Learning Personalized Photographic Style from Pairwise User Preferences

Supplementary Material

A. Web Application for Data Collection

We developed a web-based application to collect pairwise preference data at scale. This section describes the interface design and data collection procedures. The main survey procedure is described in Section 3.3 of the main paper.

A.1. Overall Procedure

The data collection process consists of five stages:

1. **Study Guide.** Participants review an overview explaining the study's purpose, estimated completion time (10-20 minutes), and data privacy policy. The interface clarifies that responses are used for academic research and that aggregated results may be publicly released without personal identifiers.
2. **User Registration.** Participants provide demographic information and viewing environment details. An auto-generated user ID is assigned to each participant, which can be modified if needed. A seed number is also provided to enable reproducible comparison ordering for participants who wish to repeat the survey.
3. **Pre-Survey.** Before the main task, participants complete a display characterization survey to measure their viewing environment and color sensitivity. This stage captures display-specific properties and individual perceptual characteristics.
4. **Tutorial.** Participants review step-by-step instructions for the comparison task. The tutorial uses a sequential disclosure format where each instruction becomes visible only after acknowledging the previous one.
5. **Main Survey.** Participants complete 90 pairwise comparisons, selecting their preferred image based on color and tone. The survey enforces a minimum 2-second viewing time per pair to encourage careful evaluation. Details of this stage are provided in Section 3.3.

On average, participants completed the entire procedure in 11.6 minutes. This duration was designed to minimize visual fatigue and maintain attention throughout the survey.

A.2. User Registration

The registration form collects information about participants' background. Participants specify their eyewear type (none, glasses, glasses with blue-light filter, or contact lenses), current vision quality (good, average, or poor), and color vision status (normal, red-green deficiency (protan), red-green deficiency (deutan), blue-yellow deficiency (tritan), achromatopsia, or unsure). They also indicate their display device type (desktop monitor, laptop, tablet, or mobile) and optionally provide the display model details.

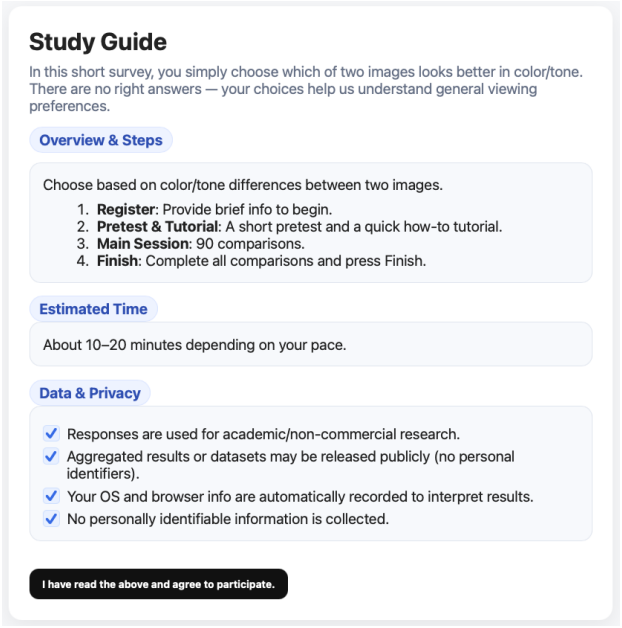


Figure A. Capture image of study guide page (stage 1).

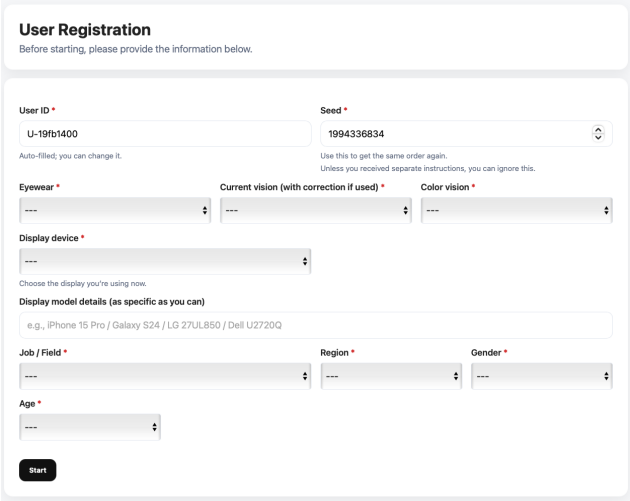


Figure B. Capture image of user registration page (stage 2).

Additionally, participants report their job or field, region, gender, and age group. The job field options include student, engineer, designer, researcher, artist, photographer, videographer, image editor, colorist, color scientist, display engineer, UI/UX designer, marketing professional, gamer, and others. In practice, most participants selected the "others" category, indicating diverse backgrounds.

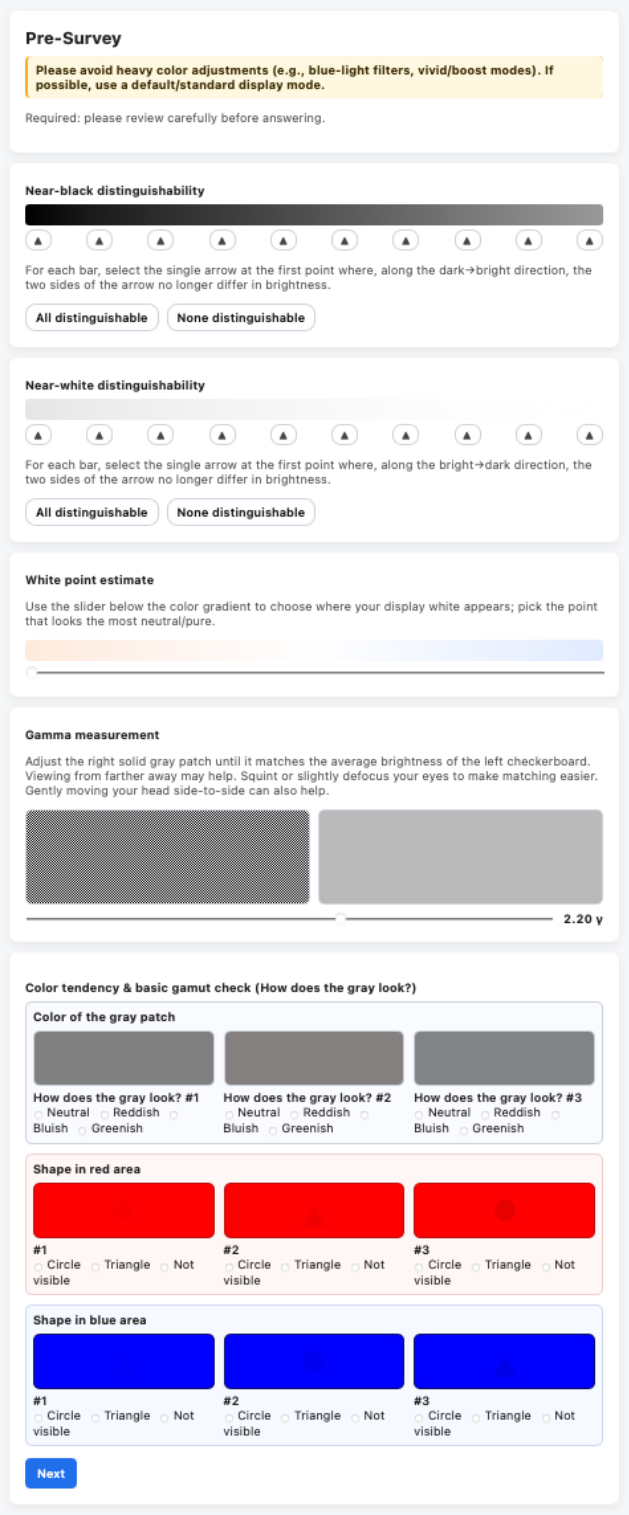


Figure C. Capture image of pre-survey page (stage 3).

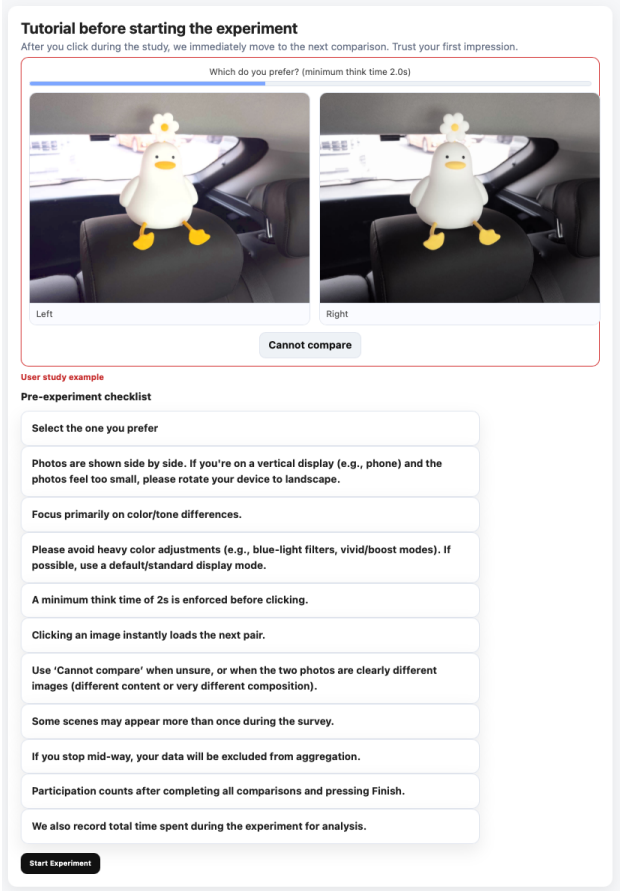


Figure D. Capture image of tutorial page (stage 4).

A.3. Pre-Survey

The pre-survey characterizes each participant's display and color perception. This stage consists of four measurements:

- **Near-Black Distinguishability.** Participants view a series of gradually darkening bars and identify the point where brightness differences become indistinguishable. This measures the display's black level and the participant's sensitivity to dark tones.
- **Near-White Distinguishability.** Using a similar approach with brightening bars, this test measures the display's white level handling and the participant's sensitivity to highlight details.
- **White Point Estimate.** Participants select a point on a color gradient that appears most neutral or pure white on their display. This captures the display's white point characteristics and any color tinting present in the viewing environment.
- **Gamma Measurement.** Participants adjust a solid gray patch to match the average brightness of a checkerboard pattern. This estimates the effective gamma curve of their display system, which affects how brightness levels are distributed across the tonal range.

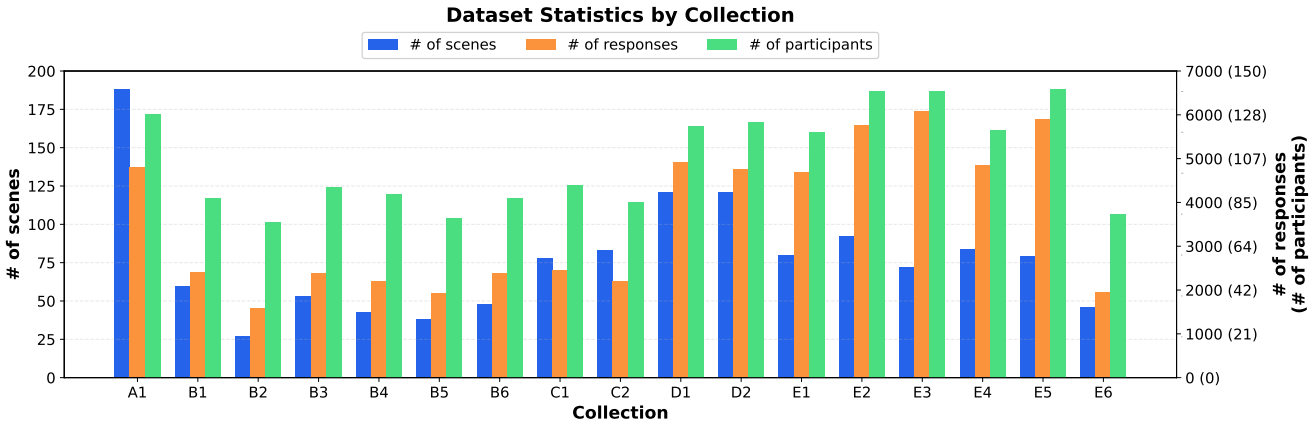


Figure E. **Dataset statistics by collection.** Distribution of scenes (blue bars), responses (orange bars), and participants (green bars) across all 17 subcollections in PPSD. The right y-axis shows the number of responses and participants. Collection A1 contains the most scenes (188), while content-specific collections (E2-E5) attracted the highest participation with over 6,000 responses each. The dataset achieves broad coverage across all collection types, with each receiving substantial user engagement.

- **Color Tendency and Basic Gamut Check.** Participants evaluate three gray patches and indicate whether each appears neutral or exhibits a color cast (reddish, bluish, or greenish). They also identify shapes within solid red and blue areas. These tests assess color tinting in the display and the participant's sensitivity to subtle color variations.

While these measurements were not directly used in the baseline experiments presented, they are included in the released dataset. Analysis of how personal display settings and color sensitivity correlate with aesthetic preferences represents a promising direction for future work.

A.4. Tutorial

The tutorial stage ensures participants understand the comparison task and viewing guidelines. Instructions are presented sequentially, with each item revealed only after the participant acknowledges understanding of the previous one. This design prevents participants from skipping important information and ensures all guidelines are reviewed before beginning the main survey.

B. Dataset Statistics

B.1. Details on Image Collections

Our dataset comprises five main categories (Types A through E) as described in Section 3.2. Here we provide additional details on the subcollections within each type and analyze their distribution in the dataset.

Collection Structure. Each main type is further divided into subcollections based on specific sources or processing methods. Type A consists of a single collection (A1) from Adobe-MIT 5K [6]. Type B is divided into six subcollections (B1-B6) representing different device pairs or dataset sources: B1 from HKUST-ISP dataset [13], B2 from the Raw-to-Raw [9] dataset, and B3-B6 from our own captures using various smartphone and camera combinations (Galaxy S25 Ultra vs. iPhone 15 Pro, Galaxy S24+ vs. Galaxy S20, Galaxy Note10 vs. iPhone 16 Pro, Canon EOS R50 vs. iPhone 16+). Type C includes two subcollections from different generative models: C1 using FLUX-krea [5] and C2 using Qwen-Images [12]. Type D is split into two subcollections: D1 with simple color manipulations (four style variations per scene) and D2 with LUT-based retouching (four LUTs applied per scene). Type E is divided into six content-specific subcollections (E1-E6), where E1 uses PPR10K images and E2-E6 focus on specific subjects (dogs, nature, architecture, food, and duck dolls, respectively). For E2 through E6, each base image generated by FLUX-krea was processed with four consistent retouching approaches: saturation reduction, brightness increase, and two different LUTs.

Dataset Statistics by Collection. Figure E presents the distribution of scenes, responses, and participants across all subcollections. During data collection, each participant was randomly assigned 1 to 6 collections from the 17 available subcollections (A1-E6). Collection A1 contains the largest number of scenes (188) and received over 6,000 responses from 128 participants. Among device-based collections (B1-B6), the distribution varies based on source availability, with B1 from HKUST-ISP containing 60 scenes and our self-captured collections (B2-B6) ranging from 27 to 53 scenes. The generative model collections (C1-C2) show balanced coverage with 78 and 83 scenes respectively. The controlled synthetic collections (D1-D2) each contain 121 scenes. Content-specific collections (E2-E6) received substantial engagement, with E3 (nature) gathering over 6,000 responses from 138 participants. The variation in response counts across collections reflects both the number of

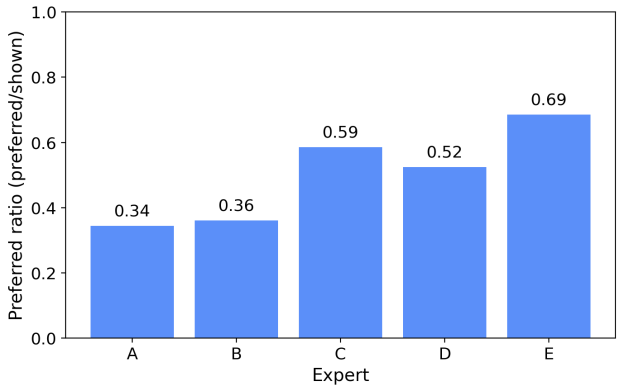


Figure F. Preferred ratio for each Adobe-MIT 5K expert style.

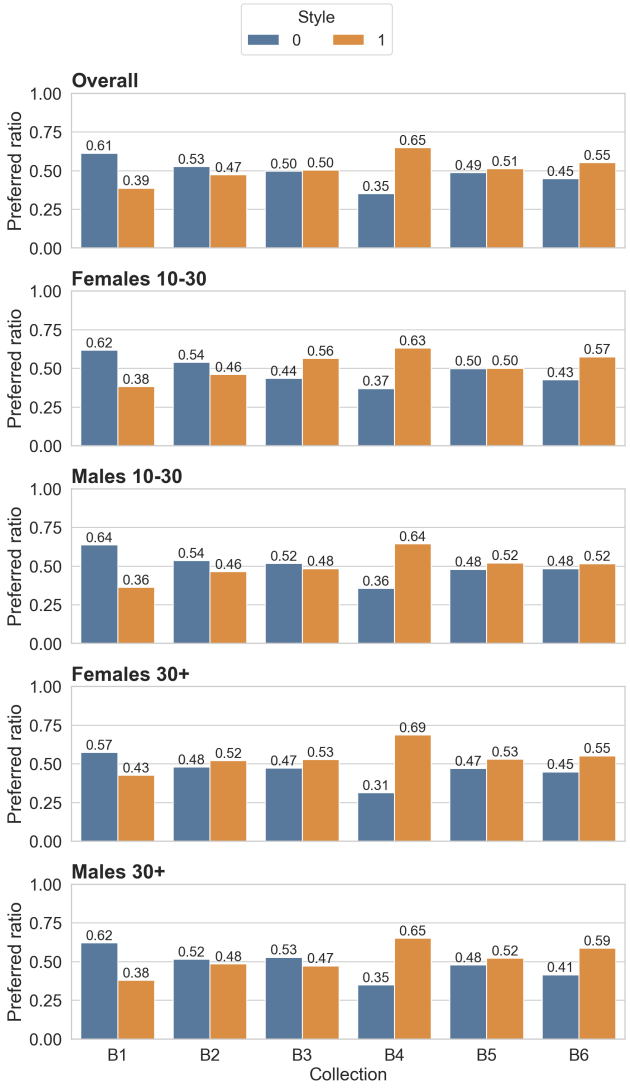
scenes and the random assignment of collections to participants. Overall, the dataset achieved broad coverage across all collection types, ensuring diverse representation of photographic style variations.

B.2. Preference Pattern Analysis

We analyze several interesting patterns that emerge from user responses. This analysis provides insights into how different retouching sources and styles are perceived across our user base.

Which Adobe-MIT 5K expert is most preferred? The Adobe-MIT 5K dataset provides five expert retouching results (A, B, C, D, E) for each image. We examine which expert's aesthetic receives the highest preference rate when compared against other experts or alternative retouching methods. Figure F shows the preferred ratio for each expert style in our dataset, where the ratio is calculated as the number of times each style was preferred divided by the number of times it was shown to users. Expert E (style 4) shows the highest preference rate at 0.69, followed by Expert C (style 2) at 0.59, while Experts A and B (styles 0 and 1) show lower rates of 0.34 and 0.36. This could suggest that certain expert aesthetics tend to resonate more with our user base. However, we note that these observations are based on a subset of Adobe-MIT 5K images, and preference rates may vary depending on which style pairs are presented together.

Galaxy vs. iPhone: Do young women really prefer iPhone captures? Different smartphone manufacturers employ distinct image processing pipelines, which may lead to varying user preferences. We examine whether demographic factors such as age and gender influence preferences for images captured with Samsung Galaxy and Apple iPhone devices. Figure G shows preferred ratios across six device pair collections (B1-B6), where style 0 and style 1 represent different device outputs in each collection. The overall results show varied preferences depending on the



	B1	B2	B3	B4	B5	B6
0	iPhone (?)	Galaxy S9	Galaxy S25 Ultra	Galaxy S24+	Galaxy Note 10	Canon EOS R50
1	Mi (?)	iPhone X	iPhone 15 Pro	Galaxy S20	iPhone 16 Pro	iPhone 16+

Figure G. Preferred ratios for individual devices by age and gender.

specific device pair, with no universal dominance of either manufacturer. However, demographic analysis reveals interesting patterns. In collection B2 (Galaxy S9 vs. iPhone X), females show a notably higher preference for iPhone captures (0.54) compared to males (0.54 vs. 0.46), and this preference is strongest among younger females (10-30: 0.54 vs. 30+: 0.48). Within the female group, younger users consistently show higher iPhone preference rates across multiple collections. These patterns suggest that demographic factors may influence device preference, though the extent varies by specific device models.

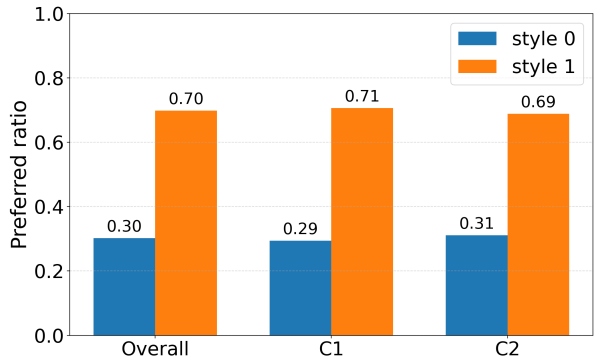


Figure H. Preferred ratio for generated image before and after generative editing.

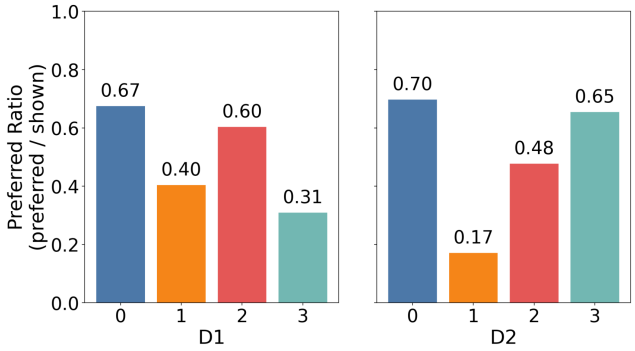


Figure I. Preferred ratios for original images and synthetic retouching styles.

Generated images: Initial output vs. refined edits For images created by generative models, we compare user preferences between the initial generated output and subsequently edited versions. This reveals whether users favor the original generative aesthetic or prefer additional refinements. Figure H shows preferred ratios for generated images before (style 0) and after (style 1) editing. Across both collections (C1: Flux-krea, C2: Qwen-Image), the edited versions consistently receive higher preference rates, with overall rates of 0.70 and 0.69 for edited images compared to 0.30 and 0.31 for initial outputs. In our pipeline, edited versions are created by extracting prompts from initial outputs using LLaVA and regenerating with Flux-kontext [4]. This suggests that users tend to prefer refined outputs over initial generations, regardless of the base generation model. Future work could explore editing capabilities of recent model-consistency-preserving approaches such as Nano Banana [7, 8] and other specialized editing models [3, 10, 11].

Are there universally preferred synthetic retouching styles? We analyze whether certain synthetic retouching approaches (e.g., specific filter combinations, adjustment patterns) consistently receive higher preference rates across diverse users and image types, or if preferences remain highly individualized. Figure I shows preferred ratios for two collections (D1, D2) with identical scene content, where style 0 represents the original image from public repositories (Pixabay [1], Unsplash [2]) or generation models. In collection D1, we apply synthetic edits: reduced saturation (style 1), increased brightness (style 2), and a flash effect with center brightness enhancement and noise (style 3). Collection D2 applies different public LUTs (4K-ALEXA-LOG-C-LTL.cube, ot_1.A032_07211259_C029.cube, SATICI_PİRABOK_1.C0026.cube) to styles 1, 2, and 3 respectively. Both collections show high preference for original images (0.67 in D1, 0.70 in D2), suggesting that images from public sharing platforms maintain good quality. In D1, brightness enhancement (style 2) achieves a comparable preference rate (0.60) to the original, while saturation reduction (style 1: 0.40) and flash effects (style 3: 0.31) receive lower preferences. This indicates that certain global adjustments may align better with general preferences. For D2, the original image receives the highest preference, suggesting that scene-agnostic LUT applications provide limited aesthetic improvement. However, style 3 shows a notably higher preference rate (0.65) compared to styles 1 (0.17) and 2 (0.48), indicating that some LUTs may be more universally appealing than others.

Do semantic categories influence retouching preferences? We examine whether image content (e.g.,animals landscapes, architecture, food) correlates with specific retouching preference patterns. Figure J shows preferred ratios across four semantic categories: animals (E2), landscapes (E3), architecture (E4), and food (E5). All images are generated using Flux-krea (style 0) and then edited with saturation reduction (style 1), brightness enhancement (style 2), LUT1 (4K-ALEXA-LOG-C-LTL.cube, style 3), and LUT2 (ot_1.A032_07211259_C029.cube, style 4). Across all categories, we observe general tendencies: original images and brightness-enhanced versions receive higher preferences, LUT2 is generally preferred over LUT1, and saturation reduction consistently receives lower preference rates. Interestingly, for food images (E5), the preference against saturation reduction intensifies, while brightness enhancement receives substantially higher preference. This may reflect the importance of color saturation in conveying food freshness and appeal. These patterns suggest that while certain retouching tendencies exist across categories, semantic content can influence the degree of preference for specific adjustments.

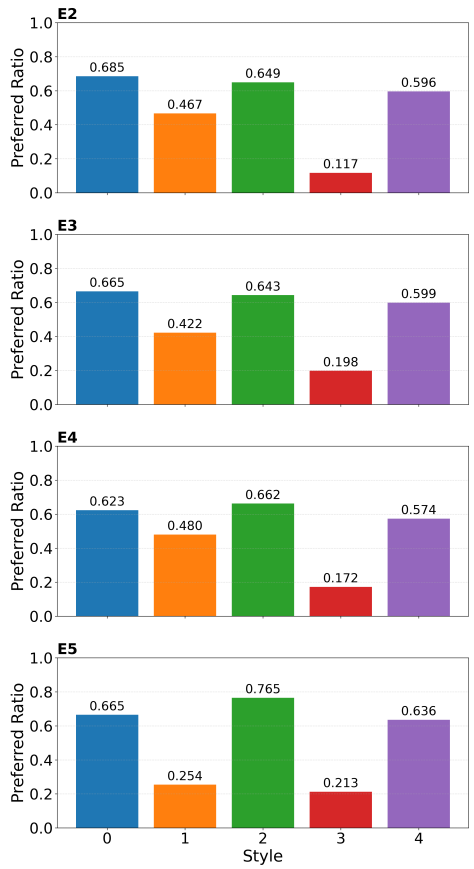


Figure J. Preferred ratios for styles by semantic category.

Table A. Impact of training weights on model performance.

ΔE_{00}	w=(0.5, 1.5)	(0.5, 1.0)	(1.0, 2.0)
vs P.	4.88	4.95	5.47
vs N.	5.71	5.52	6.55
BFS	0.189	0.191	0.167
CMR	0.078	0.054	0.090
CQS	**0.204**	0.202	0.182
PSNR	w=(0.5, 1.5)	(0.5, 1.0)	(1.0, 2.0)
vs P.	36.7	36.5	32.7
vs N.	33.5	34.5	29.6
BFS	35.05	35.49	31.11
CMR	0.046	0.028	0.050
CQS	**36.66**	36.49	32.66

Table B. Quantitative results on individual image subsets.

Δ_{00}	Type A	B	C	D	E
vs P.	4.54	6.78	5.26	3.92	3.77
vs N.	4.95	7.25	5.64	4.72	4.35
BFS↑	0.211	0.143	0.184	0.242	0.247
CMR↑	0.043	0.034	0.035	0.052	0.071
CQS↑	0.220	0.147	0.190	0.254	0.264
PSNR	Type A	B	C	D	E
vs P.	36.65	34.63	37.15	37.38	37.72
vs N.	34.91	33.07	36.19	33.91	34.25
BFS↑	35.77	33.84	36.67	35.60	35.94
CMR↑	0.024	0.023	0.013	0.049	0.048
CQS↑	36.64	34.62	37.15	37.33	37.67

C. Additional Experiments

This section presents additional empirical results for the proposed baseline methods. We note that our work focuses on establishing a foundation for personalized photographic style learning, demonstrating proof-of-concept for different technical approaches that can be further developed.

C.1. Ablation Study on Training Weight

Table A shows the impact of training weights w assigned to preferred and non-preferred samples. Comparing $(0.5, 1.0)$ and $(0.5, 1.5)$, both configurations produce outputs closer to preferred images, as indicated by the vs P. metrics. The $(0.5, 1.0)$ setting achieves better BFS scores (0.189 vs 0.191 for Δ_{01}), suggesting stronger preference alignment. However, $(0.5, 1.5)$ shows better CMR performance (0.078 vs 0.054) and achieves higher CQS (0.204 vs 0.202). Since CQS balances both aspects of quality, we adopt $(0.5, 1.5)$ as the default. When increasing the weight on preferred samples to $(1.0, 2.0)$, CMR improves further (0.050 for Δ_{00}), but BFS degrades notably (0.167). This results in lower CQS (0.182), indicating that over-emphasizing preferred samples can harm overall performance. These results suggest that balanced weighting between preferred and non-preferred samples is important for learning personalized photographic style.

C.2. Results on Individual Subsets

We evaluated our model across the five subsets (Types A-E). Table B shows distinct performance characteristics that offer important insights. We observed that the model finds Type B (Cross-device) the most challenging. This suggests that learning preferences from pairs with significant sensor and ISP discrepancies remains a difficult task. On the other hand, the model achieved high performance on Types D and E, indicating that preferences derived from synthetic color transformations are comparatively easier to learn.

C.3. Human Evaluation

We conducted a study with 21 participants (Figure K). Across all models, participants consistently preferred the retouched results over the non-preferred originals, providing direct evidence that training on PPSD enables models to produce user-preferred photographic styles. Notably, the retouched images were also often comparable to the preferred originals, even when retouching starts from a non-preferred image. Finally, the human preference trend aligns with our

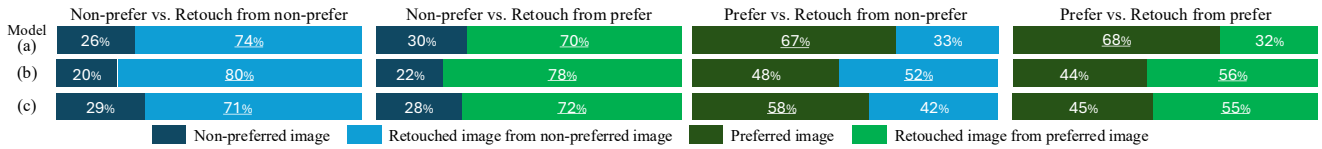


Figure K. Human evaluation study results. We conducted a pairwise preference study with 21 participants to evaluate Models (a-c).

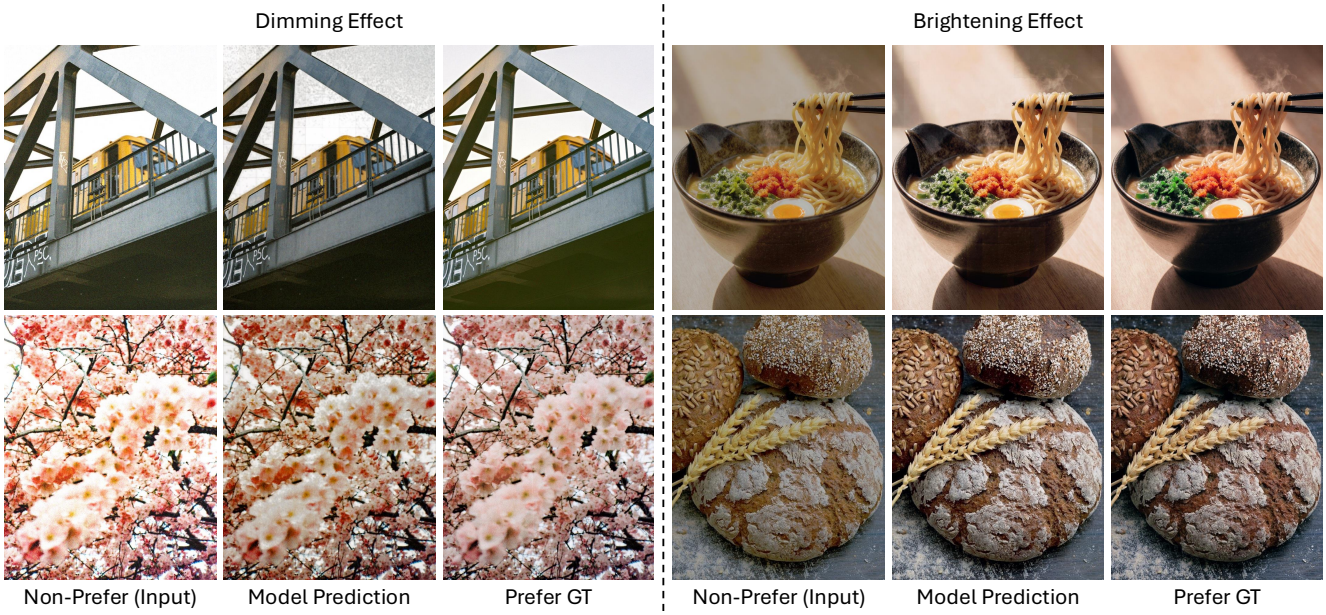


Figure L. Content-aware preference application for user gnBUxFI_desktop. The model (User Preference Embedding model) applies dimming to architecture and nature (left) while brightening food images (right).

main-paper CQS-based evaluation: Model (b) is most preferred, and Model (c) is slightly stronger than Model (a) while remaining broadly comparable, supporting CQS as a meaningful proxy for pair-wise comparison quality.

C.4. Per-User Retouching Variations

Figure L demonstrates how the trained models apply learned user preferences in a content-aware manner. The left column shows cases where the model applies dimming effects to images of architecture and nature scenes. The bridge image becomes darker with enhanced contrast, while the cherry blossom photo shifts toward muted tones. The right column shows the same user's model applying opposite adjustments to food subjects. The ramen bowl receives warmer tones and increased brightness, and the bread images are enhanced with stronger highlights.

These examples illustrate that the model does not learn a single global adjustment. Instead, it adapts the user's aesthetic preferences based on the content and context of each image. This content-aware behavior suggests that our baseline approaches can capture meaningful aspects of personal photographic style when trained on sufficient comparative preference data.

C.5. Additional Qualitative Results

Figure M(a-c) shows additional results from the three baseline approaches across diverse image types. When starting from non-preferred inputs, all methods produce outputs that move toward the preferred aesthetic. When starting from preferred inputs, the models tend to maintain the preferred color and tone rather than deviating significantly.

It is worth noting that outputs from non-preferred inputs may still remain closer to the non-preferred appearance than to the preferred ground truth. However, the important observation is that the models consistently shift in the direction of the preferred style. This demonstrates that the baseline approaches can learn meaningful aspects of personal preferences from comparative data.

Figure M(d) compares all three baseline methods on the same inputs. Different methods show varying degrees of success across images. For some cases, one method produces results closer to the expected output, while for others, a different method achieves better alignment. This variability reflects the difficulty of this problem. These results establish baseline capabilities while highlighting opportunities for methodological improvements.

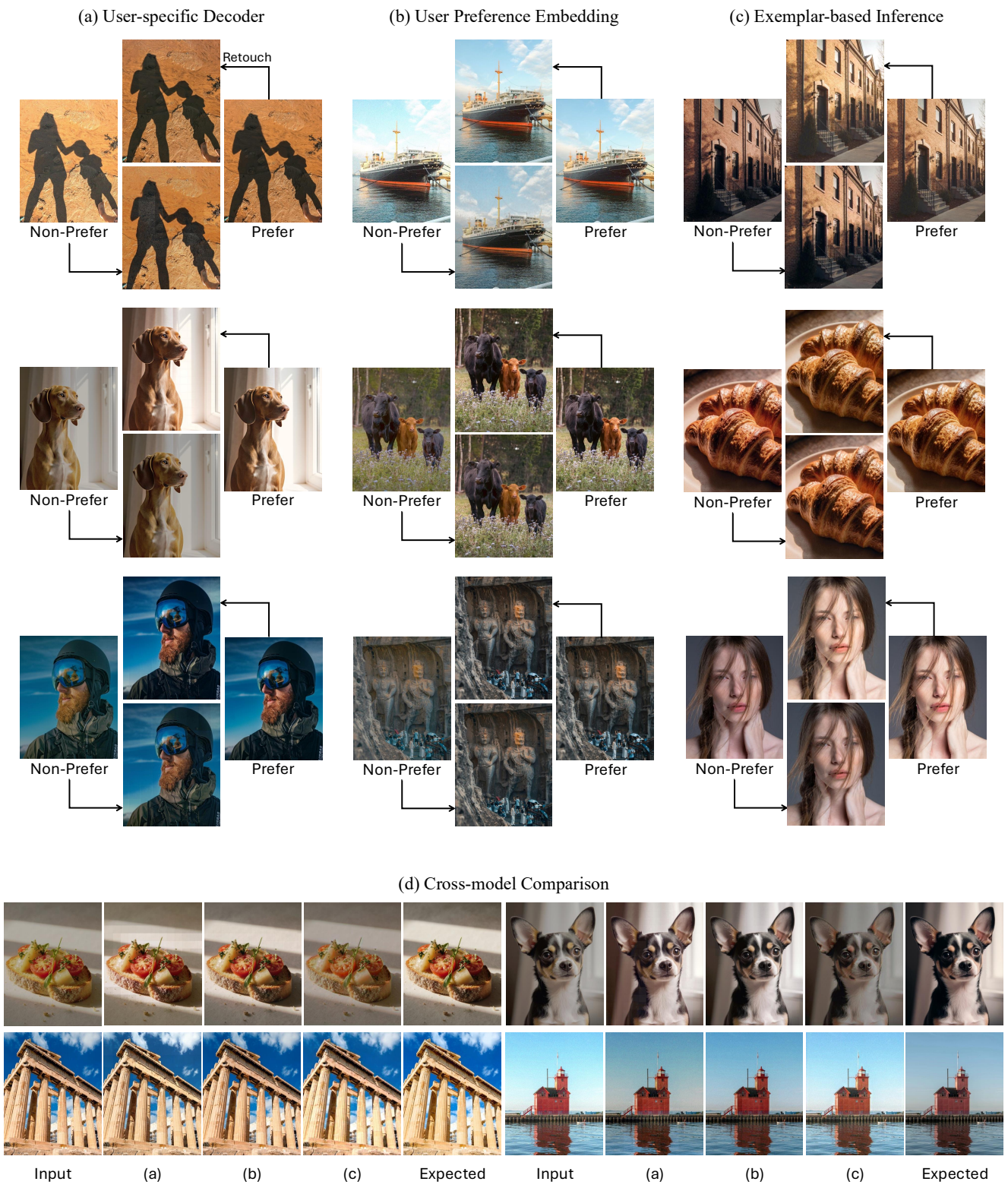


Figure M. Additional qualitative results for baseline methods. (a-c) Outputs from non-preferred and preferred inputs. (d) Cross-model comparison showing performance variation across different approaches.

D. Limitation and Discussion

Dataset Scale and Temporal Coverage. Our dataset captures an average of 70 pairwise comparisons per user, collected within a single session. While this scale demonstrates the feasibility of collecting personalized preference data and enables meaningful baseline training, it represents a limited snapshot of each user's aesthetic preferences. Personal style may require more extensive observation to fully characterize, particularly when preferences evolve over time or vary across different photographic contexts. Ideally, preference data would be collected continuously over extended periods, capturing how individual aesthetics develop and stabilize. Our work establishes a foundation at an academic scale, opening this research direction for future efforts.

Display-Agnostic Preference Modeling. The perception of photographic style is inherently tied to display characteristics. The same image array can produce different aesthetic experiences depending on display technology, calibration, and viewing conditions. While we made efforts to account for display factors by defining unique users as (user, display_type) pairs and collecting display information in our pre-survey, our approach does not fully address display-centric aspects of PPS. Furthermore, we do not capture dynamic display adaptations such as True Tone on iOS devices or adaptive color tone features on Galaxy devices, which adjust rendering based on ambient conditions. Understanding how display settings interact with learned preference models remains an important consideration for future work.

Proof-of-Concept Modeling Approaches. Our baseline methods demonstrate the viability of different learning paradigms for PPS but are not intended as optimized solutions. Each approach reveals different trade-offs and potential directions, yet we did not pursue a single, comprehensive architecture that fully addresses the challenges of this problem. Developing more sophisticated models that effectively integrate preference signals across varied photographic contexts remains an open challenge.

Evaluation Without Direct User Response. Our work establishes a reusable evaluation framework based on real-world user responses. While this method effectively validates performance against recorded preferences, we recognize the value of confirming results through live user interaction. Future work can build on this foundation by developing sustainable evaluation methods that are both low-cost and reproducible. We envision a system where offline validation is seamlessly integrated with streamlined human feedback loops.

References

[1] Pixabay. https://pixabay.com. Accessed: 2025-11-11. 5

[2] Unsplash. https://unsplash.com. Accessed: 2025-11-11. 5

[3] Higgsfield AI. Higgsfield ai, 2025. Accessed: 2025-11-11. 5

[4] Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, et al. Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space. *arXiv e-prints*, pages arXiv–2506, 2025. 5

[5] Black Forest Labs and Krea AI. Flux.1 krea [dev]. https://huggingface.co/black-forest-labs/FLUX.1-Krea-dev, 2025. Accessed: 2025-11-13. 3

[6] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input/output image pairs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 97–104. IEEE, 2011. 3

[7] Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025. 5

[8] Google. Nano banana - gemini ai image generator & photo editor, 2025. Accessed: 2025-11-11. 5

[9] Andrey Ignatov, Luc Van Gool, and Radu Timofte. Replacing mobile camera isp with a single deep learning model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 536–537, 2020. 3

[10] Inc. Midjourney. Midjourney, 2025. Accessed: 2025-11-11. 5

[11] Inc. Runway AI. Runwayml, 2025. Accessed: 2025-11-11. 5

[12] Chenfei Wu, Jiahao Li, Jingren Zhou, et al. Qwen-image technical report, 2025. 3

[13] Yazhou Xing, Changlin Li, Xuaner Zhang, and Qifeng Chen. A well-aligned dataset for learning image signal processing on smartphones from a high-end camera. In *SIGGRAPH 2022 Posters*, pages 1–2. ACM, 2022. 3