher branch exposes those motions to new HDR conditions. Under identical noise and conditioning, it compares residual-model velocity v_r with the frozen framewise teacher v_i:

$$\mathcal{L}_{\text{keep}} = 0.25\,\text{MSE}_{\text{mask}}(v_r, v_i) + \left\|\text{AvgPool}_{4\times4}(v_r - v_i)\right\|_2^2. \tag{8}$$

The pooled term limits large-scale illumination drift, while supervised video targets guide temporal changes. Teacher predictions regularize new-light behavior and are not independent ground truth. Training and inference both use 65 frames. Together, the branches connect controlled lighting, real motion, and retention of novel-light responses; Appendix F provides configuration details and the three-model visual comparison.

4 Experiments

We evaluate three questions: **Relight**, end-to-end relighting fidelity and temporal stability; **Recover**, the quality of ILC supervision; and **Resolve**, directional control through ORBIT.

4.1 Experimental Setup

Datasets. We evaluate controlled relighting on FaceOLAT and BecomingLit, and in-the-wild generalization on held-out portrait videos. Our paired FaceOLAT benchmark contains 480 relighting pairs from 12 subjects and 40 target illuminations. The BecomingLit benchmark contains 20 environment-conditioned video references synthesized from measured OLAT sequences. For evaluating ILC independently of the final relighter, we use a subject- and environment-disjoint evaluation with 20 unseen subjects and 527 held-out HDR environments; training uses 100 subjects and 3,000 environments. Additional split details and complete evaluation protocols are provided in Appendix D.

Baselines. We compare LumiForge with representative image and video relighting methods, including IC-Light (Zhang et al., 2025), Light-X (Liu et al., 2025b), DiffusionRenderer (Liang et al., 2025), Neural-Gaffer (Jin et al., 2024), Lumen (Zeng et al., 2025), RelightVid (Fang et al., 2025), UniLumos (Liu et al., 2025a), TC-Light (Liu et al., 2025c), DPR (Zhou et al., 2019), and LCVD (Guo et al., 2025). Methods retain their native lighting interfaces. Our final model configuration is denoted as **LumiForge**; internal training variants are only used for ablations.

Metrics. We report foreground-normalized PSNR, SSIM, and LPIPS (Zhang et al., 2018) for appearance fidelity. For video, framewise scores are first averaged within each sequence and then across sequences. Temporal Δ-error and Warped Δ-error measure foreground temporal-change discrepancies against reference videos, without and with motion alignment, respectively (Appendix D.2).

4.2 Relight: Accurate Portrait Video Relighting

Figure 4 shows diverse identities, expressions, poses, and environments. Under a fixed target environment, LumiForge updates highlights, shading, and color with facial motion. Figure 5 tests *correct*, not merely plausible, lighting against paired references: LumiForge follows warm/cool FaceOLAT environments and BecomingLit color/intensity changes more faithfully than the compared methods.

As shown in Table 1, LumiForge reaches 27.28 dB and 0.0262 LPIPS on FaceOLAT. Its PSNR exceeds the strongest external baseline by 9.18 dB while also substantially reducing perceptual error. On BecomingLit, LumiForge reaches 25.38 dB and 0.1053 LPIPS, improving PSNR by 8.06 dB over the strongest external baseline on this metric. Together with the paired visual comparisons, these results show that the advantage comes from following the requested illumination rather than merely preserving the input appearance.

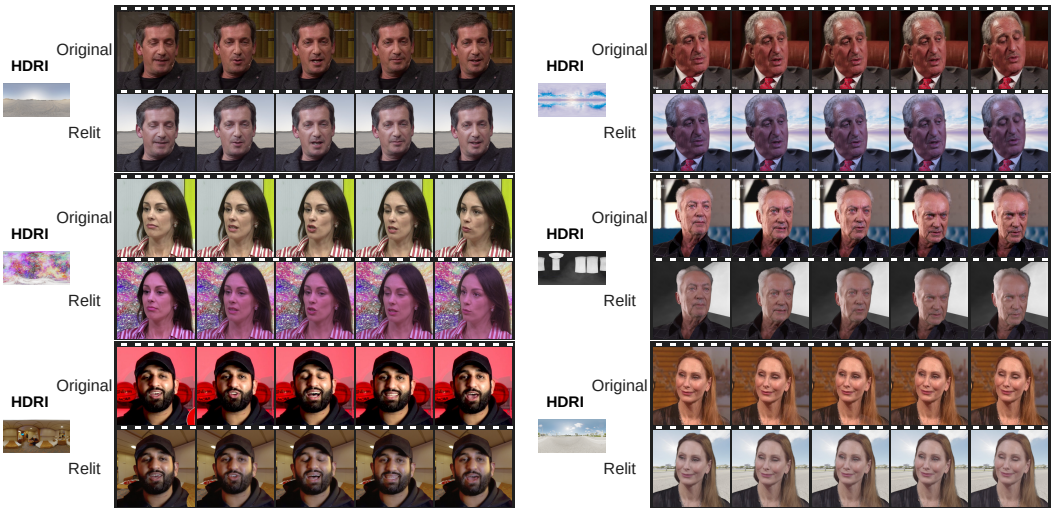

Figure 4: **Portrait video relighting across subjects and environments.** Five evenly spaced frames show the original and relit performances under each target environment, retaining expression and head motion.

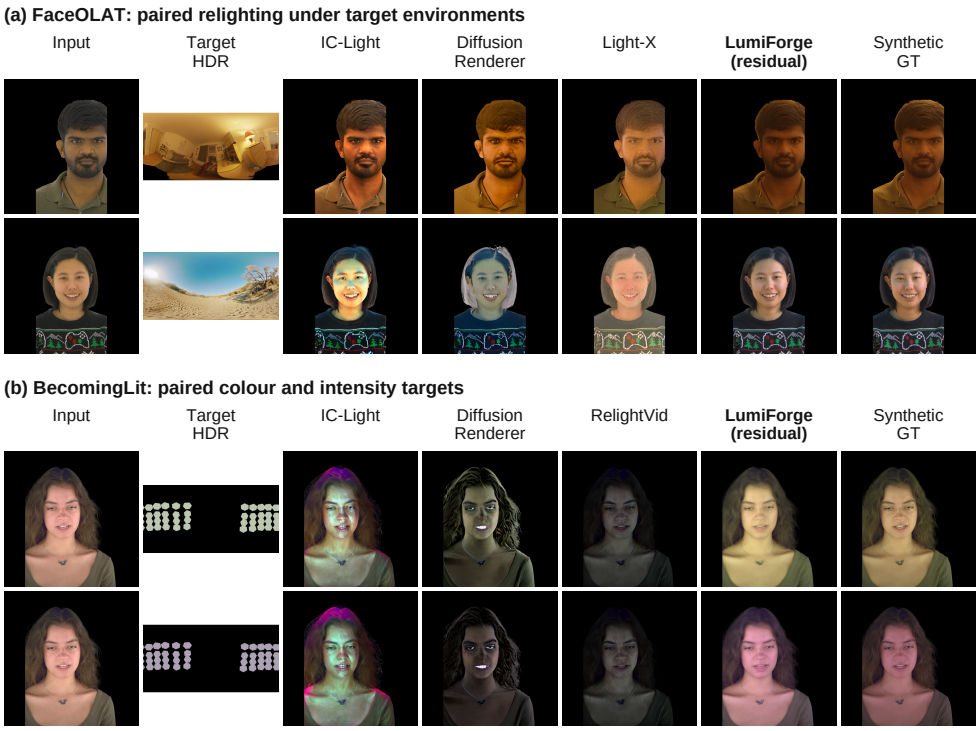

Figure 5: **Relighting fidelity against paired targets.** FaceOLAT evaluates target-environment relighting on held-out portraits, while BecomingLit evaluates controlled color and intensity changes in dynamic performances. Complete comparisons are provided in Appendix G.

Temporal stability. Table 2 measures agreement with reference temporal changes. LumiForge achieves the lowest errors among the evaluated models: 1.2834 Temporal Δ-error and 1.2804 Warped Δ-error, reductions of 21.0% and 8.8% over the Joint Video Relighter, respectively. Together with the framewise results in Table 1, these scores support improved temporal agreement and relighting fidelity; low temporal error alone does not establish accurate relighting.

Table 1: **Relighting fidelity against paired targets.** FaceOLAT contains 480 pairs and BecomingLit contains 20 video references. Bold denotes the best relighting model.

	FaceOLAT		BecomingLit	
Method	PSNR↑	LPIPS↓	PSNR↑	LPIPS↓
IC-Light	17.82	0.0862	15.66	0.1431
Light-X	15.62	0.0967	15.96	0.1427
DiffusionRenderer	18.10	0.0983	16.04	0.1534
Neural-Gaffer	14.75	0.1586	17.32	0.1500
Lumen	18.00	0.1154	17.01	0.1217
RelightVid	16.25	0.1006	14.91	0.1748
UniLumos	12.18	0.2320	10.08	0.3544
TC-Light	13.04	0.2219	8.48	0.3611
DPR	4.96	0.2312	6.32	0.2883
LCVD	16.16	0.1469	15.21	0.1792
LumiForge	**27.28**	**0.0262**	**25.38**	**0.1053**

Table 2: **Reference-based temporal errors.** 20 videos, 65 frames each; 0–255 RGB units. Lower is better; bold marks the best result.

Method	Temporal Δ-error↓	Warped Δ-error↓
Light-X	1.7942	1.5870
DiffusionRenderer	1.9640	1.8152
Image Relighter	1.6803	1.5441
Joint Video Relighter	1.6252	1.4044
LumiForge	**1.2834**	**1.2804**

Table 3: **Recovered illumination.** Estimates render through the same fixed portrait OLAT basis; the video relighter is not involved.

Method	Ang. Err.↓	PSNR↑	SSIM↑
DiffusionLight	5.4047	21.34	0.7989
StyleLight	6.2116	22.46	0.8293
LuxDiT	4.5583	25.03	0.8520
LumiForge	**2.0785**	**32.19**	**0.9489**

Lighting-response retention. Under matched target environments, Figure 6 shows that LumiForge retains the image prior's color and shading changes, while joint adaptation produces a more uniform appearance (mechanism in Appendix F).

4.3 RECOVER: DOES ILC RECOVER USEFUL LIGHTING SUPERVISION?

ILC uses recovered source illumination to turn real videos into relighting supervision. We isolate illumination recovery from video generation by rendering every estimator's output through the same measured portrait OLAT responses, with the light-transport basis fixed.

Table 3 shows that our estimator improves both illumination direction and the resulting facial appearance. Relative to LuxDiT, angular error decreases from 4.56 to 2.08 and portrait rendering PSNR increases from 25.03 to 32.19 dB. The recovered environment thus better reproduces the observed portrait's illumination under a fixed response, supporting the first step of ILC.

Recovered supervision in historical training controls. Figure 7 shows same-light reconstruction on 100 VFHQ clips: PSNR rises from 20.34 to 25.77 dB after direct wild-video adaptation and to 26.93 dB with studio-to-wild training. These controls evaluate recovered supervision separately from the final LumiForge model (protocols in Appendix E).

Complete panorama, sphere-rendering, and estimator ablations are provided in Appendix A.

4.4 RESOLVE: DOES ORBIT PRESERVE DIRECTIONAL CONTROL?

We test ORBIT's directional control by intervening on light-token groups with the portrait input and diffusion randomness fixed. In Figure 3, removing an active colored region's tokens selectively suppresses its contribution while retaining other regions. Complementary edits reverse the effect; inactive-region edits have smaller effects. Directional token identity therefore remains accessible during generation rather than acting as a global style code.



Figure 6: **Lighting-response retention during video adaptation.** Across target environments (columns), LumiForge retains the image prior's color and shading response, which joint adaptation attenuates in this example.

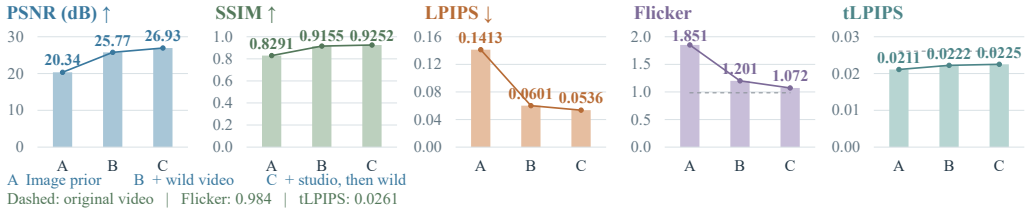

Figure 7: **Historical training controls on 100 VFHQ clips.** Same-light reconstruction and inter-frame change; see Appendix E for protocols.

Token-component ablations (direction encoding, Fourier features, and learned slots) appear in Appendix B; localized light-strength experiments appear in Appendix C. Appendix F details the temporal-adaptation architecture, single-frame retention mechanism, and parameter counts.

5 Conclusion

We presented **LumiForge** for precise portrait video relighting under explicit environment illumination. Our key idea is to *recover rather than capture*: **Inverse Lighting Curation (ILC)** turns in-the-wild portrait videos into direct IBL supervision and enables the construction of **LumiWild** at scale. We further *resolve rather than compress* illumination through **ORBIT**, which preserves directional lighting structure together with subject-dependent facial responses inside the video model. Extensive experiments demonstrate accurate environment following, strong generalization to natural performances, and fine-grained lighting control. Together, these results suggest that scalable recovered supervision and light-resolved conditioning provide a practical route toward high-fidelity and controllable portrait video relighting.

Author Contributions

Acknowledgments

AI use statement

(This section is **required** and does not count toward the page limit.)

In this work, we used generative AI tools for [tasks with required disclosure]. We have not used generative AI tools for [other tasks with required disclosure], and [the rest of the required disclosure tasks] are not applicable to this work. Additionally, we used generative AI tools for [tasks with recommended disclosure]. We have reviewed all AI-assisted work. [Elaborate.] We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

Ethics statement

(This section is **recommended** and does not count toward the page limit.)

If authors feel that their paper submission raises questions regarding the Code of Ethics, they are encouraged to include a paragraph of Ethics Statement (at the end of the main text before references) to address potential concerns where appropriate. This statement should not be more than 1 page.

Reproducibility statement

(This section is **recommended** and does not count toward the page limit.)

It is important that the work published in ICLR is reproducible. Authors are strongly encouraged to include a paragraph-long Reproducibility Statement at the end of the main text (before references) to discuss the efforts that have been made to ensure reproducibility. This paragraph should not itself describe details needed for reproducing the results, but rather reference the parts of the main paper, appendix, and supplemental materials that will help with reproducibility.

References

Ziqi Cai, Kaiwen Jiang, Shu-Yu Chen, Yu-Kun Lai, Hongbo Fu, Boxin Shi, and Lin Gao. Real-time 3d-aware portrait video relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6221–6231, 2024.

Sumit Chaturvedi, Mengwei Ren, Yannick Hold-Geoffroy, Jingyuan Liu, Julie Dorsey, and Zhixin Shu. SynthLight: Portrait Relighting with Diffusion Model by Learning to Re-render Synthetic Faces. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 369–379, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Chaturvedi_SynthLight_Portrait_Relighting_with_Diffusion_Model_by_Learning_to_Re-render_CVPR_2025_paper.html.

Paul E. Debevec, Tim Hawkins, Chris Tchou, Haarm-Pieter Duiker, Westley Sarokin, and Mark Sagar. Acquiring the reflectance field of a human face. In *Proceedings of the 27th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2000, New Orleans, LA, USA, July 23-28, 2000*, pp. 145–156. ACM, 2000. doi: 10.1145/344779.344855. URL https://doi.org/10.1145/344779.344855.

Jiankang Deng, Jia Guo, Jing Yang, Niannan Xue, Irene Kotsia, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10):5962–5979, 2022. doi: 10.1109/TPAMI.2021.3087709. URL https://doi.org/10.1109/TPAMI.2021.3087709.

Ye Fang, Zeyi Sun, Shangzhan Zhang, Tong Wu, Yinghao Xu, Pan Zhang, Jiaqi Wang, Gordon Wetzstein, and Dahua Lin. RelightVid: Temporal-Consistent Diffusion Model for Video Relighting, 2025. URL https://arxiv.org/abs/2501.16330. arXiv:2501.16330.

Mingtao Guo, Guanyu Xing, and Yanli Liu. High-fidelity relightable monocular portrait animation with lighting-controllable video diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 228–238, 2025.

Andrew Hou, Ze Zhang, Michel Sarkis, Ning Bi, Yiying Tong, and Xiaoming Liu. Towards high fidelity face relighting with realistic shadows. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14719–14728, 2021.

Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion. In *Advances in Neural Information Processing Systems*, 2024.

Hoon Kim, Minje Jang, Wonjun Yoon, Jisoo Lee, Donghyun Na, and Sanghyun Woo. 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 (CVPR)*, pp. 25096–25106, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Kim_SwitchLight_Co-design_of_Physics-driven_Architecture_and_Pre-training_Framework_for_Human_CVPR_2024_paper.html.

Ruofan Liang, Zan Gojcic, Huan Ling, Jacob Munkberg, Jon Hasselgren, Chih-Hao Lin, Jun Gao, Alexander Keller, Nandita Vijaykumar, Sanja Fidler, and Zian Wang. Diffusion Renderer: Neural Inverse and Forward Rendering with Video Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 26069–26080, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Liang_Diffusion_Renderer_Neural_Inverse_and_Forward_Rendering_with_Video_Diffusion_CVPR_2025_paper.html.

Min-Hui Lin, Mahesh Reddy, Guillaume Berger, Michel Sarkis, Fatih Porikli, and Ning Bi. Edgerelight360: Text-conditioned 360-degree hdr image generation for real-time on-device video portrait relighting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 831–840, 2024.

Pengwei Liu, Hangjie Yuan, Bo Dong, Jiazheng Xing, Jinwang Wang, Rui Zhao, Weihua Chen, and Fan Wang. UniLumos: Fast and Unified Image and Video Relighting with Physics-Plausible Feedback. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 82052–82080. Curran Associates, Inc., 2025a. doi: 10.52202/085713-2744. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/75f1a165c7561e028c41d42fa6286a76-Paper-Conference.pdf.

Tianqi Liu, Zhaoxi Chen, Zihao Huang, Shaocong Xu, Saining Zhang, Chongjie Ye, Bohan Li, Zhiguo Cao, Wei Li, Hao Zhao, and Ziwei Liu. Light-X: Generative 4D Video Rendering with Camera and Illumination Control, 2025b. URL https://arxiv.org/abs/2512.05115. arXiv:2512.05115.

Yang Liu, Chuanchen Luo, Zimo Tang, Yingyan Li, Yuran Yang, Yuanyong Ning, Lue Fan, Junran Peng, and Zhaoxiang Zhang. TC-Light: Temporally Coherent Generative Rendering for Realistic World Transfer, 2025c. URL https://arxiv.org/abs/2506.18904. arXiv:2506.18904.

Yiqun Mei, He Zhang, Xuaner Zhang, Jianming Zhang, Zhixin Shu, Yilin Wang, Zijun Wei, Shi Yan, HyunJoon Jung, and Vishal M. Patel. LightPainter: Interactive Portrait Relighting With Freehand Scribble. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 195–205, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Mei_LightPainter_Interactive_Portrait_Relighting_With_Freehand_Scribble_CVPR_2023_paper.html.

Yiqun Mei, Yu Zeng, He Zhang, Zhixin Shu, Xuaner Zhang, Sai Bi, Jianming Zhang, HyunJoon Jung, and Vishal M. Patel. Holo-relighting: Controllable volumetric portrait relighting from a single image. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4263–4273, 2024.

Yiqun Mei, Mingming He, Li Ma, Julien Philip, Wenqi Xian, David M George, Xueming Yu, Gabriel Dedic, Ahmet Levent Taşel, Ning Yu, Vishal M. Patel, and Paul Debevec. Lux Post Facto: Learning Portrait Performance Relighting with Conditional Video Diffusion and a Hybrid Dataset. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5510–5522, 2025. URL https://openaccess.thecvf.com/content/CVPR2025/html/Mei_Lux_Post_Facto_Learning_Portrait_Performance_Relighting_with_Conditional_Video_CVPR_2025_paper.html.

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

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning Robust Visual Features without Supervision. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Mallikarjun B R, Ayush Tewari, Tae-Hyun Oh, Tim Weyrich, Bernd Bickel, Hans-Peter Seidel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, and Christian Theobalt. Monocular reconstruction of neural face reflectance fields. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4791–4800, 2021.

Pramod Rao, Abhimitra Meka, Xilong Zhou, Gereon Fox, Mallikarjun B R, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Thabo Beeler, Mohamed Elgharib, Marc Habermann, and Christian Theobalt. 3DPR: Single Image 3D Portrait Relighting with Generative Priors. In *ACM SIGGRAPH ASIA 2025 Conference Proceedings*, 2025. URL https://vcai.mpi-inf.mpg.de/projects/3dpr/.

Jonathan Schmidt, Simon Giebenhain, and Matthias Niessner. BecomingLit: Relightable Gaussian Avatars with Hybrid Neural Shading. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 89394–89417. Curran Associates, Inc., 2025. doi: 10.52202/085713-2991. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/811449856c80b13a1818ea3d777961eb-Paper-Conference.pdf.

Soumyadip Sengupta, Angjoo Kanazawa, Carlos D. Castillo, and David Jacobs. Sfsnet: Learning shape, reflectance and illuminance of faces in the wild, 2018. URL https://arxiv.org/abs/1712.01261.

Tiancheng Sun, Kai-En Lin, Sai Bi, Zexiang Xu, and Ravi Ramamoorthi. Nelf: Neural light-transport field for portrait view synthesis and relighting. In *Eurographics Symposium on Rendering*, 2021.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, Jianyuan Zeng, Jiayu Wang, Jingfeng Zhang, Jingren Zhou, Jinkai Wang, Jixuan Chen, Kai Zhu, Kang Zhao, Keyu Yan, Lianghua Huang, Mengyang Feng, Ningyi Zhang, Pandeng Li, Pingyu Wu, Ruihang Chu, Ruili Feng, Shiwei Zhang, Siyang Sun, Tao Fang, Tianxing Wang, Tianyi Gui, Tingyu Weng, Tong Shen, Wei Lin, Wei Wang, Wei Wang, Wenmeng Zhou, Wente Wang, Wenting Shen, Wenyuan Yu, Xianzhong Shi, Xiaoming Huang, Xin Xu, Yan Kou, Yangyu Lv, Yifei Li, Yijing Liu, Yiming Wang, Yingya Zhang, Yitong Huang, Yong Li, You Wu, Yu Liu, Yulin Pan, Yun Zheng, Yuntao Hong, Yupeng Shi, Yutong Feng, Zeyinzi Jiang, Zhen Han, Zhi-Fan Wu, and Ziyu Liu. Wan: Open and Advanced Large-Scale Video Generative Models, 2025. URL https://arxiv.org/abs/2503.20314. arXiv:2503.20314.

Junying Wang, Jingyuan Liu, Xin Sun, Krishna Kumar Singh, Zhixin Shu, He Zhang, Jimei Yang, Nanxuan Zhao, Tuanfeng Y. Wang, Simon S. Chen, Ulrich Neumann, and Jae Shin Yoon. Comprehensive relighting: Generalizable and consistent monocular human relighting and harmonization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 380–390, 2025.

Youjia Wang, Taotao Zhou, Minzhang Li, Teng Xu, Minye Wu, Lan Xu, and Jingyi Yu. Neural relighting and expression transfer on video portraits. *arXiv preprint arXiv:2107.14735*, 2021.

Liangbin Xie, Xintao Wang, Honglun Zhang, Chao Dong, and Ying Shan. VFHQ: A High-Quality Dataset and Benchmark for Video Face Super-Resolution. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 656–665. IEEE, 2022. doi: 10.1109/cvprw56347.2022.00081. URL https://doi.org/10.1109/cvprw56347.2022.00081.

Yu-Ying Yeh, Koki Nagano, Sameh Khamis, Jan Kautz, Ming-Yu Liu, and Ting-Chun Wang. Learning to relight portrait images via a virtual light stage and synthetic-to-real adaptation. *ACM Transactions on Graphics*, 41(6):1–21, November 2022. ISSN 1557-7368. doi: 10.1145/3550454.3555442. URL http://dx.doi.org/10.1145/3550454.3555442.

Jianshu Zeng, Yuxuan Liu, Yutong Feng, Chenxuan Miao, Zixiang Gao, Jiwang Qu, Jianzhang Zhang, Bin Wang, and Kun Yuan. Lumen: Consistent video relighting and harmonious background replacement with video generative models, 2025. URL https://arxiv.org/abs/2508.12945.

Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling In-the-Wild Training for Diffusion-based Illumination Harmonization and Editing by Imposing Consistent Light Transport. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=u1cQYxRI1H.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018*, pp. 586–595. Computer Vision Foundation / IEEE Computer Society, 2018. doi: 10.1109/CVPR.2018.00068. URL http://openaccess.thecvf.com/content_cvpr_2018/html/Zhang_The_Unreasonable_Effectiveness_CVPR_2018_paper.html.

Hao Zhou, Sunil Hadap, Kalyan Sunkavalli, and David W. Jacobs. Deep Single-Image Portrait Relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 7194–7202, 2019. URL https://openaccess.thecvf.com/content_ICCV_2019/html/Zhou_Deep_Single-Image_Portrait_Relighting_ICCV_2019_paper.html.

Yujie Zhou, Jiazi Bu, Pengyang Ling, Pan Zhang, Tong Wu, Qidong Huang, Jinsong Li, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Anyi Rao, Jiaqi Wang, and Li Niu. Light-A-Video: Training-free Video Relighting via Progressive Light Fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13315–13325, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhou_Light-A-Video_Training-free_Video_Relighting_via_Progressive_Light_Fusion_ICCV_2025_paper.html.

A Environment Estimation for Inverse Lighting Curation

This section evaluates the illumination recovered by Inverse Lighting Curation (ILC) independently of the final video relighter. The central question is whether the estimated environment produces the correct facial appearance when applied to a fixed, measured light-transport basis.

A.1 Evaluation Protocol

Our environment estimator combines visible scene context with portrait-reference features and is trained with rendering-based preference alignment. Training uses 100 subjects and 3,000 environments, while evaluation uses disjoint sets of 20 subjects and 527 environments. We compare against DiffusionLight, StyleLight, and LuxDiT.

We evaluate estimated illumination at three levels: (i) direct HDR panorama agreement, (ii) rendering on diffuse, matte, and mirror spheres, and (iii) rendering through a fixed measured portrait OLAT basis. For the portrait evaluation, all estimated environments illuminate the same basis $B = \{B_i\}$:

$$\mathcal{L}_{\mathrm{OLAT}}(\hat{E}; B) = \sum_i w_i(\hat{E}) \odot B_i. \tag{9}$$

Holding B fixed isolates the effect of the estimated illumination from the LumiForge video generator.

A.2 Full Environment-Estimation Results

Table 4: **Environment-estimator evaluation.** Metrics compare HDR panoramas and portrait OLAT renders under estimated illumination. LumiForge denotes its ILC environment estimator here; portrait rendering uses the same measured basis for all methods.

Method	si-RMSE↓	Ang. Err.↓	N-RMSE↓	PSNR↑	SSIM↑
(a) HDR panorama					
DiffusionLight	19.1128	8.7134	2.7444	9.2396	0.4945
StyleLight	19.0760	11.3186	2.7463	9.0099	0.3907
LuxDiT	18.7746	8.2269	2.9873	8.7136	0.5765
LumiForge (cross-attn.)	14.0573	4.2870	2.5023	12.3162	0.7107
LumiForge (w/o alignment)	**12.9084**	3.8918	2.4056	12.8669	0.7495
LumiForge	13.0174	**3.8420**	**2.3274**	**13.3969**	**0.7523**
(b) Portrait OLAT rendering					
DiffusionLight	0.1239	5.4047	0.1609	21.3372	0.7989
StyleLight	0.1102	6.2116	0.1443	22.4622	0.8293
LuxDiT	0.0806	4.5583	0.1163	25.0298	0.8520
LumiForge (cross-attn.)	0.0663	2.8826	0.0778	29.7832	0.9175
LumiForge (w/o alignment)	0.0429	2.0900	0.0588	31.8476	0.9482
LumiForge	**0.0415**	**2.0785**	**0.0580**	**32.1927**	**0.9489**

The rendering-based evaluation is particularly relevant to ILC. Relative to LuxDiT, our estimator reduces portrait angular error from 4.56 to 2.08 and increases portrait-rendering PSNR from 25.03 to 32.19 dB. These gains indicate that the recovered environment better reproduces the illumination effect observed on the portrait, which is precisely the condition needed for forward relighting supervision.

A.3 MATERIAL RESPONSE

Table 5: **Sphere rendering under estimated illumination.** Lower is better for all metrics.

	si-RMSE			Angular error			Normalized RMSE		
Method	Diff.	Matte	Mirror	Diff.	Matte	Mirror	Diff.	Matte	Mirror
DiffusionLight	0.1629	0.3992	0.6529	2.8563	3.7001	6.2559	0.2503	0.3884	0.4639
StyleLight	0.1613	0.3734	0.6140	3.9444	5.1178	8.7320	0.2519	0.4119	0.4956
LuxDiT	0.1342	0.3132	0.5591	2.6427	3.6097	6.0421	0.2101	0.3097	0.4255
LumiForge (cross-attn.)	0.0736	0.1834	0.3803	1.3900	1.8030	4.1779	0.1147	0.1745	0.2529
LumiForge (w/o alignment)	0.0508	**0.1424**	**0.3300**	1.0438	1.4655	**3.8415**	0.0855	**0.1387**	**0.2193**
LumiForge	**0.0506**	0.1437	0.3324	**1.0167**	**1.4535**	3.8487	**0.0836**	0.1390	0.2218

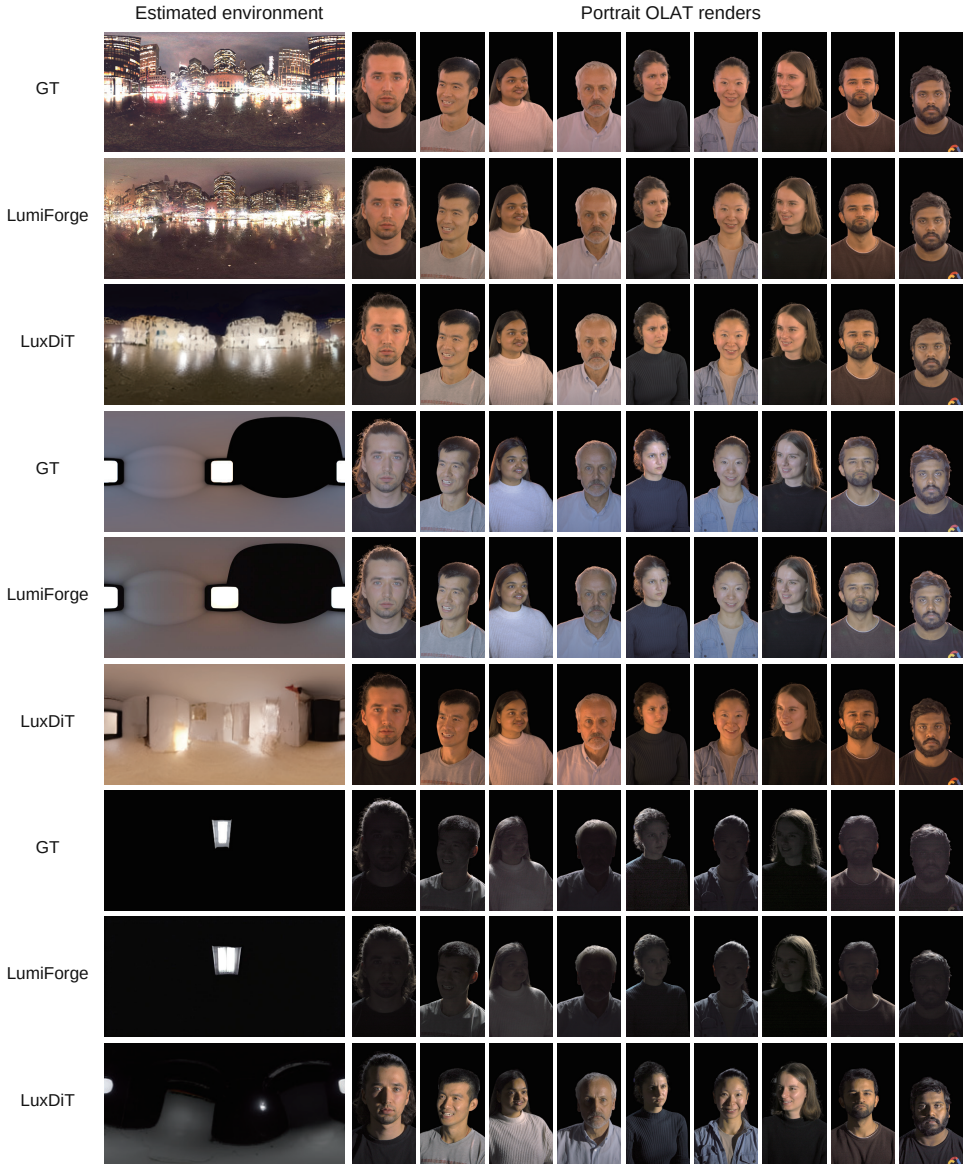

Figure 8: **Recovered illumination evaluated through a fixed portrait response.** Ground truth, the LumiForge ILC estimator, and LuxDiT illuminate the same measured OLAT basis. Shading differences reflect recovered illumination; the video relighter is not involved.

B Additional Analysis of ORBIT

The main paper evaluates selective directional control through ORBIT. Here we study the internal light-token representation in greater detail.

B.1 Light-Token Encoding Components

Table 6: **Controlled ORBIT encoding study.** Six unseen subjects, 12 unseen HDRIs, and two inference seeds are used for each variant. All variants share the same initialization and training budget.

Light encoding	FG PSNR↑	SSIM↑	LPIPS↓
Full light encoding	**21.21**	0.9284	**0.0573**
Without learned slot	21.20	**0.9286**	0.0574
Without explicit direction	21.07	0.9275	0.0578
Without Fourier features	21.17	0.9282	0.0577
RGB + direction	21.05	0.9276	0.0581

Removing explicit direction causes the largest PSNR decrease among the individual component removals, while removing learned slots or Fourier features has smaller effects. These comparisons probe encoding choices at fixed $K = 331$; they do not isolate directional tokenization from global environment conditioning.

B.2 Held-Out Light Combinations

Table 7: **Generalization to held-out multi-light combinations.** Twenty untrained combinations are evaluated with two inference seeds.

Light encoding	FG PSNR↑	SSIM↑	LPIPS↓
Full light encoding	17.43	0.8176	0.2025
Without learned slot	17.80	0.8194	0.2065
Without explicit direction	16.92	0.8139	0.2021
Without Fourier features	17.65	0.8155	0.2095
RGB + direction	16.90	0.8146	0.2037

This probe complements the controlled HDRI evaluation by testing unseen combinations of directional responses.

C Localized Lighting Control

ORBIT exposes individual directional components to the video model. We further test whether these components can be edited locally at inference time.

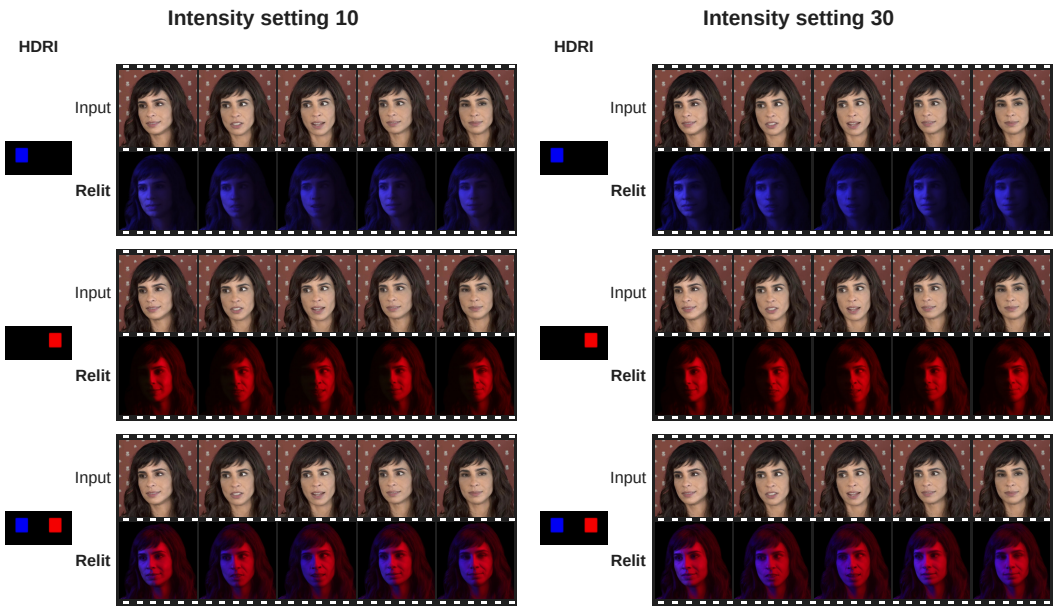

Figure 9: **Localized light activation and strength.** Rows activate two colored lighting regions separately and jointly, while columns vary their requested strength. Each example shows the target environment, flat-lit input, and relit video at five evenly spaced times.

When the performance is fixed, the two colored illumination components remain separately controllable and continue to combine when activated together. This complements the token-intervention experiment in the main paper: editing the radiance coefficients changes the requested illumination, while suppressing encoded tokens probes the representation used to realize that request.

Additional radiometric diagnostics. Historical image-domain probes yield log–log response slopes of 0.953 for differential-light gain and 0.028 for whole-environment scaling, indicating substantially weaker sensitivity to global intensity changes in this configuration.

D Evaluation Protocols

D.1 Paired-Reference Benchmarks

The main paired benchmark contains 480 FaceOLAT pairs from 12 subjects and 40 target illuminations per subject, together with 20 BecomingLit environment-reference videos. FaceOLAT references are synthesized from measured directional responses under target HDR environments. BecomingLit references are synthesized from its measured dynamic light responses under controlled color and intensity changes.

Foreground PSNR normalizes reconstruction error by foreground area. LPIPS uses identically masked full-frame images. Video metrics are first averaged over frames within each sequence and then averaged over sequences.

D.2 Reference-Based Temporal Metrics

The two metrics in Table 2 compare predicted temporal changes with those of the reference video. Let $\widehat{Y}_t$ and Y_t denote predicted and reference RGB frames normalized to $[0, 1]$, with pixel index p, color channel $c \in \{1, 2, 3\}$, and sequence length T. Define the framewise prediction error as

$$e_t(p, c) = \widehat{Y}_t(p, c) - Y_t(p, c). \tag{10}$$

Temporal-difference error. Let $U_t(p) \in \{0,1\}$ indicate the intersection of the eroded foreground masks of frames t and $t+1$, without motion alignment. We compute

$$\mathcal{E}_\Delta = \frac{255}{T-1} \sum_{t=1}^{T-1} \frac{\sum_{p,c} U_t(p)\,|e_{t+1}(p,c) - e_t(p,c)|}{3\sum_p U_t(p)}. \tag{11}$$

Here $e_{t+1} - e_t = (\widehat{Y}_{t+1} - \widehat{Y}_t) - (Y_{t+1} - Y_t)$. Thus the metric penalizes discrepancies in temporal change at fixed image coordinates, rather than penalizing reference motion or lighting changes by themselves.

Warped temporal-difference error. Let b_t denote backward optical flow from frame $t+1$ to frame t. Backward gathering uses bilinear interpolation:

$$\mathcal{W}_t(X_t)(p,c) = X_t\big(p + b_t(p), c\big). \tag{12}$$

Let $V_t(p) \in \{0,1\}$ indicate valid foreground correspondences after checking image bounds, warped foreground overlap, and forward–backward flow consistency. The motion-aligned error is

$$\mathcal{E}_{\text{warp}\Delta} = \frac{255}{T-1} \sum_{t=1}^{T-1} \frac{\sum_{p,c} V_t(p)\,|e_{t+1}(p,c) - \mathcal{W}_t(e_t)(p,c)|}{3\sum_p V_t(p)}. \tag{13}$$

Because the same bilinear warp is applied to prediction and reference, $e_{t+1} - \mathcal{W}_t(e_t)$ equals $[\widehat{Y}_{t+1} - \mathcal{W}_t(\widehat{Y}_t)] - [Y_{t+1} - \mathcal{W}_t(Y_t)]$. This compares temporal changes along valid estimated correspondences.

Evaluation protocol. Bidirectional Farneback flow is estimated on the pseudo-flat-lit input, independently of the evaluated predictions, and shared across methods. Foreground masks are thresholded at 0.5 and eroded with a 7×7 kernel. Each sequence contains $T = 65$ frames. Each metric first averages over its 64 transitions, then equally over the $N = 20$ environment-reference sequences:

$$\overline{\mathcal{E}}_\Delta = \frac{1}{N}\sum_{n=1}^{N} \mathcal{E}_\Delta^{(n)}, \qquad \overline{\mathcal{E}}_{\text{warp}\Delta} = \frac{1}{N}\sum_{n=1}^{N} \mathcal{E}_{\text{warp}\Delta}^{(n)}. \tag{14}$$

The factor 255 expresses errors in display-RGB code values and is omitted if input images already use the $[0, 255]$ scale. Lower values are better. We interpret both metrics alongside framewise fidelity and lighting response: temporally consistent prediction errors can still yield low scores, so temporal agreement alone does not establish accurate relighting.

D.3 Baseline Interfaces

Different baselines expose different lighting-control interfaces. For completeness, Table 8 summarizes the information provided to each system.

Table 8: **Lighting interfaces used by evaluated methods.**

Method	Lighting input
IC-Light	HDRI-derived background + lighting text
Light-X	HDR environment
DiffusionRenderer	HDR environment
RelightVid	HDRI-derived background + lighting text
TC-Light	Lighting text
UniLumos	HDRI-derived background + lighting text
LumiForge	HDR environment

E Historical Training Controls

The following experiments record earlier matched and partially matched training configurations. They are useful for understanding the role of controlled initialization and wild-video adaptation, but the final LumiForge results are reported separately in the main paper.

Table 9: **Historical same-light reconstruction on 100 VFHQ clips.** Each clip contains 65 frames at 512×512.

Method	Input	PSNR↑	SSIM↑	LPIPS↓	Flicker	tLPIPS
IC-Light (per frame) (Zhang et al., 2025)	B+T	17.93	0.8036	0.1400	4.328	0.0391
RelightVid (Fang et al., 2025)	B+T	17.48	0.8153	0.1460	3.129	0.0329
Light-A-Video (Zhou et al., 2025)	D+T	16.98	0.7854	0.1687	1.548	0.0245
TC-Light (Liu et al., 2025c)	T	13.84	0.6112	0.3271	0.976	0.0198
UniLumos (Liu et al., 2025a)	B+T	13.39	0.6472	0.3908	0.716	0.0136
DiffusionRenderer (Liang et al., 2025)	H	17.98	0.7387	0.1976	1.549	0.0268
Light-X (Liu et al., 2025b)	H	20.54	0.8076	0.1459	1.473	0.0247
Image Relighter (per frame)	H	20.34	0.8291	0.1413	1.851	0.0211
Joint Video Relighter (wild)	H	25.77	0.9155	0.0601	1.201	0.0222
Joint Video Relighter (studio → wild)	H	**26.93**	**0.9252**	**0.0536**	1.072	0.0225

Controlled studio initialization improves same-light reconstruction from 25.77 to 26.93 dB over direct wild adaptation, supporting the use of a controlled video stage before scaling to ILC-curated performances.

Historical metric definitions. The historical flicker statistic measures change in mask-weighted mean RGB, while tLPIPS is adjacent-frame LPIPS. They quantify inter-frame change rather than motion-compensated temporal error and are therefore retained only as historical diagnostics.

F Temporal Video Adaptation and Image-Prior Retention

The final LumiForge model uses the Temporal Residual Relighter, which preserves the frozen frame-wise image pathway and learns explicit temporal residual corrections.

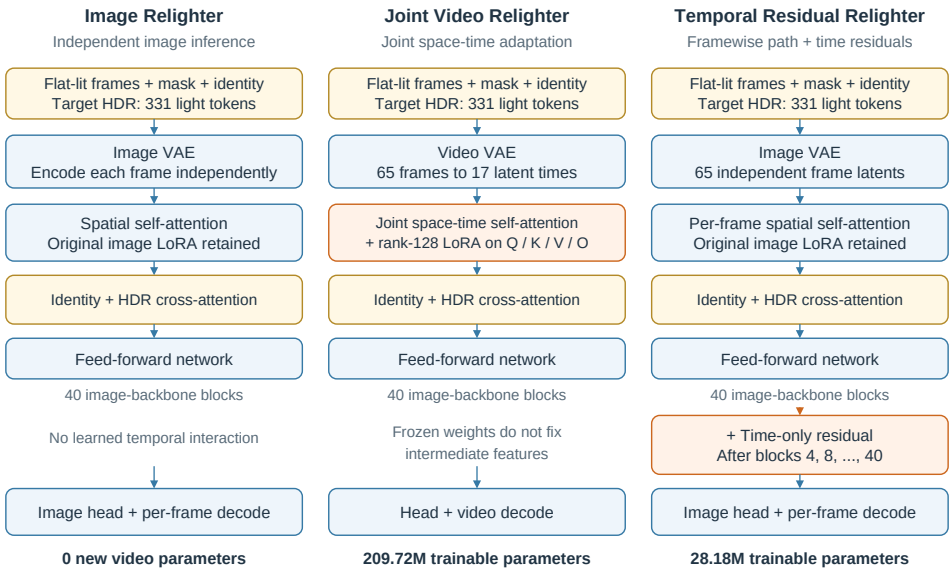

Figure 10: **From image relighting to video adaptation.** The image pathway is preserved while lightweight temporal residual modules introduce cross-frame interaction. Trainable modules are highlighted in orange.

Figure 6 in the main text compares the lighting responses of the image prior, joint adaptation, and the final residual model.

F.1 Residual Temporal Operator

The image relighter independently processes individual frames. A conventional joint video adaptation adds rank-128 LoRA modules to the Q/K/V/O projections of the 40 joint spatio-temporal blocks. Our residual configuration instead freezes the image VAE, image relighting modules, conditioning projections, and prediction head, and inserts ten lightweight temporal modules after blocks $4, 8, \ldots, 40$.

For target feature $H_{t,p} \in \mathbb{R}^{5120}$ at frame t and patch p, we first project to a 256-dimensional temporal space:

$$Z_{t,p} = W_d \,\mathrm{LN}(H_{t,p}), \qquad W_d \in \mathbb{R}^{256\times 5120}. \tag{15}$$

Four attention heads operate over the temporal dimension at each fixed patch using temporal RoPE. Let $A_{t,p}$ denote the attention output and $V_{t,p}$ the corresponding value feature. The residual update is

$$U_{t,p} = A_{t,p} - V_{t,p}, \qquad \widetilde{U}_{t,p} = U_{t,p} - \frac{1}{T}\sum_{s} U_{s,p}, \qquad H'_{t,p} = H_{t,p} + W_u \widetilde{U}_{t,p}. \tag{16}$$

Only target tokens are directly modified by these modules; reference tokens bypass the residual operator.

F.2 Single-Frame Retention

The residual branch is explicitly bypassed when $T = 1$, yielding

$$\boxed{T = 1: \qquad f^{T=1}_{\text{residual}} = f^{T=1}_{\text{image}}}. \tag{17}$$

This preserves the exact image inference path while allowing the model to introduce video-specific interactions for $T > 1$.

F.3 Training Objective and Capacity

Training mixes: 50% VFHQ reconstruction, 25% FaceOLAT HDR/basis supervision, and 25% novel-HDR teacher matching.

The supervised objective is

$$\mathcal{L}_{\text{sup}} = \mathcal{L}_{\text{FM}} + 0.1\mathcal{L}_{\Delta t}. \tag{18}$$

The image-prior retention objective is

$$\mathcal{L}_{\text{keep}} = 0.25\,\mathrm{MSE}_{\text{mask}}(v_{\text{res}}, v_{\text{img}}) + \left\|\mathrm{AvgPool}_{4\times 4}(v_{\text{res}} - v_{\text{img}})\right\|_2^2. \tag{19}$$

The ten temporal modules introduce 28.18M trainable parameters, compared with 209.72M parameters for the joint rank-128 video-LoRA adaptation.

G Extended Qualitative Comparisons

G.1 FaceOLAT Target-Environment Relighting

Figures 11 and 12 provide the complete FaceOLAT comparisons corresponding to the compact main-paper figure.

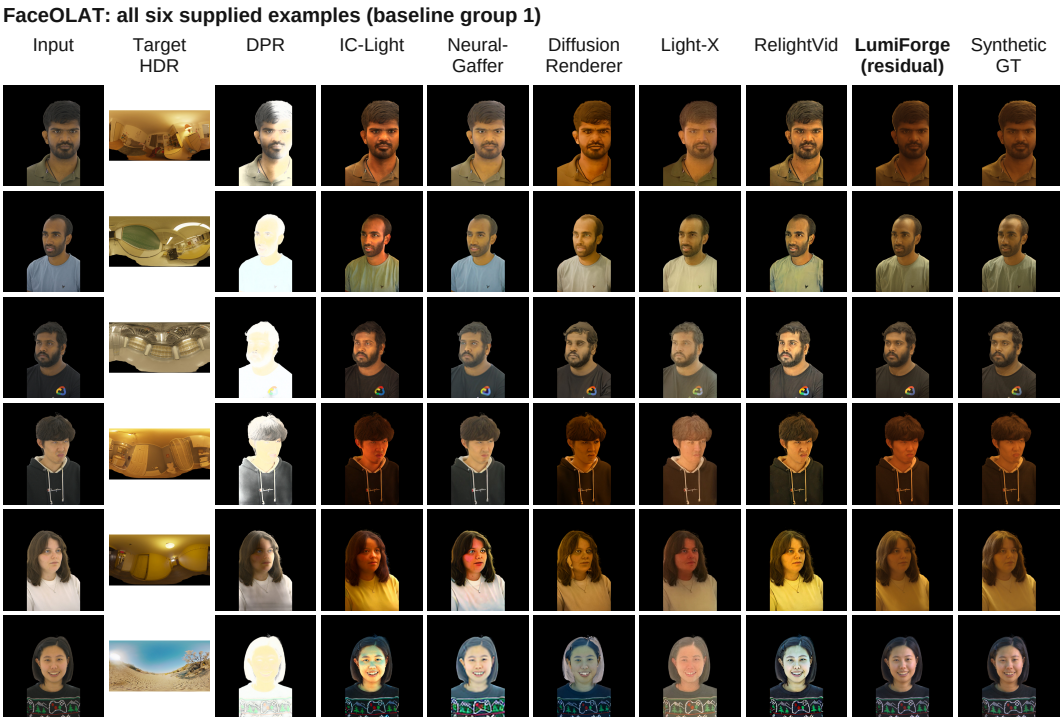

Figure 11: **FaceOLAT comparison, first baseline group.** Each example shows the target HDR environment, competing predictions, LumiForge, and the corresponding synthetic reference.

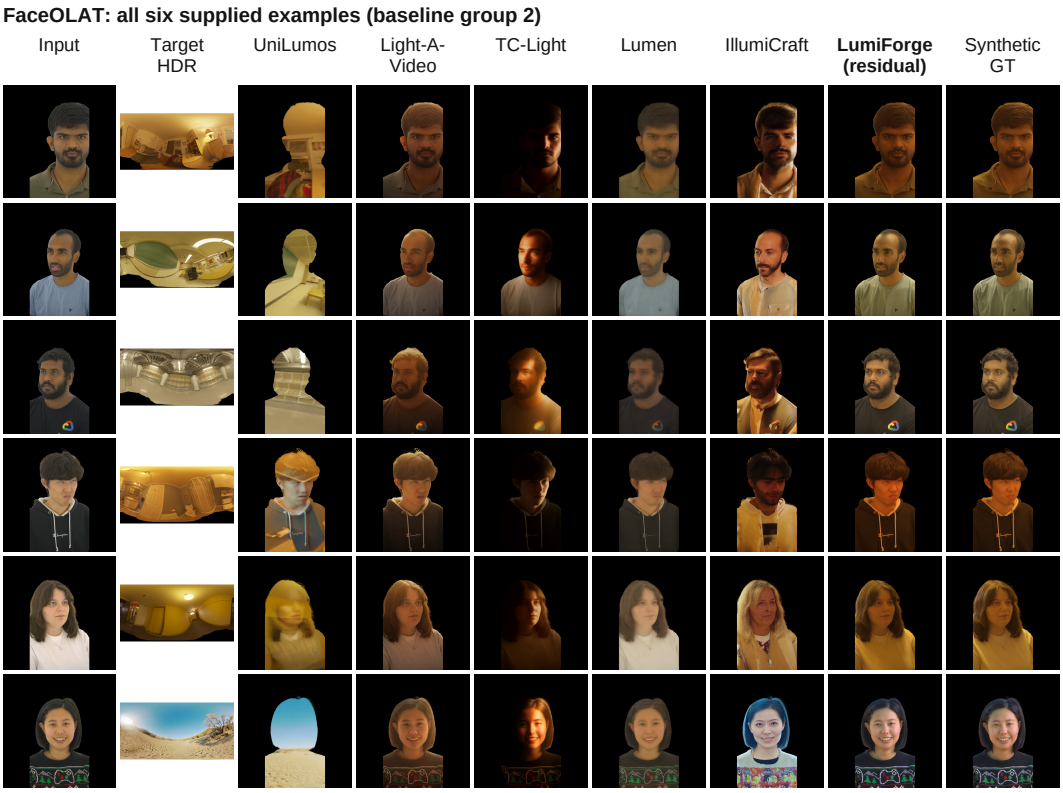

Figure 12: **FaceOLAT comparison, remaining baselines.** Lighting errors are visible as incorrect facial color, highlight placement, excess brightness, or attenuated shading variation.

G.2 BECOMINGLIT DYNAMIC RELIGHTING

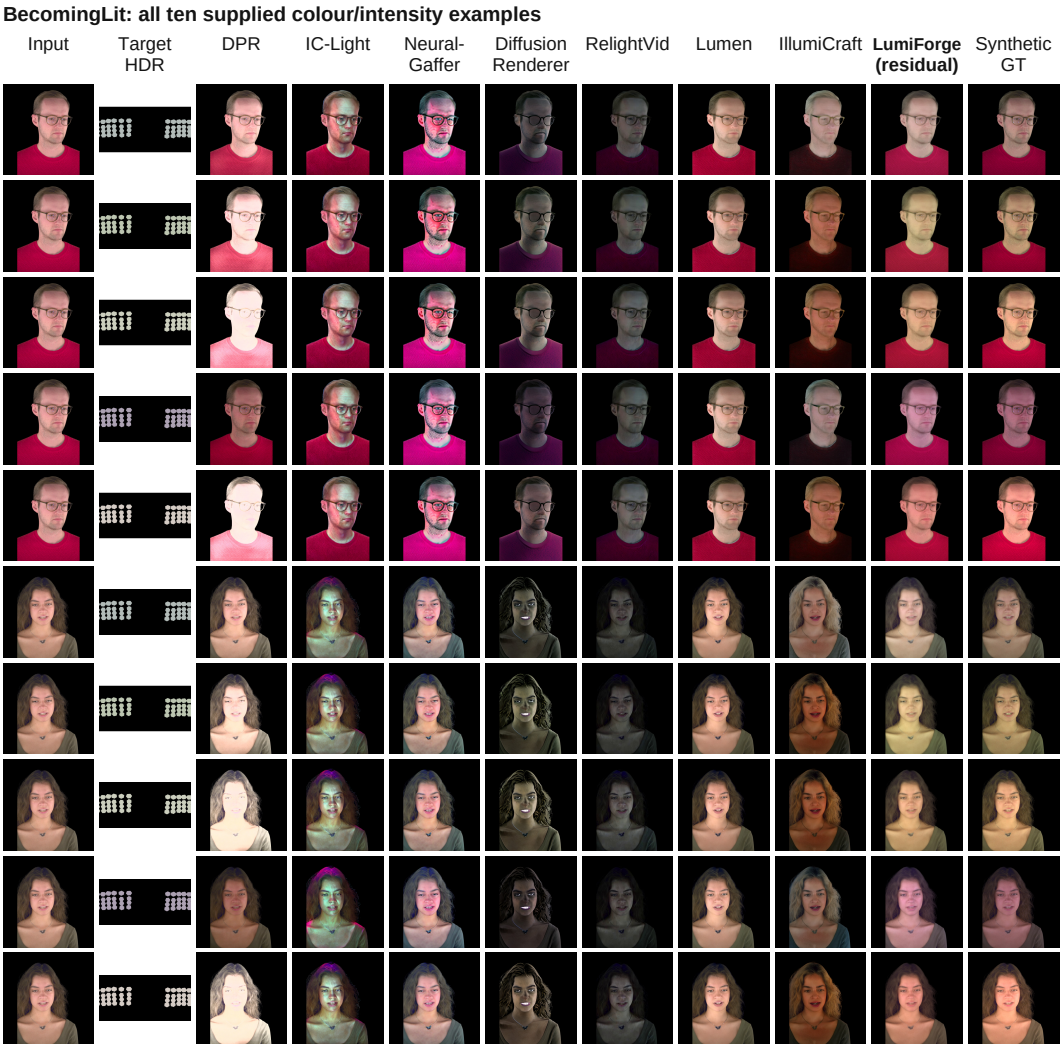

Figure 13: **BecomingLit comparison across controlled dynamic illumination conditions.** The targets synthesize controlled color and intensity changes using measured light-stage responses.

G.3 Cross-Subject In-the-Wild Gallery



Figure 14: **Cross-subject portrait video relighting.** Ten in-the-wild performances are relit under diverse target environments. Each example shows five evenly spaced frames.

H Implementation Details

H.1 LumiForge Backbone

Our primary implementation uses Wan2.1-Fun-14B-Control with hidden dimension $D = 5120$. The model receives a 16-channel noisy video latent, a 16-channel flat-lit reference latent, and four foreground-mask channels, yielding a 36-channel input. The joint-video predecessor compresses a 65-frame 512×512 sequence into 17 latent temporal positions. The final Temporal Residual Relighter instead processes 65 independently encoded frame latents, as described in Section 3.4.

H.2 ORBIT

The default ORBIT configuration uses $K = 331$ directional light tokens and $M = 4$ identity tokens. Each light token encodes log-compressed regional HDR radiance, camera-coordinate direction, Fourier directional features, and a learned slot embedding. Identity tokens are obtained from pretrained DINOv2 and ArcFace features.

H.3 Inverse Lighting Curation

The ILC environment estimator uses a Wan2.1-Fun-14B-Control latent diffusion transformer for masked HDR panorama completion. Visible background context and a visibility mask condition the noisy panorama latent, while portrait-reference tokens are concatenated with target tokens for self-attention. The video delighting teacher extends a frozen image delighting prior with a video-adaptation module trained on BecomingLit and does not require an HDR environment map as input.

H.4 Optimization

Table 10: **Training configuration.**

Setting	Value
Optimizer	AdamW
Learning rate	1e-5
Batch size	2
Gradient accumulation	8
Video length	65 frames
Resolution	512×512
Joint-model video LoRA rank	128
Diffusion / flow steps	50
CFG / guidance	1

I Limitations

ILC recovers illumination from appearance and therefore inherits the ambiguity between incident lighting, reflectance, exposure, and camera processing. Our environment evaluation measures this ambiguity through rendered appearance, while absolute radiometric calibration remains an open problem.

ORBIT represents illumination using a finite set of directional regions. Increasing the number of regions improves angular resolution at additional conditioning cost, while extremely narrow light sources may require a finer representation than the default $K = 331$ configuration.

Finally, LumiForge currently focuses on portrait-scale human performances. Generalization to broader body motion, extreme occlusion, rapidly changing illumination within a sequence, and substantially different capture domains remains future work.