SEEING THROUGH THE GLARE: A MULTI-SOURCE BENCHMARK AND OCULAR-ADAPTIVE PIXEL MEANFLOW FOR EYEGLASS REFLECTION REMOVAL

Anonymous authors
Paper under double-blind review



Figure 1: Eyeglass reflection removal on synthetic samples. Dashed lines separate input (Reflection) and restored (Removal) regions, with complementary comparisons across both rows.

ABSTRACT

Eyeglass reflection removal is important across smartphone imaging, video conferencing, and other face-centric visual applications. The task is challenging because reflections range from mild photometric contamination to severe ocular occlusion, requiring selective correction and plausible reconstruction without altering identity or natural appearance. Existing datasets cover limited reflection conditions, constraining generalization to complex real-world scenes and systematic evaluation. We introduce **OcuBench**, a multi-source benchmark comprising 10,280 controllable synthetic pairs, 732 real-input pseudo-pairs, and 458 independent real-world test images, supporting both paired evaluation and assessment beyond generated supervision. We further propose **OcuFlow**, an ocular-adaptive pixel MeanFlow (pMF) framework for efficient, detail-preserving restoration. It combines geometry-adaptive representation with one-step pMF to focus reconstruction on reflection-obscured ocular regions, together with native-resolution frequency-preserving synthesis to retain reliable observed details. Experiments across diverse reflection conditions demonstrate that OcuFlow achieves consistent advantages in reflection removal quality, ocular fidelity, and efficiency. In a blind user study, it receives 67.32% of selections, $6.2\times$ the next-best share. Both the code and dataset will be released.

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.

1 INTRODUCTION

Reflections from windows or monitors on eyeglasses can obscure the eyes while leaving most of a portrait intact. Although spatially confined, this interference can noticeably change facial appearance in photographs and video communication. Its effects range from color shifts over visible structures to severe occlusion of the iris, eyelids, and eyelashes (Zou et al., 2024; Yi et al., 2025). A single image may therefore require photometric correction in some regions, content reconstruction in others, and faithful preservation elsewhere. Accommodating these different restoration needs is essential to removing reflections without unnecessarily changing the portrait.

Despite progress in eyeglass reflection removal (Sandhan & Young Choi, 2017; Zou et al., 2024; Yi et al., 2025), challenges remain in obtaining reliable supervision and evaluation data and balancing restoration accuracy with inference efficiency. At the data level, limitations are threefold: (1) head motion, blinking, and illumination changes compromise pixel-level correspondence in captured pairs (Zou et al., 2024); (2) synthetic pairs provide controlled correspondence and degradation but cannot alone establish performance under natural reflections; and (3) real-input pseudo-pairs retain natural reflections, but the true appearance of obscured ocular content cannot be verified from reflected inputs alone. Therefore, reliable assessment requires complementary sources and real-world evaluation independent of generated targets. At the model level, reconstructing obscured ocular structures requires expressive priors, while applications demand efficient inference. Diffusion models (Ho et al., 2020), DiT (Peebles & Xie, 2023), and SiT (Ma et al., 2024) demonstrate strong image synthesis capabilities. These models typically rely on iterative sampling with multiple network evaluations. MeanFlow (Geng et al., 2026) learns average velocities over time intervals for single-evaluation generation, while pMF (Lu et al., 2026) combines image prediction with velocity-space learning for one-step generation without a latent autoencoder.

Although pMF provides an efficient foundation for one-step generation, applying it to eyeglass reflection removal requires aligning generative computation with spatially heterogeneous restoration demands. Under a limited token budget, uniform tokenization allocates computation to largely unaffected regions while providing insufficient resolution for fine ocular structures. Within the lenses, similar reflection intensities may leave different amounts of ocular structure visible, indicating that reflection strength alone cannot determine restoration requirements. Therefore, effective reconstruction requires guidance from both local observations and ocular geometry. In addition, compact patch representations can discard native photographic details, whereas directly preserving observed high-frequency content risks retaining reflection artifacts.

To address these challenges, we introduce **OcuBench**, a multi-source benchmark for eyeglass reflection removal. OcuBench comprises 10,280 synthetic pairs, 732 real-input pseudo-pairs, and 458 independently collected real-world test images. Synthetic pairs vary in subject, imaging, and reflection factors. Real-input pseudo-pairs combine real reflected inputs with generated clean targets. The real-world set evaluates performance beyond generated supervision, with clean references for a subset. These sources support development and evaluation under controlled and real-world conditions. Based on OcuBench, we develop **OcuFlow**, an efficient ocular-adaptive pMF framework for eyeglass reflection removal. OcuFlow uses geometry-adaptive supports under a fixed token budget to concentrate resolution on ocular structures while preserving coarse context. The supports retain position and scale information to preserve spatial correspondence. A geometry-guided pMF integrates local reflected observations, ocular priors, and cross-scale features to restore pixel patches in one forward pass. A native-resolution frequency-preserving renderer projects predictions to the supplied input resolution. The renderer selectively blends predictions with observations across frequency bands to preserve photographic details and reduce unnecessary modifications.

Experiments show that OcuFlow achieves superior overall restoration quality on the held-out paired benchmark compared with existing methods. OcuFlow runs in 56 ms per image on an RTX 4090 GPU, providing 4–$66\times$ speedups over the compared methods under the same evaluation protocol. Independent real-world evaluation further shows competitive no-reference performance, and OcuFlow achieves the highest user preference in the user study. Beyond image-level restoration metrics, we evaluate downstream iris localization using a fixed detector without fine-tuning. OcuFlow achieves the best both-eyes localization accuracy among the compared method, demonstrating that its restoration benefits extend to downstream ocular analysis. These results demonstrate the potential of OcuFlow for efficient and detail-preserving portrait restoration.

Table 1: Comparison of representative datasets for eyeglass reflection removal. "Syn." denotes synthetic supervision, "Real" denotes real reflected inputs, and "Real Test" denotes an independent real-world evaluation set.

Dataset	Year	Syn.	Real	Real Test	Scale	Access
EwE (Sandhan & Young Choi, 2017)	2017	✓	–	–	540	Legacy (Link Unavailable)
Watanabe et al. (Watanabe & Hasegawa, 2021)	2021	–	✓	–	1,639	No Public Link
De-Glared (Sadik-E-Tawheed et al., 2023)	2023	✓	–	–	2,541	No Public Link
ReyeR (Zou et al., 2024)	2024	–	✓	–	13,610	No Public Link
ReyeR+ (Yi et al., 2025)	2025	–	✓	–	14,328	No Public Link
OcuBench	2026	✓	✓	✓	**10,280 / 732 / 458**	**Public / Verified**

2 Related Work

Eyeglass Reflection Removal and Benchmarks. EwE (Sandhan & Young Choi, 2017) overlays reflection patches on frontal portraits, while De-Glared (Sadik-E-Tawheed et al., 2023) constructs 2,541 synthetic pairs from generated portraits. ReyeR (Zou et al., 2024) provides filtered, color-corrected real pairs across varied conditions. Synthetic realism and real-pair alignment remain challenges. Methods span constrained optimization (Sandhan & Young Choi, 2017), supervised restoration (Watanabe & Hasegawa, 2021; Sadik-E-Tawheed et al., 2023), and generative facial reconstruction (Chandramouli & Gandikota, 2019). ER^{2}Net (Zou et al., 2024) fuses reflection removal and inpainting by reflection intensity, while DL2G (Yi et al., 2025) combines degradation correction, local structure-aware diffusion, and global refinement. We introduce a multi-source benchmark and efficient restoration for portrait editing and video communication.

Spatially Adaptive Visual Representations. Adaptive visual representations improve efficiency through token aggregation, conditional computation, and spatially varying resolution. TokenLearner (Ryoo et al., 2021) aggregates dense features into fewer adaptive tokens, while A-ViT (Yin et al., 2022) dynamically halts token processing at different network depths. Mixed-Resolution Tokenization (Ronen et al., 2023) constructs variable-sized patches through saliency-guided quadtree partitioning, whereas MSViT (Havtorn et al., 2023) uses learned gating to select patch scales with input-dependent token counts. For eyeglass restoration, spatial allocation must resolve fine ocular structures while maintaining correspondence with the original image. OcuFlow uses input-derived ocular geometry to optimize quadtree partitioning under a fixed token budget, concentrating resolution on critical eye structures. Each token retains its position and scale for geometry-guided restoration and native-resolution reconstruction.

Generative Restoration. Generative priors support restoration from incomplete observations. Diffusion (Ho et al., 2020) and Flow Matching (Lipman et al., 2023) typically require iterative sampling. DifFace (Yue & Loy, 2024) reconstructs faces with pretrained diffusion, while DiffBIR (Lin et al., 2024) combines degradation removal with diffusion-based detail regeneration. GFP-GAN (Wang et al., 2021) enables single-pass restoration using pretrained facial GAN priors. CodeFormer (Zhou et al., 2022) predicts discrete codes with a controllable fidelity–quality trade-off. Dereflection Any Image (Hu et al., 2026) adopts one-step diffusion for reflection removal. MeanFlow (Geng et al., 2026) learns average velocities over time intervals for single-evaluation generation. pMF (Lu et al., 2026) combines image prediction with velocity-space learning for one-step pixel-space generation. We adapt it to conditional restoration over ocular-adaptive patches. One-step restoration addresses *efficiency*, but not *what to restore and preserve*. WaveFace (Miao et al., 2024) uses diffusion for low-frequency wavelet components and a feed-forward network for high-frequency recovery. Our renderer blends predictions with native-resolution observations by frequency, selectively retaining reliable high-frequency details beyond compact patches.

3 OcuBench: A Multi-Source Benchmark

Tab. 1 compares representative eyeglass reflection datasets by construction, pairing, scale, and access. Controlled real capture provides physically valid paired observations, while synthetic composition enables scalable and controllable variation in reflection patterns. However, large-scale real paired facial data are difficult to collect and release due to acquisition constraints and privacy concerns, whereas synthetic data alone cannot fully capture naturally occurring reflections. OcuBench therefore combines controllable synthetic pairs, real-input pseudo-pairs, and independent real-world

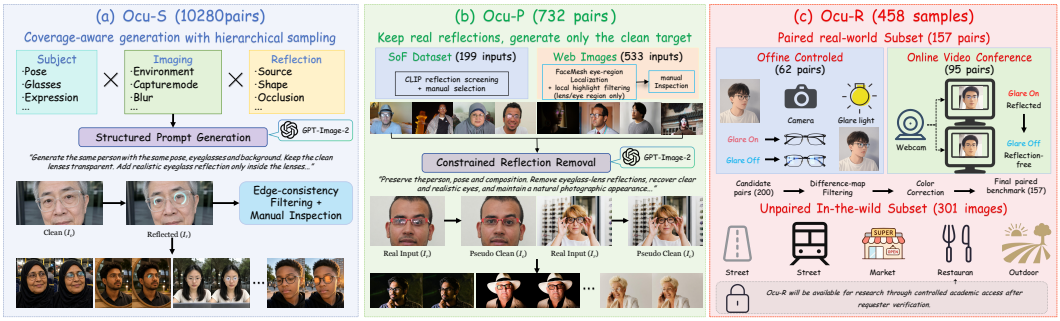

Figure 2: Construction of OcuBench. Ocu-S is generated from hierarchically sampled subject, imaging, and reflection factors. Ocu-P retains original reflected images and generates only clean targets. Ocu-R combines captured reference pairs and unpaired images from unconstrained scenes.

samples, leveraging their complementary strengths for controlled degradation analysis, realistic reflection coverage, and independent real-world evaluation. The following subsections describe data collection, generation, cleaning, and quality control, with further procedural details in § A.

3.1 Data Collection and Generation

Controllable Synthetic Pairs (Ocu-S). We adopt *structured hierarchical sampling* over subject factors $\mathbf{s}$ (pose, eyeglasses, expression, and gaze), imaging factors $\mathbf{i}$ (capture mode, scene mode, environment, background blur, and camera distance), and reflection factors $\mathbf{r}$ (source, shape, color, transparency, strength, and target eye occlusion), with their primary high-level values summarized in Tab. 5:

$$p(\mathbf{s}, \mathbf{i}, \mathbf{r}) = p(\mathbf{s})\, p(\mathbf{i} \mid \mathbf{s})\, p(\mathbf{r} \mid \mathbf{s}, \mathbf{i}). \tag{1}$$

We sample subject factors first, then imaging factors conditioned on the subject, followed by reflection factors conditioned on both. This hierarchy promotes compatible combinations spanning mild photometric contamination to severe ocular occlusion. Structured prompts encode sampled configurations for GPT-Image-2 (OpenAI, 2026) to generate reflected–clean candidates $(\widetilde{I}_r, \widetilde{I}_c)$, requesting consistent identity, pose, eyeglasses, background, and composition while varying only lens reflections. These prompts are provided in § A.1.

Real-input Pseudo-pairs (Ocu-P). We collect naturally reflected images from SoF (Afifi & Abdelhamed, 2019) and Web search, with the latter yielding 11,520 initial candidates. Following input screening, we preserve each selected reflected image and generate only its missing clean target as $\widetilde{I}_c = \mathcal{E}(I_r^{\text{real}}, \mathbf{p}_{\text{remove}})$. The operator $\mathcal{E}$ denotes GPT-Image-2 editing, and $\mathbf{p}_{\text{remove}}$ requests reflection removal while preserving identity, pose, eyeglasses, composition, and photographic appearance. The input I_r^{real} remains unchanged, retaining naturally observed reflection patterns.

Independent Real-world Samples (Ocu-R). We acquire 200 candidate reflected–clean pairs through controlled offline photography and online video conferencing, with the reflection-inducing light enabled and disabled. We collect 301 unpaired images from unconstrained environments, covering varied illumination, viewpoints, eyeglasses, and reflection conditions. All reflected inputs and clean references are acquired without image generation and reserved for real-world evaluation.

3.2 Data Cleaning and Quality Control

Ocu-P Input Screening. For the SoF dataset, CLIP-based reflection screening (Radford et al., 2021) followed by manual inspection yields 199 images. For the Web data, FaceMesh-guided local-highlight screening in estimated eye regions yields 1,740 candidates from 11,520 retrieved images, with manual inspection further reducing them to 533. The resulting 732 images are used to generate pseudo-clean targets.

Ocu-S and Ocu-P Validation. GPT-Image-2 generates synthetic pairs under shared scene constraints, whereas real-input editing must preserve the photograph while inferring obscured content. Both risk *pairwise structural drift* in facial or ocular geometry. An *edge-consistency filter* is applied to detect contour changes unrelated to reflections. Four reviewers assess realism, identity and ocular

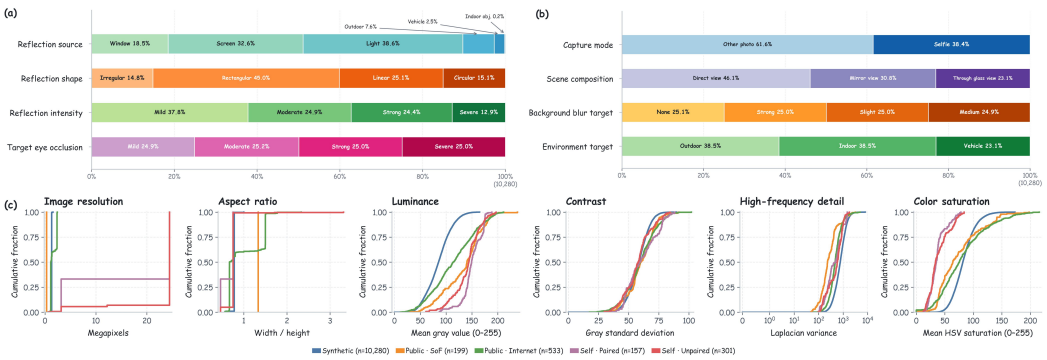

Figure 3: Dataset analysis of OcuBench. (a) Ocu-S reflection configurations. (b) Ocu-S imaging configurations. (c) Empirical cumulative distributions of image properties across OcuBench subsets.

consistency, and preservation of unaffected content, accepting pairs by consensus. We retain 10,280 samples in Ocu-S ($\mathcal{D}_{\text{syn}}$) and 732 in Ocu-P ($\mathcal{D}_{\text{pseudo}}$). Pseudo-clean supervision remains approximate because obscured ocular appearance cannot be verified from reflected inputs alone. Sampling configurations, prompt templates, and filtering criteria appear in § A.

Ocu-R Reference-pair Processing. Following ReyeR (Zou et al., 2024), we filter candidate captures with large spatial discrepancies using image differences and apply color correction to reduce capture-induced illumination shifts. Of the 200 candidate pairs, 157 are retained, comprising 62 offline and 95 video-conferencing pairs. Residual local misalignment remains, so the clean captures serve as visual references rather than pixel-accurate ground truth. Together with the 301 unpaired images, these form the 458-image Ocu-R.

3.3 Dataset Analysis

Reflection and Occlusion Distributions. Fig. 3(a) summarizes requested reflection configurations in Ocu-S. Screen and light reflections account for 32.6% and 38.6% of samples, respectively, while rectangular patterns are the most common shape at 45.0%. Target ocular occlusion is approximately balanced across mild, moderate, strong, and severe levels, each representing 24.9–25.2% of samples. Reflection strength follows a distinct distribution, with mild and severe cases accounting for 37.8% and 12.9%, respectively. These statistics describe sampled generation settings.

Imaging Conditions. Fig. 3(b) summarizes the imaging configurations. Selfies account for 38.4% of Ocu-S. Scene compositions include direct, mirror, and through-glass views across indoor, outdoor, and vehicle settings. The four target background-blur levels are approximately balanced, spanning sharp to defocused backgrounds. These configurations vary the photographic context around reflected eye regions.

Cross-source Image Statistics. Fig. 3(c) compares the data sources through empirical cumulative distributions of resolution, aspect ratio, luminance, contrast, Laplacian variance, and color saturation. Ocu-R reaches substantially higher resolutions than Ocu-S and Ocu-P, while Ocu-S generally exhibits lower luminance. The three sources also differ in aspect ratio, high-frequency detail, and saturation, but show more similar contrast distributions. These differences highlight source-dependent variation and motivate separate evaluation on independently captured images. **Dataset Splits, Evaluation, and Access.** We separately split Ocu-S ($\mathcal{D}_{\text{syn}}$) and Ocu-P ($\mathcal{D}_{\text{pseudo}}$) into 60% training and 40% test subsets, then merge the respective splits to form the final training and test sets. Held-out paired evaluation is performed on the merged test set. All 458 Ocu-R inputs are evaluated using no-reference metrics, with 30 randomly selected for a blind user study, while the 157 clean captures support qualitative analysis. We plan to release the synthetic and SoF components directly, release the Web-sourced component upon permission, and provide self-collected data through verified academic access (Tab. 1).

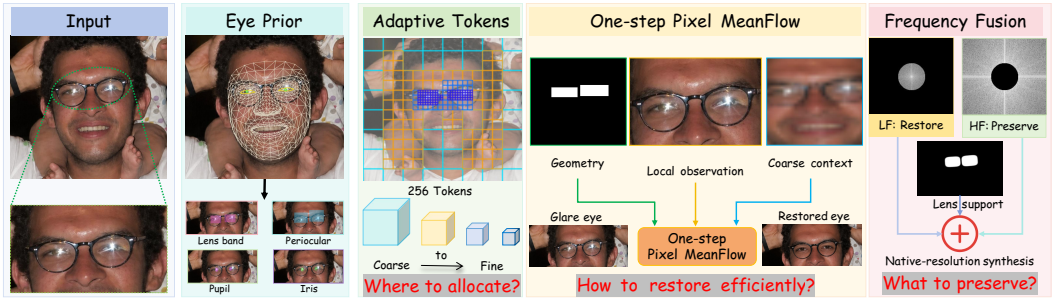

Figure 4: Overview of OcuFlow. Ocular priors guide allocation of 256 tokens. Geometry, reflected observations, and coarse context condition one-step patch restoration. Frequency-preserving synthesis blends predictions with native details within soft lens support.

4 METHOD

Given an arbitrary-resolution image with eyeglass reflections $I_g \in \mathbb{R}^{H\times W\times 3}$, OcuFlow estimates a reflection-free image I_{out} at the same spatial resolution as the input, we denote the clean training target by I_0. As illustrated in Fig. 4, the method coordinates three decisions: *where to allocate capacity, how to restore efficiently, and what to preserve*. Ocular geometry determines the spatial supports of a fixed number of pixel patches. A geometry-guided pMF restores the patches in one pass, while a frequency-preserving renderer fuses the predictions with native-resolution observations.

4.1 OCULAR-PRIOR-GUIDED ADAPTIVE REPRESENTATION

Ocular regions require finer detail than largely unaffected portrait regions. We therefore use finer patches around the eyes and coarser patches elsewhere, while keeping the sequence length fixed. Facial landmarks from I_g define geometry-derived ocular priors on a 256×256 logical domain to guide partitioning, while patches are sampled directly from the input-resolution image. We realize this variable-resolution partition as a quadtree forest initialized by an 8×8 grid of 32×32 root cells. Each leaf $q_i = (\mathbf{c}_i, s_i)$ has center $\mathbf{c}_i$, side length $s_i \in \{4, 8, 16, 32\}$, and support Ω_i. We define the ocular-importance mass of each leaf as

$$M_i = \int_{\Omega_i} w(\mathbf{x})\, d\mathbf{x}, \tag{2}$$

where $w(\mathbf{x}) \geq 0$ assigns larger importance to the periocular, eye, iris, eyelid, and lens regions. Under the fixed budget $N = 256$, we select

$$\mathcal{Q}^* = \arg \min_{\substack{\mathcal{Q}\in\mathfrak{T} \\ |\mathcal{Q}|=N}} \sum_{q_i\in\mathcal{Q}} \rho(s_i) M_i, \tag{3}$$

where $\mathfrak{T}$ denotes the set of admissible quadtree leaf partitions and $\rho(s_i)$ is a scale penalty that increases with leaf size. Thus, high-importance ocular regions are represented by smaller leaves, while largely unaffected regions retain coarser supports. The resulting fixed-budget adaptive representation is used by the one-step restoration model described next. Implementation details are in § B.1.

4.2 GEOMETRY-GUIDED ONE-STEP PIXEL MEANFLOW

The adaptive representation determines where the fixed token budget is concentrated; we next use these geometry-aware tokens to reconstruct clean ocular appearance from the reflected observation. We combine reflected observations with ocular geometry and multi-scale context into a conditioned representation $\mathbf{h}$, with leaf-wise ocular priors supplied separately to the backbone.

Observation-anchored Source. We construct $\tilde{\mathbf{g}}$ by adding small ocular-prior-weighted Gaussian perturbations to the reflected adaptive patches $\mathbf{g}$. This keeps the source close to the observation while emphasizing geometrically important ocular regions. The modulation rule is given in § B.2.

One-step Pixel MeanFlow. Following pMF (Lu et al., 2026), we define a linear path from the clean adaptive patches $\mathbf{x}_0$ to the observation-anchored source:

$$\mathbf{z}_t = (1-t)\mathbf{x}_0 + t\tilde{\mathbf{g}}, \qquad t \in [0, 1]. \tag{4}$$

The predictor $\widehat{\mathbf{x}}_\theta(\mathbf{z}_t, \Delta; \mathbf{h})$ estimates the clean endpoint conditioned on the current path state, interval length Δ, and geometry-aware context $\mathbf{h}$. At inference, setting $t = 1$ and $\Delta = 1$ yields the restored patches in one network evaluation:

$$\widehat{\mathbf{x}} = \widehat{\mathbf{x}}_{\bar{\theta}}(\widetilde{\mathbf{g}}, 1; \mathbf{h}), \tag{5}$$

where $\bar{\theta}$ denotes exponential-moving-average parameters. The source modulation, MeanFlow parameterization, conditioning architecture, and time sampling are detailed in § B.2.

4.3 Native-Resolution Frequency-Preserving Synthesis

Canonical patches provide a compact representation, but resampling can discard photographic details. Conversely, retaining input detail indiscriminately can preserve reflection artifacts. We therefore combine predicted and observed content through spatially varying frequency mixing.

Absolute and Residual Reconstruction. For the fixed partition $\mathcal{Q}^*$, let $\mathcal{R}$ denote linear projection from canonical patches to the input-resolution canvas. It resamples patches onto their native supports and combines them using normalized raised-cosine windows whose weights sum to one. We construct two candidates:

$$I_{\text{abs}} = \mathcal{R}(\widehat{\mathbf{x}}), \quad I_{\text{res}} = I_g + \mathcal{R}(\widehat{\mathbf{x}} - \mathbf{g}). \tag{6}$$

The absolute candidate reconstructs the image from predicted patches, whereas the residual candidate applies their estimated correction to the native input. By linearity, $I_{\text{res}} - I_{\text{abs}} = I_g - \mathcal{R}(\mathbf{g})$, which represents the input information not preserved during patch sampling and projection. It may contain both useful photographic details and reflection artifacts, which motivates selective fusion.

Frequency-dependent Fusion. Let $\mathcal{L}$ be a low-pass operator, $\mathcal{H}(I) = I - \mathcal{L}(I)$ the high-frequency component, and $A_{\text{lens}} \in [0, 1]^{H \times W}$ a soft lens-support map. Spatial maps α and β provide location-dependent control over the mixing of low- and high-frequency components:

$$L_{\text{out}} = \alpha \odot \mathcal{L}(I_{\text{abs}}) + (1 - \alpha) \odot \mathcal{L}(I_{\text{res}}), \tag{7}$$

$$H_{\text{out}} = (1 - A_{\text{lens}}) \odot \mathcal{H}(I_g) + A_{\text{lens}} \odot \Big[\beta \odot \mathcal{H}(I_g) + (1 - \beta) \odot \mathcal{H}(I_{\text{abs}})\Big], \tag{8}$$

$$I_{\text{render}} = L_{\text{out}} + H_{\text{out}}. \tag{9}$$

All maps are broadcast over RGB channels. α balances absolute and residual reconstructions, while larger β retains more observed high-frequency detail. A lightweight gate refines α and β through lens-masked residuals, with $\alpha = 0$ and $\beta = 1$ where $A_{\text{lens}} = 0$. Final compositing gives

$$I_{\text{out}} = (1 - A_{\text{lens}}) \odot I_g + A_{\text{lens}} \odot I_{\text{render}}. \tag{10}$$

Details of native projection, lens-support construction, and spatial gate construction are in § B.3.

5 Experiments

5.1 Metrics and Implementation Details

Paired evaluation uses Mask-PSNR/SSIM, Mixed-PSNR/SSIM, Eye-ROI LPIPS, and Eye-NME/Iris-NME. Eyeglass masks are initialized with Grounded-SAM-2 (Ren et al., 2024) and manually corrected, while NME is computed from MediaPipe landmarks (Lugaresi et al., 2019). Evaluation on Ocu-R comprises NIQE (Mittal et al., 2012), CLIP-IQA+ (Wang et al., 2023), and a blind user study. All methods use 1008×1296 inputs for paired inference and efficiency evaluation on the same RTX 4090. Runtime is measured as the steady-state average wall-clock processing time per image under sustained inference, with each method evaluated using the same throughput-oriented protocol. Paired metrics are computed at the original paired resolutions, while Ocu-R is evaluated at its original resolution. Metric definitions, baseline configurations, training objectives, and efficiency measurement protocols are provided in § C.1.

5.2 Evaluation

Quantitative Results. In Tab. 2, OcuFlow shows clear advantages in restoration accuracy and efficiency. It ranks first on six of the seven restoration and ocular geometry metrics and achieves

Table 2: Quantitative comparison on the held-out paired OcuBench test set. Best and second-best results are **bold** and underlined, respectively.

Method	Restoration Quality					Ocular Geometry		Efficiency	
	Mask PSNR ↑	Mixed PSNR ↑	Mask SSIM ↑	Mixed SSIM ↑	Eye-ROI LPIPS ↓	Eye-NME (%) ↓	Iris-NME (%) ↓	Memory (GB) ↓	Runtime (ms/image) ↓
Stable Diffusion 1.5 (Rombach et al., 2022)	19.1926	21.2009	0.5446	0.6375	0.1187	1.7369	2.1161	10.31	2709
DSRNet (Hu & Guo, 2023)	20.1502	21.8836	0.6256	0.6845	0.1831	2.0755	2.4358	6.55	3430
ER^{2}Net (Zou et al., 2024)	19.9652	22.1108	0.5761	0.6742	0.1520	1.7455	2.1130	18.20	240
Stable Diffusion 3.5 (Esser et al., 2024)	18.4196	20.6247	0.5249	0.6346	0.2029	2.0579	2.5207	20.06	1454
DSIT (Hu et al., 2024)	21.5400	23.2821	0.6478	**0.7078**	0.1461	1.5363	1.6998	2.96	3094
RDNet (Zhao et al., 2025)	21.0688	22.8433	0.6300	0.6931	0.1161	1.5624	1.7536	1.49	2707
DL2G (Yi et al., 2025)	19.8145	21.7636	0.5747	0.6599	0.1788	1.7772	2.1993	1.62	3747
ReflexSplit (Lee et al., 2026)	20.7927	22.4125	0.6275	0.6838	0.1350	1.7792	1.9592	13.15	1840
Ours	**22.1054**	**23.4710**	**0.6633**	0.7074	**0.1035**	**1.4030**	**1.5863**	**1.04**	**56**

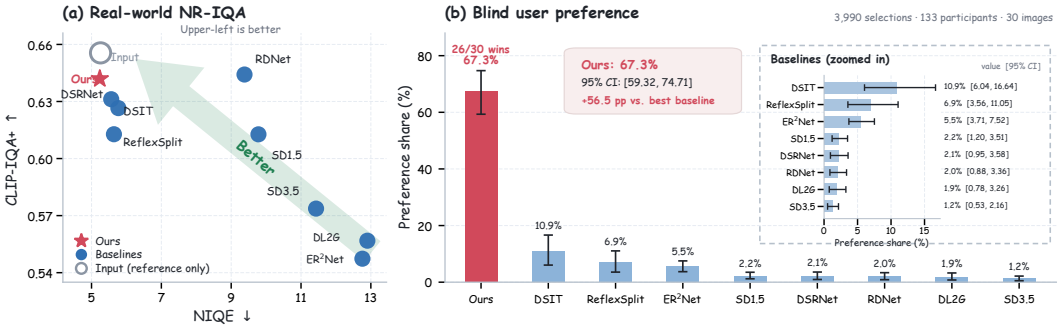

Figure 5: Real-world evaluation on OcuBench. **(a)** NR-IQA results on 458 real-world images, where the arrow indicates the direction of better performance. **(b)** Blind user preference comparison; error bars denote 95% cluster-bootstrap confidence intervals.

a $4.3\times$ speedup in steady-state processing over ER^{2}Net, the second-fastest method under the same evaluation protocol. On Ocu-R (Fig. 5(a)), OcuFlow achieves the lowest NIQE (5.2369) and second-highest CLIP-IQA+ (0.6419) among restoration methods. Since the reflected input itself obtains a higher CLIP-IQA+ score (0.6556), we treat no-reference metrics as auxiliary indicators. In the blind user study (Fig. 5(b)), OcuFlow receives 67.32% of valid selections and ranks first on 26 of the 30 images. The 95% cluster-bootstrap confidence interval for its preference share is $[59.32\%, 74.71\%]$. Beyond image-quality evaluation, downstream iris localization further shows that reflection removal improves the reliability of ocular analysis, with OcuFlow achieving strong bilateral iris-localization performance. Additional analyses and downstream task results are provided in § C.2 and C.3.

Qualitative Results. Fig. 6 compares results under localized glare and broad blue reflections. RDNet and ReflexSplit retain lens haze in the first example. In the second example, DSIT and ReflexSplit produce softer, lower-contrast eyes, while RDNet retains a blue-purple cast. DL2G introduces irregular details around the rims and eyelids. OcuFlow suppresses dominant reflections in both cases, recovering clearer iris and eyelid contours closer to the references while preserving reliable facial details. Additional results appear in § C.4.

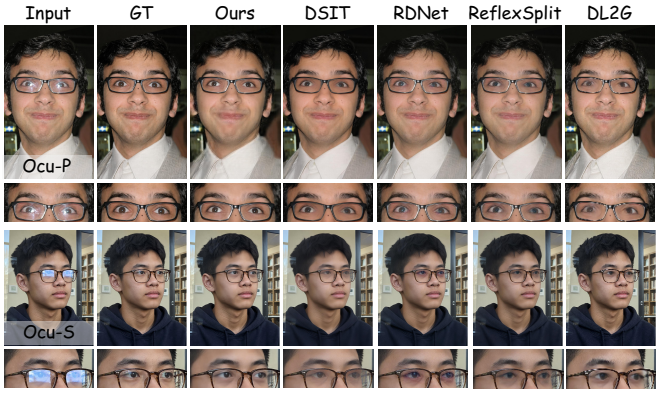

Figure 6: Qualitative comparison. Close-ups highlight reflection removal, ocular reconstruction, and detail preservation.

5.3 Ablation Study

We evaluate the three components of OcuFlow in Tab. 3 and the effect of training data composition in Tab. 4. See § C.5 for more results.

Table 3: Ablation studies of OcuFlow. Out-PSNR and Out-HF-PSNR measure RGB and high-frequency agreement outside the evaluation mask. Detailed definitions are provided in § C.1.3.

Setting	Mask PSNR ↑	Mask SSIM ↑	Eye-ROI LPIPS ↓	Eye-NME (%) ↓	Iris-NME (%) ↓	Out PSNR ↑	Out-HF PSNR ↑	Time (ms) ↓
(a) Where to allocate capacity								
Uniform Grid	20.8082	0.6171	0.1811	2.0670	2.5891	26.4132	31.2027	**49**
w/o Geometry Embedding	21.9823	0.6586	0.1129	1.4367	1.6297	26.4076	31.2065	+8
Ocular-Adaptive Grid (Ours)	**22.1054**	**0.6633**	**0.1035**	**1.4030**	**1.5863**	**26.4147**	**31.2118**	+7
(b) How to restore efficiently								
w/o Geometry-Aware Source	21.9896	0.6592	0.1127	1.4449	1.6336	26.4040	31.2059	**48**
w/o Coarse-to-Fine	21.9891	0.6591	0.1133	1.4384	1.6250	26.4079	31.2076	+10
2-NFE	21.4412	0.6329	**0.0826**	1.4695	1.7152	26.3898	31.1961	+12
1-NFE (Ours)	**22.1054**	**0.6633**	0.1035	**1.4030**	**1.5863**	**26.4147**	**31.2118**	+8
(c) What to preserve								
Direct PMF	21.2437	0.6287	0.1107	1.4352	1.5988	25.9394	30.8842	59
Residual Rendering	21.7838	0.6461	0.1076	1.4389	1.6252	26.2933	31.0292	<3
Frequency Renderer (Ours)	**22.1054**	**0.6633**	**0.1035**	**1.4030**	**1.5863**	**26.4147**	**31.2118**	<3

Where to Allocate Capacity. Tab. 3(a) shows that ocular-adaptive allocation improves masked reconstruction, Eye-ROI LPIPS, and both ocular NME measures over a uniform grid. Out-PSNR and Out-HF-PSNR remain nearly unchanged, indicating that the gains primarily concern ocular restoration. Removing geometry embedding reduces reconstruction and ocular fidelity, supporting explicit encoding of token position and scale. The uniform grid remains faster, so adaptive allocation trades additional processing time for improved restoration quality.

How to Restore Efficiently. Removing either the geometry-aware source or coarse-to-fine conditioning worsens masked reconstruction and all eye-specific metrics in Tab. 3(b). These results support combining local observations with geometric and contextual guidance. A single network function evaluation (1-NFE) achieves better masked reconstruction and ocular geometry than two evaluations (2-NFE), while requiring less runtime. Although 2-NFE achieves lower Eye-ROI LPIPS, its additional computation does not consistently improve restoration. The 1-NFE setting therefore provides a favorable balance between reconstruction accuracy, ocular fidelity, and efficiency.

What to Preserve. Residual rendering improves masked reconstruction and outside-mask agreement over Direct PMF but slightly worsens ocular geometry. The frequency-preserving renderer performs best across this block, with higher Out-PSNR and Out-HF-PSNR indicating closer outside-mask agreement, including in high-frequency content. These results support selective frequency fusion for balancing restoration and preservation.

Table 4: Ablation studies of data scale and composition.

Ocu-S_{train} used (%)	Ocu-P_{train}	Paired Benchmark (Ocu-S_{test} + Ocu-P_{test}) Mask PSNR ↑	Mixed PSNR ↑	Mask SSIM ↑	Mixed SSIM ↑	Eye-NME (%) ↓	Iris-NME (%) ↓	Real-World (Ocu-R) CLIP-IQA+ ↑
10	–	21.1233	22.7431	0.6375	0.6944	1.6321	1.9898	0.6410
30	–	21.6538	23.1349	0.6546	0.7030	1.4955	1.7323	0.6369
70	–	21.9224	23.3336	0.6597	0.7055	1.4264	1.6076	0.6313
100	–	21.9834	23.3767	0.6617	0.7065	1.4119	**1.5842**	0.6356
100	✓	**22.1054**	**23.4710**	**0.6633**	**0.7074**	**1.4030**	1.5863	**0.6419**

Effect of Multi-source Supervision. Tab. 4 shows that increasing the synthetic training fraction consistently improves paired reconstruction and ocular geometry. However, real-world CLIP-IQA+ does not improve monotonically with synthetic data scale. Adding real-input pseudo-pairs (*i.e.*, Ocu-P_{train}) further improves the paired PSNR/SSIM metrics and Eye-NME, and yields the highest CLIP-IQA+ on the independent real-world set. Iris-NME increases slightly relative to the 100% Ocu-S-only setting. The combined supervision thus provides broader benefits across paired and real-world evaluation, although these gains do not extend to every metric.

6 Conclusion

We present OcuBench and OcuFlow, a multi-source benchmark and an ocular-adaptive pMF framework for eyeglass reflection removal. OcuBench combines synthetic pairs, real-input pseudo-pairs, and independent real-world evaluation beyond generated targets. OcuFlow aligns a fixed token budget with ocular geometry and couples one-step restoration with frequency-preserving synthesis to reconstruct obscured eyes while retaining reliable native details. Experiments demonstrate improved reflection removal and ocular-structure reconstruction, leading to more accurate downstream iris localization, while achieving favorable human perceptual preferences on independent real-world images at substantially lower inference cost. Our findings suggest that adapting representation to local needs and selectively preserving observations support effective one-step portrait restoration. Future work could extend this approach to temporally consistent video restoration.

AI USE STATEMENT

In this work, we used GPT-Image-2 for dataset construction. Specifically, it was used to generate controllable synthetic clean–reflection image pairs from structured prompts and to generate the missing clean targets for real-input pseudo-pairs while keeping the original reflected images unchanged. All generated samples were subject to automated consistency filtering and manual inspection, and samples exhibiting unintended identity changes, geometric distortions, or other non-reflection modifications were excluded.

AI tools were also used solely to assist with language polishing, expression refinement, and textual organization of the manuscript, and were not used to generate research content, experimental results, or scientific conclusions. All AI-generated data used in this work were reviewed and validated by the authors before being included in the dataset or experiments. The authors take full responsibility for the design and execution of the study, the validation and selection of AI-generated data, the reported experimental results, and the final content of this work.

ETHICS STATEMENT

This work involves facial-image data and a human-participant user study. We considered informed consent, privacy, data provenance and licensing, controlled data sharing, and potential misuse throughout data collection, dataset preparation, and release.

Our dataset contains synthetic images, images from existing datasets and public online sources, and facial images collected by our research team. Synthetic images do not correspond to real participants. The SoF component is used according to the conditions of the original dataset. For the Pexels-sourced images, we have requested explicit permission regarding redistribution of processed derivatives. Until such permission is obtained, neither the original Pexels images nor their processed derivatives will be publicly released. If permission is granted, the processed derivatives will be released only under the specified conditions.

For our self-collected facial images, all participants provided written informed consent covering academic research, publication, model training and evaluation, and controlled sharing with external researchers. These images will therefore be provided only through controlled academic access after verification of the requester's identity, affiliation, and research purpose. Names, contact information, and other unnecessary personal identifiers are not distributed. The data are not intended for identity recognition, surveillance, commercial exploitation, or other unrelated purposes.

Participation in our perceptual user study was voluntary, informed consent was obtained, and responses were collected without direct personal identifiers and used only for aggregate research analysis. As facial images may remain identifiable after metadata removal, we use source-specific release and access policies to mitigate potential privacy and misuse risks.

REPRODUCIBILITY STATEMENT

We provide the information necessary to facilitate reproducibility throughout the main paper and appendix. Dataset construction, composition, quality control, and data splits are described in § 3 and 3.3, with additional details on sampling, prompt construction, retrieval, and filtering provided in the appendix. The OcuFlow architecture, adaptive representation, one-step pMF formulation, frequency-preserving synthesis, and training objectives are described in § 4, with implementation and optimization details provided in the appendix. Evaluation metrics, protocols, input resolutions, and computational setup are specified in § 5.1, while detailed baseline configurations, training settings, efficiency measurement protocols, and source-specific evaluations are provided in the appendix.

REFERENCES

Mahmoud Afifi and Abdelrahman Abdelhamed. Afif4: Deep gender classification based on adaboost-based fusion of isolated facial features and foggy faces. *Journal of Visual Communication and Image Representation*, 62:77–86, 2019.

Paramanand Chandramouli and Kanchana Vaishnavi Gandikota. Blind single image reflection suppression for face images using deep generative priors. In *ICCV Workshops*, pp. 3315–3323, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *ICML*, 2024.

Zhengyang Geng, Mingyang Deng, Xingjian Bai, Zico Kolter, and Kaiming He. Mean flows for one-step generative modeling. In *NeurIPS*, pp. 75460–75482, 2026.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. Geneval: An object-focused framework for evaluating text-to-image alignment. In *NeurIPS*, pp. 52132–52152, 2023.

Jakob Drachmann Havtorn, Amélie Royer, Tijmen Blankevoort, and Babak Ehteshami Bejnordi. Msvit: Dynamic mixed-scale tokenization for vision transformers. In *ICCV Workshops*, pp. 838–848, 2023.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, pp. 6840–6851, 2020.

Jichen Hu, Chen Yang, Zanwei Zhou, Jiemin Fang, Qi Tian, and Wei Shen. Dereflection any image with diffusion priors and diversified data. In *AAAI*, pp. 4860–4868, 2026.

Qiming Hu and Xiaojie Guo. Single image reflection separation via component synergy. In *ICCV*, pp. 13092–13101, 2023.

Qiming Hu, Hainuo Wang, and Xiaojie Guo. Single image reflection separation via dual-stream interactive transformers. In *NeurIPS*, pp. 55228–55248, 2024.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2i-compbench: A comprehensive benchmark for open-world compositional text-to-image generation. In *NeurIPS*, pp. 78723–78747, 2023.

Abhinav Kumar, Tim K Marks, Wenxuan Mou, Ye Wang, Michael Jones, Anoop Cherian, Toshiaki Koike-Akino, Xiaoming Liu, and Chen Feng. Luvli face alignment: Estimating landmarks' location, uncertainty, and visibility likelihood. In *CVPR*, pp. 8233–8243, 2020.

Chia-Ming Lee, Yu-Fan Lin, Jin-Hui Jiang, Yu-Jou Hsiao, Chih-Chung Hsu, and Yu-Lun Liu. Reflexsplit: Single image reflection separation via layer fusion-separation. In *CVPR*, pp. 1300–1309, 2026.

Xinqi Lin, Jingwen He, Ziyan Chen, Zhaoyang Lyu, Bo Dai, Fanghua Yu, Yu Qiao, Wanli Ouyang, and Chao Dong. Diffbir: Toward blind image restoration with generative diffusion prior. In *ECCV*, pp. 430–448, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *ICLR*, 2023.

Yiyang Lu, Susie Lu, Qiao Sun, Hanhong Zhao, Zhicheng Jiang, Xianbang Wang, Tianhong Li, Zhengyang Geng, and Kaiming He. One-step latent-free image generation with pixel mean flows. In *ICML*, 2026.

Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, et al. Mediapipe: A framework for building perception pipelines. *arXiv preprint arXiv:1906.08172*, 2019.

Nanye Ma, Mark Goldstein, Michael S Albergo, Nicholas M Boffi, Eric Vanden-Eijnden, and Saining Xie. Sit: Exploring flow and diffusion-based generative models with scalable interpolant transformers. In *ECCV*, pp. 23–40, 2024.

Yunqi Miao, Jiankang Deng, and Jungong Han. Waveface: Authentic face restoration with efficient frequency recovery. In *CVPR*, pp. 6583–6592, 2024.

Anish Mittal, Rajiv Soundararajan, and Alan C Bovik. Making a "completely blind" image quality analyzer. *IEEE SPL*, 20(3):209–212, 2012.

OpenAI. GPT-Image-2. https://developers.openai.com/api/docs/models/gpt-image-2, 2026. OpenAI API documentation, released April 21, 2026; accessed September 20, 2026.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *ICCV*, pp. 4172–4182, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, pp. 8748–8763, 2021.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *ICLR*, pp. 28085–28128, 2025.

Tianhe Ren, Shilong Liu, Ailing Zeng, Jing Lin, Kunchang Li, He Cao, Jiayu Chen, Xinyu Huang, Yukang Chen, Feng Yan, et al. Grounded sam 2: Ground and track anything in videos, 2024.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pp. 10674–10685, 2022.

Tomer Ronen, Omer Levy, and Avram Golbert. Vision transformers with mixed-resolution tokenization. In *CVPR Workshops*, pp. 4613–4622, 2023.

Michael Ryoo, AJ Piergiovanni, Anurag Arnab, Mostafa Dehghani, and Anelia Angelova. Tokenlearner: Adaptive space-time tokenization for videos. In *NeurIPS*, pp. 12786–12797, 2021.

Md Sadik-E-Tawheed, Tasnia Ahmed, Masud Bhuiyan, Abir Hossain, and Swakkhar Shatabda. Deglared: Eyeglasses glare and reflection removal using deep neural networks. In *International Conference on Computer and Information Technology*, pp. 1–6, 2023.

Tushar Sandhan and Jin Young Choi. Anti-glare: Tightly constrained optimization for eyeglass reflection removal. In *CVPR*, pp. 1241–1250, 2017.

Jianyi Wang, Kelvin CK Chan, and Chen Change Loy. Exploring clip for assessing the look and feel of images. In *AAAI*, pp. 2555–2563, 2023.

Xintao Wang, Yu Li, Honglun Zhang, and Ying Shan. Towards real-world blind face restoration with generative facial prior. In *CVPR*, pp. 9164–9174, 2021.

Zhou Wang and Alan C Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009.

Sota Watanabe and Makoto Hasegawa. Reflection removal on eyeglasses using deep learning. In *International Technical Conference on Circuits/Systems, Computers and Communications*, pp. 1–4, 2021.

Zhilv Yi, Xiao Lu, Hong Ding, Jingbo Hu, Zhi Jiang, and Chunxia Xiao. Dl2g: Degradation-guided local-to-global restoration for eyeglass reflection removal. In *CVPR*, pp. 16061–16070, 2025.

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. A-vit: Adaptive tokens for efficient vision transformer. In *CVPR*, pp. 10799–10808, 2022.

Zongsheng Yue and Chen Change Loy. Difface: Blind face restoration with diffused error contraction. *IEEE TPAMI*, 46(12):9991–10004, 2024.

Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *CVPR*, pp. 586–595, 2018.

Hao Zhao, Mingjia Li, Qiming Hu, and Xiaojie Guo. Reversible decoupling network for single image reflection removal. In *CVPR*, pp. 26430–26439, 2025.

Shangchen Zhou, Kelvin Chan, Chongyi Li, and Chen Change Loy. Towards robust blind face restoration with codebook lookup transformer. In *NeurIPS*, pp. 30599–30611, 2022.

Wentao Zou, Xiao Lu, Zhilv Yi, Ling Zhang, Gang Fu, Ping Li, and Chunxia Xiao. Eyeglass reflection removal with joint learning of reflection elimination and content inpainting. *IEEE TCSVT*, 34(10):10266–10280, 2024.

APPENDIX CONTENTS

A Dataset Construction Details **14**
A.1 Ocu-S Construction Details . . . 14
A.1.1 Hierarchical Factor Sampling . . . 14
A.1.2 Pair Generation . . . 14
A.1.3 Pair Filtering . . . 16
A.2 Ocu-P Construction Details . . . 18
A.2.1 Real Reflected-input Collection . . . 18
A.2.2 Clean-target Generation . . . 19
A.2.3 Pseudo-pair Filtering . . . 19
A.3 Ocu-R Construction Details . . . 19
A.3.1 Reference-pair Capture and Processing . . . 19
A.3.2 In-the-wild Test Images . . . 20
A.3.3 Real-world Test-set Characteristics . . . 20

B Method Details **21**
B.1 Ocular-Prior-Guided Adaptive Representation . . . 21
B.1.1 Ocular Prior and Structural Support Construction . . . 21
B.1.2 Exact-Budget Quadtree Allocation . . . 23
B.2 Geometry-Guided One-Step pMF . . . 26
B.2.1 Geometry and Cross-Scale Conditioning . . . 26
B.2.2 Observation-Anchored Source . . . 26
B.2.3 pMF Parameterization . . . 26
B.2.4 MeanFlow Time Sampling . . . 27
B.3 Native-Resolution Frequency-Preserving Synthesis . . . 27
B.3.1 Native Projection and Lens Support . . . 27
B.3.2 Spatial Gate Construction . . . 28

C Experimental Details **29**
C.1 Implementation and Evaluation . . . 29
C.1.1 Training Setup and Objectives . . . 29
C.1.2 Inference Details . . . 31
C.1.3 Evaluation Protocols . . . 31
C.1.4 Baseline Configurations . . . 33
C.1.5 Efficiency Protocol . . . 34
C.2 Additional Quantitative Analysis . . . 34
C.3 Downstream Iris Localization . . . 36
C.4 Additional Qualitative Results . . . 36
C.4.1 Qualitative comparison. . . . 38
C.4.2 Ocular-prior-guided selective restoration. . . . 39
C.4.3 Failure cases. . . . 40
C.5 Additional Ablation and Inference Analysis . . . 40
C.5.1 Training and Conditioning Ablations . . . 40
C.5.2 Frequency-Renderer Gate Ablation . . . 41
C.5.3 Multi-step Inference . . . 41
C.5.4 Inference-time Breakdown and Pipeline Parallelism . . . 42

Table 5: Primary high-level factor space used for Ocu-S generation.

Group	Factor	Values / Range
Subject $\mathbf{s}$	Pose	frontal; yaw $\pm 15^\circ, \pm 30^\circ$; pitch/roll $\pm 10^\circ$
	Eyeglasses	rectangular; round; square; oval;semi-rimless; browline; acetate; frameless
	Expression	neutral; slight smile; calm; serious;natural; relaxed; focused; slightly tired
	Gaze	camera/center; left; right; up; down;away; phone-directed; screen-directed
Imaging $\mathbf{i}$	Capture mode	selfie; photographed by another person
	Scene mode	direct; through-glass; mirror view
	Environment	indoor; vehicle; outdoor
	Background blur	none; slight; medium; strong
	Camera distance	0.35–2.5 m
Reflection $\mathbf{r}$	Source	window; screen; light; outdoor; vehicle
	Shape	irregular; rectangular; linear; circular
	Color	neutral; warm; cool; mixed
	Transparency	15%–95%
	Strength	mild; moderate; strong; severe
	Target eye occlusion	mild; moderate; strong; severe

A Dataset Construction Details

This section provides additional details of the multi-source eyeglass-reflection dataset, including the controllable synthetic pairs (Ocu-S), real-input pseudo-pairs (Ocu-P), and independent real-world test set (Ocu-R).

A.1 Ocu-S Construction Details

A.1.1 Hierarchical Factor Sampling

We instantiate the hierarchical sampling in Eq. (1) using the primary factor space summarized in Tab. 5. Subject factors are sampled first, after which imaging conditions are selected through compatibility-aware conditional rules. In particular, capture mode conditions the scene configuration, environment, background blur, and camera distance, while the reflection source is further conditioned on the sampled imaging setting. Reflection shape, color, strength, transparency, and target ocular-occlusion level are then sampled according to the selected source and imaging configuration. Eyeglass styles are grouped into eight high-level frame families, with color, material, size, and related appearance variants instantiated within each family during prompt generation. Additional prompt attributes, including time of day, lighting, background complexity, camera style, reflection location and edge character, eye-area occlusion percentage, and eye-visibility targets, are instantiated from the sampled factors using deterministic mappings and factor-conditioned sampling rules.

The target ocular-occlusion levels are explicitly scheduled to be approximately balanced across mild, moderate, strong, and severe cases, accounting for about 25% of the samples at each level. In contrast, the sampled reflection-strength levels are intentionally non-uniform. We distinguish this prompt-level strength control from the realized visual intensity in the generated image, which may additionally depend on reflection source, transparency, spatial extent, and other scene-dependent factors, consistent with the imperfect attribute binding and compositional control observed in text-to-image generation (Huang et al., 2023; Ghosh et al., 2023).

A.1.2 Pair Generation

For each sampled factor tuple, we instantiate two structured prompts: a clean-image prompt and a reflection-edit prompt. All prompts are generated offline in advance using a fixed random seed of 2026 and stored before image synthesis. This decouples factor sampling from image generation and allows interrupted runs to resume from the stored prompt list without resampling or duplicating previously generated factor combinations.

GPT-Image-2 first generates a clean candidate $\widetilde{I}_c$ from the clean-image prompt. The resulting image is then used as *Image A* and edited with the corresponding reflection prompt to produce the matched reflection candidate $\widetilde{I}_r$. The two prompt templates are shown below, where green bracketed terms denote values instantiated from the sampled subject, imaging, and reflection factors.

Clean-image prompt template

```
Generate a realistic high-detail clean portrait candidate for a paired
eyeglass-reflection dataset.
SUBJECT / IDENTITY
Subject: <subject>
Eyeglasses: <eyeglass type>
Expression: <expression>
Face pose: <face pose>
Gaze: <gaze direction>
IMAGING CONDITION
Capture mode: <capture mode>
Scene mode: <scene mode>
Environment: <environment> (<environment category>)
Time of day: <time of day>
Camera-to-subject distance: <distance>
Lighting: <lighting type>
Background complexity: <background complexity>
Background blur: <background blur>
Camera style/composition: <camera style>
<scene-mode instruction>
PAIRING CONSTRAINTS
This image will be used as the clean member of a matched clean--reflection
pair. The eyeglass lenses must be transparent and optically clean, with no
lens glare or lens reflection. Both eyes should be visible as allowed by
the pose and gaze. Do not add text, borders, watermarks, collages, or extra
panels.
Render at <image resolution> and keep a realistic camera-photo appearance with
high detail.
```

Reflection-edit prompt template

```
Edit Image A (the clean image) into the reflected member of the SAME matched
pair.
STRICT PAIRWISE CONSISTENCY
Preserve the same identity, face geometry, ocular geometry, expression, face
pose, gaze direction, hairstyle, skin texture, clothing, eyeglass frame
geometry, lens geometry, environment, background, camera viewpoint, framing,
crop, depth of field, and overall composition.
Do not change any facial or ocular contours. The intended change is ONLY the
addition of realistic optical reflection inside the eyeglass lenses.
REFLECTION CONTROL
Source: <reflection source>
Shape: <reflection shape>
Color: <reflection color>
Strength: <reflection strength>
Approximate opacity / reflection visibility: <reflection transparency>
Spatial location: <reflection position>
Edge character: <reflection edge>
Target ocular-occlusion level: <occlusion level>
Target eye-area occlusion: approximately <eye occlusion percentage>%
Eye visibility target: <eye visibility>
PHYSICAL / LOCALITY CONSTRAINTS
The reflection must look optically plausible and conform to the lens
perspective and curvature. Keep the added reflection strictly inside the
transparent lens regions. Do not paint reflection onto the frame, skin, eyes
```

```
themselves, hair, clothing, or background.  Do not introduce new objects or
alter the environment to explain the reflection.  Do not change scene lighting
outside the lens reflections.  Do not add text, borders, watermarks, collages,
or extra panels.
Return only the edited reflected portrait, aligned to Image A.
```

Each prompt produces a single 1008×1296 image using GPT-Image-2. If a generation call fails, the request is retried after a 20-second delay, with at most five total attempts per prompt. If all five attempts fail, the corresponding sample is discarded and generation proceeds to the next prompt. In practice, no prompt reached the maximum number of attempts, and all scheduled samples were successfully generated. Generating the complete synthetic set required approximately 17 days.

A.1.3 Pair Filtering

We apply a two-stage quality-control procedure consisting of automatic edge-consistency filtering followed by manual inspection.

Edge-consistency Filtering To tolerate small generation-induced pixel shifts while screening for pairwise structural drift, we first register the reflected candidate $\widetilde{I}_r$ to its clean counterpart $\widetilde{I}_c$ using a similarity transformation restricted to translation, rotation, and isotropic scaling. We allow at most 8 pixels of translation, 1.5° of rotation, and an isotropic scale within $[0.985, 1.015]$. Local non-rigid deformations are assessed during subsequent manual inspection.

After alignment, structural consistency is evaluated over three regions: facial structure outside the eyeglass area, visible periocular structure, and the approximate eyeglass-frame/bridge region. Pixels dominated by the intended lens-reflection change are excluded from the corresponding regional comparisons. For each available region $k \in \{\mathrm{face}, \mathrm{eye}, \mathrm{frame}\}$, we compute a symmetric distance-transform-based edge distance

$$d_k = \frac{1}{2}\left[\frac{1}{|E_c^k|}\sum_{p\in E_c^k} D_{E_r^k}(p) + \frac{1}{|E_r^k|}\sum_{q\in E_r^k} D_{E_c^k}(q)\right], \tag{11}$$

where E_c^k and E_r^k denote the valid clean and aligned-reflection edge sets within region k, respectively, and $D_E(\cdot)$ denotes the Euclidean distance to the nearest edge in E. A regional distance is treated as unavailable when insufficient valid edge support remains after excluding the intended reflection-affected area.

A pair is rejected automatically if any available regional edge distance exceeds 3 pixels. When the periocular or frame metric is unavailable because too little valid structure remains after masking strong reflections, the pair is forwarded to subsequent manual inspection. Pairs passing the automatic edge-consistency filter are manually inspected for local non-rigid deformation and other structural changes.

Fig. 7 illustrates representative outcomes of the filtering procedure. The accepted example in Fig. 7(a) exhibits closely overlapping facial, ocular, frame, and background structures, indicating that the reflected image differs primarily by the intended lens reflection. In contrast, Fig. 7(b) shows a predominantly periocular failure, where the eye-region edges become spatially displaced despite relatively stable facial and frame structure. Fig. 7(c) illustrates eyeglass-frame deformation, characterized by pronounced disagreement along the frame and lens boundaries, while Fig. 7(d) exhibits widespread edge discrepancies extending beyond the eyeglass region, indicating an unintended edit to the face, subject placement, or background.

In the structural-edge overlays, yellow indicates edges shared by the clean and aligned reflected candidates, whereas green and red indicate edges appearing only in the clean and reflected images, respectively. Strong yellow overlap indicates structural consistency, while separated red–green responses indicate spatial drift or deformation. Cyan boxes mark the primary periocular/eyeglass inspection region, with enlarged local crops shown below. The displayed distances are measured in pixels, with lower values indicating stronger structural agreement. For visualization, we additionally report a background edge distance outside the expanded face region as a diagnostic measure; it is not used as a hard acceptance criterion. An unavailable eye-region value indicates that insufficient valid periocular edge support remains after excluding the intended reflection-affected area.

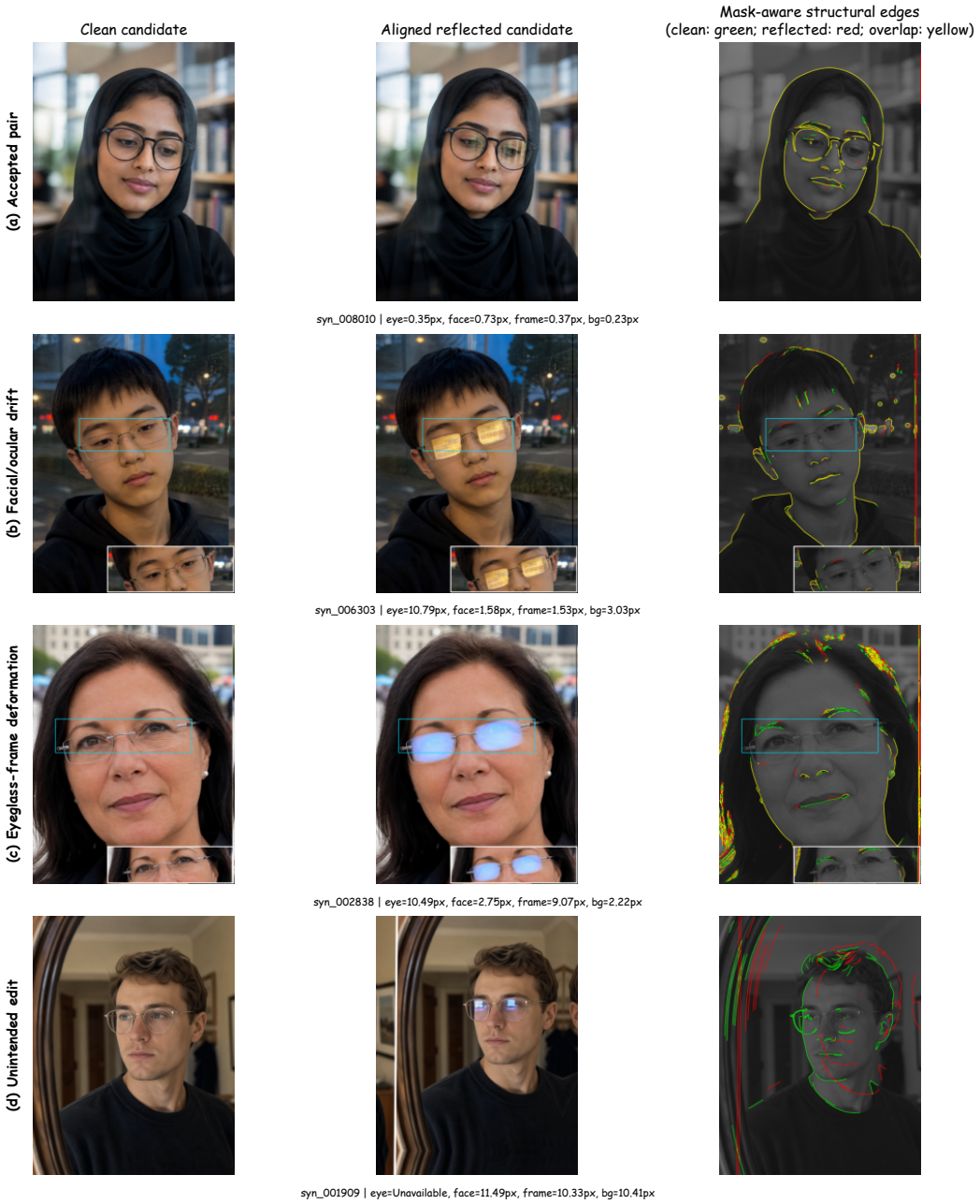

Figure 7: Representative examples of mask-aware edge-consistency filtering. Each row shows a clean candidate, its similarity-aligned reflected counterpart, and the corresponding structural-edge overlay. The examples include one accepted pair and three typical failure modes: facial/ocular drift, eyeglass-frame deformation, and unintended edits beyond the lens-reflection region.

Manual Inspection

Pairs passing the automatic filter are further reviewed by four reviewers. The reviewers assess realism, identity and ocular consistency, eyeglass geometry, and preservation of unaffected image content. Samples are rejected if they exhibit identity drift, facial or ocular deformation, eyeglass-frame distortion, changes in pose, gaze, expression, background, crop, or composition, unintended modifications outside the lens-reflection regions, or other generation artifacts that compromise pair-

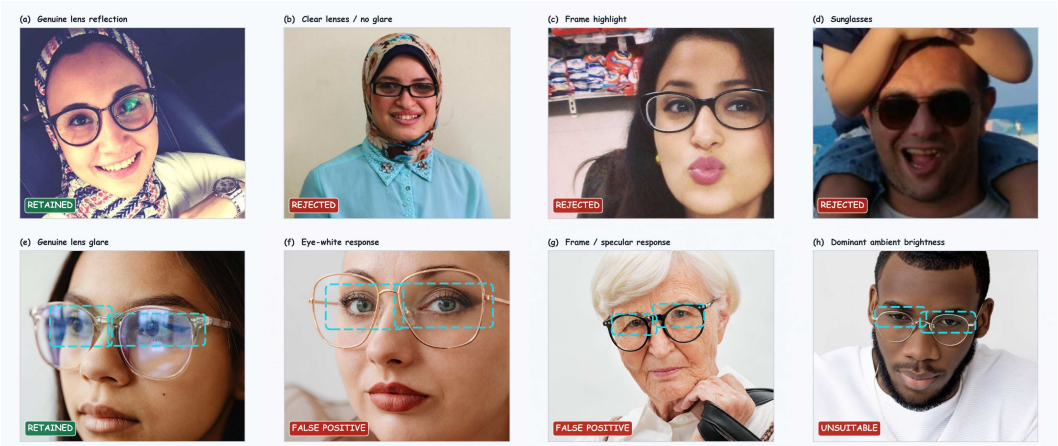

Figure 8: Representative examples of real reflected-input screening. The top row illustrates SoF samples under CLIP-based screening, including a genuine lens reflection and typical negatives such as clear lenses, frame highlights, and sunglasses. The bottom row illustrates Web samples under FaceMesh-guided local screening, including a retained lens-glare example and representative false positives caused by eye whites, frame/specular highlights, and dominant ambient brightness. Cyan boxes indicate the estimated lens regions used for local glare analysis.

wise correspondence. A pair is retained only when accepted by reviewer consensus. The complete quality-control procedure retains 10,280 clean–reflection pairs for Ocu-S.

A.2 Ocu-P Construction Details

A.2.1 Real Reflected-input Collection

For the SoF subset, we start from 2,662 images and perform CLIP-based (Radford et al., 2021) reflection screening using CLIP ViT-B/32. To emphasize localized lens reflections while retaining global semantic context, we combine the full-image and periocular-region CLIP features with weights of 0.4 and 0.6, respectively. The text representation is constructed from a multi-prompt ensemble covering five semantic categories: lens reflection, clear eyeglasses, frame highlight, sunglasses, and no glasses. For each image, we compute a reflection margin as the cosine similarity to the reflection prototype minus the highest similarity among the four negative categories, and rank all images accordingly. The top 400 candidates are manually reviewed, yielding 199 images with genuine eyeglass-lens reflections. These 199 images are used for clean-target generation and retained in the final SoF pseudo-pair subset.

For the Web subset, we retrieve 11,520 images from Pexels using eyeglass-reflection-related portrait queries. MediaPipe FaceMesh (Lugaresi et al., 2019) is used to localize the face and estimate the left and right lens regions from periocular landmarks. Within the estimated lens regions, we compute a local glare score that combines the proportions of low-saturation bright pixels and highly bright pixels, the number and area of valid highlight connected components, and the peak intensity. We retain candidates with a glare score of at least 0.58, a white-highlight ratio of at least 0.006, a bright-pixel ratio of at least 0.010, and at least one valid highlight component. This automatic screening reduces the pool from 11,520 to 1,740 images. We then manually remove false positives caused by eye whites, skin or eyeglass-frame highlights, and other non-lens bright regions, as well as images without clear lens reflections, with opaque or heavily tinted lenses, or with insufficient facial or eyeglass detail. This results in 533 Web images, all of which are retained in the final pseudo-pair set after target generation and pair-level quality con-

Table 6: Construction of the real reflected-input subset. Automatic screening uses CLIP ranking for SoF and FaceMesh-guided reflection screening for Web images.

Source	Initial	Automatic	Manual	Final
SoF	2,662	400	199	199
Web	11,520	1,740	533	533
Total	14,182	2,140	732	732

trol. Tab. 6 summarizes the screening statistics, while Fig. 8 shows representative retained examples and typical false positives from the two screening pipelines.

A.2.2 Clean-target Generation

For each retained real reflected input I_r^{real}, we keep the original image unchanged and use GPT-Image-2 to generate the missing reflection-free target $\widetilde{I}_c$ at the same resolution as the input. If an editing call fails, the request is retried after a 20-second delay, with at most five total attempts per input. If all five attempts fail, the corresponding sample is discarded and processing proceeds to the next input. In practice, no input exhausted the five-attempt budget, and all retained real reflected inputs were successfully paired with generated clean targets. The original reflected image is used directly as the input of each pseudo-pair, while GPT-Image-2 is applied only to obtain its missing clean counterpart.

The editing instruction removes lens reflections while preserving identity, facial and ocular structure, pose, eyeglass geometry, background, composition, and overall photographic appearance. A representative prompt template is shown below.

Reflection-removal prompt template

```
Preserve the person, pose, eyeglasses, background, and composition.
Remove reflections only from the eyeglass lenses, recover clear and realistic
eyes, and maintain a natural photographic appearance.
Do not modify identity, facial geometry, eyeglass geometry, or any
non-reflection content.
```

The generated image is used exclusively as the pseudo-clean target $\widetilde{I}_c$; the reflected member of the pair remains the original real image I_r^{real} without generative modification. This asymmetric construction ensures that the reflection distribution in the pseudo-pairs is inherited directly from real observations rather than from the image generator.

A.2.3 Pseudo-pair Filtering

We apply the same registration-aware, mask-aware edge-consistency filtering procedure described in § A.1.3 to assess structural agreement between each real reflected input and its generated pseudo-clean target. Pairs are subsequently reviewed by four reviewers for identity drift, facial or ocular deformation, eyeglass-frame distortion, changes in pose or gaze, background or composition changes, incomplete reflection removal, and unintended modifications outside the lens regions. A pseudo-pair is retained only with unanimous agreement among all four reviewers.

The 732 reflected inputs retained after input screening, comprising 199 images from SoF and 533 from the Web subset, are used for clean-target generation. In practice, none of the resulting pseudo-pairs were rejected during pair-level quality control, and all 732 were retained in the final Ocu-P set.

A.3 Ocu-R Construction Details

A.3.1 Reference-pair Capture and Processing

We collect 200 candidate reflected–clean pairs through two complementary acquisition protocols, comprising 100 offline pairs captured under controlled indoor conditions and 100 online pairs collected through remote video sessions.

For the offline subset, images are captured with an iPhone in a dim indoor environment under relatively soft ambient illumination. A clean image is first captured without an active reflection source. Controlled reflection sources, including a candle, flashlight, desk lamp, smartphone screen, and LED light stick, are then introduced while keeping the camera and subject approximately fixed. The reflection source is repositioned around the subject to produce spatially diverse lens reflections. Subject pose is also varied using near-frontal views, left/right head turns, and upward/downward head rotations. These controlled combinations yield 100 candidate reflected–clean pairs.

For the online subset, data are collected through remote WeChat video sessions. Participants prepare common controllable light sources, such as a computer screen and desk lamp, and switch or reposition them to produce visible eyeglass reflections. Clean and reflected frames are captured from the video stream under approximately unchanged subject pose and camera viewpoint, yielding another 100 candidate reflected–clean pairs.

Following the paired-data processing strategy of ReyeR (Zou et al., 2024), we screen the candidate pairs using geometric and appearance-based consistency checks followed by manual inspection. Samples with noticeable head motion, eye or gaze changes, eyeglass-frame displacement, background shift, camera shake, or other local structural inconsistencies are removed. We then apply color-consistency correction using white-balance adjustment, histogram matching, and color transfer. After filtering and correction, 62 offline pairs and 95 online pairs are retained, resulting in 157 reference pairs.

The retained clean images are used as visual references for qualitative analysis rather than as pixel-level ground truth. Their corresponding 157 reflected images are included in Ocu-R together with 301 independently collected unpaired reflected images, forming the full 458-image real-world evaluation set. From these 458 reflected inputs, 30 images are randomly sampled for the blind user study. The clean reference images are not shown to participants and are not used in the user-study evaluation.

A.3.2 In-the-wild Test Images

We additionally collect 301 unpaired reflected images in unconstrained real-world environments. The images are self-captured across diverse everyday scenes, including shopping malls, subway stations, streets and intersections, supermarkets, and other outdoor locations. We use opportunistic capture: subjects wear their own eyeglasses, and each image is acquired after visually confirming the presence of lens reflection at capture time. Reflections arise naturally from surrounding illumination, including nearby artificial light sources and outdoor natural light such as low-angle sunlight around sunset.

No corresponding clean image is captured for this subset. The resulting 301 images form the unpaired real-world subset of Ocu-R. The subset follows the naturally occurring variation induced by opportunistic capture, including differences in illumination, viewpoint, head pose, eyeglass appearance, reflection source, and reflection severity. These images contribute to the real-world evaluation on Ocu-R, including no-reference assessment and blind user evaluation. Representative self-collected face images are omitted from the dataset overview figure.

A.3.3 Real-world Test-set Characteristics

We characterize the independently collected real-world test set using aggregate image statistics over all 458 reflected inputs, comprising 157 paired-reference reflected images and 301 unpaired in-the-wild images. The corresponding clean captures from the paired-reference subset serve as visual references and are not included in this analysis.

We measure image resolution, aspect ratio, luminance, contrast, high-frequency detail, and color saturation. For an image of width W and height H, resolution is measured in megapixels as $HW/10^6$, and aspect ratio is defined as W/H. The RGB image is converted to luminance using

$$Y = 0.299R + 0.587G + 0.114B, \tag{12}$$

from which luminance and contrast are computed as the mean and standard deviation of Y, respectively. Color saturation is measured as the mean saturation value in HSV space.

High-frequency detail is quantified by the mean absolute Laplacian response of the luminance image,

$$H_{\text{freq}} = \frac{1}{|\Omega|} \sum_{p \in \Omega} |L(p)|\,, \tag{13}$$

where L denotes the Laplacian response of the luminance image and Ω denotes the image domain.

To standardize the spatial scale used for photometric and high-frequency measurements, images whose longer side exceeds 512 pixels are resized with preserved aspect ratio such that the longer side is 512 pixels. Resolution and aspect ratio are always computed from the original images.

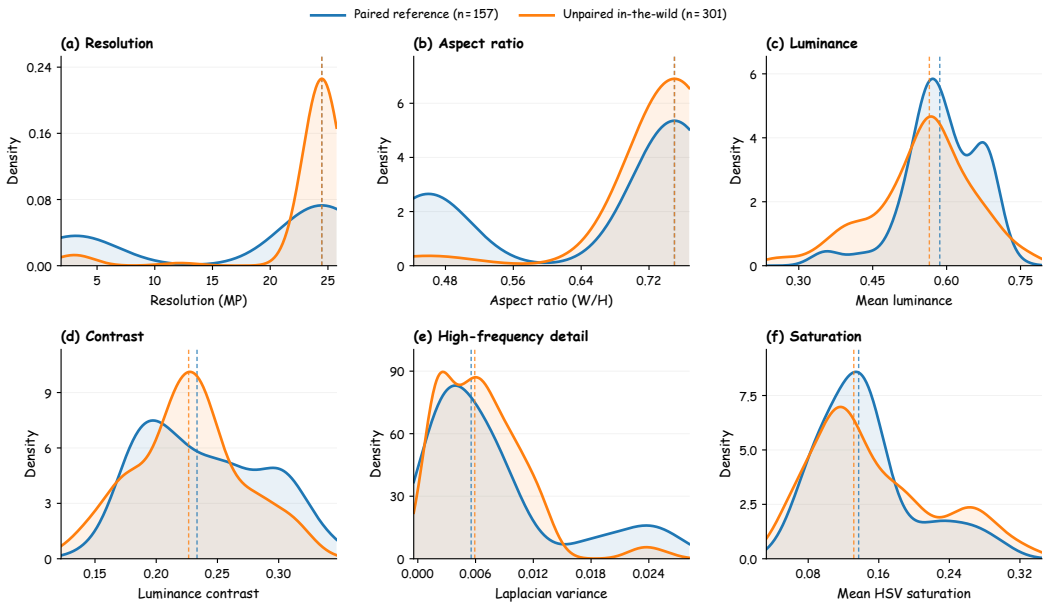

Figure 9: Image-level characteristics of the independently collected real-world test set. Smoothed distributions are shown separately for the 157 paired-reference reflected inputs and the 301 unpaired in-the-wild inputs in terms of (a) resolution, (b) aspect ratio, (c) luminance, (d) contrast, (e) high-frequency detail, and (f) color saturation. Dashed vertical lines indicate the median of each subset.

Fig. 9 summarizes the resulting image-level characteristics. The paired-reference and in-the-wild subsets exhibit broad and distinct distributions across resolution, aspect ratio, luminance, contrast, high-frequency detail, and color saturation, characterizing the range of capture conditions represented in Ocu-R.

B Method Details

B.1 Ocular-Prior-Guided Adaptive Representation

We provide additional details of the ocular-prior-guided representation introduced in § 4.1. As illustrated in Fig. 10, the reflected input is first converted into a geometry-aware importance field, from which an exact-budget quadtree determines the spatial support of each token. The resulting leaves retain both native image content and continuous geometric information for subsequent restoration.

B.1.1 Ocular Prior and Structural Support Construction

All geometric cues are extracted exclusively from the reflected observation I_g; the clean target I_0 is never used for prior construction. We detect facial landmarks with MediaPipe and retain the eye contours, eyebrow landmarks, and iris landmarks of both eyes. These cues are mapped to a common 256×256 logical domain and used to construct a spatial importance field together with structural supports for the constrained quadtree solver.

Logical-coordinate mapping. Let the original image size be $W \times H$, $S = \max(W, H)$, and $\delta_x = (S - W)/2$, $\delta_y = (S - H)/2$. A landmark with normalized image coordinates $\boldsymbol{\ell} = (u, v)$ is mapped to

$$T(\boldsymbol{\ell}) = \frac{256}{S} \left(\delta_x + uW,\, \delta_y + vH\right). \tag{14}$$

Rasterized masks are aligned to the corresponding padded square canvas and resized to the logical domain using nearest-neighbor resampling, keeping all geometry aligned with the quadtree independently of input aspect ratio.



Figure 10: Construction of the ocular-prior-guided adaptive representation. Face and iris landmarks extracted from the reflected input are used to construct geometry-derived ocular supports, which define the spatial importance field and structural constraints. An exact-budget heap solver allocates $N = 256$ quadtree leaves across four spatial scales, concentrating fine supports on ocular and lens regions while retaining coarse context elsewhere. The resulting leaf supports are mapped back to the native-resolution image for canonical patch sampling, while leaf geometry and pooled semantic priors provide conditioning for subsequent restoration.

Primitive ocular supports. For each eye, 16 eye-contour landmarks together with the corresponding eyebrow landmarks define an axis-aligned periocular box with spans d_x^k and d_y^k. We expand it by $0.22d_x^k$ horizontally, $0.35d_y^k$ upward, and $0.245d_y^k$ downward; the union of both boxes forms M_{peri}. The filled eye-contour polygons define M_{eye}. For four iris-contour landmarks $\{\mathbf{r}_j\}_{j=1}^4$, we compute

$$\mathbf{c}_{\text{iris}} = \frac{1}{4}\sum_{j=1}^{4}\mathbf{r}_j, \quad r_{\text{iris}} = 1.35\,\frac{1}{4}\sum_{j=1}^{4}\|\mathbf{r}_j - \mathbf{c}_{\text{iris}}\|_2\,. \tag{15}$$

A filled disk with center $\mathbf{c}_{\text{iris}}$ and radius r_{iris} is rasterized for each eye and dilated by one logical pixel with an elliptical structuring element to obtain M_{iris}. The eyelid support is derived directly from the eye mask as $M_{\text{lid}} = \text{Grad}_{r=2}(M_{\text{eye}})$, where $\text{Grad}_{r=2}$ is a binary morphological gradient using a 5×5 elliptical structuring element.

Face-context and spectacle-lens supports. A weak face-context support is obtained from the bounding box covering both periocular regions. If its spans are d_x and d_y, we expand it by $0.65d_x$ to both sides, $1.40d_y$ upward, and $3.00d_y$ downward, followed by clipping to the logical canvas, yielding M_{face}. For each eye contour, let (c_x^k, c_y^k) denote its bounding-box center and (d_x^k, d_y^k) its spans. We estimate a spectacle-lens support with

$$w_{\text{lens}}^k = (1+2\eta_x)d_x^k, \quad h_{\text{lens}}^k = \max\big((1+2\eta_y)d_y^k,\, 0.55w_{\text{lens}}^k\big)\,, \tag{16}$$

where $\eta_x = 0.45$ and $\eta_y = 1.00$. The corresponding axis-aligned rectangles are united to form M_{lens}; if detailed eye contours are unavailable, M_{peri} is used as a conservative fallback. This binary support is used only for adaptive allocation and is distinct from the soft renderer support A_{lens} in § 4.3.

Sparse structural anchors. We further preserve structurally critical ocular locations through sparse anchors. These include the two iris centers, the inner and outer corner of each eye, and regularly subsampled eye-contour landmarks. With the 16-point contours used in our implementation, every second point is selected, producing eight contour anchors per eye. Each anchor is rasterized as a disk of radius one logical pixel, and their union defines M_{anc}.

Spatial importance field. The geometry-derived supports are combined additively into the non-negative importance field used in Eq. (3):

$$\begin{aligned} w(\mathbf{x}) = \lambda_{\text{c}} + \lambda_{\text{f}}M_{\text{face}}(\mathbf{x}) + \lambda_{\text{p}}M_{\text{peri}}(\mathbf{x}) + \lambda_{\text{e}}M_{\text{eye}}(\mathbf{x}) \\ + \lambda_{\text{i}}M_{\text{iris}}(\mathbf{x}) + \lambda_{\text{l}}M_{\text{lid}}(\mathbf{x}) + \lambda_{\text{g}}M_{\text{lens}}(\mathbf{x}), \end{aligned} \tag{17}$$

with $(\lambda_c, \lambda_f, \lambda_p, \lambda_e, \lambda_i, \lambda_l, \lambda_g) = (0.02,\ 0.20,\ 1.00,\ 3.00,\ 8.00,\ 5.00,\ 4.00)$. These weights are designed as relative importance levels rather than independently optimized parameters. We use the periocular region as a reference level and progressively increase the weight for finer ocular structures, with the iris receiving the highest priority, followed by the eyelid and spectacle-lens regions. The face and canvas terms provide weaker global coverage, with a small positive canvas weight preventing zero-cost regions. The relative scale was chosen by coarse empirical tuning on training data and then fixed for all experiments. The supports intentionally overlap, so geometrically important locations accumulate contributions from multiple terms. The positive canvas weight avoids zero-cost regions, while the larger ocular weights progressively bias capacity toward fine eye structures and likely spectacle regions. The field w determines the soft refinement preference. For structural constraints, we define $S_{\text{iris}} = \text{supp}(M_{\text{iris}})$, $S_{\text{lid}} = \text{supp}(M_{\text{lid}})$, $S_{\text{lens}} = \text{supp}(M_{\text{lens}})$, and $S_{\text{anc}} = \text{supp}(M_{\text{anc}})$.

B.1.2 Exact-Budget Quadtree Allocation

The 256×256 logical domain is initialized by an 8×8 forest of 32×32 roots, giving $R = 64$ independent quadtrees. Each node can remain a leaf or be divided into four equal children down to the minimum size $s_{\min} = 4$. We use $\rho(s) = \log_2 \frac{s}{4}$, and define the importance mass and leaf cost of a node q as $M(q) = \int_{\Omega_q} w(\mathbf{x})\, d\mathbf{x}$ and $C(q) = \rho(s_q)M(q)$. For a selected leaf q_i, this notation coincides with the main-text definition $M_i = M(q_i)$. All masses are queried in constant time from an integral image of w.

Structural admissibility. The importance field softly favors refinement, while selected structural supports impose a maximum admissible leaf size. For a hard-support family $\mathcal{H} \subseteq \{S_{\text{iris}}, S_{\text{lid}}, S_{\text{lens}}, S_{\text{anc}}\}$, a leaf is admissible iff

$$\text{adm}(q; \mathcal{H}) = \bigwedge_{S \in \mathcal{H}} \left[\Omega_q \cap S = \varnothing \ \vee \ s_q \le 8\right]. \tag{18}$$

Thus, any leaf larger than 8×8 that intersects an active iris, eyelid, lens, or anchor support must be split.

Exact dynamic-programming formulation. A generic exact solution can be obtained by defining $D_q(k)$ as the minimum cost of representing the subtree rooted at q with exactly k admissible leaves:

$$D_q(k) = \begin{cases} C(q), & k = 1 \text{ and } \text{adm}(q; \mathcal{H}), \\ \min\limits_{\substack{k_1+\cdots+k_4=k \\ k_j \ge 1}} \sum\limits_{j=1}^{4} D_{q_j}(k_j), & s_q > 4, \\ +\infty, & \text{otherwise}, \end{cases} \tag{19}$$

where $\{q_j\}_{j=1}^4$ are the four children of q. The resulting root budget curves can be combined by a second knapsack dynamic program (DP) under the global constraint $\sum_{r=1}^{R} k_r = N$. This formulation is exact but requires maintaining and convolving budget tables throughout the forest.

Exact heap specialization. Our choice of ρ admits a substantially simpler exact solver. Splitting a node q of size s into four children of size $s/2$ decreases the objective by

$$\Delta(q) = C(q) - \sum_{q' \in \text{ch}(q)} C(q') = \rho(s)M(q) - \rho(s/2) \sum_{q' \in \text{ch}(q)} M(q') \tag{20}$$

$$= \big[\rho(s) - \rho(s/2)\big] M(q) = M(q), \tag{21}$$

because the child supports partition Ω_q and $\rho(s) - \rho(s/2) = 1$. Therefore, if $\mathcal{S}$ denotes the set of internal nodes that are split,

$$J(\mathcal{Q}) = J_{\text{root}} - \sum_{q \in \mathcal{S}} M(q), \tag{22}$$

where J_{root} is the cost of the initial 64-root forest. Since every split replaces one leaf by four, it increases the leaf count by exactly three. Hence any N-leaf solution contains exactly

$$K = \frac{N - R}{3} \tag{23}$$

internal splits; with $R = 64$ and $N = 256$, $K = 64$. Minimizing Eq. (22) is therefore equivalent to selecting exactly K feasible splits with maximum total importance mass.

Crucially, $w(\mathbf{x}) \geq 0$ implies

$$M(q') \leq M(q), \qquad \forall q' \in \mathrm{ch}(q), \tag{24}$$

so split gains are non-increasing along every root-to-leaf path. Any feasible split set is ancestor-closed because a node can be split only after its ancestors have been split. Consequently, in the unconstrained case, repeatedly selecting the currently available node with the largest $M(q)$ produces a globally non-increasing feasible ordering of split gains and maximizes the total gain for any fixed number of splits.

The same argument applies after imposing hard structural constraints. Let $\mathcal{S}_{\mathrm{hard}}$ denote the ancestor-closed set of mandatory splits required to make all current leaves admissible. Conditioned on including $\mathcal{S}_{\mathrm{hard}}$, its induced leaf frontier forms a new forest whose remaining split gains retain the monotonicity in Eq. (24). Therefore, selecting the next $K - K_{\mathrm{hard}}$ splits by the same max-heap rule is optimal among all exact-budget partitions satisfying the current hard constraints. Equal-mass ties may yield different leaf layouts but identical objective values.

Hard constraints and feasibility. We use the following ordered hard-support schedule:

$$\begin{aligned} \mathcal{H}_1 &= \{S_{\mathrm{iris}}, S_{\mathrm{lid}}, S_{\mathrm{lens}}, S_{\mathrm{anc}}\}, \\ \mathcal{H}_2 &= \{S_{\mathrm{iris}}, S_{\mathrm{lid}}, S_{\mathrm{anc}}\}, \\ \mathcal{H}_3 &= \{S_{\mathrm{anc}}\}, \\ \mathcal{H}_4 &= \varnothing. \end{aligned} \tag{25}$$

For each relaxation level $\mathcal{H}_j$, the solver starts from a fresh copy of the original $R = 64$ root forest. It recursively splits every inadmissible leaf until all frontier leaves satisfy Eq. (18). Let $\mathcal{S}_{\mathrm{hard}}(\mathcal{H}_j)$ denote the resulting set of mandatory internal splits and

$$K_{\mathrm{hard}} = |\mathcal{S}_{\mathrm{hard}}(\mathcal{H}_j)| \,. \tag{26}$$

If $K_{\mathrm{hard}} > K$, that relaxation level is infeasible and is discarded entirely. Otherwise, the remaining $K - K_{\mathrm{hard}}$ splits are selected from the induced leaf frontier using a max-heap keyed by $M(q)$. The first feasible relaxation level is used.

Importantly, mandatory splits from an infeasible stricter level are not carried into the next level; the forest and split count are reinitialized before each solve. Relaxing a hard support does not remove its contribution from w. For example, the lens region retains its soft refinement weight even after its hard 8×8 maximum-leaf-size constraint is relaxed.

Geometry fallback. We use all available ocular cues without requiring complete iris detection. For each detected face, missing iris landmarks remove only the corresponding iris prior, while available eye and periocular supports are retained. For multi-face inputs, the eye, iris, periocular, and lens supports from the detected faces are jointly incorporated into the spatial prior and renderer support. Adaptive allocation falls back to the uniform 16×16 grid only when no detected face provides reliable bilateral periocular geometry and eye centers; available partial region priors are still retained for conditioning and renderer support. If face geometry is unavailable altogether, the region prior becomes background-only and the lens support is set to zero, so the final renderer preserves the input. **Complexity.** Let $A = 256^2$ be the logical-canvas area. Integral images for the importance field and structural supports require $O(A)$ time and memory. At each of the four fixed relaxation levels, mandatory refinement is terminated once more than K splits would be required, so only $O(N)$ quadtree nodes need to be materialized. The active frontier contains at most $R + 3K = N$ leaves, and each optional split performs one heap extraction and at most four insertions, each in $O(\log N)$ time. Since the number of relaxation levels is constant, the specialized solver runs in $O(A + N \log N)$ time and $O(A + N)$ memory. In contrast, our direct budget-table DP implementation maintains up to N states per quadtree node and combines child tables by min-plus convolution, giving a worst-case cost of $O(|V|N^2)$ and $O(|V|N)$ storage for a forest with $|V|$ possible nodes. The heap solver therefore exploits the particular scale-linear objective to remove the budget dimension while preserving exact optimality. Alg. 1 summarizes the solver used in practice. We retain the dynamic-programming implementation as an exact regression verifier: at each tested hard-support level, both solvers are required to agree on feasibility and objective value, although tied constant-importance regions may lead to different but equally optimal leaf layouts. **Native-coordinate patch**

Algorithm 1 Exact-budget quadtree allocation

Require: Importance field w; supports $S_{\text{iris}}, S_{\text{lid}}, S_{\text{lens}}, S_{\text{anc}}$; target budget N
Ensure: Partition $\mathcal{Q}^*$ with $|\mathcal{Q}^*| = N$

Initialize the $R = 64$ root forest $\mathcal{Q}_0$
Build integral images for w and all structural supports
if $N < R$ or $(N - R) \bmod 3 \neq 0$ **then**
    return infeasible
end if
$K \leftarrow (N - R)/3$
Define $\mathcal{H}_1, \ldots, \mathcal{H}_4$ as in Eq. (25)
for $\mathcal{H} \in [\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3, \mathcal{H}_4]$ **do**
    $\mathcal{Q} \leftarrow$ fresh copy of $\mathcal{Q}_0$
    $\mathcal{S}_{\text{hard}} \leftarrow \varnothing$
    while there exists $q \in \mathcal{Q}$ with $\neg\,\text{adm}(q; \mathcal{H})$ **do**
        Select an inadmissible frontier leaf q
        Replace q in $\mathcal{Q}$ by $\text{ch}(q)$
        $\mathcal{S}_{\text{hard}} \leftarrow \mathcal{S}_{\text{hard}} \cup \{q\}$
        if $|\mathcal{S}_{\text{hard}}| > K$ **then**
            break
        end if
    end while
    if $|\mathcal{S}_{\text{hard}}| > K$ **then**
        continue
    end if
    $\mathcal{S} \leftarrow \mathcal{S}_{\text{hard}}$
    Initialize an empty max-heap PQ
    for all $q \in \mathcal{Q}$ with $s_q > 4$ **do**
        Insert q into PQ with key $M(q)$
    end for
    while $|\mathcal{S}| < K$ **do**
        if PQ is empty **then**
            break
        end if
        $q \leftarrow \text{ExtractMax}(\text{PQ})$
        Replace q in $\mathcal{Q}$ by $\text{ch}(q)$
        $\mathcal{S} \leftarrow \mathcal{S} \cup \{q\}$
        for all $q' \in \text{ch}(q)$ with $s_{q'} > 4$ **do**
            Insert q' into PQ with key $M(q')$
        end for
    end while
    if $|\mathcal{S}| = K$ **then**
        return $\mathcal{Q}$ as $\mathcal{Q}^*$
    end if
end for
return infeasible

sampling. After obtaining $\mathcal{Q}^*$, each logical leaf support is mapped through the same square-canvas geometry back to the native image coordinate system. RGB content is sampled directly from the native-resolution image and resampled to a canonical 24×24 patch, yielding one patch per adaptive leaf. Thus, the 256×256 domain determines only the spatial partition, rather than imposing a 256×256 image-resolution bottleneck. **Leaf-wise semantic priors.** Given the adaptive partition $\mathcal{Q}^*$, we construct four semantic maps corresponding to background, periocular, eye, and iris regions. For each leaf q_i with support Ω_i, average pooling over these maps yields the leaf-wise semantic prior

$$\mathbf{p}_i = \begin{bmatrix} p_{i,\text{bg}} \\ p_{i,\text{peri}} \\ p_{i,\text{eye}} \\ p_{i,\text{iris}} \end{bmatrix} \in [0,1]^4. \tag{27}$$

These pooled occupancies provide ocular-region conditioning to the restoration backbone.

B.2 Geometry-Guided One-Step pMF

B.2.1 Geometry and Cross-Scale Conditioning

We further specify the conditioning representation $\mathbf{h}$ used in § 4.2. For each adaptive leaf $q_i = (\mathbf{c}_i, s_i)$, the current path state $\mathbf{z}_{t,i}$ defined by Eq. (4) and the reflected observation patch $\mathbf{g}_i$ are encoded separately into a state token $\mathbf{e}_i$ and an observation token $\mathbf{o}_i$, respectively. We additionally embed the continuous leaf geometry as $\mathbf{e}_i^{\text{geo}} = \mathrm{PE}(\mathbf{c}_i/256, \log s_i)$, where PE encodes the continuous position and scale of each adaptive leaf. The geometry embedding is added to the query and key features of the conditioning attention, while values contain appearance features only. Thus, geometry controls spatial correspondence without being treated as image content.

We first inject reflected-observation features into the restoration state. Define the active token set as $\mathcal{A} = \{i : 1 - p_{\text{bg},i} > 10^{-4} \ \vee \ s_i \le 8\}$. The first condition selects leaves with non-background ocular support, whereas the second retains all fine leaves with $s_i \in \{4, 8\}$, including those allocated around lens boundaries. Using geometry-aware cross-attention, the intermediate state tokens are

$$\mathbf{e}'_i = \mathbf{e}_i + \mathbf{1}_{\{i\in\mathcal{A}\}}\, \boldsymbol{\gamma}_{\text{obs}} \odot \mathrm{Attn}_{\text{geo}}(\mathbf{e}_i,\, \{\mathbf{o}_j\}_{j\in\mathcal{A}})\,, \tag{28}$$

where $\boldsymbol{\gamma}_{\text{obs}}$ is a learned channel-wise residual gate. Consequently, only geometry-relevant restoration tokens are updated by this branch, while their observation features are associated through continuous position and scale. We then supply larger-context information specifically to high-resolution ocular tokens. Let $\mathcal{F} = \{i : s_i \le 8\}$ and $\mathcal{C} = \{i : s_i \ge 16\}$ denote the fine and coarse token sets. A second geometry-aware attention uses coarse state tokens as context and updates only fine tokens:

$$\mathbf{h}_i = \mathbf{e}'_i + \mathbf{1}_{\{i\in\mathcal{F}\}}\, \boldsymbol{\gamma}_{\text{ctx}} \odot \mathrm{Attn}_{\text{geo}}\big(\mathbf{e}'_i,\, \{\mathbf{e}'_j\}_{j\in\mathcal{C}}\big)\,. \tag{29}$$

Hence, fine tokens with $s_i \in \{4, 8\}$ retain access to the broader facial appearance and illumination represented by coarse supports with $s_i \in \{16, 32\}$, while coarse tokens are left unchanged by the cross-scale branch. The collection $\mathbf{h} = \{\mathbf{h}_i\}_{i=1}^N$ constitutes the conditioned representation used by the restoration backbone. The leaf-wise semantic prior is additionally supplied to the backbone as ocular-region conditioning. Both attention branches use learned residual gates, and their output projections are initialized to zero so that the added conditioning paths begin as identity-preserving residual branches.

B.2.2 Observation-Anchored Source

Let $\mathbf{g} = \{\mathbf{g}_i\}_{i=1}^N$ denote the adaptive patches sampled from the reflected observation. For each leaf, we define the modulation weight

$$m_i = \max\big(1 - p_{\text{bg},i}, \tau\, \mathbf{1}_{\{s_i\le 8\}}\big)\,, \tag{30}$$

where $p_{\text{bg},i}$ is the leaf-wise background occupancy and $\tau = 0.75$ provides a minimum modulation level for fine leaves. The observation-anchored source is then

$$\widetilde{\mathbf{g}} = \mathbf{g} + \sigma \mathbf{m} \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}), \tag{31}$$

where $\mathbf{m}$ stacks the token-wise modulation weights and is broadcast over the corresponding RGB patches. We use $\sigma = 0.05$.

The term $1 - p_{\text{bg},i}$ increases perturbation according to non-background ocular occupancy, while the floor τ prevents fine leaves, including those near lens boundaries, from receiving negligible perturbation when their pooled occupancy is small. For fine leaves, $\tau = 0.75$ guarantees a modulation weight of at least 0.75, while $\sigma = 0.05$ keeps the perturbed source close to the reflected observation. Both values are fixed for all experiments.

B.2.3 pMF Parameterization

Along the linear path defined in Eq. (4), the clean adaptive patches $\mathbf{x}_0$ correspond to $t = 0$, while the observation-anchored source $\widetilde{\mathbf{g}}$ corresponds to $t = 1$.

For an interval $0 \le r \le t \le 1$, let $\Delta = t - r$. Given the current path state $\mathbf{z}_t$, the image predictor $\widehat{\mathbf{x}}_\theta(\mathbf{z}_t, \Delta; \mathbf{h})$ induces the average velocity

$$\mathbf{u}_\theta(\mathbf{z}_t, r, t; \mathbf{h}) = \frac{\mathbf{z}_t - \widehat{\mathbf{x}}_\theta(\mathbf{z}_t, \Delta; \mathbf{h})}{t}, \qquad t > 0. \tag{32}$$

The corresponding update from t to the earlier endpoint r is

$$\widehat{\mathbf{z}}_r = \mathbf{z}_t - (t - r)\mathbf{u}_\theta(\mathbf{z}_t, r, t; \mathbf{h}) = \frac{r}{t}\mathbf{z}_t + \frac{t - r}{t}\widehat{\mathbf{x}}_\theta(\mathbf{z}_t, \Delta; \mathbf{h}). \tag{33}$$

For one-step inference, setting $(r, t) = (0, 1)$ gives $\Delta = 1$ and $\mathbf{z}_1 = \widetilde{\mathbf{g}}$. The update therefore reduces to

$$\widehat{\mathbf{x}} \equiv \widehat{\mathbf{z}}_0 = \widehat{\mathbf{x}}_{\bar{\theta}}(\widetilde{\mathbf{g}}, 1; \mathbf{h}), \tag{34}$$

which is the one-evaluation restoration rule used in § 4.2.

B.2.4 MeanFlow Time Sampling

During training, we sample the MeanFlow interval endpoints through two independent logit-normal draws. Specifically,

$$\xi_t, \xi_r \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1), \quad \widetilde{t} = \mathrm{sigmoid}(\mu + \sigma_t \xi_t), \quad \widetilde{r} = \mathrm{sigmoid}(\mu + \sigma_t \xi_r). \tag{35}$$

To jointly expose the model to finite MeanFlow intervals and their instantaneous limit, we draw $b \sim \mathrm{Bernoulli}(p_{\mathrm{FM}})$ and collapse the pair when $b = 1$:

$$\widetilde{r} \leftarrow \begin{cases} \widetilde{t}, & b = 1, \\ \widetilde{r}, & b = 0. \end{cases} \quad t = \max(\widetilde{t}, \widetilde{r}), \quad r = \min(\widetilde{t}, \widetilde{r}). \tag{36}$$

The sorting guarantees $0 \le r \le t \le 1$. Collapsed samples satisfy $r = t$ and therefore have zero interval length, whereas the remaining samples provide finite-interval MeanFlow conditioning. We use $\mu = -0.4$, $\sigma_t = 1.0$, and $p_{\mathrm{FM}} = 0.5$ in our default restoration setting.

The restoration backbone is conditioned by the current path time t and the interval length $\Delta = t - r$. Thus, the endpoint r is represented implicitly by (t, Δ). At inference, the one-step setting used in § 4.2 corresponds to $(r, t) = (0, 1)$ and hence $\Delta = 1$.

B.3 Native-Resolution Frequency-Preserving Synthesis

We provide additional details of the native-resolution synthesis module introduced in § 4.3. As illustrated in Fig. 11, the restored canonical patches are first projected back to their native spatial supports. We then construct complementary absolute and residual reconstruction candidates and combine them in the frequency domain. A lightweight residual gate locally adjusts the frequency mixing, while the final lens-supported composition preserves the observed image exactly outside the editable region.

B.3.1 Native Projection and Lens Support

We first detail the projection operator $\mathcal{R}$ used in § 4.3. For each adaptive leaf q_i, let $\mathcal{W}_i(\mathbf{y}_i)$ denote bicubic resampling of its canonical patch $\mathbf{y}_i$ onto the corresponding native-resolution support. Each support is extended by a small proportional halo and weighted by a separable raised-cosine window w_i that equals one over the interior and smoothly decays to zero across the halo. Overlapping predictions are then combined by normalized partition-of-unity blending:

$$\mathcal{R}(\mathbf{y})(\mathbf{p}) = \frac{\sum_i w_i(\mathbf{p})\, \mathcal{W}_i(\mathbf{y}_i)(\mathbf{p})}{\sum_i w_i(\mathbf{p}) + \varepsilon}. \tag{37}$$

This normalized overlap suppresses seams between neighboring adaptive supports while preserving the linearity of $\mathcal{R}$, which is used by both absolute and residual reconstruction in Eq. (6).

The soft lens-support map A_{lens} is constructed independently of the adaptive partition. Spectacle geometry derived from the ocular landmarks is rasterized on the logical canvas, expanded to cover the lens region, and feathered before being mapped to the native resolution. As a result, the final compositing support is spatially smooth and does not inherit quadtree boundaries from $\mathcal{Q}^*$.

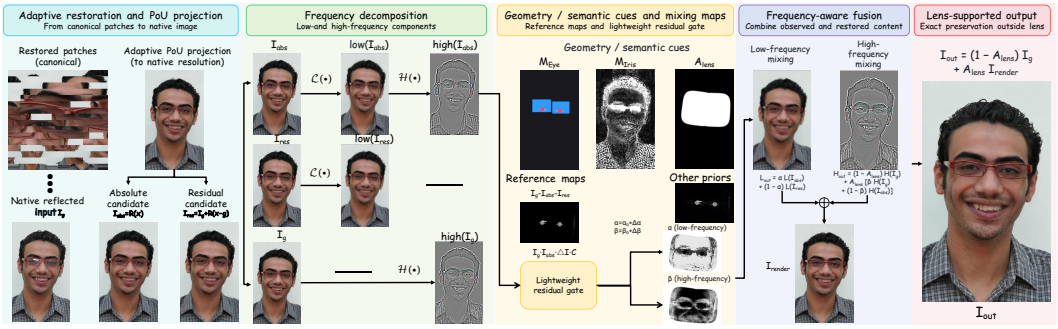

Figure 11: Overview of native-resolution frequency-preserving synthesis. Projected patch predictions form absolute and residual reconstructions. Spatial maps α and β mix $\mathcal{L}(I_{\text{abs}})$ with $\mathcal{L}(I_{\text{res}})$ and $\mathcal{H}(I_g)$ with $\mathcal{H}(I_{\text{abs}})$, respectively. Final lens-supported compositing preserves the observed input outside the spectacle region.

B.3.2 Spatial Gate Construction

The frequency gates distinguish regions that require reconstruction from those where native photographic information can be safely retained. We implement the native-resolution low-pass operator $\mathcal{L}$ as a 15×15 box filter, and define the corresponding high-frequency component as in § 4.3.

To estimate where the restored patches consistently request a change to the observation, we first form the projected correction

$$\Delta I = \mathcal{R}(\hat{\mathbf{x}} - \mathbf{g}), \quad C = \text{mean}_{\text{RGB}} \left| \mathcal{L}_c(\Delta I) \right|, \tag{38}$$

where $\mathcal{L}_c$ is an 11×11 box filter. We convert this magnitude into a soft reconstruction-need map

$$P_{\text{corr}} = s\left(\frac{C - 0.03}{0.10 - 0.03} \right), \quad s(u) = \bar{u}^{\,2}(3 - 2\bar{u}), \quad \bar{u} = \text{clip}(u, 0, 1). \tag{39}$$

Thus, correction magnitudes below 0.03 receive little reconstruction evidence, while values above 0.10 saturate the gate.

We additionally obtain eye and iris supports from the ocular prior. Their binary maps are smoothed independently with a 7×7 box filter and passed through the same smoothstep function, yielding soft maps $\widetilde{M}_{\text{eye}}$ and $\widetilde{M}_{\text{iris}}$. To identify strong near-white reflections, we convert the native observation from $[-1, 1]$ to $[0, 1]$ as $I_g^{01} = (I_g + 1)/2$ and define

$$B = \text{mean}_{\text{RGB}}\, I_g^{01}, \quad \chi = \max_{\text{RGB}} I_g^{01} - \min_{\text{RGB}} I_g^{01}. \tag{40}$$

Here B is a brightness cue and χ measures local chroma spread. The soft highlight response is

$$S = A_{\text{lens}}\, s\left(\frac{B - 0.82}{0.96 - 0.82} \right) s\left(\frac{0.16 - \chi}{0.16} \right). \tag{41}$$

A response therefore requires both high brightness and low chroma within the lens support.

These signals define analytical initializations for the two frequency gates. For the low-frequency branch,

$$\alpha_0 = A_{\text{lens}}\, \max\left(0,\, 0.35\widetilde{M}_{\text{eye}},\, 0.55\widetilde{M}_{\text{iris}},\, 0.90S \right). \tag{42}$$

Thus, ordinary lens regions initially favor residual reconstruction, whereas eye, iris, and strong-highlight regions receive progressively larger contributions from the absolute clean candidate.

For the high-frequency branch, let $\text{lerp}(a, b; m) = (1 - m)a + mb$. We first define the within-lens native-detail weight

$$\widetilde{\beta}_0 = \text{lerp}\Big(\text{lerp}\left(0.90,\, 0.30;\, \widetilde{M}_{\text{eye}} \right), 0.15;\, \widetilde{M}_{\text{iris}} \Big) \times (1 - 0.75P_{\text{corr}})\,(1 - S), \tag{43}$$

and then extend it to the full image as

$$\beta_0 = (1 - A_{\text{lens}}) + A_{\text{lens}}\widetilde{\beta}_0. \tag{44}$$

Accordingly, native high-frequency information receives an initial weight of 0.90 over ordinary lens regions, 0.30 over the eye, and 0.15 over the iris, before attenuation by correction evidence and the highlight cue. Outside the soft lens support, $\alpha_0 = 0$ and $\beta_0 = 1$.

A lightweight learned gate refines these analytical maps rather than predicting the rendered image itself. Its input comprises $\left[I_g,\, I_{\text{abs}},\, \Delta I,\, C,\, \widetilde{M}_{\text{eye}},\, \widetilde{M}_{\text{iris}},\, A_{\text{lens}},\, \alpha_0,\, \beta_0\right]$, forming 15 channels. Before entering the gate, ΔI is scaled by 0.5 and clipped to $[-1, 1]$, while C is normalized by 0.10 and clipped to $[0, 2]$. All inputs are bilinearly resized to a fixed 256×256 gate canvas.

The gate uses a 3×3 convolutional stem with width 32, followed by three depthwise-separable residual blocks with dilation rates 1, 2, and 4. Each block consists of a depthwise 3×3 convolution, a pointwise 1×1 convolution, GroupNorm, and SiLU activation. A final 1×1 convolution predicts two residual maps. The entire gate contains only $8{,}578$ trainable parameters. Its output head is initialized to zero, so optimization starts exactly from the analytical gates α_0 and β_0.

Let $(\delta_\alpha, \delta_\beta) = \tanh G_\phi\left(I_g, I_{\text{abs}}, \Delta I, C, \widetilde{M}_{\text{eye}}, \widetilde{M}_{\text{iris}}, A_{\text{lens}}, \alpha_0, \beta_0\right)$ denote the predicted residuals after bilinear upsampling to the native image resolution. The final gates are

$$\alpha = \text{clip}\left(\alpha_0 + A_{\text{lens}}\delta_\alpha,\, 0, 1\right), \quad \beta = \text{clip}\left(\beta_0 + A_{\text{lens}}\delta_\beta,\, 0, 1\right). \tag{45}$$

They are used directly in Eq. (9). The learned module therefore adjusts only the two spatial mixing maps, while the frequency decomposition and final lens-supported composition remain unchanged.

C Experimental Details

C.1 Implementation and Evaluation

C.1.1 Training Setup and Objectives

Training data handling. Ocu-S pairs are stored at 1008×1296. For Ocu-P, each real reflected input retains its original resolution, and the corresponding pseudo-clean target is generated at the same resolution. During training, adaptive patches are extracted directly from the stored paired images at their respective resolutions.

OcuFlow is trained with a velocity-space restoration objective together with image-space constraints for reconstruction, ocular fidelity, and preservation. The native-resolution renderer is optimized with a separate frequency-aware objective.

Velocity-space restoration. For the conditional path in § 4.2, let $\mathbf{v}^\star = \widetilde{\mathbf{g}} - \mathbf{x}_0$ denote the instantaneous path velocity, and let (r, t) be sampled as described in § B.2. The image-valued prediction induces the average velocity $\mathbf{u}_\theta$. We evaluate its total derivative along the ground-truth path using a Jacobian–vector product,

$$D_t\mathbf{u}_\theta = \text{JVP}_{(\mathbf{z},t,r)}[\mathbf{u}_\theta; (\mathbf{v}^\star, 1, 0)], \qquad \mathbf{U}_\theta = \mathbf{u}_\theta + (t - r)\,\text{sg}[D_t\mathbf{u}_\theta], \tag{46}$$

where $\text{sg}[\cdot]$ denotes stop-gradient. We supervise $\mathbf{U}_\theta$ toward $\mathbf{v}^\star$ and retain a training-only instantaneous-velocity head $\mathbf{v}_\theta^{\text{aux}}$. We apply detached per-image adaptive normalization

$$\text{AN}(E) = \frac{E}{\text{sg}[(E + \epsilon_{\text{n}})^p]}, \qquad p = 1, \qquad \epsilon_{\text{n}} = 0.01. \tag{47}$$

For image b, we define

$$E_U^{(b)} = \kappa \sum_{i=1}^{N} \left\| \mathbf{U}_{\theta,i}^{(b)} - \mathbf{v}_i^{\star,(b)} \right\|_2^2, \tag{48}$$

$$E_V^{(b)} = \kappa \sum_{i=1}^{N} \left\| \mathbf{v}_{\theta,i}^{\text{aux},(b)} - \mathbf{v}_i^{\star,(b)} \right\|_2^2, \qquad \kappa = \left(\frac{16}{24}\right)^2. \tag{49}$$

The factor κ compensates for the increased canonical pixel count relative to the 16×16 reference scale. The velocity-space objective is

$$\mathcal{L}_{\mathrm{vel}} = \frac{1}{B} \sum_{b=1}^{B} \left[\mathrm{AN}(E_U^{(b)}) + \mathrm{AN}(E_V^{(b)}) \right]. \tag{50}$$

The auxiliary velocity branch is used only during training.

Restoration regularization. We impose reconstruction fidelity at medium and coarse canonical scales. Let $\mathcal{D}_k$ denote area resampling to $k \times k$. We define

$$\mathcal{E}_k = \frac{1}{BN} \sum_{b=1}^{B} \sum_{i=1}^{N} \frac{16^2}{k^2} \left\| \mathcal{D}_k(\widehat{\mathbf{x}}_{0,i}^{(b)}) - \mathcal{D}_k(\mathbf{x}_{0,i}^{(b)}) \right\|_2^2, \qquad \mathcal{L}_x = \mathcal{E}_{12} + 0.5\, \mathcal{E}_6. \tag{51}$$

Semantic priors weight first-order gradient supervision over ocular structures:

$$w_i^{\mathrm{eye}} = 0.05\, p_{i,\mathrm{peri}} + p_{i,\mathrm{eye}} + 2\, p_{i,\mathrm{iris}}, \tag{52}$$

$$e_i^{\nabla} = \frac{1}{2} \left(\|\nabla_x \widehat{\mathbf{x}}_{0,i} - \nabla_x \mathbf{x}_{0,i}\|_1^{\mathrm{mean}} + \|\nabla_y \widehat{\mathbf{x}}_{0,i} - \nabla_y \mathbf{x}_{0,i}\|_1^{\mathrm{mean}} \right), \tag{53}$$

and

$$\mathcal{L}_{\mathrm{eye}} = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_i w_i^{\mathrm{eye},(b)} e_i^{\nabla,(b)}}{\sum_i w_i^{\mathrm{eye},(b)} + \varepsilon}. \tag{54}$$

Using the geometry-guided restoration support m_i defined in § 4.2, let

$$d_i = \|\widehat{\mathbf{x}}_{0,i} - \mathbf{g}_i\|_1^{\mathrm{mean}}. \tag{55}$$

Leaves are weighted by their corresponding spatial area:

$$\mathcal{L}_{\mathrm{pres}} = \frac{1}{B} \sum_{b=1}^{B} \frac{\sum_i s_i^2 (1 - m_i) d_i}{\sum_i s_i^2 (1 - m_i) + \varepsilon}. \tag{56}$$

The complete restoration objective is

$$\mathcal{L}_{\mathrm{rest}} = \mathcal{L}_{\mathrm{vel}} + 0.05\, \mathcal{L}_x + 0.005\, \mathcal{L}_{\mathrm{eye}} + 0.01\, \mathcal{L}_{\mathrm{pres}}. \tag{57}$$

Renderer objective. The frequency-preserving renderer is supervised over the full-resolution lens support. Let $A = A_{\mathrm{lens}}$ and define the lens-normalized average of a C-channel field F as

$$\langle F \rangle_A = \frac{\sum_{\mathbf{x},c} A(\mathbf{x}) F_c(\mathbf{x})}{C \sum_{\mathbf{x}} A(\mathbf{x}) + \varepsilon}. \tag{58}$$

We use reconstruction, low-frequency, and first-order gradient terms:

$$\mathcal{L}_{\mathrm{rec}} = \left\langle \sqrt{(I_{\mathrm{out}} - I_0)^2 + \epsilon_c^2} \right\rangle_A, \qquad \epsilon_c = 10^{-3}, \tag{59}$$

$$\mathcal{L}_{\mathrm{low}} = \langle |\mathcal{L}(I_{\mathrm{out}}) - \mathcal{L}(I_0)| \rangle_A, \tag{60}$$

$$\mathcal{L}_{\mathrm{grad}} = \frac{1}{2} \sum_{d \in \{x,y\}} \langle |\nabla_d I_{\mathrm{out}} - \nabla_d I_0| \rangle_{A_d}, \tag{61}$$

where $\mathcal{L}$ is the same 15×15 low-pass operator used by the renderer and A_d denotes the valid lens support for the corresponding finite-difference direction.

Let

$$\mathbf{r}_\phi = [A_{\mathrm{lens}} \Delta\alpha_\phi, A_{\mathrm{lens}} \Delta\beta_\phi] \tag{62}$$

denote the lens-masked residual gate fields. Their magnitude and spatial variation are regularized by

$$\mathcal{L}_{\mathrm{gate}} = \|\mathbf{r}_\phi\|_1^{\mathrm{mean}}, \qquad \mathcal{L}_{\mathrm{TV}} = \frac{1}{2} \left(\|\nabla_x \mathbf{r}_\phi\|_1^{\mathrm{mean}} + \|\nabla_y \mathbf{r}_\phi\|_1^{\mathrm{mean}} \right). \tag{63}$$

The complete renderer objective is

$$\mathcal{L}_{\text{rend}} = \mathcal{L}_{\text{rec}} + 0.5\,\mathcal{L}_{\text{low}} + 0.1\,\mathcal{L}_{\text{grad}} + 0.01\,\mathcal{L}_{\text{gate}} + 0.01\,\mathcal{L}_{\text{TV}}. \tag{64}$$

Optimization. We initialize the restoration backbone from pretrained **pMF-B-16** weights (Lu et al., 2026) and train OcuFlow for 60k steps with batch size 6. Fixed-size adaptive patch representations are processed jointly across the batch, while native-resolution renderer supervision is evaluated sample by sample and accumulated before each optimizer update.

The first 45k steps train the restoration backbone and adaptive conditioner with the renderer frozen. We then freeze the restoration network and optimize the renderer gate for 5k steps, followed by 10k steps of joint fine-tuning. The renderer objective is linearly ramped to its full weight during the first 2k joint steps.

We use AdamW with $(\beta_1, \beta_2) = (0.9, 0.95)$ and weight decay 10^{-4}. The backbone and conditioner use learning rate 2×10^{-5} during the first stage and 2×10^{-6} and 10^{-5} during joint fine-tuning, respectively. The renderer uses learning rate 10^{-4} whenever it is optimized. Training uses BF16 precision, gradient clipping at global norm 5, and EMA decay 0.999.

C.1.2 Inference Details

For all reported results, we use the final EMA checkpoint. Given an input image of size $H \times W$, OcuFlow edge-pads the shorter spatial dimension to a square canvas of side $S = \max(H, W)$. The estimated ocular geometry is mapped to the 256×256 logical domain, where the adaptive quadtree produces $N = 256$ leaves.

Each leaf is mapped to its corresponding support on the $S \times S$ image canvas, and its RGB content is bicubically resampled into a 24×24 canonical patch. The observation-anchored source perturbation described in § 4.2 is applied using a fixed evaluation seed. Restoration uses the endpoint setting $(r, t) = (0, 1)$ and requires one restoration-network evaluation (1-NFE).

The restored canonical patches are projected onto their corresponding full-resolution supports and combined with the padded observation by the frequency-preserving renderer. The padded region is then removed to recover the supplied $H \times W$ image extent. Inference uses BF16 precision.

C.1.3 Evaluation Protocols

Common paired-evaluation protocol. For the paired benchmark, all methods receive inputs standardized to 1008×1296 using the same aspect-ratio-preserving resize-and-pad transform. Each method then applies its own internal preprocessing as required. Predictions are mapped back through the inverse external transform to the original paired-image resolution, where all full-reference metrics are computed.

Evaluation rationale. Eyeglass reflections typically affect a relatively small portion of the image while directly degrading the appearance and local structure of the ocular region (Yi et al., 2025). Consequently, full-image PSNR can be dominated by the much larger unaffected facial and background regions: because PSNR is derived from spatially averaged mean-squared error, errors within a small but task-critical eye region may contribute only weakly to the final score. This limitation is consistent with the well-established shortcomings of MSE-based fidelity measures for perceptually important image distortions (Wang & Bovik, 2009). We therefore report Mask-PSNR and Mask-SSIM within the eyeglass region, together with Mixed-PSNR and Mixed-SSIM to measure restoration inside the affected region while accounting for preservation elsewhere.

Regional pixelwise agreement alone does not fully characterize the quality of ocular restoration. PSNR and SSIM can remain insensitive to perceptually meaningful changes in local appearance and structure (Zhang et al., 2018). We therefore additionally report Eye-ROI LPIPS for perceptual similarity and landmark-based errors for ocular geometry. This task-oriented evaluation is also motivated by prior eyeglass-reflection-removal studies, which have shown that reflections affect iris localization and facial landmark detection and have evaluated restoration through downstream iris or landmark accuracy (Sandhan & Young Choi, 2017; Zou et al., 2024). Eye-NME and Iris-NME directly quantify displacement of eye-contour and iris landmarks relative to the clean reference, complementing the regional reconstruction and perceptual metrics. Tab. 7 summarizes the paired and real-world evaluation metrics.

Table 7: Summary of evaluation metrics.

Metric	Region / Purpose	Better
Mask-PSNR / SSIM	Eyeglass region; local restoration fidelity	↑
Mixed-PSNR / SSIM	Eyeglass and non-eyeglass regions; restoration–preservation balance	↑
Out-PSNR	Outside eyeglass region; RGB preservation	↑
Out-HF-PSNR	Outside eyeglass region; high-frequency detail preservation	↑
Eye-ROI LPIPS	Binocular ROI; perceptual ocular fidelity	↓
Eye-NME	Eye contours; ocular geometric fidelity	↓
Iris-NME	Iris landmarks; iris geometric fidelity	↓
NIQE (Mittal et al., 2012)	Full real image; no-reference naturalness	↓
CLIP-IQA+ (Wang et al., 2023)	Full real image; perceptual quality	↑
User preference	Real-world study set; visual preference	↑

Masks and regional reconstruction metrics. RGB values are normalized to $[0, 1]$. Binary eyeglass masks are initialized using Grounded-SAM-2 (Ren et al., 2024): Grounding DINO uses the text prompt "glasses frames" with box and text thresholds of 0.25, and the detected components are segmented using the SAM 2.1 Hiera-Large checkpoint (Ravi et al., 2025). All detected eyeglass components are merged and manually corrected before evaluation. The resulting evaluation mask $\mathcal{M}$ defines the local region used for regional reconstruction metrics and is distinct from the soft renderer support A_{lens} in § 4.3. Let $\overline{\mathcal{M}}$ denote its complement within the valid image support.

Mask-PSNR is computed from the RGB mean-squared error within $\mathcal{M}$. To additionally account for preservation outside the eyeglass region, Mixed-PSNR combines the regional errors in the MSE domain:

$$E_{\text{mix}} = \gamma E_{\text{mask}} + (1-\gamma)E_{\text{out}}, \qquad \text{Mixed-PSNR} = 10\log_{10}\frac{1}{E_{\text{mix}}}, \qquad \gamma = 0.5, \tag{65}$$

where E_{mask} and E_{out} are the RGB mean-squared errors over $\mathcal{M}$ and $\overline{\mathcal{M}}$, respectively.

For the preservation-oriented ablations, we additionally report Out-PSNR and Out-HF-PSNR over the valid image region outside the eyeglass mask. Out-PSNR is obtained directly from the outside-region RGB error:

$$\text{Out-PSNR} = 10\log_{10}\frac{1}{E_{\text{out}}}, \tag{66}$$

where E_{out} is defined above.

To specifically evaluate preservation of fine image detail, Out-HF-PSNR compares high-frequency residuals outside the eyeglass mask. For an image I, we define

$$H_{\text{eval}}(I) = I - G_\sigma(I), \tag{67}$$

where G_σ denotes Gaussian filtering with a 7×7 kernel and $\sigma = 1.5$. Let Ω_{out} denote the valid image support outside $\mathcal{M}$. The high-frequency error is

$$E_{\text{out}}^{\text{HF}} = \frac{1}{3|\Omega_{\text{out}}|}\sum_{\mathbf{x}\in\Omega_{\text{out}}} \|H_{\text{eval}}(I_{\text{out}})(\mathbf{x}) - H_{\text{eval}}(I_0)(\mathbf{x})\|_2^2, \tag{68}$$

and we report

$$\text{Out-HF-PSNR} = 10\log_{10}\frac{1}{E_{\text{out}}^{\text{HF}}}. \tag{69}$$

Higher values indicate closer preservation of RGB appearance and high-frequency content, respectively, outside the eyeglass region.

For SSIM, we compute a full grayscale SSIM map using a 7×7 window. The mask and complementary support are each eroded by three pixels before regional averaging so that the local SSIM window remains within the corresponding region. Mask-SSIM is the mean SSIM over the eroded eyeglass support. Let S_{mask} and S_{out} denote the regional SSIM means; we report

$$\text{Mixed-SSIM} = \gamma S_{\text{mask}} + (1-\gamma)S_{\text{out}}, \qquad \gamma = 0.5. \tag{70}$$

Ocular perceptual and geometric fidelity. Regional PSNR and SSIM quantify reconstruction agreement within the task-relevant area, while Eye-ROI LPIPS and landmark-based errors characterize perceptual and geometric changes to the ocular structures themselves. Eye-ROI LPIPS is evaluated over a binocular region defined from clean-reference MediaPipe landmarks (Lugaresi et al., 2019). The joint eye bounding box is enlarged by 25% horizontally and 35% vertically, and the same crop is applied to the prediction and reference. Both crops are aspect-preservingly resized and reflect-padded to 128×128, after which LPIPS is computed using the AlexNet backbone.

For geometric fidelity, MediaPipe landmarks with iris refinement are independently detected from the clean reference and restored image. Eye-NME measures displacement of the left- and right-eye contour landmarks, while Iris-NME measures displacement of the refined iris landmarks. Following standard face-alignment evaluation practice (Kumar et al., 2020), we report normalized mean error (NME) using an inter-ocular normalization distance. For landmark set $\mathcal{K}$,

$$\mathrm{NME}_{\mathcal{K}} = \frac{100}{|\mathcal{K}|} \sum_{j \in \mathcal{K}} \frac{\|\widehat{\mathbf{p}}_j - \mathbf{p}_j\|_2}{\|\mathbf{p}_{33} - \mathbf{p}_{263}\|_2}, \tag{71}$$

where $\mathbf{p}_j$ and $\widehat{\mathbf{p}}_j$ are landmarks detected from the clean reference and restored image, respectively. The normalization distance is measured between reference landmarks 33 and 263, corresponding to the two outer eye corners. Eye-NME and Iris-NME are reported as percentages.

Downstream iris localization. Building on the iris-localization evaluation used in Anti-Glare (Sandhan & Young Choi, 2017), we evaluate iris localization on the reflected inputs and restoration outputs. For each paired sample, iris landmarks detected from the clean image are converted into left- and right-iris circles and used as reference targets. The same iris-localization pipeline is independently applied to the reflected input and to the output of each restoration method.

For each eye, localization accuracy is measured by the intersection-over-union (IoU) between the predicted and reference iris circles. An eye is considered correctly localized when its circle IoU is at least τ. At the image level, a sample is counted as successful when both eyes satisfy the threshold. We report the bilateral localization success rate across IoU thresholds and the minimum IoU of the two eyes. All methods are evaluated using the same downstream detector without downstream fine-tuning or method-specific adaptation.

Real-world metrics and blind user study. We evaluate all 458 Ocu-R images at their native resolutions using NIQE (Mittal et al., 2012) and CLIP-IQA+ (Wang et al., 2023). Both metrics are computed using the default PyIQA configurations from the final restored images, without additional manual resizing or normalization.

We additionally conduct a blind multi-way preference study on 30 images randomly sampled from Ocu-R using a fixed random seed of 2026. All 133 participants evaluate the same 30 images, yielding $133 \times 30 = 3{,}990$ selections. For each trial, the original reflected input is displayed above the outputs of all nine compared methods. Method identities are hidden, and the presentation order of the nine outputs is independently randomized for each participant and image. Participants select exactly one result, with no tie option, based jointly on reflection-removal effectiveness and preservation of the original image content. All 30 trials are required for submission, and all 133 submitted questionnaires are complete. Informed consent is obtained before participation.

We report each method's overall preference share over all 3,990 selections and its image wins, defined as the number of test images on which it receives the largest number of participant selections. No image-level ties occur. Uncertainty in the overall preference share is estimated using a two-way bootstrap over participants and images. For each of $5{,}000$ bootstrap replicates, 133 participants and 30 images are independently resampled with replacement, and the preference share is recomputed over the resulting crossed responses. The 2.5th and 97.5th percentiles define the reported 95% confidence interval.

C.1.4 Baseline Configurations

Training and checkpoint selection. All learned restoration baselines are retrained or fine-tuned on the OcuBench training data using their respective training recipes and initialization strategies. When a method defines a validation protocol, its checkpoint is selected according to that protocol using a validation subset drawn from the OcuBench training data. Otherwise, the final available checkpoint is used.

Table 8: Method-specific inference configurations. Spatial dimensions follow the W×H convention.

Method	Internal input handling	Key inference settings	Precision
Stable Diffusion 1.5	Full-image resize to 992×768; masked inpainting	DDIM; 50 configured / 35 executed; strength 0.7; CFG 6.0	FP16
ER^{2}Net	Short-side resize to 256; dimensions aligned to multiples of 16	Single-pass restoration	FP32
DSRNet	Dimensions aligned to multiples of 32; 768×768 tiles	64-px overlap; overlapping-tile reconstruction	FP32
Stable Diffusion 3.5	Full-image resize to 512×512; masked inpainting	FlowMatchEulerDiscreteScheduler; 28 configured / 15 executed; strength 0.55; CFG 4.5	BF16
DSIT	Reflect-padding to multiples of 32; 384×384 tiles	64-px overlap; overlapping-tile reconstruction	FP32
RDNet	320×320 tiles; boundary padding when required	64-px overlap; Hann-weighted fusion	FP32
ReflexSplit	Reflect-padding to multiples of 32; 768×768 tiles	64-px overlap; overlapping-tile reconstruction	FP32
DL2G	Full-image resize to 256×256	DEM–LSDM–GCR; 64×64 LSDM patches, stride 16; 100 DDIM steps	FP16 autocast
OcuFlow	Edge padding; 256×256 logical domain; 24×24 canonical patches	$N = 256$ adaptive leaves; 1-NFE; native-resolution synthesis	BF16 autocast

We compare against Stable Diffusion 1.5 (Rombach et al., 2022), ER^{2}Net (Zou et al., 2024), DSRNet (Hu & Guo, 2023), Stable Diffusion 3.5 (Esser et al., 2024), DSIT (Hu et al., 2024), RDNet (Zhao et al., 2025), ReflexSplit (Lee et al., 2026), and DL2G (Yi et al., 2025). Their method-specific inference configurations are summarized in Tab. 8.

Diffusion baseline details. For Stable Diffusion 1.5, we use the pretrained Stable Diffusion 1.5 weights together with our ReGlass-IP Lite adapter. The positive prompt is *"Remove the reflection on the eyeglasses."*, and the negative prompt is *"low quality, worst quality, blurry, distorted face, deformed eyes, extra eyes, bad anatomy, artifacts, watermark, text"*.

For Stable Diffusion 3.5, we use the pretrained Stable Diffusion 3.5 Medium weights with our task-adapted ControlNet. The positive prompt is *"Remove the reflection on the eyeglasses."*, and the negative prompt is *"reflection, glare, specular highlight, white streak, bright spot, lens reflection"*. The task-adapted ControlNet is evaluated at its 512×512 training resolution. Both diffusion baselines use a fixed base seed of 42 with deterministic per-image seed offsets.

C.1.5 Efficiency Protocol

All methods are benchmarked on the same NVIDIA RTX 4090 using the standardized paired-benchmark inputs defined in § C.1.3. The common benchmark preprocessing is completed before timing. Timing covers the complete method-specific inference pipeline, including internal resizing, padding, tiling, sampling, preprocessing, and postprocessing. Three warm-up runs are excluded.

Runtime is reported as the average wall-clock time per image under sustained inference. Each method uses its normal CPU–GPU parallelism and asynchronous execution during benchmarking. File encoding and disk I/O are excluded.

OcuFlow is evaluated with batch size 1, BF16 precision, and 1-NFE. Its timed pipeline includes ocular-geometry extraction, adaptive-grid construction, native-resolution patch sampling, canonical resampling, one-step pMF restoration, partition-of-unity projection, frequency-preserving synthesis, and final cropping. Ocular-geometry extraction uses four asynchronous CPU workers and runs concurrently with GPU inference.

C.2 Additional Quantitative Analysis

Complementary restoration and ocular gains. The results in Tab. 2 show that OcuFlow improves restoration quality, ocular perceptual fidelity, and geometric consistency simultaneously rather than optimizing a single reconstruction criterion. Among the seven paired-reference metrics, OcuFlow ranks first on six and second only on Mixed-SSIM, where it differs from the best result by 0.0004 (0.7074 versus 0.7078). Compared with DSIT, the strongest competing baseline across the four reconstruction metrics, OcuFlow improves Mask-PSNR by 0.5654 dB, Mixed-PSNR by 0.1889 dB, and Mask-SSIM by 0.0155, while reducing Eye-ROI LPIPS by 29.2%, Eye-NME by 8.7%, and Iris-NME by 6.7%. Relative to RDNet, the closest competitor in Eye-ROI LPIPS, OcuFlow further reduces Eye-ROI LPIPS by 10.9%, Eye-NME by 10.2%, and Iris-NME by 9.5%, while improving Mask-PSNR and Mixed-PSNR by 1.0366 dB and 0.6277 dB, respectively. These results show that the gains in regional reconstruction are accompanied by improved ocular appearance and geometry rather than obtained at their expense.

Quality–efficiency trade-off. OcuFlow requires 56.4551 ms/image of steady-state wall-clock time and 1.04 GB of GPU memory. Relative to ER^{2}Net, the second-fastest baseline, OcuFlow is approximately $4.3\times$ faster. Compared with DSIT, OcuFlow is approximately $54.8\times$ faster and uses 64.9% less GPU memory while improving six of the seven paired-reference metrics. Relative to RDNet,

Table 9: No-reference evaluation on 458 independent real-world images. The reflected input is included for reference but excluded from restoration-method ranking. Best and second-best restoration results are shown in **bold** and underlined, respectively.

Method	NIQE ↓	CLIP-IQA+ ↑
Input	5.2583	0.6556
Stable Diffusion 1.5	9.7805	0.6127
ER^{2}Net	12.8999	0.5568
DSRNet	<u>5.5638</u>	0.6312
Stable Diffusion 3.5	11.4307	0.5737
DSIT	5.7683	0.6265
RDNet	9.3812	**0.6442**
ReflexSplit	5.6417	0.6128
DL2G	12.7582	0.5473
Ours	**5.2369**	<u>0.6419</u>

Table 10: Blind user study on 30 images randomly sampled from the 458-image real-world set with seed 2026. All 133 participants completed all 30 trials, yielding 3,990 selections. Image Wins counts images on which a method receives the largest number of selections.

Method	Votes	Preference (%) ↑	Image Wins ↑
DSIT	433	10.85	3/30
ReflexSplit	276	6.92	1/30
ER^{2}Net	220	5.51	0/30
Stable Diffusion 1.5	89	2.23	0/30
DSRNet	84	2.11	0/30
RDNet	78	1.95	0/30
DL2G	75	1.88	0/30
Stable Diffusion 3.5	49	1.23	0/30
Ours	**2686**	**67.32**	**26/30**

it is approximately $47.9\times$ faster and uses 30.2% less GPU memory while outperforming RDNet on all seven paired restoration and ocular metrics. These results reflect the fixed adaptive-token budget, one-step pMF restoration, and lightweight native-resolution synthesis used by OcuFlow.

Real-world evaluation. The main paper summarizes no-reference image quality and human preference on the independent real-world set in Fig. 5. For completeness, Tabs. 9 and 10 provide the corresponding numerical results. The evaluation protocols, including the sampling and statistical analysis of the user study, are detailed in § C.1.3.

No-reference image quality. As shown in Tab. 9, OcuFlow achieves the lowest NIQE among all restoration methods, improving from 5.5638 for the second-best DSRNet to 5.2369. For CLIP-IQA+, RDNet obtains the highest restoration score of 0.6442, while OcuFlow ranks second at 0.6419, a difference of only 0.0023. Notably, the reflected input itself receives a CLIP-IQA+ score of 0.6556, higher than any restored output despite retaining the target reflection artifacts. This observation illustrates that generic full-image no-reference metrics do not directly measure reflection-removal correctness. We therefore treat NIQE and CLIP-IQA+ as complementary indicators of global perceptual quality rather than task-specific measures of successful ocular restoration.

Blind user preference. The human study in Tab. 10 provides substantially stronger separation between methods. OcuFlow receives $2{,}686$ of the $3{,}990$ selections, corresponding to an overall preference share of 67.32%. The next most selected method, DSIT, receives 10.85%, giving OcuFlow a 56.47 percentage-point margin and approximately $6.2\times$ its preference share. OcuFlow receives the largest number of selections on 26 of the 30 images (86.7%), compared with three images for DSIT and one for ReflexSplit. A two-way participant-by-image bootstrap with $5{,}000$ replicates gives a 95% confidence interval of $[59.32\%, 74.71\%]$ for OcuFlow's overall preference share. Its lower bound remains 9.32 percentage points above 50%, while the $26/30$ image-win rate shows that the aggregate preference is distributed broadly across the evaluated examples rather than concentrated on only a small subset.

Cross-protocol consistency. The paired-reference benchmark, generic no-reference metrics, and blind user study capture complementary aspects of restoration. The paired benchmark measures regional reconstruction fidelity, ocular perceptual similarity, and landmark consistency against paired references; NIQE and CLIP-IQA+ characterize global perceptual quality on the independent real-world set; and the blind study measures task-level human preference under simultaneous consideration of reflection removal and content preservation. OcuFlow performs strongly across these protocols, while the relatively modest separation under generic NR-IQA and the substantially larger separation in human preference demonstrate that no single metric family fully characterizes eyeglass-reflection removal. Together with the Eye-ROI LPIPS and ocular NME results, these findings motivate reporting complementary regional, geometric, no-reference, and human-perceptual criteria.

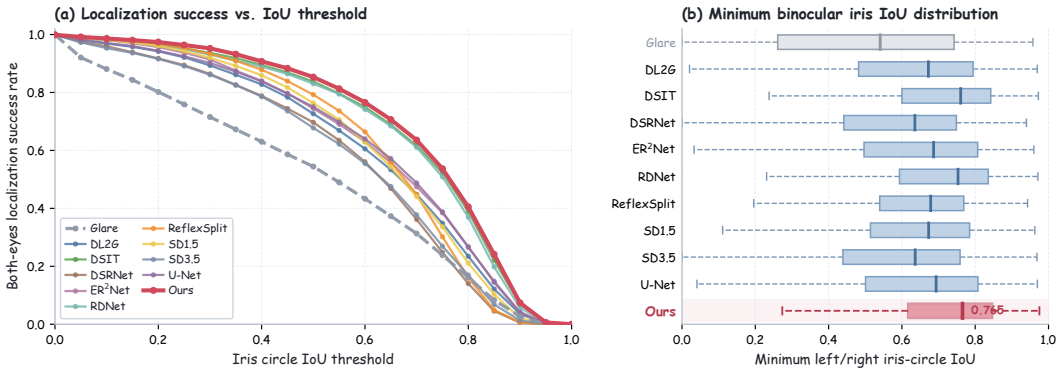

Figure 12: Downstream iris-localization evaluation. **(a)** Both-eyes localization success rate under different iris-circle IoU thresholds. **(b)** Distribution of the minimum left/right iris-circle IoU for each method. Glare denotes the original reflected input without reflection removal, while OcuFlow is highlighted in red.

C.3 Downstream Iris Localization

Beyond restoration fidelity, we further examine whether eyeglass-reflection removal improves the reliability of a downstream ocular-analysis task. Specifically, we evaluate iris localization using a fixed iris-landmark detector, without method-specific adaptation or downstream fine-tuning. The goal is to assess whether the restored ocular structures are not only visually plausible, but also sufficiently consistent for subsequent vision processing.

For each paired sample, iris circles estimated from the clean image are used as pseudo-ground truth. The same detector is independently applied to the reflected input and to the output of each restoration method. Let $\mathrm{IoU}_{\mathrm{left}}$ and $\mathrm{IoU}_{\mathrm{right}}$ denote the circle IoUs of the two eyes. We define the image-level localization score as $s_{\mathrm{iris}} = \min\left(\mathrm{IoU}_{\mathrm{left}}, \mathrm{IoU}_{\mathrm{right}}\right)$, such that an image is considered successfully localized at threshold τ only when $s_{\mathrm{iris}} \geq \tau$. This conservative criterion requires both eyes to be accurately localized.

As shown in Fig. 12(a), eyeglass reflections substantially degrade downstream iris localization, with the original reflected input exhibiting a markedly lower success rate over a broad range of IoU thresholds. Reflection removal recovers a large portion of this lost localization performance, while OcuFlow remains among the strongest methods throughout the moderate-to-high threshold range. At the default threshold $\tau = 0.5$, OcuFlow achieves a both-eyes localization success rate of 85.30%, compared with 54.43% for the reflected input and 83.78% for the strongest competing method, DSIT. This corresponds to a 30.87 percentage-point improvement over the unprocessed reflected input. Fig. 12(b) further examines the distribution of the minimum left/right iris IoU, which reflects the localization quality of the more difficult eye in each image. OcuFlow exhibits a clear rightward shift relative to the reflected input and achieves a median minimum IoU of 0.765. This distributional improvement indicates that the downstream gain is not restricted to a single operating threshold, but instead reflects more consistently recovered iris geometry across the evaluation set.

These results complement Eye-NME and Iris-NME by measuring functional downstream utility rather than geometric error alone. While the landmark-based metrics quantify agreement with the clean reference, the downstream localization experiment directly evaluates whether a fixed ocular-analysis pipeline can reliably operate on the restored images. The results therefore show that reflection removal restores task-relevant iris structure, with OcuFlow providing strong overall downstream localization performance.

C.4 Additional Qualitative Results

We provide additional paired qualitative comparisons to complement Fig. 6. The examples cover diverse subjects, eyeglass styles, reflection intensities, and ocular appearances. Each comparison includes the reflected input, clean reference, OcuFlow, and the competing methods evaluated in the



Figure 13: Additional qualitative comparisons on held-out paired OcuBench examples. Enlarged ocular regions are shown below the corresponding full-image comparisons to highlight residual reflections, iris and eyelid structure, and local detail preservation.



Figure 14: Additional qualitative comparisons on held-out paired OcuBench examples. Methods are shown in the same order as in Fig. 13.

main paper. Enlarged ocular crops are shown to facilitate inspection of residual reflections, ocular structure, and local appearance preservation.

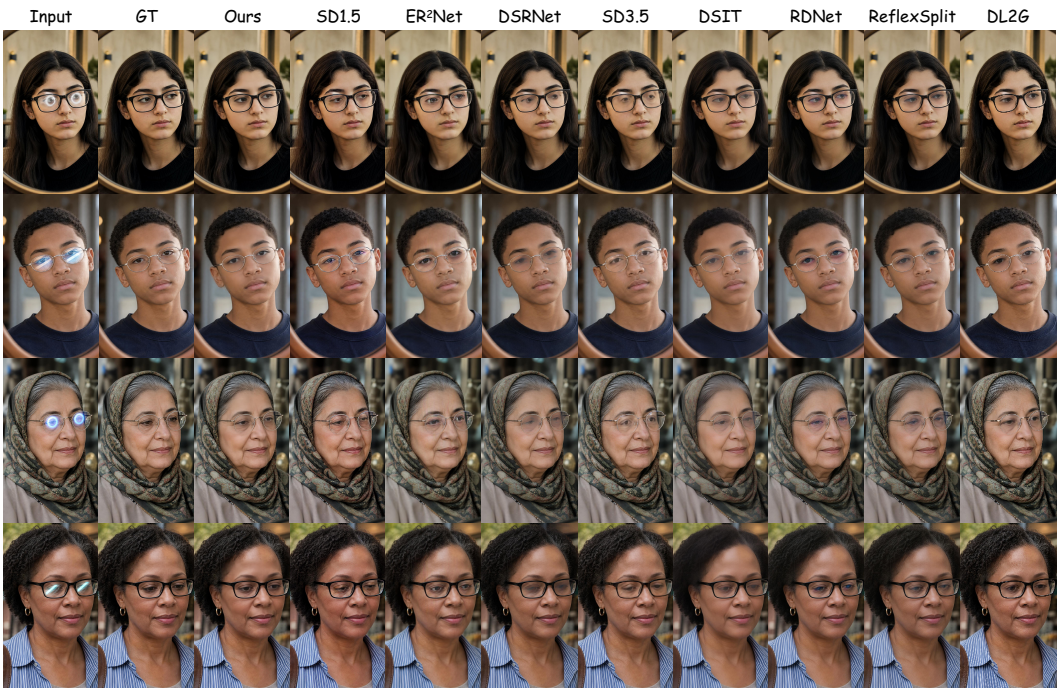

Figure 15: Further qualitative comparisons on held-out paired OcuBench examples, covering additional subjects, reflection patterns, and ocular appearances.

C.4.1 Qualitative comparison.

Across Figs. 13–15, the competing methods exhibit several recurring types of restoration error. ER^{2}Net (Zou et al., 2024) and DSRNet (Hu & Guo, 2023) often suppress the strongest reflection response while leaving residual glare or reduced contrast within the lens region. Under stronger occlusion, the recovered iris and eyelid boundaries can remain less distinct than in the clean reference. DSIT (Hu et al., 2024) and ReflexSplit (Lee et al., 2026) generally remove a larger fraction of the visible reflection, but several examples exhibit smoother ocular appearance and attenuated local contrast. RDNet (Zhao et al., 2025) frequently retains sharper local structure, although blue or purple chromatic residuals remain around the iris or lens region in multiple cases. Stable Diffusion 1.5 (Rombach et al., 2022) and Stable Diffusion 3.5 (Esser et al., 2024) can substantially alter local color, contrast, or facial appearance beyond the original reflection pattern. DL2G (Yi et al., 2025) preserves global facial appearance well in several examples but can introduce locally inconsistent structures around the eyeglass rims and eyelid boundaries.

These differences become particularly visible for saturated highlights, broad colored reflections, and glare overlapping the pupil or iris. In such cases, incomplete suppression leaves visible contamination, while stronger correction can introduce smoothing, chromatic shifts, or local structural changes. Across the shown examples, OcuFlow more closely follows the clean references in both the spatial extent of the correction and the recovered ocular appearance. Localized glare is removed while eyeglass boundaries and surrounding facial texture remain stable, and strongly colored reflections are restored with fewer persistent chromatic residuals. Under severe ocular occlusion, the recovered iris position, eyelid contour, and local eye contrast remain closer to the clean reference than the smoother or more strongly altered outputs observed in several competing methods. These comparisons illustrate the restoration–preservation trade-off in eyeglass reflection removal: effective suppression of corrupted content should be accompanied by preservation of ocular geometry and reliable local detail.



Figure 16: Ocular-prior-guided selective restoration. In the upper example, the eyeglasses are displaced away from the visible eyes, and OcuFlow leaves the lens appearance largely unchanged. In the lower example, a strong reflection overlaps the eye; OcuFlow applies stronger correction to the ocularly relevant region while making smaller changes to surrounding lens reflections.

C.4.2 Ocular-prior-guided selective restoration.

Fig. 16 illustrates the spatial selectivity induced by the ocular prior. In the upper example, the eyeglasses are displaced above the eyes, so the visible lens reflections do not overlap the ocular structures. OcuFlow consequently leaves the input largely unchanged. Several competing methods modify the lens regions more extensively, producing partial reflection suppression, local contrast changes, or irregular residual structures. In this example, DSIT and ReflexSplit introduce eye-like ghosting patterns within the lenses, while DL2G produces noticeable local texture distortion.

The lower example presents the complementary case. A strong bright reflection overlaps the left eye and obscures useful ocular detail, whereas the larger reflections near the upper parts of the lenses lie mostly outside the visible eye regions. OcuFlow applies stronger restoration around the affected eye while making smaller changes to the surrounding lens appearance. Several competing methods show broader or less localized changes: ER^{2}Net and DSRNet retain visible degradation around the ocular region, SD3.5 substantially alters the overall lens appearance, and DL2G leaves a strong bright residual near the eye together with local structural artifacts.

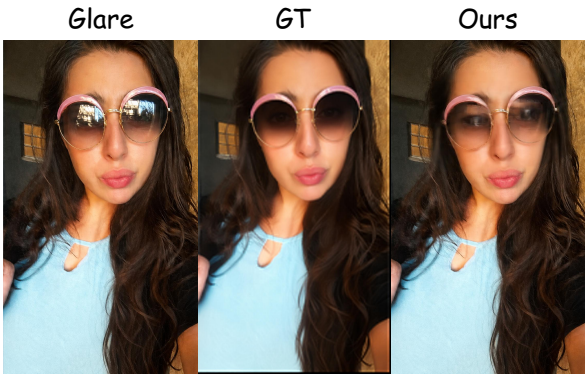

Figure 17: Failure case on sunglasses. OcuFlow reconstructs visible eye content behind the intentionally dark lenses, deviating from the clean reference.

Table 11: Additional ablations of data sampling, auxiliary supervision, and semantic conditioning. All variants are evaluated on the complete paired OcuBench test set.

Setting	Mask-PSNR ↑	Mixed-PSNR ↑	Mask-SSIM ↑	Mixed-SSIM ↑	Eye-ROI LPIPS ↓	Eye-NME (%) ↓	Iris-NME (%) ↓
Balanced Source Sampling	21.9821	23.3859	0.6588	0.7050	0.1125	1.4409	1.6286
w/o Eye Gradient	21.9915	23.3927	0.6592	0.7052	0.1125	1.4374	1.6241
w/o Semantic Prior	21.9883	23.3920	0.6593	0.7053	0.1122	1.4305	1.6171
w/o Multi-scale Reconstruction ($\mathcal{L}_x$)	21.3276	22.9074	0.6367	0.6939	**0.1015**	1.6258	1.9470
w/o Preserve Loss	21.9931	23.3945	0.6591	0.7052	0.1128	1.4360	1.6255
Ours (Natural Sampling)	**22.1054**	**23.4710**	**0.6633**	**0.7074**	0.1035	**1.4030**	**1.5863**

These examples illustrate how the ocular prior modulates the spatial extent of restoration according to the relationship between lens reflections and visible ocular structures. The resulting behavior reduces unnecessary changes to lens regions that do not interfere with eye visibility while concentrating stronger correction on reflections that overlap the eyes.

C.4.3 Failure cases.

OcuFlow can fail on sunglasses, where reduced ocular visibility is caused by intrinsic lens attenuation rather than reflection alone. As shown in Fig. 17, the ocular-restoration prior encourages reconstruction of visible eye content behind the dark lenses, leading to a result that deviates from the clean reference. This failure mode indicates that future models should better distinguish reflection-induced occlusion from intrinsic lens properties such as dark tint or strong attenuation.

C.5 Additional Ablation and Inference Analysis

The main paper studies the three high-level components of OcuFlow and the effect of training-data composition. We further examine training-time sampling and auxiliary objectives, semantic conditioning, the learned residual gates of the frequency-preserving renderer, and the behavior of multi-step inference. Unless otherwise specified, all variants are evaluated on the complete held-out paired OcuBench test set using the same metrics as in the main paper.

C.5.1 Training and Conditioning Ablations

Tab. 11 complements the component study in Tab. 3 by isolating several training-time choices not included in the main-paper ablation.

Natural sampling outperforms balanced source sampling across all seven reported metrics, indicating that retaining the empirical mixture of the training data is beneficial for the final model. Removing the eye-gradient objective, semantic prior, or preservation objective results in smaller but consistent degradation across reconstruction, perceptual, and ocular-geometry metrics. These components therefore provide complementary structural and preservation supervision in addition to the primary restoration objective.

Table 12: Ablation of the frequency-renderer gates. Both variants use the same native-resolution frequency decomposition and synthesis; the learned variant predicts content-adaptive residual corrections to the fixed mixing maps.

Setting	Mask-PSNR ↑	Mixed-PSNR ↑	Mask-SSIM ↑	Mixed-SSIM ↑	Eye-ROI LPIPS ↓	Eye-NME (%) ↓	Iris-NME (%) ↓	Out-PSNR ↑
Fixed Mixing Maps	21.8975	23.3077	0.6502	0.6967	**0.0999**	1.4075	**1.5821**	26.2977
Learned Residual Gates (Ours)	**22.1054**	**23.4710**	**0.6633**	**0.7074**	0.1035	**1.4030**	1.5863	**26.4147**

Table 13: Effect of the number of pMF network function evaluations (NFE).

NFE	Mask-PSNR ↑	Mixed-PSNR ↑	Mask-SSIM ↑	Mixed-SSIM ↑	Eye-ROI LPIPS ↓	Eye-NME (%) ↓	Iris-NME (%) ↓	Time (ms) ↓
1	**22.1054**	**23.4710**	**0.6633**	**0.7074**	0.1035	**1.4030**	**1.5863**	**56.4551**
2	21.4412	22.9939	0.6329	0.6917	0.0826	1.4695	1.7152	60.3832
4	21.3870	22.9516	0.6311	0.6907	**0.0822**	1.4638	1.7025	62.0009
8	21.3066	22.8885	0.6279	0.6888	0.0826	1.4614	1.7057	72.6149

The multi-scale reconstruction term $\mathcal{L}_x$ has a larger effect. Removing it reduces Mask-PSNR from 22.1054 to 21.3276 dB and Mixed-PSNR from 23.4710 to 22.9074 dB, while Mask-SSIM decreases from 0.6633 to 0.6367. Eye-NME and Iris-NME increase from 1.4030% and 1.5863% to 1.6258% and 1.9470%, respectively. Eye-ROI LPIPS, in contrast, decreases from 0.1035 to 0.1015. This divergence shows that improvement in a single perceptual metric does not necessarily correspond to better reference-aligned ocular reconstruction. The multi-scale reconstruction term contributes substantially to pixel-level fidelity and ocular geometric consistency.

C.5.2 Frequency-Renderer Gate Ablation

The main-paper ablation progresses from direct pMF prediction to residual rendering and frequency-preserving synthesis. Here, we isolate the content-adaptive residual gates within the frequency renderer. Both variants use the same frequency decomposition, lens support, and native-resolution synthesis. The fixed variant uses the handcrafted mixing maps directly, while the learned variant predicts residual corrections to these maps.

For this ablation, we additionally report Out-PSNR, defined as PSNR computed over the complement of the eyeglass evaluation mask, $\overline{\mathcal{M}}$.

Learning the residual gates improves Mask-PSNR by 0.2079 dB and Mixed-PSNR by 0.1633 dB, while increasing Mask-SSIM and Mixed-SSIM by 0.0131 and 0.0107, respectively. Eye-NME decreases from 1.4075% to 1.4030%, and Out-PSNR increases from 26.2977 to 26.4147 dB. Overall, the learned residual gates improve six of the eight reported metrics, indicating that content-adaptive refinement of the frequency mixing maps benefits both restoration within the eyeglass region and preservation outside it.

The fixed variant obtains slightly lower Eye-ROI LPIPS (0.0999 versus 0.1035) and Iris-NME (1.5821% versus 1.5863%). Together with the training ablations above, these results show that individual perceptual and landmark metrics can favor slightly different reconstruction trade-offs. The learned gates provide stronger masked and mixed reconstruction fidelity while retaining comparable ocular perceptual and geometric quality.

C.5.3 Multi-step Inference

OcuFlow uses a single pMF network function evaluation by default. We additionally evaluate repeated MeanFlow updates with $K \in \{1, 2, 4, 8\}$ evaluations. For K steps, the interval $[0, 1]$ is divided uniformly as

$$t_k = 1 - \frac{k}{K}, \qquad r_k = 1 - \frac{k+1}{K}, \qquad k = 0, \ldots, K-1, \tag{72}$$

and the state is updated according to

$$\mathbf{z}_{r_k} = \mathbf{z}_{t_k} - (t_k - r_k)\mathbf{u}_\theta(\mathbf{z}_{t_k}, r_k, t_k). \tag{73}$$

Tab. 13 shows that additional MeanFlow evaluations do not progressively improve the reference-aligned restoration metrics. The 1-NFE setting achieves the highest Mask/Mixed PSNR and SSIM

and the lowest Eye-NME and Iris-NME. Increasing to 2-NFE lowers Eye-ROI LPIPS from 0.1035 to 0.0826, but Mask-PSNR decreases by 0.6642 dB and both ocular NME measures increase. The 4-NFE setting yields the lowest Eye-ROI LPIPS at 0.0822, while PSNR and SSIM continue to decrease as the number of evaluations increases.

These results suggest that repeated MeanFlow updates change the reconstruction trade-off rather than progressively refining the 1-NFE solution. Multi-step inference produces lower LPIPS values but weaker pixel-level agreement and ocular geometric consistency with the clean reference. Runtime also increases with the number of evaluations, from 56.4551 ms/image at 1-NFE to 60.3832, 62.0009, and 72.6149 ms/image at 2-, 4-, and 8-NFE, respectively. We therefore use 1-NFE for the reported OcuFlow results.

C.5.4 Inference-time Breakdown and Pipeline Parallelism

The steady-state profile shows that the one-step pMF predictor accounts for 12.5873 ms (22.3%) of the measured 56.4551 ms/image processing interval, while the native-resolution renderer accounts for 14.1714 ms (25.1%). The largest individual contribution is the 17.6631 ms (31.3%) of CPU preparation that remains exposed after asynchronous overlap. Within the renderer, partition-of-unity projection accounts for 66.1% of the rendering cost, compared with 33.6% for frequency/decode operations and 0.3% for lens resizing, making projection the dominant rendering operation.

Table 14: Steady-state runtime breakdown of OcuFlow under sustained pipelined execution (56.4551 ms/image).

Component	Time (ms)	Share (%)
Exposed CPU-preparation wait	17.6631	31.2869
Native-resolution renderer	14.1714	25.1021
pMF	12.5873	22.2962
Patch preparation	5.3085	9.4030
PIL materialization	2.8555	5.0580
Batch packing	2.4048	4.2597
H2D transfer	0.5426	0.9612
Quantization + D2H	0.4070	0.7208
Other pipeline overhead	0.5150	0.9122
Steady-state total	**56.4551**	**100.0000**

Pipeline parallelism substantially reduces the exposed cost of ocular-geometry and preparation work. The aggregate CPU preparation work corresponds to 220.3822 ms/image. With four asynchronous preprocessing workers, the corresponding idealized preparation interval is approximately 55.0956 ms/image. The remaining inference-stage work accounts for approximately 38.7920 ms/image. Without CPU–GPU overlap, these two stages would yield an estimated processing interval of 93.8876 ms/image.

Under sustained pipelined execution, CPU preparation overlaps with GPU inference, reducing the measured steady-state processing interval to 56.4551 ms/image. Under ideal full overlap, throughput is bounded by the slower of the two pipeline stages, giving an idealized interval of 55.0956 ms/image. The measured interval is therefore within approximately 2.5% of this idealized bound.