LIGHTDIAL: PROGRESSION-CONDITIONED ATTENTION FOR CONTINUOUS PORTRAIT RELIGHTING

Anonymous authors
Paper under double-blind review

ABSTRACT

Portrait relighting plays an important role in portrait enhancement and creative editing by reshaping how illumination defines facial structure, subject appearance, and visual atmosphere. Despite recent advances in instruction-based image editing, current models still struggle to produce consistently plausible and visually coherent portrait lighting across diverse effects, while offering limited control over how strongly each effect is expressed. A key challenge is that increasing lighting strength is not a uniform change in edit magnitude: different effects evolve through distinct, spatially structured changes in illumination. In this work, we present LightDial, which formulates lighting strength through a continuous progression variable. This progression-based formulation turns strength control from magnitude scaling into learning structured changes along an ordered lighting progression. By modulating the interactions among text, reference-image, and noisy-output representations according to the requested progression state, and by explicitly constraining response changes between neighboring states during training, LightDial enables smooth and structured evolution of lighting effects while preserving identity and scene content. To provide the effect diversity and ordered progression supervision required by this formulation, we construct PortraitLight-50K, a large-scale portrait relighting dataset covering structurally diverse lighting effects with multi-level progression sequences. Extensive experiments demonstrate high-quality portrait relighting and strong identity preservation, together with smoother and more consistent progression responses than existing continuous-control approaches.

1 INTRODUCTION

Portrait relighting changes the spatial distribution of illumination to shape facial appearance and scene atmosphere. Contour highlights, projected shadows, localized bands, and directional light shafts produce different photographic effects while the subject's identity and scene content should remain fixed. Instruction-based image editing makes these changes accessible through natural language (Brooks et al., 2023; Wu et al., 2025; Black Forest Labs, 2025). An interactive editor must additionally control how strongly the requested lighting effect is expressed.

This control is difficult because lighting strength is effect specific. Hair lighting develops along the subject contour; Slit lighting strengthens a localized illumination band; Rembrandt lighting changes the balance between lit and shadowed facial regions. Tyndall, Dappled, and Sunset lighting follow still different spatial developments. Increasing strength therefore cannot be assumed to scale one fixed edit direction. Each intermediate state must realize the intended lighting structure while preserving the portrait. A full-strength target specifies the final appearance, but does not specify which intermediate states or local changes should connect it to the source.

Existing continuous editing methods parameterize edit magnitude through learned modulation, parameter or token directions, and inference-time guidance (Gandikota et al., 2024; Parihar et al., 2026; Zarei et al., 2026; Zhang et al., 2025b). These mechanisms can provide continuous control, but endpoint supervision alone does not define the effect-specific path between the source and final edit. We instead formulate portrait lighting strength as a continuous *progression variable* s: for a fixed source

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.



Figure 1: **LightDial for continuous portrait relighting.** The six effects span different spatial illumination patterns. Each row illustrates an effect-specific sequence at increasing progression states; rows may use different source portraits.

and instruction, each value requests one state along an ordered lighting progression. The progression variable is sequence relative rather than a calibrated physical or perceptual intensity shared across portraits.

Learning this formulation requires supervision for both state appearance and local state relationships. PortraitLight-50K supplies seven-level sequences for six structurally distinct effects. Its retained intermediate targets are selected from complementary source–endpoint constructions and endpoint-conditioned transitions, followed by frame- and sequence-level verification of lighting structure, preservation, and order. The dataset thus provides multi-state supervision that is absent from endpoint pairs alone.

LightDial introduces the progression state where the instruction, preserved source content, and evolving output representation interact. TriDial applies group-specific orthogonal rotations to attention queries and keys, so s changes relative cross-group compatibility without directly scaling query or key norms. Strength-Aware Alignment (SAA) adds matched-path finite-difference supervision between adjacent targets. It leaves the pointwise flow-matching target unchanged while regularizing how neighboring prediction responses vary. One adapted model supports all six effects and directly predicts the requested state from the source, instruction, and s (Figure 1).

Our contributions are threefold. First, we formulate portrait relighting strength as an effect-specific ordered progression and instantiate it through progression-conditioned relative interaction modeling and local neighboring-response supervision. Second, we construct PortraitLight-50K, a verified multi-level dataset spanning six spatially distinct lighting effects. Third, we evaluate endpoint effect fidelity, image quality, identity and content preservation, unseen intermediate control, automatic response behavior, and human-perceived sequence quality. This evidence separates whether the lighting effect is correct from whether its response is ordered, regular, and perceptually smooth.

2 Related Work

2.1 Instruction-based Image Editing

Instruction-based editing uses natural language to specify changes to an input image. InstructPix2Pix established scalable paired instruction supervision (Brooks et al., 2023), and subsequent systems expanded editing coverage and content preservation (Sheynin et al., 2024; Yu et al., 2025; Wu et al., 2025; Black Forest Labs, 2025). A complementary direction learns image-to-image transformations in generative latent spaces; LBM, for example, applies latent bridge matching to translation tasks including relighting (Chadebec et al., 2025). LightDial builds on a flow-based editor but supervises

multiple ordered states for each source–instruction pair rather than only a source–target transformation.

2.2 Portrait Relighting

Portrait relighting methods use illumination representations, shading cues, and paired observations to modify lighting while retaining subject appearance (Zhou et al., 2019; Pandey et al., 2021; Ponglertnapakorn et al., 2023). Controllable Light Diffusion adjusts the softness of portrait shadows and highlights (Futschik et al., 2023); IntrinsicEdit and Latent Intrinsics operate through explicit or learned intrinsic representations (Lyu et al., 2025; Zhang et al., 2024). IC-Light uses light-transport consistency (Zhang et al., 2025a), while CFT supervises velocity relations across image pairs that share an illumination transformation (Zhang et al., 2026). Synthetic and captured datasets support lighting control under varied conditions (Chaturvedi et al., 2025; Magar et al., 2025; Chen et al., 2026). LightDial focuses on ordered development within the same source–instruction sequence. SAA relates neighboring states of that sequence; this pairing differs from, but is complementary to, CFT's cross-image transformation consistency.

2.3 Continuous Editing Control

Concept Sliders and SAEdit expose continuous attributes through learned parameter or token directions (Gandikota et al., 2024; Kamenetsky et al., 2025). Kontinuous Kontext maps a scalar strength to learned modulation, and SliderEdit provides instruction-aware continuous control through low-rank adapters (Parihar et al., 2026; Zarei et al., 2026). Text-embedding interpolation, CFG and LoRA scaling, GRAG, and FlowSlider provide embedding-, parameter-, guidance-, or attention-space control at inference (Ekin & Gandelsman, 2026; Zhang et al., 2025b; Endo et al., 2026). FreeMorph and video generation can construct transitions when an endpoint is available (Cao et al., 2025; Team Wan, 2025). These approaches demonstrate that continuous editing does not require a single mechanism. LightDial specifically learns effect-dependent spatial lighting progressions from ordered multi-state supervision: s conditions relative interactions among text, reference, and output streams, and adjacent states supervise local response changes. At inference, it requires the source, instruction, and requested state, but not a full-strength endpoint.

3 Methodology

3.1 Proposed Method

Lighting progression couples the requested effect, the source structure to preserve, and the appearance being generated. LightDial conditions the interactions among these representations and learns local changes from adjacent targets. TriDial provides the interaction parameterization; SAA aligns neighboring velocity responses with their target differences (Figure 2).

Let x be the source portrait, p the relighting instruction, and $s \in [0, 1]$ the continuous progression variable. We learn

$$\hat{y}_s = F_\phi(x, p, s; \epsilon), \tag{1}$$

where ϵ denotes generation noise. Each s specifies a requested progression state for the same source and instruction. The training anchors are the source at $s = 0$ and the accepted full-strength edit at $s = 1$. The labels encode within-sequence order, not a common physical or perceptual scale across portraits. Each state is generated directly, with s held fixed while flow time t evolves.

Progression-conditioned flow matching. For target latent z_s, flow matching (Lipman et al., 2023) constructs the noisy latent state and target velocity as

$$z_t = (1 - t)z_s + t\epsilon, \qquad v_s^* = \epsilon - z_s, \qquad \mathcal{L}_{\mathrm{FM}} = \mathbb{E}\|v_\phi(z_t, t, x, p, s) - v_s^*\|_2^2. \tag{2}$$

This objective supervises the appearance at each sampled progression state. TriDial introduces s into attention, and SAA adds supervision for the difference between adjacent states.

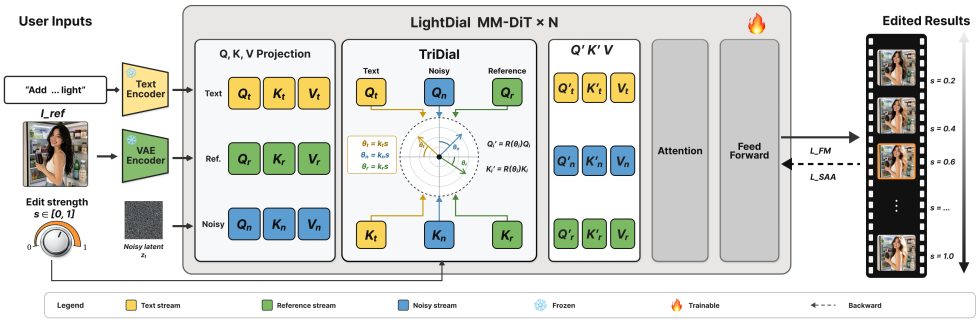

Figure 2: **LightDial framework.** TriDial conditions text, reference, and noisy-output interactions on the progression variable s. SAA supervises adjacent velocity responses during training. Flow time t evolves within each generation, while s specifies the requested lighting state.

TriDial. Text tokens describe the intended light, reference tokens encode source content, and noisy-output tokens represent the evolving generated appearance. Lighting progression changes how these streams should interact: the instruction specifies what lighting structure to form, the reference constrains what should be preserved, and the output stream realizes the requested state. TriDial provides an interaction-level inductive bias for this three-way relation. Rather than globally scaling an existing edit direction, it uses s to vary relative cross-stream compatibility while preserving each stream's query/key norm and same-stream pre-softmax compatibility. Orthogonal relative rotation is the parameterization studied here; other conditioning mechanisms remain compatible with the progression formulation. Let $\bar{q}_i, \bar{k}_i \in \mathbb{R}^d$ be the query and key after native rotary positional encoding (Su et al., 2024), with token group $c_i \in \{\text{text}, \text{output}, \text{reference}\}$. We apply

$$\tilde{q}_i = R(s\alpha_{c_i})\bar{q}_i, \qquad \tilde{k}_i = R(s\alpha_{c_i})\bar{k}_i, \tag{3}$$

where R contains two-dimensional orthogonal rotation blocks and α_c is a fixed group-specific angle. The operation leaves values unchanged and yields

$$\tilde{q}_i^\top \tilde{k}_j = \bar{q}_i^\top R\big(s(\alpha_{c_j} - \alpha_{c_i})\big)\,\bar{k}_j. \tag{4}$$

For fixed incoming features, same-group pre-softmax logits remain unchanged, while cross-group logits vary continuously with s. Softmax normalization can consequently redistribute both cross- and same-group attention weights. We set $\alpha_c = Km_c$, with $K = 0.5$ radians and $m = (1, 2, 3)$ for text, output, and reference groups. Only angle differences affect Equation 4: if all groups share a common angle, the added rotations cancel from every query–key logit. The common-angle ablation in Table 4 tests this distinction between arbitrary query/key rotation and progression-dependent relative interaction. Appendix B gives the implementation and common-offset equivalence.

SAA: local progression supervision. Adjacent targets describe the local change of the same lighting effect. SAA supervises this relationship in addition to individual target fitting. For adjacent progression states with latents z_a, z_b and values $s_a < s_b$, the two branches share source, instruction, augmentation, noise ϵ, and flow time t:

$$z_t^r = (1-t)z_r + t\epsilon, \qquad \hat{v}_r = v_\phi(z_t^r, t, x, p, s_r), \quad r \in \{a, b\}. \tag{5}$$

The local response difference is evaluated along these matched training paths:

$$\Delta\hat{v}_{a,b} = v_\phi(z_t^b, t, x, p, s_b) - v_\phi(z_t^a, t, x, p, s_a), \qquad \Delta v^*_{a,b} = z_a - z_b. \tag{6}$$

The shared noise cancels from $v_b^* - v_a^* = (\epsilon - z_b) - (\epsilon - z_a)$. We align the predicted and target differences using

$$\mathcal{L}_{\text{SAA}} = \text{SmoothL1}_\beta(\Delta\hat{v}_{a,b},\, z_a - z_b). \tag{7}$$

Both s and the noisy latent state change across the pair. For branch errors $e_r = \hat{v}_r - v_r^*$,

$$(\Delta\hat{v}_{a,b} - \Delta v^*_{a,b}) = e_b - e_a. \tag{8}$$

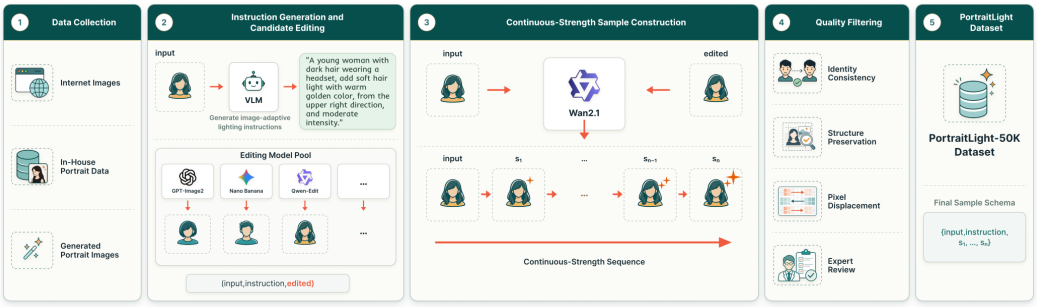

Figure 3: **PortraitLight-50K construction.** Source-specific instructions guide endpoint generation. Intermediate candidates are constructed from source–endpoint differences and endpoint-conditioned transitions; frame- and sequence-level screening then verifies lighting structure, preservation, and progression order. The diagram visualizes the endpoint-conditioned branch of this candidate pool.

SAA does not alter the pointwise flow-matching target: if both branch errors vanish, its residual also vanishes. Instead, it regularizes how prediction responses vary between neighboring progression states under matched stochastic conditions. This matched-path finite-difference supervision emphasizes local progression consistency without introducing a separate ground-truth target independent of flow matching. The complete objective is

$$\mathcal{L} = \tfrac{1}{2}(\mathcal{L}^a_{\mathrm{FM}} + \mathcal{L}^b_{\mathrm{FM}}) + \lambda_{\mathrm{SAA}}\mathcal{L}_{\mathrm{SAA}}. \tag{9}$$

Only LoRA parameters are optimized (Hu et al., 2022). The rotation parameters are fixed, and SAA is used only during training. Inference uses one adapted checkpoint for all effects and requested progression states.

3.2 Dataset Construction

The supervision must cover different spatial lighting structures and provide ordered states for each structure. PortraitLight-50K includes Hair, Tyndall, Dappled, Rembrandt, Sunset, and Slit lighting: contour illumination, volumetric rays, projected light patches, asymmetric facial shading, warm directional illumination, and narrow-band lighting. Each sequence supplies state targets for flow matching and adjacent pairs for local progression supervision:

$$\mathcal{G} = \left(x, p, e, \{(s_k, y_k)\}_{k=0}^{6}\right), \qquad s_k = k/6, \quad y_0 = x, \tag{10}$$

where e is the effect and y_6 is the accepted full-strength target. The five intermediate targets describe an accepted development of that effect from the source. The labels index within-sequence order; they are not calibrated physical intensity values or perceptually equal steps.

The pipeline in Figure 3 has four selection stages. First, it starts from licensed internet portraits, consented captures, and synthetic images. A vision-language model (VLM) evaluates each source–effect combination from visible facial and scene geometry. Hair lighting requires a separable contour, Tyndall lighting requires spatial support for directional rays, and Rembrandt lighting depends on visible face pose. For each compatible pair, the VLM constructs an instruction specifying the effect, light direction, target region, and content to preserve.

Second, five editors generate independent full-strength candidates: GPT-Image 2, Nano Banana 2, Qwen-Image-Edit, FLUX.1 Kontext, and FLUX.2-klein-9B (OpenAI, 2026; Google DeepMind, 2026; Wu et al., 2025; Black Forest Labs, 2025; 2026). Automatic screening combines MUSIQ, ArcFace, and DINO similarity with VLM assessment of lighting correctness and preservation. Expert review resolves low-margin or conflicting cases, yielding one accepted endpoint for each source–effect pair.

Third, we construct an intermediate candidate pool with two complementary source–endpoint strategies. A difference-based construction produces intermediate candidates by progressively applying the observed source–endpoint change. Wan 2.1 (Team Wan, 2025) contributes endpoint-conditioned transition candidates. Wan is therefore one candidate source rather than the definition of the retained progression. We assign progression states only after candidates have been screened and ordered.

Finally, frame-level checks assess perceptual quality, identity, content preservation, and the requested lighting structure. Sequence-level VLM and expert review then assess effect consistency, adjacent-state continuity, and progression order. The retained intermediate states are selected from this multi-source pool according to lighting-structure, preservation, and ordering criteria, rather than copied as supervision from a single transition model. PortraitLight-50K contains approximately 50,000 strength-indexed images, or roughly 7.1K seven-level sequences, with identity-disjoint partitions. Appendix A details the source composition, effect definitions, instruction construction, candidate screening, and sequence verification.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Implementation. We use FLUX.2-klein-9B (Black Forest Labs, 2026) with a frozen backbone and rank-32 LoRA adapters ($\alpha = 32$). Training uses 1024×1024 images on eight NVIDIA H20 GPUs. We employ AdamW with a learning rate of 10^{-5}, weight decay of 0.01, 500 warm-up steps, and cosine decay. Each device processes a batch of four with two gradient-accumulation steps. Adjacent target pairs share augmentation, noise, and flow time. Appendix C provides the remaining implementation details.

Compared methods. The endpoint comparison includes eight general-purpose editors. The progression benchmark covers three interface classes. Kontinuous Kontext and SliderEdit are retrained on the same PortraitLight-50K partitions as learned controllers. Text-embedding interpolation, CFG scaling, LoRA scaling, and GRAG control task-adapted editors at inference. FreeMorph and Wan 2.1 construct transitions from a supplied full-strength endpoint. LightDial instead receives the source, instruction, and requested progression state. This grouping makes the different input requirements explicit; shared data for the retrained controllers do not imply matched architectures or objectives.

Evaluation. The progression benchmark contains 200 identity-disjoint groups queried at $s \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, giving 1,000 outputs per method. The four interior values are absent from the training grid $k/6$, so they test interpolation rather than retrieval of labeled training states. MUSIQ, CLIP-IQA, MANIQA, NIQE, and an aesthetic predictor assess image quality (Ke et al., 2021; Wang et al., 2023; Yang et al., 2022; Mittal et al., 2013; Schuhmann, 2022). ArcFace measures identity similarity, and DINO-Sim measures feature-level content preservation (Deng et al., 2019; Oquab et al., 2024).

We separately characterize progression response. A guided filter estimates the illumination layer L (He et al., 2013). Let q_k be the trimmed mean absolute log-illumination difference from the source, $d_k = q_{k+1} - q_k$, and $M = 5$:

$$\text{Mono} = \frac{1}{M} \sum_{k=0}^{M-1} \mathbf{1}[\text{sign}(d_k) = \text{sign}(q_M - q_0)], \qquad \text{CV} = \frac{\text{std}(d_k)}{|\,\text{mean}(d_k)| + \delta}. \tag{11}$$

Mono describes the ordering of source-relative illumination deviation, while CV describes the regularity of successive increments. Neither metric identifies the requested lighting effect or verifies its spatial structure. We therefore treat four properties as complementary: endpoint effect-fidelity judgments assess lighting-state correctness; Mono/CV diagnose automatic response behavior; human sequence ratings assess perceived smoothness and ordering; ArcFace/DINO assess identity and content preservation.

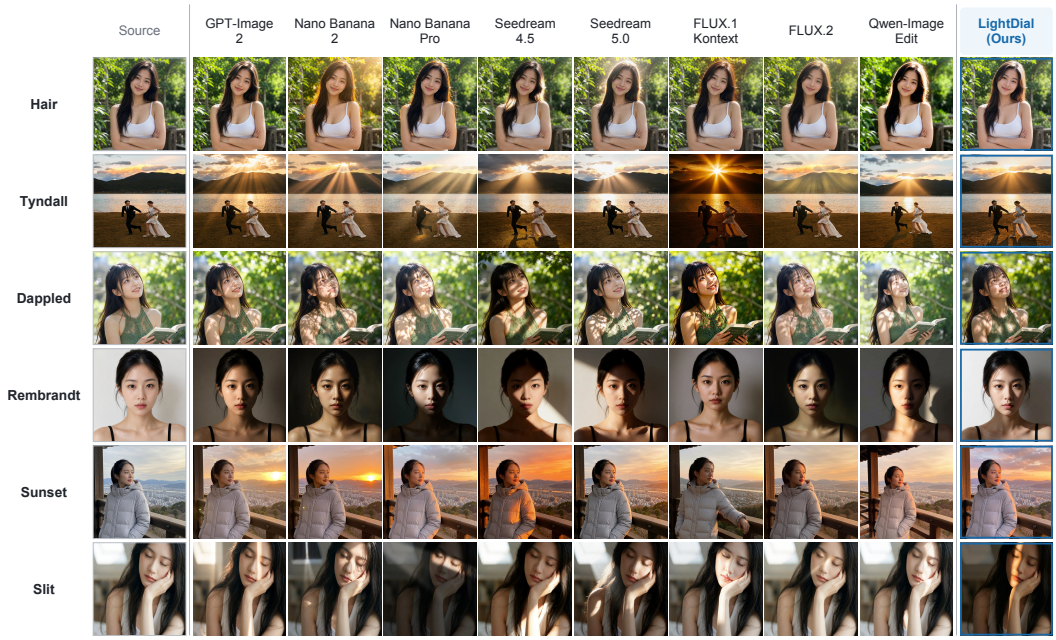

Figure 4: **Endpoint portrait relighting.** Source images and full-strength edits compare the requested lighting structure and content preservation across six effects.

4.2 Endpoint Relighting Quality

Table 1: **Full-strength portrait relighting.** Scores are averaged across six effects. LightDial is task-adapted; other editors are evaluated without task adaptation.

Method	MUSIQ↑	CLIP-IQA↑	MANIQA↑	NIQE↓	Aesthetic↑	ArcFace↑	DINO-Sim↑
FLUX.1 Kontext	74.17	0.715	0.593	4.11	6.28	0.924	0.943
FLUX.2-klein-9B	74.46	0.648	0.564	4.59	6.41	0.943	0.952
Nano Banana 2	73.16	0.682	0.543	3.79	6.39	0.931	0.923
Nano Banana Pro	68.18	0.596	0.500	4.30	6.07	0.933	0.905
GPT-Image 2	75.21	0.663	0.607	4.11	6.44	0.943	0.969
Qwen-Image-Edit	74.38	0.634	0.576	4.82	6.34	0.905	0.940
Seedream 4.5	73.61	0.712	0.579	4.49	6.29	0.915	0.909
Seedream 5.0 Lite	75.42	0.667	0.599	4.46	6.45	0.917	0.958
LightDial	75.61	0.750	0.592	3.73	6.99	0.981	0.972

Table 2: **Endpoint effect-fidelity preferences.** Leading individual selection share for each effect; shares are multi-candidate preferences, not pairwise win rates.

Effect	Leading method	Share (%)
Hair	LightDial	62.5
Tyndall	LightDial	62.5
Dappled	GPT-Image 2	50.0
Rembrandt	LightDial	75.0
Sunset	GPT-Image 2	37.5
Slit	LightDial	75.0

Endpoint quality and effect fidelity. Table 1 compares full-strength edits across the six effects. LightDial obtains the highest MUSIQ, CLIP-IQA, aesthetic, ArcFace, and DINO-Sim values and the lowest NIQE. Its ArcFace score is 0.981 and its DINO-Sim is 0.972, indicating strong identity and content preservation. MANIQA gives a different ranking: GPT-Image 2, Seedream 5.0 Lite, and FLUX.1 Kontext score above LightDial. The table establishes that LightDial retains strong endpoint quality alongside progression control. LightDial is task-adapted, whereas the listed editors are evaluated without this adaptation, so the comparison is a system-capability evaluation rather than component attribution.

Generic quality and preservation scores do not establish whether the requested lighting structure is correct. Table 2 therefore reports the endpoint effect-fidelity preferences from the anonymous multi-candidate study. LightDial receives the largest individual selection share for Hair, Tyndall, Rembrandt, and Slit, whereas GPT-Image 2 leads Dappled and Sunset. The category variation prevents the endpoint result from being summarized as uniform dominance across effects.

4.3 PROGRESSION BEHAVIOR

Comparison with continuous-control alternatives. Table 3 evaluates quality, preservation, automatic response diagnostics, and human sequence ratings. LightDial attains Mono 0.87 and CV 0.43, together with the highest human smoothness and ordering ratings, 5.00 and 4.88. Its ArcFace and DINO-Sim values, 0.951 and 0.974, jointly indicate strong identity and content preservation across the progression benchmark.

Table 3: **Progression quality and preservation.** Mono and CV are automatic source-relative illumination-response diagnostics. H-Smooth and H-Ord. are mean human ratings (1–5) for perceived smoothness and monotonic ordering in the ordered-slider study. These columns characterize progression behavior; endpoint effect fidelity is evaluated separately in Table 2.

Method	MUSIQ↑	Aesth.↑	ArcFace↑	DINO↑	Mono↑	CV↓	H-Smooth↑	H-Ord.↑
Controllers retrained on PortraitLight-50K								
Kontinuous Kontext	74.41	6.27	0.928	0.949	0.77	1.98	3.50	3.13
SliderEdit	73.71	6.00	0.939	0.923	0.63	2.39	1.00	1.38
Inference-time controls on task-adapted editors								
Text-emb. interpolation	71.93	5.86	0.871	0.900	0.61	1.96	1.50	1.63
CFG scaling	74.96	6.21	0.892	0.957	0.66	1.37	4.13	4.13
LoRA scaling	74.97	6.25	0.916	0.947	0.72	2.00	1.63	1.88
GRAG	73.77	6.15	0.906	0.941	0.77	1.05	3.63	3.13
Endpoint-conditioned transitions								
FreeMorph	70.25	5.02	0.381	0.720	0.59	3.97	1.25	1.13
Wan 2.1	75.25	6.11	0.939	0.961	0.84	0.52	3.25	3.13
LightDial	**75.59**	**6.89**	**0.951**	**0.974**	**0.87**	**0.43**	**5.00**	**4.88**

Common magnitude-based controls produce weaker ordering or less regular responses. Text interpolation, CFG scaling, LoRA scaling, and GRAG obtain Mono values between 0.61 and 0.77 and CV values between 1.05 and 2.00. The learned controllers trained on the same progression data also remain behind LightDial: Kontinuous Kontext obtains Mono 0.77, CV 1.98, and human ordering 3.13, while SliderEdit obtains 0.63, 2.39, and 1.38. These method-level comparisons support learning effect-specific progression behavior rather than treating strength only as a generic magnitude control. The component analysis below connects this result to the relative-interaction parameterization within LightDial.

The measurements also expose disagreements that a single aggregate would hide. CFG scaling receives a relatively high human ordering rating of 4.13 despite weaker automatic response diagnostics and lower preservation than LightDial. Conversely, GRAG reaches Mono 0.77 and CV 1.05 but receives a human ordering rating of 3.13. Mono/CV and human ratings therefore provide related but non-identical views of progression behavior.

Wan 2.1 is the strongest endpoint-conditioned transition baseline, with Mono 0.84 and CV 0.52. It also contributes one class of intermediate candidates during dataset construction. The retained training sequences are selected from a multi-source candidate pool through frame- and sequence-level verification. At inference, Wan requires a full-strength endpoint, whereas LightDial internalizes the verified progression supervision and directly predicts the requested state from the source, instruction, and s.

4.4 RELATIVE INTERACTION CONDITIONING AND LOCAL ALIGNMENT

We next isolate local progression supervision from progression-conditioned relative interaction.



Figure 5: **Continuous control for Hair lighting.** Increasing control progressively develops contour illumination.

Table 4: **Relative interaction conditioning and SAA.** Multipliers follow text:output:reference, with $K = 0.5$ rad. The common-angle setting analytically cancels the progression-dependent relative rotation, so response diagnostics are not applicable.

Configuration	ArcFace↑	DINO-Sim↑	Mono↑	CV↓
Angles $(1, 2, 0)$	0.939	0.965	0.81	0.69
Angles $(0, 2, 3)$	0.948	0.973	0.82	0.64
Angles $(1, 0, 3)$	0.952	0.976	0.75	0.88
Without SAA	0.952	0.975	0.81	0.71
Common angle $(1, 1, 1)$	0.954	0.977	–	–
Angles $(1, 3, 5)$	0.934	0.958	0.84	0.60
Full model $(1, 2, 3)$	0.951	0.974	0.87	0.43

Local progression supervision. Removing SAA reduces Mono from 0.87 to 0.81 and raises CV from 0.43 to 0.71, while ArcFace and DINO-Sim remain close to the full model. This matches the error-difference interpretation in Equation 8: SAA regularizes neighboring responses without changing the pointwise flow-matching target.

Relative interaction geometry. When $\alpha_{\text{text}} = \alpha_{\text{output}} = \alpha_{\text{reference}}$, the added relative rotation cancels from every query–key logit. Because this setting contains no progression-dependent TriDial path, Mono and CV are not reported for the common-angle control. The remaining settings show a response–preservation trade-off, with $(1, 2, 3)$ providing the strongest balance among the evaluated configurations.

4.5 Human Evaluation and Unseen Progression States

Human study. Fifty participants completed anonymous endpoint and sequence tasks with randomized candidate, effect, and method order. The endpoint task selects among nine candidates on four criteria; the sequence task rates ordered five-level sliders at method-normalized positions. LightDial receives 5.00, 4.88, 4.63, and 4.75 for smoothness, monotonicity, preservation, and usable range. Table 3 and Appendix E report the comparative ratings and protocol; the aggregate records support descriptive results but not item-level inference.

Unseen progression states. The four interior benchmark values are absent from the training grid $k/6$, so Table 3 and Figure 5 test interpolation. Appendix H examines extrapolation beyond $s = 1$.

Limitations. One variable couples multiple attributes; generated supervision may retain bias, automatic diagnostics omit semantics, and user-study aggregates lack item-level uncertainty. Appendix I shows extrapolation failures.

5 Conclusion

LightDial treats portrait relighting strength as an effect-specific ordered progression. TriDial conditions text–reference–output interactions, SAA regularizes neighboring responses, and PortraitLight-50K supplies verified multi-level supervision. Across endpoint, off-grid progression, preservation, and perceptual evaluations, LightDial produces ordered, smooth edits while retaining identity and content. Independent attribute control remains beyond the single progression variable.

AI USE STATEMENT

Generative AI and vision-language models were used to create synthetic portraits, assess source–effect suitability, generate image-adaptive instructions, produce endpoint and intermediate-state candidates, and support candidate filtering, pairwise progression ordering, and semantic review. An LLM-based assistant supported conceptual framing, methodology and experiment feedback, interpretation of existing results, manuscript structuring and drafting, language revision, literature search and checking, reference formatting, and internal consistency checks. The authors reviewed all AI-assisted data decisions and outputs through the screening and expert-review procedures described in the paper, verified claims, statistics, and citations against the underlying records and sources, and take full responsibility for the submission.

ETHICS STATEMENT

The dataset construction procedure includes source-license checks, consent for in-house captures, and removal of unsuitable or sensitive images. The user study was voluntary and anonymous, with consent obtained before participation. Portrait relighting can change the perceived appearance of a person; data release and use should respect the applicable permissions and disclose substantive manipulation where appropriate.

REPRODUCIBILITY STATEMENT

Section 3 specifies the conditioning mechanism and training objective, and Section 3.2 describes dataset construction. Section 4 defines the evaluation interfaces and metrics. The appendices document implementation details, optimization settings, effect definitions, candidate screening, complete automatic metrics, and the human-study protocol.

REFERENCES

Black Forest Labs. FLUX.1 Kontext: Context-aware image generation and editing. https://docs.bfl.ai/kontext/kontext_overview, 2025. Accessed 2026-07-28.

Black Forest Labs. FLUX.2-klein-9B. https://huggingface.co/black-forest-labs/FLUX.2-klein-9B, 2026. Accessed 2026-07-28.

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18392–18402, 2023.

Yukang Cao, Chenyang Si, Jinghao Wang, and Ziwei Liu. Freemorph: Tuning-free generalized image morphing with diffusion model. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 18111–18120, 2025.

Clément Chadebec, Onur Tasar, Sanjeev Sreetharan, and Benjamin Aubin. LBM: Latent bridge matching for fast image-to-image translation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/papers/Chadebec_LBM_Latent_Bridge_Matching_for_Fast_Image-to-Image_Translation_ICCV_2025_paper.pdf.

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*, pp. 369–379, 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. URL https://openaccess.thecvf.com/content/CVPR2026/html/Chen_POLAR_A_Portrait_OLAT_Dataset_and_Generative_Framework_for_Illumination-Aware_CVPR_2026_paper.html.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, 2019.

Yigit Ekin and Yossi Gandelsman. The unreasonable effectiveness of text embedding interpolation for continuous image steering. *arXiv preprint arXiv:2603.17998*, 2026. doi: 10.48550/arXiv.2603.17998.

Taichi Endo, Guoqing Hao, and Kazuhiko Sumi. FlowSlider: Training-free continuous image editing via fidelity-steering decomposition. *arXiv preprint arXiv:2604.02088*, 2026. doi: 10.48550/arXiv.2604.02088. URL https://arxiv.org/abs/2604.02088.

David Futschik, Kelvin Ritland, James Vecore, Sean Fanello, Sergio Orts-Escolano, Brian Curless, Daniel Sýkora, and Rohit Pandey. Controllable light diffusion for portraits. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8412–8421, 2023. URL https://arxiv.org/abs/2305.04745.

Rohit Gandikota, Joanna Materzyńska, Tingrui Zhou, Antonio Torralba, and David Bau. Concept sliders: LoRA adaptors for precise control in diffusion models. In *European Conference on Computer Vision*, pp. 172–188, 2024.

Google DeepMind. Nano Banana 2. https://deepmind.google/models/gemini-image/flash/, 2026. Accessed 2026-07-28.

Kaiming He, Jian Sun, and Xiaoou Tang. Guided image filtering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1397–1409, 2013. doi: 10.1109/TPAMI.2012.213.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Ronen Kamenetsky, Sara Dorfman, Daniel Garibi, Roni Paiss, Or Patashnik, and Daniel Cohen-Or. SAEdit: Token-level control for continuous image editing via sparse autoencoder. *arXiv preprint arXiv:2510.05081*, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5148–5157, 2021. doi: 10.1109/ICCV48922.2021.00510.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations*, 2023.

Linjie Lyu, Valentin Deschaintre, Yannick Hold-Geoffroy, Miloš Hašan, Jae Shin Yoon, Thomas Leimkühler, Christian Theobalt, and Iliyan Georgiev. IntrinsicEdit: Precise generative image manipulation in intrinsic space. *ACM Transactions on Graphics*, 44(4), 2025. URL https://intrinsic-edit.github.io/.

Nadav Magar, Amir Hertz, Eric Tabellion, Yael Pritch, Alex Rav-Acha, Ariel Shamir, and Yedid Hoshen. Lightlab: Controlling light sources in images with diffusion models. In *ACM SIGGRAPH 2025 Conference Papers*, pp. 106:1–106:11, 2025. doi: 10.1145/3721238.3730696.

Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013. doi: 10.1109/LSP.2012.2227726.

OpenAI. GPT-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

OpenAI. GPT Image 2. https://developers.openai.com/api/docs/models/gpt-image-2, 2026. Accessed 2026-07-28.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud 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. *Transactions on Machine Learning Research*, 2024.

Rohit Pandey, Sergio Orts-Escolano, Chloe LeGendre, Christian Häne, Sofien Bouaziz, Christoph Rhemann, Paul E. Debevec, and Sean Ryan Fanello. Total relighting: Learning to relight portraits for background replacement. *ACM Transactions on Graphics*, 40(4):43:1–43:21, 2021. doi: 10.1145/3450626.3459872.

Rishubh Parihar, Or Patashnik, Daniil Ostashev, Venkatesh Babu Radhakrishnan, Daniel Cohen-Or, and Kuan-Chieh Jackson Wang. Kontinuous kontext: Continuous strength control for instruction-based image editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 37929–37939, 2026.

Puntawat Ponglertnapakorn, Nontawat Tritrong, and Supasorn Suwajanakorn. DiFaReli: Diffusion face relighting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22646–22657, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Ponglertnapakorn_DiFaReli_Diffusion_Face_Relighting_ICCV_2023_paper.html.

Christoph Schuhmann. LAION-Aesthetics. https://laion.ai/blog/laion-aesthetics/, 2022. Accessed 2026-07-25.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8871–8879, 2024.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024. doi: 10.1016/j.neucom.2023.127063.

Team Wan. Wan: Open and advanced large-scale video generative models. *arXiv preprint arXiv:2503.20314*, 2025.

Jianyi Wang, Kelvin C. K. Chan, and Chen Change Loy. Exploring CLIP for assessing the look and feel of images. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(2):2555–2563, 2023. doi: 10.1609/aaai.v37i2.25353.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Sidi Yang, Tianhe Wu, Shuwei Shi, Shanshan Lao, Yuan Gong, Mingdeng Cao, Jiahao Wang, and Yujiu Yang. MANIQA: Multi-dimension attention network for no-reference image quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1191–1200, 2022.

Qifan Yu, Wei Chow, Zhongqi Yue, Kaihang Pan, Yang Wu, Xiaoyang Wan, Juncheng Li, Siliang Tang, Hanwang Zhang, and Yueting Zhuang. Anyedit: Mastering unified high-quality image editing for any idea. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26125–26135, 2025.

Arman Zarei, Samyadeep Basu, Mobina Pournemat, Sayan Nag, Ryan A. Rossi, and Soheil Feizi. Slideredit: Continuous image editing with fine-grained instruction control. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14430–14439, 2026.

Bohan Zhang, Huanwei Liang, Yuhan He, Hongteng Xu, Quxiao Chao, Luoqi Liu, Dixin Luo, and Ting Liu. Consistent feature transport for image relighting. *arXiv preprint arXiv:2607.17833*, 2026. doi: 10.48550/arXiv.2607.17833. URL https://arxiv.org/html/2607.17833v1.

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*, 2025a. URL https://openreview.net/forum?id=u1cQYxRI1H.

Xiao Zhang, William Gao, Seemandhar Jain, Michael Maire, David A. Forsyth, and Anand Bhattad. Latent intrinsics emerge from training to relight. In *Advances in Neural Information Processing Systems*, volume 37, pp. 96775–96796, 2024. doi: 10.52202/079017-3068. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/af79bcbd1d229a527c8f10d8d41c589d-Abstract-Conference.html.

Xuanpu Zhang, Xuesong Niu, Ruidong Chen, Dan Song, Jianhao Zeng, Penghui Du, Haoxiang Cao, Kai Wu, and An-an Liu. Group relative attention guidance for image editing. *arXiv preprint arXiv:2510.24657*, 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.

APPENDIX OVERVIEW

The appendix documents PortraitLight-50K construction, TriDial implementation, SAA training, evaluation protocols, human evaluation, additional qualitative results, progression interpolation and extrapolation, and failure cases.

A PORTRAITLIGHT-50K CONSTRUCTION DETAILS

A.1 CONSTRUCTION OVERVIEW

We construct PortraitLight-50K in four selection stages. We first assemble a diverse portrait source pool and remove images that fail provenance, safety, or basic quality requirements. GPT-4o then assesses source–effect compatibility and constructs an image-adaptive instruction for each accepted pair. Five instruction-based systems generate candidate full-strength edits, from which automatic metrics, blinded VLM review, and expert adjudication select one endpoint. We next build a multi-source pool of intermediate candidates using source–endpoint difference construction and Wan 2.1 endpoint-conditioned transitions. Frame-level screening and sequence-level VLM and expert review select the final ordered targets.

A.2 LIGHTING-EFFECT DEFINITIONS

We organize PortraitLight-50K around six semantic portrait lighting effects. The progression variable indexes the accepted ordering within a source–instruction pair. Increasing perceptual prominence is the intended behavior, not a calibrated label scale. We define Hair lighting through localized contour illumination along the hair and upper silhouette. For Tyndall lighting, we require directional shafts formed by volumetric scattering. We represent Dappled lighting with spatially varying patches of light and shadow. We define Rembrandt lighting through asymmetric facial illumination and an illuminated cheek region on the shadowed side. We characterize Sunset lighting as warm, low-angle directional illumination. For Slit lighting, we require a narrow directional band across the portrait or its surroundings. We use these operational definitions during instruction generation and data review.

A.3 SOURCE COLLECTION AND COMPOSITION

We assemble an initial pool of approximately 62,385 portrait candidates from three channels. We collect 24,211 images from the internet, accounting for 38.8% of the pool. We retain only content whose recorded license permits research use and store the source URL, retrieval date, license statement, and content hash. We remove images with unverifiable provenance, visible watermarks, sensitive personal information, unsafe content, severe compression, or ambiguous portrait ownership.

We add 7,707 consented in-house images, accounting for 12.4% of the pool. We link each capture session to written research consent while separating participant identity from the training identifier. We generate the remaining 30,467 portraits, which account for 48.8%. For these images, we retain the generator, prompt, generation parameters, available seed, and creation date. Before VLM screening, we remove near duplicates within and across all three channels.

We retain approximately 50,000 strength-indexed images in the final dataset, equivalent to roughly 7.1K complete seven-level sequences. In Figure 6, we separately report the source pool and final effect allocation. We assign the largest share to Hair lighting because it is compatible with a broad range of portraits. We allocate smaller shares to Tyndall and Slit lighting because they require more specific scene geometry.

A.4 VLM SUITABILITY SCREENING

We employ GPT-4o as the VLM (OpenAI, 2024). In the first pass, we determine whether each image is a usable portrait. We require a valid source to contain a clearly resolvable primary subject, sufficient visible facial or upper-body structure, and enough image quality for identity-preserving editing. We reject severe occlusion, extreme blur, dominant text or watermarks, conflicting pre-

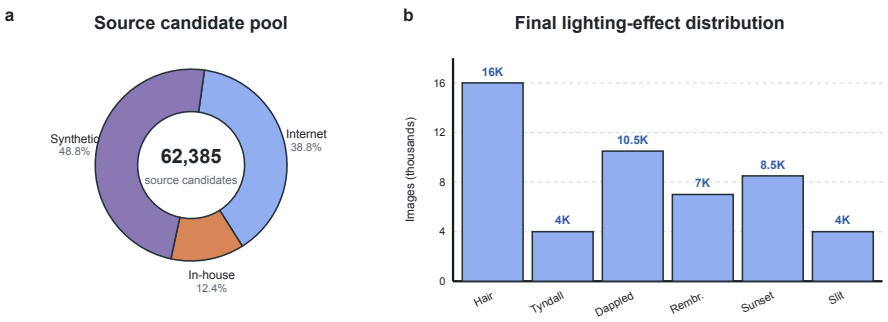

Figure 6: **Data composition of PortraitLight-50K. (a)** Distribution of the initial source candidate pool across internet collection, consented in-house capture, and synthetic portraits. **(b)** Approximate distribution of roughly 50,000 final strength-indexed images across the six lighting effects. The two panels describe the pre-filter source pool and final dataset, respectively.

existing effects, and compositions in which a requested lighting pattern would have no plausible spatial support.

For each accepted portrait, we evaluate every lighting category independently rather than assigning one global label. We record the source decision, rejection codes, compatible effects, suitability, light direction, and image-adaptive instruction. We also record target regions, protected regions, and relevant scene cues. We record suitability as `reject`, `marginal`, or `suitable`. This fixed schema separates portrait acceptance from effect compatibility and spatial planning.

We instruct the VLM to use only visible image evidence, assess all six effects independently, and abstain when the required spatial pattern is ambiguous. We also prohibit changes to identity, expression, clothing, pose, camera viewpoint, and unrelated scene content. We send only source-effect pairs labeled `suitable` to candidate generation.

We determine effect compatibility from the spatial support visible in each portrait. For Hair lighting, we require a separable hair or upper-body contour. For Tyndall lighting, we require sufficient scene depth for directional rays. We assign Dappled lighting when visible surfaces can receive coherent projected patterns. We use face pose to assess support for asymmetric Rembrandt lighting. We assign Sunset lighting to compositions that support warm, low-angle light. For Slit lighting, we require a plane that can receive a narrow light band. We then specify the light direction, affected regions, effect geometry, protected content, and source-specific failures to avoid.

A.5 Image-Adaptive Instruction Generation

We assemble each instruction from the structured VLM analysis rather than from a fixed effect name. We let e denote the effect, d the inferred direction, r^+ the regions that should receive the effect, r^- the protected regions, g the relevant scene geometry, and c the source-specific preservation constraints. We compose

$$p = \mathrm{Compose}(e, d, r^+, r^-, g, c). \tag{12}$$

We use the following natural-language template:

> Apply [lighting effect] to this portrait. Introduce the dominant light from [direction] so that [effect-specific structure] appears over [target regions] and remains consistent with [scene cues]. Preserve [protected facial and scene content]. Avoid [source-specific failure modes].

We populate the bracketed fields independently for every source. We require the preservation constraints to cover identity, facial geometry, expression, skin texture, hair shape, pose, clothing, viewpoint, and unrelated background content. We prohibit global recoloring unless the Sunset effect requires it. This construction makes each instruction specific enough to describe the desired spatial pattern while preventing the editor from treating the task as unconstrained portrait restyling.

A.6 Full-Strength Candidate Generation

We submit each accepted source-effect assignment to GPT-Image 2, Nano Banana 2, Qwen-Image-Edit, FLUX.1 Kontext, and FLUX.2-klein-9B (OpenAI, 2026; Google DeepMind, 2026; Wu et al., 2025; Black Forest Labs, 2025; 2026). We combine hosted and weight-available instruction editors to increase candidate diversity. We access GPT-Image 2 and Nano Banana 2 as hosted systems and use available model weights for the remaining editors. We do not treat any editor as ground truth. We store each candidate with the source identifier, effect, exact instruction, editor identifier, model version, output resolution, generation parameters, seed or request identifier when available, and generation date.

We define the five-candidate set as

$$\mathcal{C}(x, p, e) = \{G_j(x, p)\}_{j=1}^{5}, \tag{13}$$

where G_j denotes one candidate editor. We neither average nor blend the candidates. In the next stage, we retain one full-strength edit for each source-effect assignment.

A.7 Candidate Filtering and Selection

We perform filtering from inexpensive deterministic checks to semantic review. We first remove failed API responses, invalid files, incorrect aspect ratios, empty or duplicated outputs, safety refusals, and images in which the primary face cannot be detected. We evaluate the remaining candidates with MUSIQ (Ke et al., 2021), ArcFace (Deng et al., 2019), and DINO-Sim (Oquab et al., 2024). We use these metrics to measure perceptual quality, face-embedding identity similarity, and feature-level content preservation, respectively.

For compact notation, we let $Q(y)$, $I(x, y)$, and $C(x, y)$ denote the MUSIQ, ArcFace, and DINO-Sim measurements. For candidate y, the automatic acceptance gate we apply is

$$\begin{aligned} A(y) = &\mathbf{1}[Q(y) \geq \tau_Q]\ \mathbf{1}[I(x, y) \geq \tau_I] \\ &\mathbf{1}[C(x, y) \geq \tau_C]\ \mathbf{1}[V(x, p, e, y) = \text{pass}], \end{aligned} \tag{14}$$

where we imply the product of the four indicators. We use the VLM decision V to check that the requested effect is recognizable, its spatial pattern is compatible with the portrait, the result remains natural, and unrelated content is unchanged. We hide editor names during this semantic review to avoid generator-specific preference.

We discard candidates that fail any hard gate. When multiple candidates remain, we use the VLM for pairwise comparison based on effect correctness, naturalness, identity preservation, and instruction consistency. We send low-margin or conflicting comparisons to expert review. This staged procedure prevents a high no-reference quality score from compensating for an incorrect lighting effect, while ArcFace and DINO-Sim aim to reduce content-changing results. These filters do not guarantee correctness and can bias evaluation with the same metrics.

A.8 Intermediate Candidate Construction

For each accepted source–endpoint pair, we construct a pool of candidate intermediate states through two complementary procedures. The difference-based branch progressively applies the observed source–endpoint change to form candidates between the boundary images. The endpoint-conditioned branch provides the same source and selected full-strength edit to Wan 2.1 (Team Wan, 2025) and samples transition frames. These branches provide candidate states rather than fixed training labels: no single generator or construction rule determines the retained progression.

After frame- and sequence-level screening, we select five intermediate states to form

$$\mathcal{G} = \left(x, p, e, \{(s_k, y_k)\}_{k=0}^{6}\right), \qquad s_k = \frac{k}{6}, \tag{15}$$

where $y_0 = x$ and y_6 is the accepted full-strength result. We assign the normalized labels after verifying the accepted order. They index progression within a source–instruction sequence and are neither physical radiometric measurements nor a claim of perceptually equal spacing.

A.9 Sequence-Level Filtering and Expert Review

We evaluate candidate states at both frame and sequence levels. Frame-level screening covers perceptual quality, identity, content preservation, and the requested lighting structure. We then assemble and assess complete seven-image sequences for effect consistency, adjacent identity or background drift, and agreement with the ordering $(s_0, \ldots, s_6)$. Pairwise VLM comparisons support ordering decisions. The benchmark illumination readout remains a diagnostic and is not used as the sole definition of strength.

Expert reviewers inspect each source, instruction, proposed sequence, and automatic rejection flag without seeing candidate provenance. They assess effect correctness, lighting naturalness, subject and scene preservation, full-strength adequacy, and the order and continuity of intermediate states. Only sequences that pass the core criteria are retained; disagreements are adjudicated and rejected cases receive failure codes. The final targets are therefore selected from a multi-source pool rather than copied directly from one transition model.

B TriDial Implementation Details

B.1 Integration into Joint Attention

We use the three token groups in each FLUX.2-klein-9B joint attention sequence: text, noisy output, and reference-image tokens. We integrate TriDial without adding an attention layer or altering the backbone's value projections. Instead, we wrap the shared attention-dispatch function. We intercept each full-sequence joint attention call after native rotary positional encoding (RoPE) (Su et al., 2024). We then call the original attention implementation with the rotated queries and keys.

We denote the post-RoPE query, key, and value tensors as $\bar{\mathbf{Q}}, \bar{\mathbf{K}}, \mathbf{V} \in \mathbb{R}^{B\times S\times H\times D}$. Here, B is the batch size, S the sequence length, H the number of heads, and D the head dimension. We construct one angle for every token position and broadcast it across the batch and attention heads. We use the same token layout and strength for all batch elements within one forward call. For paired SAA branches, we invoke the model separately at the corresponding strengths.

B.2 Token-Group Identification

We recover token groups from native text and image identifiers rather than fixed sequence offsets. We assign every row of `txt_ids` to the text group. For an image token with identifier $\mathbf{u}_i$, the first identifier coordinate separates noisy output and reference-image tokens:

$$c_i = \begin{cases} \text{output}, & u_{i,0} \le \tau_{\text{id}}, \\ \text{reference}, & u_{i,0} > \tau_{\text{id}}, \end{cases} \qquad \tau_{\text{id}} = 0.5. \tag{16}$$

We name these image groups `noisy` and `cond`, corresponding to the paper's output and reference groups. We record the resulting length-S group-index vector with a forward pre-hook. This token-wise assignment remains valid when the image groups are not stored as contiguous segments. When token metadata are unavailable, we return the original attention computation. We do the same when the attention sequence length differs from the recorded joint sequence. This guard prevents us from applying a full-sequence angle vector to an incompatible key-value subsequence.

B.3 Strength-Dependent Rotation

For token group c, we define the rotation angle as

$$\begin{aligned} \theta_c(s) &= K\, s\, m_c, \qquad K = 0.5 \text{ rad}, \\ m_{\text{text}} : m_{\text{output}} : m_{\text{reference}} &= 1 : 2 : 3. \end{aligned} \tag{17}$$

Here, we use s for the requested strength of the current forward pass. This notation corresponds to $\alpha_c = K m_c$ in the main paper. We use $s \in [0, 1]$ in the main experiments. We do not clamp this value, which permits the interpolation and extrapolation analyses described later. However, we do not assume stable behavior outside the training range. We fix both K and the group multipliers and do not optimize or store them as learned weights.

For the complete model, we set the three absolute angles to $0.5s$, $1.0s$, and $1.5s$ radians. Only their differences affect query-key interactions, so we equivalently represent them as $(-0.5s, 0, 0.5s)$. The adjacent group separation is therefore $0.5s$ radians, and the maximum separation between text and reference-image tokens is $1.0s$ radians.

For an even head dimension D, we define

$$\mathbf{R}_2(\theta) = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}, \qquad \mathbf{R}(\theta) = \mathrm{diag}(\mathbf{R}_2(\theta), \ldots, \mathbf{R}_2(\theta)) \,. \tag{18}$$

We use the same angle for every adjacent channel pair $(2r, 2r+1)$ within a token and attention head. We apply

$$\widetilde{\mathbf{q}}_i = \mathbf{R}(\theta_{c_i}(s))\, \bar{\mathbf{q}}_i, \qquad \widetilde{\mathbf{k}}_i = \mathbf{R}(\theta_{c_i}(s))\, \bar{\mathbf{k}}_i, \tag{19}$$

while leaving $\mathbf{V}$ unchanged. We construct rotation angles in 32-bit floating-point precision and cast their sine and cosine values to the query and key data type. We pass tensor shapes, the attention mask, and all remaining arguments to the original attention backend unchanged.

B.4 Effect on Attention Geometry

For tokens i and j, we use orthogonality to obtain

$$\widetilde{\mathbf{q}}_i^\top \widetilde{\mathbf{k}}_j = \bar{\mathbf{q}}_i^\top \mathbf{R}\big(Ks(m_{c_j} - m_{c_i})\big)\, \bar{\mathbf{k}}_j. \tag{20}$$

From Equation 20, we obtain three implementation properties. First, we modulate attention through relative group angles. A common offset therefore leaves every query-key logit unchanged. Second, we preserve same-group pre-softmax logits because their relative rotation is zero. Cross-group logits vary continuously with s. Third, we preserve query and key norms through orthogonal rotation, avoiding direct feature scaling.

We do not claim that same-group attention weights remain fixed. Cross-group logit changes alter the softmax denominator and can redistribute all normalized attention weights. At $s = 0$, we set every additional angle to zero and recover the same adapted checkpoint's attention computation without the added rotation, not the original pretrained model or the source image. We use TriDial to condition cross-group interactions on the progression variable. This is suited to lighting progression because increasing strength changes how instruction semantics, preserved source structure, and generated appearance should interact; it need not scale every stream uniformly. Ordered image-space changes remain training goals to be evaluated, not guarantees of rotation or SAA.

B.5 Forward Procedure and Computational Cost

For each forward pass, we first store the requested progression variable. We then convert the native token identifiers into a length-S group-index vector. We construct the corresponding angle vector from Equation 17 and broadcast it over B and H. We rotate queries and keys pairwise along the head dimension. Finally, we pass $(\widetilde{\mathbf{Q}}, \widetilde{\mathbf{K}}, \mathbf{V})$ to the original scaled dot-product attention function.

We add $O(BSHD)$ arithmetic and do not introduce another $S \times S$ attention matrix. We allocate rotated query and key tensors with the same shapes and data types as their inputs. We preserve the attention backend, mask, and output interface. When we disable the controller, we bypass the wrapper and evaluate the same checkpoint without TriDial.

B.6 Exact Ablation Configurations

For activated-group ablations, we retain each active group's full-model multiplier and assign zero angle to the excluded group. We report the resulting triplets in Table 5. These settings change several relative angles simultaneously. They do not isolate the necessity of applying a nonzero rotation to each group; the full setting has an equivalent representation with one zero multiplier.

With the common-angle configuration $(1, 1, 1)$, we apply the same additional rotation to every token. All relative angles cancel, so we recover the unmodified attention computation in exact arithmetic. We center the full setting $(1, 2, 3)$ as $(-1, 0, 1)$ and center $(1, 3, 5)$ as $(-2, 0, 2)$. We use these

Table 5: **Exact TriDial configurations used in the ablations.** A zero entry means no direct additional rotation for that group; its relative interactions can still change. All settings use $K = 0.5$ rad.

Configuration	m_{text}	m_{output}	$m_{\text{reference}}$
Text + output	1	2	0
Output + reference	0	2	3
Text + reference	1	0	3
Common angle	1	1	1
Expanded separation	1	3	5
Full model	1	2	3

settings to test the scale of relative-angle separation, not different allocation patterns. We use $1:2:3$ to obtain nonzero cross-group rotations with equal adjacent separation. We do not claim that this choice is universally optimal.

B.7 Common-Angle Equivalence

For a common-angle setting, $\theta_{\text{text}} = \theta_{\text{output}} = \theta_{\text{reference}} = \gamma$, every relative angle is zero. Consequently, $R(\gamma)^\top R(\gamma) = I$, and the added operation leaves all query–key logits unchanged for fixed incoming features. This exact equivalence motivates the common-angle control in Table 8: it retains a common orthogonal rotation but removes progression-dependent relative interaction. Accordingly, the main component table leaves its Mono and CV entries undefined instead of treating stochastic variation as a progression response.

C Strength-Aware Alignment and Training Details

C.1 Motivation

Standard flow matching applies a sample-wise objective at each progression state, with shared model parameters across samples. Although it supervises the desired output at each level, it does not explicitly prioritize changes between neighboring values. We distinguish these changes from global brightness adjustments. For example, we expect stronger Rembrandt lighting to rebalance illuminated and shadowed facial regions. By contrast, we expect stronger Hair lighting to develop a localized contour highlight.

We therefore use the observed difference between adjacent targets as the SAA supervisory signal. We require no illumination decomposition, region mask, effect classifier, or manually defined photometric direction. Instead, we derive the meaning of each strength change from its multi-level sequence.

C.2 Matched Adjacent-Strength Pair

We sample a seven-level data group and an adjacent pair indexed by $k \in \{0, \dots, 5\}$:

$$s_a = \frac{k}{6}, \qquad s_b = \frac{k+1}{6}. \tag{21}$$

We denote the two target images by I_a and I_b, and define $z_a = \mathcal{E}(I_a)$ and $z_b = \mathcal{E}(I_b)$ as their latent representations. We use the same source portrait x, lighting instruction p, stochastic augmentation, Gaussian noise ϵ, and flow-matching timestep t in both branches. We vary only the target latent and the strength supplied to TriDial.

We match stochastic conditions to reduce unrelated variation; the input state still changes with the target. We define the two flow-path states as

$$z_t^a = (1-t)z_a + t\epsilon, \qquad z_t^b = (1-t)z_b + t\epsilon, \tag{22}$$

with target velocities

$$v_a^* = \epsilon - z_a, \qquad v_b^* = \epsilon - z_b. \tag{23}$$

We use the shared network to predict

$$\begin{aligned} \widehat{v}_a &= v_\phi(z_t^a, t, x, p, s_a), \\ \widehat{v}_b &= v_\phi(z_t^b, t, x, p, s_b). \end{aligned} \tag{24}$$

By sharing noise and timesteps, we avoid mixing the desired strength response with unrelated stochastic variation.

C.3 Local Progression Supervision

Under this matched construction, we cancel the random term from the target velocity difference:

$$\Delta v^* = v_b^* - v_a^* = (\epsilon - z_b) - (\epsilon - z_a) = z_a - z_b. \tag{25}$$

We therefore use Δv^* to represent the data-defined local change between adjacent lighting targets under the same source and instruction. We align this target with the predicted difference $\Delta\widehat{v} = \widehat{v}_b - \widehat{v}_a$:

$$\mathcal{L}_{\mathrm{SAA}} = \mathrm{SmoothL1}(\Delta\widehat{v}, \Delta v^*)\,. \tag{26}$$

For residual r, we define the elementwise SmoothL1 penalty with transition parameter $\beta > 0$ as

$$\rho_\beta(r) = \begin{cases} \dfrac{r^2}{2\beta}, & |r| < \beta, \\ |r| - \dfrac{\beta}{2}, & \text{otherwise.} \end{cases} \tag{27}$$

We use the quadratic region to retain sensitivity to small response mismatches. We use the linear region to limit the influence of occasional imperfect adjacent targets.

We also retain an individual flow-matching objective for each branch:

$$\mathcal{L}_{\mathrm{FM}}^r = \|\widehat{v}_r - v_r^*\|_2^2\,, \qquad r \in \{a, b\}. \tag{28}$$

We define the complete training objective as

$$\mathcal{L} = \frac{1}{2}\left(\mathcal{L}_{\mathrm{FM}}^a + \mathcal{L}_{\mathrm{FM}}^b\right) + \lambda_{\mathrm{SAA}}\mathcal{L}_{\mathrm{SAA}}. \tag{29}$$

We use the two flow-matching terms to preserve accuracy at the sampled strengths. For branch errors $e_r = \widehat{v}_r - v_r^*$, the SAA residual is

$$(\Delta\widehat{v} - \Delta v^*) = e_b - e_a. \tag{30}$$

SAA therefore does not introduce a different pointwise flow-matching optimum: if both errors are zero, the SAA residual is also zero. Under finite LoRA optimization, it changes the inductive bias by directly penalizing variation of prediction errors between neighboring states along matched stochastic paths. The mechanistic ablation separates adjacency from path matching.

C.4 Optimization Procedure and Scope

We perform each SAA update in five operations. We first sample one multi-level group and select two adjacent targets. We apply the same spatial augmentation to the source-target triplet before latent encoding. We then reuse one flow-matching timestep and one Gaussian noise tensor in both branches. We evaluate both branches at their respective strengths and combine their losses in one update. We share model weights across branches and update only LoRA.

We use SAA only during training and add no inference module or parameter. We encourage the velocity-response change to follow the observed adjacent target difference. We do not define a physical lighting scale or claim a mathematical guarantee of monotonic image-space variation. We assess continuity and monotonicity empirically with the strength-control benchmark.

C.5 Backbone and Trainable Parameters

We build LightDial on FLUX.2-klein-9B (Black Forest Labs, 2026). We freeze the pretrained backbone and optimize only LoRA parameters (Hu et al., 2022). We keep TriDial parameter-free by fixing its angle scale and token-group multipliers during training. We also implement SAA without a trainable module because it operates on the two predicted velocity fields. Consequently, we use one adapted model for all six effects and all supported strengths.

Table 6: **LightDial training and inference configuration.** The table reports the shared settings; each ablation changes only its named factor.

Item	Setting
Backbone	FLUX.2-klein-9B
Trainable adaptation	LoRA; backbone frozen
LoRA configuration	Rank 32; $\alpha = 32$
Training resolution	1024×1024
Training hardware	Eight NVIDIA H20 GPUs
Optimizer	AdamW; weight decay 0.01
Learning-rate schedule	1×10^{-5}; 500-step warm-up; cosine decay
Batching	4 per device; 2 gradient accumulation steps
Training strength grid	$s_k = k/6,\ k \in \{0, \ldots, 6\}$
Pair construction	Adjacent levels from the same sequence
Shared pair conditions	Source, instruction, augmentation, noise, and t
Training objective	Paired flow matching $+\lambda_{\mathrm{SAA}}$ SAA
TriDial angle scale	$K = 0.5$ rad
Token-group multipliers	$1 : 2 : 3$ (text:output:reference)
Inference range in main evaluation	$s \in [0, 1]$
SAA inference overhead	None

C.6 Data Sampling and Optimization

We sample adjacent pairs from the seven-level PortraitLight-50K sequences and balance the six lighting-effect categories. Within each pair, we share the source, instruction, augmentation, noise, and flow-matching timestep. We pass the corresponding scalar strength to each branch through TriDial. We weight both branches equally in the paired flow-matching term. Across ablations, we retain the same data partition, resolution, and optimization protocol.

We train at 1024×1024 resolution using rank-32 LoRA adapters with $\alpha = 32$. We employ AdamW with an initial learning rate of 1×10^{-5} and weight decay of 0.01. After 500 warm-up steps, we reduce the learning rate with cosine decay. We use a per-device batch size of 4 and accumulate gradients for 2 steps before each optimizer update.

We summarize the training settings in Table 6.

C.7 Inference

At inference, we receive a source portrait, a lighting instruction, and a progression variable $s \in [0, 1]$. We convert s into three group angles through TriDial and do not activate the training-only SAA objective. We use one LoRA-adapted checkpoint for all six effects without checkpoint interpolation, parameter scaling, or strength-specific adapters. In the main evaluation, we use four values absent from the training grid and $s = 1$ to test interpolation.

Matched-path response differences. SAA compares $v_\phi(z_t^b, t, x, p, s_b)$ with $v_\phi(z_t^a, t, x, p, s_a)$. Both the progression variable and the noisy latent state vary across the pair; noise, time, instruction, source, and augmentation are matched. The difference is a local response along these paired training paths, not a partial derivative with respect to s at fixed z_t. No division by the progression interval is introduced. The zero-strength target is the source portrait.

D Evaluation Protocol

D.1 Baseline Conditions

Kontinuous Kontext and SliderEdit are retrained using the same identity-disjoint partitions and resolution as LightDial. Text-embedding interpolation, CFG scaling, LoRA scaling, and GRAG use editors adapted to the portrait-lighting data. FreeMorph and Wan 2.1 additionally receive a full-strength endpoint. Direct progression control and endpoint-conditioned transitions therefore have

different input requirements. Method-specific control ranges are mapped to the displayed progression and are not assumed to represent a shared physical scale.

The progression benchmark contains 200 identity-disjoint groups. Each method produces five outputs at $s \in \{0.2, 0.4, 0.6, 0.8, 1.0\}$, giving 1,000 edited images. Training uses the grid $s_k = k/6$, so the four interior evaluation positions are unseen. Images are standardized to 1024×1024; image-level scores are averaged within each group and then across groups. Sequence diagnostics include the source as the zero-state reference and do not measure a separately generated $s = 0$ reconstruction.

D.2 Image Quality and Preservation

MUSIQ, CLIP-IQA, and MANIQA evaluate image quality (Ke et al., 2021; Wang et al., 2023; Yang et al., 2022). NIQE provides a naturalness measure (Mittal et al., 2013), and the aesthetic predictor scores visual appeal (Schuhmann, 2022). ArcFace measures face-embedding identity similarity, while DINO-Sim measures feature-level content similarity (Deng et al., 2019; Oquab et al., 2024). These measures are interpreted together: preservation alone can favor a weak edit, and an appealing image can fail to express the requested lighting structure.

D.3 Illumination-Response Diagnostics

The response readout converts each image to linear-RGB luminance and estimates its illumination layer using a guided filter with radius $0.06 \min(H, W)$ and regularization 10^{-2}. We compute

$$r_k = \log(L(y_k) + 10^{-4}) - \log(L(x) + 10^{-4}), \qquad q_k = \mathrm{TrimMean}_{1\%} |r_k|. \tag{31}$$

The trimming removes one percent from each tail. With $d_k = q_{k+1} - q_k$ and $M = 5$, the sequence measures are

$$\mathrm{Mono} = \frac{1}{M} \sum_{k=0}^{M-1} \mathbf{1}[\mathrm{sign}(d_k) = \mathrm{sign}(q_M - q_0)], \qquad \mathrm{CV} = \frac{\mathrm{std}(d_k)}{|\,\mathrm{mean}(d_k)| + \delta}. \tag{32}$$

Here $\delta > 0$ is distinct from the generation noise ϵ. Mono summarizes whether source-relative illumination deviation follows a consistent order, and CV summarizes the regularity of successive increments. The readout does not identify the effect type, its direction, or its spatial location.

These limitations admit degenerate cases. For example, a nearly unchanged sequence can obtain favorable response diagnostics if its small readout changes are sufficiently ordered and regular. We therefore do not use Mono or CV as semantic effect-fidelity metrics. Endpoint human judgments assess whether the requested structure is realized; ordered-slider ratings assess perceived smoothness and monotonicity; ArcFace and DINO-Sim assess identity and content preservation. The four evidence types answer different questions.

D.4 Complete Quantitative Results

The main paper presents the endpoint results and compact progression and component tables. Tables 7 and 8 provide the remaining automatic metrics while preserving the same method groupings and configurations.

Common-angle configuration. For a common group angle, $\alpha_{\mathrm{text}} = \alpha_{\mathrm{output}} = \alpha_{\mathrm{reference}}$, the added relative rotation in every attention logit is zero. The common-angle row is therefore an internal control for progression-dependent relative interaction. Since s has no effect through TriDial in this configuration, Mono and CV are not interpreted as progression diagnostics and are omitted from Table 4.

Table 7: **Additional progression-benchmark quality metrics.** The main paper reports MUSIQ, aesthetics, preservation, response diagnostics, and human ratings; this table provides only the complementary quality measures.

Method	CLIP-IQA↑	MANIQA↑	NIQE↓
Controllers retrained on PortraitLight-50K			
Kontinuous Kontext	0.602	0.529	3.75
SliderEdit	0.685	0.588	4.30
Inference-time controls on task-adapted editors			
Text-embedding interpolation	0.708	0.557	4.20
CFG scaling	0.672	0.576	3.76
LoRA scaling	0.674	0.578	3.69
GRAG	0.684	0.568	3.86
Endpoint-conditioned transitions			
FreeMorph	0.542	0.442	5.19
Wan 2.1	0.722	0.570	3.59
LightDial	**0.760**	**0.595**	3.63

Table 8: **Additional quality metrics for the component analysis.** Table 4 reports preservation and progression diagnostics; this table provides only the complementary quality measures.

Configuration	MUSIQ↑	CLIP-IQA↑	MANIQA↑	NIQE↓	Aesthetic↑
Angles $(1, 2, 0)$	75.18	0.751	0.591	3.70	6.78
Angles $(0, 2, 3)$	75.05	0.739	0.584	3.74	6.70
Angles $(1, 0, 3)$	74.82	0.731	0.577	3.83	6.61
Without SAA	75.64	0.754	**0.596**	**3.62**	6.84
Common angle $(1, 1, 1)$	75.21	0.724	0.587	3.67	6.69
Angles $(1, 3, 5)$	74.93	**0.767**	0.581	3.88	6.73
Full model $(1, 2, 3)$	75.59	0.760	0.595	3.63	**6.89**

E Human Evaluation

E.1 Evaluation Questions

The study separates two perceptual questions. The endpoint task evaluates a single full-strength result: effect fidelity asks whether the requested spatial lighting structure is realized, visual appeal concerns aesthetic preference, content preservation concerns identity and unrelated scene content, and overall quality asks for the preferred result considering the complete edit. The sequence task evaluates progression behavior: smoothness concerns gradual changes between displayed states, monotonicity concerns consistent increase of the effect, preservation concerns stability of the subject and scene across the sequence, and usable range concerns whether the slider spans distinct and useful states. Endpoint correctness and sequence behavior are reported separately.

E.2 Interface and Protocol

We conduct an anonymous web study with 50 completed responses. Participation is voluntary, and we collect no names, email addresses, or direct identifiers. After consent, we record only self-reported editing experience and viewing-device category.

In Part I, participants evaluate full-strength editing for all six lighting effects. Each trial displays the source portrait and nine anonymous candidates from eight general-purpose editors and LightDial. Participants select one candidate independently for effect fidelity, visual appeal, content preservation, and overall quality; the same candidate may be selected for multiple criteria. The source remains visible for direct comparison. Figure 7 shows the interface.

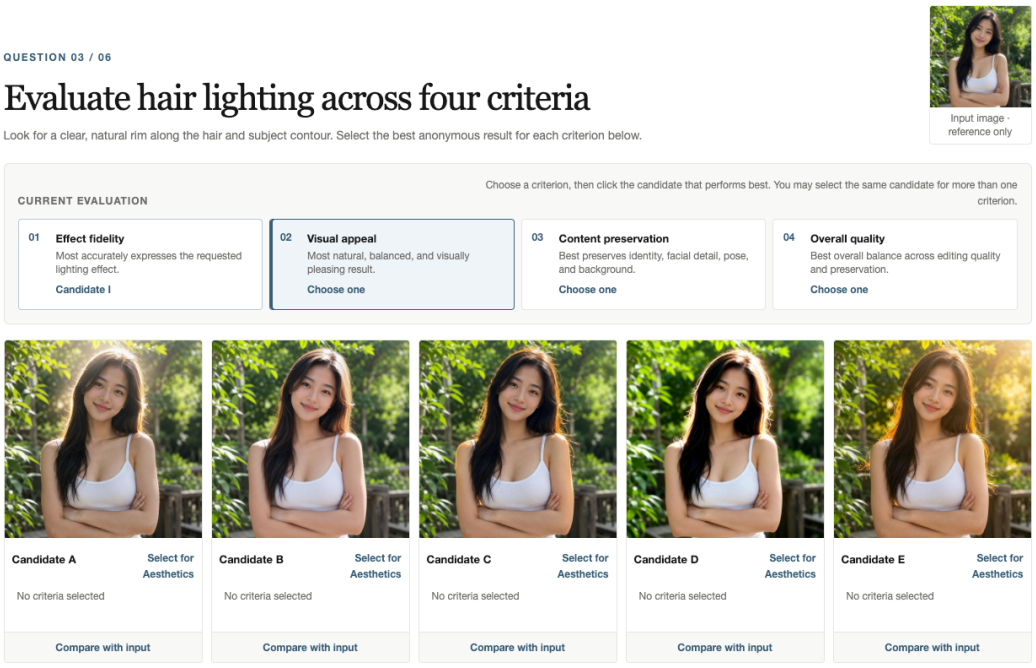

Figure 7: **Interface for endpoint evaluation.** The source remains visible while participants inspect anonymous full-strength candidates and select one result independently for each of four criteria.

In Part II, participants inspect one anonymous five-level sequence at a time through a slider with normalized positions from 0 to 1 in increments of 0.25. These positions index each method's own displayed control range and do not represent equal physical lighting increments across methods. Participants may toggle the source image and then rate smoothness, monotonicity, preservation, and usable range on a five-point scale. Figure 8 shows the interface.

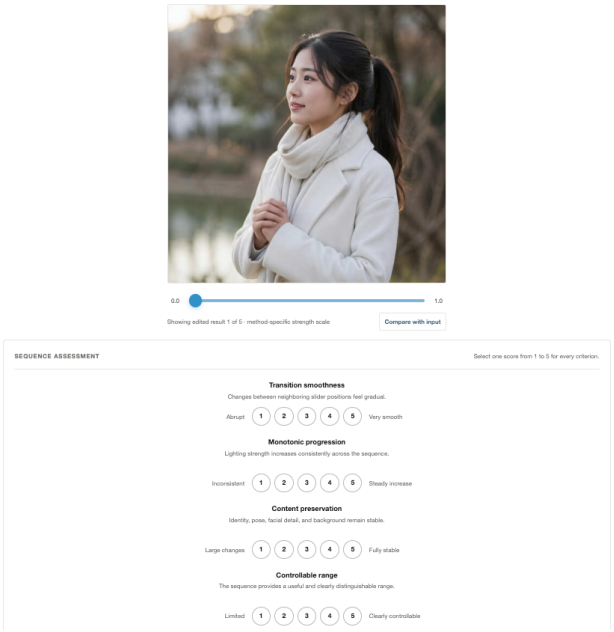

Figure 8: **Interface for progression evaluation.** Participants use the slider to inspect five ordered outputs from one anonymous method and assign four sequence-level ratings from 1 to 5.

E.3 Blinding, Randomization, and Aggregation

The browser receives anonymized method codes. For each response, it independently randomizes the six endpoint-effect trials, the candidate order within each trial, and the nine progression-sequence trials. All endpoint candidates for one effect use the same source and instruction. The sequence states retain their weak-to-strong order because smoothness and monotonicity are the rated properties, while method order remains randomized.

The endpoint study compares LightDial with GPT-Image 2, Nano Banana 2, Nano Banana Pro, Seedream 4.5, Seedream 5.0 Lite, FLUX.1 Kontext, FLUX.2-klein-9B, and Qwen-Image-Edit. The sequence study compares LightDial with Kontinuous Kontext, SliderEdit, text-embedding interpolation, FreeMorph, Wan 2.1, CFG scaling, LoRA scaling, and GRAG. Method identities are recovered only during aggregation.

For Part I, selections are counted independently for each effect–criterion pair. The available aggregate chart records the leading individual method and its selection share; it does not recover complete per-method distributions. For Part II, ratings are averaged separately for each method and criterion. Because item-level responses are not available in the aggregate records, we report descriptive shares and means without confidence intervals or statistical significance tests.

E.4 Results

For endpoint effect fidelity, LightDial receives the largest individual share for Hair, Tyndall, Rembrandt, and Slit lighting, whereas GPT-Image 2 leads Dappled and Sunset. Figure 9 also shows visual appeal, content preservation, and overall quality. Across these panels, the preferred individual editor remains effect dependent; the "Others combined" segment is an aggregate of all non-leading methods and is not one competitor.

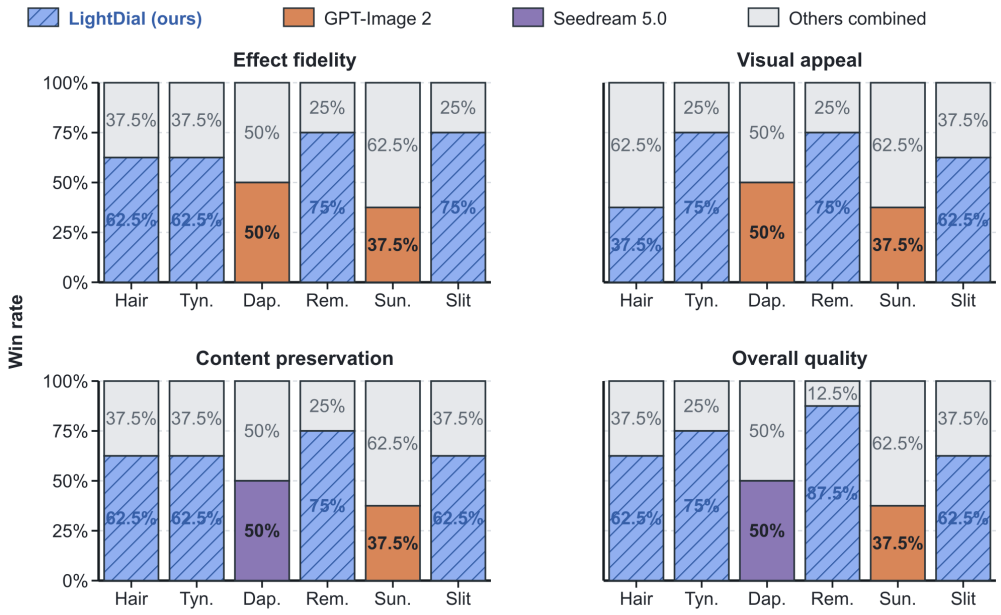

Figure 9: **Endpoint relighting preferences.** Each panel reports the selection share of the leading individual method for one criterion and effect; all remaining selections are grouped as "Others combined." The displayed shares are multi-candidate preferences, not pairwise win rates.

LightDial obtains mean ratings of 5.00 for smoothness, 4.88 for monotonicity, 4.63 for preservation, and 4.75 for usable range. These values complement the automatic Mono/CV diagnostics: the automatic measures summarize source-relative illumination response, whereas participants judge the perceived behavior of complete image sequences. Neither sequence evaluation substitutes for the endpoint effect-fidelity question above.

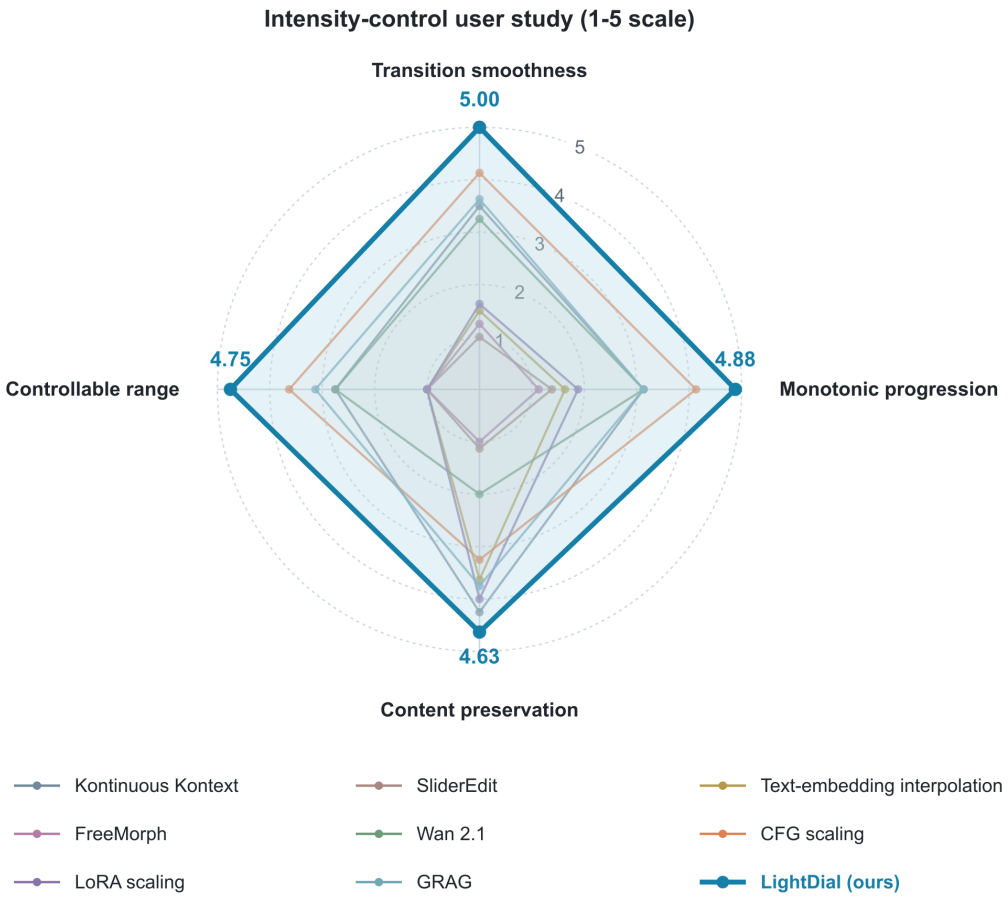

Figure 10: **Progression user study.** Mean five-point ratings compare perceived smoothness, monotonicity, preservation, and usable control range for the nine methods.

Table 9: **Complementary human progression ratings (1–5).** Table 3 reports smoothness and monotonicity; this table reports the remaining preservation and usable-range criteria.

Method	Preservation↑	Usable range↑
Kontinuous Kontext	4.25	2.75
SliderEdit	1.13	1.00
Text-embedding interpolation	3.63	1.00
FreeMorph	1.00	1.00
Wan 2.1	2.00	2.75
CFG scaling	3.25	3.63
LoRA scaling	4.00	1.00
GRAG	3.75	3.13
LightDial	**4.63**	**4.75**

The human evaluation is independent of the intermediate candidate-generation process: participants judge anonymous model outputs and are not asked to match Wan transitions, difference-based candidates, or dataset state labels. This provides perceptual evidence about the resulting editor while leaving the candidate-source limitations of the training data explicit.

F Additional Results

F.1 More Fine-Grained Strength-Control Results

We provide three additional cases for every lighting effect in Figures 11–16. We use one source portrait per row and show five increasing strengths after the source. The examples illustrate the range of lighting structures produced for different subjects and scenes.



Figure 11: **Additional LightDial results for Dappled lighting.** Each row is a different portrait. From left to right, localized patches of light and shadow become increasingly pronounced as strength changes from $s = 0.20$ to $s = 1.00$.



Figure 12: **Additional LightDial results for Hair lighting.** Selected queries at specified progression values show contour illumination around the hair and silhouette across three portraits.



Figure 13: **Additional LightDial results for Rembrandt lighting.** LightDial progressively strengthens asymmetric facial illumination while maintaining the subject's appearance across the displayed strength range.



Figure 14: **Additional LightDial results for Slit lighting.** Narrow directional light bands become more prominent with increasing strength while the portrait and surrounding scene remain consistent.



Figure 15: **Additional LightDial results for Sunset lighting.** Selected warm-lighting outputs are shown at five increasing values of s across three scenes.



Figure 16: **Additional LightDial results for Tyndall lighting.** Each row presents a different portrait. Directional volumetric light shafts become progressively more prominent as strength changes from $s = 0.20$ to $s = 1.00$, while the subject and surrounding scene remain visually stable.

G Qualitative Component Analysis

G.1 Relative-Angle Configurations and SAA

We compare relative-angle configurations and the SAA objective. The SAA comparison retains TriDial and removes the differential alignment term.

Figure 17 compares three two-group configurations with the full model. The multiplier settings change relative-angle signs and spans. These selected examples cannot establish that all three groups require nonzero rotation. We use output for the noisy output group and reference for the source-image group.

The SAA examples visualize the response with and without neighboring-level supervision (Figure 18).

G.2 Rotation-Angle Scale

Figure 19 varies four maximum rotation-angle scales while retaining the full-model multipliers. Small scales respond weakly, whereas larger scales introduce the effect earlier. The main model uses $K = 0.5$ rad; larger values introduce more pronounced appearance changes in these examples.



Figure 17: **Relative-angle configuration comparison.** Three two-group configurations are compared with the full model.



Figure 18: **SAA ablation.** TriDial w/o SAA is compared with the full model in two matched examples.



Figure 19: **Qualitative sensitivity to the maximum rotation-angle scale.** Rows vary K while preserving the full-model multiplier pattern. Each sequence contains the source and five representative strength levels.

H Generalization Evaluations

H.1 Cross-Backbone Evaluation

We apply the same relative query-key rotation principle to FLUX.1 Kontext and Qwen-Image-Edit, with token-group identification and rotation placement adapted to each architecture. Figure 20 shows ordered contour-light increases under the same Hair-lighting instruction. We treat this as qualitative evidence rather than a controlled architecture comparison.



Figure 20: **Cross-backbone qualitative results.** The same source and Hair-lighting instruction are evaluated with LightDial instantiated on FLUX.1 Kontext, FLUX.2-klein-9B, and Qwen-Image-Edit. Each row contains six images spanning increasing strength.

H.2 Progression Interpolation and Extrapolation

We evaluate s from 0.2 to 2.0 in steps of 0.2, using unseen values inside $[0, 1]$ for interpolation and $s > 1$ for extrapolation. In Figure 21, three sequences show ordered Hair-lighting changes at off-grid settings and continued strengthening beyond $s = 1$. Larger values can introduce appearance deviations, so these examples are exploratory and do not establish a reliable extended control range.



Figure 21: **Progression interpolation and extrapolation.** The displayed values $s = 0.40$ and $s = 0.80$ are absent from the seven-level training grid, $s = 1.00$ is the trained upper boundary, and $s = 1.40$ and $s = 2.00$ are extrapolated settings. Each row contains six images from one sequence.

I Limitations and Failure Cases

I.1 Supervision and Comparison Scope

We select full-strength targets from multiple general-purpose editors, some of which also serve as endpoint baselines. Intermediate candidates come from both source–endpoint difference construction and Wan 2.1 endpoint-conditioned transitions, followed by automatic, VLM, and expert verification. This multi-source procedure reduces dependence on one transition model but does not remove biases inherited from generated endpoints or intermediate candidates. The endpoint comparison also contrasts specialized training with general-purpose editing rather than matched backbones.

I.2 Evaluation and Control Scope

The evaluation combines automatic measures, qualitative comparisons, and a 50-response user study. The aggregate user-study records support descriptive means and selection shares but not item-level uncertainty. Mono and CV measure source-relative response order and regularity rather than semantic lighting correctness. Effect-wise endpoint reporting exposes category variation, but low-frequency failures may remain underrepresented. One scalar controls the coupled progression of direction, softness, color temperature, contrast, and spatial extent rather than disentangling these attributes. Its value is sequence relative, not a globally calibrated physical intensity.

I.3 Failure Cases

Figure 22 uses high-strength extrapolation to illustrate the learned control boundary. At $s = 2.00$, lighting may continue to increase while facial appearance and bright-region fidelity deteriorate.

Evaluation scope. The examples at $s = 2$ illustrate extrapolation failures and do not estimate in-range failure rates. The automatic control diagnostics summarize illumination deviation, while



Figure 22: **Failure cases under excessive strength extrapolation.** At $s = 2.00$, the upper example shows facial drift and the lower shows highlight saturation and a global lighting shift; both are weaker at $s = 1.00$.

endpoint judgments assess requested effect structure and the ordered-slider study assesses perceived progression behavior. Consequently, the results should not be interpreted as a calibrated physical lighting scale. The relative-angle and SAA analyses characterize the studied backbone and training setup rather than establishing universal superiority across architectures.