# AFTER A DECADE: BRINGING SHADOW REMOVAL INTO THE REAL WORLD WITH AGENTIC TRAINING DATA

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Shadow removal looks nearly solved on established benchmarks, yet remains brittle in the real world. Models have advanced; the paired training data they rely on have barely changed in nearly a decade. The reason is simple: obtaining a shadow-free target requires removing the occluder while keeping the scene, camera, and illumination otherwise unchanged, making diverse paired data difficult to capture. Meanwhile, large shadow detection datasets already contain diverse real-world images and masks, but no shadow-free targets. To turn this abundant but incomplete data into paired supervision, we propose an offline agentic workflow combining physics-motivated generation, failure detection, feedback-driven retry, candidate selection, and deterministic correction. Using this workflow, we construct **AgenticShadow**, a dataset of 17,138 image-mask-target triplets spanning general scenes, faces, and remote sensing. Our construction workflow reduces Color Distribution Difference by 50.5% over previous shadow removal work, while training existing shadow removal models on AgenticShadow reduces cross-domain LAB RMSE by 19.7–37.5%.

## 1 INTRODUCTION

Shadow removal appears increasingly mature on standard paired benchmarks. On ISTD+ (Wang et al., 2018; Le & Samaras, 2019), several recent methods report relatively small differences among leading approaches with visually near-perfect results (Guo et al., 2023b; Xiao et al., 2024; Xu et al., 2025b; Lee et al., 2026). Yet this performance is limited to the relatively simple cases represented in the dataset and does not carry over reliably to diverse real-world images. Large cast shadows in outdoor scenes, mixed illumination in street imagery, shadows across faces, or aerial scenes with broad occlusions frequently lead to incomplete removal, color shifts, and structural artifacts. The reason is simple: the conditions encountered in practice are much broader than the paired data available for training.

Collecting a valid pair requires removing the shadow while keeping the camera, scene geometry, materials, exposure, and other illumination conditions unchanged. This is manageable in controlled capture, but difficult to scale to streets, faces, complex outdoor scenes, or aerial imagery. SRD (Qu et al., 2017) and ISTD (Wang et al., 2018) remain the principal paired datasets used for shadow removal. Later efforts improve different aspects of this data: ISTD+ (Le & Samaras, 2019) corrects color inconsistencies, while SRD+ (Inoue & Yamasaki, 2021) cleans and re-splits SRD. WSRD and WSRD+ (Vasluianu et al., 2023; 2024) provide high-resolution paired imagery through calibrated, close-range indoor capture. Nevertheless, paired acquisition still constrains both scale and domain coverage. The bottleneck is therefore not simply obtaining more shadow images, but obtaining their corresponding shadow-free targets under sufficiently varied conditions.

The main contribution of this work is to bridge this decade-old data gap. We do so with a physics-motivated agentic workflow using multimodal foundation models that converts real shadow image-mask pairs into paired training data. Rather than accepting a single prompted edit, the workflow treats generation as a proposal: it evaluates an initial candidate, detects major failures, triggers feedback-guided retries when needed, filters invalid outputs, and selects among valid candidates using visual and structural cues. A final deterministic regional correction reduces residual photometric



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

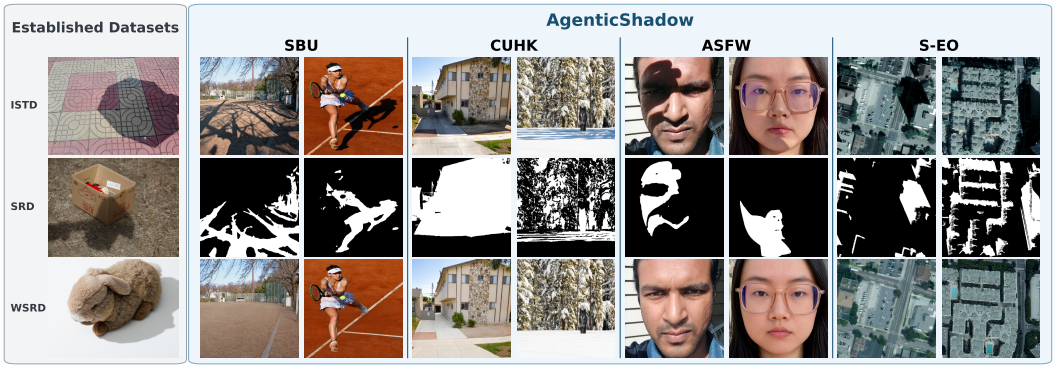



Figure 1: **From constrained paired capture to diverse real-world supervision.** ISTD (Wang et al., 2018) and SRD (Qu et al., 2017) largely reflect constrained paired acquisition with limited scene diversity, while WSRD (Vasluianu et al., 2023) remains centered on controlled, close-range object and surface interactions. AgenticShadow draws from SBU (Vicente et al., 2016), CUHK-Shadow (Hu et al., 2021), ASFW (Luo et al., 2026), and S-EO (Masquil et al., 2025), providing substantially broader scenes, viewpoints, shadow appearances, and application domains.

inconsistencies. This generate–evaluate–retry–select–correct loop produces pseudo shadow-free targets at scale while explicitly controlling the failure modes of generative editing. This process yields **AgenticShadow**: 17,138 image–mask–target triplets spanning general scenes, faces, and remote sensing. Crucially, the resulting dataset is not simply larger. It expands the distribution of paired supervision to larger and more spatially distributed shadows, lower-contrast cases, and substantially more heterogeneous correction requirements than established paired benchmarks.

Simply put, our proposed AgenticShadow dataset provides supervision for shadow distributions where paired capture is impractical. Our pseudo targets are not physical ground truth. In unconstrained scenes, however, an exact shadow-free counterpart may be inherently unobtainable: one cannot remove a building to eliminate its shadow for data collection, for example. Even conventional paired capture (Wang et al., 2018) provides only an approximation, since removing or repositioning an occluder can perturb illumination, alignment, or exposure. Our goal is therefore not exact physical reconstruction, but a controlled approximation that preserves scene content while restoring illumination. We enforce this through shadow-formation grounding in generation and evaluation, rejection of candidates that alter scene content or retain shadow-like structure, structural checks for added or lost detail, and regional photometric correction for residual brightness and color errors. These constraints do not guarantee a unique physical solution, but they make the constructed targets sufficiently controlled to serve as supervision. As we show in Section 5.1, the ablations support this design directly: adding these controls progressively improves target quality. Under generic prompting, retry and selection already reduce Color Distribution Difference (CDD), which compares RGB color distributions sampled along the shadow and non-shadow sides of annotated boundaries. Shadow-formation grounding improves every corresponding stage, and structural diagnostics and regional correction further reduce the final error.

We validate both the fidelity of the constructed targets and their value as training data. Without training on ISTD+ or access to the references during construction, our agentic workflow produces targets that closely match the color-corrected ISTD+ ground truth, achieving a LAB RMSE of 3.85. On the benchmark introduced in ShadowRemovalRefine (Hu et al., 2025), the workflow reduces CDD by 50.5% relative to the strongest previous method. More importantly, this supervision transfers across architectures, reducing cross-domain macro LAB RMSE by 19.7–37.5% for ShadowDiffusion (Guo et al., 2023b), HomoFormer (Xiao et al., 2024), and PhaSR (Lee et al., 2026).

Our main contributions are:

- We introduce AgenticShadow, to our knowledge the largest paired shadow removal benchmark constructed from diverse real-world imagery, with 17,138 triplets spanning general, facial, and remote-sensing domains.
- We develop a physics-motivated agentic data-construction workflow that converts shadow datasets with image-mask pairs into paired removal supervision through failure detection, conditional retry,

candidate filtering and selection, structural diagnostics, and deterministic regional correction, making AgenticShadow readily extensible to new data sources.
- We show that AgenticShadow supervision substantially improves cross-domain generalization across existing architectures and enables models with semantic and geometric priors—including a mask-free variant—to operate effectively under substantially broader real-world supervision.

## 2 Related Work

**Shadow Removal.** Early methods removed shadows using illumination invariance (Finlayson et al., 2002), matting (Wu et al., 2007), texture consistency (Liu & Gleicher, 2008), and relationships between paired shadowed and lit regions (Guo et al., 2013). Learning-based approaches later formulated the task as image restoration through multi-context reasoning (Qu et al., 2017), joint detection and removal (Wang et al., 2018), shadow-image decomposition (Le & Samaras, 2019; 2022), and hierarchical feature aggregation (Cun et al., 2020). Recent methods adopt transformer architectures for global-context or homogenized restoration (Guo et al., 2023a; Xiao et al., 2024), inpainting formulations (Li et al., 2023a), diffusion models (Guo et al., 2023b; Xu et al., 2025b), and semantic or geometric priors (Lee et al., 2026). Although these methods improve performance on established benchmarks, their generalization across substantially different scenes and shadow configurations remains underexplored.

**Shadow Datasets and Benchmarks.** Paired benchmarks including SRD (Qu et al., 2017) and ISTD (Wang et al., 2018) provide real shadow and shadow-free pairs but remain modest in scale and limited in their diversity of scenes, viewpoints, and shadow configurations. WSRD (Vasluianu et al., 2023) adds high-resolution pairs but uses a controlled, close-range setup focused on object and surface interactions. Synthetic datasets increase scale through generated shadows (Liu et al., 2021) or fully rendered virtual scenes under direct and indirect illumination (Xu et al., 2025a), but retain a domain gap with real images. Real-world datasets span general scenes in SBU (Vicente et al., 2016) and CUHK-Shadow (Hu et al., 2021), object-shadow associations in SOBA (Wang et al., 2020), documents (Das et al., 2020; Li et al., 2023b), faces (Liu et al., 2022; Luo et al., 2026), videos in ViSha (Chen et al., 2021), and remote-sensing imagery in S-EO (Masquil et al., 2025). However, large detection-oriented collections typically provide only shadow images and masks because capturing shadow-free counterparts is usually impractical. AgenticShadow bridges this gap by combining broad real-world coverage with unified image-mask-target supervision.

**Generative Editing for Data Construction.** Instruction-guided editors enable localized image modifications from natural-language requests (Brooks et al., 2023; Zhang et al., 2023; Sheynin et al., 2024; Zhao et al., 2024), but do not explicitly model shadow formation; their stochastic outputs may preserve shadows, alter scene content, or introduce physically inconsistent changes. Recent work couples image generation with model-based validation and correction for visual data construction (Huang et al., 2026; Park et al., 2026). We adapt this general paradigm to shadow removal through physics-motivated candidate generation and evaluation, failure-driven retry, candidate selection, and deterministic regional correction.

## 3 Agentic Dataset Construction

Given an image $I$ and shadow mask $M$, we construct a shadow-free target $\hat{I}$ while preserving scene geometry, surface detail, and non-shadow appearance. As summarized in Figure 2, GPT Image 2 (OpenAI, 2026) produces candidate edits, while GPT-5 mini (Singh et al., 2026) serves as the vision-language model (VLM) that evaluates major failures, provides feedback for retry, and selects among valid candidates. Detailed prompt designs and pipeline settings are provided in Appendix B.1. Direct validation against color-corrected ISTD+ references is provided in Appendix B.2.

**Shadow-formation grounding.** A shadow represents reduced illumination caused by light occlusion, not a change in the underlying material or scene content. We encode this distinction in both the generation and evaluation prompts. The editor is asked to recover the appearance of the same surface under normal illumination, while the evaluator rejects candidates that reinterpret the shadow as an object, material boundary, or missing content. We compare this physics-motivated formulation with an otherwise identical pipeline using generic shadow removal prompts.

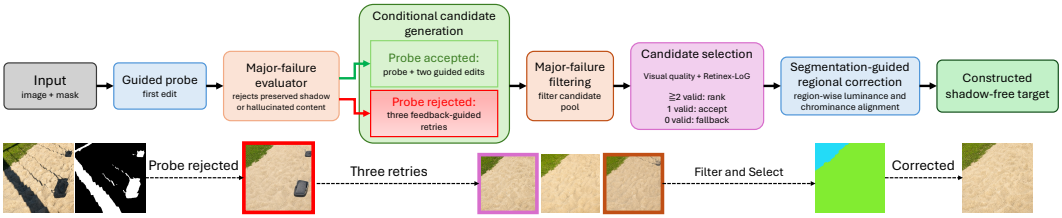



Figure 2: **Agentic dataset-construction pipeline.** A VLM evaluates an initial mask-guided probe, triggers additional guided edits or feedback-guided retries, filters major failures, and selects a candidate based on visual quality and Retinex-LoG diagnostics. Segmentation-guided regional correction produces the final shadow-free target. The bottom row illustrates a rejected-probe example.

**Probe-guided conditional generation.** The pipeline begins with a mask-guided probe. A major-failure evaluator determines whether the probe retains the shadow by treating it as object-like content or hallucinates new scene content within the masked region. If the probe passes, we retain it and generate two additional guided candidates. If it fails, we discard it and generate three feedback-guided retry candidates from the original image and mask.

**Candidate filtering and selection.** Every candidate entering the selection pool must pass the major-failure evaluator. When multiple candidates remain, the VLM selector compares their shadow removal quality, scene fidelity, detail preservation, and artifact severity, supported by Retinex-LoG structural diagnostics. Because raw gradients are sensitive to illumination, we apply Retinex (Jobson et al., 1997) in the log-luminance domain to suppress slowly varying illumination and use a Laplacian-of-Gaussian (LoG) (Marr & Hildreth, 1980) operator to measure the remaining local structure. For an image $X$ with luminance $Y_X$, its Retinex-LoG response is

$$F(X) = \left|\mathrm{LoG}\left(\log(Y_X + \epsilon) - G_\sigma * \log(Y_X + \epsilon)\right)\right|. \tag{1}$$

Here, $G_\sigma$ estimates smooth illumination, and subtracting it from log-luminance produces a reflectance representation. Let $\Omega_M$ denote the valid pixels obtained by excluding a narrow band around the shadow boundary. Before comparison, both response maps are divided by the 95th-percentile response of the input over $\Omega_M$ and clipped to $[0, 1]$, yielding $\widetilde{F}(I)$ and $\widetilde{F}(C)$. We then measure the structural discrepancy between candidate $C$ and input $I$ as

$$D_{\text{Retinex-LoG}}(C, I) = \mathrm{MAE}_{\Omega_M}\left(\widetilde{F}(C), \widetilde{F}(I)\right). \tag{2}$$

A lower discrepancy indicates better local structure preservation. We also measure the mean magnitudes of response decreases and increases to indicate lost and added detail. These diagnostics support, but do not replace, the selector's visual judgment. If one candidate passes, it is selected directly; if none pass, a fallback selector returns the least damaging candidate from the generated set.

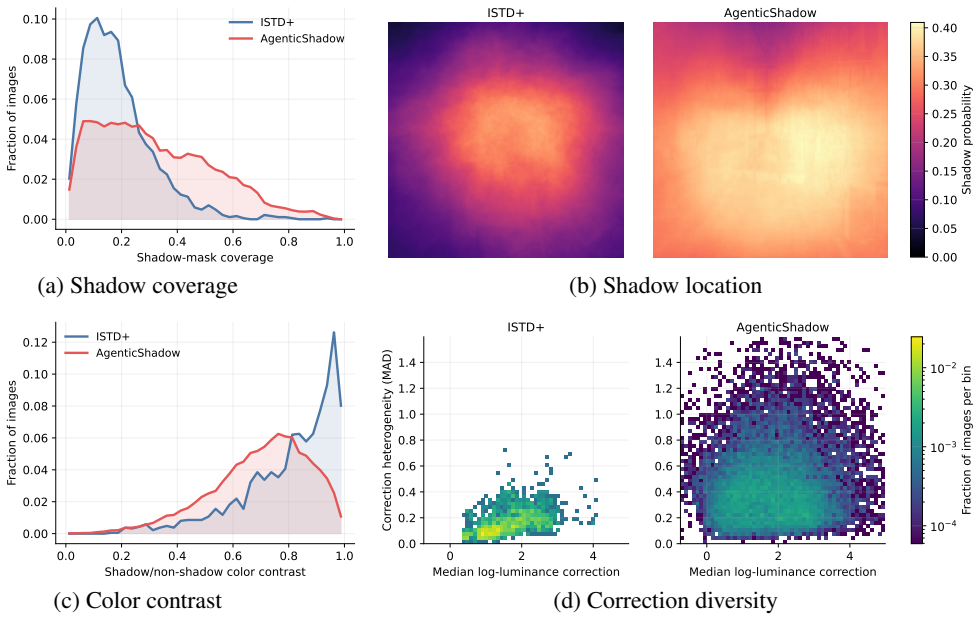
**Segmentation-guided regional correction.** The selected proposal may remove the shadow successfully while exhibiting broad or regional color shifts. We segment the proposal into class-agnostic regions using CropFormer (Qi et al., 2023). For each region, we compare the input and proposal only at reliable pixels outside a dilated shadow mask and estimate a constant offset from the channel-wise median of their YCbCr differences after outlier filtering. A local offset is applied when the region contains sufficient reliable pixels and improves low-frequency color agreement; otherwise, a global offset is used. Because this stage adjusts only brightness and chrominance, it does not explicitly copy textures, edges, or shadow structure from the input. The corrected image becomes the final target $\hat{I}$.

## 4 AGENTICSHADOW BENCHMARK

**Dataset composition.** AgenticShadow combines four complementary sources of real shadow images: SBU (Vicente et al., 2016), CUHK-Shadow (Hu et al., 2021), ASFW (Luo et al., 2026), and S-EO (Masquil et al., 2025). We apply the agentic construction workflow described in Section 3

Table 1: **Composition of AgenticShadow.** The four source domains provide complementary scene types and together form a benchmark of 17,138 image-mask-target triplets.

| | SBU | CUHK | ASFW | S-EO | **Total** |
|---|---|---|---|---|---|
| Domain | General | General | Facial | Remote sensing | – |
| Train | 3,974 | 8,335 | 928 | 880 | 14,117 |
| Test | 635 | 2,087 | 153 | 146 | 3,021 |
| **Total** | **4,609** | **10,422** | **1,081** | **1,026** | **17,138** |




(a) Shadow coverage (b) Shadow location

(c) Color contrast (d) Correction diversity

Figure 3: **Dataset diversity compared with ISTD+.** (a) Distribution of shadow-mask coverage. (b) Average spatial distribution of shadow masks. (c) Color contrast between shadow and non-shadow regions. (d) Joint distribution of median log-luminance correction and within-shadow heterogeneity. AgenticShadow contains larger and more spatially distributed shadows, more low-contrast cases, and substantially broader and more heterogeneous correction requirements.

to construct a shadow-free target for each retained image-mask pair. As summarized in Table 1, the resulting benchmark contains 17,138 image-mask-target triplets, divided into 14,117 training and 3,021 test examples. SBU and CUHK-Shadow provide diverse general scenes, while ASFW and S-EO extend the benchmark to facial shadows and remote-sensing imagery. Source processing, cleaning, splits, and retained counts are detailed in Appendix A.1.

**Benchmark diversity.** We compare AgenticShadow with ISTD+ (Le & Samaras, 2019) using four complementary statistics computed over each complete benchmark. Shadow coverage is the fraction of pixels inside the shadow mask, while the location heatmap averages masks after resizing them to common image coordinates. Following the analysis of CUHK-Shadow (Hu et al., 2021), color contrast is measured by the $\chi^2$ distance between normalized RGB histograms of shadow and non-shadow pixels. To characterize correction diversity, we compute the signed log-luminance change from each shadow image to its target within the shadow region; its median measures the typical correction, while its median absolute deviation (MAD) measures spatial heterogeneity.

As shown in Figure 3, AgenticShadow contains substantially larger shadows, with median coverage of $28.6\%$ versus $16.1\%$ for ISTD+. Its heatmap shows broader shadow locations, while its lower median color contrast ($0.724$ versus $0.849$) indicates greater coverage of subtle, low-contrast shadows. Finally, its median log-luminance correction increases from $1.51$ to $1.83$, while correction heterogeneity increases from $0.141$ to $0.395$. These results demonstrate that AgenticShadow covers more diverse

**(a) Quantitative comparison**

| Method | Mean | Std | Min | Max |
|---|---|---|---|---|
| SID (Le & Samaras, 2019) | 38.00 | 45.05 | 1.66 | 353.64 |
| SF (Guo et al., 2023a) | 33.63 | 52.74 | 0.41 | 427.34 |
| SD (Guo et al., 2023b) | 89.72 | 130.54 | 0.29 | 650.26 |
| I4S (Li et al., 2023a) | 31.57 | 48.39 | 0.43 | 402.44 |
| SRR (Hu et al., 2025) | 14.06 | 26.08 | 0.24 | 215.40 |
| **Ours-Physics** | **6.96** | **18.59** | **0.01** | **158.40** |

**(b) Pipeline ablation**

| Prompt | Variant | CDD |
|---|---|---|
| Generic | Probe | 12.07 |
| | Major-retry | 11.77 |
| | Visual selector | 10.33 |
| Physics | Probe | 9.51 |
| | Major-retry | 7.78 |
| | Visual selector | 7.45 |
| | Retinex-LoG | 7.07 |
| | Regional correction | **6.96** |

Table 2: **Agentic target-construction quality on SRR (Hu et al., 2025).** (a) Comparison with previous shadow removal methods. (b) Ablation of the agentic construction pipeline. CDD values are multiplied by $10^3$, agentic results are averaged over two independent runs, and lower is better.

and spatially varying shadow removal requirements than ISTD+. Additional construction and semantic-diversity statistics are provided in Appendix A.2.

**Evaluation protocol.** We evaluate shadow removal methods separately on the SBU, CUHK-Shadow, ASFW, and S-EO test sets. Because their sizes differ substantially, we report per-domain results and a macro average that assigns equal weight to each domain. We report ISTD+ separately as an established benchmark. Reconstruction quality is measured using PSNR, SSIM, MAE, and LAB RMSE. Checkpoint, inference, and metric details are provided in Appendix C.1.

## 5 Experiments and Results

**Implementation Details.** We implement the restoration experiments in PyTorch (Paszke et al., 2019) and train on four NVIDIA RTX A6000 GPUs. Our PF baseline combines a pretrained HomoFormer (Xiao et al., 2024) with frozen DINOv2-L/14 (Oquab et al., 2024) and Depth Anything V2-L (Yang et al., 2024) models, using input resolutions of 448 and 518 pixels, respectively. Unless otherwise stated, AgenticShadow-trained models use the natural concatenation of ISTD+ and AgenticShadow. PF is trained for 100 epochs with a total batch size of 16 using Adam (Kingma & Ba, 2017). The backbone and prior-module learning rates start at $10^{-4}$ and $3 \times 10^{-4}$ and decay to $10^{-6}$ under cosine scheduling. PF-MF follows the same configuration. Additional training details are provided in Appendix C.2, with model complexity and runtime reported in Appendix C.3.

**Evaluations.** We evaluate agentic-construction quality on the 399-image ShadowRemovalRefine benchmark (Hu et al., 2025) using Color Distribution Difference (CDD), which measures color discrepancy across annotated shadow boundaries; lower is better. We compare with SID (Le & Samaras, 2019), ShadowFormer (Guo et al., 2023a), ShadowDiffusion (Guo et al., 2023b), Inpaint4Shadow (Li et al., 2023a), and ShadowRemovalRefine. Because generative editing is stochastic, we evaluate each agentic configuration over two independent full-set runs and average the results. For cross-domain evaluation, we test official ISTD+ checkpoints of SID, ShadowFormer, Inpaint4Shadow, StableSR (Xu et al., 2025b), ShadowDiffusion, HomoFormer, and PhaSR (Lee et al., 2026) across all four AgenticShadow domains. We additionally compare ShadowDiffusion, HomoFormer, and PhaSR with counterparts trained on ISTD+ and AgenticShadow. We report per-domain and macro-average PSNR, SSIM, MAE, and LAB RMSE, with ISTD+ reported separately.

### 5.1 Agentic Target-Construction Quality

We assess both the final target quality and the contribution of each construction stage on the common normalized SRR benchmark described in Appendix B.3. As shown in Table 2(a), our final physics-motivated pipeline achieves a mean CDD of 0.006960, a 50.5% reduction relative to the strongest previous method, SRR. It also yields the lowest standard deviation and lowest maximum CDD, indicating that the workflow improves average target quality while reducing severe construction failures. See Appendix B.4 for qualitative analyses and Appendix B.5 for limitations.

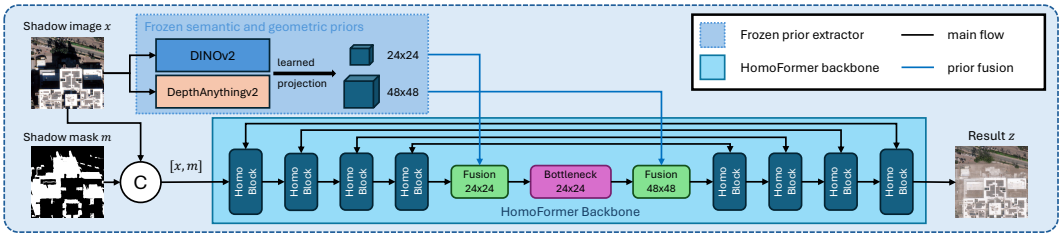



Figure 4: **Prior-Fusion (PF) baseline.** Frozen DINOv2 (Oquab et al., 2024) and Depth Anything V2 (Yang et al., 2024) extract semantic and geometric priors from the shadow image. Learned projections align these priors with intermediate features, and zero-initialized residual fusion incorporates them at the $24 \times 24$ bottleneck input and the first $48 \times 48$ decoder stage.

**(a) Dataset–prior interaction**

| | ISTD+ | | AgenticShadow | |
|---|---|---|---|---|
| Model | PSNR↑ | LAB↓ | PSNR↑ | LAB↓ |
| HomoFormer* | 32.53 | 1.92 | 18.85 | 8.43 |
| PF* | 32.95 | 1.87 | 18.70 | 8.55 |
| HomoFormer | 32.92 | 1.85 | 21.95 | 5.87 |
| PF | 32.74 | 1.86 | 22.50 | 5.55 |

**(b) Prior-component ablation**

| | AgenticShadow | | | |
|---|---|---|---|---|
| Variant (E50) | PSNR↑ | SSIM↑ | MAE↓ | LAB↓ |
| HomoFormer | 21.90 | 0.7161 | 15.36 | 5.912 |
| Depth only | 22.13 | 0.7259 | 14.96 | 5.769 |
| Semantic only | 22.46 | 0.7306 | 14.42 | 5.585 |
| PF | **22.47** | **0.7312** | **14.40** | **5.577** |

Table 3: **Analysis of the Prior-Fusion baseline.** (a) Interaction between training data and prior fusion, comparing ISTD+-only training with combined ISTD+ and AgenticShadow training; * denotes the former. (b) Matched E50 prior ablation initialized from HomoFormer E100 (Xiao et al., 2024). AgenticShadow results are four-domain macro averages.

Table 2(b) isolates the contributions of the construction pipeline. Under generic prompting, major-failure retry and visual selection reduce mean CDD from 0.012073 to 0.010334. Shadow-formation grounding consistently outperforms its generic counterpart at the probe, retry, and selection stages, showing that explicitly treating shadows as illumination effects improves both candidate generation and evaluation. Retinex-LoG evidence and regional correction further reduce CDD to 0.006960. Together, these results validate both the agentic control flow and its physics-motivated grounding.

### 5.2 Prior-Fusion Baseline

To examine whether broader paired supervision enables more effective use of external priors, we construct a **Prior-Fusion (PF)** baseline based on HomoFormer (Xiao et al., 2024). Given a shadow image $x$ and mask $m$, PF predicts a shadow-free image in one feed-forward pass while incorporating frozen semantic and geometric features, as illustrated in Figure 4. This deliberately simple design provides a controlled way to evaluate how training supervision affects the utility of external priors.

**Prior extraction and fusion.** We extract semantic features using DINOv2 (Oquab et al., 2024) and relative depth using Depth Anything V2 (Yang et al., 2024), keeping both models frozen. A learned $1 \times 1$ projection and spatial resizing align the semantic features, while a lightweight convolutional pyramid projects the depth map. Let $P_r^{\text{sem}}$ and $P_r^{\text{geo}}$ denote the resulting features at resolution $r \times r$. They are fused with HomoFormer tokens $H_r$ through

$$\widetilde{H}_r = H_r + \alpha_r \, \phi(P_r^{\text{sem}}) + \beta_r \, \phi(P_r^{\text{geo}}), \qquad r \in \{24, 48\}, \tag{3}$$

where $\phi(\cdot)$ converts feature maps into tokens, and $\alpha_r$ and $\beta_r$ are learned scalar strengths initialized to zero. Fusion is applied before the $24 \times 24$ bottleneck and at the first $48 \times 48$ decoder stage.

**Training and variants.** The prior extractors remain fixed, while the backbone, projections, and fusion strengths are optimized using the Charbonnier loss (Charbonnier et al., 1994). We additionally evaluate semantic-only and depth-only variants. PF-MF removes the mask input and its reinjection at the final decoder and output stages while retaining the same prior-fusion design.

Table 4: **Cross-domain shadow removal evaluation.** Official ISTD+ checkpoints are evaluated without adaptation, while † denotes training with both ISTD+ and AgenticShadow. We report per-domain LAB RMSE and equal-weight macro averages across the four AgenticShadow domains. ISTD+ is reported separately. Best and second-best results are **bolded** and underlined, respectively.

| | ISTD+ | AgenticShadow: LAB RMSE ↓ | | | | AgenticShadow: Macro | | | |
|---|---|---|---|---|---|---|---|---|---|
| Method | LAB ↓ | SBU | CUHK | ASFW | S-EO | PSNR ↑ | SSIM ↑ | MAE ↓ | LAB ↓ |
| *Official ISTD+ checkpoints* | | | | | | | | | |
| SID (Le & Samaras, 2019) | 2.89 | 7.25 | 10.86 | 7.48 | 10.23 | 18.28 | 0.6404 | 23.77 | 8.96 |
| ShadowFormer (Guo et al., 2023a) | 1.95 | 6.63 | 10.01 | 7.35 | 10.74 | 18.57 | 0.6592 | 23.36 | 8.68 |
| Inpaint4Shadow (Li et al., 2023a) | 2.77 | 7.12 | 9.72 | 7.58 | 10.08 | 18.59 | 0.6119 | 23.06 | 8.63 |
| StableSR (Xu et al., 2025b) | 2.03 | 7.32 | 8.84 | 10.01 | 10.93 | 17.90 | 0.6506 | 25.23 | 9.28 |
| ShadowDiffusion (Guo et al., 2023b) | 1.94 | 6.83 | 9.09 | 7.47 | 10.50 | 18.72 | 0.6563 | 22.67 | 8.47 |
| PhaSR (Lee et al., 2026) | 2.13 | 7.32 | 8.71 | 10.36 | 10.92 | 17.90 | 0.6556 | 24.75 | 9.33 |
| HomoFormer (Xiao et al., 2024) | 1.92 | 6.78 | 9.08 | 7.00 | 10.87 | 18.85 | 0.6597 | 22.36 | 8.43 |
| *With AgenticShadow supervision* | | | | | | | | | |
| ShadowDiffusion† | 2.06 | 5.84 | 7.40 | 5.97 | 7.98 | 20.47 | 0.6952 | 18.58 | 6.80 |
| PhaSR† | 2.11 | 4.92 | 6.19 | 5.31 | 6.90 | 22.06 | 0.7309 | 15.44 | 5.83 |
| HomoFormer† | **1.85** | 4.90 | 6.23 | 5.33 | 7.02 | 21.95 | 0.7181 | 15.25 | 5.87 |
| **PF** | 1.86 | **4.73** | **6.05** | **4.67** | **6.76** | **22.50** | **0.7314** | **14.36** | **5.55** |
| PF-MF | 1.95 | 4.75 | 6.11 | 4.97 | 6.78 | 22.36 | 0.7303 | 14.67 | 5.65 |

Table 3(a) shows that prior fusion alone is insufficient for generalization. Under ISTD+-only training, PF improves in-domain performance but not AgenticShadow results. With combined ISTD+ and AgenticShadow supervision, the same priors provide clear AgenticShadow gains while retaining comparable ISTD+ performance, indicating that broader supervision is needed to translate external priors into transferable improvements. Table 3(b) provides a matched E50 component analysis initialized from HomoFormer E100. Both priors independently improve the backbone across all metrics. Their combination performs best overall: semantic information provides the larger gain, while relative depth contributes complementary geometric context.

### 5.3 Cross-Domain Benchmark Evaluation

We first evaluate official ISTD+ checkpoints on the four AgenticShadow domains without adaptation, exposing their cross-domain limitations. We then retrain ShadowDiffusion, HomoFormer, and PhaSR using ISTD+ and AgenticShadow to measure the effect of broader supervision. PF follows the same combined-data setting as its HomoFormer backbone.

As shown in Table 4, official checkpoints obtain macro LAB RMSE values of 8.43–9.33. AgenticShadow supervision reduces this range to 5.83–6.80 for the three retrained architectures, demonstrating a substantial generalization benefit from the additional data. PF performs best across every AgenticShadow domain and all macro metrics while remaining competitive on ISTD+. Complete per-domain and cross-dataset results appear in Appendix D.1 and Appendix D.2.

Figure 5 provides a visual counterpart to these results. Official ISTD+ checkpoints frequently retain shadows or introduce color shifts on unfamiliar scenes, while AgenticShadow-trained counterparts produce substantially cleaner outputs across general, facial, and remote-sensing imagery. PF further improves target agreement in the selected cases, consistent with the quantitative evaluation.

### 5.4 Mask-Free and Generalization

We evaluate PF-MF to assess whether AgenticShadow supervision and external priors reduce reliance on annotated masks. As shown in Table 4, it closely matches PF, trailing by only 0.14 dB PSNR and 0.10 LAB RMSE. Figure 6 shows video generalization: PF produces cleaner removal than pretrained HomoFormer on ViSha (Chen et al., 2021) using masks, while PF-MF operates directly on SBU-TimeLapse (Le & Samaras, 2022) frames without masks and outperforms SP+M+I-Net. These results support video use where frame-level shadow masks are unavailable; complete videos and further qualitative comparisons are provided in Appendix D.3.

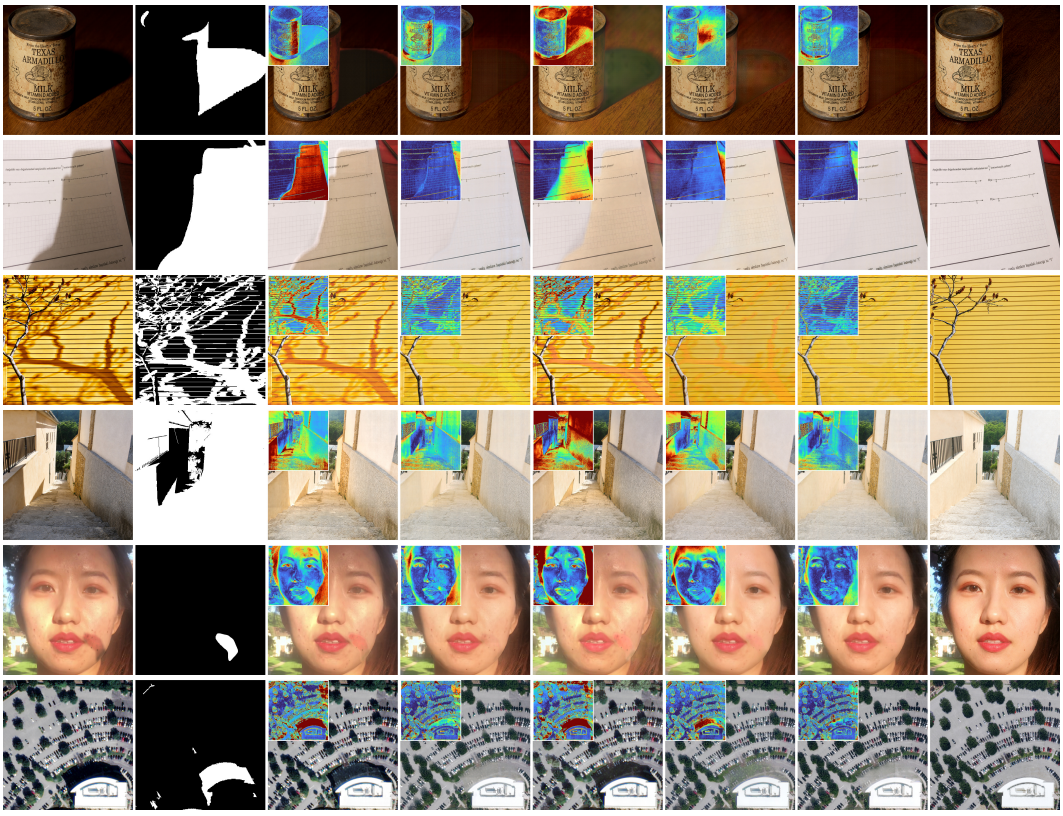
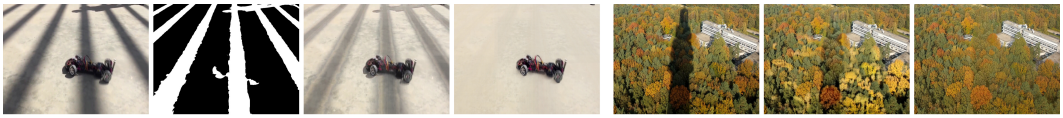



Figure 5: **Qualitative cross-domain comparison.** We compare HomoFormer (Xiao et al., 2024) and PhaSR (Lee et al., 2026), their counterparts trained with AgenticShadow (†), and PF. Models pretrained on ISTD+ (Le & Samaras, 2019) often leave residual shadows or color shifts on diverse AgenticShadow scenes, while training with AgenticShadow substantially improves removal. PF consistently produces outputs closer to the targets than the AgenticShadow-trained counterparts. Colored insets show mean absolute RGB error using a shared scale.




(a) PF on ViSha (b) PF-MF on SBU-TimeLapse

Figure 6: **Generalization to shadow videos.** (a) PF is compared with pretrained HomoFormer (Xiao et al., 2024) on ViSha (Chen et al., 2021) using the provided masks. (b) PF-MF is compared with SP+M+I-Net on SBU-TimeLapse (Le & Samaras, 2022) without input masks. In both settings, PF and PF-MF produce cleaner and more complete shadow removal than the compared methods.

## 6 Conclusion

We introduced AgenticShadow, a benchmark of 17,138 image-mask-target triplets constructed from diverse real shadow images through a physics-motivated agentic workflow. Our evaluation validates the constructed targets, shows that AgenticShadow offers substantially greater cross-domain diversity than existing paired benchmarks, and demonstrates that its supervision consistently improves representative shadow removal architectures. We also study a simple Prior-Fusion baseline, which achieves the strongest overall performance across the four AgenticShadow domains; its mask-free variant remains competitive and extends shadow removal to settings without annotated masks. Together, these results establish AgenticShadow as a resource for robust real-world shadow removal.

### AI USE STATEMENT

Generative AI is an explicit component of the proposed dataset-construction workflow. GPT Image 2 is used for image normalization and candidate target generation, while GPT-5 mini performs failure evaluation, feedback-guided retry, and candidate selection.

Beyond this role, we used generative AI tools to provide feedback on experimental setup, assist with implementing author-designed methods, support qualitative analysis, and interpret results. We did not use generative AI to develop the conceptual framework, formulate mathematical claims, or propose or refine the research hypotheses. Theoretical modeling, mathematical proofs, and translation were not applicable. Additionally, we used generative AI tools to suggest experimental parameters; create or edit supporting code, figures, tables, and other artifacts; verify information and format references; and assist with drafting and polishing the manuscript.

We reviewed all AI-assisted work. The authors inspected and executed AI-assisted code, verified quantitative results against recorded outputs, checked citations against their original sources, and reviewed all generated figures and text. We take responsibility for the final content of this work, including text, claims, code, data, and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

AgenticShadow is constructed from four published research datasets; no new human-subject data or private data were collected. We record the source dataset for each constructed sample and will release the constructed data in accordance with the licenses and terms of the original datasets. Because the targets are generated rather than physically captured, they may contain artifacts and should not be treated as physical ground truth; these limitations are documented in the appendix.

### REPRODUCIBILITY STATEMENT

The main paper describes the dataset-construction workflow, benchmark, training, and evaluation protocol. The appendix provides source processing and cleaning details, prompts and pipeline settings, implementation and inference configurations, and complete experimental results. We will release the construction and evaluation code, dataset manifests and targets, and trained checkpoints, subject to the licenses of the source datasets.

## REFERENCES

Tim Brooks, Aleksander Holynski, and Alexei A. Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 18392–18402, 2023. doi: 10.1109/CVPR52729.2023.01764.

Pierre Charbonnier, Laure Blanc-Féraud, Gilles Aubert, and Michel Barlaud. Two deterministic half-quadratic regularization algorithms for computed imaging. In *Proceedings of the IEEE International Conference on Image Processing*, volume 2, pp. 168–172, 1994. doi: 10.1109/ICIP.1994.413553.

Zhihao Chen, Liang Wan, Lei Zhu, Jia Shen, Huazhu Fu, Wennan Liu, and Jing Qin. Triple-cooperative video shadow detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 2714–2723, 2021. doi: 10.1109/CVPR46437.2021.00274.

Xiaodong Cun, Chi-Man Pun, and Cheng Shi. Towards ghost-free shadow removal via dual hierarchical aggregation network and shadow matting GAN. In *Proc. AAAI Conf. Artif. Intell.*, volume 34, pp. 10680–10687, 2020. doi: 10.1609/aaai.v34i07.6695.

Sagnik Das, Hassan Ahmed Sial, Ke Ma, Ramon Baldrich, Maria Vanrell, and Dimitris Samaras. Intrinsic decomposition of document images in-the-wild. In *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2020. doi: 10.5244/C.34.188.

Graham D. Finlayson, Steven D. Hordley, and Mark S. Drew. Removing shadows from images. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 823–836, 2002. doi: 10.1007/3-540-47979-1_55.

Lanqing Guo, Siyu Huang, Ding Liu, Hao Cheng, and Bihan Wen. ShadowFormer: Global context helps shadow removal. In *Proc. AAAI Conf. Artif. Intell.*, volume 37, pp. 710–718, 2023a. doi: 10.1609/aaai.v37i1.25148.

Lanqing Guo, Chong Wang, Wenhan Yang, Siyu Huang, Yufei Wang, Hanspeter Pfister, and Bihan Wen. ShadowDiffusion: When degradation prior meets diffusion model for shadow removal. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 14049–14058, 2023b. doi: 10.1109/CVPR52729.2023.01350.

Ruiqi Guo, Qieyun Dai, and Derek Hoiem. Paired regions for shadow detection and removal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2956–2967, 2013. doi: 10.1109/TPAMI.2012.214.

Shilin Hu, Hieu Le, ShahRukh Athar, Sagnik Das, and Dimitris Samaras. Shadow removal refinement via material-consistent shadow edges. In *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pp. 2631–2641, 2025. doi: 10.1109/WACV61041.2025.00261.

Xiaowei Hu, Tianyu Wang, Chi-Wing Fu, Yitong Jiang, Qiong Wang, and Pheng-Ann Heng. Revisiting shadow detection: A new benchmark dataset for complex world. *IEEE Trans. Image Process.*, 30:1925–1934, 2021. doi: 10.1109/TIP.2021.3049331.

Jing-En Huang, I-Sheng Fang, Tzuhsuan Huang, Yu-Lun Liu, Chih-Yu Wang, and Jun-Cheng Chen. Gen-n-val: Agentic image data generation and validation. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Findings (CVPR Findings)*, pp. 8786–8795, 2026.

Naoto Inoue and Toshihiko Yamasaki. Learning from synthetic shadows for shadow detection and removal. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(11):4187–4197, 2021. doi: 10.1109/TCSVT.2020.3047977.

Daniel J. Jobson, Zia ur Rahman, and Glenn A. Woodell. Properties and performance of a center/surround retinex. *IEEE Transactions on Image Processing*, 6(3):451–462, 1997. doi: 10.1109/83.557356.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

Hieu Le and Dimitris Samaras. Shadow removal via shadow image decomposition. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 8577–8586, 2019. doi: 10.1109/ICCV.2019.00867.

Hieu Le and Dimitris Samaras. Physics-based shadow image decomposition for shadow removal. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(12):9088–9101, 2022. doi: 10.1109/TPAMI.2021.3124934.

Chia-Ming Lee, Yu-Fan Lin, Yu-Jou Hsiao, Jin-Hui Jiang, Yu-Lun Liu, and Chih-Chung Hsu. PhaSR: Generalized image shadow removal with physically aligned priors. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 22679–22688, 2026.

Xiaoguang Li, Qing Guo, Rabab Abdelfattah, Di Lin, Wei Feng, Ivor Tsang, and Song Wang. Leveraging inpainting for single-image shadow removal. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 13009–13018, 2023a. doi: 10.1109/ICCV51070.2023.01200.

Zinuo Li, Xuhang Chen, Chi-Man Pun, and Xiaodong Cun. High-resolution document shadow removal via a large-scale real-world dataset and a frequency-aware shadow erasing net. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 12449–12458, 2023b.

Feng Liu and Michael Gleicher. Texture-consistent shadow removal. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 437–450, 2008. doi: 10.1007/978-3-540-88693-8_32.

Yaojie Liu, Andrew Hou, Xinyu Huang, Liu Ren, and Xiaoming Liu. Blind removal of facial foreign shadows. In *Proc. Brit. Mach. Vis. Conf. (BMVC)*, 2022.

Zhihao Liu, Hui Yin, Xinyi Wu, Zhenyao Wu, Yang Mi, and Song Wang. From shadow generation to shadow removal. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 4927–4936, 2021. doi: 10.1109/CVPR46437.2021.00489.

Tailong Luo, Yihang Dong, Jiesong Bai, Jinyang Huang, Junyu Xia, Wangyu Wu, and Xuhang Chen. Beyond shadows: A large-scale benchmark and multi-stage framework for high-fidelity facial shadow removal. In *ICASSP 2026 - 2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, May 2026. doi: 10.1109/ICASSP55912.2026.11462271.

David Marr and Ellen Hildreth. Theory of edge detection. *Proceedings of the Royal Society of London. Series B, Biological Sciences*, 207(1167):187–217, 1980. doi: 10.1098/rspb.1980.0020.

Elías Masquil, Roger Marí, Thibaud Ehret, Enric Meinhardt-Llopis, Pablo Musé, and Gabriele Facciolo. S-EO: A large-scale dataset for geometry-aware shadow detection in remote sensing applications. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, pp. 2408–2418, 2025.

OpenAI. Models. `https://platform.openai.com/docs/models`, 2026. OpenAI API documentation, accessed July 12, 2026.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024. ISSN 2835-8856. URL `https://openreview.net/forum?id=a68SUt6zFt`.

Jaehyun Park, Minyoung Ahn, Minkyu Kim, Jonghyun Lee, Jae-Gil Lee, and Dongmin Park. See and fix the flaws: Enabling VLMs and diffusion models to comprehend visual artifacts via agentic data synthesis. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 35810–35820, 2026.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 32, 2019. doi: 10.48550/arXiv.1912.01703.

Lu Qi, Jason Kuen, Tiancheng Shen, Jiuxiang Gu, Wenbo Li, Weidong Guo, Jiaya Jia, Zhe Lin, and Ming-Hsuan Yang. High quality entity segmentation. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 4047–4056, 2023.

Liangqiong Qu, Jiandong Tian, Shengfeng He, Yandong Tang, and Rynson W. H. Lau. DeshadowNet: A multi-context embedding deep network for shadow removal. In *Proc. IEEE Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 2308–2316, 2017. doi: 10.1109/CVPR.2017.248.

Shelly Sheynin, Adam Polyak, Uriel Singer, Yuval Kirstain, Amit Zohar, Oron Ashual, Devi Parikh, and Yaniv Taigman. Emu edit: Precise image editing via recognition and generation tasks. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 8871–8879, 2024. doi: 10.1109/CVPR52733.2024.00847.

Aaditya Singh et al. OpenAI GPT-5 System Card, 2026. URL `https://arxiv.org/abs/2601.03267`.

Florin-Alexandru Vasluianu, Tim Seizinger, and Radu Timofte. WSRD: A novel benchmark for high resolution image shadow removal. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, pp. 1826–1835, 2023.

Florin-Alexandru Vasluianu, Tim Seizinger, Zhuyun Zhou, Zongwei Wu, Cailian Chen, Radu Timofte, et al. NTIRE 2024 image shadow removal challenge report. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, pp. 6547–6570, 2024. doi: 10.1109/CVPRW63382.2024.00654.

Tomás F. Yago Vicente, Le Hou, Chen-Ping Yu, Minh Hoai, and Dimitris Samaras. Large-scale training of shadow detectors with noisily-annotated shadow examples. In *Proc. Eur. Conf. Comput. Vis. (ECCV)*, pp. 816–832, 2016. doi: 10.1007/978-3-319-46466-4_49.

Jifeng Wang, Xiang Li, and Jian Yang. Stacked conditional generative adversarial networks for jointly learning shadow detection and shadow removal. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 1788–1797, 2018. doi: 10.1109/CVPR.2018.00192.

Tianyu Wang, Xiaowei Hu, Qiong Wang, Pheng-Ann Heng, and Chi-Wing Fu. Instance shadow detection. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 1877–1886, 2020. doi: 10.1109/CVPR42600.2020.00195.

Tai-Pang Wu, Chi-Keung Tang, Michael S. Brown, and Heung-Yeung Shum. Natural shadow matting. *ACM Trans. Graph.*, 26(2):8, 2007. doi: 10.1145/1243980.1243982.

Jie Xiao, Xueyang Fu, Yurui Zhu, Dong Li, Jie Huang, Kai Zhu, and Zheng-Jun Zha. HomoFormer: Homogenized transformer for image shadow removal. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 25617–25626, 2024.

Jiamin Xu, Zelong Li, Yuxin Zheng, Chenyu Huang, Renshu Gu, Weiwei Xu, and Gang Xu. OmniSR: Shadow removal under direct and indirect lighting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 8887–8895, 2025a. doi: 10.1609/aaai.v39i8.32961.

Jiamin Xu, Yuxin Zheng, Zelong Li, Chi Wang, Renshu Gu, Weiwei Xu, and Gang Xu. Detail-preserving latent diffusion for stable shadow removal. In *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recog. (CVPR)*, pp. 7592–7602, 2025b.

Han Yang, Tianyu Wang, Xiaowei Hu, and Chi-Wing Fu. SILT: Shadow-aware iterative label tuning for learning to detect shadows from noisy labels. In *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, pp. 12641–12652, 2023. doi: 10.1109/ICCV51070.2023.01165.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Advances in Neural Information Processing Systems*, volume 37, pp. 21875–21911, 2024. doi: 10.52202/079017-0688.

Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 36, pp. 31428–31449, 2023. doi: 10.52202/075280-1365.

Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In *Adv. Neural Inf. Process. Syst. (NeurIPS)*, volume 37, pp. 3058–3093, 2024. doi: 10.52202/079017-0100.

# A AgenticShadow Dataset Details

## A.1 Source Processing, Cleaning, and Splits

AgenticShadow is constructed from four existing shadow-image sources: SBU (Vicente et al., 2016), CUHK-Shadow (Hu et al., 2021), ASFW (Luo et al., 2026), and S-EO (Masquil et al., 2025). Because these sources differ in organization, annotation format, and image quality, we process each separately before target construction.

For SBU, we pair the original images with the refined training masks and relabeled test masks released by SILT (Yang et al., 2023). For CUHK-Shadow, we retain only pairs listed in the official training, validation, and test manifests. Duplicate review excludes 102 SBU and 70 CUHK-Shadow examples.

To place the heterogeneous SBU and CUHK-Shadow images on a common resolution and quality level, we apply the image-normalization procedure described in Appendix B.3. Seven SBU images are excluded because normalization is blocked by content moderation. ASFW already provides aligned $512 \times 512$ image-mask pairs and therefore requires no normalization.

The S-EO release provides RGB images and Min-DSM-derived shadow annotations in separate archives. We match them by AOI and view identifiers, obtaining 19,162 pairs from 20,275 RGB images and 19,163 complete mask groups. Because many S-EO views contain large empty or invalid zero-valued regions within the AOI, we prioritize candidate crops with the highest valid-image

Table 5: **AgenticShadow processing and retention.** Eligible denotes matched source image-mask pairs, selected denotes examples retained after duplicate review or crop selection, usable denotes inputs ready for target construction after source-specific preprocessing, and generated denotes successful target generation before regional correction.

| | | | | | Final triplets | | |
|---|---|---|---|---|---|---|---|
| Source | Eligible | Selected | Usable | Generated | Train | Test | Total |
| SBU | 4,723 | 4,621 | 4,614 | 4,609 | 3,974 | 635 | 4,609 |
| CUHK | 10,500 | 10,430 | 10,430 | 10,425 | 8,335 | 2,087 | 10,422 |
| ASFW | 1,081 | 1,081 | 1,081 | 1,081 | 928 | 153 | 1,081 |
| S-EO | 19,162 | 1,026 | 1,026 | 1,026 | 880 | 146 | 1,026 |
| **Total** | **35,466** | **17,158** | **17,151** | **17,141** | **14,117** | **3,021** | **17,138** |

coverage for visual review, aided by the released uncertainty, unseen-pixel, and vegetation annotations. We visually inspect 1,119 candidate crops and retain 1,026 after rejecting candidates with poor image or shadow mask quality and removing overlapping selections.

As summarized in Table 5, we produce 17,151 usable image-mask pairs before target generation. Candidate generation excludes five SBU and five CUHK-Shadow examples, while regional correction excludes three additional CUHK-Shadow examples, yielding the final 17,138 triplets.

SBU and ASFW retain their source splits. For CUHK-Shadow, the official validation set is assigned to training, while the test split is preserved. S-EO uses an AOI-disjoint split with 152 training and 25 test AOIs, ensuring that all crops and source views from each AOI remain in the same split. Duplicate review is completed before split finalization to prevent train-test overlap.

### A.2 Construction and Diversity Statistics

**Construction outcomes.** We retain the generation and evaluation decisions for every constructed case. As shown in Table 6, the initial probe passes the major-failure evaluator for 97.93% of generated targets, while 2.07% enter the feedback-guided retry branch. After candidate filtering, 99.52% of cases contain at least two valid candidates for Retinex-LoG-supported selection. Only 68 cases contain a single passing candidate, and 15 cases use the least-damaging fallback because no candidate passes the major-failure evaluator. For the largest source, CUHK-Shadow, we track \$822.06 in API cost for generating 10,425 targets, including retries, averaging approximately \$0.079 per target. Subsequent segmentation, correction, and auditing are performed locally.

**Semantic scene diversity.** We compare the semantic diversity of AgenticShadow and ISTD+ (Le & Samaras, 2019) using normalized CLS embeddings from frozen DINOv2-L/14 (Oquab et al., 2024). Figure 7(a) jointly projects all 14,117 AgenticShadow and 1,330 ISTD+ training images using PCA. AgenticShadow occupies a visibly broader embedding space, with facial and remote-sensing imagery extending into distinct semantic regions. Because visual coverage grows with dataset size, we use a size-controlled comparison in Figure 7(b). We sample 1,000 images without replacement from each dataset and compute the mean pairwise cosine distance in the original DINOv2 embedding space, repeating this procedure 100 times. A larger distance means that two randomly selected images are less semantically similar and therefore indicates greater scene diversity. AgenticShadow obtains $0.9413 \pm 0.0020$, compared with $0.7487 \pm 0.0032$ for ISTD+, where values denote the mean and standard deviation across 100 resamples, corresponding to a 25.7% increase.

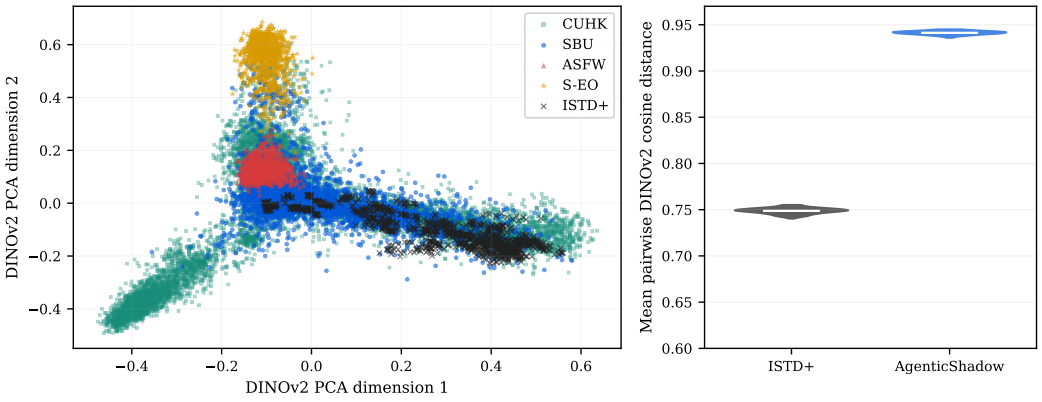
## B Agentic Construction Details

### B.1 Prompt Design and Pipeline Settings

The pipeline uses GPT Image 2 (OpenAI, 2026) for candidate generation and GPT-5 mini (Singh et al., 2026) for evaluation and selection. Each case starts with one mask-guided probe. Passing probes are retained and followed by two additional candidates, for three editing calls total; rejected

Table 6: **Dataset-wide target-construction outcomes.** Pass and retry describe the initial routing decision. Multi, single, and fallback are mutually exclusive selection outcomes. Percentages are computed within each source and overall.

| | | Probe routing | | Final selection | | |
|---|---|---|---|---|---|---|
| Source | Generated | Pass | Retry | Multi | Single | Fallback |
| SBU | 4,609 | 4,500 (97.64%) | 109 (2.36%) | 4,588 (99.54%) | 19 (0.41%) | 2 (0.04%) |
| CUHK | 10,425 | 10,193 (97.77%) | 232 (2.23%) | 10,364 (99.41%) | 48 (0.46%) | 13 (0.12%) |
| ASFW | 1,081 | 1,081 (100%) | 0 (0%) | 1,081 (100%) | 0 (0%) | 0 (0%) |
| S-EO | 1,026 | 1,013 (98.73%) | 13 (1.27%) | 1,025 (99.90%) | 1 (0.10%) | 0 (0%) |
| **Total** | **17,141** | **16,787 (97.93%)** | **354 (2.07%)** | **17,058 (99.52%)** | **68 (0.40%)** | **15 (0.09%)** |




(a) DINOv2 embedding space

(b) Size-controlled diversity

Figure 7: **Semantic scene diversity compared with ISTD+.** (a) Joint PCA visualization of DINOv2-L/14 embeddings for all training images. (b) Mean pairwise DINOv2 cosine distance over 100 size-matched resamples of 1,000 images; higher values indicate greater semantic diversity. Distances are computed in the original embedding space rather than the two-dimensional projection.

probes are replaced by three retries using evaluator feedback, for four calls total. Every candidate undergoes the same major-failure evaluation. A single passing candidate is selected directly, multiple passing candidates are compared using visual assessment supported by Retinex-LoG diagnostics, and if none pass, the least-damaging fallback selector is used.

The two prompt families differ only in their shadow-formation grounding, as summarized in Figure 8. The generic prompts request shadow removal while preserving scene structure, object identity, materials, texture, markings, and local color. The physics-motivated prompts retain these requirements and additionally distinguish cast shadows on receiving surfaces from attached or self-shadows on the same object. This grounding is applied consistently to generation, retry, evaluation, and selection. The evaluator schema remains unchanged and rejects only two major failures: preserving a substantial shadow as coherent scene content or hallucinating semantic content inside the mask. This controlled design isolates the effect of shadow-formation grounding from the agentic control flow.

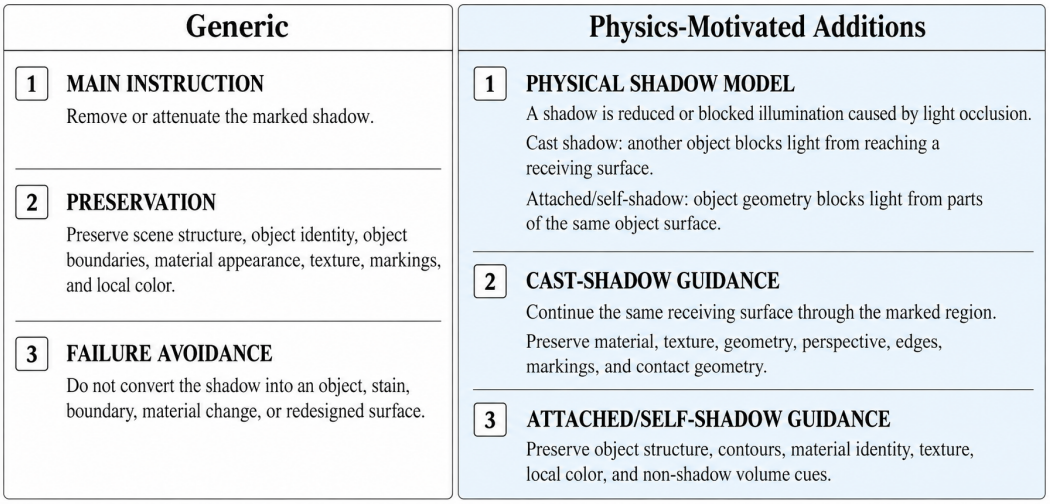



Figure 8: **Generic prompting and physics-motivated additions.** Both settings use identical inputs, preservation requirements, evaluator schemas, and agentic control flow. Physics-motivated prompting additionally defines cast and attached shadows as illumination effects caused by light occlusion and provides type-specific guidance for preserving the underlying surface or object.

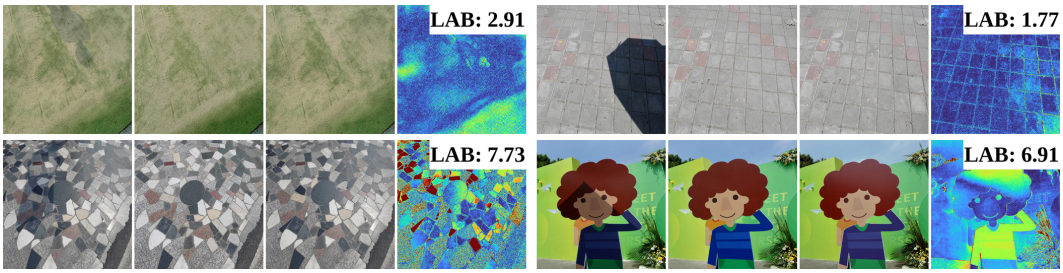



Figure 9: **Construction against captured ISTD+ references.** Each group shows the shadow input, our constructed target, the color-corrected ISTD+ reference, and the mean absolute RGB error map using a shared 0–50 scale. The top row presents two low-error cases, while the bottom row presents two higher-error cases that expose remaining texture, detail, and color discrepancies.

### B.2 Validation against Captured ISTD+ References

We first test whether the construction workflow can recover known paired targets by applying it to all 540 ISTD+ test images and masks (Le & Samaras, 2019). The workflow receives only the shadow image and mask, uses no ISTD+ reference during construction, and receives no task-specific training on ISTD+. We apply the same candidate generation, evaluation, Retinex-LoG-supported selection, and segmentation-guided regional correction used to construct AgenticShadow. Against the color-corrected ISTD+ references, the resulting targets achieve 25.11 PSNR, 0.7380 SSIM, 10.95 MAE, and 3.85 LAB RMSE. This direct comparison shows that the workflow can approximate the color-corrected ISTD+ references without using them during construction. Representative cases are shown in Figure 9. The top row presents low-error cases with shallow and stronger shadows, showing close agreement with the references across different shadow intensities. The bottom row presents higher-error cases with larger differences in local color and fine appearance. Nevertheless, the dominant shadows are removed while the overall scene content remains consistent.




Original | Normalized 512 × 512 | Original | Normalized 512 × 512

Figure 10: **Image normalization examples.** The first three pairs show low-resolution inputs standardized to $512 \times 512$, while the last pair shows a higher-resolution input normalized to the same resolution. The preprocessing pass improves resolution and image quality while preserving the scene and original shadow appearance.

Table 7: **API usage for complete SRR runs.** Each run processes all 399 evaluation cases. Pass and retry indicate the routing decision for the initial probe. Edits include the probe and subsequent guided or retry candidates, and each edit receives one major-failure evaluation. Selector counts exclude cases with exactly one passing candidate, which are selected directly.

| Prompt setting | Cases | Pass | Retry | Edits | Evaluations | Selectors |
|---|---|---|---|---|---|---|
| Generic | 399 | 391 | 8 | 1,205 | 1,205 | 394 |
| Physics-motivated | 399 | 389 | 10 | 1,207 | 1,207 | 397 |

### B.3 Image Normalization and API Usage

Because the source images vary substantially in resolution and image quality, we use a preprocessing-only GPT Image 2 pass to produce standardized $512 \times 512$ images while preserving the scene and original shadow appearance. The normalization prompt explicitly prohibits shadow removal, relighting, reframing, scene modification, or object addition and removal. For the ShadowRemovalRefine (SRR) benchmark (Hu et al., 2025), we apply this procedure to each image and spatially align its mask to the normalized output.

SRR originally contains 400 images with diverse resolutions and aspect ratios. One image is excluded after an API moderation block, leaving 399 evaluation cases. As illustrated in Figure 10, each image is normalized before candidate generation. All compared methods and agentic configurations receive the same normalized images and aligned masks, ensuring consistent quantitative comparisons.

Table 7 summarizes the API calls required by complete generic and physics-motivated runs. Their similar usage reflects the identical control flow, with shadow-formation grounding as the principal difference. The agentic workflow is intended for offline target construction rather than direct deployment: a complete 399-image run takes approximately 14 hours and incurs an estimated API cost of about $29, whereas specialized learned models operate on the order of one second per image. General-purpose generative editors also remain stochastic and can misinterpret shadow structure, motivating the evaluation, filtering, and correction stages. The resulting targets can instead be constructed once and used to train efficient specialized models for deployment.

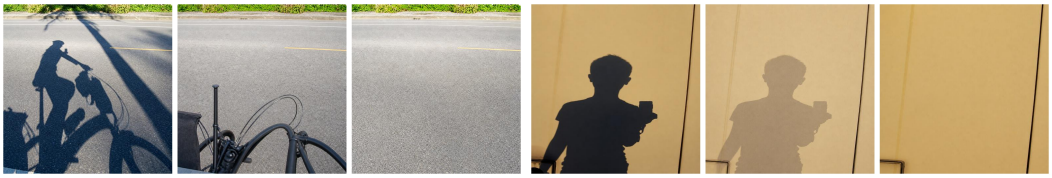



Figure 11: **Effect of major-failure retry.** The evaluator rejects probes with severe residual shadow structure or object-like hallucination in the masked region. Feedback-guided retry produces cleaner candidates before final selection.

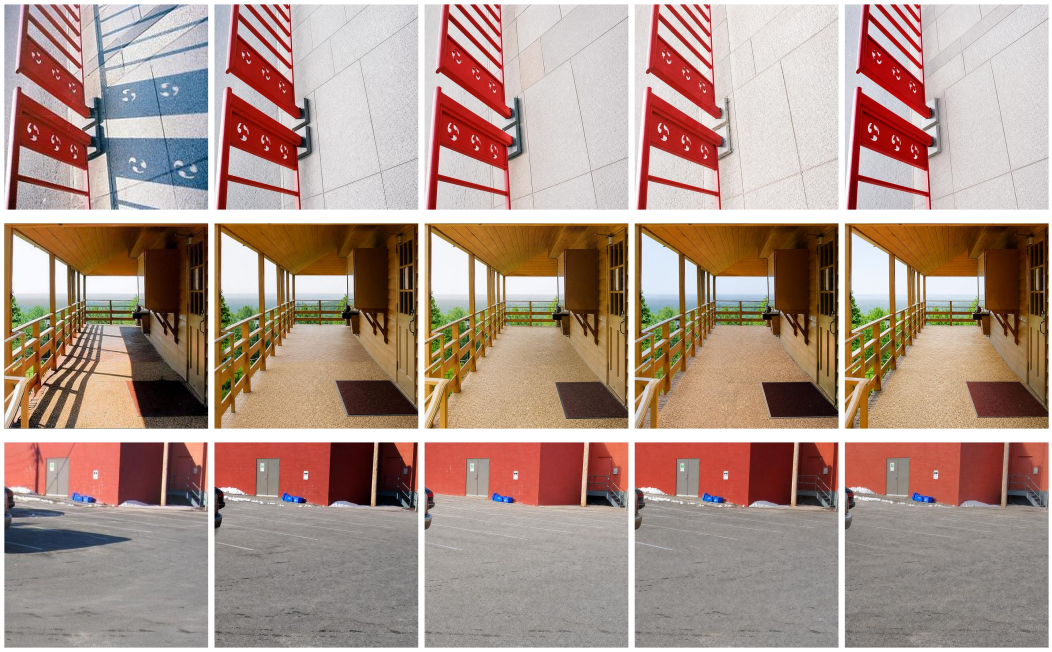



Figure 12: **Effect of shadow-formation grounding and visual candidate selection.** Physics-motivated prompting reduces task-level errors and produces stronger candidate edits than generic prompting. All three physics-motivated candidates are shown, with the visually selected output in the final column.

### B.4 Additional Qualitative Examples

**Major-failure retry.** Figure 11 shows cases in which the initial probe preserves the shadow as coherent scene content or introduces an object-like interpretation of the masked region. The major-failure evaluator rejects these probes and supplies feedback for retry generation, producing candidates with more complete shadow removal and better scene preservation.

**Shadow-formation grounding and visual selection.** Figure 12 compares generic outputs with physics-motivated candidates. Generic prompting can leave residual shadows, introduce boundary artifacts, or reinterpret dark regions as material or object structure. Shadow-formation grounding produces a stronger candidate pool, and the visual selector chooses the candidate that best balances complete shadow removal with scene, material, and structural consistency. The examples also illustrate why comparing multiple stochastic edits is more reliable than accepting a single output.

**Retinex-LoG-supported selection.** Figure 13 compares visual-only selection with the final selection supported by Retinex-LoG evidence. Although both candidates may appear plausible, visual

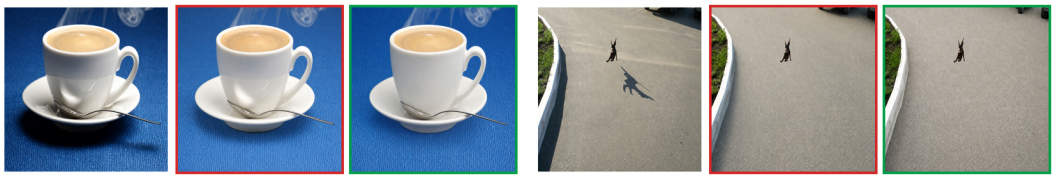



Figure 13: **Effect of Retinex-LoG-supported selection.** Red frames indicate candidates selected through visual assessment alone, while green frames indicate the final selections obtained with Retinex-LoG structural evidence. The additional evidence helps reject candidates with residual shadows and select candidates that better preserve scene structure.




Figure 14: **Effect of segmentation-guided regional correction.** Selected proposals remove the dominant shadows but can retain spatially varying brightness or color shifts. Regional correction improves agreement with the surrounding scene while preserving image structure.




Figure 15: **Representative construction failures.** Each pair shows the mask-overlay input and constructed target. The first example exhibits texture smoothing and loss of valid scene detail. The second illustrates the recurring residual-shadow problem beneath vehicles. The third shows incomplete removal of an attached shadow.

assessment alone can favor incomplete shadow removal. By exposing residual, lost, and added local structure after suppressing low-frequency illumination, Retinex-LoG helps the selector achieve more complete removal while preserving the underlying texture.

**Segmentation-guided regional correction.** Figure 14 shows the input, selected proposal, and regionally corrected output. Candidate selection removes the dominant shadow, but the proposal may retain broad or region-specific brightness and chrominance shifts. We partition the proposal into class-agnostic regions using CropFormer (Qi et al., 2023). For each region, the correction stage estimates offsets from reliable pixels outside the dilated shadow mask, improving local color consistency without reintroducing shadow structure from the input.

### B.5 Failure Cases and Limitations

**Failure cases.** Although the evaluation, selection, and correction stages reduce major construction errors, several failure modes remain. As shown in Figure 15, generative editing can smooth fine texture or remove valid scene detail. Residual shadows beneath vehicles represent a recurring failure mode. Attached shadows may also be only partially removed.

Table 8: **Checkpoint and inference settings.** Combined denotes training with ISTD+ and AgenticShadow, while ISTD+ / INS (Xu et al., 2025a) denotes separate official checkpoints. Outputs are restored to the target resolution before evaluation.

| Method | Checkpoint condition | Mask | Inference |
|---|---|---|---|
| SID (Le & Samaras, 2019) | Official ISTD+ | Yes | Resize to $512 \times 512$ |
| ShadowFormer (Guo et al., 2023a) | Official ISTD+ | Yes | Resize to $640 \times 480$ |
| Inpaint4Shadow (Li et al., 2023a) | Official ISTD+ | Yes | Resize to $256 \times 256$ |
| StableSR (Xu et al., 2025b) | Official ISTD+ / INS | No | Native resolution; 20 steps<br>Padded to a multiple of 16 |
| ShadowDiffusion (Guo et al., 2023b) | Official ISTD+<br>Combined E400 | Yes | Full $512 \times 512$ inference<br>25 DDIM steps |
| HomoFormer (Xiao et al., 2024) | Official ISTD+<br>Combined E200 | Yes | $384 \times 384$ tiles<br>30-pixel overlap |
| PhaSR (Lee et al., 2026) | Official ISTD+ / INS<br>Combined E200 | No | Native resolution<br>Padded to a multiple of 128 |

**Limitations.** CDD measures color consistency across the shadow boundary but does not capture changes elsewhere in the image. We therefore additionally measure RGB RMSE between the input and output over non-shadow pixels on the normalized SRR benchmark, using the 0–255 scale. Averaged over two independent runs, the final workflow obtains a mean RMSE of $23.13$, compared with $3.44$–$18.39$ for the evaluated shadow removal methods (Le & Samaras, 2019; Guo et al., 2023a;b; Li et al., 2023a; Hu et al., 2025). Unlike methods that largely preserve non-shadow pixels, a general-purpose editor regenerates the image and may not remain perfectly aligned with the input at the pixel level. The relatively high RMSE of ShadowDiffusion ($18.39$) is consistent with this effect. Regional correction reduces the mean non-shadow RMSE of the Retinex-LoG-selected outputs from $25.48$ to $23.13$, but the remaining gap reflects both pixel-level misalignment and unintended changes outside the target region, including color shifts, texture smoothing, and subtle structural alterations.

# C Implementation and Evaluation Details

## C.1 Baseline Checkpoints and Evaluation Protocol

Table 8 records the exact checkpoint and inference settings used for evaluation. We evaluate all 540 ISTD+ and 3,021 AgenticShadow test images, restoring every prediction to its target resolution before computing metrics. PSNR, SSIM, and MAE are computed in RGB space with values in $[0, 255]$; SSIM is averaged across the three color channels. LAB RMSE is computed after converting the prediction and target from sRGB to LAB. We first average each metric over images within each domain, then compute the AgenticShadow macro average as the unweighted mean of the four domain-level results. ISTD+ is reported separately.

## C.2 PF and PF-MF Implementation and Training

Combined-data training uses the natural concatenation of 1,330 ISTD+ and 14,117 AgenticShadow examples without source reweighting. Image, mask, and target triplets receive the same random $384 \times 384$ crop and synchronized horizontal-flip and $90^\circ$-rotation augmentations. The Charbonnier loss (Charbonnier et al., 1994) uses $\epsilon = 10^{-3}$.

The component analysis compares semantic-only, depth-only, and full PF variants after 50 additional epochs from the same HomoFormer E100 checkpoint. The final PF model is initialized from HomoFormer E100, while PF-MF transfers the compatible RGB weights from the same checkpoint. Both models are then trained for 100 epochs; their backbones therefore receive 200 cumulative epochs of combined-data training, while the newly added prior modules are trained for 100 epochs.

Table 9: **Model complexity and inference efficiency.** We compare ShadowDiffusion (Guo et al., 2023b), HomoFormer (Xiao et al., 2024), PhaSR (Lee et al., 2026), PF, and PF-MF. Online parameters include all networks executed during inference. PhaSR uses precomputed geometry, whereas PF and PF-MF execute both frozen prior extractors online. Runtime and peak VRAM are measured on an NVIDIA RTX A6000 during end-to-end processing of $512 \times 512$ images.

| Method | Online params (M) | Trainable params (M) | Time/image (s) | Peak VRAM (GiB) |
|---|---|---|---|---|
| ShadowDiffusion | 55.52 | 55.52 | 2.558 | 1.51 |
| HomoFormer | 17.81 | 17.81 | 0.505 | 0.92 |
| PhaSR | 323.37 | 19.00 | 0.445 | 2.80 |
| PF | 658.93 | 19.25 | 0.999 | 3.33 |
| PF-MF | 658.61 | 18.92 | 0.978 | 3.32 |

Table 10: **Full evaluation on the SBU and CUHK-Shadow domains.** Official ISTD+ checkpoints are evaluated without adaptation, while † denotes training with ISTD+ and AgenticShadow. Best and second-best results are **bolded** and underlined, respectively.

| | SBU | | | | CUHK-Shadow | | | |
|---|---|---|---|---|---|---|---|---|
| Method | PSNR↑ | SSIM↑ | MAE↓ | LAB↓ | PSNR↑ | SSIM↑ | MAE↓ | LAB↓ |
| *Official ISTD+ checkpoints* | | | | | | | | |
| SID (Le & Samaras, 2019) | 20.28 | 0.6258 | 18.26 | 7.25 | 16.32 | 0.5643 | 30.43 | 10.86 |
| ShadowFormer (Guo et al., 2023a) | 21.02 | 0.6506 | 16.99 | 6.63 | 17.09 | 0.6008 | 28.50 | 10.01 |
| Inpaint4Shadow (Li et al., 2023a) | 20.51 | 0.5883 | 18.30 | 7.12 | 17.54 | 0.5406 | 26.43 | 9.72 |
| StableSR (Xu et al., 2025b) | 20.34 | 0.6345 | 18.83 | 7.32 | 18.38 | 0.6173 | 23.94 | 8.84 |
| ShadowDiffusion (Guo et al., 2023b) | 20.75 | 0.6401 | 17.73 | 6.83 | 17.88 | 0.6026 | 25.22 | 9.09 |
| HomoFormer (Xiao et al., 2024) | 20.85 | 0.6430 | 17.41 | 6.78 | 18.01 | 0.6085 | 24.69 | 9.08 |
| PhaSR (Lee et al., 2026) | 20.28 | 0.6405 | 18.72 | 7.32 | 18.47 | 0.6240 | 23.44 | 8.71 |
| *With AgenticShadow supervision* | | | | | | | | |
| ShadowDiffusion† | 21.85 | 0.6578 | 16.12 | 5.84 | 19.80 | 0.6618 | 19.96 | 7.40 |
| HomoFormer† | 23.61 | 0.6783 | 12.79 | 4.90 | 21.53 | 0.6926 | 15.89 | 6.23 |
| PhaSR† | 23.66 | **0.6924** | 12.92 | 4.92 | 21.62 | **0.7073** | 15.93 | 6.19 |
| PF | **23.93** | 0.6888 | 12.39 | **4.73** | **21.80** | 0.7030 | **15.38** | **6.05** |
| PF-MF | 23.90 | 0.6896 | **12.38** | 4.75 | 21.72 | 0.7039 | 15.61 | 6.11 |

For the dataset-prior analysis, the PF restoration network is trained from scratch on ISTD+ for 400 epochs following the HomoFormer training regimen, while both prior extractors remain frozen.

### C.3 Model Complexity and Runtime

We measure efficiency on one NVIDIA RTX A6000 at batch size 1 over 40 fixed $512 \times 512$ test images, ten from each AgenticShadow domain. Each model processes five warm-up images before timing. Mean time per image includes model-specific preprocessing, online prior extraction, tiled inference and stitching where applicable, and output conversion, but excludes model loading and disk I/O. Peak VRAM is the maximum allocated GPU memory during measurement.

As shown in Table 9, PF and PF-MF require approximately one second per image. Their larger online parameter counts and memory usage arise primarily from running both frozen prior extractors, while their trainable parameter counts remain close to those of HomoFormer and PhaSR.

## D Additional Results

### D.1 Full Per-Domain Evaluation

Table 10 and Table 11 provide the complete per-domain results underlying the macro averages in the main paper. AgenticShadow supervision improves all four metrics for each retrained architecture in

Table 11: **Full evaluation on the ASFW and S-EO domains.** Official ISTD+ checkpoints are evaluated without adaptation, while † denotes training with ISTD+ and AgenticShadow. Best and second-best results are **bolded** and underlined, respectively.

| | ASFW | | | | S-EO | | | |
|---|---|---|---|---|---|---|---|---|
| Method | PSNR↑ | SSIM↑ | MAE↓ | LAB↓ | PSNR↑ | SSIM↑ | MAE↓ | LAB↓ |
| *Official ISTD+ checkpoints* | | | | | | | | |
| SID (Le & Samaras, 2019) | 20.02 | 0.7687 | 19.38 | 7.48 | 16.50 | 0.6027 | 27.02 | 10.23 |
| ShadowFormer (Guo et al., 2023a) | 20.18 | 0.7801 | 19.36 | 7.35 | 16.01 | 0.6052 | 28.57 | 10.74 |
| Inpaint4Shadow (Li et al., 2023a) | 19.90 | 0.7523 | 19.88 | 7.58 | 16.43 | 0.5665 | 27.62 | 10.08 |
| StableSR (Xu et al., 2025b) | 17.14 | 0.7367 | 28.16 | 10.01 | 15.75 | 0.6140 | 29.98 | 10.93 |
| ShadowDiffusion (Guo et al., 2023b) | 20.10 | 0.7704 | 19.61 | 7.47 | 16.17 | 0.6119 | 28.12 | 10.50 |
| HomoFormer (Xiao et al., 2024) | 20.57 | 0.7789 | 18.54 | 7.00 | 15.95 | 0.6083 | 28.77 | 10.87 |
| PhaSR (Lee et al., 2026) | 17.00 | 0.7454 | 28.03 | 10.36 | 15.87 | 0.6125 | 28.82 | 10.92 |
| *With AgenticShadow supervision* | | | | | | | | |
| ShadowDiffusion† | 21.81 | 0.8030 | 15.79 | 5.97 | 18.40 | 0.6583 | 22.46 | 7.98 |
| HomoFormer† | 23.03 | 0.8177 | 13.37 | 5.33 | 19.65 | 0.6839 | 18.95 | 7.02 |
| PhaSR† | 23.11 | 0.8240 | 14.29 | 5.31 | 19.87 | **0.7000** | 18.63 | 6.90 |
| PF | **24.33** | **0.8348** | **11.59** | **4.67** | 19.96 | 0.6990 | **18.08** | **6.76** |
| PF-MF | 23.81 | 0.8298 | 12.55 | 4.97 | **19.99** | 0.6979 | 18.14 | 6.78 |

Table 12: **Synthetic-to-real transfer.** We compare official ISTD+ and INS (Xu et al., 2025a) checkpoints of StableSR (Xu et al., 2025b) and PhaSR (Lee et al., 2026). AgenticShadow results are four-domain macro averages.

| | | ISTD+ | | | | AgenticShadow | | | |
|---|---|---|---|---|---|---|---|---|---|
| Model | Checkpoint | PSNR↑ | SSIM↑ | MAE↓ | LAB↓ | PSNR↑ | SSIM↑ | MAE↓ | LAB↓ |
| StableSR | ISTD+ | 32.23 | 0.9272 | 4.84 | 2.03 | 17.90 | 0.6506 | 25.23 | 9.28 |
| | INS | 22.76 | 0.8986 | 12.14 | 5.82 | 16.55 | 0.6350 | 28.93 | 11.12 |
| PhaSR | ISTD+ | 32.12 | 0.9330 | 4.82 | 2.13 | 17.90 | 0.6556 | 24.75 | 9.33 |
| | INS | 26.03 | 0.9129 | 8.93 | 4.32 | 17.24 | 0.6415 | 26.90 | 10.37 |

every domain, demonstrating consistent benefits across general, facial, and remote-sensing scenes. PF and PF-MF rank first or second in PSNR, MAE, and LAB RMSE across all four domains, with PF-MF remaining close despite requiring no input mask, while PhaSR† achieves the highest SSIM on SBU, CUHK-Shadow, and S-EO.

### D.2 Cross-Dataset and Prior Analysis

**Synthetic-to-real transfer.** Large-scale synthetic data provide another route to paired supervision: OmniSR (Xu et al., 2025a) introduced INS with over 30,000 rendered shadow and shadow-free image pairs spanning diverse objects and direct and indirect illumination. We therefore evaluate official INS checkpoints of StableSR and PhaSR to test whether this scale transfers to real imagery. As shown in Table 12, both INS checkpoints underperform their ISTD+ counterparts across all metrics on ISTD+ and AgenticShadow, revealing a persistent synthetic-to-real gap and underscoring the importance of large-scale supervision constructed from diverse real-world images.

**Complementary paired supervision.** Table 13 examines how controlled and constructed pairs complement one another. Training with only AgenticShadow after ISTD+ initialization reduces macro LAB RMSE from 8.43 to 5.94 but increases ISTD+ LAB RMSE from 1.92 to 2.74, indicating that controlled captures remain a useful anchor for photometric fidelity. Combined training performs best on both evaluations, showing that AgenticShadow adds substantial real-world diversity while controlled pairs help preserve precise restoration behavior.

Table 13: **Effect of HomoFormer training data.** The AgenticShadow and Combined conditions are initialized from the official ISTD+ checkpoint; Combined denotes training with ISTD+ and AgenticShadow. AgenticShadow results are four-domain macro averages.

| | ISTD+ | | AgenticShadow | | | |
|---|---|---|---|---|---|---|
| Training data | PSNR↑ | LAB↓ | PSNR↑ | SSIM↑ | MAE↓ | LAB↓ |
| ISTD+ | <u>32.53</u> | <u>1.92</u> | 18.85 | 0.6597 | 22.36 | 8.43 |
| AgenticShadow | 27.96 | 2.74 | <u>21.85</u> | <u>0.7152</u> | <u>15.45</u> | <u>5.94</u> |
| Combined | **32.92** | **1.85** | **21.95** | **0.7181** | **15.25** | **5.87** |

Table 14: **Per-domain PF component analysis.** All values are LAB RMSE. Each variant is initialized from HomoFormer E100 and trained for 50 additional epochs; Macro is the unweighted average of the four AgenticShadow domains.

| Variant | ISTD+ | SBU | CUHK | ASFW | S-EO | Macro |
|---|---|---|---|---|---|---|
| HomoFormer | 1.860 | 4.907 | 6.266 | 5.434 | 7.041 | 5.912 |
| Depth only | 1.859 | 4.835 | 6.155 | 5.122 | 6.964 | 5.769 |
| Semantic only | **1.857** | 4.731 | 6.051 | 4.750 | 6.807 | 5.585 |
| PF | 1.861 | **4.723** | **6.045** | **4.737** | **6.804** | **5.577** |

**Prior components.** Table 14 shows that semantic and depth priors independently improve every AgenticShadow domain, with semantic information providing the larger gain. Combining both priors performs best on all four domains and the macro average, while ISTD+ performance remains nearly unchanged across the matched variants. This indicates that prior fusion is most beneficial under the broader scene and shadow diversity of AgenticShadow.

### D.3 Qualitative Results and Videos

**Videos.** We provide complete video results in the supplementary `video.zip`, comparing PF with pretrained HomoFormer (Xiao et al., 2024) on ViSha (Chen et al., 2021) using the provided masks, and PF-MF with SP+M+I-Net (Le & Samaras, 2022) on SBU-TimeLapse without input masks. All methods are applied independently to individual frames without temporal modeling. Although our baselines are not optimized for sequential data, PF and PF-MF produce cleaner and more complete shadow removal than the compared methods across the videos.

**Additional qualitative comparisons.** We provide three complementary comparisons: (1) Figure 16 completes qualitative coverage of the official ISTD+ checkpoints; (2) Figure 17 compares ISTD+ and INS checkpoints to illustrate the synthetic-to-real generalization gap; and (3) Figure 18 contrasts ISTD+-pretrained models with their AgenticShadow-trained counterparts and PF-MF, demonstrating the benefit of broader paired supervision and the effectiveness of the mask-free variant.

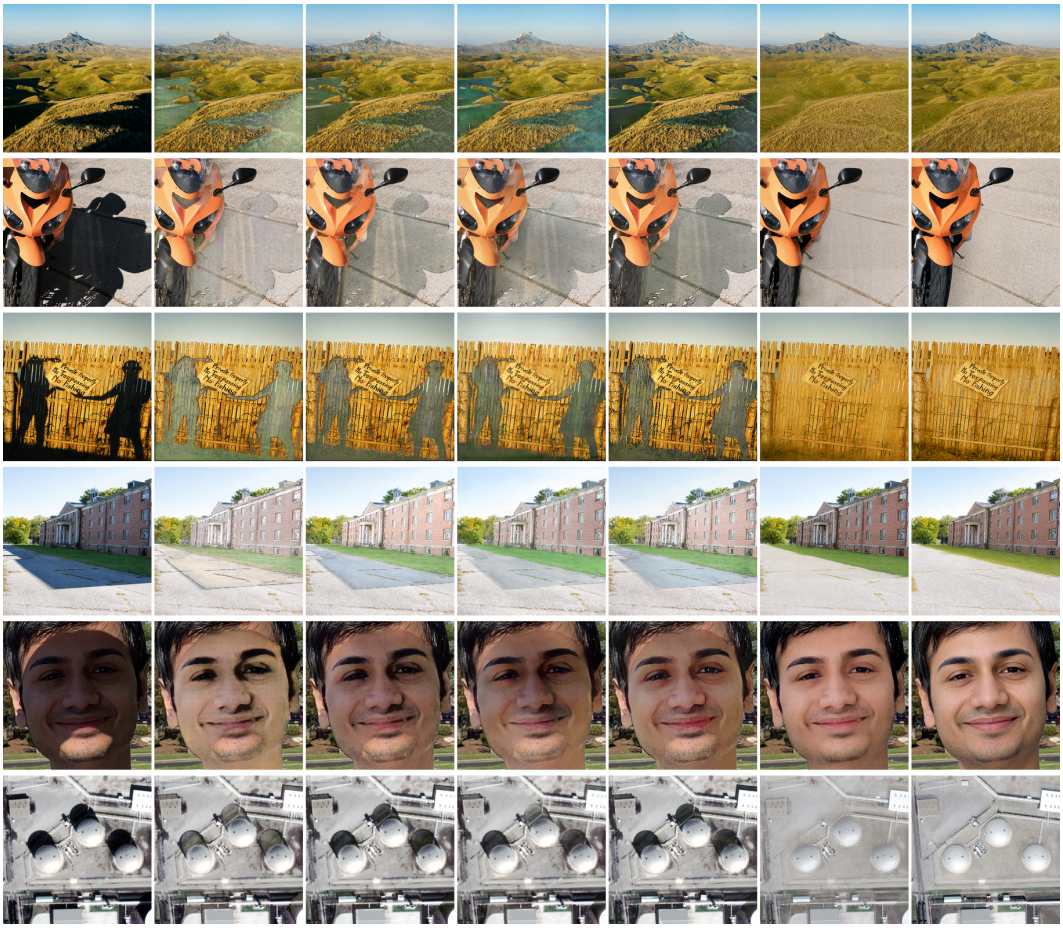



Figure 16: **Qualitative comparison with official ISTD+ checkpoints.** We compare SID (Le & Samaras, 2019), ShadowFormer (Guo et al., 2023a), Inpaint4Shadow (Li et al., 2023a), and ShadowDiffusion (Guo et al., 2023b) with PF across the four AgenticShadow domains. The official checkpoints frequently leave residual shadows or introduce color shifts, whereas PF produces more complete removal and outputs closer to the constructed targets.

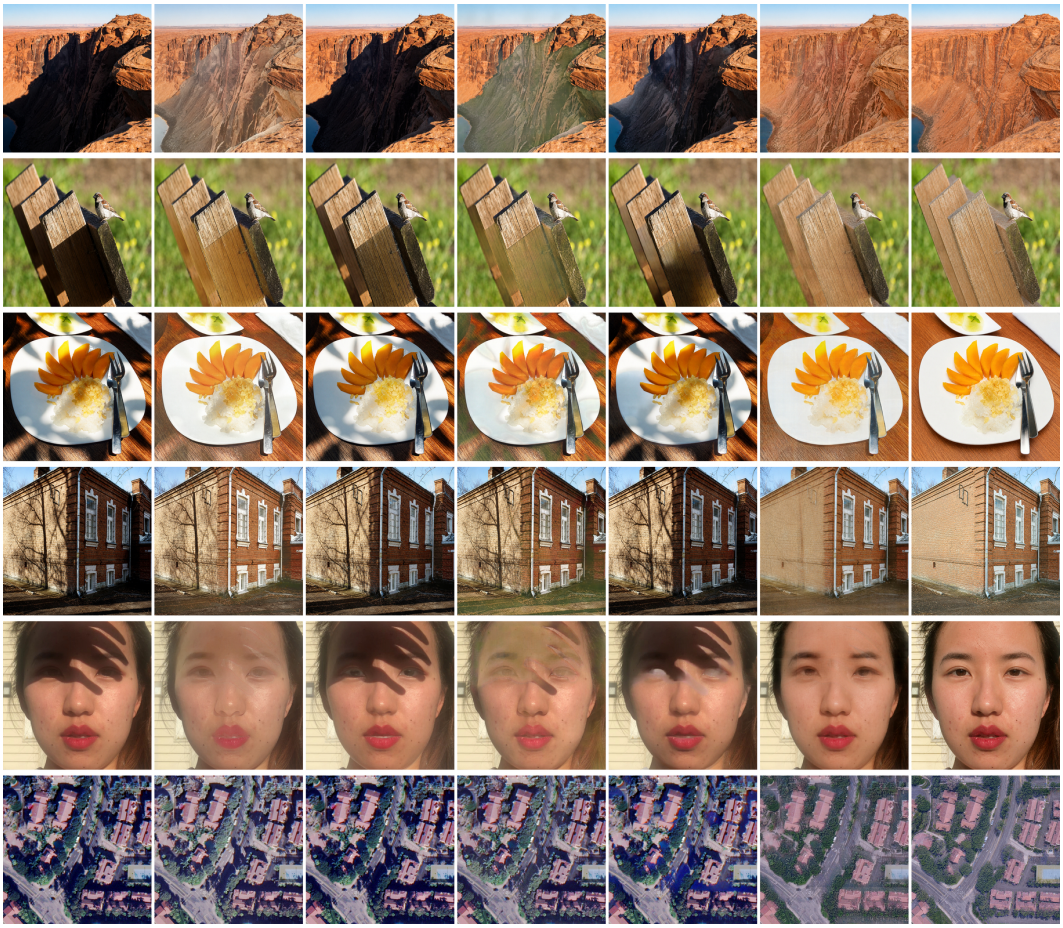



Figure 17: **Qualitative comparison of ISTD+ and INS checkpoints.** We compare the official ISTD+ and synthetic INS (Xu et al., 2025a) checkpoints of StableSR (Xu et al., 2025b) and PhaSR (Lee et al., 2026) with PF. The INS-trained models continue to exhibit residual shadows and appearance shifts on real-world AgenticShadow scenes, illustrating that large-scale synthetic supervision does not eliminate the synthetic-to-real gap.

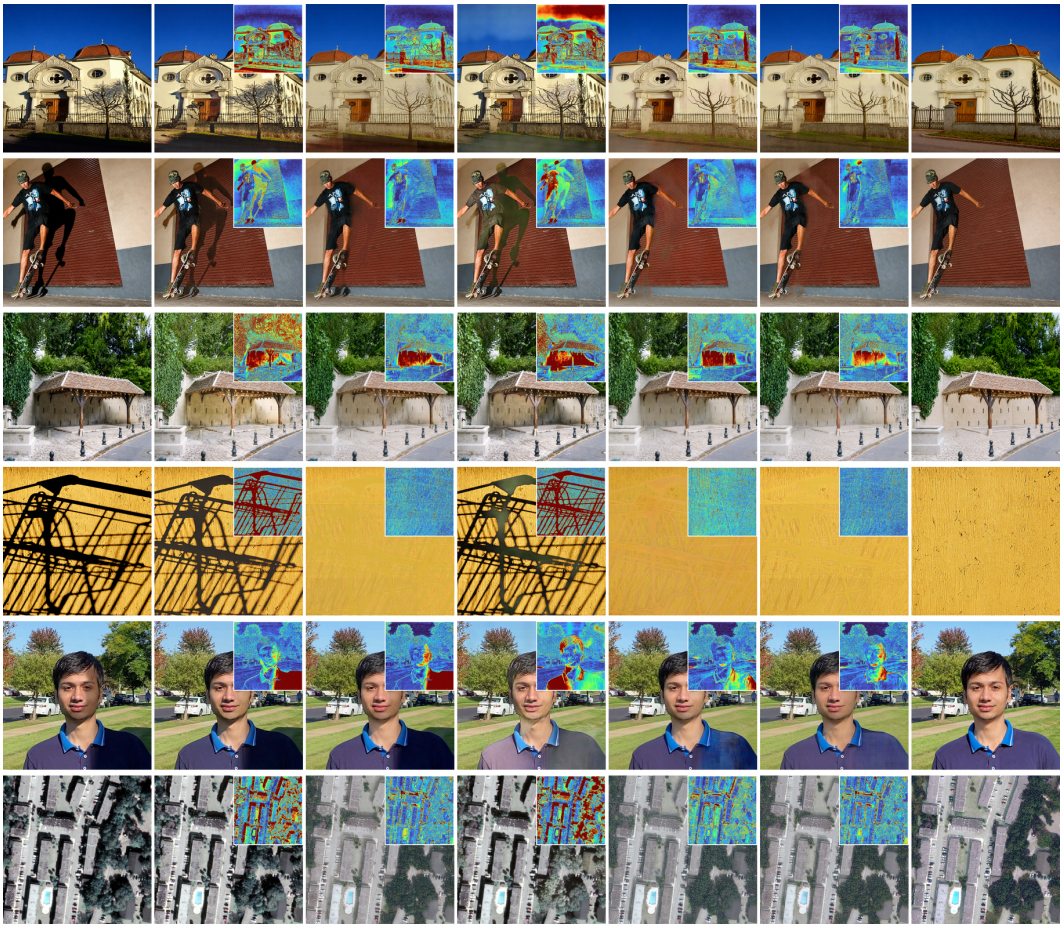



Figure 18: **Additional qualitative comparison of AgenticShadow supervision.** We compare the official ISTD+ checkpoints of HomoFormer (Xiao et al., 2024) and PhaSR (Lee et al., 2026), their counterparts trained with ISTD+ and AgenticShadow (†), and PF-MF. AgenticShadow supervision produces cleaner and more complete shadow removal, while PF-MF achieves strong results without input masks. Colored insets show mean absolute RGB error using a shared scale.