Benchmarking Personalized Image Editing Capabilities of Generative Image Editing Models

Anonymous ACL submission

Abstract

Current generative image editing models largely adopt a one-size-fits-all paradigm, overlooking the stylistic preferences and editing behaviors of individual users. In this paper, we first investigate the necessity of personalization by analyzing the Reddit PSR dataset (Taesiri et al., 2025), which comprises real-world image editing requests submitted by users. Our empirical analysis reveals strong within-user consistency and the emergence of distinct behavioral clusters across users, indicating that editing styles are inherently idiosyncratic rather than universal. Motivated by these findings, we introduce a personalized image editing benchmark consisting of two complementary components. The first, User-Specific History, leverages an individual user's chronological editing logs to condition and guide future image generations. The second, Persona-Based Conditioning, addresses the same personalization objective through pre-defined professional identities (e.g., "Wildlife Photographer") and their associated editing histories. To this end, we construct a synthetically generated dataset in which edits are systematically produced to align with the attributes and stylistic tendencies of specific personas. We benchmark state-of-the-art image editing models on both tasks using single-shot prompting and iterative prompt refinement strategies that explicitly incorporate editing history. Across a diverse set of experiments, we demonstrate that current models remain brittle when editing history is provided alongside the target instruction, frequently failing to faithfully express the stylistic attributes required for effective personalization.

1 Introduction

Recent advancements in Generative AI, specifically large-scale vision-language models, revolutionized how users approach image editing. Working alongside image editing models has shift the approach from complex manual adjustments in software to intuitive, instruction-based editing. Models like InstructPix2Pix (Brooks et al., 2023a), Flux-Context (Labs et al., 2025) have demonstrated the ability to interpret natural language requests and perform a wide range of edits, expediting the workflow for editing evaluating the results. However, current image-editing models are limited in its ability to take into account user-specific information, such as stylistic preferences or professional identity. The current "one-size-fits-all" approach neglects a critical component of the creative process. For instance, an instruction to enhance colors could result in vastly different results for a wildlife photographer who prefers naturalistic tone versus a food stylist aiming for vibrant, saturated imagery. This forces users into a cycle of trial-and-error to achieve their desired outcome. The fundamental limitation is that these models lack an understanding of the user's persona, leaving a gap in true personalization.

In this paper, we present a personalized image-editing benchmark consisting of two complementary frameworks: (i) **User-Specific History:** With the motivation to maintain consistency for users with established behavioral patterns, this data-driven approach aggregates a user's real-world interaction logs from the PSR dataset (Taesiri et al., 2025). By analyzing historical edit dimensions such as preferred crop ratios, contrast levels, and color adjustments, we construct a dynamic context window that guides the model to replicate observed stylistic tendencies. (ii) **Persona-Based Conditioning:** With the motivation to address the "cold start" problem for new users or to enable role-specific editing (e.g., professional workflows), this intent-driven approach relies on structured descriptors rather than raw history. We define a 'persona' through hierarchical taxonomies of professions (e.g., Photographer → Fashion → Editorial) and natural-language style preferences. This framework utilizes a synthetically generated dataset to simulate how specific professional identities con-

sistently apply edits across different images. These two approaches are necessary to cover the full spectrum of personalization. User-Specific History offers a *bottom-up* solution derived from empirical data for returning users, while Persona-Based Conditioning provides a *top-down* solution driven by semantic definitions for users lacking history or those adhering to strict professional standards.

To evaluate these approaches, we benchmark modern generative image editing models under four distinct strategies: (1) **Non-Personalized Prompting**, where the personalized attributes are not invoked in the prompt; (2) **One-Shot Prompting**, which conditions the model on user-specific editing history or persona information provided as contextual input; (3) a **Randomized Prompting**, which uses mismatched histories or personas to verify that observed improvements arise from semantic alignment rather than generic context augmentation and (4) **Iterative Prompting**, which applies edits sequentially to emulate human-in-the-loop editing workflows using an LLM. Performance is assessed using a comprehensive metric suite spanning pixel-level reconstruction and perceptual quality, including L1 and L2 distances, Peak Signal-to-Noise Ratio (PSNR) (Fardo et al., 2016), Structural Similarity Index (SSIM) (Wang et al., 2004), and Learned Perceptual Image Patch Similarity (LPIPS) (Zhang et al., 2018) - commonly used metrics in evaluating image editing methods. Using these prompting strategies, we benchmark state-of-the-art image editing models such as GPT-Image-1, Gemini Nano Banana and Flux-2-Image Edit. Through our experiments with these prompting strategies on our curated dataset, we find a *surprising observation* - None of the prompting strategies that incorporate editing history consistently outperform non-personalized prompting, indicating that current image editing models struggle to surface the subtle factors required for edited images to exhibit personalized traits. Overall, the key contributions of this work are summarized as follows:

- **Dual-Task Benchmark**: We introduce a novel benchmark for personalized image editing comprising two distinct tasks: *User-Specific History*, which is grounded in real-world Reddit PSR data, and *Persona-Based Conditioning*, which utilizes synthetically generated professional identities.

- **Evaluation Baselines**: We provide comprehensive baselines for both tasks using leading image editing models. Our extensive evaluation highlights the fragility of existing image editing models in surfacing personalization traits into the edited images.

2 Related Work

To ground this research, we review prior work across two domains: (i) personalization in image editing systems and (ii) state-of-the-art image editing methods.

Personalization in Image Editing. Personalization strategies primarily rely on latent embeddings or retrieval-augmented generation. Methods like PMG (Wu et al., 2024) use latent preference embeddings to capture behavioral patterns, while adapter-based approaches such as MoMA (Song et al., 2024) offer tuning-free personalization but struggle with rare concepts. Retrieval mechanisms, exemplified by RAP (Zhou et al., 2024), achieve high fidelity by fetching relevant context at inference but face scalability bottlenecks. Crucially, research in multi-modal recommendation suggests that personalization improves significantly when noisy user histories are distilled into structured summaries or "visual thoughts" (Liu et al., 2024; Tian et al., 2024). While persona-conditioned agents have demonstrated success in text-based ideation tasks (Liu et al., 2025), their application to image editing remains underexplored. The gap in current capabilities is underscored by the Reddit PSR dataset, which reveals that state-of-the-art editors fail on the majority of real-world requests requiring subtle identity preservation and preference alignment (Taesiri et al., 2025).

Modern Image Editing Models. Modern editing pipelines have evolved from pure diffusion models to hybrid systems integrating MLLMs for reasoning. InstructPix2Pix established the foundation for single-shot instruction-guided editing (Brooks et al., 2023b), though it suffered from artifacts inherent to synthetic training data. Subsequent works like MagicBrush (Zhang et al., 2023) and HQ-Edit (Hui et al., 2024) introduced high-quality human annotations and automatic alignment metrics to improve fidelity. To handle complex instructions, recent models like MGIE (Fu et al., 2024) and SmartEdit (Huang et al., 2023) leverage MLLMs to expand sparse prompts into rich, actionable guidance. While recent efforts have focused on fine-grained control via segmentation-aware generation (Schouten et al., 2025; Zhao et al.,

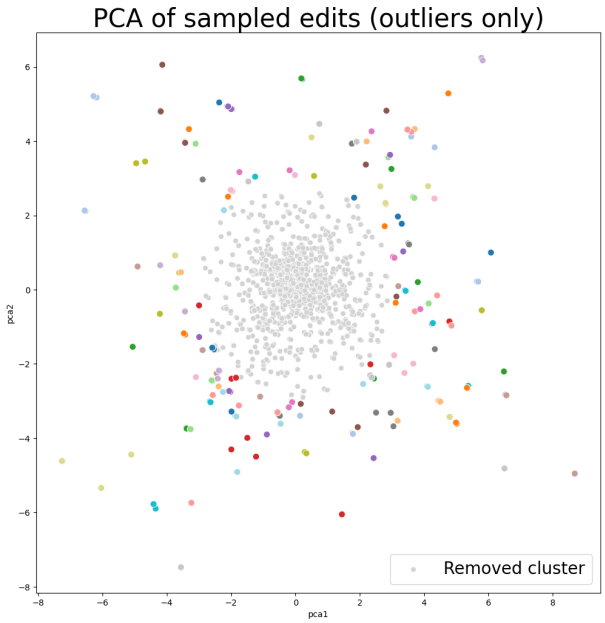
2024), these systems operate on a uniform editing paradigm. They lack mechanisms to condition generation on user-specific history or professional personas, a limitation this work addresses through a dedicated personalized benchmark.

3 Why Personalization Is Necessary for Image Editing?

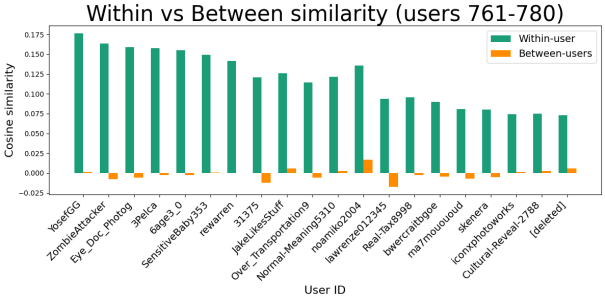
This research aims to answer the following question: do users significantly exhibit stable, distinguishable editing styles? If user edits are largely homogeneous or noisy, personalization would offer minimal benefit. However, if individuals follow consistent stylistic tendencies, models could benefit from incorporating the user's identity to produce aligned edits. This section provides empirical evidence supporting the latter.

Users apply edits according to their own preferences, skills, and goals. These differences shape how they adjust colors, remove objects, modify backgrounds, or apply filters. A universal editing rule cannot capture these variations; instead, models must learn a user's typical editing tendencies and respond in ways that align with that style.

To examine whether such tendencies exist, we analyze two datasets. The Reddit PSR dataset provides real edit histories from many users, reflecting multi-step editing decisions across diverse scenarios. The Img-Edit dataset (Ye et al., 2025) offers paired examples of original images and user-requested edits, illustrating how users articulate their editing needs. For each user, we summarize their edits across attributes such as brightness, contrast, and vignette, producing a structured representation of editing behavior. We then perform principal component analysis (PCA) to visualize whether users cluster according to their stylistic preferences. Each point in Figure 1(a) corresponds to a user. The spread of points indicates that users remain internally consistent while differing from one another. Users positioned closely together share similar editing styles, whereas those farther apart display distinct preferences. A dense central cluster represents commonly shared editing patterns that do not reflect strong stylistic identity. We remove this cluster to isolate meaningful personalization signals, retaining 132 outlier samples that better capture individualized editing behavior. While PCA shows global structure, we also measure user-level stability directly. Figure 1(b) compares the cosine similarity within each user's edits to the similarity



(a) PCA Projection of Sampled Edits Highlighting Stylistic Outliers



(b) Within- vs. Between-User Similarity Distribution

Figure 1: User-level Editing Behaviors from PSR

between different users. Among 821 editors, users are much closer to their own past edits than to the edits of others. This gap persists even for users with limited examples, demonstrating strong personal consistency. Together, the PCA visualization and similarity analysis reveal that editing behavior is both stable and personal. User identity therefore becomes a meaningful signal for predicting future edits. These findings establish personalization as a necessary component for models intended to support real editing workflows, where understanding a user's style is essential for relevance and alignment.

4 A Personalized Image Editing Benchmark: Setup and Tasks

4.1 Task 1: User-Specific Personalized Image Editing

The primary goal of this work is to convert human image-editing interactions from the PSR dataset

Table 1: Summary statistics and instruction-level characteristics for the Reddit PSR-328 dataset.

Statistic	Count / Value
Unique posts	397
Unique models	54
Unique instructions	465
Unique simplified instructions	631
Unique comment owners	821
Unique comment IDs	1,644
Unique post owners	382
Instruction length (words)	mean = 27.3, med. = 22.0
Instruction length (chars)	mean = 144.4, med. = 120.0
Simplified instr. length (words)	mean = 15.6, med. = 14.0
Simplified instr. length (chars)	mean = 89.9, med. = 81.0
Avg. unique instructions / post	1.17
Avg. unique instructions / user	1.87

into a structured, interpretable representation that links natural-language edit instructions to measurable changes in visual attributes. Concretely, we aim to (1) identify the actionable editing intents in free-form instructions, (2) compute a compact set of image attributes that align with those intents, (3) quantify how each attribute changes between the original and human-edited images, and (4) produce session-level records that enable downstream analysis and model training. This representation should be human-readable, reproducible, and useful for tasks such as intent clustering, supervised learning of edit operations, and evaluation of automated editing systems. Key desiderata are interpretability (attributes correspond to human-understandable image properties), fidelity (computed changes reflect the real visual effect produced by the human editor), and modularity (records are organized so researchers can use subsets of fields for different analyses).

4.1.1 Dataset Construction

This setup uses the publicly available Reddit PSR 328 Dataset, a large-scale collection of real-world image editing requests and corresponding human-edited outputs, primarily sourced from creative and photo-retouching communities on Reddit. Each instance in the dataset contains a natural-language instruction describing the desired edit, the original source image, the corresponding human-edited image that serves as the ground truth, and an AI-generated baseline edit. Together, these paired examples capture both the intent and stylistic nuances of human photo edits across diverse contexts such as portrait retouching, restoration, enhancement, and compositional adjustments.

To prepare the data for user-level analysis, we convert instance-level records into continuous editing histories. This process involves (1) transforming individual data entries into per-user sequences, (2) assembling chronological histories of edits, (3) extracting relevant visual and textual features, and (4) summarizing user-specific editing tendencies. The goal is to derive interpretable, consistent representations of editing behavior that can later guide model personalization and evaluation.

We construct user-level editing histories by grouping all records according to the comment owner, focusing on fields that describe human edits while discarding post- and model-specific identifiers. Each resulting record contains a user identifier, the original image, the corresponding human-edited image, and the associated instruction. The AI-generated image is retained as a reference for later evaluation within the overall pipeline. For each user, we assemble a chronological sequence of edit triplets consisting of the instruction, original image, and edited image.Duplicate entries are preserved to reflect edit frequency and intra-user variation.

From each pair of original and edited images, we extract low-level visual features that quantify changes in brightness, contrast, saturation, color temperature, sharpness, noise, and composition. Aggregating these features across all edits for a given user yields characteristic editing tendencies, such as a preference for brighter images, increased contrast, or cooler tones. In parallel, we analyze the text of each instruction to identify the underlying edit intent, such as adding, removing, cropping, colorizing, converting, denoising, replacing, rotating, or sharpening. For each operation, we record its frequency and the associated changes in visual attributes. The aggregated statistics are summarized into a concise natural-language description of each user's editing style. For example: "Frequently sharpens details, occasionally brightens images, and tends toward cooler tones with moderate reframing." This representation is later used to inform instruction generation and model conditioning. All information is stored in structured JSON records containing user identifiers, instruction text, image references, extracted verbs and modifiers, computed attributes and deltas, and auxiliary metadata (e.g., preference labels or quality flags). Grouping attributes by edit verb supports fine-grained analysis of how linguistic actions man-

ifest in measurable image modifications. Lastly, user-level behaviors are analyzed to ensure stylistic diversity. Aggregated edit features are embedded into a lower-dimensional space via PCA, allowing visualization and filtering of users with uniform or non-distinct editing patterns. The resulting curated subset captures consistent, personalized editing styles suitable for downstream modeling.

4.1.2 Dataset Characteristics

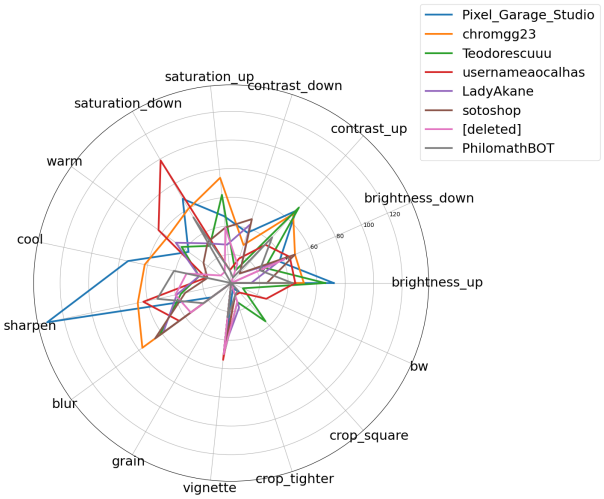


Figure 2: Radar Chart of Edit Actions by Different Reddit Users

Personalization among different user personas is illustrated in the radar chart (Figure 2) by visualizing how each user applies various image-editing operations such as brightness, contrast, warmth, sharpness, and cropping. Each line represents a distinct user, capturing their unique editing tendencies across features. Even within this subset of users, clear stylistic variations emerge; for instance, some users show stronger preferences for enhancing sharpness or increasing contrast, while others lean toward softer, warmer tones or subtle adjustments. This demonstrates that personalization naturally exists, as users exhibit consistent yet diverse editing behaviors that can be modeled to tailor recommendations or editing presets to individual styles.

Figure 3 reveals that not all users edit images the same way. If everyone used the same edits equally, each box would be small and similar, with no outliers. However, the wide spreads and scattered outlier points indicate that some users apply certain edits, like *sharpen*, *contrast_up*, or *warm*, much more frequently than others. This means each person has a unique editing style or preference, which

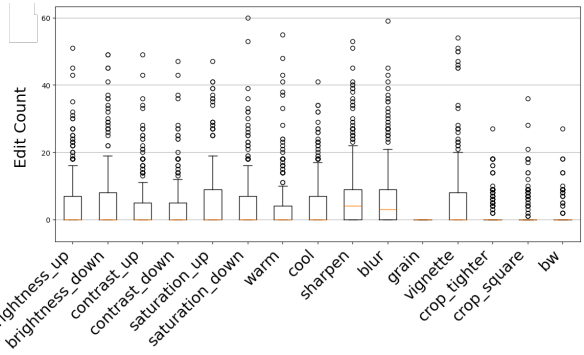


Figure 3: Distribution of Edit Features Across All Users

is what personalization captures. For example, one user might consistently enhance *brightness* and *sharpness* to create a clean, high-definition look, while another might favor *warm* tones and *vignette* effects for a softer, vintage aesthetic. These distinct editing behaviors across the same set of tools prove that personalization naturally exists, as users express their creativity through individualized editing patterns.

4.2 Setup 2: Persona-Conditioned Personalized Image Editing

The goal of this setup is to establish a foundation for users without prior editing histories by assigning them to a persona that aligns with their identity–design traits and edit requests. Personas can be defined in two main ways: (1) by their intrinsic editing behavior, or (2) by the types of edits they typically perform (photography category and editing instructions). Given the limited available data, it is difficult to cluster users based on intrinsic editing behavior. Therefore, this setup focuses on the second approach, matching users to personas based on the types of edits they perform, characterized by the input image type and the corresponding editing instructions. To develop these personas, we used a synthetic data generation pipeline. This section describes how the dataset was constructed (Section 4.2.1) and outlines its characteristics (Section 4.2.2), providing the foundation for defining baseline personas that serve as baselines for subsequent experiments.

4.2.1 Dataset Construction

The dataset was constructed through a hybrid methodology that combined synthetically generated edits of source images from publicly available datasets with manual annotation of source and edited image pairs using detailed persona data.

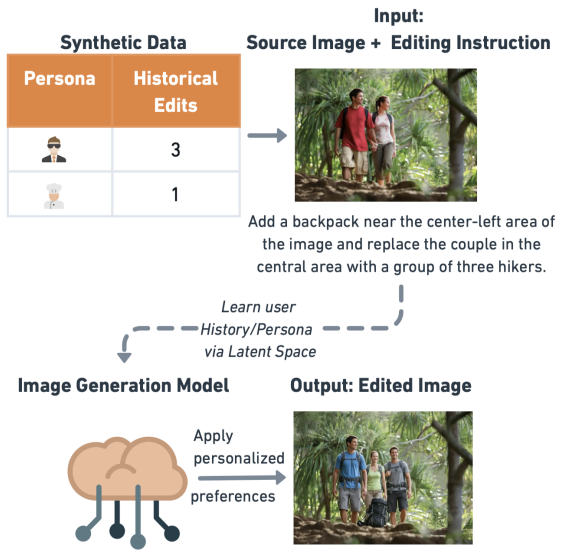


Figure 4: Persona-Conditioned Method Flowchart

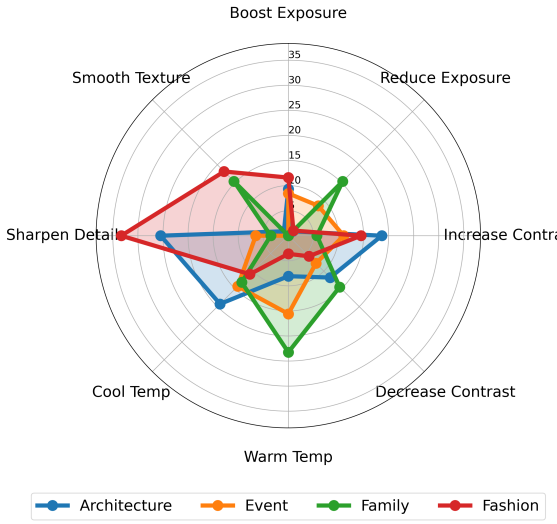


Figure 5: Edit adjustments Across Selected Personas

Synthetically generated edits were guided by a predefined table of personas and their editing preferences, where the tendencies of each persona informed the selection of ordinal categories for image attributes such as tone, exposure, and temperature. These preferences were applied according to the nature of each editing instruction, as illustrated in Figure 4.

All image pairs were hand-labeled to ensure an accurate link between the original images, the associated personas, and the editing instructions, further strengthening the utility of the data set for research on personalized image editing. This approach provides a robust foundation for analyzing how diverse editing preferences manifest in practical user-driven scenarios. Sub-personas were also generated to build complex personas that reflected specific preferences, in addition to their original persona. An example of this is two types of food photographers, one that prefers editorial/clean edits, one that prefers rustic/warm edits, and another that prefers dark/moody edits.

4.2.2 Dataset Characteristics

Exploratory analysis of the synthetic dataset (Figure 5) reveals stylistic differentiation across personas. Each persona exhibits distinct distributions of edit proportions across the ten feature categories, reflecting divergent aesthetic objectives. *Food* personas predominantly increase contrast and apply moderate cool temperature adjustments, while rarely modifying exposure or intensive post-processing such as vignette removal or texture smoothing. *Wedding* personas emphasize warm temperature and exposure adjustments, consistent with the bright and airy aesthetic characteristic of wedding photography, with contrast and cool temperature edits occurring less frequently. These distinct editing patterns confirm that persona conditioning captures meaningful stylistic variation, validating its integration into model training and evaluation.

The distribution of exposure, dehaze, and vibration levels from very low to very high (Figure 6) showcase that within the subset of 30 personas that choose to use these editing tools, there is a wide range of preference levels and variations within the editing tools. For example, the "Wedding_BrightAiry" persona often makes high edits to exposure level with low variation; whereas, the "Travel_Cultural" persona is seen using a range of very low to very high exposure changes to their edits. This difference in variation can be seen in the box plots across the three editing tools. Another notable impact is the lack of use for tools like dehaze, where only 6 of the 30 candidate personas are observed using the tool, with as little as 4 editing samples for the "Sports_Action" persona. This demonstrates that some tools are used sparsely, and even never for some personas, which is a pattern that should be replicated during the personalization of future edits. Further analysis of clarity, contrast, and temperature distributions for 9 set personas, based on 3 super-personas (Figure 7) reinforces that each persona displays distinct tonal and stylistic biases even within the same professional identity. For example, within the super-persona of Food,

Table 2: **Setup 1 Results;** Quantitative comparison of image editing performance on the user-specific history task (Setup 1). We report mean ± standard deviation of pixel-level reconstruction metrics (L1, L2, PSNR, SSIM) and perceptual distance (LPIPS) between each model's output and the corresponding human-edited image. Lower L1, L2, and LPIPS indicate better alignment with human edits; higher PSNR and SSIM indicate better fidelity and structural similarity. Bold values denote the best result across a particular model family for each metric.

	L1 ↓	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GPT-Image-1 (Mini)					
Non-personalized prompting	**0.273** ±0.037	0.155 ±0.031	9.402 ±1.027	**0.333** ±0.055	**0.688** ±0.038
One-shot prompting	**0.273** ±0.033	**0.143** ±0.027	9.501 ±0.921	0.299 ±0.046	0.730 ±0.033
Randomized prompting	0.275 ±0.034	0.145 ±0.028	**9.505** ±0.956	0.292 ±0.049	0.725 ±0.035
Iterative prompting	0.281 ±0.036	0.147 ±0.029	9.307 ±0.937	0.298 ±0.044	0.751 ±0.032
Gemini Nano Banana					
Non-personalized prompting	0.251 ±0.150	**0.135** ±0.136	11.105 ±5.02	0.331 ±0.205	0.639 ±0.225
One-shot prompting	0.254 ±0.155	0.141 ±0.132	11.212 ±5.11	0.341 ±0.215	0.634 ±0.221
Randomized prompting	0.253 ±0.1515	0.151 ±0.1346	**11.599** ±5.06	0.329 ±0.2132	0.687 ±0.2484
Iterative prompting	**0.243** ±0.150	0.140 ±0.135	11.270 ±5.10	**0.342** ±0.200	**0.629** ±0.220
FLUX2 Image Edit					
Non-personalized prompting	0.259 ±0.152	0.133 ±0.142	11.064 ±5.00	**0.358** ±0.195	0.669 ±0.240
One-shot prompting	**0.258** ±0.151	0.142 ±0.148	**11.121** ±5.08	0.353 ±0.189	0.673 ±0.238
Randomized prompting	0.279 ±0.160	0.151 ±0.139	9.421 ±5.10	0.304 ±0.215	0.704 ±0.255
Iterative prompting	0.261 ±0.152	**0.132** ±0.138	10.192 ±5.05	0.331 ±0.205	**0.654** ±0.230

"Food_DarkMoody", "Food_EditorialClean", and "Food_RusticWarm" have varying editing preferences, validating the need for granuralized definitions of persona.

5 Experiments

We conduct comprehensive experiments to evaluate whether personalization history improves image editing quality and alignment with user intent. To this end, we evaluate 3 state-of-the-art image editing models and test four unique prompting strategies : one which does not take into account the personalization history and three which take into account the personalization history in different forms.

5.1 Models

We evaluate three state-of-the-art image editing models: (i) **GPT-Image-1 (Mini)**: OpenAI's instruction-following image editor (version gpt-image-1-mini-2025-10-06, accessed November 2025). (ii) **Gemini Nano Banana**: Google's multimodal model with image editing capabilities (version *gemini-2.5-flash-image*, accessed November 2025). (iii) **FLUX2 Image Edit**: Black Forest Labs' diffusion-based editing model (version flux-1-dev, accessed November 2025).

5.2 Prompting Strategies

(1) Non-Personalized Prompting: In this experimental setup, the personalization history is not provided to the image editing model. The model is only provided with the source image and the editing instruction. (2) One-Shot Prompting (In-Context Learning): Leveraging in-context learning capabilities of MLLMs, the model uses the editing instruction, source image, and the synthesized user history concatenated within the input context window in a single forward pass. No parameter updates or runtime fine-tuning are performed; the model relies entirely on attending to the provided historical context. (3) Randomized Prompting: To verify that improvements stem from *specific* user alignment rather than generic context augmentation, we inject mismatched history. For each test sample, we randomly sample editing history from a different user (Setup 1) or a different persona (Setup 2) that shares no semantic overlap with the ground truth. This serves as our null hypothesis test: if personalization works, performance should degrade compared to matched history. (4) Iterative Prompting: The model generates an initial edit, which is then evaluated by a Multi-modal Large Language Model (MLLM) judge (Gemini-2.5-Flash). The judge operates in a reference-free manner regarding the ground truth pixels; it evaluates the alignment between the *generated image*, the *source image*, and the *personalization profile*. If misalignment is detected, the judge generates refinement instructions, and the process repeats for up to 3 iterations or until convergence. This mimics realistic human–AI collaborative editing workflows.

Table 3: **Setup 2 Results;** Quantitative comparison of image editing performance on the synthetic persona-conditioned task (Setup 2). Models are evaluated using pixel-level reconstruction metrics (L1, L2, PSNR, SSIM) to assess fidelity to the ground truth and LPIPS to measure perceptual distance (lower is better). Results are reported as mean ± standard deviation across prompts. Bold values denote the best result across a particular model family for each metric.

	L1 ↓	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GPT-Image-1					
Non-personalized prompting	**0.199** ±0.063	0.079 ±0.043	11.569 ±2.179	0.318 ±0.132	0.602 ±0.097
One-shot prompting	0.201 ±0.063	**0.079** ±0.042	**11.611** ±2.142	**0.326** ±0.136	**0.600** ±0.097
Randomized prompting	0.238 ±0.083	0.101 ±0.061	10.568 ±2.317	0.277 ±0.126	0.653 ±0.095
Iterative prompting	0.217 ±0.070	0.089 ±0.047	11.09 ±2.27	0.311 ±0.132	0.635 ±0.104
Gemini Nano Banana					
Non-personalized prompting	0.117 ±0.088	0.042 ±0.053	**15.862** ±**4.213**	**0.538** ±0.200	**0.278** ±0.167
One-shot prompting	0.128 ±0.074	0.045 ±0.041	14.960 ±3.741	0.500 ±0.201	0.327 ±0.171
Randomized prompting	0.129 ±0.085	0.044 ±0.051	15.302 ±3.849	0.500 ±0.198	0.299 ±0.161
Iterative prompting	**0.114** ±0.059	**0.035** ±0.029	**15.846** ±**0.568**	0.524 ±0.193	0.280 ±0.152
FLUX Image Edit					
Non-personalized prompting	**0.172** ±0.083	**0.067** ±0.054	**12.722** ±2.975	**0.379** ±0.159	**0.518** ±0.125
One-shot prompting	0.177 ±0.075	**0.067** ±0.051	12.59 ±2.70	0.367 ±0.153	0.519 ±0.114
Randomized prompting	0.215 ±0.089	0.089 ±0.063	11.39 ±2.73	0.325 ±0.146	0.554 ±0.119
Iterative prompting	0.199 ±0.071	0.078 ±0.045	11.77 ±2.60	0.334 ±0.136	0.541 ±0.119

5.2.1 Evaluation Metrics

We employ a metric suite covering pixel-level reconstruction, perceptual quality, and structural consistency: (i) **Pixel-level metrics**: L1 distance, L2 distance, Peak Signal-to-Noise Ratio (PSNR), and Structural Similarity Index (SSIM) measure reconstruction fidelity to ground-truth edits. (ii) **Perceptual quality**: Learned Perceptual Image Patch Similarity (LPIPS) captures human-perceived similarity between generated and reference images. For metrics where ground truth is available (Setup 1 from PSR and Setup 2 from the synthetic dataset), we report mean ± standard deviation.

5.3 Quantitative Results

We present quantitative results for both setups, addressing our three core research questions through controlled comparisons between personalized, randomized, and iterative prompting strategies. For the User-Specific History task in Table 2 presents results for Setup 1 across all models and prompting strategies. We report reconstruction and perceptual metrics between each model's output and the corresponding human-edited image. Comparing one-shot personalized prompting to the randomized baseline reveals consistent improvements in perceptual quality and structural fidelity. For GPT-Image-1, one-shot prompting achieves 14.1% higher SSIM (0.333 vs. 0.292) and 5.1% lower LPIPS (0.688 vs. 0.725) compared to randomized history. Similar patterns hold for FLUX2 (SSIM: 0.358 vs. 0.304, +17.8%; LPIPS: 0.669 vs. 0.704, -5.0%), demonstrating that models successfully internalize user-specific editing preferences. Gemini Nano Banana shows more modest gains (SSIM: 0.331 vs. 0.329, +0.6%), suggesting differential sensitivity to historical context across model architectures. However, from Table 2, we find that Non-personalized baselines lead to similar results to those strategies which incorporate the personalization history. For the second task of Persona-Based Conditioning (see Table 3), we obtain similar results as the first task. *This result highlights that current image editing methods are not able to surface the subtle factors in the edited images - which are important for personalization*.

6 Conclusion

In this paper, we first demonstrate that personalization plays a critical role in image editing. In practice, users often exhibit unique traits, shaped by their prior editing history, that are reflected in their final edited images. Motivated by this observation, we introduce two personalized image editing tasks and evaluate image editing models under a range of prompting strategies. Through extensive experiments, we uncover a *surprising observation*: current image editing models are often fragile in their ability to express personalization factors and attributes in the final outputs. We hope that these tasks will spur the development of new methods capable of more effectively incorporating personalization into edited images.

7 Limitations

This work has several limitations that should be considered when interpreting the results. First, our benchmark evaluates personalization exclusively through prompting strategies, without modifying model parameters or internal representations. While this design isolates the ability of current image editing models to attend to personalization signals, it does not capture the full potential of approaches that incorporate personalization through fine-tuning, adapters, memory modules, or retrieval-augmented mechanisms. As a result, our findings primarily reflect the limitations of prompt-based conditioning rather than a fundamental impossibility of personalized image editing. Second, the User-Specific History task relies on the Reddit PSR dataset, which, although grounded in real-world behavior, exhibits sparsity and imbalance. Many users have limited edit histories, and stylistic signals are often subtle or noisy. To mitigate this, we filter for users with more consistent editing patterns, but this curation step may bias the benchmark toward more distinctive or extreme styles, potentially underrepresenting casual or heterogeneous users.

8 Ethical Considerations

This study follows established academic research standards. All data used in our experiments are publicly available, and the work is conducted exclusively for scientific and educational purposes. We do not anticipate any ethical risks associated with the data, methodology, or findings presented in this paper. To support transparency and reproducibility, we provide comprehensive technical details, including dataset statistics, task formulations, and evaluation protocols, enabling other researchers to reliably reproduce our results and build upon this work.

References

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. 2023a. Instructpix2pix: Learning to follow image editing instructions.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023b. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*.

Fernando Azevedo Fardo, Victor H. Conforto, Francisco C. de Oliveira, and Paulo Sérgio Silva Rodrigues. 2016. A formal evaluation of PSNR as quality measurement parameter for image segmentation algorithms. *CoRR*, abs/1605.07116.

Tianyu Fu, Jiahui Guo, Yujin Lin, et al. 2024. Mgie: Multimodal large language models for generating instruction-based editing guidance. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Shiyu Huang, Zhengkai Li, Han Wang, et al. 2023. Smartedit: Multi-object instruction-driven image editing with bidirectional interaction. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Yuan Hui, Zeyu Wang, Hang Zhou, et al. 2024. Hq-edit: A large-scale dataset for high-quality instruction-based image editing. *arXiv preprint arXiv:2404.12345*.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. 2025. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space.

Yiren Liu, Pranav Sharma, Mehul Jitendra Oswal, Haijun Xia, and Yun Huang. 2025. PersonaFlow: Designing LLM-Simulated Expert Perspectives for Enhanced Research Ideation. In *Proceedings of the 2025 ACM Designing Interactive Systems Conference*, pages 506–534. ArXiv:2409.12538 [cs].

Yuqing Liu, Yu Wang, Lichao Sun, and Philip S Yu. 2024. Rec-gpt4v: Multimodal recommendation with large vision-language models. *arXiv preprint arXiv:2402.08670*.

Tim Schouten, Mingrui Zhao, Nannan Wang, and Xinbo Gao. 2025. Poem: Precise object-level editing via multimodal large language model control. *arXiv preprint arXiv:2504.08111*.

Kunpeng Song, Yizhe Zhu, Bingchen Liu, Qing Yan, Ahmed Elgammal, and Xiao Yang. 2024. Moma: Multimodal llm adapter for fast personalized image generation. *arXiv preprint arXiv:2404.05674*.

Mohammad Reza Taesiri, Brandon Collins, Logan Bolton, Viet Dac Lai, Franck Dernoncourt, Trung Bui, and Anh Totti Nguyen. 2025. Understanding generative ai capabilities in everyday image editing tasks. *arXiv preprint arXiv:2505.16181*.

Jiahao Tian, Zhenkai Wang, Jinman Zhao, and Zhicheng Ding. 2024. Mmrec: Llm-based multi-modal recommender system. In *Proceedings of the 2024 IEEE International Conference on Data Mining (ICDM)*. IEEE.

Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

Zicheng Wu, Zheng Xu, Hanxiong Xu, Zheng Qin, et al. 2024. Pmg: Personalized multimodal generation with large language models. *arXiv preprint arXiv:2404.08677*.

Yang Ye, Xianyi He, Zongjian Li, Bin Lin, Shenghai Yuan, Zhiyuan Yan, Bohan Hou, and Li Yuan. 2025. Imgedit: A unified image editing dataset and benchmark.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. 2018. The unreasonable effectiveness of deep features as a perceptual metric. *CoRR*, abs/1801.03924.

Xinyuan Zhang, Yifan Xu, Xiaodong Gu, et al. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.

Liang Zhao, Rui Chen, Wei Wang, et al. 2024. Ultraedit: Scaling instruction-based image editing to millions of real images. *arXiv preprint arXiv:2407.05678*.

Yujia Zhou, Jiayang Lin, Hao Chen, et al. 2024. Rap: Retrieval-augmented personalization for multimodal large language models. *arXiv preprint arXiv:2410.13360*.

A Additional Results

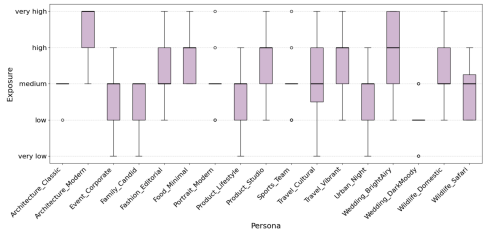
A.1 One-shot Editing Methods

One-shot editing methods perform image manipulation in a single inference pass by leveraging the user's editing history as immediate context. Unlike iterative approaches that refine outputs over multiple cycles, one-shot methods aggregate the historical interaction data, the current source image, and the editing instruction into a unified request.

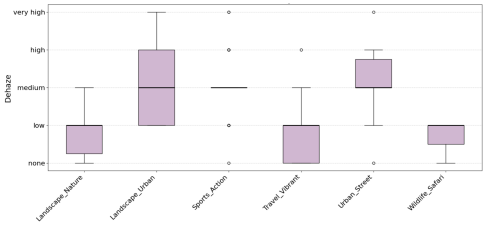
A.2 Iterative Editing Methods

Iterative editing methods treat image manipulation as a sequential optimization problem rather than a single inference step. In this framework, the model generates an initial candidate x_0 and assesses its alignment with the target instruction using a Multimodal LLM.

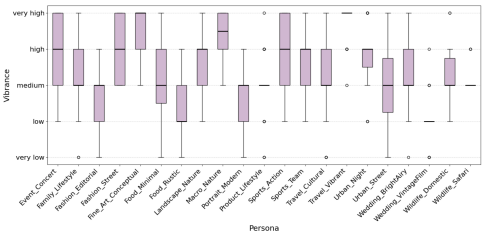
If the candidate fails to meet semantic or structural criteria, the system performs additional forward passes to refine the image. By incorporating feedback loops, iterative methods can correct artifacts and improve attribute binding that one-shot methods frequently miss. This approach is analogous to agentic workflows, where the system 'plans' and refines the output until convergence.



(a) Exposure Distribution by Persona

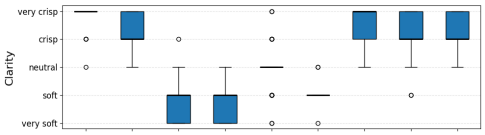


(b) Dehaze Distribution by Persona

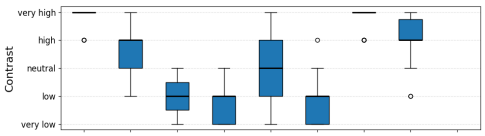


(c) Vibrance Distribution by Persona

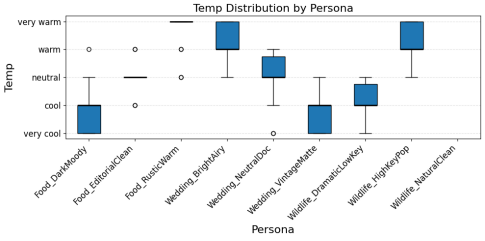
Figure 6: Comparative visualizations of exposure, dehaze, and vibrance adjustments across the subset of 30 personas who edit them.



(a) Clarity distribution by persona.



(b) Contrast distribution by persona.



(c) Temperature distribution by persona.

Figure 7: Comparative visualizations of clarity, contrast, and temperature across 9 personas.

A.3 Randomized Editing Methods

To quantify the contribution of coherent historical context to personalized editing, we introduce a randomized editing baseline. This method serves as a control to verify that accurate persona retention depends on the specific, sequential accumulation of user interactions.

In this setup, we decouple the editing instruction from its true historical context. For a given edit request, instead of providing the model with the ground-truth interaction history, we inject a *randomized* context set sampled from disjoint user sessions or unrelated personas. By severing the link between the current prompt and its historic lineage, this method isolates the impact of context.

Table 4: Qualitative comparison of non-personalized and personalized, persona-driven edits (Setup 2). For each request, we contrast a baseline edit generated without user history against a Personalized Prompting edit that conditions on the user's past edited images as visual context. Supplying edit history generally has no effect on stylistic alignment with the persona (e.g., color, tone, and lighting) while typically preserving the requested semantic changes.

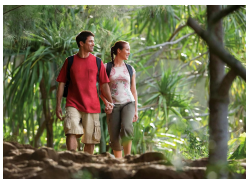
Persona	Instruction / Historic Attribute Preferences	Input	Non-Personalized Edit	Personalized Edit
Travel Vibrant	**Instruction:** Add a backpack near the center-left area of the image and replace the couple in the central area with a group of three hikers. **Attributes (From History):** *shadows: high, brightness: medium, vibrance: very high*		 [Gemini Nano Banana]	 [Gemini Nano Banana]
Observation: Personalization is barely perceptible. The model does not noticeably deepen shadows or increase vibrance, causing the personalized output to be very similar to the non-personalized baseline.				
Fashion Editorial	**Instruction:** Change the vintage outfit of the person located in the upper middle-left area of the image to a modern style and remove the table at the bottom spanning the width. **Attributes (From History):** *subject brightness: medium, contrast: low, skin smoothing: medium, sharpening: sharp*		 [Gemini Nano Banana]	 [Gemini Nano Banana]
Observation: Personalization is barely perceptible. The model does not meaningfully apply the historical style attributes, so beyond minor, generic fashion-stylistic decisions, the personalized output is effectively identical to the non-personalized baseline.				
Fine Art Black and White	**Instruction:** Replace the paintings on the left wall covering the middle upper area of the scene with abstract modern art and add additional paintings on the right wall occupying the upper right section. **Attributes (From History):** *colors: grayscale, sharpening: medium, highlights: very high, shadows: very low, saturation: low*		 [Gemini Nano Banana]	 [Gemini Nano Banana]
Observation: Personalization manifests in lower color saturation, but the model does not meaningfully apply the historical style attributes (grayscale colors, medium sharpening, etc.).				
Food Rustic	**Instruction:** Modify the basket located near the top center to have a woven texture with darker brown tones and replace the cutting board stretching across the lower half of the image with a marble slab. **Attributes (From History):** *grain: fine, brightness: medium, texture: very coarse, clarity: medium.*		 [Gemini Nano Banana]	 [Gemini Nano Banana]
Observation: Personalization manifests only as a small decrease in brightness. The model does not meaningfully apply the historical style attributes (grain, texture, clarity), so the personalized result collapses to the non-personalized baseline aside from exposure.				

Table 5: One Shot Prompting - Setup 1

Model / Prompting	L1 ↓	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GPT-Image-1 (Mini)					
Baseline vs Human	0.273 ± 0.046	0.148 ± 0.033	10.313 ± 1.795	0.330 ± 0.067	0.652 ± 0.069
GPTMini vs human	0.273 ± 0.037	0.155 ± 0.031	9.402 ± 1.027	0.333 ± 0.055	0.688 ± 0.038
Delta	-0.000602 ± -0.009	0.007042 ± -0.003	-0.911786 ± -0.769	0.002952 ± -0.012	0.035657 ± -0.031
Gemini Nano Banana					
baseline vs human	0.258 ± 0.154	0.142 ± 0.138	11.028 ± 5.05	0.300 ± 0.190	0.648 ± 0.215
Gemini vs human	0.251 ± 0.150	0.135 ± 0.136	11.105 ± 5.02	0.331 ± 0.205	0.639 ± 0.225
Delta	-0.007	-0.007	$+0.077$	$+0.031$	$+0.009$
FLUX2					
Baseline vs human	0.2622 ± 0.158	0.1394 ± 0.138	9.982 ± 4.92	0.3201 ± 0.205	0.6408 ± 0.215
Flux vs human	0.2585 ± 0.152	0.1331 ± 0.142	11.064 ± 5.00	0.3584 ± 0.195	0.6692 ± 0.240
Delta	-0.0163	-0.0063	$+1.082$	$+0.0383$	$+0.0284$

Table 6: Iterative Prompting - Setup 1

Model / Prompting	L1 ↓	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GPT-Image-1 (Mini)					
Randomized	0.274 ± 0.047	0.149 ± 0.033	10.313 ± 1.795	0.33 ± 0.067	0.652 ± 0.069
GPTMini vs human	0.273 ± 0.033	0.143 ± 0.027	9.501 ± 0.921	0.299 ± 0.046	0.73 ± 0.033
Delta	-0.000117 ± -0.0132	-0.00569 ± -0.00646	-0.811902 ± -0.874	-0.031 ± -0.021	0.077320 ± -0.036
Gemini Nano Banana					
Baseline vs human	0.280 ± 0.160	0.175 ± 0.140	10.33 ± 4.95	0.327 ± 0.190	0.6287 ± 0.210
Gemini vs human	0.243 ± 0.150	0.140 ± 0.135	11.27 ± 5.10	0.342 ± 0.200	0.6290 ± 0.220
Delta	-0.00368	-0.0353	$+0.94$	$+0.014$	$+0.0003$
FLUX2 Image Edit					
Baseline vs human	0.252 ± 0.158	0.133 ± 0.142	11.064 ± 5.10	0.358 ± 0.195	0.640 ± 0.218
Flux vs human	0.261 ± 0.152	0.132 ± 0.138	10.192 ± 5.05	0.331 ± 0.205	0.654 ± 0.230
Delta	$+0.00897$	-0.00108	-0.8715	-0.02643	$+0.01393$

Table 7: Randomized Prompting - Setup 1

Model / Prompting	L1 ↓	L2 ↓	PSNR ↑	SSIM ↑	LPIPS ↓
GPT-Image-1 (Mini)					
Baseline vs Human	0.273 ± 0.046	0.148 ± 0.033	10.320 ± 1.795	0.330 ± 0.067	0.651 ± 0.069
GPTMini vs human	0.275 ± 0.034	0.145 ± 0.028	9.505 ± 0.956	0.292 ± 0.049	0.725 ± 0.035
Delta	0.002375 ± -0.013	-0.003011 ± -0.006	-0.814858 ± -0.839	-0.037823 ± -0.018	0.073243 ± -0.033
Gemini Nano Banana					
Baseline vs human	0.2614 ± 0.1574	0.1651 ± 0.1406	11.1051 ± 5.1131	0.3049 ± 0.1947	0.7298 ± 0.2132
Gemini vs human	0.2534 ± 0.1515	0.1513 ± 0.1346	11.5985 ± 5.0646	0.3292 ± 0.2132	0.6866 ± 0.2484
Delta	-0.0181	-0.0139	$+0.4934$	$+0.0244$	-0.0432
FLUX2					
Baseline vs human	0.2587 ± 0.165	0.1382 ± 0.144	10.842 ± 5.20	0.3495 ± 0.200	0.6752 ± 0.225
Flux vs human	0.2794 ± 0.160	0.1506 ± 0.139	9.421 ± 5.10	0.3038 ± 0.215	0.7038 ± 0.255
Delta	$+0.0207$	$+0.0124$	-1.421	-0.0457	$+0.0286$