SIMSTAGE: HARNESSING VIDEO DIFFUSION FOR CONSISTENT VIRTUAL LIGHT STAGE SYNTHESIS

Anonymous authors
Paper under double-blind review

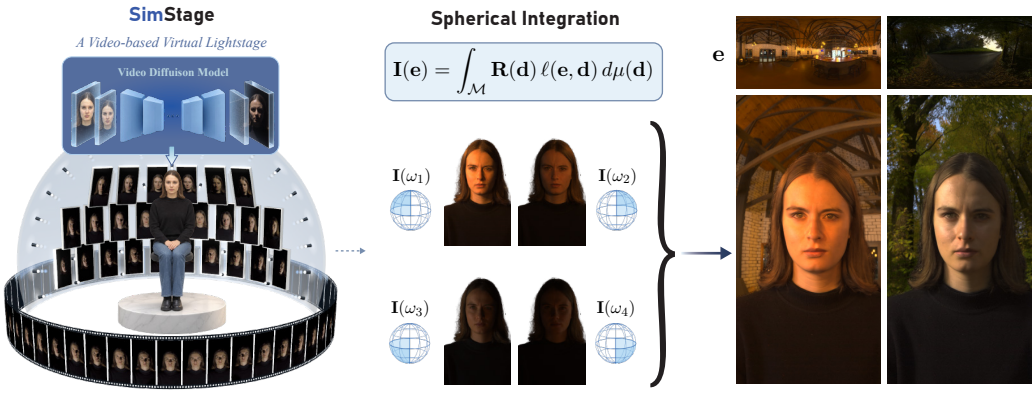

Figure 1: SimStage re-binds a video diffusion model's temporal axis to the manifold of light directions: a single portrait becomes a full 331-image HDR-linear OLAT reflectance field, and relighting under any environment is one linear combination.

ABSTRACT

One-light-at-a-time (OLAT) prediction provides explicit and reusable control for portrait relighting, but its promise depends on recovering an entire lighting basis with both *Basis Fidelity* and *Basis Coherence*: each response must preserve accurate HDR appearance, while all responses must remain structurally aligned so that details survive composition. We introduce **SimStage**, a *Full-Basis OLAT Generator* that directly predicts all 331 lighting responses within a single generation window. Our key insight is *Light-Axis Rebinding*: rather than treating OLATs as independent predictions, we redesign the joint-generation paradigm of video diffusion so that its sequence axis represents illumination instead of time. This is non-trivial—native temporal position modeling fails to reliably learn even substantially shorter OLAT windows. We therefore introduce *LiRoPE*, which replaces temporal relative positioning with calibrated light-direction geometry and enables full-basis joint generation. Together with an HDR radiance encoding, SimStage produces an integration-ready OLAT basis from a single portrait. Experiments on FaceOLAT improve composited relighting PSNR by 9.16 dB over 3DPR, while preserving substantially sharper details and accurately controlling shading, cast shadows, and specular highlights.

1 INTRODUCTION

OLAT offers a particularly attractive route to controllable portrait relighting. Instead of regenerating a portrait for every target illumination, OLAT methods predict its responses to individual lights and explicitly combine them according to the desired environment (Debevec et al., 2000; Rao et al., 2025). This provides two advantages that direct relighting does not naturally offer: (i) *explicit control*, because every light contributes through calibrated weights; and (ii) *reuse*, because the same responses can support many target illuminations without repeated generative inference. The ultimate

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.

goal is therefore a *Full-Basis OLAT Generator*: given one portrait, directly recover the complete lighting-response basis and turn an ordinary photograph into a virtual light stage.

A useful OLAT basis is governed by two stringent requirements. First, *Basis Fidelity*: every response must retain fine structures, directional shadows, specular highlights, and the HDR radiance needed for subsequent composition. A single portrait, however, provides incomplete evidence about geometry and reflectance, making these unseen lighting responses intrinsically difficult to infer (Kim et al., 2024). Second, *Basis Coherence*: the entire response set must preserve exactly the same underlying subject despite large changes in illumination. Small shifts in facial contours, hair strands, or texture can remain almost invisible in individual OLATs, yet turn sharp details into blur or ghosting when hundreds of responses are combined (Rao et al., 2025). These two requirements cannot be traded against each other: suppressing variation improves apparent consistency but destroys correct shadows and highlights, whereas independently sharp responses can still fail after integration. The problem is therefore not to generate 331 good images, but to generate *one good* 331*-image basis*.

Our key finding is that this problem can be addressed through *Full-Basis Generation*: the complete OLAT basis can be modeled as a single joint generative object rather than hundreds of separate predictions. Video diffusion provides a compelling foundation for this formulation because it already combines high-fidelity appearance synthesis with joint generation over multiple correlated observations. Yet directly transferring a video model does not work. Its native generation window is designed for much shorter sequences, and, more fundamentally, its interactions are organized by *temporal geometry*. OLATs are instead indexed by directions in illumination space. We find that this mismatch prevents naive video-style training from reliably learning even substantially shorter lighting windows, making full 331-response generation unattainable. The central problem is therefore not simply extending a video to more frames, but *rebinding the sequence axis from time to light*.

We introduce **SimStage** to realize this *Light-Axis Rebinding*. Starting from a pretrained video diffusion backbone, SimStage is redesigned and trained on our constructed OLAT data as a dedicated full-basis generator. Its key component, *LiRoPE*, replaces temporal relative positioning with relative light-direction geometry, allowing cross-response attention to organize itself according to illumination rather than frame order. SimStage then jointly denoises all 331 responses within one window, directly targeting *Basis Coherence*, while the strong pretrained appearance prior supports *Basis Fidelity*. Finally, an HDR radiance encoding preserves high-dynamic-range responses inside the video VAE and returns predictions to linear radiance before composition. The result is an integration-ready OLAT basis that can be reused for new target environments through explicit weighted summation, without further diffusion sampling.

Experiments on FaceOLAT show that SimStage improves composited relighting PSNR by 9.16 dB over 3DPR (Rao et al., 2025). Its full-basis predictions retain sharp facial contours, hair strands, and fine textures after integration, while accurately reproducing target-dependent shading, cast shadows, and specular highlights. We further analyze how generation-window size affects both individual OLAT quality and final composition, and show that LiRoPE is critical to scaling joint generation toward the full lighting basis. Our contributions are:

- **Full-Basis Generation.** We formulate the complete OLAT basis as a single joint generative object and show that a video-generation paradigm can be redesigned to directly synthesize all 331 lighting responses within one window.
- **Light-Axis Rebinding.** We identify the mismatch between temporal geometry and illumination geometry as the key obstacle to full-basis generation, and introduce *LiRoPE* to replace temporal relative positioning with calibrated light-direction relationships.
- **SimStage.** We combine LiRoPE, full-basis joint denoising, and HDR radiance encoding to achieve high *Basis Fidelity* and *Basis Coherence*, producing sharp, accurate, and precisely controllable relighting through explicit OLAT composition.

2 Related Work

Inverse-rendering methods estimate geometry, reflectance, and illumination (Wang et al., 2008; Nestmeyer et al., 2020; Wang et al., 2020; Hou et al., 2022; Pandey et al., 2021; Kim et al., 2024), while direct methods synthesize target-lit images (Sun et al., 2019; Zhou et al., 2019; Hou et al., 2021; Ponglertnapakorn et al., 2026; Ren et al., 2024; Chaturvedi et al., 2025; Liu et al., 2025).

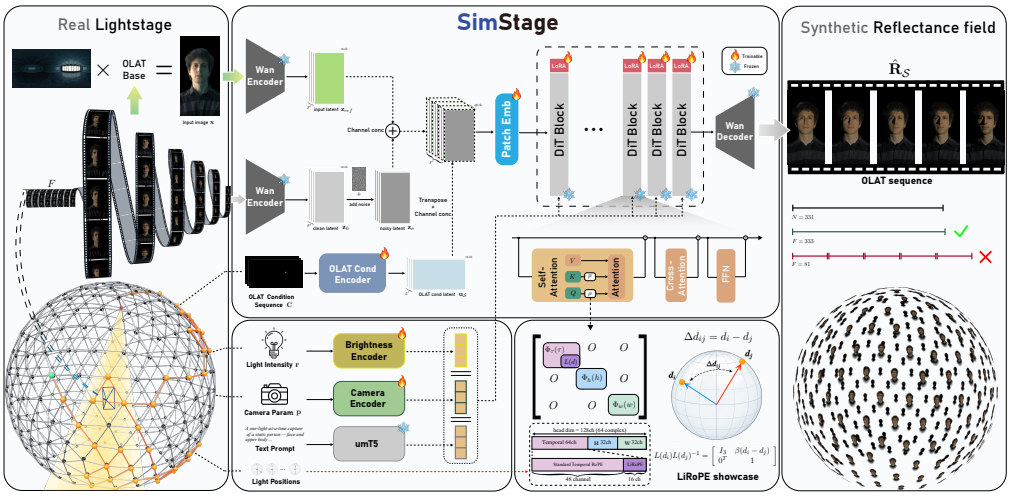

Figure 2: **SimStage pipeline.** The portrait, camera, intensity, and per-light conditions guide joint OLAT generation. Training uses encoded OLATs with added noise; inference starts from Gaussian noise. LiRoPE supplies relative light geometry, while HDR encoding connects linear radiance to the frozen video VAE. The recovered basis supports subsequent relighting through linear composition.

IC-Light (Zhang et al., 2025b) additionally incorporates light-transport consistency during training. OLAT prediction instead recovers reusable lighting responses for explicit composition (Debevec et al., 2000). Deep Reflectance Fields (Meka et al., 2019) use gradient-illumination captures; in the single-portrait setting, 3DPR (Rao et al., 2025) combines 3D-generative inversion with a triplane reflectance network, while POLAR (Chen et al., 2026) learns pairwise latent bridges from uniformly lit portraits (Chadebec et al., 2025). These methods query responses individually, leaving cross-light misalignment a source of composition error (Rao et al., 2025). SimStage jointly predicts the full basis with reversible HDR encoding, coordinating responses for composition in linear radiance. Built on Wan (Wan et al., 2025), SimStage jointly predicts all 331 responses from one portrait, using light-relative conditioning to organize cross-frame interactions by illumination rather than time.

3 Method

3.1 SimStage: Full-Basis Light Diffusion

SimStage models the complete OLAT basis as a *single diffusion sample*. Given a portrait $\mathbf{x}$, camera parameters $\mathbf{p}$, and N=331 calibrated lighting conditions $\mathbf{C} = (\mathbf{c}_1, \ldots, \mathbf{c}_N)$ with directions $\mathbf{d}_l \in \mathbb{S}^2$, it jointly predicts the HDR responses $\hat{\mathbf{R}} = (\hat{\mathbf{R}}_1, \ldots, \hat{\mathbf{R}}_N)$. Initialized from Wan 2.2 (Wan et al., 2025), SimStage replaces the temporal sequence semantics with a *light axis*: the subject and viewpoint remain fixed while illumination varies. All responses evolve within one denoising process, combining the pretrained appearance prior with cross-light self-attention. One generated sample therefore represents a complete virtual light stage (Fig. 2).

For a sampled light sequence $\mathcal{S} = (l_1, \ldots, l_F)$, let $\mathbf{R}_\mathcal{S}$, $\mathbf{C}_\mathcal{S}$, and $\mathbf{D}_\mathcal{S}$ denote its targets, conditions, and directions. The frozen video VAE maps F image frames to $F' = 1 + (F-1)/4$ latent frames. The portrait is encoded once and broadcast as $\mathbf{z}_{\text{ref}}$, while a trainable convolutional encoder g_ϕ maps per-light environment maps to spatially aligned features $\mathbf{u}_\mathcal{S}$. These features are concatenated with the noisy OLAT latent:

$$\mathbf{h}_{\text{in}} = \text{Concat}_{\text{ch}}[\mathbf{z}_\sigma, \mathbf{u}_\mathcal{S}, \mathbf{z}_{\text{ref}}]. \tag{1}$$

Each input contributes 16 channels, yielding 48 before the rebuilt patch embedding. Global context enters through cross-attention: camera extrinsics and normalized intrinsics are mapped to 32 tokens using Fourier features and an MLP, and sample-wide intensity is mapped to four tokens using a log-Fourier encoder. These tokens are concatenated with the cached fixed-prompt embedding from the frozen UMT5 encoder.

The remaining design addresses three mismatches: (i) temporal positioning versus illumination geometry; (ii) short video windows versus full-basis coupling; and (iii) bounded LDR inputs versus the HDR radiance required for composition.

3.2 LIGHT-AXIS CONSTRUCTION WITH LIROPE

Wan's temporal RoPE encodes relative frame offsets, but an OLAT sequence has no corresponding temporal geometry. The same offset can connect different angular separations, while nearby light directions may be distant in the ordering. Retaining this representation makes OLAT learning difficult even at substantially shorter windows and prevents direct scaling to the complete basis.

LiRoPE replaces temporal relative position with *relative light geometry* in selected attention channels. For a calibrated direction $\mathbf{d} \in \mathbb{S}^2$, define

$$L(\mathbf{d}) = \begin{bmatrix} \mathbf{I}_3 & \beta\mathbf{d} \\ \mathbf{0}^\top & 1 \end{bmatrix}. \tag{2}$$

Applying $L(\mathbf{d}_i)$ to a query and $L(\mathbf{d}_j)^{-\top}$ to a key introduces the relative factor $L(\mathbf{d}_i)L(\mathbf{d}_j)^{-1}$, whose translation is $\beta(\mathbf{d}_i - \mathbf{d}_j)$. The replaced subspace therefore encodes relative light displacement rather than frame-index difference. Following the band-replacement strategy of ReRoPE (Li et al., 2026), LiRoPE modifies the final 16 real channels of the temporal positional band and retains the other temporal and spatial channels. These channels carry four associated light directions as homogeneous 4-vectors, matching the VAE's four-frame compression. Exact grouping and derivation are provided in App. B.2.

Training sequences are graph walks on a spherical Delaunay graph of calibrated light directions. Adjacent positions tend to correspond to nearby lights, while randomized starts cover different regions of the rig. Walks may revisit lamps, so sequence length need not equal the number of distinct lights. At inference, the calibrated lights follow the rig's spiral ordering, with their relative geometry supplied by LiRoPE.

3.3 NATIVE FULL-BASIS JOINT DIFFUSION

SimStage trains and samples directly at the full-basis window. Separate windows prevent cross-window interaction during denoising, whereas length extrapolation exposes attention to a sequence scale absent from training. Native full-window generation instead couples responses within one denoising trajectory at the trained length, allowing structural agreement to develop before composition.

The rig contains 331 calibrated lights. To satisfy the frozen VAE constraint $F \equiv 1 \pmod 4$, the generation window contains 333 image positions; two padding outputs are discarded to recover the complete basis. The enlarged window serves cross-light coherence rather than video duration: LiRoPE supplies illumination geometry, and the shared portrait anchors the jointly evolving responses to a common subject. Split-window and length-extrapolation alternatives are compared in Sec. 4.3.

3.4 HDR-LINEAR BASIS REPRESENTATION

OLAT composition requires radiometric fidelity as well as structural coherence. Captured responses contain bright, spatially sparse highlights, whereas the frozen video VAE expects bounded LDR inputs. Clipping or irreversible tone mapping alters these responses' contributions to the final image. SimStage therefore separates its *generative representation* from the *linear-radiance readout*.

After scaling captured radiance by $\sqrt{N}$, each nonnegative value r is mapped into the VAE input range:

$$\tilde{r} = T(r) = 2\, f_{\mathrm{sRGB}}\left(\frac{\log\big(1 + \alpha\gamma \min(r, r_{\max})\big)}{\log(1 + \alpha\gamma r_{\max})}\right) - 1, \tag{3}$$

where f_{sRGB} is the sRGB transfer function. The mapping admits an analytic inverse for $0 \le r \le r_{\max}$; values above $r_{\max}$ are clipped. After VAE decoding, we apply the inverse and remove the $\sqrt{N}$ scaling, ensuring that composition operates in recovered linear radiance rather than VAE or display space. Constants, the inverse mapping, and reconstruction analysis are given in App. B.4.

Table 1: **OLAT reconstruction on held-out subjects.** PSNR-μ/SSIM-μ are measured in the companded HDR domain; remaining metrics use a common display-space representation.

Method	PSNR-μ ↑	SSIM-μ ↑	PSNR↑	SSIM↑	L1↓	LPIPS↓	DISTS↓
3DPR (Rao et al., 2025)	23.22	0.397	18.76	0.312	0.052	0.505	0.332
SimStage	**32.68**	**0.900**	**28.60**	**0.832**	**0.021**	**0.143**	**0.147**

Table 2: **Relighting quality on held-out subjects.** All methods use the same subjects and target environments. The table evaluates displayed appearance; HDR radiometric accuracy is evaluated separately. †Without person mask.

Method	PSNR↑	SSIM↑	LPIPS↓	DISTS↓
Square canvas (512×512 aligned face window)				
3DPR (Rao et al., 2025)	18.43	0.691	0.339	0.182
IC-Light (Zhang et al., 2025b)	17.25	0.845	0.254	0.158
Lite2Relight (Rao et al., 2024)	16.28	0.839	0.346	0.182
NeuralGaffer (Jin et al., 2024)	15.96	0.797	0.294	0.208
RITW (Tajima et al., 2021)	14.39	0.737	0.307	0.248
SimStage	**27.59**	**0.901**	**0.109**	**0.102**
Full-frame canvas (256×512 model domain)				
IC-Light† (Zhang et al., 2025b)	19.20	**0.895**	0.167	0.143
SimStage	**28.19**	0.863	**0.074**	**0.090**

3.5 Training and Light-Stage Readout

For a sampled sequence $\mathcal{S}$ and global intensity b, the frozen VAE encoder $\mathcal{E}$ produces the clean latent

$$\mathbf{z}_0 = \mathcal{E}\Big(T(b\sqrt{N}\,\mathbf{R}_{\mathcal{S}})\Big). \tag{4}$$

The noisy latent follows standard flow interpolation,

$$\mathbf{z}_\sigma = (1-\sigma)\mathbf{z}_0 + \sigma\boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}), \tag{5}$$

and the conditioned DiT predicts the velocity using

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}\left[\frac{1}{N_z}\,\|v_\theta - (\boldsymbol{\epsilon} - \mathbf{z}_0)\|_2^2\right]. \tag{6}$$

We optimize the DiT LoRA adapters, rebuilt patch embedding, and conditioning modules, keeping the base DiT, VAE, and text encoder frozen. Both noise experts are trained separately in their respective noise regimes using 333-frame sequences. Architectural and optimization details are provided in App. A.

At inference, SimStage jointly denoises the complete sequence from Gaussian noise in 50 steps, switching experts according to noise level. VAE decoding and inverse HDR mapping recover linear-radiance responses, with the two padding outputs discarded:

$$\hat{\mathbf{R}} = \frac{1}{\sqrt{N}}\,T^{-1}(\mathcal{D}(\hat{\mathbf{z}}_0)). \tag{7}$$

For a target environment $\mathbf{e}$, calibrated weights $w_l(\mathbf{e})$ give

$$\hat{\mathbf{I}}(\mathbf{e}) = \sum_{l=1}^{N} w_l(\mathbf{e})\,\hat{\mathbf{R}}_l. \tag{8}$$

This readout follows the linearity of light transport in incident illumination (App. B.1). The recovered basis is reused without further diffusion sampling: *one full-basis generation, many lighting readouts*.

4 Experiments

4.1 Experimental Setup

Data and implementation. We train and evaluate on FaceOLAT (Rao et al., 2025), which contains 139 subjects captured from multiple calibrated cameras under 331 point lights. We hold out 13 subjects by identity and train on the remaining 126. Quantitative evaluation uses the near-frontal Cam06



Figure 3: **Relighting comparison.** Held-out subjects under target environments. SimStage preserves substantially sharper facial and hair details while reproducing target-dependent shading, shadows, and highlights.

view, where captured OLAT ground truth is available. SimStage is initialized from Wan 2.2-Fun-A14B-Control (Wan et al., 2025) and predicts the complete basis in one 333-frame pass, including two padding positions required by the frozen VAE and discarded after generation. Training details are provided in App. A.

Metrics and baselines. We evaluate the representation at two levels. For individual OLATs, we report PSNR-μ and SSIM-μ in the companded HDR domain, together with PSNR, SSIM, L1, LPIPS, and DISTS in display space. For final relighting, predicted OLATs are combined with the same calibrated environment weights used to construct the captured ground truth, and we report display-space PSNR, SSIM, LPIPS, and DISTS. The common display protocol uses the same canvas, foreground support, and exposure-selection rule for prediction and ground truth; HDR-linear accuracy is reported separately where the prediction representation permits it. We compare OLAT reconstruction with 3DPR (Rao et al., 2025), and final relighting with 3DPR and direct relighters including IC-Light (Zhang et al., 2025b), Lite2Relight (Rao et al., 2024), NeuralGaffer (Jin et al., 2024), and RITW (Tajima et al., 2021).

4.2 Full-Basis OLAT and Relighting Quality

Basis fidelity. Table 1 compares predicted OLAT responses against captured ground truth. 3DPR predicts in a tone-mapped domain whose released preprocessing clips bright radiance before display mapping; its HDR-domain scores are therefore computed after approximately inverting this transform. SimStage instead predicts through the reversible radiance representation described in Sec. 3.4. SimStage improves PSNR-μ by 9.46 dB and display-space PSNR by 9.84 dB over 3DPR, while substantially reducing perceptual error. The improvement is particularly visible around hair, facial contours, and localized highlights, where a basis must preserve both fine structure and strong illumination-dependent changes.

Basis readout. For each environment, we combine the predicted OLATs using exactly the calibrated light weights used to construct the captured target. Table 2 compares this readout with 3DPR and direct relighting methods. On the square portrait domain, SimStage exceeds 3DPR by more

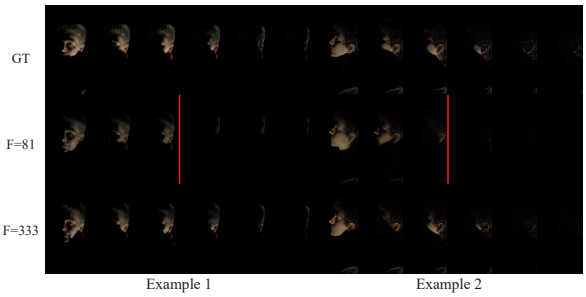

Figure 4: **Split-window seams.** Two neighborhoods around boundaries between independently generated windows. Each short-window prediction is locally plausible, but structural discrepancies appear where no cross-window interaction is available.

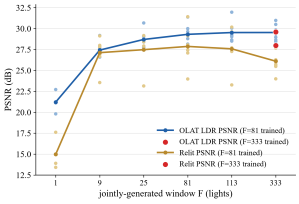

Figure 5: **Full-basis coupling must be learned, not extrapolated.** An 81-frame-trained model improves as its inference context grows within the familiar regime, but degrades when extrapolated toward the complete basis. The native full-window model avoids this reversal.

Table 3: **Native full-basis training versus split-window generation.** Same backbone and matched training samples.

	F=81 / split	F=333 / full
OLAT reconstruction		
PSNR-μ ↑	33.51	**34.22**
SSIM-μ ↑	0.894	**0.898**
LDR PSNR↑	29.20	**29.91**
LDR L1↓	0.0177	**0.0171**
LDR LPIPS↓	0.1174	**0.1041**
LDR DISTS↓	0.1563	**0.1415**
Composite relighting		
PSNR↑	25.97	**26.40**
LPIPS↓	0.0881	**0.0838**
DISTS↓	0.1022	**0.1011**

than 9 dB in PSNR and substantially improves LPIPS and DISTS. Compared with direct relighters, the advantage is not only visual quality but also representation reuse: SimStage generates the basis once, and subsequent environments require only a new linear readout.

4.3 Full-Basis Joint Generation

Joint generation v.s. split windows. We compare native F=333 generation against an F=81 model that assembles all 331 responses from independent windows, with the same backbone, data, and matched training samples. Native full-window training improves every reported OLAT and relighting metric (Table 3). Split-window outputs exhibit structural discrepancies at window boundaries (Fig. 4).

Full-basis coupling must be learned, not extrapolated. We fix the F=81 model weights and increase its inference window. Reconstruction initially improves but degrades near the full basis; native F=333 training avoids this reversal (Fig. 5). Relative to the full calibrated basis, native training reduces the light-pair separation gap from 0.39 to 0.14 in total-variation distance and from 18.3° to 6.5° in Wasserstein distance compared with 81-light training walks. These results support training at the full window rather than relying on test-time extrapolation.

Compressed history is not a substitute for the full basis. We additionally test FramePack (Zhang et al., 2025a) as a representative packed-context strategy using Wan 2.2 5B. Under the matched-step study, FramePack obtains composite PSNR/SSIM/LPIPS of 24.04/0.844/0.169, compared with 25.05/0.867/0.149 for the independent-window reference. It also trails in per-OLAT PSNR on 70.8% of the evaluated responses. This comparison does not rule out every cache design, but indicates that compressing the latent history is poorly matched to a reconstruction problem in which fine response errors remain visible after composition. Detailed distributions are provided in App. C.5.

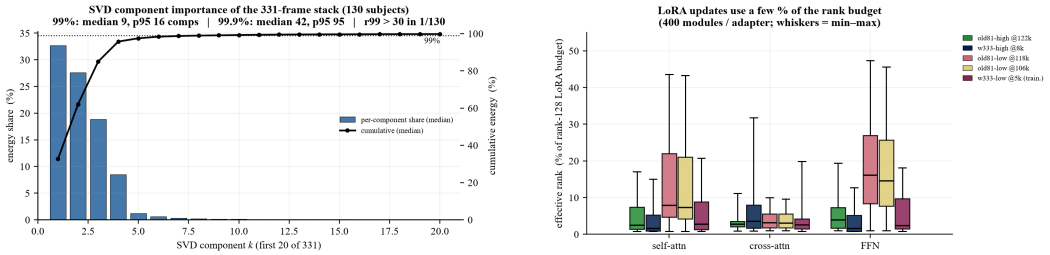

Figure 6: **Feasibility of full-basis adaptation.** Left: spectral dimension of the captured OLAT fields. Right: effective rank of the learned adapters. The basis is long in samples but compact in variation.

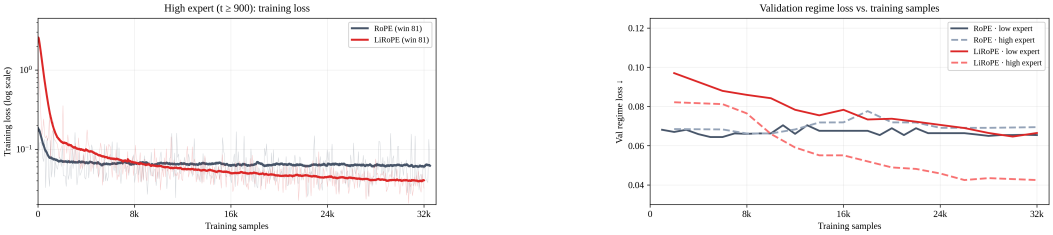

Figure 7: **LiRoPE makes the light axis learnable.** Standard temporal RoPE quickly reaches a plateau and shows little further validation improvement, whereas LiRoPE continues to optimize and reaches a substantially better validation regime under the same training budget.

Feasibility of full-basis adaptation. The complete basis is long in samples but comparatively compact in variation. Subject identity, pose, viewpoint, and geometry remain fixed while illumination changes. Across the captured fields, a median of only 42.5 spectral components explains 99.9% of the energy in the 331-response stack, while the smooth diffuse component saturates around spherical-harmonic degree 4–5. The learned adaptation is similarly concentrated: the effective ranks of the LoRA updates are only 2–5 within a rank-128 parameterization, and the 81- and 333-frame adapters exhibit strong top-direction subspace overlap. These observations help explain why the generation window can be expanded dramatically with comparatively modest adaptation: more positions largely sample additional responses of the same underlying scene rather than introducing the unconstrained state changes of a general long video.

4.4 Effect of Light-Axis Rebinding

We compare temporal RoPE and LiRoPE at $F{=}81$ with the same Wan 2.2-Fun-A14B-Control backbone, training recipe, and per-light environment-map conditioning. Temporal RoPE quickly plateaus, whereas LiRoPE continues to improve and achieves lower validation loss for both noise experts (Fig. 7). Both variants receive the same absolute lighting conditions, so this comparison supports the benefit of explicit relative light geometry. Qualitative queries at uncalibrated directions are provided in App. C.3; ground-truth OLATs between calibrated lamps are unavailable.

4.5 HDR Representation Fidelity

We evaluate the HDR encoding through a frozen-VAE round trip on 24 validation windows (1,944 frames). Reconstruction reaches 47.9 ± 3.6 dB PSNR-μ and 0.983 SSIM per frame. With the VAE and clipping ceiling fixed, PSNR-μ varies little over practical compression slopes, but aggressive compression introduces high-frequency artifacts in dark regions (Fig. 8). Reconstructed high-frequency energy reaches $1.7\times$ ground truth at $\alpha{=}1000$, versus near parity at $\alpha \leq 30$; we therefore use $\alpha{=}10$.

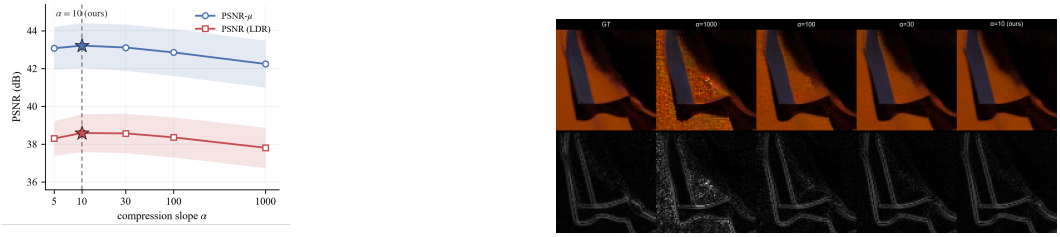

Figure 8: **HDR representation through the frozen VAE.** Reconstruction quality and high-frequency behavior as the compression slope α varies.

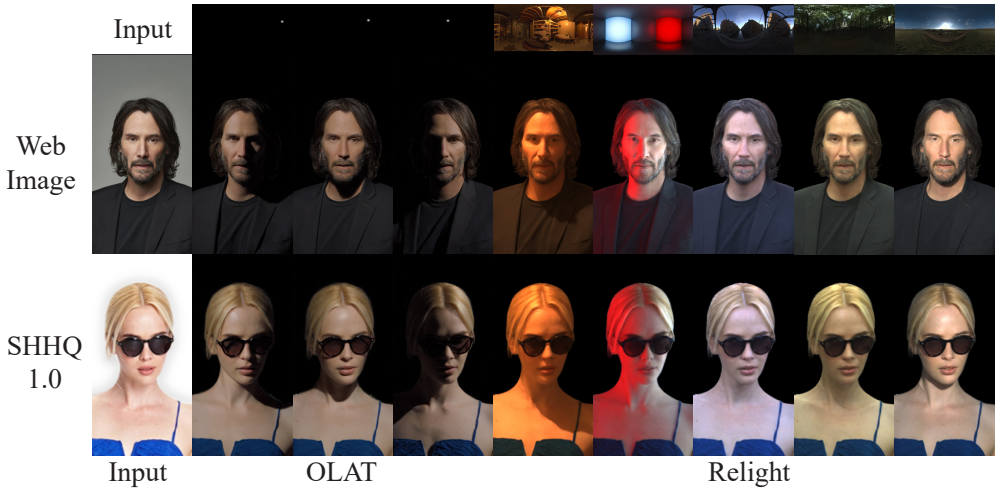

Figure 9: **In-the-wild virtual light stages.** Web and SHHQ 1.0 (Fu et al., 2022) inputs, recovered OLAT responses, and relit results; full grids in App. C.7.

4.6 In-the-Wild Virtual Light Stages

We valuate SimStage on web and SHHQ 1.0 (Fu et al., 2022) portraits. Figure 9 shows stable identity and appearance across relit outputs, with illumination-dependent shading, shadows, and highlights. Full grids and environment-subset decompositions are provided in Fig. 23 (App. C.7).

5 Conclusion

SimStage redesigns video diffusion into a dedicated full-basis OLAT generator, treating the complete lighting basis as *one diffusion sample*. The key is *Light-Axis Rebinding*: LiRoPE supplies relative illumination geometry to cross-response attention, overcoming the learning plateau observed with native temporal positioning. This enables joint generation of all 331 calibrated responses within one window, while HDR encoding connects the pretrained video VAE to linear-radiance composition. Together, these designs target both *basis fidelity* and *basis coherence*, preserving detailed lighting responses and their structural agreement under integration. Experiments demonstrate improved OLAT reconstruction and sharper composited relighting, with controlled comparisons supporting the roles of light-relative conditioning and native full-window training. The result is a reusable virtual light stage rather than a single lighting edit: *one full-basis generation, many explicitly controlled lighting readouts*, without further diffusion sampling.

Author Contributions

Acknowledgments

AI use statement

In this work, we used generative AI tools for tasks with recommended disclosure: editing for readability, identifying relevant literature, and formatting references. We have not used generative AI tools for developing theoretical models or conceptual frameworks, formulating or assisting in the proof of mathematical claims, proposing or refining hypotheses, designing or providing feedback on methodology or experiments, implementing methods, or interpreting results

Ethics statement

This work builds on FaceOLAT (Rao et al., 2025), a research light-stage library of volunteers captured under calibrated illumination. We use the dataset as released by its authors for academic research, keep every evaluation subject out of training (App. B.1), and show dataset imagery only to document relighting behavior. The in-the-wild study (Sec. 4.6) additionally uses web-sourced and SHHQ 1.0 (Fu et al., 2022) portraits for qualitative documentation. Our identity-consistency protocol (App. C.1) computes ArcFace cosine similarity between predicted and captured portraits of the same subject; this biometric embedding serves strictly as an accuracy metric and is never used to identify or search for individuals. Because photorealistic relighting could be repurposed to fabricate imagery of real people, we develop SimStage for graphics and vision research, relight only captured or provided portraits, and support provenance disclosure and watermarking so that rendered relightings remain distinguishable from photographs. Predicted reflectance inherits the appearance priors of the pretrained backbone and the composition of the capture library; we report per-subject statistics in App. C.6 to keep these limits visible.

Reproducibility statement

SimStage is initialized from the publicly released Wan 2.2-Fun-A14B-Control checkpoint (Wan et al., 2025), with the VAE and text encoder used exactly as released; all baselines run zero-shot from their released checkpoints (App. C.1). Complete architecture, LoRA, optimization, and schedule settings, including the two-expert regime split, the fixed subject-level data splits, and the training seed, are specified in App. A.1; the HDR encoding and its exact round trip in App. B.4; the LiRoPE construction in App. B.2. Every comparison follows a documented protocol: automatic-exposure relighting metrics (App. C.1), the coupling-window sweep (App. C.2), densified queries (App. C.3), the controlled FramePack study (App. C.5), spectral analyses with fixed definitions (App. C.6), and in-the-wild grids (App. C.7). We will release code, LoRA adapters, and evaluation scripts upon publication.

References

Junuk Cha, Mengwei Ren, Krishna Kumar Singh, He Zhang, Yannick Hold-Geoffroy, Seunghyun Yoon, HyunJoon Jung, Jae Shin Yoon, and Seungryul Baek. Text2relight: Creative portrait relighting with text guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1980–1988, 2025.

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. Lbm: Latent bridge matching for fast image-to-image translation. In *International Conference on Computer Vision (ICCV)*, 2025. arXiv:2503.07535.

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*, 2025.

Zhuo Chen, Chengqun Yang, Zhuo Su, Zheng Lv, Jingnan Gao, Xiaoyuan Zhang, Xiaokang Yang, and Yichao Yan. Polar: A portrait olat dataset and generative framework for illumination-aware face modeling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 28871–28881, 2026.

Paul 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*, pp. 145–156, 2000.

Jianglin Fu, Shikai Li, Yuming Jiang, Kwan-Yee Lin, Chen Qian, Chen Change Loy, Wayne Wu, and Ziwei Liu. Stylegan-human: A data-centric odyssey of human generation. *arXiv preprint arXiv:2204.11823*, 2022.

Mingming He, Pascal Clausen, Ahmet Levent Taşel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgert, Ning Yu, et al. Diffrelight: Diffusion-based facial performance relighting. In *SIGGRAPH Asia 2024 conference papers*, pp. 1–12, 2024.

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.

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

Haian Jin, Yiming Li, Zhi Li, Yu Hong, Fujun Luan, Noah Snavely, Sai Bi, Kalyan Sunkavalli, Miloš Hašan, Yang-Tian Sun, Rui Chen, Zhan Bu, and Shuang Zhao. Neural gaffer: Relighting any object via diffusion. In *Advances in Neural Information Processing Systems (NeurIPS)*, 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*, pp. 25096–25106, 2024.

Chunyang Li, Yuanbo Yang, Jiahao Shao, Hongyu Zhou, Katja Schwarz, and Yiyi Liao. Rerope: Repurposing rope for relative camera control. *arXiv preprint arXiv:2602.08068*, 2026.

Yong Liu, Wenpeng Xiao, Qianqian Wang, Junlin Chen, Shiyin Wang, Yitong Wang, Xinglong Wu, and Yansong Tang. Dreamlight: Towards harmonious and consistent image relighting. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025. arXiv:2506.14549.

Abhimitra Meka, Christian Haene, Rohit Pandey, Michael Zollhöfer, Sean Fanello, Graham Fyffe, Adarsh Kowdle, Xueming Yu, Jay Busch, Jason Dourgarian, et al. Deep reflectance fields: high-quality facial reflectance field inference from color gradient illumination. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.

Thomas Nestmeyer, Jean-François 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.

Rohit Pandey, Sergio Orts-Escolano, Chloe Legendre, Christian Haene, Sofien Bouaziz, Christoph Rhemann, Paul E Debevec, and Sean Ryan Fanello. Total relighting: learning to relight portraits for background replacement. *ACM Trans. Graph.*, 40(4):43–1, 2021.

Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. Difareli++: Diffusion face relighting with consistent cast shadows. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 48:5068–5082, 2026.

Pramod Rao, Gereon Fox, Abhimitra Meka, Mallikarjun BR, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, Mohamed Elgharib, et al. Lite2relight: 3d-aware single image portrait relighting. In *ACM SIGGRAPH 2024 Conference Papers*, pp. 1–12, 2024.

Pramod Rao, Abhimitra Meka, Xilong Zhou, Gereon Fox, Mallikarjun BR, Fangneng Zhan, Tim Weyrich, Bernd Bickel, Hanspeter Pfister, Wojciech Matusik, et al. 3dpr: Single image 3d portrait relighting with generative priors. In *Proceedings of the SIGGRAPH Asia 2025 Conference Papers*, pp. 1–12, 2025.

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.

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

Daichi Tajima, Yoshihiro Kanamori, and Yuki Endo. Relighting humans in the wild: Monocular full-body human relighting with domain adaptation. *Computer Graphics Forum*, 40(7):205–216, 2021. doi: 10.1111/cgf.14414.

Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Yang Wang, Lei Zhang, Zicheng Liu, Gang Hua, Zhen Wen, Zhengyou Zhang, and Dimitris Samaras. Face relighting from a single image under arbitrary unknown lighting conditions. *IEEE transactions on pattern analysis and machine intelligence*, 31(11):1968–1984, 2008.

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

Lvmin Zhang, Shengqu Cai, Muyang Li, Gordon Wetzstein, and Maneesh Agrawala. Frame context packing and drift prevention in next-frame-prediction video diffusion models. *arXiv preprint arXiv:2504.12626*, 2025a.

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 *International Conference on Learning Representations (ICLR)*, 2025b.

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*, pp. 7194–7202, 2019.

A Implementation Details

A.1 Architecture and Training

SimStage fine-tunes the Wan 2.2-Fun-A14B-Control video DiT (dual-expert, 40 transformer blocks, hidden dim 5120, 40 heads), loaded as `ti2v` with `add_ref_conv=False`. The Wan 2.1 VAE (16-ch latent, $8\times$ spatial / $4\times$ temporal) and the UMT5 text encoder are frozen. We apply LoRA adapters (rank $r_{\text{LoRA}}{=}128$, $\alpha_{\text{LoRA}}{=}128$, Gaussian init) to `self_attn`, `cross_attn`, `ffn.0`, `ffn.2`; the patch embedding is rebuilt to 48 input channels (warm-started from the pretrained 52-channel weights by reusing the noisy- and control-video filters), and the OLAT environment-map encoder, camera encoder, and brightness encoder are trainable; LiRoPE uses fixed calibrated directions and has no learned geometry parameters. Following Wan 2.2, a high-noise expert ($\sigma{\geq}0.9$, timestep $t{\geq}900$) and a low-noise expert ($\sigma{<}0.9$) are fine-tuned separately within their timestep regimes and routed by noise level at inference; the two experts are independent checkpoints with separate LoRA adapters, loaded jointly and switched per timestep. We use the FlowMatchEulerDiscrete scheduler (shift 5.0, 1000 train timesteps, 50 inference steps), AdamW under DeepSpeed ZeRO-2 (bf16) on $8\times$A100-80GB, learning rate 10^{-5} for both experts, with gradient checkpointing. The low-noise expert trains for up to $5{\times}10^4$ steps and the high-noise expert for $\sim 5{\times}10^3$ steps, with validation every 1000 steps. Video frames are resized to $512{\times}256$; training uses a per-GPU micro-batch of 1 across 8 GPUs, seed 42.

The reference portrait is encoded to a single-frame VAE latent and broadcast across the chunk; each per-light environment map ($512{\times}1024$ Gaussian-spill latlong, one per calibrated light) is encoded by a lightweight convolutional encoder g_ϕ into a latent-aligned feature. Both are channel-concatenated

with the noisy latent (48 channels total: $[\mathbf{z}_\sigma \parallel \mathbf{u}_\mathcal{S} \parallel \mathbf{z}_{\text{ref}}]$) and fed to the rebuilt patch embedding. Camera parameters $\mathbf{p}$ (a 4×4 extrinsic matrix plus normalized intrinsics) are mapped by Fourier pose features and an MLP (32 tokens, hidden dim 256, output dim 5120) and prepended to the text context of every cross-attention layer; UMT5 text embeddings $\mathbf{c}_{\text{text}}$ are pre-cached once with a fixed descriptive prompt. An independent brightness encoder maps the sample-wide intensity b through log-Fourier features and an MLP to four tokens; the context order is brightness, camera, then text (Sec. 3.1). The same b scales all target OLATs before HDR encoding, while the environment-map conditions remain at unit intensity. A per-camera 4×4 color-calibration matrix is applied to captured frames before encoding. Camera parameters come from the calibrated rig; applying SimStage to uncalibrated photographs would require estimating them, whose effect on quality we have not quantified; the in-the-wild study of Sec. 4.6 operates under estimated cameras. We concatenate the reference rather than prepending it as a token: a prepended token would occupy the temporal slot that anchors LiRoPE's manifold geometry, breaking the one-to-one latent-to-manifold-point mapping.

The native model uses 333-frame training windows (84 latent frames after the $4\times$ temporal VAE), sampled as graph walks over the spherical Delaunay graph of the N=331 calibrated lights, with brightness augmentation in $[0.3, 1.0] \cdot \sqrt{331}$. A walk may revisit lights, so its 333 positions need not contain every distinct lamp. The separate 81-frame checkpoints used for the window sweep are described in App. B.5. LiRoPE replaces only the last 16 real channels of the temporal RoPE band with $L(\mathbf{d})$ at β=0.5 (all spatial and higher-frequency temporal channels unchanged); raw light directions are obtained by a least-squares sphere fit to the calibrated lamp positions. The 331 frames are ordered by the calibration spiral and padded to $333 = 1 + 4 \times 83$ by repeating the final two frames in reverse order (the temporal VAE requires $F \equiv 1 \bmod 4$); the two padding frames are discarded at readout and enter the repurposed channels only as zero-displacement repeats. Grouping matches the $4\times$ temporal VAE exactly: latent 0 carries frame 0's direction replicated across its four homogeneous slots (an anchor for every raw slot), and latent k>0 carries the four directions of frames $1{+}4(k{-}1)$ to $4k$, one homogeneous 4-vector each—the 16 repurposed channels are four lights $\times$ four homogeneous coordinates. Ground-truth OLATs are linear HDR radiance; we scale each by $\sqrt{331}$, then apply the encoding of Eq. 3 (α=10, $\gamma \approx 3.6$, $r_{\max}$=28) into the VAE's $[-1, 1]$ range—invertible with an exact round trip on unclamped values (App. B.4).

B Background and Derivations

B.1 The Reflectance Field of Skin

This section reviews, following Debevec et al. (2000), the physical background summarized in Sec. 3.1. A fixed-view image under an environment map can be written as an integral over incident directions,

$$\mathbf{I}(\mathbf{e}) = \int_{\mathbb{S}^2} \mathbf{R}(\mathbf{d})\, \ell(\mathbf{e}, \mathbf{d})\, d\mu(\mathbf{d}), \tag{9}$$

whose calibrated discrete readout is Eq. 8. A reflectance function is the *complete* transport response of the subject to a single light: it encodes diffuse reflection, specular reflection, self-shadowing, translucency, mutual illumination, and subsurface scattering. For skin, these phenomena organize into two additive components: a *specular* component—Fresnel reflection at the air–oil interface of the skin surface, a sharp view-dependent lobe that produces catchlights and strengthens toward grazing angles—and a *diffuse subsurface* component—light that crosses the surface, scatters within the epidermal and dermal layers, and re-emerges as a broad, nearly Lambertian lobe:

$$\mathbf{R}(\mathbf{d}) \;=\; \mathbf{R}^{\text{spec}}(\mathbf{d}) \;+\; \mathbf{R}^{\text{sub}}(\mathbf{d}),$$

a separation Debevec et al. (2000) obtain physically with cross-polarization. All light paths ending in diffuse reflection, including interreflection between parts of the face and translucency, are folded into the subsurface component, and self-shadowing acts by driving both components to zero in occluded directions. Because radiance is additive in the incident field, the decomposition superposes component-wise under any environment, so a recovered field relights both phenomena—and everything folded into them—exactly.

Counts and splits. A FaceOLAT *entry* is one (subject, expression, camera) capture. The library's 139 subjects are split by subject with a fixed seed: 13 held-out test subjects (never trained; every

evaluation table) and 126 training subjects, of which 13 form the validation split; checkpoints are selected by the per-expert in-distribution regime loss on validation windows, and the coupling-window sweep uses seven subjects (App. C.2). Full-library spectral statistics use the 130 subjects with a first-expression capture. The FramePack walkthrough covers 282 entries; its golden composite protocol draws 256 maps by brightness stratification from the same 6,651-map pool as App. C.1.

B.2 LiRoPE: Derivation and Properties

Mechanics. In rotary position embeddings, attention between a query at position i and a key at position j depends on the positions only through their difference, because the transforms applied to the query and the key are mutual inverses at equal positions. LiRoPE obtains the same relative behavior on the light sphere by replacing the last 16 real channels of the temporal RoPE band—four groups of four channels, each carrying a homogeneous 4-vector—with transforms of the calibrated directions: treating each homogeneous block as a row vector, the query channels of frame i are right-multiplied by $L(\mathbf{d}_i)$ and the key channels of frame j by $L(\mathbf{d}_j)^{-\top}$ (equivalently, column vectors are left-multiplied by the transposes), with L defined in Eq. 2. The attention logit between the two frames then contains the bilinear form

$$L(\mathbf{d}_i)\, L(\mathbf{d}_j)^{-1} \;=\; \begin{bmatrix} \mathbf{I}_3 & \beta(\mathbf{d}_i - \mathbf{d}_j) \\ \mathbf{0}^\top & 1 \end{bmatrix}, \tag{10}$$

using $L(\mathbf{d}_j)^{-1} = \left[\begin{smallmatrix} \mathbf{I}_3 & -\beta\mathbf{d}_j \\ \mathbf{0}^\top & 1 \end{smallmatrix}\right]$: the rotational block cancels to the identity and the translational block becomes the displacement $\beta(\mathbf{d}_i - \mathbf{d}_j)$.

Properties. (i) Identical directions give the identity *on the repurposed sub-band*: two frames under the same light attend as the same position there (the unrepurposed temporal channels still encode their sequence offset). (ii) *Relative-displacement parameterization.* The bilinear form of Eq. 10 depends on $(\mathbf{d}_i, \mathbf{d}_j)$ only through the displacement $\mathbf{d}_i - \mathbf{d}_j$: no absolute origin and no trajectory position enters the repurposed channels, and adding a common offset to every direction—re-anchoring the frame about which directions are expressed—leaves every logit exactly unchanged. Which displacements a window contains, not where it sits, is all the sub-band ever sees; in particular, uncalibrated directions queried at inference (Sec. 4.4) enter attention through their displacements from the calibrated directions, requiring no absolute anchor. What this parameterization does *not* imply: the chord norm $\|\mathbf{d}_i - \mathbf{d}_j\| = 2\sin(\theta_{ij}/2)$, with θ_{ij} the geodesic angle on $\mathbb{S}^2$, is strictly monotone in the separation and invariant under global rotations, but the logit is a function of the displacement *vector*, not only of its norm, and a bilinear form need not be rotation-invariant (a rotation carrying a displacement from the x-axis to the y-axis changes $q\, L(\mathbf{d}_i)L(\mathbf{d}_j)^{-1}k^\top$ for generic q, k). Insensitivity of the *learned* kernel to displacement orientation is therefore an empirical property of the trained model—fostered by the random-walk training distribution, which traverses all orientations (App. A.1)—not a structural guarantee. (iii) Only 16 channels of an existing RoPE band are repurposed; no new parameters, no architectural change.

Sphere fit. Manifold points are obtained once, offline: given calibrated lamp positions $\mathbf{s}_l \in \mathbb{R}^3$, solve $\min_{\mathbf{o},\rho} \sum_l \big(\|\mathbf{s}_l - \mathbf{o}\| - \rho\big)^2$ for the center $\mathbf{o}$ and radius ρ, and set $\mathbf{d}_l = (\mathbf{s}_l - \mathbf{o})/\|\mathbf{s}_l - \mathbf{o}\|$ (per-point normalization, matching the implementation).

B.3 LiRoPE: The Design Space of Relative Encodings on $\mathbb{S}^2$

This subsection justifies the two decisions implicit in Eq. 2: representing light directions as ambient points of $\mathbb{R}^3$ rather than as angles, and borrowing the abelian translation group rather than the rotation group of the sphere.

Angle encodings conflate geometry with coordinates. The seemingly natural alternative is to feed the spherical coordinates (θ, ϕ) of each direction into the standard RoPE frequencies, so that attention depends on the coordinate differences $(\Delta\theta, \Delta\phi)$. But coordinate differences are not a function of geodesic separation: $\cos\theta_{ij} = \cos\theta_i \cos\theta_j + \sin\theta_i \sin\theta_j \cos(\Delta\phi)$ depends on where on the sphere the pair sits, so two pairs with identical coordinate differences can have arbitrarily different separations (a fixed $\Delta\phi$ spans a large arc at the equator and almost none near the poles, where ϕ

moreover degenerates—all longitudes meet—and wraps modulo 2π elsewhere). Attention would again carry absolute-position information—precisely what a relative encoding must exclude—and the coupling learned on one region of the sphere would not transfer to another, breaking property (ii) of App. B.2 at its root.

Rotation encodings cannot isolate a relative quantity. A second alternative lifts each direction to a rotation $R(\mathbf{d}) \in SO(3)$ with $R(\mathbf{d})\,\mathbf{e}_3 = \mathbf{d}$ and lets the logit contain $R(\mathbf{d}_i)\,R(\mathbf{d}_j)^\top$. Two obstructions are fatal. First, the lift requires choosing, for every $\mathbf{d}$, an additional rotation about $\mathbf{d}$; no continuous such choice exists on all of $\mathbb{S}^2$ (it would yield a continuous unit tangent field $\mathbf{v}(\mathbf{d}) = R(\mathbf{d})\mathbf{e}_1$, excluded by the hairy-ball theorem). Any practical lift is therefore discontinuous or position-dependent, and re-gauging $\mathbf{d} \mapsto S(\mathbf{d})R(\mathbf{d})$ with $S(\mathbf{d})$ a rotation about $\mathbf{d}$ changes the relative element in a position-dependent way—gauge artifacts, not geometry, would enter the logits. Second, even under an ideally covariant gauge $R(Q\mathbf{d}) = Q\,R(\mathbf{d})$, a global rotation Q of the entire calibration conjugates the relative element, $R(Q\mathbf{d}_i)R(Q\mathbf{d}_j)^\top = Q\,R(\mathbf{d}_i)R(\mathbf{d}_j)^\top Q^\top$: the pair's encoding moves nontrivially with global orientation instead of remaining a difference in a fixed vector space as in Eq. 10. The root cause is non-commutativity. A relative encoding built from mutual-inverse transforms collapses to a difference only in abelian families—$\Phi(i)^{-1}\Phi(j) = \Phi(j{-}i)$ in standard RoPE, $L(\mathbf{d}_i)L(\mathbf{d}_j)^{-1}$ in Eq. 10—whereas a product of rotations is conjugated by the global action, $Q\,R(\mathbf{d}_i)R(\mathbf{d}_j)^\top Q^\top$, and never collapses to a fixed difference. The only gauge-invariant content of a light pair is the separation angle θ_{ij}; no bilinear form of the lifted matrices isolates it (extracting a rotation angle is nonlinear, and the $q^\top k$ of attention is linear).

The obstruction is intrinsic to $\mathbb{S}^2$. Among spheres, only $\mathbb{S}^0$, $\mathbb{S}^1$, and $\mathbb{S}^3$ carry a group structure. Standard RoPE lives on $\mathbb{S}^1 \cong U(1)$, where each frequency pair is a point of the circle and rotation-by-inverse realizes the difference natively; $\mathbb{S}^3 \cong SU(2)$ (unit quaternions) would likewise admit the canonical relative $q_i q_j^{-1}$.[1] $\mathbb{S}^2$ admits no Lie-group structure at all—the hairy-ball theorem already excludes a nonvanishing tangent field, a fortiori a group law—so *no intrinsic difference of two directions exists*; any RoPE-style scheme on the light sphere must borrow a group acting around, rather than on, the manifold.

Translations are the simplest adequate borrowed group. The construction of Eq. 2 asks the borrowed family to meet exactly three requirements, each forced by the preceding paragraphs: (i) it is abelian, so the relative transform collapses to a difference (Eq. 10); (ii) that difference faithfully encodes separation—the chord norm $\|\mathbf{d}_i - \mathbf{d}_j\| = 2\sin(\theta_{ij}/2)$ is strictly monotone in the geodesic angle; and (iii) norm-based quantities are exactly invariant under global rotations of the calibration, since all displacements rotate jointly. The one deliberate trade-off is that the *direction* of a displacement is frame-dependent: the parameterization removes the origin, not the orientation of the frame. The random-walk training distribution (walks start at random lights and traverse all orientations, App. A.1) exposes the kernel to every orientation, so the learned coupling is *expected* to depend on separation more than on orientation—an empirical property, not a structural guarantee. This is also where the construction departs from ReRoPE (Li et al., 2026), whose camera block injects a full projective transform: a non-commutative family whose relative form does not collapse to a difference and which is not norm-preserving (they stabilize training by normalizing the translation component). LiRoPE inherits only the band-replacement finding; the transform content is an abelian translation, with the scale β spanning the repurposed channels' dynamic range.

Training does not converge without the rebinding. Fig. 10 complements the loss curves of Fig. 7 with validation grids from the same ablation pair under identical backbone, data, and training budget: the model with the light-axis rebinding produces coherent single-light predictions, whereas the temporal-RoPE variant without it fails to converge to comparable outputs, mirroring its stalled validation loss.

[1] Unit quaternions multiply within the sphere, so $q_i q_j^{-1}$ is itself a point of $\mathbb{S}^3$ representing the displacement from j to i—an intrinsic difference requiring no gauge and no ambient space. Had the light manifold been $\mathbb{S}^3$, group multiplication alone would provide the relative encoding.

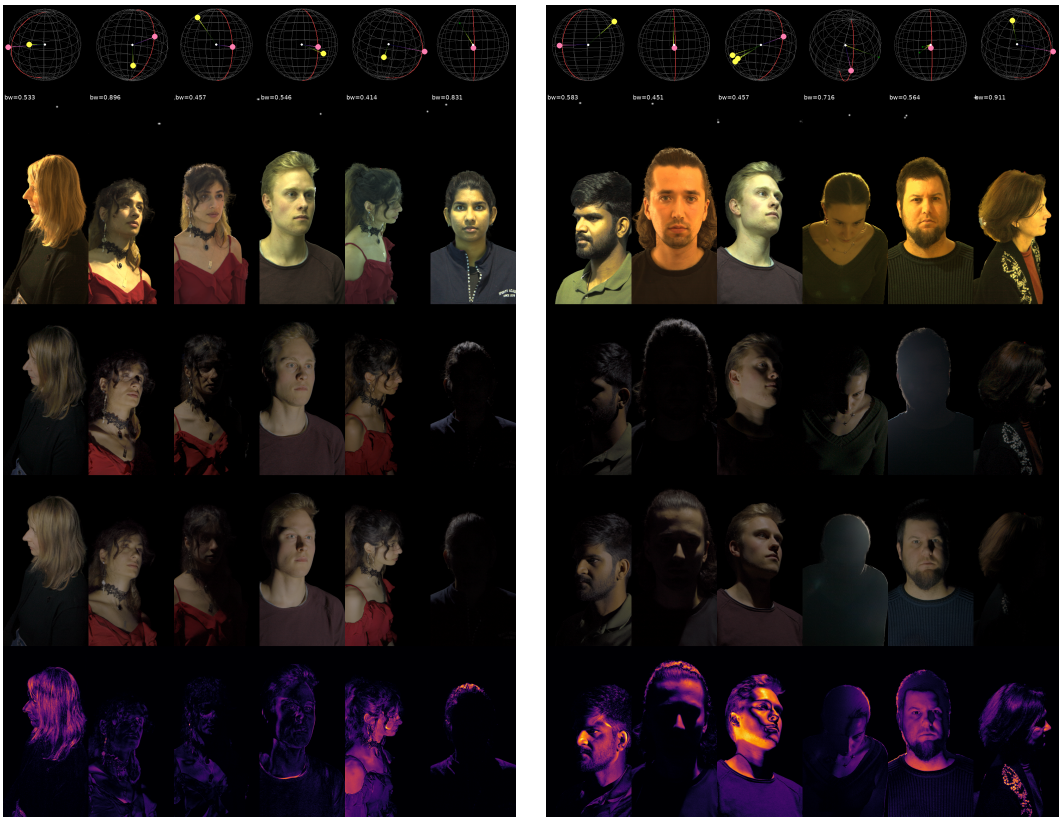

Figure 10: **Light-axis rebinding is necessary for convergence.** Validation grids of the ablation pair of Fig. 7 (same backbone, data, and training budget): with LiRoPE (left) the model converges to coherent OLAT predictions; with temporal RoPE (right) it does not.

B.4 HDR Encoding: Explicit Inverse

The encoding of Eq. 3 composes two maps. The inner map compresses radiance, $\tau(r) = \log\big(1 + \alpha\gamma\ \min(r, r_{\max})\big)\ /\ \log(1 + \alpha\gamma\, r_{\max}) \in [0, 1]$, and the outer map is the sRGB OETF,

$$f_{\mathrm{sRGB}}(\tau)\ =\ \begin{cases} 12.92\,\tau, & \tau \le 0.0031308, \\ 1.055\,\tau^{1/2.4} - 0.055, & \text{else}, \end{cases}$$

a strictly increasing bijection $[0, 1] \to [0, 1]$ (continuous at the junction with a bounded derivative; the small derivative kink is that of the sRGB standard itself), so the composed map is strictly increasing and hence invertible.

Explicit inverse. Given an encoded value $\tilde{r} \in [-1, 1]$, let $s = (\tilde{r} + 1)/2$ and recover

$$\tau\ =\ f_{\mathrm{sRGB}}^{-1}(s)\ =\ \begin{cases} s\,/\,12.92, & s \le 0.04045, \\ \big((s + 0.055)/1.055\big)^{2.4}, & \text{else}, \end{cases}$$

then decode the radiance

$$\hat{r}\ =\ \frac{\exp\big(\tau\ \log(1 + \alpha\gamma\, r_{\max})\big) - 1}{\alpha\gamma}.$$

The round trip $\hat{r}(\tilde{r}(r)) = r$ is exact for all $0 \le r \le r_{\max}$; values above $r_{\max}$ saturate to $\tilde{r} = 1$ by design. This is the compression ceiling, and $r_{\max}$=28 is chosen to cover the $\sqrt{331}$-scaled OLAT captures of App. A.1 (including specular highlights) with margin.

Jacobian. On $0 < r < r_{\max}$,

$$\frac{\partial \tilde{r}}{\partial r} \;=\; 2\, f'_{\mathrm{sRGB}}\big(\tau(r)\big) \cdot \frac{\alpha\gamma}{\big(1+\alpha\gamma\, r\big)\,\log(1+\alpha\gamma\, r_{\max})},$$

which is strictly positive and bounded away from zero and infinity (with the bounded kink of the sRGB standard at its junction): the round trip is numerically well-conditioned, with no vanishing or exploding slopes at either end of the range. The encoding enters training only as an offline transform of the ground truth—the diffusion loss is computed in latent space—so its Jacobian matters for round-trip conditioning, not for optimization. Constants: α=10, $\gamma\approx$3.6, $r_{\max}$=28.

Dynamic range and clipping. Across the full capture library (103 subjects, 75,137 frames, $\sqrt{331}$-scaled radiance), per-frame maxima have median 0.11, p_{99} 1.03, $p_{99.9}$ 4.35, and a global maximum of 26.2; only 0.09% of frames exceed 5. The compression ceiling $r_{\max}$=28 therefore covers the entire library with margin, and the clipping branch of Eq. 3 is empty in practice.

VAE reconstruction upper bound. Invertibility of the encoding does not by itself show that the frozen VAE carries the encoded signal. We measure the representation ceiling directly: ground-truth frames pass through the full training encoding chain ($\sqrt{331}$ scaling, color calibration, Eq. 3), then through the frozen Wan 2.1 VAE encoder and decoder, and back through the inverse encoding. Over 24 validation windows (1,944 frames, sampled exactly as training validation), this round trip attains 47.9±3.6 dB PSNR-μ / 0.983 SSIM and 42.8±3.2 dB / 0.949 SSIM in the LDR display domain, per frame; window means span 42.7–54.4 dB, so subject–camera pairs differ in their inherent ceilings. This certifies that the representation *contains* the information the field requires; the gap between this ceiling and the model's attained quality then measures the generation-side headroom (per-frame; composition averages part of the VAE error).

Choice of α. The compression slope α trades μ-domain fidelity against spurious high frequencies in the VAE reconstruction. A sweep at $r_{\max}$=28 shows PSNR-μ nearly flat across $\alpha \in [5, 100]$ (43.1 → 42.9 dB, 43.2 at α=10) with the LDR-domain PSNR equally flat (38.3–38.6 dB); both degrade only at α=1000 (42.2 / 37.8 dB; Fig. 8), where the reconstruction also develops periodic "fake texture" absent from the capture—its high-frequency energy rises to 1.7× the ground-truth level, against near-parity at $\alpha \le 30$. We operate at α=10: the flat region, with ground-truth-level high-frequency energy.

B.5 Pairwise-Displacement Statistics of Training Windows

The 81-light-trained checkpoints used in the window sweep and extrapolation analysis (Sec. 4.3) were trained on 81-point forward-biased random walks on the spherical Delaunay graph of the 331 calibrated lights (App. A.1). Because the repurposed channels depend only on relative displacement (App. B.2), what that training exposes them to is the *distribution of pairwise displacements* within windows. We measure it directly (Fig. 11): simulating the training sampler (20,000 windows of 81 lights), the within-window geodesic-separation distribution D_{train} differs markedly from the all-pairs distribution D_{full} of the 331-point calibration—total-variation distance 0.39, Wasserstein-1 18.3°. The forward bias is deliberate—consecutive training frames must be nearby, mutually consistent lights—so windows over-weight local pairs (28.3% vs. 6.4% of pairs below 30°) and under-weight distant ones (15.4% vs. 50.3% beyond 90°; 1.0% vs. 6.9% beyond 150°). Random starts give D_{train} support across the full separation range—no displacement is unseen—but exposure is heavily reweighted toward local geometry.

Two consequences follow. First, at the full-field length the repurposed channels operate mostly at displacements the kernel encountered at far lower frequency during training. Second, displacement magnitude and attention distance grow together, so the extrapolated regime leaves the trained distribution *jointly* in attention range and displacement statistics; the two cannot be separated in our data. The reversal measured beyond the trained length (Sec. 4.3) is consistent with this joint shift. Native whole-manifold training addresses both at once: training windows become 333-light graph walks, which cover on average half the sphere each yet bring the pair statistics close to the full field (TV 0.39 → 0.14, W_1 18.3° → 6.5°, with 0.5% zero-separation self-pairs from revisits; Fig. 11), and the attention range matches the inference window exactly.

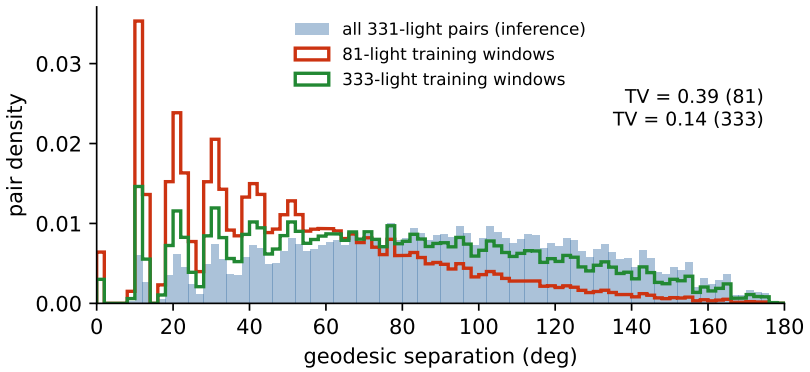

Figure 11: **Pairwise geodesic-separation distributions.** All pairs of the 331-point calibration (inference population) vs. within-window pairs of the two training regimes: 81-light walks (20k simulated, TV=0.39) and 333-light walks (3k simulated, mean 49% sphere coverage per window, TV=0.14). Both walks support the full range but over-weight local pairs; the 333-light window largely closes the gap.

B.6 An Integration Error Model

Setup. Let the predicted field be $\hat{\mathbf{R}}(\mathbf{d}_l) = \mathbf{R}(\mathbf{d}_l) + \boldsymbol{\epsilon}_l$. Relighting (Eq. 8) is linear, so the relit error under an environment with fitted weights $w_l \geq 0$ is exactly

$$\Delta \;=\; \sum_l w_l\, \boldsymbol{\epsilon}_l \;:$$

per-light error passes into the relit image weighted by the environment, regardless of how the field was produced. Write $w_l = s\,\bar{w}_l$ with $\sum_l \bar{w}_l = 1$: the environment's total energy s (spanning four orders of magnitude across the battery, App. C.1) scales signal and error alike and is absorbed by display anchoring, while the combination of the $\boldsymbol{\epsilon}_l$ is governed by their correlation structure under $\bar{w}$. For a random-error model with means $\mathbf{m}_l = \mathbb{E}[\boldsymbol{\epsilon}_l]$ and cross-light covariance C_{lm},

$$\mathbb{E}\|\Delta\|^2 \;=\; s^2\Big(\big\|\textstyle\sum_l \bar{w}_l\, \mathbf{m}_l\big\|^2 + \sum_{l,m} \bar{w}_l \bar{w}_m\, \mathrm{tr}(C_{lm})\Big)\,.$$

Two regimes bracket this decomposition. If the $\boldsymbol{\epsilon}_l$ were zero-mean and uncorrelated, composition would *average* them ($\mathbb{E}\|\Delta\|^2 = s^2\sigma^2/L$ at uniform $\bar{w}$): relighting would be self-correcting, and no coupling would be needed. The regime of this paper is the opposite one: per-light errors of independent generators are *structured*—small misalignments and gains, nonzero-mean and correlated through shared conditioning—so they survive composition, and their heterogeneity is what smears the sum. Per-image quality controls $\|\boldsymbol{\epsilon}_l\|$; *consistency across lights* controls the cross terms $\bar{w}_l \bar{w}_m\, \mathrm{tr}(C_{lm})$.

Shared bias vs. idiosyncratic drift. Write $\boldsymbol{\epsilon}_l = \mathbf{b} + \boldsymbol{\delta}_l$: a shared component $\mathbf{b}$ (an identity-, geometry-, or calibration-consistent bias common to the whole field) and an idiosyncratic drift $\boldsymbol{\delta}_l$. Two facts follow. (1) The shared bias does not accumulate: $\sum_l \bar{w}_l \mathbf{b} = \mathbf{b}$ contributes one coherent shift (scaled by s) to every relit image alike. (2) The drift is invisible per image but visible in the integral. To first order, model per-light inconsistency as a geometric misalignment $\mathbf{t}_l$ and a radiometric gain a_l: $\hat{\mathbf{R}}_l \approx (1{+}a_l)\mathbf{R}_l + \nabla\mathbf{R}_l \cdot \mathbf{t}_l$. Each term can be small in per-image metrics, yet

$$\Delta \;=\; \sum_l w_l\big(\nabla\mathbf{R}_l \cdot \mathbf{t}_l + a_l\, \mathbf{R}_l\big) + O(\|\mathbf{t}\|^2)\,:$$

heterogeneous $\mathbf{t}_l, a_l$ accumulate into blur and ghosting—each light contributes a differently placed shadow edge or a differently scaled highlight, and their weighted sum smears exactly the high-frequency structure (shadow boundaries, catchlights) that relighting exists to reproduce. Under full coupling the drifts are shared, $\mathbf{t}_l = \mathbf{t}$ and $a_l = a$ for all l, and Δ collapses to a single coherent warp and gain, $\big(\sum_l w_l \nabla\mathbf{R}_l\big)\cdot\mathbf{t} + a\sum_l w_l \mathbf{R}_l$—one shared bias rather than an accumulated one. This is

the analytic form of the observations that motivate our design: misalignments that "may accumulate into noise or flicker when linearly combined" (Rao et al., 2025), and the smoothed detail and global shading bias POLAR observes across independently predicted OLAT sets (Chen et al., 2026).

Predictions. The model predicts the trend measured in Sec. 4.3: (i) per-OLAT quality is nearly flat in the coupling window F across in-range windows (the $\boldsymbol{\epsilon}_l$ are individually small regardless); (ii) relit quality rises with F, as raising F replaces heterogeneous drifts by shared ones; (iii) cross-light identity consistency—which measures the dispersion of $\mathbf{t}_l, a_l$ across lights—rises with F; and (iv) the running weighted-mean error $\big(\sum_{l\le k} w_l \boldsymbol{\epsilon}_l\big)/\big(\sum_{l\le k} w_l\big)$ is flat in k under full coupling—it *is* the shared bias—whereas under independent generation it sits above that floor and fluctuates with the accumulated idiosyncratic component. It further predicts (v): a generator whose per-light errors are structured rather than zero-mean—compressed-context inference being the canonical case—shows a composite deficit *exceeding* its per-OLAT deficit, since composition removes no correlated or misaligned component (uncorrelated zero-mean error would instead average *down*; Sec. 4.3).

C Experimental Protocols

C.1 Single-Image Relighting Comparison (Sec. 4.2)

Subjects and viewpoint. All relighting comparisons use the 13 held-out subjects, evaluated exclusively at camera Cam06. Cam06 is close to frontal, matching the frontal-portrait input assumption shared by IC-Light, Lite2Relight, NeuralGaffer, and RITW; restricting evaluation to this single viewpoint removes viewpoint as a confound and is the setting most favorable to the baselines. All baselines run from their released checkpoints, zero-shot on our test set, with no FaceOLAT adaptation.

Ground-truth synthesis. Relighting ground truth is synthesized from the OLAT bases by linear composition: for a target environment map, per-lamp weights are fitted from the map (whose center corresponds to the $+Z$ frontal incident-light direction), and the ground truth is the weighted sum of the 331 captured single-light images in linear HDR. The same weights and the same environment-map alignment define the conditioning of every method, rendered to each method's native format (for SimStage and the envmap-conditioned baselines, the composited equirectangular map), so conditioning and ground truth share one definition and cannot diverge.

Environment battery. All 6,651 retained candidate maps are used (the 6,808-map pool with 157 corrupted entries removed); their total composited energy spans 6.3 to $\sim 1.2\times10^5$, over four orders of magnitude. The same maps are used for all subjects.

Scoring canvases. Two canvases are scored: the full-frame 256×512 canvas (the model's input domain, to which the native 270×512 capture is resized) and the 512×512 five-point-aligned face window (the square-portrait domain the baseline relighters are designed for). On the square canvas, prediction, ground truth, and person mask are registered by one shared per-subject five-point alignment onto a single sampling grid—the same transform on both sides of every comparison.

Metrics. *SimStage rows*: PSNR/SSIM in LDR display space under the golden chain—HDR-to-LDR anchored to the ground truth (in-mask p_{99} luminance $\to 0.5$, white point $p_{99.99}$, extended Reinhard, $\gamma\,2.2$), person mask applied after tonemapping (App. C.5). *Baseline rows*: each method is scored in its official output convention—its released LDR rendering as-is, against the ground-truth LDR render, both normalized by their own whole-image p_{99} (a self-anchored display: global exposure errors are forgiven rather than penalized; no exposure alignment to the ground truth, no color calibration), person mask multiplied in the linear domain. The two chains differ in anchor and masking domain, so only same-chain comparisons are drawn. Identity consistency is the ArcFace cosine between the predicted and ground-truth relit portraits (8-bit sRGB, five-point-aligned 112×112).

IC-Light full-frame protocol note. The daggered IC-Light full-frame entry of Table 2 is scored in the protocol of its own output: the whole image (no person mask) in PNG display encoding, anchored to the image's own p_{99}. Because both the anchor and the masking differ from every other entry, it is not comparable to any of them—under that protocol its per-subject SSIM also exceeds SimStage's on a minority of subjects, which is why we flag the mismatch rather than average it away.

Table 4: **Baselines without runnable code: results as reported on each method's own benchmark.** Not directly comparable across rows or to our test set: SynthLight reports on its real light-stage test set (its synthetic test set is in-distribution for SynthLight); DifFRelight trains per subject and is evaluated on held-out novel lighting, with backgrounds masked before scoring; Text2Relight is text-conditioned, scored against simulated ground truth.

Method	LPIPS↓	SSIM↑
SynthLight (Chaturvedi et al., 2025)	0.165	0.813
DifFRelight (He et al., 2024)	0.175	0.821
Text2Relight (Cha et al., 2025)	0.401	0.546



Figure 12: **OLAT predictions versus ground truth.** One subject (camera Cam13): five groups of calibrated lights, each group shown as a ground-truth row and a prediction row over 22 ordered lights. Qualitative counterpart of Table 1.

C.2 Coupling-Window Sweep (Sec. 4.3)

The sweep varies only how many lights are generated jointly, holding weights, conditioning, and the environment readout fixed. **Variant A (independent windows).** The 331 lights are split into consecutive, non-overlapping windows along the manifold trajectory, each denoised by a separate pass: window size $F \in \{1, 5, 9, 17, 25, 41, 57, 81, 113, 161, 209, 257, 333\}$ (the seven evaluation subjects at $\{9, 25, 81, 113, 333\}$, three of them at $F=1$; the remaining sizes on a single training subject, so those points carry subject rather than population weight, and consecutive points differ in subject composition). $F=1$ reproduces independent per-frame generation—the loosest form of the per-light paradigm of 3DPR and POLAR, which couple lights only through a shared latent or source portrait—using our own model; note that it is simultaneously the most length-extrapolated point of

the sweep for these weights. $F \in [9, 81]$ stay within the training window and carry the coupling-vs-length isolation; $F > 81$ probe the extrapolated regime of Sec. 4.3. Changing F also changes the per-pass latent grouping of the temporal VAE, a path shared by every windowed method; Variant B removes it. **Variant B (attention masking).** A single 331-frame pass with one shared initial noise; the temporal attention receptive field is attenuated beyond bandwidth F (additive log-λ bias, applied identically to self-attention in every block and every head; reference tokens are exempt). This holds fixed everything Variant A varies—sampling, noise, and the VAE path—leaving attention bandwidth the only moving part. Two limits remain: masked attention can still relay information across larger spans through intermediate tokens, so the network's effective receptive field exceeds the per-layer bandwidth; and the $4\times$ temporal VAE mixes neighboring frames within each latent in all variants. Against F we report: (i) per-OLAT quality (PSNR-μ); (ii) relit quality (PSNR-μ/SSIM-μ, LPIPS) under a fixed battery of test environments; and (iii) cross-light identity consistency (ArcFace cosine over all pairs of generated lights). The prefix-sum diagnostic plots $\| \sum_{l \le k} w_l \, \epsilon_l \|$ against k along a fixed light ordering for $F{=}1$ vs. $F{=}333$.

C.3 Densification (Sec. 4.4)

Evaluation at uncalibrated directions. Ground truth exists only at the calibrated lamps, and every lamp belongs to the training graph of the released checkpoints—a strict training-level hold-out would require retraining, which we defer. Per-light conditioning draws only on the direction (its environment map), so conditioning-side leakage is excluded; quantitative evaluation at calibrated directions therefore measures the consistency of the uncalibrated inference path on training-seen directions. Between lamps, the evidence is qualitative: OLATs generated at uncalibrated directions, shown beside their nearest calibrated neighbors, with spherical interpolation (slerp) and an SH fit of the neighboring captures as ground-truth-only reference lines. **Densified relighting (convergence).** We upsample the field to 662 and 993 directions ($2\times$ and $3\times$; new directions refine the partition between the calibrated ones). The densified trajectory is generated as independent 333-frame native windows—each a full native-length pass, coupling the new directions with the calibrated lights it contains—and we relight with each densified combination, re-fitting per-direction weights from each environment map over the densified direction set by the same procedure as the calibration. Because the light stage provides no ground truth between its lamps, Relit@N is scored against the ground-truth 331-light composite—the operational ground truth of the task—and we additionally report the mutual convergence of the 331, 662, and 993 composites as a stability diagnostic: composites that stabilize indicate that refinement injects no inconsistency, whereas refinement that injects inconsistent OLATs drifts as the density grows. 3DPR's triplane likewise supports continuous light queries and reports 1324-OLAT synthesis (Rao et al., 2025); we run the identical densification on 3DPR for a head-to-head comparison, evaluating both the added OLATs' consistency with the rest of the field and the relit quality of each densified combination.

C.4 HDR Fidelity Metrics (Sec. 4.5)

Highlight metrics. Reconstruction is evaluated inside the top 0.1% and 1% ground-truth-luminance regions (where catchlights live), and dynamic-range fidelity is reported as the ratio of 99.9th-percentile to median radiance, predicted vs. ground truth. LDR or tonemapped prediction caps this ratio structurally—3DPR's supervision hard-clips radiance at $5\times$ before tonemapping (App. A). **Energy linearity.** We relight under $\mathbf{e}$, $2\mathbf{e}$, and $4\mathbf{e}$ and measure each method's deviation from $I(c\,\mathbf{e}) = c\,I(\mathbf{e})$; a linear field satisfies it exactly by construction, while saturation in any non-linear pipeline breaks it. The test requires no additional ground truth; it certifies the prediction domain—linear vs. saturating—while basis accuracy is carried by the ground-truth-anchored metrics. **Sparse, extreme environments.** Under environments dominated by a few narrow, bright sources, highlight fidelity dominates perceptual quality; we evaluate relighting on such HDRIs, following the sparse-lighting protocol of Rao et al. (2025).

C.5 Long-Window Exclusion Study (Sec. 4.3)

Controlled comparison. Both systems fine-tune the Wan 2.2 5B base with the same LoRA recipe for the same number of steps (243k): *independent windows*—our pipeline trained with 65-frame windows, whose field is composited by linearly combining all 331 OLATs under each

test environment—and *FramePack* (Zhang et al., 2025a)—the same base adapted to next-frame-section prediction with a packed, importance-weighted context. An independent-window control 10k steps earlier (233k) matches the 243k control to within 0.03 dB on every relit metric (LDR PSNR/SSIM/LPIPS 25.02/0.866/0.149 vs. 25.05/0.867/0.149), ruling out training budget as the cause. The FramePack side necessarily differs in training objective (next-frame-section prediction), packing layout, sampling schedule, and per-step token counts—base, LoRA recipe, and step count are held fixed, and the comparison spans the full adaptation recipe. **Walkthrough protocol (per-OLAT).** 282 entries $\times$ 331 lamps (93,342 frames per system). LDR metrics use `linear_to_display` at a fixed exposure (OLAT frames carry no white point to anchor); PSNR-μ and SSIM-μ are computed in HDR-linear space. Frames are scored without masks, as in training validation; the *non-black* subset (89,671 lamps) restricts to lamps whose ground-truth display-domain p_{99} luminance is at least 0.10. **Golden protocol (composite relighting).** 282 entries $\times$ 256 brightness-stratified environment maps drawn from the pool of App. C.1 (72,192 composites per system). Predictions and ground truth are each linearly composited over the 331 frames with cached environment weights; metrics are computed in the companded domain and in an LDR domain obtained by luminance-domain Reinhard tonemapping anchored per image to the ground truth's p_{99}/white point, after multiplying by a linear person mask. Fig. 13 shows the full results.

C.6 Feasibility Spectral Analysis (Sec. 4.3)

Protocol. All spectra are computed on the matrix whose rows are the 331 OLAT frames (frame-id order) and whose columns are person-region pixel–channel values ($\sim$40k pixels, max-frame luminance $> 10^{-3}$, HDR-linear domain), at model resolution. Light directions are obtained by least-squares sphere fitting of the calibrated lamp positions (center $[606, 1551, 536]$ mm, radius 1707 mm, residual std 3.2 mm); normalizing about the world origin instead introduces direction errors of median 43° and inflates cross-validation scores by compressing angular distances—all direction-dependent statistics reported here use the fitted center. GT statistics cover the full 130-subject library (first expression per subject); generated-field statistics use the 7-subject sweep subset with per-subject matched GT. Definitions: r_{99}/r_{999} = number of SVD components covering $99/99.9\%$ of squared energy; L^* = smallest SH degree reaching 99% of the best validation R^2 (ridge regression, 5-fold over held-out light directions); *tail* = $1 - \mathrm{cum}_{\mathrm{pred}}[r_{99}(\mathrm{GT})]$, the energy left outside the GT dimension budget—a spectral quantity (the prediction's own cumulative spectrum, truncated at the GT's dimension count) that measures complexity match, not alignment with the GT subspace, since differently oriented subspaces can share a spectrum; *relative erank* = $\mathrm{erank}_{\mathrm{pred}}/\mathrm{erank}_{\mathrm{GT}}$ (exponential-entropy effective rank); adapter statistics come from the closed-form task vector $\Delta W = (\alpha/r)BA$ (rank 128, α/r=1, 400 modules per adapter), with singular values via a QR identity and random baselines from 200 Gaussian subspace pairs.

Projection residual. The tail is spectral by construction, so we additionally report a direct subspace quantity: with U_r the top-$r_{99}(\mathrm{GT})$ left singular vectors of the GT matrix, $\mathrm{resid} = 1 - \|U_r^\top X_{\mathrm{pred}}\|_F^2 / \|X_{\mathrm{pred}}\|_F^2$—the prediction's energy outside the GT row space. The two measures part ways exactly where the spectral one is blind: the collapsed F=1 field shows the *smallest* tail (0.003 median over its three subjects—a rank-deficient spectrum fits any dimension budget) but the *largest* residual (0.38—its energy points away from the GT subspace), and across the sweep the residual falls monotonically with the window ($0.25 \rightarrow 0.12$ median from F=9 to 333; per-subject 0.026–0.999 at F=333, the outlier being a subject whose GT subspace is exceptionally small, r_{99}=4, with prediction energy concentrated in the GT's sub-r_{99} weak directions). The residual thus carries the window trend as a subspace quantity, complementing the tail.

Task structure. The energy separates into a sparse high-energy layer and a band-limited diffuse layer: *specular-proxy* columns (per pixel–channel spikiness ≥ 0.5—a statistical proxy for the cross-polarization specular layer, not a physical separation) are a median 36% of columns but carry a median 98% of squared energy (minimum 70.7%), and are not predictable by low-order SH ($R^2 \approx 0$), while the diffuse layer's SH bandwidth saturates at L^*=4–5 (IQR, max 6; Fig. 14): 25 coefficients capture $\sim 72\%$ of held-out diffuse energy. The SH dimension bounds only the smooth part—the gap to the total diffuse dimension is a median $2.9\times$ (Fig. 16), the non-smooth residual of shadow boundaries and lamp-ring gaps (Fig. 15).

Behavior: fixed weights, varying only the inference window. With the 81-trained weights held fixed, the tail decreases monotonically with the inference window for $7/7$ subjects over the full-

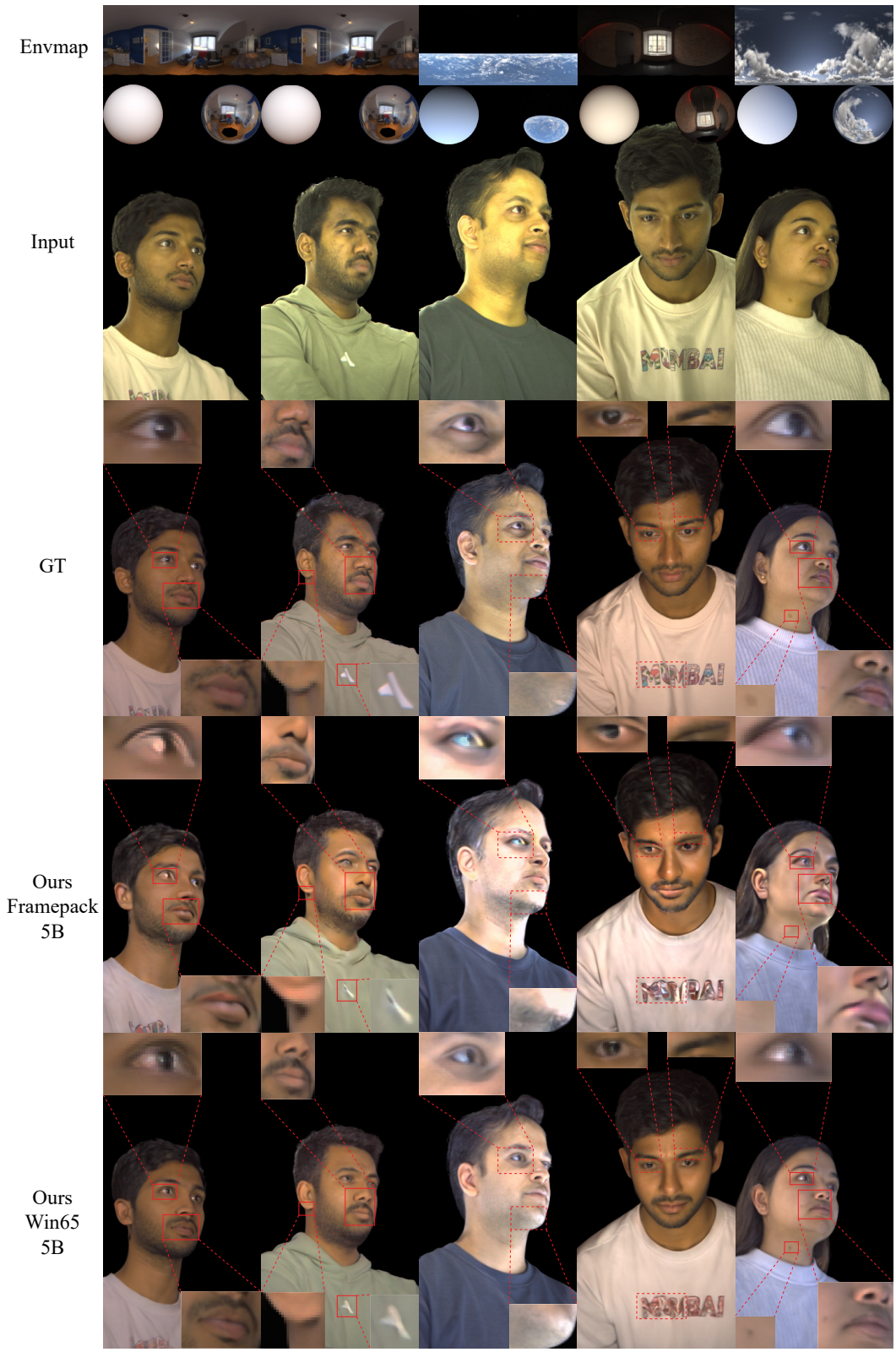

Figure 13: **Compressed-context exclusion, full results (5B, matched weights).** (a) Per-OLAT vs. composite relit PSNR in both domains; (b) frame-level per-lamp PSNR distribution.

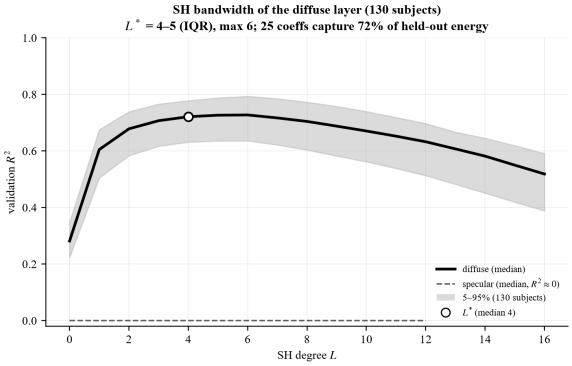

Figure 14: **SH bandwidth of the diffuse layer** (130 **subjects).** Validation R^2 vs. SH degree (median $\pm$ 5–95%): diffuse (solid) saturates at L^*=4–5; specular proxy (dashed) stays at $R^2\approx 0$.

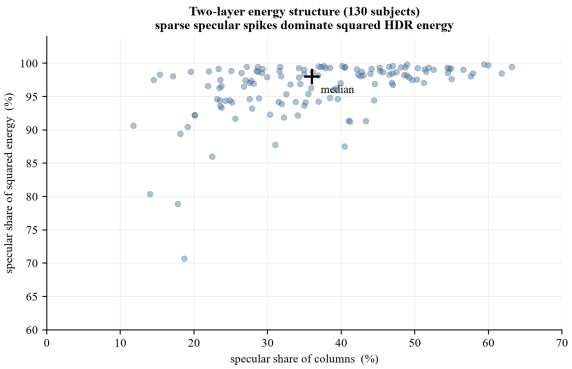

Figure 15: **Two-layer energy structure** (130 **subjects).** Specular-proxy share of squared energy vs. share of columns: sparse spikes dominate the squared HDR energy (median 98%).

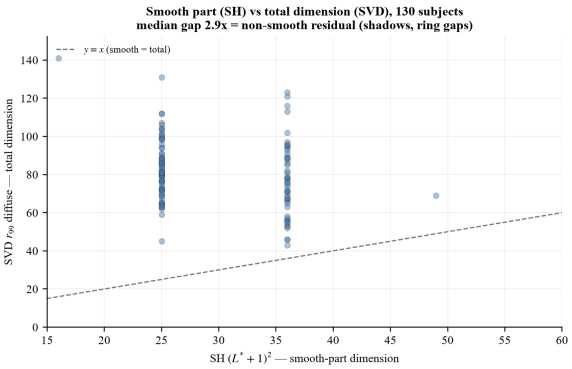

Figure 16: **Smooth part vs. total dimension.** $(L^*+1)^2$ vs. diffuse r_{99} per subject; all points lie above $y=x$ (median gap 2.9× = non-smooth residual of shadows and lamp-ring gaps).

sizes $\{9, \ldots, 333\}$ (0.088→0.036, 7-subject median; Fig. 17)—longer joint contexts denoise the spectrum (F=1 is the blind spot of the spectral tail: its collapsed spectrum fits any budget, which is why the projection residual above carries it). Relative erank declines smoothly with the window for every subject; its distance to the GT level is minimized at the training length for $5/7$ subjects, and extrapolating beyond it over-flattens typical subjects below the GT effective rank while the two noisiest subjects remain above it (Fig. 18). The single-frame regime collapses to a mean-field

predictor: win1 concentrates 97% of the energy in its first component (GT: 43%) at 0.27× the GT effective rank (Fig. 19).

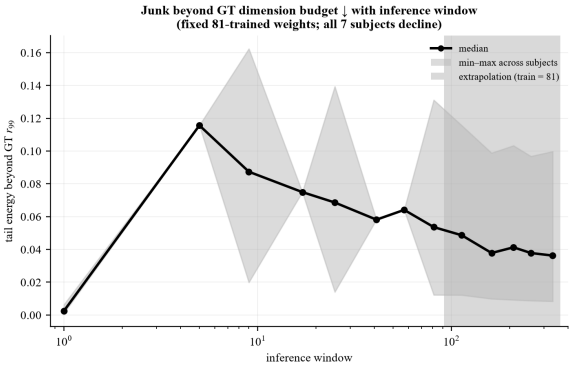

Figure 17: **Consistency vs. inference window (fixed 81-trained weights).** Tail energy beyond the GT dimension budget; median ± min–max across the 7-subject subset. Shaded region: beyond the training window.

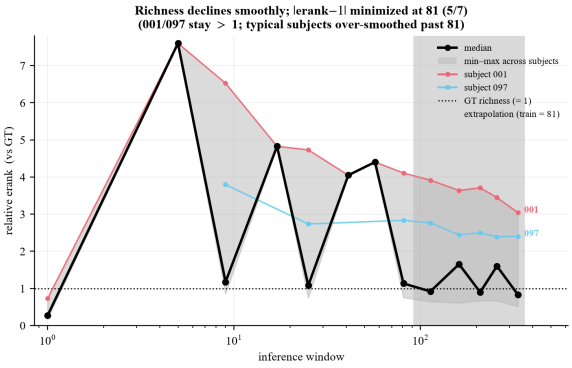

Figure 18: **Richness vs. inference window.** Relative erank; the two noisiest subjects (001, 097) are highlighted. $|\text{erank}-1|$ is minimized at the training window for $5/7$ subjects.

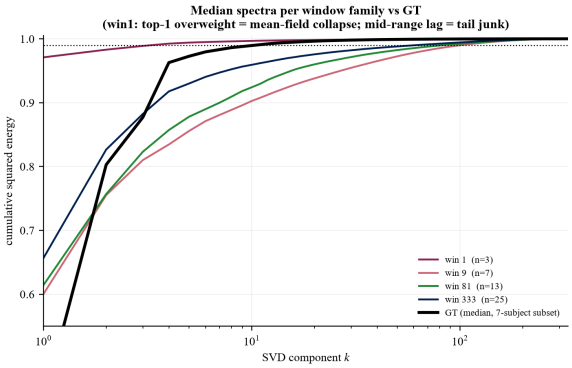

Figure 19: **Median cumulative spectra per window family vs. GT.** All generated fields overweight the first component; the mid-range lag ($k \approx 3$–50) is the tail of Fig. 17 made visible.

Weight side. The 333-light task vector matches the 81-light one in every spectral readout: effective ranks of 2–5 per module class (of 128), output-subspace top-1 overlap 0.718/0.775 for the high/low adapter pairs (random: 0.308), and task-vector alignment 3–4 orders above the random baseline,

concentrated in self-attention (Figs. 20–21). Training-step totals differ (122k vs. 8k; the latter is the validation-final checkpoint) and spectra are stable across late checkpoints, so the comparison reflects converged adaptation. The rank bound is imposed by construction; the evidence is the fraction consumed, the cross-adapter agreement, and the distance from random subspaces. Both adapters fine-tune the same frozen base, so cross-adapter subspace agreement partly reflects the base's existing feature directions—the random baseline controls dimension alone.

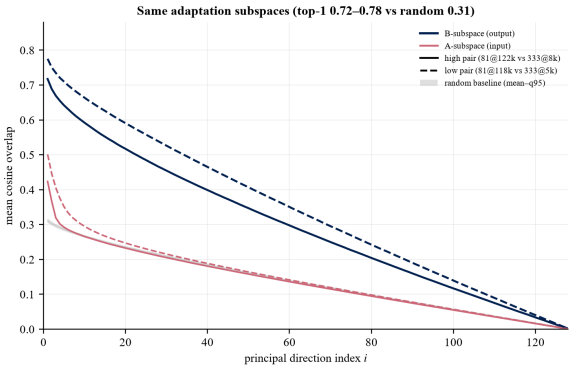

Figure 20: **Adaptation subspaces coincide.** Mean principal-angle cosine spectra of the output (B) and input (A) subspaces of ΔW for the 81 vs. 333 adapter pairs, against the random baseline band.

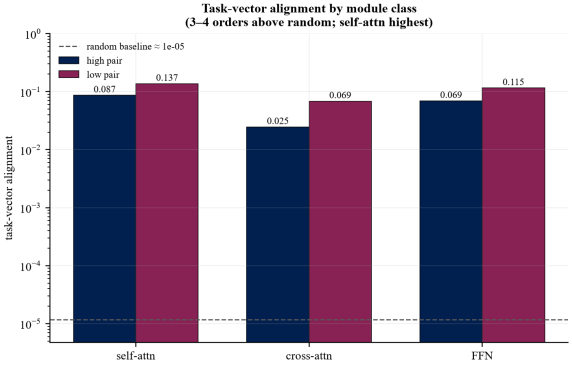

Figure 21: **Task-vector alignment by module class.** Alignment $\mathrm{tr}(\Delta W_1^\top \Delta W_2)/\|\Delta W_1\|_F \|\Delta W_2\|_F$ (log scale): 3–4 orders above random, highest in self-attention.

Mechanism: coupling length, not seam count. Window length W and window count $N{=}\lceil 331/W \rceil$ are collinear in the sweep, so we disentangle them twice. First, a fixed-count control: 113/161 lights share $N{=}3$ and 209/257 share $N{=}2$; lengthening the window at fixed N improves the tail in $8/8$ paired comparisons (the paired sizes necessarily differ in light composition and sequence positions) (relative erank: $4/4$ and $3/4$, one flat). Second, a per-frame seam test: the residual of each frame against the matched GT's top-r_{99} subspace, paired against the single-window (333) baseline to cancel per-frame difficulty, shows no near-seam concentration—the near/far contrast flips sign across the seven families ($3+/4-$), the only nominal effect ($+0.007$, $p{=}0.035$) fails a 7-family Bonferroni correction, and one family is reverse-monotone (Fig. 22). The improvement is consistent with in-window coupling length carrying the effect. This does not contradict the sliding-window exclusion above: the failure of windowed schemes on relighting metrics is not attributable to per-frame degradation at seams—it is consistent with cross-frame inconsistency between individually clean frames, which composition exposes and which the native window suppresses at generation time.

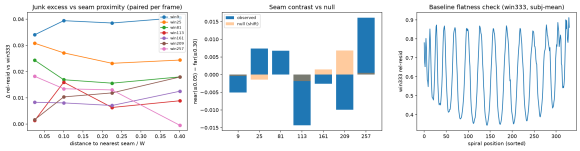

Figure 22: **Per-frame seam test.** Frame residual against the GT top-r_{99} subspace (paired minus the single-window baseline), binned by relative distance to the nearest window boundary: no near-seam concentration across families.

C.7 In-the-Wild Qualitative Study (Sec. 4.6)

Figure 23 shows the qualitative grids summarized in Sec. 4.6. Columns are subjects. The Right, Mid, and Left rows reuse a single environment, composing the basis with lights drawn only from its right, middle, and left regions, so the same identity is seen under disjoint subsets of one illumination. The remaining rows apply full image-based environments, each specified in the conditioning column by its environment map together with chrome and diffuse probes.



Figure 23: **In-the-wild virtual light stages.** One OLAT basis is recovered per input portrait (top row) and linearly composed into the target environments of each row; the conditioning column shows the environment map with its chrome and diffuse probes.