# Bridging Image Restoration and Recognition via Causal Mediated Unrolling

**Haowei Li**[1] **Qing Zhao**[1] **Yanze Li**[1] **Zhichen Chen**[1]
**Pengxu Wei**[1,3*] **Dongyu Zhang**[2,*] **Liang Lin**[1,3]

[1]School of Computer Science and Engineering, Sun Yat-sen University
[2]College of Professional and Continuing Education,
The Hong Kong Polytechnic University
[3]Peng Cheng Laboratory

{lihw59,zhaoq78,liyz53,chencc7}@mail2.sysu.edu.cn
{weipx3,linlng}@mail.sysu.edu.cn
donny.zhang@cpce-polyu.edu.hk

## Abstract

Robust visual recognition under degraded imaging conditions is critical for real-world vision systems. Task-oriented image restoration addresses this problem by enhancing degraded inputs before recognition, but it must preserve natural visual appearance without disrupting the evidence used by downstream recognizers. Existing optimization strategies struggle to satisfy these requirements simultaneously. Restoration driven mainly by visual quality may weaken task-critical cues or introduce texture shifts that are visually plausible but poorly matched to the recognizer. Directly tuning restoration networks with recognition loss is also insufficient, as the restoration model may exploit recognizer-specific shortcut cues and produce visually unnatural artifacts while improving task scores. We propose *Causal Mediated Unrolling (CaMeU)*, an optimization-based framework for task-oriented image restoration with frozen recognition models. Starting from a joint optimization objective, the proposed framework derives a Half-Quadratic Splitting formulation that separates restoration into alternating task-driven and manifold transitions. The task-driven update introduces an explicit mediator for front-door guided optimization, reducing shortcut bias induced by the recognizer. The prior-guided update performs manifold projection, pulling the intermediate result back toward the natural-image manifold to preserve visual quality. Through this mediated alternating process, the restored image is progressively optimized for both recognition performance and visual quality. Experiments on VOC Haze, VOC Dark, and degraded CUB-200-2011 show that CaMeU consistently improves downstream recognition while reducing visual artifacts. The code is available at `https://github.com/distancemay5/cameu`.

## 1 Introduction

Visual recognition is no longer a passive endpoint of a perception pipeline but a critical causal precursor for decision-making in embodied systems, serving as the foundational signals for physical interaction. To ensure reliable real-world deployment, such systems must maintain robustness under diverse environmental degradations [25, 22, 20], such as haze and low-light conditions. Consequently, the restoration-recognition cascade has emerged as a prevailing paradigm to address this issue, where

*Co-corresponding authors.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

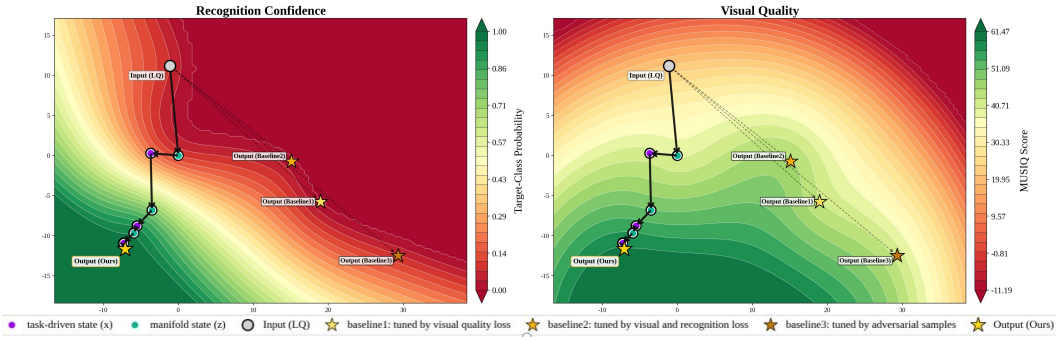


Figure 1: Optimization trajectory of the proposed method. The unrolled process alternates between the task-driven state $x$ and the manifold state $z$, moving restoration toward a region favorable to both recognition and visual quality. Compared with three task-oriented tuning baselines, i.e., using visual-quality loss, using both visual-quality and recognition losses, and using adversarial samples [29], our method iteratively approaches a shared peak with improved visual quality and recognition performance. Recognition and visual quality are measured by target-class probability and MUSIQ [14], respectively.

a restoration module mitigates image degradations before recognition extracts semantic evidence for downstream tasks [10, 28, 22].

However, the efficacy of such cascades is fundamentally limited by an objective mismatch. Conventional restoration models are typically optimized for human-centric perceptual metrics (e.g., PSNR, LPIPS) [40, 28], which tend to suppress high-frequency discriminative cues vital for machine perception recognition [28, 42, 10, 7, 33]. While joint training offers a potential remedy [17, 41, 32, 19, 42], it often suffers from the optimization conflict between two heterogeneous tasks and incurs prohibitive computational costs for joint training. Moreover, such task-specific tuning frequently leads to a performance trade-off, where gains under adverse conditions come at the expense of accuracy on clean data.

A straightforward alternative is to guide a restoration model using the gradients of a frozen recognizer [10]. Yet, this approach inadvertently forces the restoration process to overfit to a fixed and potentially biased recognition signal. Given that pretrained recognizers often rely on non-robust predictive patterns [12], optimizing for recognition loss alone can reward adversarial-like artifacts that boost task confidence without restoring true semantic fidelity [15, 4]. We identify this as a causal bias problem: the recognizer's shortcut preferences act as an unobserved confounder in the restoration-to-recognition transition. Direct optimization thus captures a biased association between restored pixels and task scores, rather than the true causal effect of semantically faithful restoration.

To address these challenges, we propose a Causal-Mediated Unrolling (CaMeU) framework, which decomposes the restoration-recognition joint alignment problem into a sequence of tractable transitions by unrolling the optimization trajectory. Rather than treating Half-Quadratic Splitting (HQS) as a mere numerical solver, we utilize its mathematical structure to establish a dual-state evolution framework. This formulation assigns the conflicting requirements of recognition utility and visual naturalness to two coupled latent states: the task-driven state $x$ and the prior-guided manifold state $z$.

*1) Task-driven Transition ($z \rightarrow x$):* By isolating this transition through HQS splitting, we gain the architectural interface to inject front-door causal intervention. We route task evidence through an observable mediator, effectively filtering out recognizer-induced shortcut biases and ensuring the update remains grounded in robust semantic features.

*2) Manifold Transition ($x \rightarrow z$):* This stage acts as a structural counterweight, projecting the task-driven state back onto the natural-image manifold. It ensures that the restoration trajectory remains anchored to physical authenticity, preventing the generation of "recognizer-pleasing" but visually distorted artifacts.

Through this iterative alternation, the restored image is progressively moved toward a region favorable to both recognition performance and visual quality, as illustrated in Fig. 1. We evaluate the proposed framework on restoration-assisted detection and fine-grained classification under haze, low-light,

and blur degradations, showing consistent improvements in both recognition performance and visual quality.

In summary, our main contributions are as follows:

- We formulate restoration-recognition alignment as a joint task-and-prior dual-state optimization problem and derive an HQS-unrolled framework that provides a principled foundation for balancing task utility with visual naturalness.
- We introduce a front-door mediated task transition for the task-driven state update, which routes task evidence through an observable mediator to reduce recognizer-induced shortcut bias under frozen recognition supervision.
- We instantiate the proposed framework as an end-to-end unrolled restoration model and demonstrate consistent improvements on restoration-assisted detection and fine-grained classification under haze, low-light, and blur degradations.

## 2 Related Work

Restoration-recognition cascades have been widely studied for robust visual recognition under degraded conditions [28, 10, 42, 7, 33, 22]. These works show that improving visual appearance alone does not necessarily improve recognition, because the restored image must also preserve the semantic evidence and feature distribution required by the downstream model [28, 10, 42]. To reduce this gap, prior methods integrate restoration modules into detection or recognition pipelines, or design recognition-driven restoration frameworks for multiple degradations [17, 41, 32, 19, 42, 34, 35]. Although effective, these methods often involve retraining or adapting the downstream recognizer, making them less suitable when the recognizer is fixed.

Task-oriented restoration under frozen recognition supervision addresses this setting by optimizing the restoration model with downstream task feedback while keeping the recognizer unchanged [10, 13, 23, 33, 42, 3]. However, directly using recognition losses as supervision exposes restoration to the biased decision rules of the frozen model. Since pretrained recognizers may rely on non-robust or shortcut features [12], task-driven restoration can amplify recognizer-preferred artifacts instead of recovering semantically faithful image content [29, 15]. Existing adversarial, alternating, or unified restoration methods mainly address the task-quality trade-off through objective design or optimization strategy [29, 15, 3]. Our method instead regulates the recognition-driven transition itself with a mediator, followed by a prior-guided manifold transition for visual naturalness. Deep unfolding is related to our implementation, as it maps iterative optimization into trainable restoration networks [26, 16, 39]. We use the HQS alternating structure to instantiate the two transitions, organizing task-oriented restoration into recognition-driven adaptation and prior-guided projection within a unified unrolled process.

## 3 Method

### 3.1 Overview

**Problem formulation.** Given a frozen downstream recognition model $\phi(\cdot)$, task-oriented restoration aims to recover an image $x$ that is both recognition-effective and visually natural. We formulate this objective as

$$\min_{x} \ \mathcal{L}_{\text{task}}(\phi(x), y_{\text{task}}) + \lambda \mathcal{R}(x), \tag{1}$$

where $\mathcal{L}_{\text{task}}$ is the downstream task loss, $y_{\text{task}}$ denotes the task annotation, $\mathcal{R}(x)$ is a natural-image prior, and $\lambda$ controls the prior strength. The recognizer $\phi(\cdot)$ is kept fixed throughout restoration training and inference.

**Dual-state transition via HQS.** To decouple task adaptation from natural-image regularization, we formulate Eq. (1) with Half-Quadratic Splitting (HQS) [9] by introducing an auxiliary variable $z$:

$$\min_{x,z} \mathcal{L}_{\text{task}}(\phi(x), y_{\text{task}}) + \lambda \mathcal{R}(z) + \frac{\beta}{2}\|x - z\|_2^2, \tag{2}$$

where $x$ denotes the task-driven state, $z$ denotes the prior-guided manifold state, and $\beta$ controls their coupling strength.

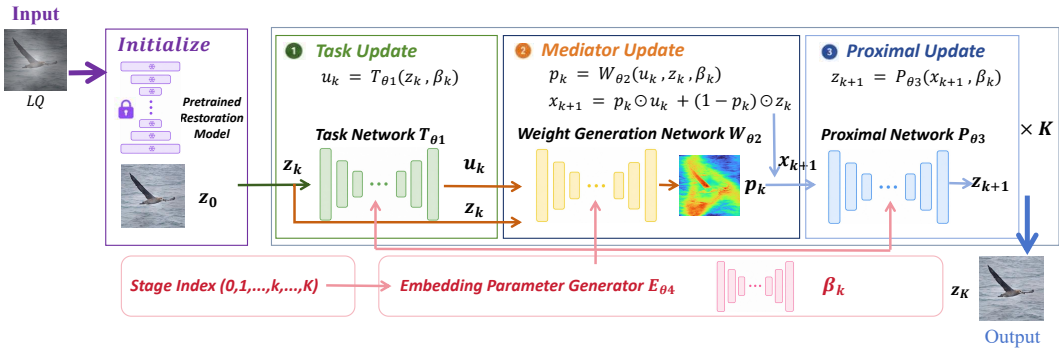


Figure 2: Framework of the proposed method. Each stage performs task update, mediator update, and proximal manifold projection.




(a) Direct task-driven update.

(b) Our front-door mediated update under intervention.

Figure 3: Direct and mediated task update. $C$ denotes the latent shortcut bias induced by the frozen recognizer. (a) Direct updating leaves the transition $Z_k \to X_{k+1}$ confounded. (b) Our redesigned graph forms a task-oriented proposal $U_k$ from the manifold anchor $Z_k$, and introduces $P_k$ as the observable mediator that regulates the transition to the next task state $X_{k+1}$.

At stage $k$, Eq. (2) yields two alternating subproblems:

$$x^{k+1} = \arg\min_x \left( \mathcal{L}_{\text{task}}(\phi(x), y_{\text{task}}) + \frac{\beta_k}{2} \|x - z^k\|_2^2 \right), \tag{3}$$

and

$$z^{k+1} = \arg\min_z \left( \lambda \mathcal{R}(z) + \frac{\beta_k}{2} \|x^{k+1} - z\|_2^2 \right). \tag{4}$$

Here, the $x$-step corresponds to the task-driven transition $(z \to x)$, which adapts the current manifold anchor toward recognition utility, while the $z$-step corresponds to the manifold transition $(x \to z)$, which projects the task-driven state back toward the natural-image manifold.

In the unrolled network, the $x$-step is not solved explicitly. Instead, it is approximated by an amortized task solver that predicts a task-oriented proposal from the current manifold state. Since this proposal is supervised through the frozen recognizer, directly accepting it may inherit recognizer-specific shortcut bias. CaMeU therefore introduces an explicit mediator and applies front-door adjustment at the stage level to regulate the task-driven transition. The mediated task state is then passed to a proximal network, which implements the manifold transition by projecting it back toward the natural-image manifold. As illustrated in Fig. 2, the task proposal, mediator update, and proximal projection together form one CaMeU stage.

### 3.2 Dual-State Transitions

The HQS decomposition in Sec. 3.1 gives two coupled transitions at each unrolled stage: a task-driven transition $(z \to x)$ and a manifold transition $(x \to z)$. The former updates the current manifold state toward recognition utility, while the latter projects the task-driven state back toward the natural-image manifold. We instantiate these two transitions as the core stage-wise update of CaMeU, as illustrated in Fig. 2.

**Task-driven transition** $(z \to x)$**.** For the $x$-subproblem, we use amortized optimization to avoid explicitly solving the task-driven update at each stage. The amortized task solver is instantiated as

$$U_k = T_{\theta_1}(Z_k; \beta_k), \tag{5}$$

where $U_k$ is the task-oriented proposal predicted from the current manifold anchor $Z_k$, and $\beta_k$ is the stage-dependent coupling parameter used for conditioning.

Since the amortized solver is optimized through the frozen recognizer $\phi(\cdot)$, the resulting task-oriented transition is influenced by the recognizer's latent shortcut preference. We denote this unobserved factor as $C$. In the original stage-wise causal graph in Fig. 3(a), $C$ confounds the task-driven transition from the current manifold anchor $Z_k$ to the next task state $X_{k+1}$, making the direct transition shortcut-prone and not directly deconfoundable. To address this issue, we introduce an observable mediator $P_k$ into the stage-wise transition, yielding an explicit intermediate variable for front-door adjustment.

As shown in Fig. 3(b), the mediated graph routes the task-driven effect from $Z_k$ to $X_{k+1}$ through the observable mediator $P_k$, which is estimated from $U_k$ and $Z_k$. Since $P_k$ is constructed from observable stage variables and the backdoor paths from $P_k$ to $X_{k+1}$ are blocked by conditioning on $Z_k$, the graph admits front-door adjustment for the task-oriented transition [27].

We therefore consider the interventional effect $P(X_{k+1} \mid do(Z_k = z_k))$. As shown in Fig. 3, the intervention $do(Z_k = z_k)$ cuts the incoming confounded edge from the latent recognizer bias $C$ to $Z_k$, thereby removing the shortcut-prone influence carried by the observational transition and isolating the task effect to be mediated through $P_k$.

Under the redesigned graph in Fig. 3, the front-door adjustment gives

$$\begin{aligned} &P(X_{k+1} \mid do(Z_k = z_k)) \\ &= \sum_p P(P_k = p \mid Z_k = z_k) \sum_{z'} P(X_{k+1} \mid P_k = p, Z_k = z') P(Z_k = z'). \end{aligned} \tag{6}$$

Since the unrolled network performs stage-wise inference along a single trajectory, the state distribution in Eq. (6) is approximated by the current iterate. We therefore use the empirical distribution

$$P(Z_k = z') \approx \delta(z' - z_k), \tag{7}$$

which collapses the inner summation over $z'$. Eq. (6) then becomes

$$P(X_{k+1} \mid do(Z_k = z_k)) \quad \approx \sum_p P(P_k = p \mid Z_k = z_k) P(X_{k+1} \mid P_k = p, Z_k = z_k). \tag{8}$$

Eq. (8) marginalizes over the mediator $P_k$. We parameterize the mediator as a conditional random variable:

$$P_k \sim q_\theta(P_k \mid U_k, Z_k; \beta_k), \tag{9}$$

where $U_k = T_{\theta_1}(Z_k; \beta_k)$ is the task-oriented proposal generated from the manifold anchor $Z_k$. The mediator marginalization can then be written as

$$P(X_{k+1} \mid do(Z_k = z_k)) \quad \approx \mathbb{E}_{p \sim q_\theta(\cdot \mid U_k, Z_k; \beta_k)} \left[ P(X_{k+1} \mid P_k = p, Z_k = z_k) \right]. \tag{10}$$

For deterministic unrolling, we use the conditional mean of the front-door-adjusted next-state distribution. We constrain this mean transition to lie in the convex hull of the task-oriented state and the manifold anchor, so that the mediator controls a spatially adaptive trade-off between task evidence and natural-image regularity:

$$\mathbb{E}[X_{k+1} \mid P_k = p, Z_k = z_k] \approx p \odot U_k + (1 - p) \odot Z_k. \tag{11}$$

Since $U_k$ and $Z_k$ are fixed at the current stage, the expectation only acts on $P_k$. By the law of total expectation,

$$\mathbb{E}[X_{k+1} \mid do(Z_k = z_k)] \approx \bar{P}_k \odot U_k + (1 - \bar{P}_k) \odot Z_k. \tag{12}$$

where

$$\bar{P}_k = \mathbb{E}_{q_\theta(P_k \mid U_k, Z_k; \beta_k)}[P_k]. \tag{13}$$

Therefore, the deterministic front-door mediated $x$-step is given by

$$X_{k+1} = \bar{P}_k \odot U_k + (1 - \bar{P}_k) \odot Z_k. \tag{14}$$

**Manifold transition** $(x \to z)$**.** After the front-door mediated task-driven transition, the manifold transition is implemented as a proximal projection:

$$Z_{k+1} = \mathrm{P}_{\theta_3}(X_{k+1}; \beta_k), \tag{15}$$

where $\mathrm{P}_{\theta_3}$ is a learnable proximal network conditioned on the stage-dependent coupling parameter $\beta_k$. This projection pulls the task-driven state back toward the natural-image manifold and provides the manifold anchor for the next stage.

As shown in Fig. 2, $T_{\theta_1}$, $W_{\theta_2}$, $\mathrm{P}_{\theta_3}$, and $E_{\theta_4}$ implement the task proposal, mediator update, proximal projection, and stage-dependent parameter generation, respectively. After $K$ stages, the final restored image is $\hat{X} = Z_K$.

### 3.3 End-to-End Training

All learnable components are trained jointly, including $T_{\theta_1}$, $W_{\theta_2}$, $\mathrm{P}_{\theta_3}$, and $E_{\theta_4}$. The overall objective is

$$\mathcal{L}_{\text{total}} = \lambda_{\text{task}}\mathcal{L}_{\text{task}} + \lambda_{\text{img}}\mathcal{L}_{\text{img}} + \lambda_x\mathcal{L}_{x\text{-cons}} + \lambda_z\mathcal{L}_{z\text{-cons}} + \lambda_{\text{kl}}\mathcal{L}_{\text{kl}}. \tag{16}$$

Here, $\mathcal{L}_{\text{rec}}$ is the downstream recognition loss computed with the frozen recognizer, and $\mathcal{L}_{\text{perc}}$ constrains perceptual fidelity. The task loss supervises recognition utility at the proposal, mediated state, and final-output levels:

$$\mathcal{L}_{\text{task}} = \frac{1}{K}\sum_{k=0}^{K-1}\mathcal{L}_{\text{rec}}(\phi(U_k), y_{\text{task}}) + \eta\frac{1}{K}\sum_{k=0}^{K-1}\mathcal{L}_{\text{rec}}(\phi(X_{k+1}), y_{\text{task}}) + \mathcal{L}_{\text{rec}}(\phi(\hat{X}), y_{\text{task}}). \tag{17}$$

Here, the first term drives $T_{\theta_1}$ to produce the task-oriented proposal for the $x$-subproblem, while the latter two terms preserve recognition utility after mediation and after the full unrolled process.

The image-related losses train the proximal $z$-step and constrain the final restoration quality:

$$\mathcal{L}_{\text{img}} = \|\hat{X} - X_{\text{gt}}\|_1 + \mathcal{L}_{\text{perc}}(\hat{X}, X_{\text{gt}}) + \sum_{k=0}^{K-1}\omega_k\left(\|Z_{k+1} - X_{\text{gt}}\|_1 + \mathcal{L}_{\text{perc}}(Z_{k+1}, X_{\text{gt}})\right). \tag{18}$$

This loss trains the proximal projection and keeps both intermediate manifold states and the final output visually faithful.

Finally, we retain the two HQS coupling terms to keep training aligned with the $x$- and $z$-subproblems:

$$\begin{aligned}\mathcal{L}_{x\text{-cons}} &= \sum_{k=0}^{K-1}\frac{\beta_k}{2}\left\|\bar{P}_k \odot (U_k - Z_k)\right\|_2^2, \\ \mathcal{L}_{z\text{-cons}} &= \sum_{k=0}^{K-1}\frac{\beta_k}{2}\left\|Z_{k+1} - X_{k+1}\right\|_2^2.\end{aligned} \tag{19}$$

The first term corresponds to the $x$-step quadratic coupling, and the second corresponds to the $z$-step quadratic coupling. We further include a KL regularization term $\mathcal{L}_{\text{kl}}$ on the conditional mediator distribution in Eq. (9), which constrains $P_k$ toward its prior and stabilizes mediator learning. Together, the task loss trains the amortized task proposal, the image loss trains the proximal manifold projection, and the two consistency terms keep the learned updates aligned with the HQS decomposition.

## 4 Experiments

### 4.1 Experimental Setup

**Datasets.** We conduct experiments on degraded PASCAL VOC 2012 [6] and CUB-200-2011 [30] for object detection and fine-grained classification. For detection, we use VOC-Haze from TogetherNet [32] and synthesize VOC-Dark from PASCAL VOC 2012 following the low-light synthesis pipeline [1, 5]. Models are trained on the corresponding degraded VOC training splits and evaluated on the VOC validation splits. We further evaluate real-world generalization on RTTS [18] for haze and ExDark [24] for low-light conditions. For classification, we use CUB-Dark synthesized with the

Table 1: Quantitative comparison of restoration-assisted object detection with frozen YOLOv10 and YOLOv8. "Direct", "Pre", "+ Adv.", "+ Alt.", and "Ours" denote direct recognition, the pre-trained restoration model, adversarial training [29], alternating optimization [15], and CaMeU, respectively. **For Direct and Pre settings, the detector is not involved in restoration; therefore, LPIPS and MUSIQ are identical across detector backbones.**

| Detector | Method | Fog | | | | | Dark | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | VOC-Haze | | | RTTS | | VOC-Dark | | | ExDark | |
| | | mAP↑ | MUSIQ↑ | LPIPS↓ | mAP↑ | MUSIQ↑ | mAP↑ | MUSIQ↑ | LPIPS↓ | mAP↑ | MUSIQ↑ |
| YOLOv10 | Direct | 49.3 | 62.76 | 0.3937 | 38.9 | 53.76 | 59.3 | 60.69 | 0.3815 | 46.2 | 46.74 |
| | Pre | 76.2 | 72.64 | 0.1281 | 45.1 | 59.56 | 65.8 | 64.19 | 0.1756 | 43.5 | 47.92 |
| | + Adv. | 77.8 | 71.88 | 0.1594 | 45.9 | 59.43 | 66.0 | 63.85 | 0.1910 | 47.2 | 46.72 |
| | + Alt. | 78.5 | 72.97 | 0.1269 | 46.3 | **60.01** | 66.9 | **64.78** | 0.1831 | 47.4 | 48.07 |
| | Ours | **80.4** | **73.01** | **0.1236** | **46.9** | 60.00 | **67.9** | 64.74 | **0.1621** | **48.2** | **48.14** |
| YOLOv8 | Direct | 52.1 | 62.76 | 0.3937 | 43.9 | 53.76 | 62.9 | 60.69 | 0.3815 | 48.1 | 46.74 |
| | Pre | 78.5 | 72.64 | 0.1281 | 48.2 | 59.56 | 66.9 | 64.19 | 0.1756 | 47.5 | 47.92 |
| | + Adv. | 79.9 | 72.00 | 0.1296 | 49.0 | 59.54 | 67.8 | 63.74 | 0.1953 | 48.8 | 46.50 |
| | + Alt. | 80.6 | 73.03 | 0.1238 | 49.5 | 59.93 | 68.4 | 65.09 | 0.1608 | 49.2 | 48.46 |
| | Ours | **81.8** | **73.16** | **0.1214** | **50.0** | **59.97** | **68.8** | **65.12** | **0.1543** | **50.1** | **48.99** |

same low-light pipeline, together with CUB-Haze and CUB-Blur from CUB-C [36, 37]. Models are trained on the degraded CUB training splits and evaluated on the corresponding test splits. These benchmarks cover haze, low-light, and blur degradations in both synthetic and real-world scenarios.

**Implementation Details.** For restoration-assisted detection, we use YOLOv10-n [31] and YOLOv8-n as frozen downstream detectors, both initialized from COCO pre-training [21] and fine-tuned on PASCAL VOC. For restoration-assisted fine-grained classification, we use a frozen ResNet-50 [11] trained on CUB-200-2011 with MMClassification. CaMeU is implemented in PyTorch with $K = 3$ unrolled stages and trained for 50 epochs using AdamW with an initial learning rate of $2 \times 10^{-4}$, weight decay of $1 \times 10^{-4}$, and cosine annealing to $1 \times 10^{-6}$. Gradient clipping with a maximum norm of 5.0 is applied for stable training. All experiments are conducted on a platform with 8 NVIDIA RTX 3090 GPUs. More design details are provided in the supplementary material.

### 4.2 Quantitative Evaluation

We evaluate CaMeU on restoration-assisted object detection and fine-grained classification under haze and low-light degradations, with additional blur evaluation for fine-grained classification. Hazy, low-light, and blurred inputs are restored by IPC-Dehaze [8], Retinexformer [2], and Restormer [38], respectively. We report mAP@50 for detection, accuracy for classification, and MUSIQ [14] and LPIPS [40] for visual quality assessment. We compare direct recognition, pre-trained restoration, adversarial training [29], alternating optimization [15], and CaMeU.

As shown in Tables 1 and 2, CaMeU improves recognition across detection and fine-grained classification while maintaining visual quality. For detection, it achieves the best mAP under foggy and low-light conditions on VOC, RTTS, and ExDark with both YOLOv10 [31] and YOLOv8, showing generalization across frozen detectors and synthetic-to-real benchmarks. For classification, CaMeU obtains the highest accuracy on CUB-Haze, CUB-Dark, and CUB-Blur, indicating its advantage for fine-grained recognition where subtle discriminative details are critical. Compared with baselines, the gains are achieved with competitive MUSIQ and best or near-best LPIPS, suggesting improved task alignment without visually destructive perturbations.

### 4.3 Analysis and Ablation Studies

We analyze CaMeU qualitatively from two aspects: visual comparison against the baseline restoration model, and whether the unrolled trajectory exhibits effective decoupling between task-driven updates and manifold projection.

**Visual comparison against the baseline.** As shown in Fig. 4, we compare CaMeU with the pre-trained restoration baseline under the same frozen recognizer. CaMeU produces restored images that are more favorable for downstream recognition, leading to corrected predictions in several challenging

Table 2: Quantitative comparison on degraded CUB-200-2011. Haze and dark settings use the same pre-trained pre-process model as Table 1, and defocus blur uses Restormer [38]. Abbreviations follow Table 1.

| Method | CUB-Haze | | | CUB-Dark | | | CUB-Blur | | |
|---|---|---|---|---|---|---|---|---|---|
| | Acc.↑ | MUSIQ↑ | LPIPS↓ | Acc.↑ | MUSIQ↑ | LPIPS↓ | Acc.↑ | MUSIQ↑ | LPIPS↓ |
| Direct | 67.22 | 47.73 | 0.2955 | 70.19 | 45.10 | 0.2998 | 66.12 | 22.13 | 0.2422 |
| Pre | 70.91 | 54.48 | 0.3133 | 73.14 | 52.86 | 0.2876 | 68.89 | 36.26 | 0.2515 |
| + Adv. | 71.79 | 48.90 | 0.3382 | 77.18 | 51.32 | 0.3174 | 69.19 | 35.12 | 0.2854 |
| + Alt. | 72.73 | 50.56 | 0.2952 | 77.82 | 54.45 | 0.2195 | 71.45 | **37.41** | 0.2451 |
| Ours | **75.61** | **54.97** | **0.2200** | **79.27** | **60.34** | **0.1420** | **73.30** | 37.28 | **0.2165** |

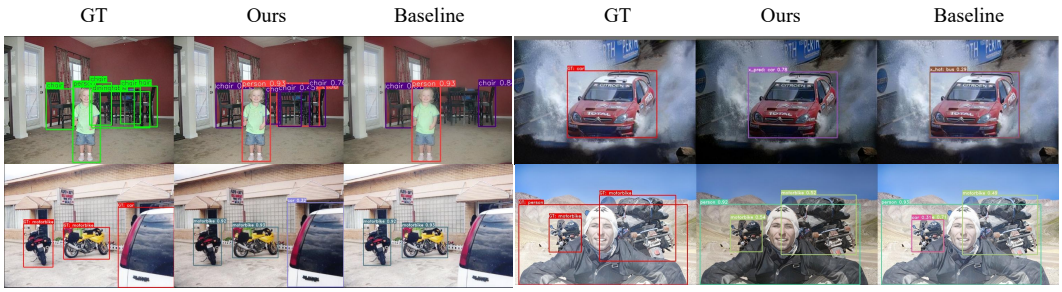


Figure 4: Visual comparison between the baseline restoration and the proposed CaMeU. From left to right: ground truth, our result and baseline without CaMeU. Compared with the pre-trained restoration baseline [2, 8], CaMeU produces outputs that are more favorable for downstream recognition. In particular, it improves object recognizability for previously ambiguous regions (e.g., the table in the upper-left example and the car in the lower-left example), and also corrects misclassified categories (e.g., *bus* → *car* in the upper-right example, and *car* → *motorbike* in the lower-right example).

cases. Notably, these recognition gains do not always correspond to obvious visual changes perceived by humans. Instead, CaMeU adjusts the restored pixel distribution in a task-aware manner, enhancing recognition-relevant cues while keeping the visual appearance close to the baseline. This demonstrates that the proposed mediated update improves recognition not merely by producing visually stronger restoration, but by steering the restoration toward a representation better aligned with the frozen downstream model.

**Dataset-level transition behavior: does the unrolling process converge with decoupled transitions?** We analyze dataset-level update statistics to examine the stability and functional separation of the unrolled optimization. As shown in Fig. 5, the residual magnitudes of both the task-oriented

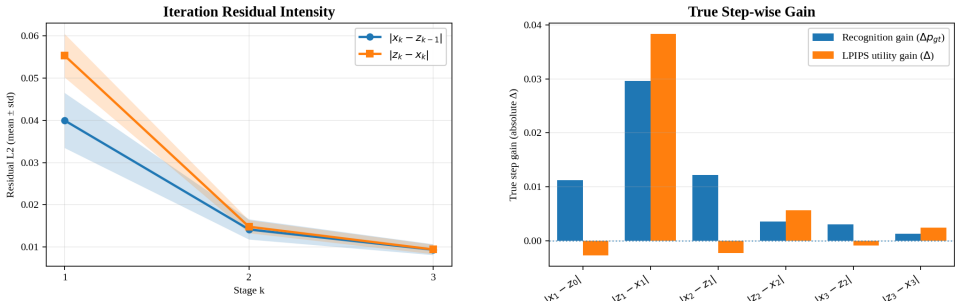


Figure 5: Dataset-level analysis of the unrolled optimization. Left: residual magnitudes of the task-oriented update, i.e., the $x$-step update $(x_k - z_{k-1})$, and the quality-oriented update, i.e., the $z$-step update $(z_k - x_k)$, decrease across stages, indicating stable convergence. Right: step-wise gains in recognition and LPIPS utility show that the $x$-step mainly improves recognition, while the $z$-step mainly improves visual quality, demonstrating the decoupled roles of the two updates.

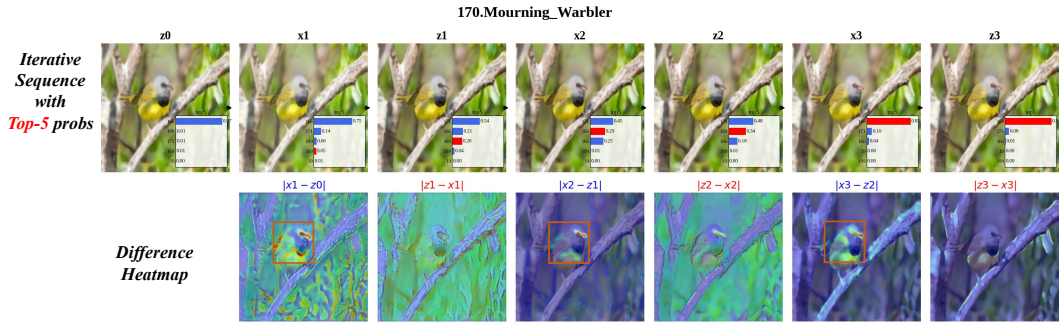


Figure 6: Sample-level trajectory of CaMeU. The iterative sequence shows increasing confidence for the correct class, while the difference heatmaps reveal semantic-focused task updates and global quality-oriented projections.

Table 3: Ablation study on CUB-Blur. We ablate the four training losses and the number of unrolled stages. Top-1 accuracy is used as the evaluation metric.

| ID | Variant / Setting | $\mathcal{L}_{task}$ | $\mathcal{L}_{img}$ | $\mathcal{L}_{x\text{-}cons}$ | $\mathcal{L}_{z\text{-}cons}$ | #Stages | Acc. (%) | $\Delta$ |
|---|---|---|---|---|---|---|---|---|
| A0 | Full Model, Stage = 3 (Ours) | ✓ | ✓ | ✓ | ✓ | 3 | **73.30** | 0.00 |
| A1 | w/o $\mathcal{L}_{task}$ | × | ✓ | ✓ | ✓ | 3 | 69.52 | -3.78 |
| A2 | w/o $\mathcal{L}_{img}$ | ✓ | × | ✓ | ✓ | 3 | 69.05 | -4.25 |
| A3 | w/o $\mathcal{L}_{x\text{-}cons}$ | ✓ | ✓ | × | ✓ | 3 | 72.66 | -0.64 |
| A4 | w/o $\mathcal{L}_{z\text{-}cons}$ | ✓ | ✓ | ✓ | × | 3 | 73.04 | -0.26 |
| A5 | Stage = 1 | ✓ | ✓ | ✓ | ✓ | 1 | 71.67 | -1.63 |
| A6 | Stage = 2 | ✓ | ✓ | ✓ | ✓ | 2 | 72.35 | -0.95 |
| A7 | Stage = 4 | ✓ | ✓ | ✓ | ✓ | 4 | 72.48 | -0.82 |

update, i.e., the $x$-step update $(x_k - z_{k-1})$, and the quality-oriented update, i.e., the $z$-step update $(z_k - x_k)$, decrease across stages, indicating that CaMeU makes progressively smaller modifications and gradually converges. The step-wise gains further show that the $x$-step produces larger recognition gains, while the $z$-step produces larger LPIPS utility gains. This demonstrates that the two updates are functionally decoupled: the $x$-step mainly performs recognition-driven adaptation, whereas the $z$-step mainly performs quality-oriented projection.

**Sample-level update trajectory: how does each step modify the image?** We further inspect a sample-level trajectory to visualize how the alternating updates refine the image during unrolling. As shown in Fig. 6, the confidence of the correct class increases monotonically along the iterative sequence, indicating that the optimization progressively improves recognition. The difference heatmaps further reveal the distinct roles of the two updates. The task-oriented update mainly modifies recognition-critical object regions, such as the bird beak and abdomen, showing a clear semantic focus. In contrast, the quality-oriented update produces more global modifications that improve the overall visual quality. Along the trajectory, the update pattern gradually changes from coarse to fine, suggesting that CaMeU performs structured refinement through the alternating $x$- and $z$-steps.

**Ablation studies.** To analyze the contribution of the training objectives and the unrolled depth, we conduct ablation studies on CUB-Blur. The full model with $K = 3$ is used as the reference configuration, and the results are summarized in Table 3. As shown in Table 3, removing any loss term degrades the performance, indicating that task supervision, image-level restoration, and the two HQS consistency constraints are all necessary for effective unrolled optimization. The stage ablation further shows that $K = 3$ provides the best accuracy, while shallower or deeper unrolling leads to inferior results.

## 5 Conclusion

We propose CaMeU, a task-oriented image restoration framework that alternates front-door mediated task updates and quality-oriented manifold projection. Starting from a joint task-and-prior objective, CaMeU derives an HQS-based unrolled optimization process that separates recognition-driven adaptation from visual-quality regularization. By introducing an explicit mediator, CaMeU reduces shortcut-prone task updates caused by frozen recognizers and guides restoration toward recognition-effective yet visually faithful outputs. Experiments on detection and fine-grained classification under haze, low-light, and blur degradations demonstrate consistent improvements in downstream recognition while preserving visual quality.

**Limitations.** CaMeU is mainly evaluated on detection and fine-grained classification with fixed downstream recognizers. Its generality to broader recognition tasks, more complex real-world degradations, and different recognizer architectures remains to be further studied.

**Broader impacts.** CaMeU can improve recognition robustness under adverse imaging conditions, which may benefit safety-critical perception systems. However, enhanced recognition in degraded scenes may also raise privacy concerns when applied to sensitive visual data.

## Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant Nos. 62376292 and 62636010; the Guangdong Provincial General Fund under Grant No. 2024A1515010208; the Guangzhou Science and Technology Program Project under Grant No. 2025A04J5465; the Research Grants Council of the Hong Kong Special Administrative Region, China, under Project No. UGC/FDS24/E02/26; and the College of Professional and Continuing Education, an affiliate of The Hong Kong Polytechnic University.

## References

[1] Tim Brooks, Ben Mildenhall, Tianfan Xue, Jiawen Chen, Dillon Sharlet, and Jonathan T Barron. Unprocessing images for learned raw denoising. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11036–11045, 2019.

[2] Yuanhao Cai, Hao Bian, Jing Lin, Haoqian Wang, Radu Timofte, and Yulun Zhang. Retinexformer: One-stage retinex-based transformer for low-light image enhancement. In *ICCV*, 2023.

[3] I Chen, Wei-Ting Chen, Yu-Wei Liu, Yuan-Chun Chiang, Sy-Yen Kuo, Ming-Hsuan Yang, et al. Unirestore: Unified perceptual and task-oriented image restoration model using diffusion prior. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 17969–17979, 2025.

[4] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2533–2542, 2021.

[5] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi You, Zenghui Zhang, and Tatsuya Harada. Multitask aet with orthogonal tangent regularity for dark object detection. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2553–2562, 2021.

[6] Mark Everingham, SM Ali Eslami, Luc Van Gool, Christopher KI Williams, John Winn, and Andrew Zisserman. The pascal visual object classes challenge: A retrospective. *International journal of computer vision*, 111(1):98–136, 2015.

[7] Yihua Fan, Yongzhen Wang, Mingqiang Wei, Fu Lee Wang, and Haoran Xie. Friendnet: Detection-friendly dehazing network, 2025.

[8] Jiayi Fu, Siyu Liu, Zikun Liu, Chun-Le Guo, Hyunhee Park, Ruiqi Wu, Guoqing Wang, and Chongyi Li. Iterative predictor-critic code decoding for real-world image dehazing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 12700–12709, 2025.

[9] Donald Geman and Chengda Yang. Nonlinear image recovery with half-quadratic regularization. *IEEE transactions on Image Processing*, 4(7):932–946, 1995.

[10] Muhammad Haris, Greg Shakhnarovich, and Norimichi Ukita. Task-driven super resolution: Object detection in low-resolution images. In *International conference on neural information processing*, pages 387–395. Springer, 2021.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[12] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *Advances in neural information processing systems*, 32, 2019.

[13] Sanket Kalwar, Dhruv Patel, Aakash Aanegola, Krishna Reddy Konda, Sourav Garg, and K Madhava Krishna. Gdip: gated differentiable image processing for object-detection in adverse conditions. *arXiv preprint arXiv:2209.14922*, 2022.

[14] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 5148–5157, 2021.

[15] Jaeha Kim, Junghun Oh, and Kyoung Mu Lee. Beyond image super-resolution for image recognition with task-driven perceptual loss. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2651–2661, 2024.

[16] Shengjiang Kong, Weiwei Wang, Xiangchu Feng, and Xixi Jia. Deep red unfolding network for image restoration. *IEEE Transactions on Image Processing*, 31:852–867, 2021.

[17] Boyi Li, Xiulian Peng, Zhangyang Wang, Jizheng Xu, and Dan Feng. Aod-net: All-in-one dehazing network. In *Proceedings of the IEEE international conference on computer vision*, pages 4770–4778, 2017.

[18] Boyi Li, Wenqi Ren, Dengpan Fu, Dacheng Tao, Dan Feng, Wenjun Zeng, and Zhangyang Wang. Benchmarking single-image dehazing and beyond. *IEEE transactions on image processing*, 28(1):492–505, 2018.

[19] Chengyang Li, Heng Zhou, Yang Liu, Caidong Yang, Yongqiang Xie, Zhongbo Li, and Liping Zhu. Detection-friendly dehazing: Object detection in real-world hazy scenes. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(7):8284–8295, 2023.

[20] Qi Li, Bo Yin, Weiqi Huang, Ruhao Liu, Bojun Zou, Runpeng Yu, Jingwen Ye, Weihao Yu, and Xinchao Wang. Vision-language-action safety: Threats, challenges, evaluations, and mechanisms. *arXiv preprint arXiv:2604.23775*, 2026.

[21] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[22] Yunlong Lin, Zixu Lin, Haoyu Chen, Panwang Pan, Chenxin Li, Sixiang Chen, Kairun Wen, Yeying Jin, Wenbo Li, and Xinghao Ding. Jarvisir: Elevating autonomous driving perception with intelligent image restoration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22369–22380, 2025.

[23] Wenyu Liu, Gaofeng Ren, Runsheng Yu, Shi Guo, Jianke Zhu, and Lei Zhang. Image-adaptive yolo for object detection in adverse weather conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2022.

[24] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Computer vision and image understanding*, 178:30–42, 2019.

[25] Yueen Ma, Zixing Song, Yuzheng Zhuang, Jianye Hao, and Irwin King. A survey on vision–language–action models for embodied ai. *IEEE Transactions on Neural Networks and Learning Systems*, 2026.

[26] Vishal Monga, Yuelong Li, and Yonina C Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021.

[27] Judea Pearl and Dana Mackenzie. *The book of why: the new science of cause and effect*. Basic books, 2018.

[28] Yanting Pei, Yaping Huang, Qi Zou, Yuhang Lu, and Song Wang. Does haze removal help cnn-based image classification? In *Proceedings of the European conference on computer vision (ECCV)*, pages 682–697, 2018.

[29] Shangquan Sun, Wenqi Ren, Tao Wang, and Xiaochun Cao. Rethinking image restoration for object detection. *Advances in Neural Information Processing Systems*, 35:4461–4474, 2022.

[30] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, Serge Belongie, et al. The caltech-ucsd birds-200-2011 dataset. Technical report, Technical Report CNS-TR-2011-001, California Institute of Technology, 2011.

[31] Ao Wang, Hui Chen, Lihao Liu, Kai Chen, Zijia Lin, Jungong Han, and Guiguang Ding. Yolov10: Real-time end-to-end object detection, 2024.

[32] Yongzhen Wang, Xuefeng Yan, Kaiwen Zhang, Lina Gong, Haoran Xie, Fu Lee Wang, and Mingqiang Wei. TogetherNet: Bridging Image Restoration and Object Detection Together via Dynamic Enhancement Learning. *Computer Graphics Forum*, 2022.

[33] Jiawei Wu and Zhi Jin. Unsupervised variational translator for bridging image restoration and high-level vision tasks. In *European Conference on Computer Vision*, pages 214–231. Springer, 2024.

[34] Zizheng Yang, Jie Huang, Jiahao Chang, Man Zhou, Hu Yu, Jinghao Zhang, and Feng Zhao. Visual recognition-driven image restoration for multiple degradation with intrinsic semantics recovery. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14059–14070, 2023.

[35] Zizheng Yang, Jie Huang, Man Zhou, Naishan Zheng, and Feng Zhao. Irvr: A general image restoration framework for visual recognition. *IEEE Transactions on Multimedia*, 26:7012–7026, 2024.

[36] Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, and Pinar Duygulu. A deep dive into adversarial robustness in zero-shot learning. In *European Conference on Computer Vision*, pages 3–21. Springer, 2020.

[37] Mehmet Kerim Yucel, Ramazan Gokberk Cinbis, and Pinar Duygulu. How robust are discriminatively trained zero-shot learning models? *Image and Vision Computing*, 119:104392, 2022.

[38] Syed Waqas Zamir, Aditya Arora, Salman Khan, Munawar Hayat, Fahad Shahbaz Khan, and Ming-Hsuan Yang. Restormer: Efficient transformer for high-resolution image restoration. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5728–5739, 2022.

[39] Kai Zhang, Luc Van Gool, and Radu Timofte. Deep unfolding network for image super-resolution. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3217–3226, 2020.

[40] Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 586–595, 2018.

[41] Zhengxi Zhang, Liang Zhao, Yunan Liu, Shanshan Zhang, and Jian Yang. Unified density-aware image dehazing and object detection in real-world hazy scenes. In *Proceedings of the Asian conference on computer vision*, 2020.

[42] Qing Zhao, Weijian Deng, Pengxu Wei, ZiYi Dong, Hannan Lu, Xiangyang Ji, and Liang Lin. Delving into cascaded instability: A lipschitz continuity view on image restoration and object detection synergy. *arXiv preprint arXiv:2510.24232*, 2025.

# A Technical appendices and supplementary material

This appendix provides additional implementation details and qualitative evidence for CaMeU. We first describe the network instantiation used in the experiments, focusing on how the unrolled formulation in the main paper is implemented for fine-grained classification and object detection. We then provide additional qualitative results, including more visual comparisons, more optimization trajectories, more sample-level unrolled trajectories, and visualizations of the mediator variable $P_k$. These results complement Sec. 3 and Sec. 4 and are intended to support the claims made in the main paper.

## A.1 Additional implementation details

**Unrolled restoration backbone.** CaMeU is implemented as a three-stage HQS-style unrolled network. At each stage, a task proposal network first predicts the task-oriented state $U_k$ from the current manifold anchor $Z_k$. A mediator generator then estimates the spatial mediator $P_k$, which produces the mediated state by adaptively combining the task proposal and the manifold anchor. Finally, a proximal network maps the mediated state to the next manifold state $Z_{k+1}$. Both the task proposal and proximal projection are implemented with lightweight NAF-style residual blocks and are conditioned on the stage index and the learnable coupling parameter $\beta_k$. This implementation follows the formulation in Eq. (14) and Eq. (15), while keeping the number of additional restoration parameters moderate.

**Mediator generator.** The mediator generator takes the task proposal and the manifold anchor as input and predicts a one-channel spatial map $P_k$, together with its complementary weight $1 - P_k$. For fine-grained classification, the mediator uses a frozen MMPretrain classifier backbone to extract multi-level features from both inputs. The feature differences between the task proposal and the anchor are fused with image-level differences to estimate the mediator. For object detection, we use the same mediated unrolling structure, but replace the MMPretrain classifier interface with a frozen YOLO feature backbone. This keeps the role of the mediator unchanged across tasks: it regulates how much task-oriented evidence from $U_k$ should enter the next state while preserving the manifold information carried by $Z_k$.

**Frozen recognition supervision.** The downstream recognizer is kept frozen during training. In the classification setting, the frozen classifier provides both task supervision and intermediate feature guidance. In the detection setting, the detector is also frozen, and the detection loss is computed through a separate raw-head detection-loss module rather than by updating the detector itself. This design is consistent with the problem setting in the main paper, where CaMeU adapts the restoration process to a fixed recognition model instead of jointly fine-tuning the recognizer.

**Initialization and training.** During training, the initial unrolled state is initialized from the output of the corresponding pre-trained restoration model. Specifically, the baseline restored image is used as $z_0$, so CaMeU learns to refine the baseline restoration rather than to restore directly from the degraded input alone. This matches the interpretation of $Z_k$ as a manifold anchor and makes the unrolled process focus on task-aware refinement. The model is trained end-to-end with the hybrid objective described in the main paper, including task loss, image-level restoration loss, mediator regularization, and stage-wise manifold supervision. The training uses distributed data parallelism, AdamW optimization, cosine learning-rate decay, and gradient clipping, following the experimental setup in Sec. 4.

## A.2 Additional visual comparison against the baseline

Fig. 7 indicates that CaMeU adjusts the restored image distribution in a task-aware manner. Compared with the baseline, the texture distribution after CaMeU is better aligned with object-level structures and less dominated by degradation-induced ambiguity. This supports the main conclusion that the proposed mediated update improves recognition utility while keeping the restored image close to the natural-image manifold. The qualitative improvement is therefore not only a perceptual enhancement, but also an alignment of restoration outputs with the evidence expected by the frozen recognizer.

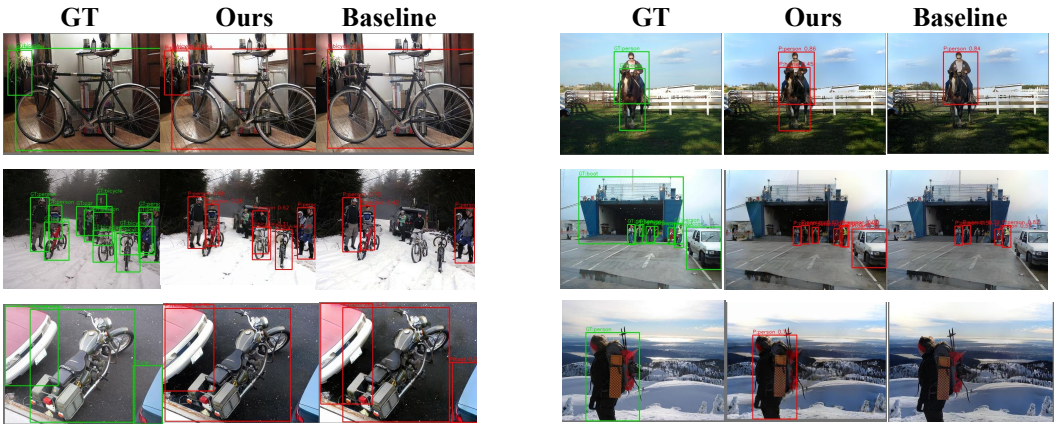


Figure 7: Additional visual comparison between the baseline restoration model and CaMeU. CaMeU preserves the overall natural appearance of the baseline while producing texture distributions that are better aligned with recognition-relevant image structures. In particular, object boundaries, local textures, and degraded regions are refined without introducing strong visually unnatural artifacts, supporting the observation that CaMeU improves downstream recognition by steering restoration toward a recognizer-aligned yet visually faithful solution.

### A.3 Additional optimization trajectory visualization

Fig. 8 provides additional trajectory visualizations following the same interpretation as Fig. 1. The trajectory reflects the optimization behavior induced by the alternating task-driven and prior-guided updates. Across examples, the intermediate states progressively move toward a region with improved recognition response while maintaining competitive visual quality. This behavior is consistent with the HQS-unrolled formulation, where the task-driven state $X_k$ adapts the restoration toward recognition utility and the manifold state $Z_k$ pulls the result back toward natural-image regularity.

### A.4 Additional sample-level trajectory of CaMeU

Fig. 9 shows more sample-level unrolled trajectories. The visual evolution is consistent with Fig. 6: CaMeU improves the restored image through structured stage-wise refinement. In early stages, the update tends to affect relatively broad regions, indicating that the model first corrects coarse degradation patterns and restores the global structure needed for recognition. In later stages, the difference maps become more localized and texture-sensitive, suggesting that the optimization gradually shifts to fine-grained evidence refinement.

The distinction between the two updates is also visible at the sample level. The task-driven transition mainly focuses on recognition-critical regions, such as object parts and discriminative textures. The following manifold transition produces more spatially coherent corrections, which helps maintain natural appearance and suppress task-induced artifacts. This observation supports the main claim that the unrolled process decomposes restoration into recognition-driven adaptation and quality-oriented projection. The additional trajectories therefore provide sample-level evidence for the functional decoupling observed in the dataset-level statistics.

### A.5 Visualization of the mediator variable $P_k$

Fig. 10 visualizes the mediator variable $P_k$ and its effect on the mediated task update. According to Eq. (14), $P_k$ controls the spatially adaptive combination between the task-oriented proposal $U_k$ and the manifold anchor $Z_k$. The visualization shows that $P_k$ does not behave as a uniform mixing coefficient. Instead, it mainly selects high-frequency components from $U_k$, while allowing the lower-frequency structure of $Z_k$ to remain dominant. This is consistent with the intuition that recognition models often rely heavily on high-frequency discriminative evidence, whereas low-frequency components are more closely related to global appearance and visual naturalness.

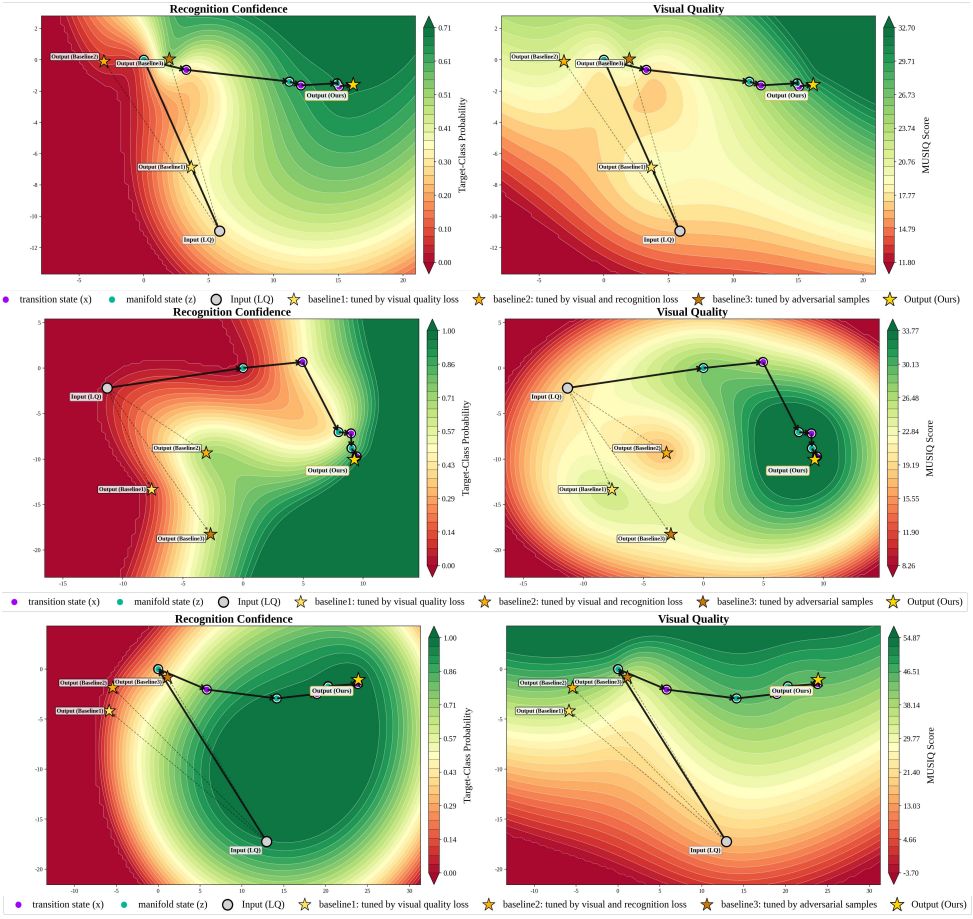


Figure 8: Additional optimization trajectory visualization. Each trajectory visualizes how the restored sample evolves along the unrolled process with respect to recognition utility and visual quality. Similar to Fig. 1, CaMeU progressively moves samples toward a region that is favorable to both downstream recognition and visual quality, while avoiding trajectories that improve one objective at the expense of the other.

Across stages, the frequency pattern selected by $P_k$ becomes progressively finer. In earlier stages, the mediator preserves relatively coarse structures from the task-oriented proposal, which helps correct large degradation-induced discrepancies. In later stages, the preserved components become more localized and high-frequency, corresponding to fine texture and boundary cues. This stage-wise behavior suggests that CaMeU performs a coarse-to-fine mediation process: it first aligns the restored image with broad recognition-relevant structures and then refines finer discriminative details.

This observation provides a more concrete interpretation of the mediator. $P_k$ acts as a stage-adaptive frequency-selective interface between recognition utility and image naturalness. It allows CaMeU to preserve high-frequency task evidence from $U_k$ while retaining the low-frequency manifold structure from $Z_k$. This behavior supports the causal-mediated design in the main paper: the mediator regulates how task-driven information enters the next state, reducing the risk of directly inheriting shortcut-prone updates from the frozen recognizer while still preserving useful recognition cues.

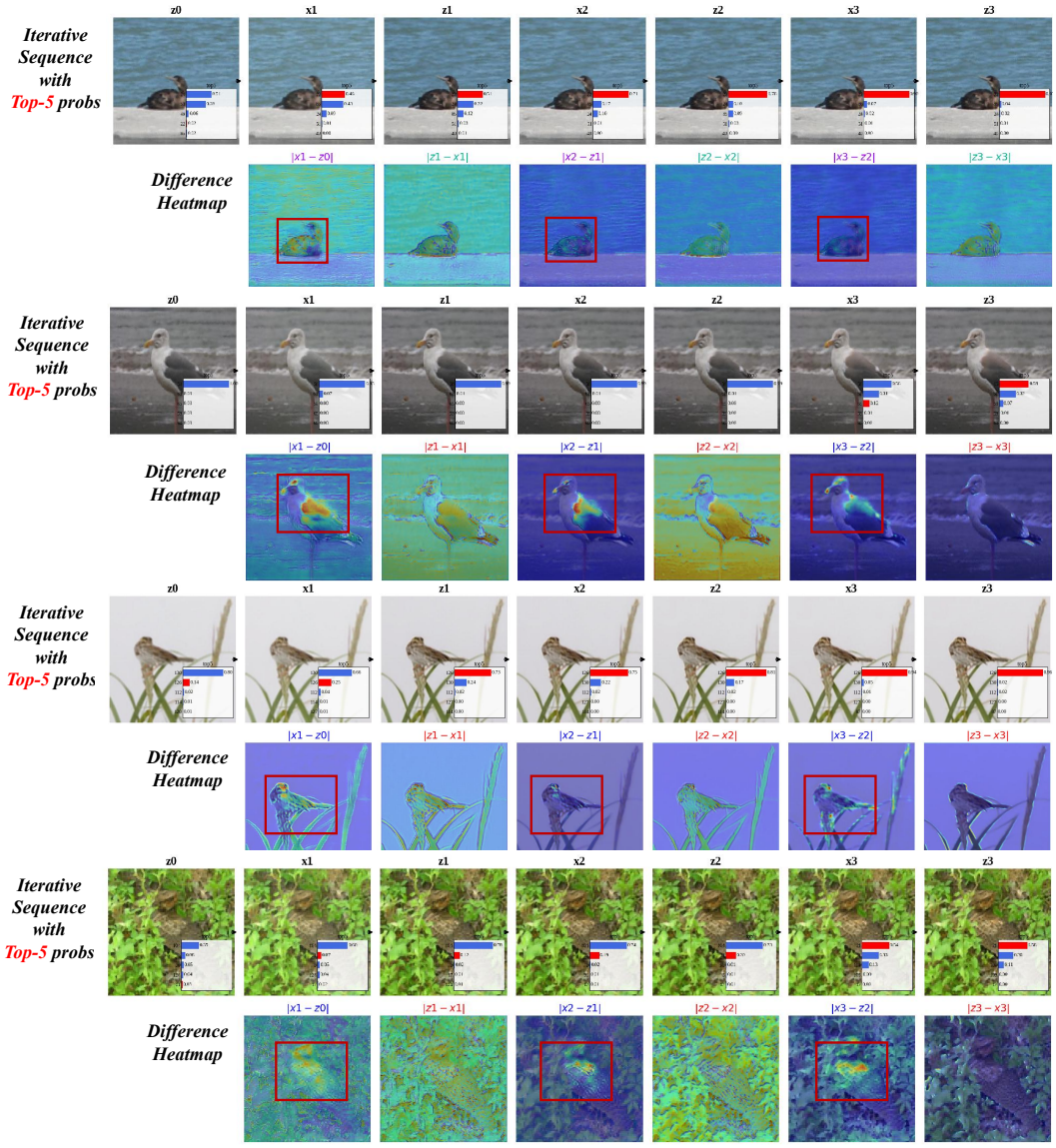


Figure 9: Additional sample-level trajectories of CaMeU. The iterative results and difference maps show how CaMeU refines the restored image across stages. The task-driven updates tend to emphasize recognition-critical local regions, whereas the manifold projections provide broader quality-oriented corrections. The update pattern gradually changes from coarse structural adjustment to finer texture refinement.

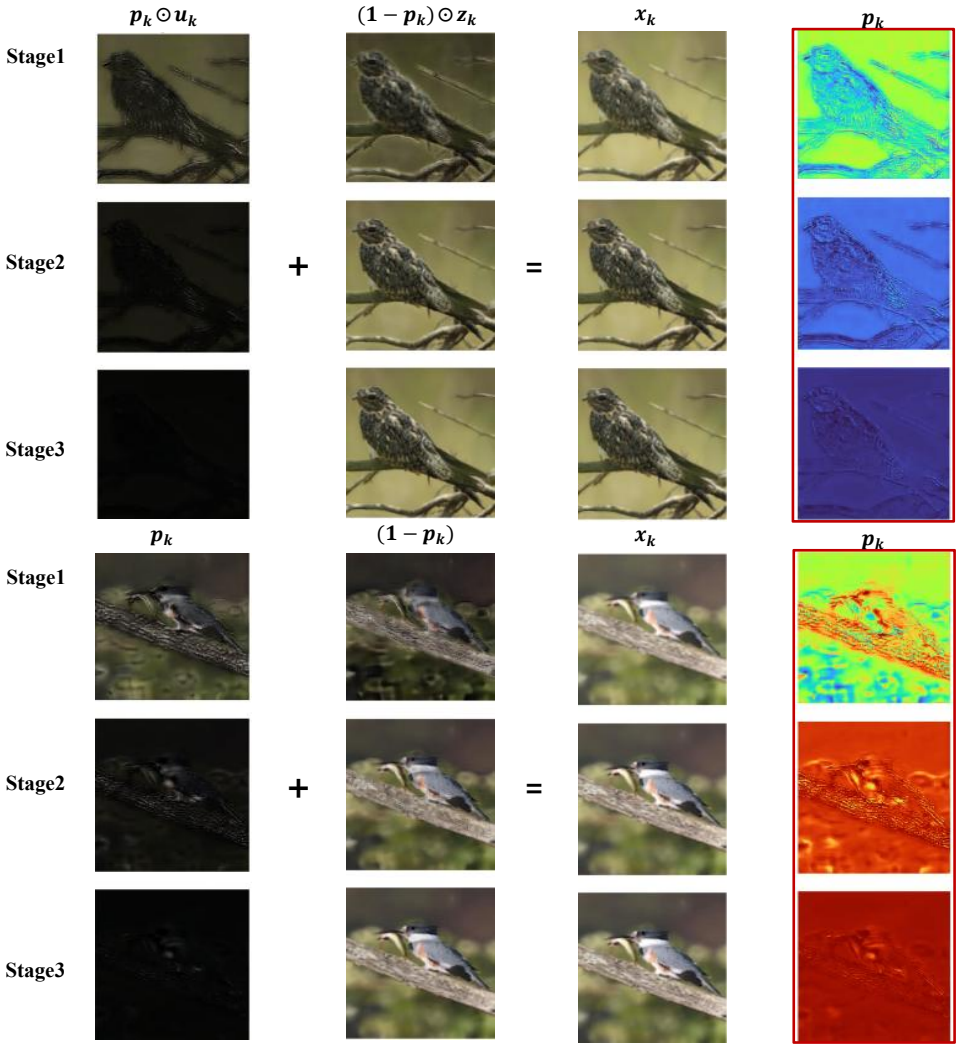


Figure 10: Visualization of the mediator variable $P_k$ across stages. The mediator mainly preserves high-frequency components from the task-oriented proposal $U_k$, while combining them with the lower-frequency structure retained by the manifold anchor $Z_k$. As the stage index increases, the preserved components become increasingly high-frequency and localized, indicating a coarse-to-fine mediation process.

## NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the claims made.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper includes a section discussing limitations.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: This paper provides the proof.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: This paper includes implementation details, and the code will be open-sourced upon acceptance.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [No]

Justification: This paper does not provide open access to data and code. The code will be open-sourced upon acceptance.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: This paper specifies all relevant training and test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: This paper does not report error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: This paper provides details on the type of compute workers (GPUs), memory requirements, and time of execution for each experiment, ensuring reproducibility.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research adheres to the NeurIPS Code of Ethics

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: The paper includes a section discussing potential positive impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: This paper cite the original paper that produced the code package or dataset.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: The paper does not include experiments and research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.