Supplementary Material for *SilverLining*: Data-First Mitigation of Spatial and Spectral Shortcuts Without Introducing New Confounders

A. Computational Cost Analysis

We benchmark SilverLining against existing shortcut mitigation methods to demonstrate its computational efficiency. All experiments were conducted on an NVIDIA L40S GPU using PneumoniaMNIST with 10,000 images (224×224 pixels), training for 20 epochs. Two-stage methods (MaskTune, JTT) use 5 additional epochs for fine-tuning.

Table 3 decomposes the computational requirements across three phases: initial training for shortcut identification, data cleaning operations, and final model training. SilverLining completes the full pipeline in 18.24 minutes, positioning it competitively among shortcut mitigation approaches.

The key insight lies in the time distribution. SilverLining dedicates only 4.70 minutes (26%) to preprocessing—2.92 minutes for debiaser training and 1.78 minutes for data cleaning—with the remaining 13.54 minutes (74%) spent on standard model training. This preprocessing creates a permanently cleaned dataset that can be reused across multiple experiments, architectures, and research teams. In contrast, two-stage methods like MaskTune spend 13.85 minutes (70% of total time) on initial training that gets largely discarded during fine-tuning, and this overhead must be repeated for every experiment.

Memory efficiency further supports practical deployment. SilverLining requires 16.77 GB peak memory, 39% less than MaskTune's 27.56 GB despite performing dual-domain analysis. This efficiency stems from operating the debiaser in inference mode during cleaning, eliminating gradient storage overhead. Baseline ERM maintain minimal memory footprint (7.18 GB) but lack shortcut mitigation capabilities while LfF requires maintaining dual networks in memory during training (14.22 GB, 23.39M parameters).

A.1. Preprocessing as a One-Time Investment

The preprocessing paradigm fundamentally transforms shortcut mitigation from a recurring computational burden into a one-time investment. Once cleaned through SilverLining's pipeline, datasets become permanent assets that can be utilized across diverse architectures, training paradigms, and downstream tasks without any architectural modifications. This model-agnostic approach proves particularly valuable in complex scenarios like object detection, where gradient-based or architecture+loss mitigation methods might fail or not translate directly.

For deployment in clinical settings, models trained on preprocessed data operate identically to standard ERM models—requiring no specialized inference pipelines, additional parameters, or runtime overhead. A hospital can deploy these models using existing infrastructure (11.69M parameters, standard forward pass) while benefiting from robust shortcut mitigation. The ability to separate the computational cost of robustness from deployment overhead is critical for resource-constrained medical environments where inference latency and memory footprint directly impact patient care. Furthermore, preprocessed datasets can be shared across institutions, enabling collaborative research on debiased data without requiring each site to implement complex mitigation strategies.

B. Hyperparameter Specifications

We provide complete hyperparameter values and sensitivity analysis to ensure reproducibility of our results.

B.1. Hyperparameter Values

Table 4 specifies all hyperparameter values used in our experiments.

Key parameter intuitions:

- **K (TopK selection)**: After computing attention weights for all patches, K determines what fraction of the highest-attended patches in a batch we treat as containing shortcuts. For example, K=1/128 means we select the top 1/128 fraction of biased patches from the batch with highest attention weights as shortcut regions. Smaller K values result in more conservative masking (fewer patches removed), while larger K values lead to more aggressive shortcut removal.
- β **(Spectral blending)**: Controls the mixing ratio when replacing shortcut frequencies with reference frequencies. β=0.95 means we retain 95% of the original frequency content and blend in 5% from random unbiased images of the same class. We do not directly mask frequency components (which would create irrecoverable artifacts when inverting the FFT), but instead perform controlled blending to preserve image integrity while reducing spectral shortcuts.
- **Patch size**: Follows standard Vision Transformer design—smaller patches (2×2 for 32×32 images) provide finer spatial control, while larger patches (14×14 for 224×224 images) balance computational efficiency with spatial resolution. Use smaller patches when shortcuts are fine-grained or localized to small regions.

Table 3. Computational cost analysis comparing SilverLining with baseline shortcut mitigation methods. Experiments conducted on PneumoniaMNIST dataset (10,000 images, 224×224 resolution) using NVIDIA L40S GPU. Time breakdown shows three phases: Initial Train (shortcut identification/debiaser training), Data Cleaning (preprocessing operations/error set identification for JTT), and Final Train (standard model training on cleaned data). Two-stage methods (MaskTune, JTT) include 5 epochs fine-tuning after 20 epochs initial training. SilverLining's preprocessing (4.70 min-2.92 min (Initial) + 1.78 min (Cleaning)) creates permanently cleaned datasets reusable across all future experiments, while other methods incur recurring overhead. Peak memory measured during entire pipeline.

Algorithm	Peak Memory	Time (minutes)			
	(GB)	Initial Train	Data Cleaning	Final Train	Total
ERM	7.18	–	–	14.00	14.00
SD	7.18	–	–	13.88	13.88
Margin Control	7.18	–	–	14.08	14.08
GroupDRO	7.18	–	–	13.82	13.82
GRL	7.18	–	–	13.86	13.86
SIFER	10.22	–	–	16.48	16.48
SubG	7.18	–	–	13.86	13.86
MaskTune	27.56	13.85	2.50	3.46	19.81
JTT	7.18	14.04	0.37	3.51	17.92
LfF†	14.22	–	–	25.97	25.97
SLVR (Ours)	16.77	2.92	1.78	13.54	18.24

†LfF uses dual networks during training but single model at inference.

Table 4. Complete hyperparameter specifications for SilverLining across all experimental datasets. Image denotes input resolution (pixels), Patch indicates patch size for spatial analysis. K Spatial and K Spectral control the fraction of top-attended patches identified as shortcuts. β controls spectral blending strength where β=0.95 retains 95% original frequency content while blending 5% from reference images. Values selected based on dataset characteristics: smaller K for subtle shortcuts, larger K for prominent artifacts.

Dataset	**Image Config**		**Shortcut Params**	
	Size	Patch	K (Spat/Spec)	β
Animals	32	2	1/32, 1/128	0.95
Pneumonia	224	14	1/128, 1/256	0.95
X-Ray	224	14	1/128, 1/256	0.95
Polyp	256	16	1/128, 1/128	0.95

Table 5. Sensitivity analysis of spatial hyperparameter K on PneumoniaMNIST dataset. K controls the aggressiveness of shortcut removal: smaller values $(1/512)$ remove fewer patches leading to incomplete shortcut mitigation, while larger values $(1/32)$ risk removing task-relevant features. Optimal $K = 1/128$ balances shortcut removal with feature preservation, achieving best spatial performance (0.9301 AUC)

K **Value**	**Pnu Spatial**	**Pnu Spectral**
1/32	0.9067	0.9712
1/64	0.9241	0.9810
1/128	**0.9301**	0.9848
1/256	0.8943	**0.9854**
1/512	0.8790	0.9835

B.2. Sensitivity Analysis

We conducted sensitivity analysis to validate the robustness of our approach to hyperparameter choices:

The results demonstrate robustness: performance remains stable (AUC $>$ 0.89) across K values spanning two orders of magnitude (1/512 to 1/32). A clear pattern emerges: when K is too large (1/32, masking 3.13% of patches), spatial performance drops to 0.907 as we remove task-relevant features along with shortcuts. When K is too small (1/512, masking only 0.20%), spatial performance decreases to 0.879 as insufficient shortcut removal allows the model to still exploit spurious correlations. The optimal $K = 1/128$ achieves the best balance (0.930 AUC) by removing enough shortcut regions (0.78% of patches) to prevent spurious learning while preserving critical task-relevant features. We keep β fixed at 0.95 throughout all experiments based on preliminary analysis.

B.3. Reference Image Selection

Reference images for spectral correction are randomly sampled from the unbiased training samples (those without shortcuts). This ensures that the performance does not depend on specific reference choices and that the method learns robust frequency patterns rather than overfitting to

particular references. For each image requiring spectral correction, we randomly select a reference from a different class to ensure the blended frequencies do not introduce class-specific patterns.

B.4. Practical Guidelines

Based on our experiments, we provide the following recommendations for applying SilverLining to new datasets:

- **Starting point**: K=1/128 and β=0.95 work well across diverse datasets
- **Adjustment strategy**:
  - For datasets with prominent, large artifacts (amount of area of artifact with respect to image size), use K=1/32 to remove more patches
  - For subtle shortcuts (e.g., texture differences), use K=1/256 for conservative masking
  - We fixed $\beta = 0.95$ across all experiments, which preserves 95% of original frequency content while effectively reducing spectral shortcuts. Lower values (e.g., 0.20) increase reference blending but may introduce visible FFT artifacts—visually inspect debiased images if adjusting.
- **Validation**: Use counter-shortcut evaluation (where shortcuts are reversed) to verify that the model has learned robust features rather than modified shortcuts

Complete configuration files are available in our codebase.

C. Effect of Data Augmentations

We conducted comprehensive experiments comparing standard augmentation techniques against our approach across different shortcut types to understand whether generic robustness methods could achieve similar debiasing effects.

Table 6. Ablation studies on Animals dataset. Comparison of augmentation strategies and masking approaches for shortcut mitigation. Models trained on 95% shortcut-contaminated data, tested on clean data. Mean and standard deviation over 3 runs. **Bold**: best results.

Method	Spatial	Spectral	Both
ERM (Baseline)	.77±.03	.74±.03	.77±.01
CutMix [34]	.79±.01	.79±.01	.79±.01
Mixup [36]	.75±.01	.80±.01	.77±.01
Color Jitter [25]	.75±.01	.78±.01	.77±.01
Brightness Jitter [25]	.76±.01	.82±.01	.79±.02
Naive masking	.79±.00	–	–
SLVR (Best)	**.87**±.00	**.87**±.00	**.83**±.02

As shown in Table 6, while augmentations provide modest robustness improvements—CutMix (0.79 AUC) and brightness jitter (0.82 AUC) outperform baseline ERM (0.77 AUC for spatial, 0.74 AUC for spectral)—they remain insufficient for comprehensive shortcut removal. The improvements are marginal: CutMix achieves only +2% AUC improvement over baseline for spatial shortcuts, while brightness jitter shows +8% for spectral shortcuts.

Most critically, when both spatial and spectral shortcuts are present simultaneously, augmentations achieve at most 0.79 AUC compared to our method's 0.83 AUC. This 4% gap demonstrates that shortcut mitigation requires targeted removal rather than generic robustness techniques. The failure of augmentations stems from their inability to specifically identify and remove shortcut regions—they apply uniform transformations that may reduce but cannot eliminate the spurious correlations that models exploit.

D. Spectral Shortcut Correction Visualization

Supplementary Figure 3 illustrates our spectral shortcut mitigation on the Pneumonia dataset with synthetic brightness shortcut. Panel (C) also shows the corrected pixel intensity histograms for the two classes pre and post corrections.

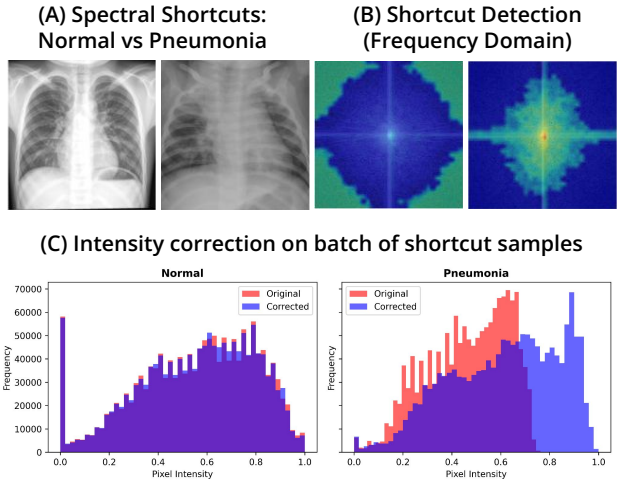

Figure 3. **Spectral shortcut mitigation on Pneumonia dataset.** Normal images are brightened, Pneumonia darkened, creating spurious intensity-diagnosis correlation. (A) Input samples showing brightness difference as a shortcut. (B) Attention on log-magnitude spectrum reveals shortcut detection model focus more on DC component (center), which encodes global intensity. (C) Pixel intensity histograms before (red) and after (blue) correction: Normal distribution shifts left (darker), Pneumonia shifts right (brighter), removing the class-correlated shortcut