Diffusing DeBias: Synthetic Bias Amplification for Model Debiasing

Massimiliano Ciranni* [1], Vito Paolo Pastore* [1,2], Roberto Di Via* [1],
Enzo Tartaglione [3], Francesca Odone [1], Vittorio Murino [2,4]

[1]MaLGa–DIBRIS, University of Genoa, Italy
[2]AI For Good (AIGO), Istituto Italiano di Tecnologia (IIT), Genoa, Italy
[3]Télécom Paris, École Polytechnique, France
[4]Department of Computer Science, University of Verona, Italy

***Equal Contribution**

NeurIPS 2025

The Thirty-Ninth Annual Conference on Neural Information Processing Systems

San Diego (CA), USA, 12/02/2025 – 12/07/2025

Bias definition in Image Classification

- Spurious correlations between class labels and samples;
- Shortcuts learned by models to minimize empirical risk;
- Present in most training samples (bias-aligned);
- Absent in a small percentage (bias-conflicting);
- A model learns these spurious correlations (instead of semantic attributes).

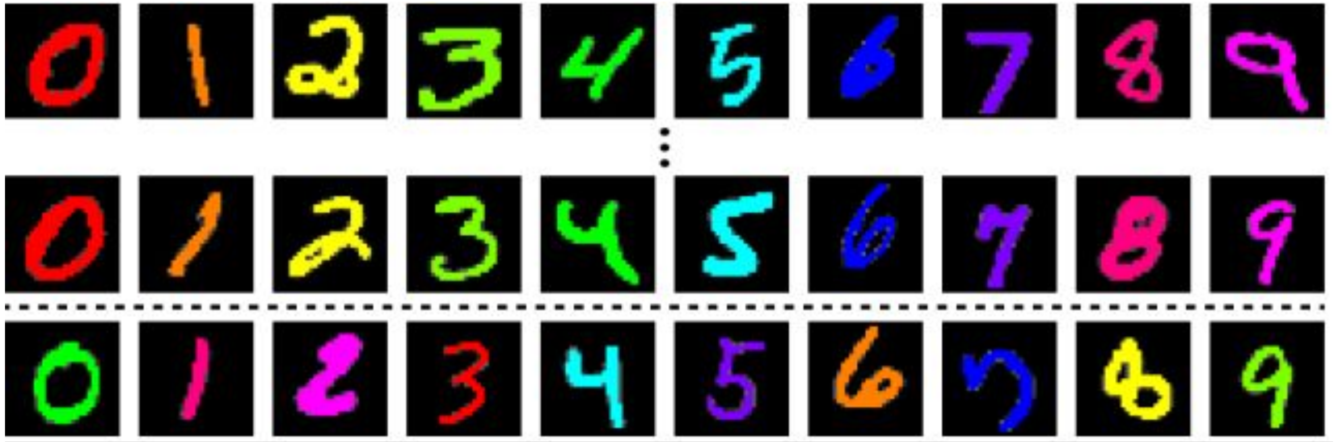
$$\mathcal{D}_{train} = \{x_i,\ y_i\ , b_i\}_{i=1}^{N}$$
$$y^{(j)},\ j \in \{1, \ldots, C\}$$
$$b^{(k)},\ k \in \{1, \ldots, B\}$$

Problematic Bias when:
Most of the samples x_i belonging to class $y^{(j)}$ share the same attribute $b^{(k)}$,
i.e. $|\mathcal{D}_{bias-aligned}| >> |\mathcal{D}_{bias-conflicting}|$.

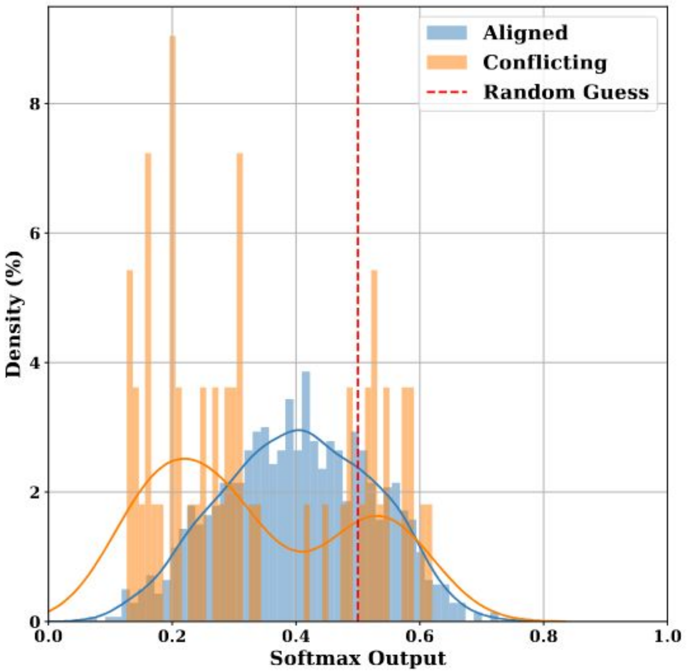
Training a model in this scenario often results in poor Generalization, i.e.

$$\text{Train Error}_{bias-conflicting} << \text{Test Error}_{bias-conflicting}$$

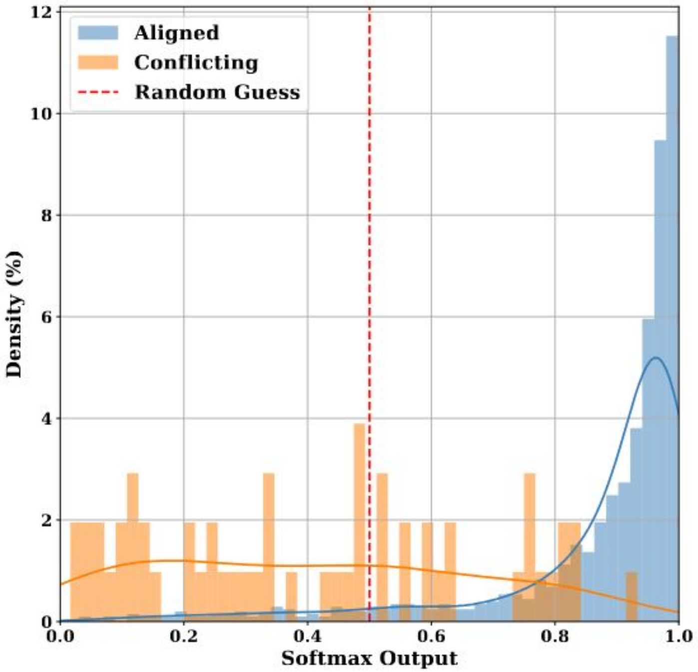
Bias-aligned/biased



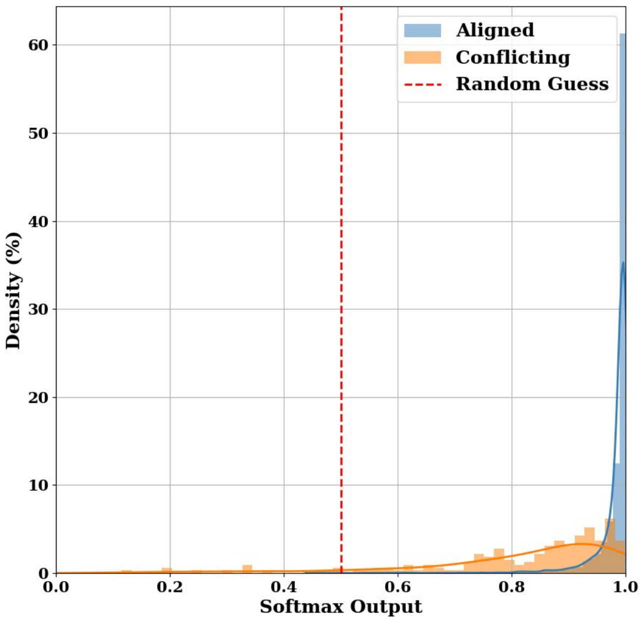
Bias-conflicting/unbiased



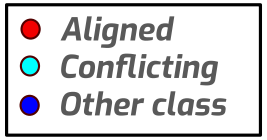
Epoch 0

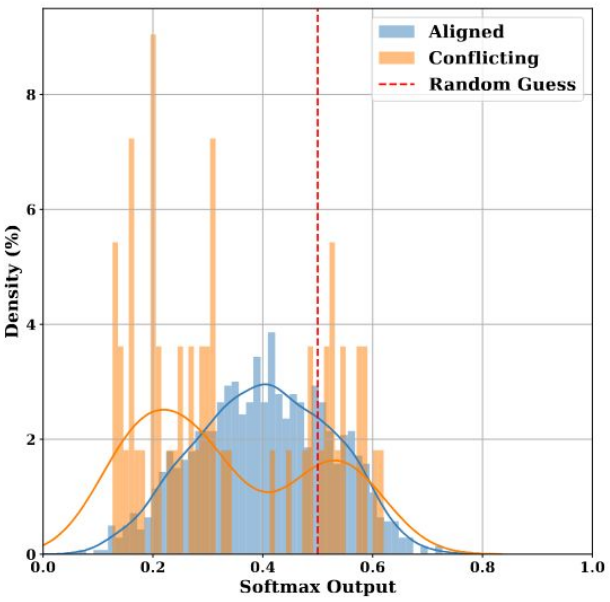


Epoch 6

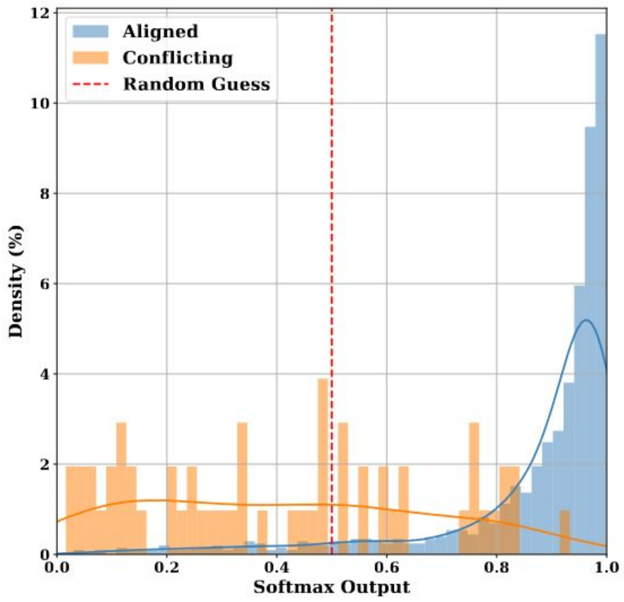


Epoch 10

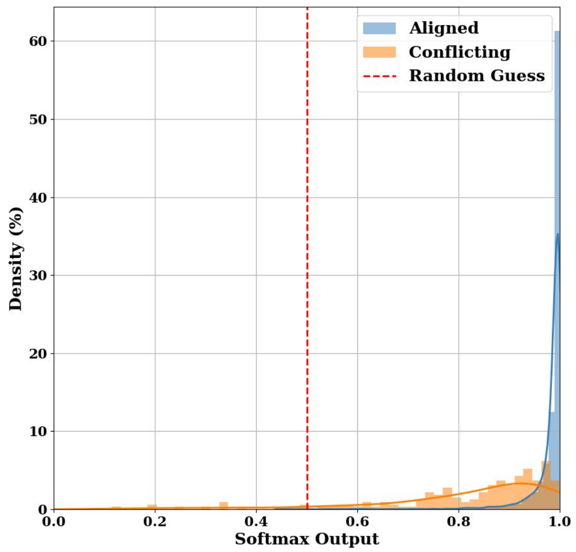




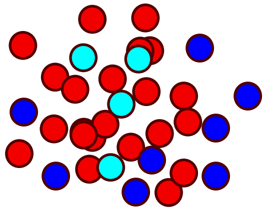
Epoch 0

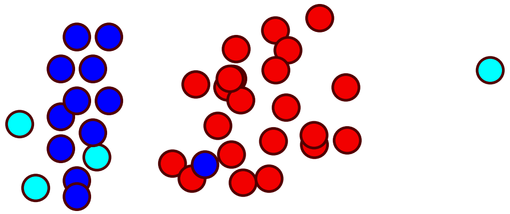


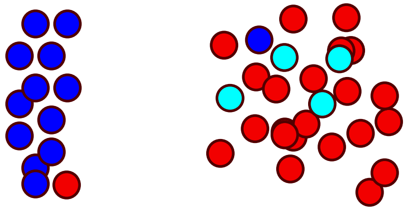
Epoch 6

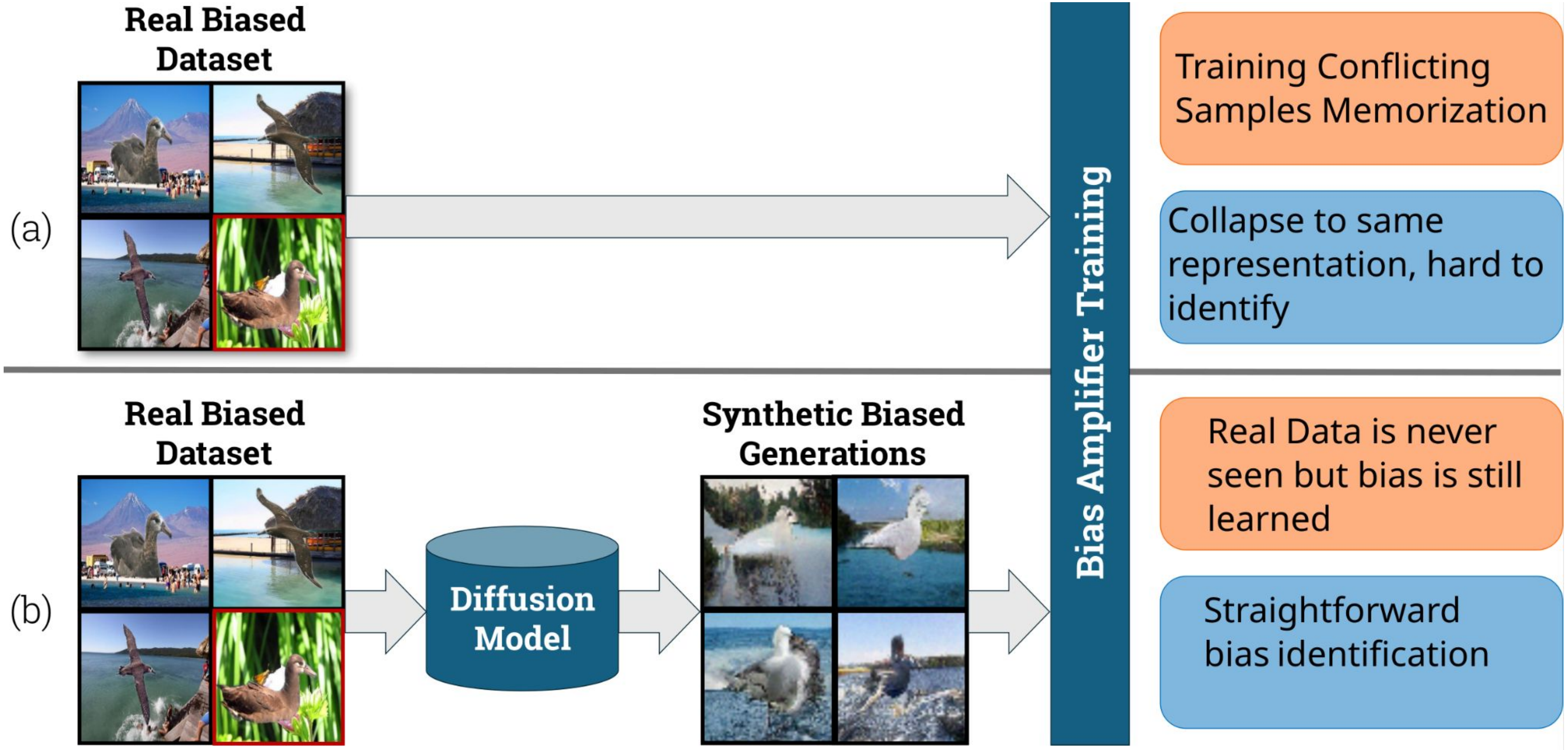


Epoch 10








Real Biased Dataset
(a)
Bias Amplifier Training
Training Conflicting Samples Memorization
Collapse to same representation, hard to identify
Real Biased Dataset
(b)
Diffusion Model
Synthetic Biased Generations
Real Data is never seen but bias is still learned
Straightforward bias identification

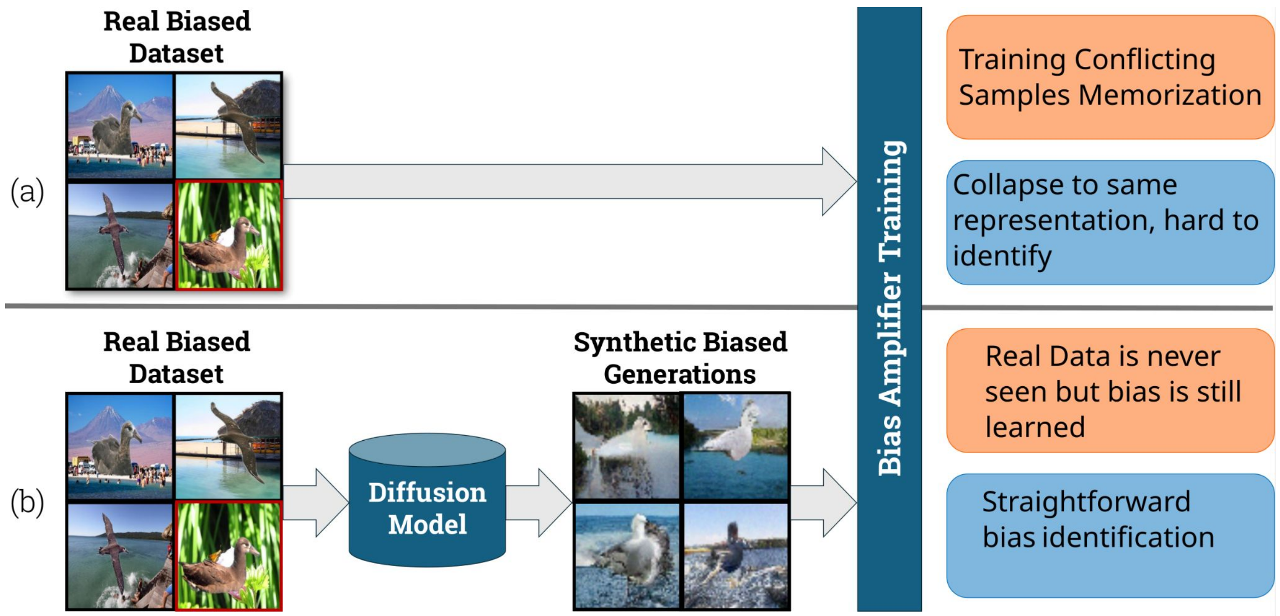
Assume

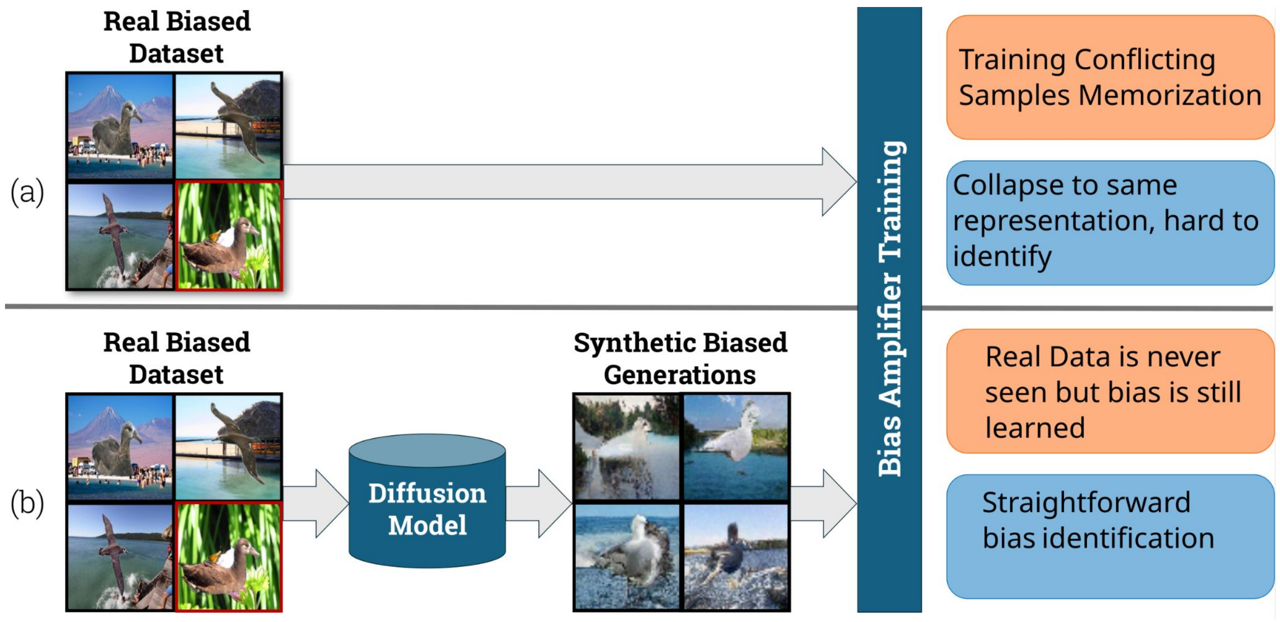
$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

$$\mathcal{D} = \mathcal{D}_{\text{unbiased}} \bigcup \mathcal{D}_{\text{biased}}$$

$$\mathcal{D}_{\text{unbiased}} \sim p_{\text{data}}$$

$$\mathcal{D}_{\text{biased}} \sim p_{\text{data}}\left(\mathbf{x}, y \mid b\right)$$





Assume

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

$$\mathcal{D} = \mathcal{D}_{\text{unbiased}} \bigcup \mathcal{D}_{\text{biased}}$$

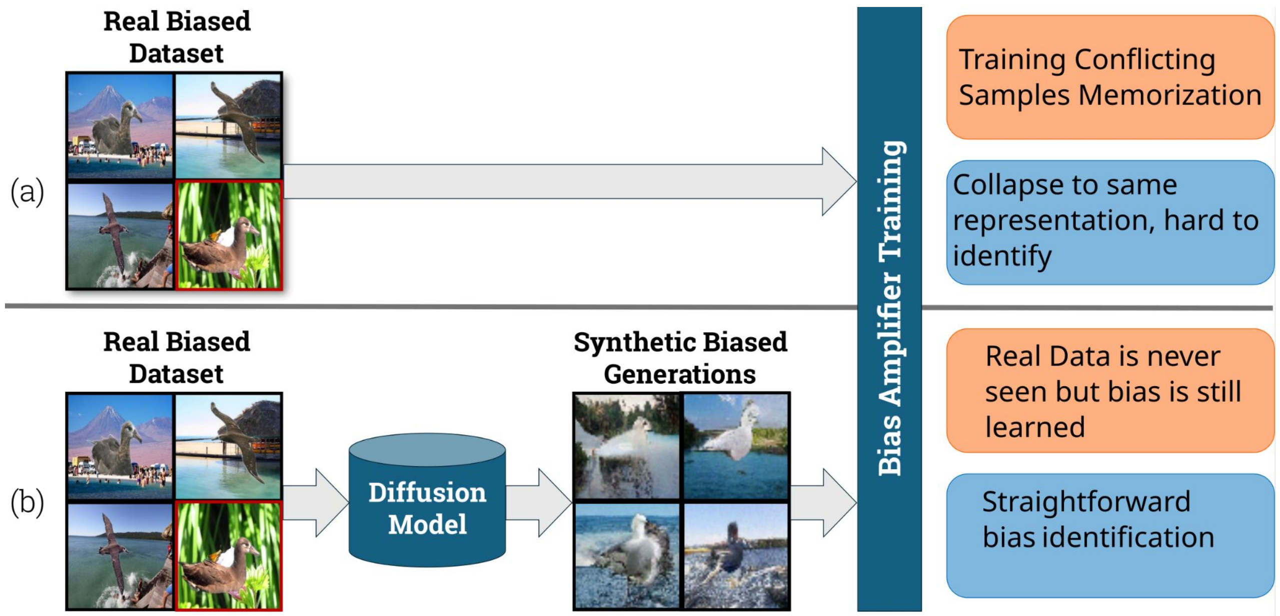
$$\mathcal{D}_{\text{unbiased}} \sim p_{\text{data}}$$

$$\mathcal{D}_{\text{biased}} \sim p_{\text{data}}\left(\mathbf{x}, y \mid b\right)$$

Our hypothesis is that a Conditional Diffusion Model can naturally approximate the biased distribution of the training set.

$$\text{If } |\mathcal{D}_{\text{biased}}| \gg |\mathcal{D}_{\text{unbiased}}|$$

$$\tilde{g}_{\phi}(\mathbf{x} \mid y) \approx p\left(\mathbf{x} \mid y\right)$$



Assume

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

$$\mathcal{D} = \mathcal{D}_{\text{unbiased}} \bigcup \mathcal{D}_{\text{biased}}$$

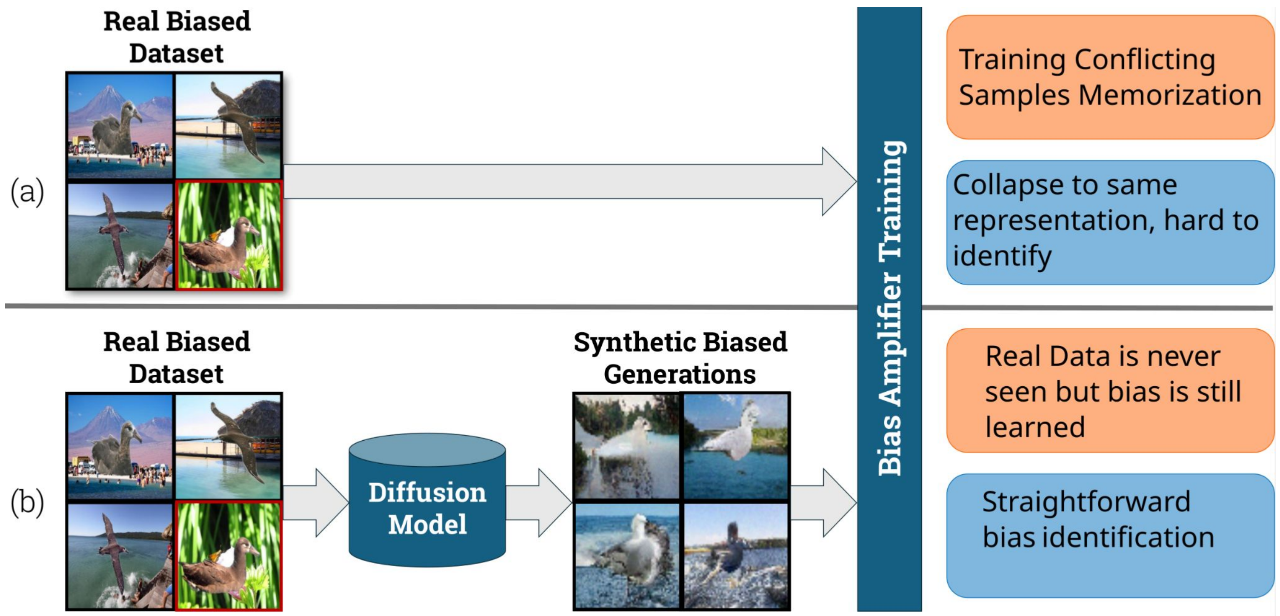
$$\mathcal{D}_{\text{unbiased}} \sim p_{\text{data}}$$

$$\mathcal{D}_{\text{biased}} \sim p_{\text{data}}\left(\mathbf{x}, y \mid b\right)$$

Our hypothesis is that a Conditional Diffusion Model can naturally approximate the biased distribution of the training set.

$$\text{If } |\mathcal{D}_{\text{biased}}| \gg |\mathcal{D}_{\text{unbiased}}|$$

$$\tilde{g}_{\phi}(\mathbf{x} \mid y) \approx p\,(\cancel{\mathbf{x} \mid y})$$



Assume

$$\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^{N}$$

$$\mathcal{D} = \mathcal{D}_{\text{unbiased}} \bigcup \mathcal{D}_{\text{biased}}$$

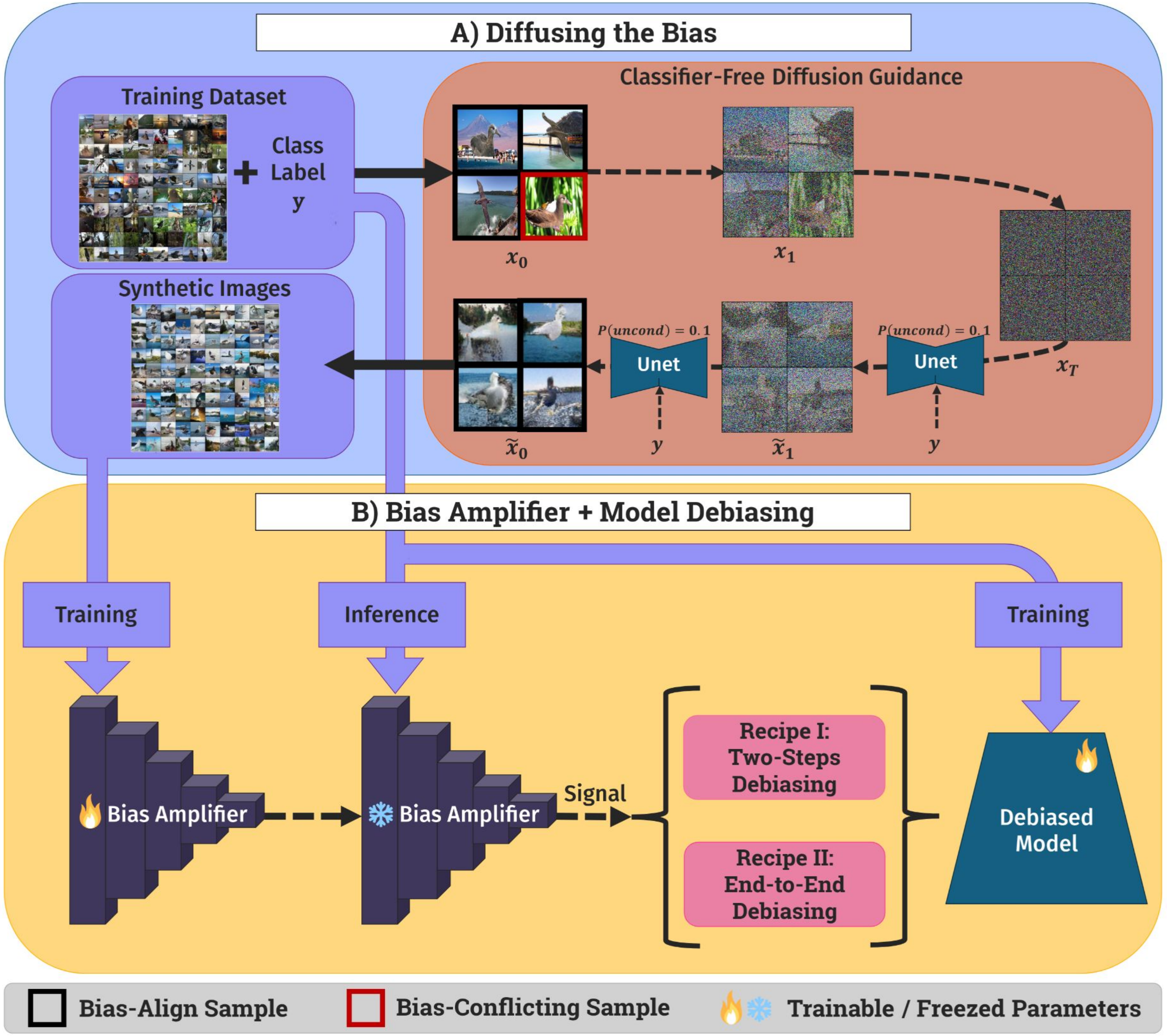
$$\mathcal{D}_{\text{unbiased}} \sim p_{\text{data}}$$

$$\mathcal{D}_{\text{biased}} \sim p_{\text{data}}\left(\mathbf{x}, y \mid b\right)$$

Our hypothesis is that a Conditional Diffusion Model can naturally approximate the biased distribution of the training set.

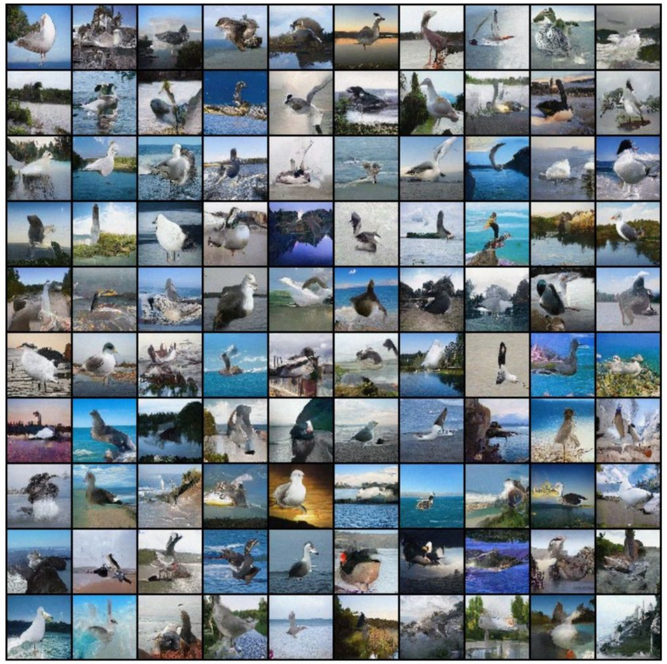
$$\text{If } |\mathcal{D}_{\text{biased}}| \gg |\mathcal{D}_{\text{unbiased}}|$$

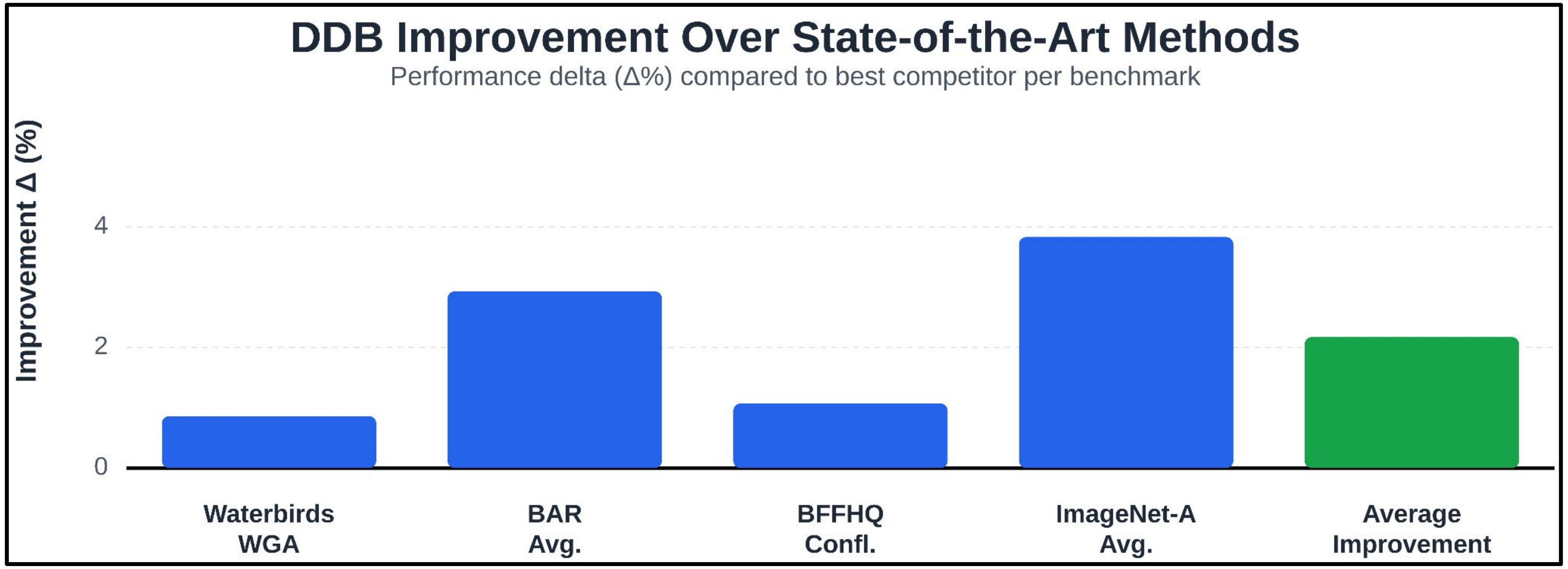
$$\tilde{g}_{\phi}(\mathbf{x} \mid y) \approx p\left(\mathbf{x} \mid b\right)$$



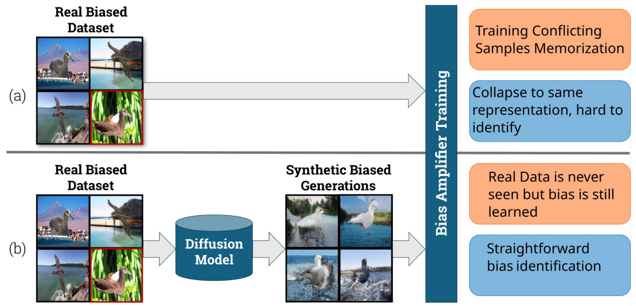
First we leverage a Conditional Diffusion Model (CDPM), to create a **substitute** of the original training set with a synthetic and bias-amplified set of images.

With this new dataset, we train a ***Bias-Amplifier***, which can act as a plug-in for existing debiasing methods.





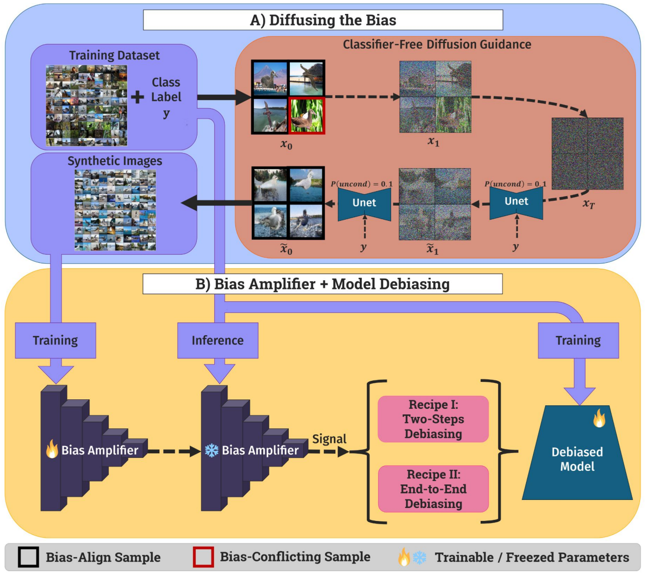
Our Bias Amplifier, regardless of the employed *recipe*, shows improvements over SOTA in all the explored benchmarks, surpassing other plug-in approaches employing our same debiasing methods.



Conclusions and takeaways

In Diffusing DeBias, we show:

- **Diffusion Models** capture **biases in the training distribution.**
- bias-conflicting memorization issue solved by construction using aligned synthetically generated images to train a bias amplifier.
- Our Bias Amplifier improves **Bias Identification capabilities**
- **Acting as a plug-in**, provides improvements over SOTA



Diffusing DeBias: Synthetic Bias Amplification for Model Debiasing

Massimiliano Ciranni* [1], Vito Paolo Pastore* [1,2], Roberto Di Via* [1], Enzo Tartaglione [3], Francesca Odone [1], Vittorio Murino [2,4]

[1]MaLGa–DIBRIS, University of Genoa, Italy
[2]AI For Good (AIGO), Istituto Italiano di Tecnologia (IIT), Genoa, Italy
[3]Télécom Paris, École Polytechnique, France
[4]Department of Computer Science, University of Verona, Italy

*Equal Contribution

GitHub Project Page

ArXiv Preprint