Debiased-CAM for bias-agnostic faithful visual explanations of deep convolutional networks

Wencan Zhang[1], Mariella Dimiccoli[2], Brian Y. Lim[1,*]
[1]Department of Computer Science, National University of Singapore, Singapore
[2]Institut de Robòtica i Informàtica Industrial, CSIC-UPC, Barcelona, Spain
*Corresponding author: brianlim@comp.nus.edu.sg

ABSTRACT

Class activation maps (CAMs) explain convolutional neural network predictions by identifying salient pixels, but they become misaligned and misleading when explaining predictions on images under bias, such as images blurred accidentally or deliberately for privacy protection, or images with improper white balance. Despite model fine-tuning to improve prediction performance on these biased images, we demonstrate that CAM explanations become more deviated and unfaithful with increased image bias. We present Debiased-CAM to recover explanation faithfulness across various bias types and levels by training a multi-input, multi-task model with auxiliary tasks for CAM and bias level predictions. With CAM as a prediction task, explanations are made tunable by retraining the main model layers and made faithful by self-supervised learning from CAMs of unbiased images. The model provides representative, bias-agnostic CAM explanations about the predictions on biased images as if generated from their unbiased form. In four simulation studies with different biases and prediction tasks, Debiased-CAM improved both CAM faithfulness and task performance. We further conducted two controlled user studies to validate its truthfulness and helpfulness, respectively. Quantitative and qualitative analyses of participant responses confirmed Debiased-CAM as more truthful and helpful. Debiased-CAM thus provides a basis to generate more faithful and relevant explanations for a wide range of real-world applications with various sources of bias.

INTRODUCTION

With the growing availability of data, Convolutional Neural Networks (CNN) and other deep neural networks are increasingly capable to achieve impressive performance in many prediction tasks, such as image recognition[1], medical image diagnosis[2], captioning[3] and dialog systems[4]. Despite their superior performance, deep learning models are complex and unintelligible, and this limits user trust and understanding[5–7]. This has driven the development of a wide variety of explainable artificial intelligence (XAI) and interpretable machine learning methods[7–10]. Saliency maps[11–14] are commonly used to provide intuitive explanations for CNN-based image prediction tasks by indicating which pixels or neurons were used for model inference. Amongst these, Class activation map (CAM)[13], Grad-CAM[12] and extensions[15,16] are particularly useful by identifying pixels relevant to specific class labels. Users can verify the correctness of each prediction by

checking whether expected pixels are highlighted. Models would be considered more trustworthy if their CAMs matched what users identify as salient.

Despite the fidelity of CAMs on clean images, real-world images are typically subjected to biases, such as image blurring or color-distortion, and these can affect what CAMs highlight. Blurring can be due to accidental motion[17] or defocus blur[18], or done deliberately to obfuscate details for privacy protection[19]. Images may also be biased with shifted color temperature[20] due to mis-set white balance. These biases decrease model prediction performance[18–20] and we further show that they also lead to deviated or biased CAM explanations that are less faithful to the original scenes. For different bias types (e.g., image blur and color temperature shift), we found that CAMs deviated more as image bias increased (Fig. 1 and Fig. 2: Biased-CAMs from RegularCNN for $\sigma > 0$). Although Biased-CAM represents what the CNN considers important in a biased image, it is misaligned with people's expectations[21], misleads users to irrelevant targets, and impedes human verification and trust[22] of the model prediction. For example, when explaining the inference of the "Fish" label for an image prediction task, Biased-CAMs select pixels of the man instead of the fish (Fig. 1). To align with user expectations, models should not only have the right predictions but also have the right reasons[23,24]; however, current approaches face challenges in achieving this goal, particularly for biased images. First, while retraining the model by fine-tuning on biased images can improve its performance[18,19], this does not seek to improve explanation faithfulness. Indeed, we found that CAMs remain deviated, unfaithful, and biased (Fig. 1, Biased-CAMs from FineTunedCNN for $\sigma > 0$). Conversely, retraining the model with attention transfer[25–27] only improves explanation faithfulness for clean images, but cannot handle biased images. Finally, evaluating human interpretability of explanations requires deep enquiry into user perception, understanding and usage[28–30], but typical evaluations of explainable AI methods involve only data simulations[14,24,31–33] or simple surveys with ratings of explanation trust[12,15,34–36]. Hence, existing methods on image explanation remain lacking to mitigate explanation deviation of biased images due to unspecific training, and limited evaluation of human interpretability.

We retrain the CNN model to improve CAM faithfulness, the similarity of the generated CAM to Unbiased-CAM, the gold-standard CAM of the unbiased image that is considered most representative. Unlike current methods which train models with reference explanations from the same unbiased image source and are prone to biased explanations of biased sources, our approach references explanations from an unbiased source to train a model that recovers unbiased explanations of predictions from a biased source. We thus developed DebiasedCNN, a multi-input, multi-task model that interprets biased images as if predicting on the unbiased form of the images and generates CAMs — Debiased-CAMs — that are more human-relatable and robust under image bias. Debiased-CAMs explain the DebiasedCNN prediction task, in a manner representative of the original unbiased scenes, i.e., Debiased-CAMs are similar to Unbiased-CAM across different bias levels (Fig. 1: DebiasedCNN CAMs for any σ). For example, Fig. 1b shows that Biased-CAMs select fewer relevant pixels (fish) and more spurious ones (person), while Unbiased-CAM and Debiased-CAM select relevant pixels of the fish.

To evaluate the developed model, we conducted four simulation studies and two user studies to address the research questions on 1) how bias in images decreases CAM faithfulness and how well debiasing mitigates this, and 2) how sensitive people are to perceiving CAM deviations and how well debiasing improves perceived CAM truthfulness and helpfulness. For generality, the simulation studies spanned three image prediction tasks and three datasets: blur-biased classification on ImageNette, blur-biased classification on NTCIR-12 wearable camera activities, blur-biased captioning on COCO, and color temperature-biased classification on NTCIR-12. Across all studies, we found that while increasing bias led to a higher CAM deviation, Debiased-CAM showed the best improvement in CAM faithfulness and task performance. Instead of trading off task performance for CAM faithfulness, our debiasing training improved both. We further demonstrated the usability and usefulness of Debiased-CAMs in two controlled user studies with 203 participants (2,547 trials) to validate that users can perceive the improved truthfulness and helpfulness of Debiased-CAMs on biased images. For each user study, we developed precise survey instruments to objectively and subjectively measure user perceptions and opinions of CAM explanations, designed scenarios to set up well-controlled condition exposure, and performed rigorous statistical analysis as well as qualitative thematic analysis to understand CAM usage and user experience. Our results showed that Debiased-CAMs are useful for privacy-preserving applications, such as blurring images captured by wearable cameras; blurring to obfuscate sensitive details in images makes recognizing other concepts more difficult, but Debiased-CAM can help users to more easily "see through the blur" to identify relevant targets. Debiased-CAM thus provides a generalizable framework to enable more faithful and relevant explanations for a wide range of real-world applications, such as other image explanations and non-image tasks, which are subjected to various sources of bias.

RESULTS

Self-Supervised Multi-Bias, Multi-Input, Multi-Task Model for Debiased-CAM

To debias predication explanations made from biased images, we developed the DebiasedCNN model trained with self-supervised learning. Fig. 3 shows the architecture of our debiasing model with a multi-input, multi-task architecture to add a CAM explanation task and bias level prediction task. DebiasedCNN has a modular design: 1) single-task (st) or multi-task (mt) to improve model training; and 2) single-bias (sb) or multi-bias (mb) to support bias-aware and bias-agnostic predictions. We denote the four DebiasedCNN variants as (sb, st), (mb, st), (sb, mt), (mb, mt). For flexibility, the model can substitute different base CNN models and primary prediction tasks.

To generate a CAM saliency map, Grad-CAM[37] computes a gradient-based weighted sum of activation maps from the final convolution filters in the CNN. Although a regularly trained CNN model (RegularCNN) can generate a truthful CAM $\widetilde{\boldsymbol{M}}$ (Unbiased-CAM) of an unbiased image $\boldsymbol{x}$, it produces a deviated CAM (Biased-CAM) of the image under bias $\boldsymbol{x}_b$ at level b, i.e., $\widetilde{\boldsymbol{M}}(\boldsymbol{x}) \neq$

$\widetilde{\boldsymbol{M}}(\boldsymbol{x}_b)$, due to the model not training on any biased images and learning spurious correlations with blurred pixels. While a fine-tuned model (FineTunedCNN) trained on biased images can improve the prediction performance on biased images, it does not significantly improve CAM faithfulness and still generates a deviated CAM $\breve{\boldsymbol{M}}$ for each biased image, as it was not trained on truthful CAMs (Fig. 1a and Fig. 2a-c: CAMs of FineTunedCNN).

We train the DebiasedCNN model to generate Debiased-CAM $\widehat{\boldsymbol{M}}$ from the biased image $\boldsymbol{x}_b$, such that it is similar to Unbiased-CAM $\widetilde{\boldsymbol{M}}$, i.e., $\widehat{\boldsymbol{M}}(\boldsymbol{x}_b) \approx \widetilde{\boldsymbol{M}}(\boldsymbol{x})$. One training approach is to minimize attention or CAM loss between CAM $\widehat{\boldsymbol{M}}$ and Unbiased-CAM $\widetilde{\boldsymbol{M}}$ for a single prediction task, i.e., $L(\boldsymbol{w}) = L_y(y, \hat{y}(\boldsymbol{w})) + \omega_M L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}})$, where $\hat{y}(\boldsymbol{w})$ is the predicted label as a function of the model weights $\boldsymbol{w}$, and L_y and L_M are the classification cross-entropy loss and CAM difference loss, respectively. Here, $L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}})$ is non-differentiable with respect to the weights $\boldsymbol{w}$, i.e., $\partial L_M / \partial \boldsymbol{w} = 0$, so this limits CAM faithfulness as $\widehat{\boldsymbol{M}}$ can still be very deviated from $\widetilde{\boldsymbol{M}}$. To overcome this limitation, we model Grad-CAM as a separate prediction task which is trained with a separate loss function. Unlike the original Grad-CAM approach which calculates a CAM as a weighted sum of activation and gradient measurements[12], our method reformulates the Grad-CAM computation as layers in a secondary prediction task in a multi-input, multi-task CNN model architecture (Fig. 3a, see details in Method 1). This new task allows us to train more specific weight updates via backpropagation into the last convolution layer and into the fully connected block based on differentiable CAM loss; i.e., $\partial L_M / \partial \boldsymbol{w}_M \neq 0$, where the training loss for the CAM task $L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}}(\boldsymbol{w}_M))$ is a function of the model weights $\boldsymbol{w}_M$ for the secondary task, and $\widetilde{\boldsymbol{M}}$ is the Unbiased-CAM from RegularCNN trained on unbiased images. While they allow backpropagation through them, the layers in the new task are frozen and not trainable, so the calculations remain faithful with Grad-CAM and do not change with the model training; hence the CAM explanations still explain the model with its activation maps based on their gradients. Furthermore, any improvement in the model performance is due to more accurately learned weights in the base CNN and not due to weights in the CAM task, since there are no weights there. The CAM task takes input $\boldsymbol{e}_c$ as the second input to the multi-input model. By specifying the prediction class c for the second input $\boldsymbol{e}_c$, the CAM $\widehat{\boldsymbol{M}}$ for class c will be predicted in the secondary task. DebiasedCNN training is generalizable and can be extended to other image prediction tasks (e.g., image captioning: Fig. 3b), different base CNN models (e.g., VGG16, Inception v3, ResNet50, Xception), and for privacy-preserving machine learning (Supplementary Fig. 1). The multi-task model is trained without human annotation using self-supervised learning[38,39] to minimize the differentiable CAM loss between Debiased-CAM $\widehat{\boldsymbol{M}}$ and Unbiased-CAM $\widetilde{\boldsymbol{M}}$ (Fig. 3c); this is unlike model fine-tuning which only trains with classification loss (Fig. 3d), or non-differentiable CAM loss for single-task models (Supplementary Fig. 2c). Furthermore, we model DebiasedCNN as bias-agnostic to predict CAMs across multiple bias levels and bias-aware to predict the bias level of the image. Since image biasing can happen sporadically at run time, the image bias level is unknown at training time. Therefore, instead of training on specific

bias levels[19] or fine-tuning with data augmentation on multiple bias levels[18], we added a tertiary bias level prediction task to DebiasedCNN to leverage on supervised learning (Fig. 3: salmon-colored layers, Method 2). In summary, DebiasedCNN has multiple capabilities to predict Debiased-CAMs at various bias levels, which we detail in Method 3, Supplementary Fig. 2 and Supplementary Table 1.

Simulation studies: DebiasedCNN improves task performance and CAM faithfulness for increasing image bias

To evaluate how well DebiasedCNN recovers CAM Faithfulness of deviated CAMs from biased images, we conducted four simulation studies. Fig. 4 shows the results of our evaluation across different prediction tasks and datasets (Supplementary Table 2) in ablation studies to investigate task performance, and CAM deviation and debiasing of different CNN models at increasing bias levels for different bias types (Method 4); Supplementary Table 3 describes in detail these improvements. For all studies, we measured model Task Performance as the area under the precision-recall curve (PR AUC, Method 5) and CAM Faithfulness as the Pearson's Correlation Coefficient (PCC) and with the Jensen-Shannon Divergence (JSD) between the CAM and Unbiased-CAM (Method 6; JSD results in Supplementary Fig. 3). DebiasedCNN showed improvements across all simulation studies with some differences which we highlight.

In Simulation Study 1, we evaluated CAMs for blur biased images of the object recognition dataset ImageNette[40]. We found that Task Performance and CAM Faithfulness decreased with increasing blur for all CNNs, but DebiasedCNN increasingly mitigated both these decreases (Fig. 4a). This indicates that model training with additional CAM loss complementarily improved model performance, instead of trading-off explainability for performance[41]. RegularCNN had the lowest Task Performance for all blur levels and the lowest CAM Faithfulness for moderate to strong blur levels ($\sigma > 8$). FineTunedCNN (sb, st) marginally improved Task Performance and CAM Faithfulness as compared to RegularCNN. In comparison, trained with differentiable CAM loss, DebiasedCNN (sb, mt) showed marked improvements to both metrics, up to 2.33x and 6.03x over FineTunedCNN's improvements, respectively. Trained with non-differentiable CAM loss, DebiasedCNN (sb, st) improved both metrics to a lesser extent than DebiasedCNN (sb, mt), confirming that separating the CAM task from the classification task in the latter variant enabled better weights update in model training. Trained with an additional bias-level task (Method 2), multi-bias DebiasedCNN (mb, mt) achieved high Task Performance and CAM Faithfulness for all bias levels that is only marginally lower than single-bias DebiasedCNN (sb, mt) which is trained at specific bias levels, because of its good regression performance for bias level prediction (Supplementary Fig. 4). Multi-bias DebiasedCNN generalizes across bias levels better than single-bias DebiasedCNN when evaluated at non-specific bias levels (Supplementary Fig. 5). Finally, all models generated more faithful CAMs when they had a higher Prediction Confidence (Supplementary Fig. 6).

In Simulation Study 2, we evaluated the impact of blur biasing with a more ecologically valid task — wearable camera activity recognition — with the NTCIR-12 dataset[42]. This task represents a real-world use case where egocentric cameras may capture blurred images accidentally due to motion or defocus, or deliberately for privacy protection. We found the same trends as for the ImageNette classification task with some differences due to the increased difficulty of classifying among more classes (Fig. 4b). In particular, the differences in Task Performance and CAM Faithfulness between RegularCNN and DebiasedCNN were amplified, indicating that debiasing is more useful for this application domain, and that RegularCNN could be overfitting to fine-grained image details. Task Performance and CAM Faithfulness decreased steeply for RegularCNN with increasing blur bias, while DebiasedCNN significantly recovered both metrics, demonstrating marginal decreases with increasing bias. FineTunedCNN marginally increased CAM Faithfulness from RegularCNN (<44%), while DebiasedCNN achieved a much larger improvement by up to 229%. We verified these trends for different base CNN models and found that more accurate models produced more faithful CAMs even for higher blur levels (Supplementary Fig. 7 and Supplementary Fig. 8). Hence, Debiased-CAM enables privacy-preserving wearable camera activity recognition with improved performance and faithful explanations.

In Simulation Study 3, we evaluated the influence of blur biasing on a different prediction task — image captioning — with the COCO[43] dataset. We found similar decreases in Task Performance and CAM Faithfulness as before, though all CNN models performed poorly at all blur levels (Fig. 4c). Furthermore, CAM Faithfulness was low for all models, even for RegularCNN at a very small blur bias ($\sigma = 1$). This could be due to captioning being a much harder task than classification, such that CAM explanations were inaccurate even for barely biased images. Yet, DebiasedCNN improved CAM Faithfulness for all blur levels by up to 224% from RegularCNN, indicating the importance of attention transfer at the model's convolution layers from Unbiased-CAM to retain CAM faithfulness that is readily lost due to bias.

In Simulation Study 4, we evaluated the influence of a different bias type — color temperature bias, due to improper white balance — for wearable camera images in NTCIR-12. This represents another realistic problem for the wearable camera use case, where the white balance may be miscalibrated. Color temperature can be bidirectionally biased towards warmer (more orange, lower values) or cooler (more blue, higher values) temperatures from neutral 6600K (details in Method 4). Furthermore, image pixel values deviate asymmetrically with larger deviations for orange than for blue biases (Method 4, Supplementary Fig. 9). Consequently, we found that orange bias led to a larger decrease in Task Performance and CAM Faithfulness than blue bias (Fig. 4d). Conversely, this larger deviation enabled multi-bias DebiasedCNN to predict bias levels more accurately for orange than blue bias (Supplementary Fig. 4d). Notably, CAM deviation was smaller across all color temperature biases than for blur biases, as indicated by the smaller decrease in CAM Faithfulness (compare Fig. 4b and d); hence, Task Performance also did not decrease as much as blur bias. FineTunedCNN had similar Task Performance but lower CAM Faithfulness than RegularCNN; this suggests that color temperature-biased images were too similar to improve

model training with classification fine-tuning, and yet this significantly degraded explanation quality. In contrast, DebiasedCNN improved Task Performance and CAM Faithfulness compared to RegularCNN. Furthermore, due to bidirectional bias, multi-bias training enabled DebiasedCNN (mb, mt) to have significantly higher Task Performance even for unbiased images ($\Delta T = 0$). These results indicate the importance of unbiased attention transfer and multi-bias training.

CAM Truthfulness User Study 1: Debiased-CAM perceived as truthful

Having found that DebiasedCNN improved CAM faithfulness, we next evaluated how well Debiased-CAM improves human interpretability as compared to Biased-CAM. We conducted user studies to evaluate the perceived truthfulness of CAMs (User Study 1) and their helpfulness (User Study 2) in an AI verification task for a hypothetical smart camera with privacy blur filter, activity label prediction and CAM explanations, i.e., the Simulation Study 1 prediction task. The experiment design had two independent variables — Blur Bias level (None $\sigma = 0$, Weak $\sigma = 16$, Strong $\sigma = 32$) and CAM type (Unbiased-CAM, Debiased-CAM, and Biased-CAM). In User Study 1, we recruited 32 participants from Amazon Mechanical Turk (AMT) who followed the experiment procedure described in Fig. 5a,b (Method 8, survey screenshots in Supplementary Figures 11-14). For each trial, the participant viewed an unblurred image, selected the perceived most important image locations regarding the label with a "grid selection" user interface (q1: CAM Truthfulness Selection Similarity, Method 7), viewed the image blurred at a random level and corresponding randomly-ordered CAMs of the three types, rated how representative each CAM type was regarding the image label (q2: CAM Truthfulness Rating), and wrote her rating rationale (q3). For external validity, we selected a variety of one image for each of 10 class labels from ImageNette (Supplementary Fig. 16) with selection criteria as described in Supplementary Method 1. For internal validity, we selected participants based on AMT qualification, a screening quiz and data quality (Supplementary Method 3). To enable participants to compare CAMs more precisely, we showed the CAMs side-by-side for each image, rather than sequentially across pages.

Fig. 5d shows results of the statistical analyses on 320 trials with significant findings at $p<.0001$ (statistical model details in Method 10, Supplementary Table 4a). The distribution of computed CAM Faithfulness (PCC) for different CAM type and Blur Bias levels (Fig. 5c) guided the hypotheses for the results. Unbiased-CAM had the highest CAM Truthfulness Selection Similarity, while Biased-CAM with most deviation had the lowest CAM Truthfulness Selection Similarity that was only 20.3-44.4% of the truthfulness of Unbiased-CAM. Debiased-CAM had significantly higher CAM Truthfulness Selection Similarity than Unbiased-CAM at 69.7-84.2% of the truthfulness of Unbiased-CAM. Similarly, for blurred images, participants rated Unbiased-CAM as the most truthful (M = 8.17 out of 10), followed by Debiased-CAM (M = 6.07 to 7.32), and Biased-CAM as the least truthful (M = 2.80 to 4.89). Our qualitative analysis of participant rating rationales found that Debiased-CAM and Unbiased-CAM were rated as more truthful because they 1) highlighted semantically relevant targets while avoiding irrelevant ones, 2) did not highlight regions that were too narrow or wide for expected objects in the domain, and 3) had

accurate shapes and edge boundaries for salient regions. More detailed quotes and interpretations are reported in Method 11. In summary, Debiased-CAM improved objective and perceived truthfulness, despite stronger blur that reduced CAM truthfulness by highlighting wrong or unexpected regions, sizes, and shapes.

CAM Helpfulness User Study 2: Debiased-CAM perceived as helpful

Having shown that Debiased-CAM was perceived as more truthful than Biased-CAM, we next investigated how helpful Debiased-CAM was to verify predictions of blur biased images. In User Study 2, we used the same image classification task as in User Study 1, and recruited another 171 AMT participants to view one of three CAM types of images blurred at the same three levels in a 3×3 factorial within-subjects experiment. The experiment procedure was similar, but with some rearrangements (Fig. 6a,b, Method 9, survey screenshots in Supplementary Figures 11-13 and 15) for participants to first view the blurred image and predicted label, verify the label with a "balls and bins" question (q1, Method 7), provide preconceived CAM truthfulness and helpfulness ratings (q2 and q3) and ratings rationale (q4), then view the unblurred image and consequently provide the ratings and rationale again (q5-7).

Fig. 6c,d shows results of the statistical analysis on 1,197 trials, and we describe statistically significant findings at $p<.0001$ (statistical model details in Method 10, Supplementary Table 4b). The mean CAM Truthfulness Selection Similarity of each image (measured in User Study 1) was moderately correlated to CAM Truthfulness and Helpfulness ratings ($|\rho| = .419$ to $.486$, all $p<.0001$, Fig. 6d), confirming that objective CAM truthfulness mediated perceived truthfulness and helpfulness. Furthermore, differences in decision quality (Labeling Correctness and Labeling Confidence) across CAM types depended on blur bias level. For None blur, decision quality was high for all CAM types (confidence M = 96.7%, correctness M = 99.8%) due to the ease of the tasks, while for Strong blur, decision quality was low for all CAM types (confidence M = 67.9%, correctness M = 80.0%), suggesting that blurring was too strong even for truthful CAMs to be useful. However, for Weak blur, Debiased-CAM reduced labeling error by 2.44x (1 – Correctness: from 16.8% to 6.8%) and improved confidence from 77.2% to 85.4% compared to Biased-CAM, such that it was not significantly different from Unbiased-CAM. We found stronger differences in preconceived ratings of CAM types. For Weak blur, participants rated Debiased-CAM as more truthful (M = 7.6 vs. 5.6 out of 10) and more helpful (M = 1.5 vs. 0.16, on 7-point Likert scale from –3 to 3) than Biased-CAM. Moreover, for Strong blur, although their decision quality did not improve, participants perceived Debiased-CAM as more truthful (M = 6.3 vs. 4.4) and helpful (M = 0.55 vs. –0.52) than Biased-CAM. These effects were similar and slightly amplified for consequent ratings (Fig. 6d, Supplementary Fig. 10), indicating that users more strongly appreciated Debiased-CAM and disliked Biased-CAM if they had foreknowledge of the unblurred scenes. Our qualitative analysis of participant rating rationales explains that truthful Debiased-CAM and Unbiased-CAM were helpful to verify classifications of unblurred or weakly blurred images, because they: sped up verification by 1) focusing user attention and 2) eliminating

irrelevant targets, 3) matched user expectations[24] of the target object shapes, 4) provided hints on which parts to study in blurred images, and 5) supported hypothesis formation and confirmation[7] of suspected objects. However, all CAMs were unhelpful for strongly blurred images because: 6) verifying the images was too difficult, 7) participants felt forced to blindly trust the CAMs, and 8) they could easily misjudge the CAMs due to confirmation bias[7]. More detailed quotes and interpretations are reported in Method 12. In summary, Debiased-CAM recovered the usefulness of deviated CAMs of moderately blurred images, and participants perceived it as helpful even for strongly blurred images.

DISCUSSION

Our results highlighted issues in explanation faithfulness when CNN models explain their predictions on biased images. To address the challenges, we developed Debiased-CAM to improve the truthfulness and helpfulness of explanations. The approach effectively enhanced CAM Faithfulness (up to 230%) even as the bias level increased in images, and across multiple prediction tasks and bias types. Furthermore, we showed that training multi-bias debiasing did not significantly decrease CAM Faithfulness, and yet supported bias-agnostic explanation, where bias levels were not known ahead of time.

While prior work exploited explanations or attention from human annotation[26], teacher models[27], multiple model layers[25], these approaches only improved predictions on unblurred images; instead, our approach to debias CAMs towards Unbiased-CAMs not only measurably improved the relevance of CAM explanations on biased images, but also enhanced performance across multiple biases (up to 300%). We achieved this by ensuring that model parameters were learned based on more important input pixels as identified by Unbiased-CAMs and on more diverse inputs due to data augmentation across multiple bias levels, and more precise training with multiple prediction tasks. Our results showed that even when images were degraded or distorted due to bias, 1) they retained sufficient useful information that DebiasedCNN could learn to recover salient locations of unbiased CAMs, and 2) these salient locations were highly relevant to the primary task such that prediction performance could be improved. Furthermore, our multi-bias, multi-task debiasing training approach can significantly improve model robustness; for DebiasedCNN trained for wearable camera activity recognition, its task performance and CAM faithfulness remained high even for images under strong blur, and its task performance improved when trained on multiple color temperature biases. Next, we discuss generalizations for explanation debiasing.

Our self-supervised debiasing approach can be applied to other Grad-CAM extensions[15] and other gradient-based attribution explanations[13,31,33] by formulating the activation, gradient or propagated terms as network layers to model a secondary prediction task. However, some saliency explanations, such as Layer-wise Relevance Propagation (LRP)[31] and Integrated Gradients[33], that produce fine-grained “edge detector” heatmaps[44] are likely to be more severely degraded with biasing, such as strong blurring, so it may be more challenging to reconstruct the original, unbiased

saliency maps. To handle this, unfreezing and retraining deeper convolutional layers can allow for more fine-grained debiasing. Also, debiasing the gradients of intermediate convolution blocks instead of input pixels could provide coarser-grained "edge detector" heatmaps that may be more amendable to debiasing. Apart from that, model-agnostic explanations, like LIME[45] and Kernel SHAP[46], can be debiased with self-supervised training by regularizing on a saliency loss metric, but this loss term will not be differentiable with respect to model weights and lead to weaker debiasing. Including some knowledge of the underlying model can improve explanation debiasing and faithfulness. Additionally, CNN explanation techniques not based on saliency maps, such as neuron and feature visualizations[34,47], attention[26] and counterfactuals[48], have higher dimensionality than saliency maps, and may have spurious explanation deviations that are harder to detect; these may require more sensitivity to debias. Performing dimensionality reduction with autoencoders or generative adversarial networks (GANs) can provide latent features and concepts that can be subsequently debiased to feasibly debias the high-dimensional explanations. Finally, concept-based explanations, like TCAV[49], are already low-dimensional and can be debiased with human annotation[24] from unbiased training datasets.

Debiased-CAM can be generalized to other bias sources and prediction tasks. We have investigated deviations in CAM explanations due to two common image biases, Gaussian blurring and color shifting. Other cases of biasing include images captured under low light[50], noisy ultrasound images[51], and with motion blur[17]. Training to debias against these can help to generate explanations which are more robust and interpretable for more contexts of use. Other than biases in images, debiasing is also necessary for explaining model predictions of other data types and behaviors, such as audio signals with noise or obfuscation[52], and human activity recognition with inertial measurement units (IMU) or other wearable sensors[53]. With the prevalence of noise and bias in real data, Debiased-CAM provides a generalizable framework to train more robust and faithful explanations that are human interpretable.

FIGURES

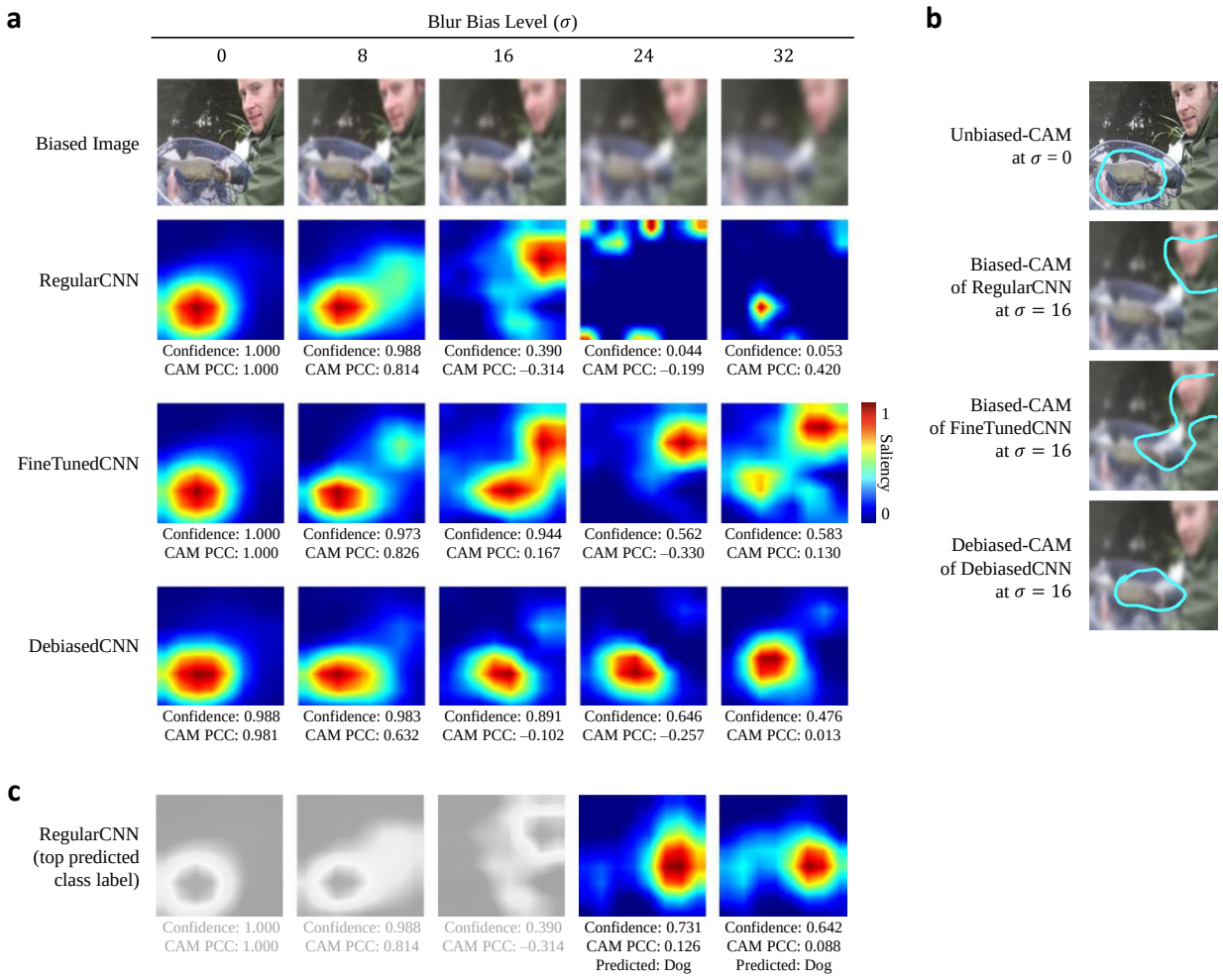


Fig. 1 | Deviated and debiased CAM explanations from different CNN models of an example blur-biased "Fish" image. a, Debiased-CAMs (from DebiasedCNN) were the most faithful to the Unbiased-CAM (from RegularCNN at $\sigma = 0$) as blur bias increased. In contrast, Biased-CAMs from RegularCNN and FineTunedCNN became significantly deviated with a much lower CAM Pearson Correlation Coefficient (PCC). **b,** Debiased-CAM selected similar important pixels of the Fish as Unbiased-CAM, while Biased-CAMs wrongly selected irrelevant pixels of the person or background instead. Important pixels shown within contour lines (cyan) were overlaid on the actual unblurred scene ($\sigma = 0$) for reference and the blurred input image at $\sigma = 16$. **c,** CAM of the top predicted class label with only RegularCNN at $\sigma = 24$ or 32 predicting the wrong prediction label "Dog". In these cases, the CAMs also do not highlight the fish and are similar to the Biased-CAM of FineTunedCNN (**a**). Furthermore, this demonstrates that even though FineTunedCNN predicted correctly, its wrong CAMs suggests that it was predicting wrongly.

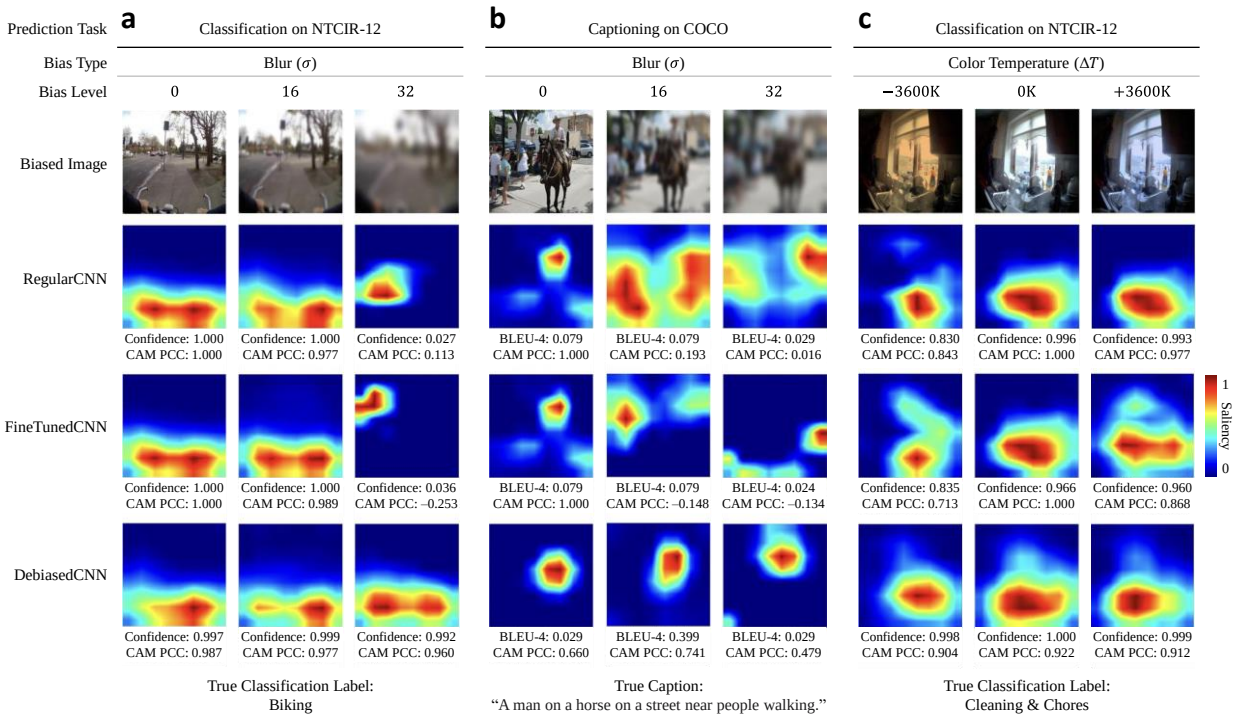


Fig. 2 | Deviated and debiased CAM explanations from CNN models trained for different image prediction tasks and types of biases at varying bias levels. a, For wearable camera activity recognition, RegularCNN and FineTunedCNN generated significantly deviated CAMs as blur bias increased, where the bicycle handlebars became less salient; whereas DebiasedCNN maintained high CAM faithfulness and prediction confidence. **b,** For image captioning, RegularCNN and FineTunedCNN generated wildly deviated CAMs as blur bias increased, where background people and other areas were highlighted; whereas DebiasedCNN produced less deviated CAMs that still selected the horse and rider. **c,** For wearable camera activity recognition, RegularCNN and FineTunedCNN generated mildly deviated CAMs which were more deviated for orange-bias (lower color temperature) than for blue-bias (higher color temperature); whereas DebiasedCNN produced CAMs that were the least deviated, highlighting the kitchen sink. **a-c,** At no bias, all CAMs from RegularCNN and FineTunedCNN are unbiased with no deviation. See Supplementary Figures 17-32 for more CAM examples for different images.

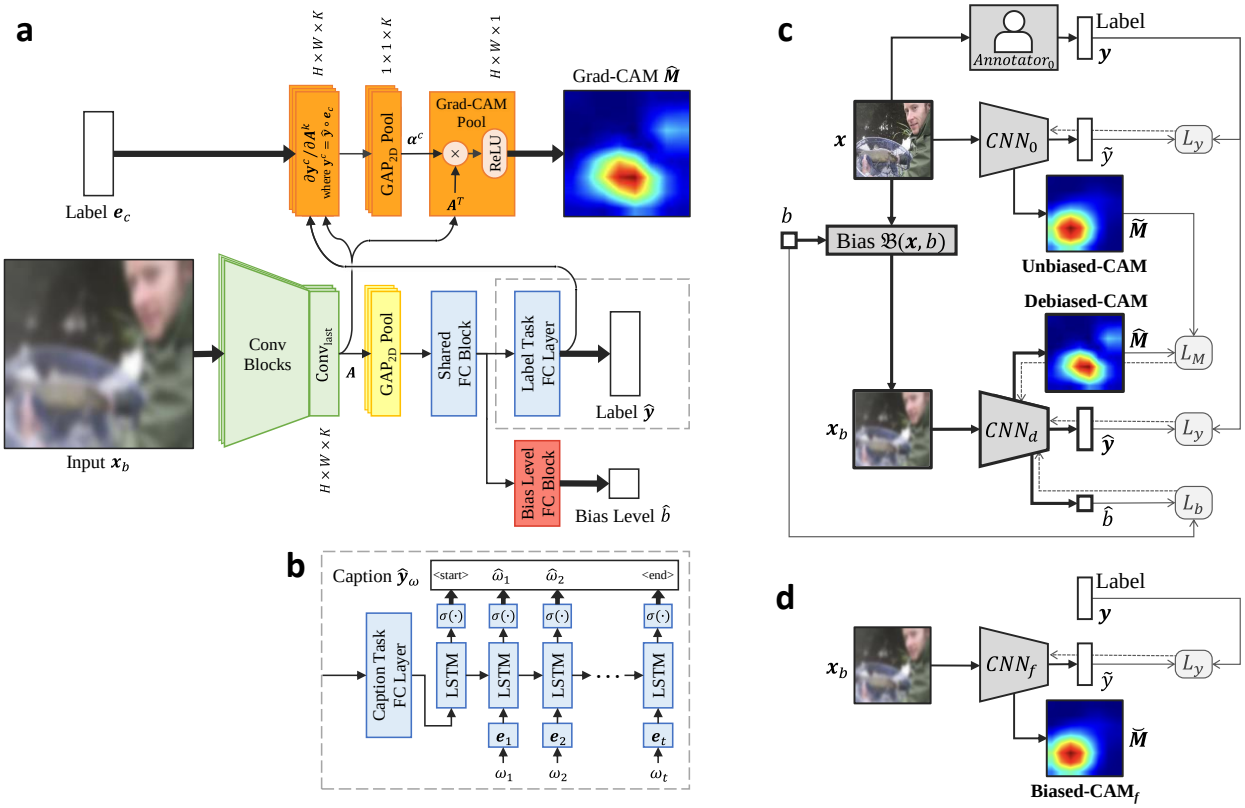


Fig. 3 | Architecture of multi-bias, multi-input, multi-task debiased CNN model and self-supervised learning to debias CAM. a, DebiasedCNN is a multi-input, multi-task deep convolutional neural network with two inputs image $\boldsymbol{x}_b$ and label $\boldsymbol{e}_c$ for CAM class c, and three tasks for primary prediction task $\hat{\boldsymbol{y}}$, CAM explanation task $\widehat{\boldsymbol{M}}$, and bias level prediction task $\hat{b}$ (Method 1). The Grad-CAM explanation is predicted from a series of layers that modeled the $H \times W \times K$ tensor for activation maps $\boldsymbol{A}$, $H \times W \times K$ tensor for activation gradients $\partial \boldsymbol{y}^c / \partial \boldsymbol{A}^k$ calculated from the label prediction probability $\boldsymbol{y}^c$, $1 \times 1 \times K$ tensor for K importance weights of activation maps $\boldsymbol{\alpha}^c$, and $H \times W \times 1$ tensor as CAM $\widehat{\boldsymbol{M}}$. Various base CNN models can be used (comparisons reported in Supplementary Fig. 8). **b,** DebiasedCNN can be trained for different primary tasks, such as image captioning with an LSTM network after the CNN encoder. **c,** Meta-architecture with self-supervised learning to minimize the CAM loss L_M between Unbiased-CAM $\widetilde{\boldsymbol{M}}$ generated from RegularCNN (CNN_0) predicting on an unbiased image $\boldsymbol{x}$ and Debiased-CAM $\widehat{\boldsymbol{M}}$ generated from DebiasedCNN (CNN_d) predicting on the biased version of the image $\boldsymbol{x}_b$ at bias level b. DebiasedCNN also learns the bias level b of the image by minimizing bias level loss L_b. The multi-bias (mb) DebiasedCNN variant is shown, but other architectures can be trained (Supplementary Fig. 2). DebiasedCNN can be trained to infer on privacy-preserving image data (Supplementary Fig. 1). **d,** A baseline FineTuneCNN (CNN_f) that is only trained to minimize the primary task loss L_y, but not CAM loss, will still a generate deviated Biased-CAM$_f$ $\breve{\boldsymbol{M}}$ on biased image $\boldsymbol{x}_b$. Note that CNN_0 will also produce a Biased-CAM on $\boldsymbol{x}_b$. **a,c,d,** CAMs for all models were generated from a classification example of a blur biased image labeled as "Fish".

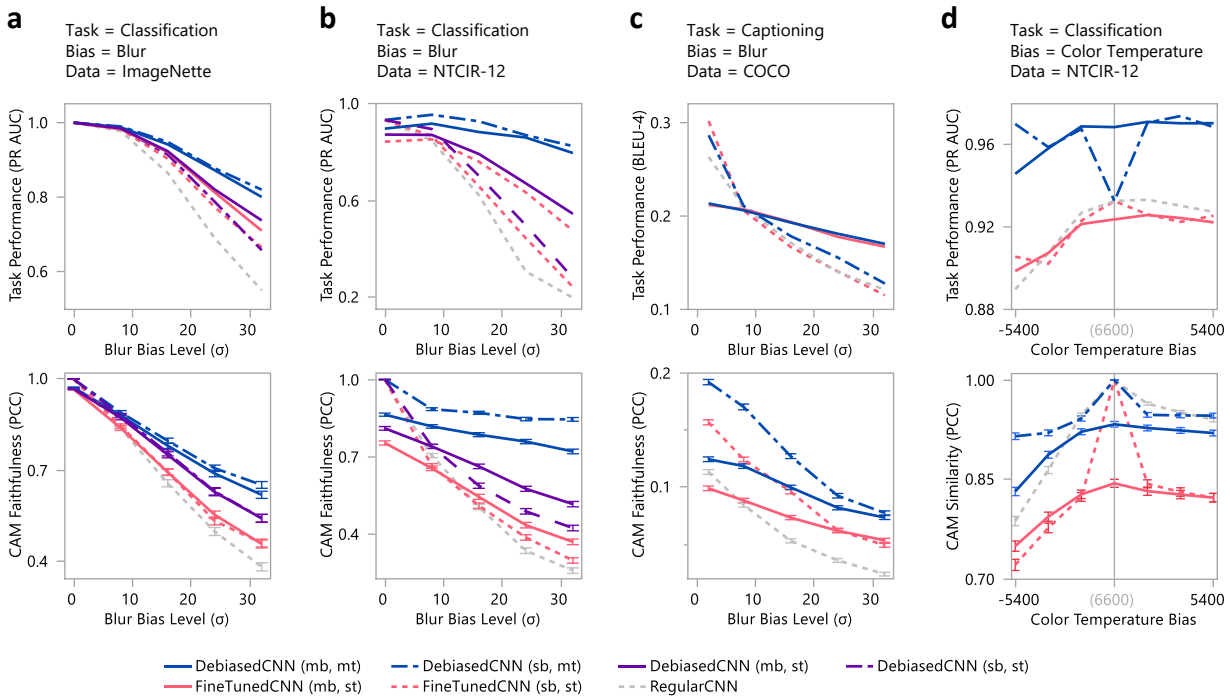


Fig. 4 | Comparisons of Task Performance and CAM Faithfulness of CNN models for different prediction tasks with increasing bias. a,b, All model Task Performance and CAM Faithfulness decreased with increasing blur, while DebiasedCNN decreased the least. DebiasedCNN (blue and violet) improved Task Performance over RegularCNN (grey) and FineTunedCNN variants (red) due to attention transfer with higher CAM Faithfulness. For DebiasedCNN variants, multi-task (blue) had the highest CAM Faithfulness and Task Performance that is higher than single-task (violet), due to differentiable CAM loss. **a,** Task Performance for no blur (σ = 0) was perfect (PR AUC = 1) because models were pre-trained on the superset ImageNet dataset. **b,** Task Performance and CAM Faithfulness decreased more sharply than **a**, since ImageNette is a cleaner dataset than NTCIR-12 with fewer label classes. In contrast, multi-bias DebiasedCNN (blue) maintained high Task Performance and CAM Faithfulness across all blur levels, indicating very effective debiasing. **c,** Task Performance and CAM Faithfulness were lower for image captioning than for classification since captioning is more complex. Despite this, DebiasedCNN improved both metrics across all bias levels. **d,** DebiasedCNN generally had the best Task Performance and CAM Faithfulness even as they decreased asymmetrically with orange or blue bias about the neutral color (6600K). **a-d,** Model variants annotated as st = single-task, mt = multi-task, sb = single-bias, mb = multi-bias (Method 3, Supplementary Table 1). Primary Task Performance (first row) was calculated as area under precision-recall curve (PR AUC) for classification and BLEU-4 score for captioning (Method 5). CAM Faithfulness (second row) was calculated as the Pearson's Correlation Coefficient (PCC) between CAM and Unbiased-CAM (Method 6; see Supplementary Fig. 3 for CAM Faithfulness calculated with Jensen-Shannon Distance). Error bars indicate 90% confidence interval. **a-c,** Image blur bias was varied with the standard deviation σ of the Gaussian blur for the normalized image (Method 4). **d,** Color temperature bias (in Kelvin) was varied with respect to neutral cloudy daylight at 6600K. **a-d,** Supplementary Table 3 describes percent improvements in Task Performance and CAM Faithfulness.

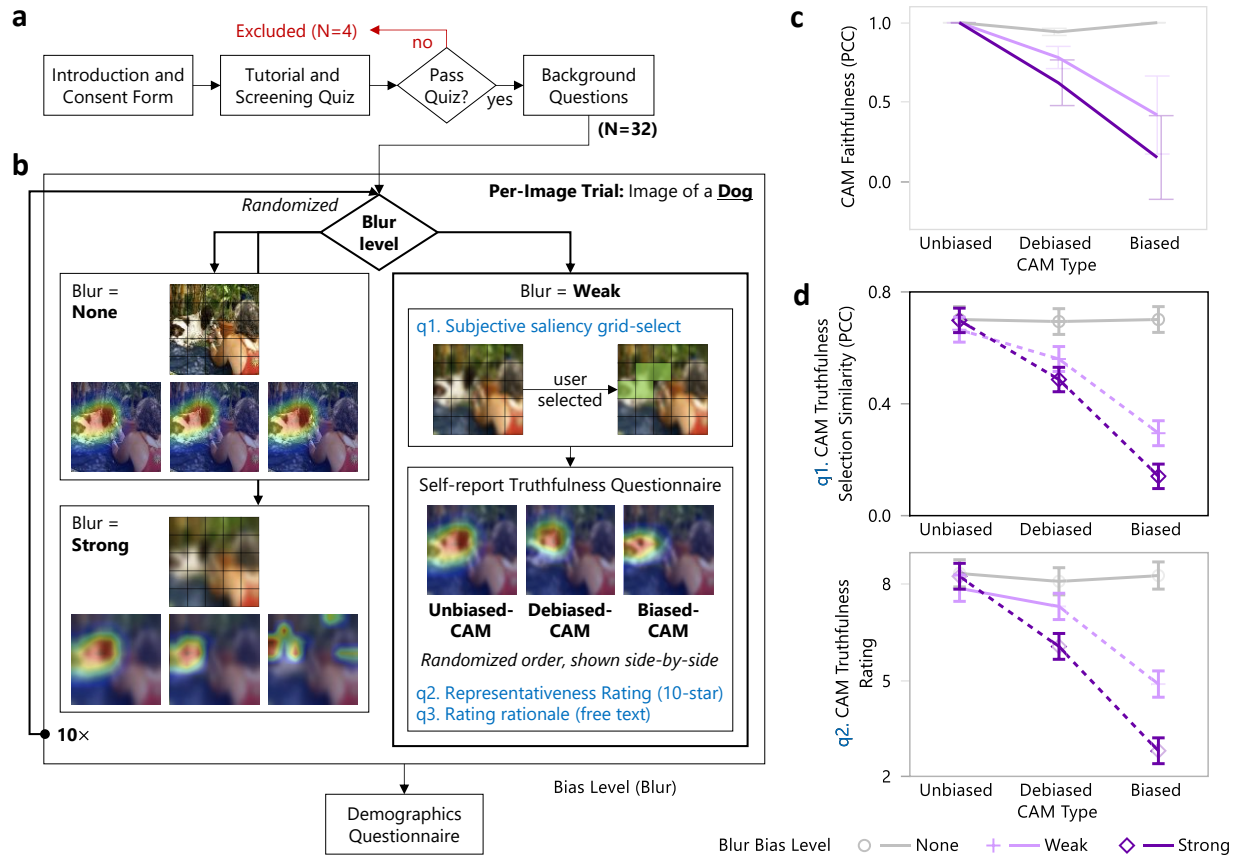


Fig. 5 | CAM Truthfulness User Study 1 with 32 participants (320 trials) to compare perceived CAM truthfulness for different CAM types (Unbiased-CAM, Debiased-CAM, Biased-CAM) under varying Blur Bias levels (None, Weak, Strong). **a,b,** Experiment procedure including a brief tutorial, screening quiz, main study, and background questions (details in Method 8). **b,** Experiment design of the main study of 10 image trials with Blur Bias level within-subjects, randomly assigned per trial, and all CAM types shown side-by-side (Method 8). Text in blue represent measures from survey questions (q1-3). **c,** Comparison of CAM Faithfulness of the selected 10 image instances. CAM Faithfulness decreased as Blur Bias increased, was the highest for Unbiased-CAM, the lowest for Biased-CAM Biased-CAM, and improved by Debiased-CAM. Error bars indicate 90% confidence interval. **d,** Least squares means estimates from linear mixed effects models (Method 10) of CAM Truthfulness Selection Similarity (q1) and CAM Truthfulness Rating (q2). Results agree with theoretical hypotheses in **c** and showed that CAM Truthfulness decreased with stronger blur and more deviated CAM biasing, but Debiased-CAM improved both Selection Similarity (PCC) and Rating. Dotted lines indicate extremely significant p<.0001 comparisons; solid lines indicate no significance at p>.01. Error bars indicate 90% confidence interval. **b-d,** Unbiased-CAM refers to the CAM from RegularCNN predicting on the unbiased image regardless of blur bias level; Debiased-CAM refers to the CAM from DebiasedCNN (mb, mt) and Biased-CAM refers to the CAM from RegularCNN at the corresponding Blur Bias levels. At the None blur level, Biased-CAM is identical to Unbiased-CAM since there is no image bias.

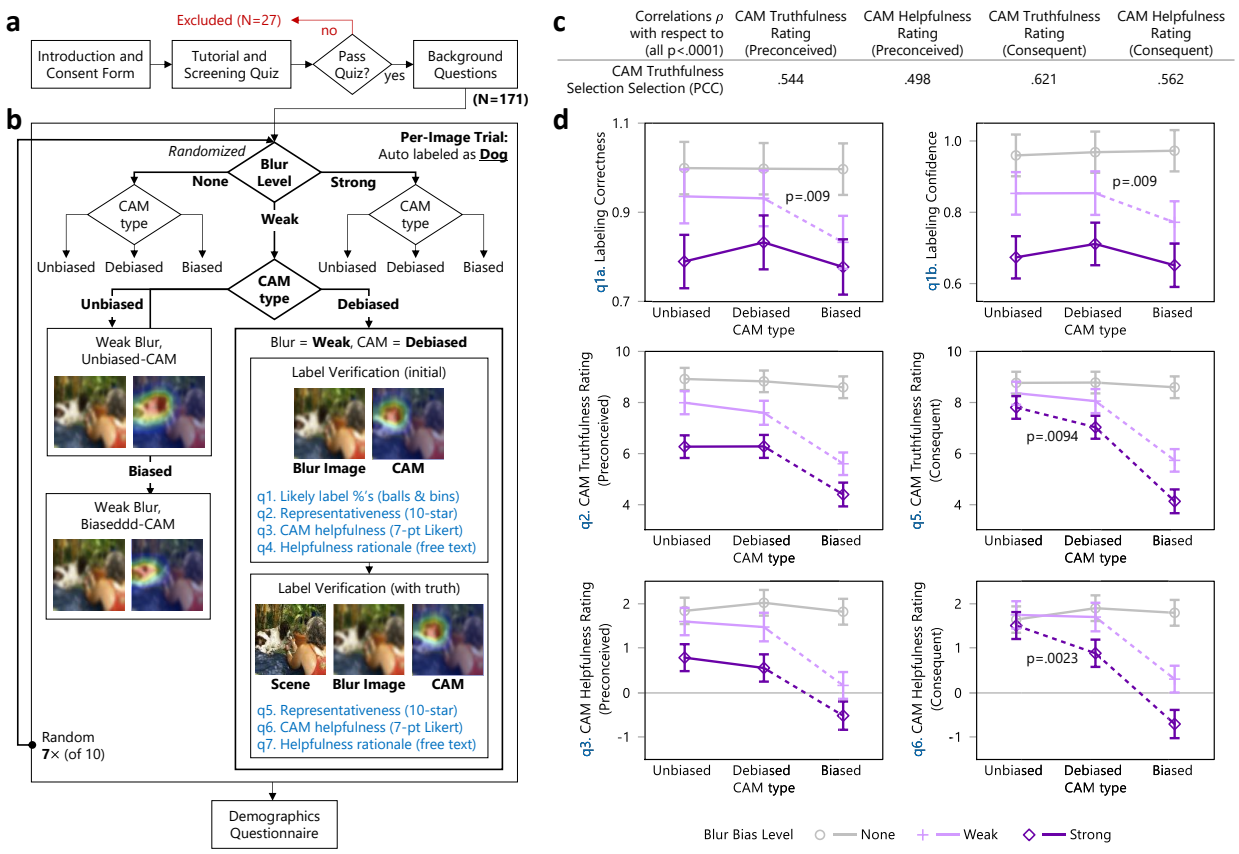


Fig. 6 | CAM Helpfulness User Study 2 with 171 participants (1,197 trials) to compare perceived truthfulness and helpfulness of CAM types under varying Blur Bias levels. a, Experiment procedure is similar to User Study 1 with the same brief tutorial, screening quiz and background questions, but with a different main study experiment design and number of trials (Method 9). **b,** Experiment design of the main study of 7 trials with both independent variables Blur Bias level and CAM type within-subjects with random assignment per trial; only one CAM is shown at a time. **c,** Results of the manipulation check showing that rating responses (q2, q3, q5, q6) were correlated with the objective CAM Truthfulness Selection Similarity metric of User Study 1 for each image instance. **d,** Results of labeling and rating responses show that labeling performance, and CAM truthfulness and helpfulness ratings decreased with stronger blur and more deviated CAM biasing; but Debiased-CAM improved all user ratings and improved labeling correctness and confidence for Weak blur. Labeling Correctness (q1a) and Labeling Confidence (q1b) were calculated from the "balls and bins" question (q1, Method 7). CAM Truthfulness Ratings were measured along a 1-10 scale, and CAM Helpfulness Ratings along a 7-point Likert scale (–3 = Strongly Disagree, 0 = Neither, +3 = Strongly Agree). Dotted lines indicate extremely significant p<.0001 comparisons, otherwise very significant as stated; solid lines indicate no significance at p>.01. Error bars indicate 90% confidence interval.

METHODS

Method 1 Debiasing CAM Prediction Task with Differentiable CAM Loss

We describe the background approach of generating a class activation map (CAM) with Grad-CAM[37], and detail our redefined approach to enable strong debiasing. Grad-CAM generates a saliency map explanation of an image prediction with regards to class c as the weighted sum of activation maps in the final convolutional layer of a CNN. Each activation map $\boldsymbol{A}^k$ indicates the activation A_{ij}^k for each grid cell (i,j) of the kth convolution filter ($k \epsilon \mathcal{K}$). The importance weight $\boldsymbol{\alpha}_k^c$ for the kth activation map is calculated by backpropagating gradients from the output y to the convolution filter and global average pooling across all grid cells in the activation map, i.e.,

$$\boldsymbol{\alpha}_k^c = \overbrace{\frac{1}{HW}\sum_{i=1}^{H}\sum_{j=1}^{W}}^{\text{Global Avg Pooling}} \underbrace{\frac{\partial \boldsymbol{y}^c}{\partial A_{ij}^k}}_{\substack{\text{gradients} \\ \text{via backprop}}} = GAP_{ij}\left(\frac{\partial \boldsymbol{y}^c}{\partial \boldsymbol{A}^k}\right) \tag{1}$$

where H and W are the height and width of activation maps, respectively; $\boldsymbol{y}^c$ is a one-hot vector indicating only the probability of class c with only the cth element non-zero, i.e., $\boldsymbol{y}^c = \boldsymbol{y} \circ \boldsymbol{e}_c$, $\boldsymbol{y}$ is the prediction probability distribution across classes, $\circ$ is the Hadamard operator and $\boldsymbol{e}_c$ is the standard basis vector (e.g., $\boldsymbol{e}_2 = (0,1,0,\ldots,0)^T$). For image captioning tasks, $\boldsymbol{y}^c$ is the average word embedding vector of all words in the predicted caption. With these weights, Grad-CAM obtains the class activation map by computing a weighted combination, followed by a ReLU transform to only show activations with positive attribution towards class c, i.e,

$$M_{GradCAM}^c(i,j) = ReLU\ \underbrace{\left(\sum_k \boldsymbol{\alpha}_k^c \boldsymbol{A}^k\right)}_{\text{linear combination}} \quad \equiv \quad \widehat{\boldsymbol{M}} = ReLU(\boldsymbol{\alpha}^c \boldsymbol{A}^{\mathrm{T}}) \tag{2}$$

which we rewrite with a matrix multiplication of all $K = |\mathcal{K}|$ importance weights $\boldsymbol{\alpha}^c = \{\alpha_k^c\}^K$ and the transpose of activation maps $\boldsymbol{A}$ along the kth axis, i.e., $\boldsymbol{A}^{\mathrm{T}} = \left\{A_{ij}^k\right\}^{K \times H \times W}$. Therefore, the CAM prediction task can be redefined as three layers (orange) in the neural network (Fig. 3a) to compute $\partial \boldsymbol{y}^c / \partial \boldsymbol{A}^k$, $\boldsymbol{\alpha}_k^c$, and $\widehat{\boldsymbol{M}}$, respectively. By reformulating Grad-CAM as a prediction task, we can train the model based on differentiable CAM loss by backpropagating through this task. This task takes $\boldsymbol{e}_c$ as the second input to the CNN architecture to specify the explanation target label. $\boldsymbol{e}_c$ is determined by the ground truth class label c at training time, and by the class c chosen by the user at test or run time. Finally, we up-scale the CAM to the input image size with bicubic interpolation.

Method 2 Multi-Bias Level Prediction Task

Since the bias level may be unknown at prediction time, we further propose a tertiary bias level prediction task to make DebiasedCNN bias level agnostic. This involves two steps:

1) Data augmentation to bias the training dataset to multiple levels. Each image is randomly biased by any continuous number within $0 \leq \sigma \leq 33$ for blur biasing, and within $-5{,}400 \leq$

$\Delta T \leq +5{,}400$ for color temperature biasing. To control the dataset size in experiments, each image in the dataset was only biased once. To train more accurate models, each original, unbiased image can be biased to multiple levels to further augment the dataset.

2) Training a tertiary task to predict the image bias level (Fig. 3a: pink block). This co-training improved the primary task and CAM task predictions due to the correlations with those tasks.

Method 3 Model Variants and Training Loss Functions

Training to improve CAM Faithfulness can be done as a single task or multi-task. For a single-task DebiasedCNN, the loss is backpropagated through the classification task as the function:

$$L(\boldsymbol{w}) = \underbrace{L_y(y, \hat{y}(\boldsymbol{w}))}_{\text{prediction loss}} + \omega_M \underbrace{L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}})}_{\text{CAM loss}} \tag{3}$$

where $L_y(\boldsymbol{w})$ is the differentiable classification loss between ground truth y and prediction $\hat{y}$, L_M is the non-differentiable CAM loss between Unbiased-CAM $\widetilde{\boldsymbol{M}}$ and the model's CAM $\widehat{\boldsymbol{M}}$, and ω_M is the CAM training hyperparameter. We found that multi-task training most improved Task Performance and CAM Faithfulness (Fig. 4a,b). For single-bias, multi-task DebiasedCNN (sb, mt), the loss function separates each loss term across different tasks, i.e.,

$$\boldsymbol{L}(\boldsymbol{w}) = \begin{pmatrix} L_y(y, \hat{y}(\boldsymbol{w})) \\ \omega_M \, L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}}(\boldsymbol{w})) \end{pmatrix} \tag{4}$$

where $\boldsymbol{L}(\boldsymbol{w})$ is a vector representing the classification loss L_y and CAM loss L_M, which are both differentiable with respect to each of their unshared tasks. With the bias level prediction task, the loss function for DebiasedCNN (mb, mt) becomes:

$$\boldsymbol{L}(\boldsymbol{w}) = \begin{pmatrix} L_y(y, \hat{y}(\boldsymbol{w})) \\ \omega_M \, L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}}(\boldsymbol{w})) \\ \omega_b \, L_b(b, \hat{b}(\boldsymbol{w})) \end{pmatrix} \tag{5}$$

where L_b is the differentiable loss between actual and predicted bias level of the input image, and ω_b is the bias level training hyperparameter. We trained regular and fine-tuned CNN models and ablated variants of DebiasedCNN to compare Task Performance and CAM Faithfulness. See Supplementary Table 1 for their training loss functions and data augmentation.

The specific loss terms are defined as follows: primary task loss L_y as cross-entropy loss for multiclass classification tasks, and as the sum of negative log likelihood[3] for each caption word for image captioning tasks; bias level loss L_b as the mean squared error (MSE), common for regression tasks; CAM loss L_M as the mean squared error (MSE), since CAM prediction can be considered a 2D regression task, and this is common for visual attention tasks[27]. Other suitable metrics for the CAM loss include: mean absolute error (MAE) which penalizes large differences less than MSE; Kullback-Leibler Divergence (KLD) or Jensen-Shannon Distance (JSD) which compare the distribution of pixel saliency between CAMs, but are more expensive to calculate; and Pearson's Correlation Coefficient (PCC) which compares the pixel-wise correlation between CAMs, but is also computationally expensive for training.

Method 4 Image Bias Types and Bias Levels

For the simulation and user studies, we varied images to bias them with blur and color temperature at different bias levels. We blurred images by applying a uniform Gaussian blur filter[19] using opencv-python v4.2.0. We scaled all images to a standardized maximum size of 1000×1000 pixels and applied Gaussian blur at various standard deviations σ. For simulation studies, we varied σ to 5 levels 0, 8, 16, 24, 32 to evaluate Task Performance and CAM Faithfulness and to random values from 0 to 32 to evaluate regression performance. For user studies, we chose three Blur Bias levels as None ($\sigma = 0$), Weak ($\sigma = 16$), and Strong ($\sigma = 32$) determined from pilot studies, such that Strong blur is generally too challenging to recognize and Weak blur is half as blurred.

Color temperature refers to the temperature of an ideal blackbody radiator as if illuminating the scene. We biased color temperature as follows. Each pixel in an unbiased image has color $(r, g, b)^T$, where R, G, B represent the red, green, and blue color values within range 0-255, respectively. Each pixel is biased from neutral temperature t by Δt_b at bias level b by multiplying a diagonal correction matrix with its color, i.e., $(r_b, g_b, b_b)^T = \text{diag}(255/R_b\ ,255/G_b, 255/B_b)(r, g, b)^T$, where $(R_b, G_b, B_b)^T = f_{CT}(t + \Delta t_b)$ are scaling factors obtained from Charity's color mapping function f_{CT} to map a blackbody temperature to RGB values[54] (Supplementary Fig. 9). We set the neutral color temperature t to 6600K, which represents cloudy/overcast daylight. Color temperature biasing is asymmetric about zero bias, because people are more sensitive to perceiving changes in orange than blue colors (Kruithof Curve[55]); and due to the non-linear monotonic relationship between blackbody temperature and modal color frequency (Wien's Displacement Law). This asymmetry explains why orange biasing led to stronger CAM deviation than blue biasing. For the simulation studies, we varied color temperature bias Δt to 7 levels –5400, –3600, –1800, 0, 1800, 3600, 5400 to evaluate Task Performance and CAM Faithfulness (Fig. 4d), and to random values between –5400 and 5400 to evaluation regression performance.

Method 5 Simulation Studies Model Task Performance Metrics

For classification tasks, we calculated model Task Performance as the area under the precision-recall curve (PR AUC) as it is robust against imbalanced data[56], and calculated the class-weighted macro average to aggregate across multiple classes. For image captioning tasks, we calculated the BLEU-4[57] score that measures how closely 4-grams in the predicted caption and actual captions matched for image captioning tasks. For bias level regression, we calculated accuracy with R^2.

Method 6 Simulation Studies CAM Faithfulness Metrics

To better compare CAMs beyond simple residual differences (e.g., MAE, MSE), we adopted other metrics from saliency map evaluations[58,59]. We calculated CAM Faithfulness as the Pearson's Correlation Coefficient (PCC) of pixel-wise saliency. We selected PCC as it closely matches human perception to favor compact locations and match the number of salient locations[58], and it fairly weights between false positive and false negatives[59]. Other saliency comparison metrics include: Area under ROC Curve (AUC) of pixel fixations; and Kullback-Leibler Divergence

(KLD) between saliency maps, which are appropriate for localization applications[59]. Since CAMs localize pixels that are important for image prediction[13], we also computed CAM Faithfulness with the Jensen-Shannon Divergence (JSD) that extends KLD to be symmetric and bounded, i.e.,

$$D_{JS}(\widetilde{\boldsymbol{P}}, \boldsymbol{P}) = \frac{1}{2} D_{KL}(\widetilde{\boldsymbol{P}}, \overline{\boldsymbol{P}}) + \frac{1}{2} D_{KL}(\boldsymbol{P}, \overline{\boldsymbol{P}}) \quad (6)$$

where $\overline{\boldsymbol{P}} = (\widetilde{\boldsymbol{P}} + \boldsymbol{P})/2$ is the average of the compared normalized CAM probabilities (e.g., $\boldsymbol{P} = \boldsymbol{M}/\sum_{ij} \boldsymbol{M}_{ij}$), and $D_{KL}(\widetilde{\boldsymbol{P}}, \overline{\boldsymbol{P}}) = \sum_{ij} \widetilde{\boldsymbol{P}}_{ij} \log(\widetilde{\boldsymbol{P}}_{ij}/\overline{\boldsymbol{P}}_{ij})$ is the KLD between normalized CAMs. Hence, we can calculate CAM Faithfulness as $1 - D_{JS}$. Results of the JSD-based metric agreed strongly with the PCC metric (compare Fig. 4 with Supplementary Fig. 3).

Method 7 User Studies Measures and Instruments

Although self-reported ratings are common in human-subjects studies of system and explanation usage[12,24], participants may poorly estimate their perceptions[60–62]. Thus, we employed objective measures of human perception and opinion, where appropriate. Specifically, we used a "grid selection" user interface to measure objective truthfulness (Supplementary Fig. 14a), and the "balls and bins" question[62] to elicit user labeling (Supplementary Fig. 15a). Employed in User Study 1 for q1, the grid selection UI overlays a clickable grid over the image for participants to select which grid cells are most important regarding the label. For usability, we limited the grid to 5×5 cells that can be selected or unselected (binary values). In the surveys, we referred to CAMs as "heatmaps", which is a more familiar term. We define User-CAM as the participant's grid selection response, and CAM as the heatmap shown. To compare User-CAM with CAM, we aggregated CAM by averaging the pixel saliency in each cell and calculated CAM Truthfulness Selection Similarity as the Pearson's Correlation Coefficient (PCC) between User-CAM and CAM.

In User Study 2 for q1, we asked the participant to indicate the likelihoods of 10 labels to be the actual image label with the "balls and bins" graphical distribution building question[62–65] to elicit her probability distribution $\boldsymbol{p} = \{p_c\}^T$ over label classes $c \in C$. This question is reliable in eliciting probabilities from lay users[62,63] and avoids priming participants with the actual label c_0, since it asks about of all labels. We calculated the participant's selected label $\acute{c}$ as the class with the highest probability, i.e., $\acute{c} = \text{argmax}_c(p_c)$, Labeling Confidence as the indicated likelihood for the actual label p_{c_0}, and Label Correctness as $[\acute{c} = c_0]$, where $[\cdot]$ is the Kronecker delta.

While CAM Truthfulness Selection Similarity was objective, we measured the CAM Truthfulness Rating as a subjective, self-reported opinion as a survey question on a unipolar 10-point star rating scale (1 to 10). We measured perceived the CAM Helpfulness Rating on a bipolar 7-point Likert scale (–3 = Strongly Disagree, –2 = Disagree, –1 = Somewhat Disagree, 0 = Neither Agree nor Disagree, +1 = Somewhat Agree, +2 = Agree, +3 = Strongly Agree). We collected the rationale of ratings as open-ended text. We used different formats for CAM Truthfulness and CAM Helpfulness to mitigate repetitive or copied responses and to allow for more precise measurement of CAM Truthfulness.

Method 8 User Study 1 Procedure and Experiment Design

For User Study 1, participants followed the experiment procedure (illustrated in Fig. 5a,b): read an introduction of the study and task, and gave consent; studied a one-page tutorial about automatic image labeling, privacy blurring, heatmap explanations, and how to interpret and answer the "balls and bins" question (Supplementary Fig. 11); answered four questions in a screening quiz to test their labeling of an unblurred and a weakly blurred image and their selection of important locations in an image and a CAM (Supplementary Fig. 12); if screening was passed, answered background questions on technology and image comprehension savviness (Supplementary Fig. 13), otherwise was directed to the survey end; performed the main study with 10 trials (Supplementary Fig. 14); and ended with demographic questions on gender, age, educational background, and occupation.

In the main study (Fig. 5b), each participant viewed 10 repeated image trials, where each trial was randomly assigned to one Blur Bias level in a within-subjects experiment design. All participants viewed the same 10 images (selection described in Supplementary Method 1), which were randomly ordered. For each trial, the participant: viewed a labeled unblurred image, indicated the most important locations on the image regarding the label with the "grid selection" UI (q1); and in the next page, viewed the image blurred by the smart camera, viewed all three CAM types generated from the blurred image and randomly-arranged side-by-side, rated on a 10-star scale how well each CAM represented the image label (q2), and wrote the rationale for her rating (q3).

Method 9 User Study 2 Procedure and Experiment Design

To carefully investigate the helpfulness of CAMs on blurred images, we modified the survey of User Study 1 to change the sequence of information shown to participants. User Study 1 focused on CAM Truthfulness to obtain the participant's saliency annotation of the unblurred image before priming the participant by showing CAMs. For User Study 2, showing the unblurred image first will invalidate the use case of verifying predictions on blurred images, since the participant would have foreknowledge of the image. Hence, participants needed to see the blurred image and model prediction first, answer questions about their perceptions, then see the actual, unblurred image. Fig. 6a,b illustrates the experiment procedure of User Study 2. Participants began with the same procedure as User Study 1 (Fig. 5a, Method 8), including the same introduction, tutorial, screening quiz, and background and demographics questions, but experienced a different main study section.

In the main study (Fig. 6b), each participant viewed 7 repeated image trials, each randomly assigned to one of 9 conditions (3 Blur Bias levels × 3 CAM types) in a within-subjects experiment design. Participants viewed 7 randomly chosen images from the same 10 images of User Study 1, instead of all 10, so that they could not easily conclude which class was the likely label for the remaining images by eliminating previous classes. For each trial, the participant performed the common explainable AI task to verify the label prediction of the model. The participant viewed a labeled image at the assigned Blur Bias level with corresponding CAM for the assigned CAM type, indicated her likelihood choice(s) for the image label with the "balls and bins" question[62] (q1, Method 7); rated how well each CAM represented the image label (q2); rated how helpful the

CAM was for verifying the label (q3), and wrote the rationale for her rating (q4). Supplementary Fig. 15a shows the first questionnaire page. In the next page, participants saw the image unblurred and answered questions q2-4 again as questions q5-7 (Supplementary Fig. 15b). This allowed us to compare between preconceived and consequent ratings and rationale (Supplementary Fig. 10).

Method 10 User Studies Statistical Analyses

For each user study, for all response variables, we fit a multivariate linear mixed effects model with Blur Bias Level, CAM Type, and Trial Number sequence as fixed effects, Blur Bias Level × CAM Type as fixed interaction effect, and Participant as random effect. For User Study 2, further we analyzed CAM Truthfulness and Helpfulness ratings with fixed main and interaction effects regarding whether users rated before or after seeing the unblurred version of the image, i.e., Unblurred Disclosure: preconceived or consequent. Supplementary Table 4a and b report the model fit (R^2) and significance of ANOVA tests for each fixed effect for User Study 1 and 2, respectively. Due to the large number of comparisons in our analysis, we consider differences with $p<.001$ as significant. This is sufficiently strict for a Bonferroni correction for 50 comparisons (significance level = .05/50). Furthermore, all results reported were significant at $p<.0001$, unless otherwise stated. We performed post-hoc contrast tests for specific differences described (dotted lines in Figures 5d and 6d). All statistical analyses were performed using JMP (v14.1.0).

Method 11 CAM Truthfulness User Study 1 Qualitative Analysis and Results

We analyzed the written rationale of participant ratings to better understand how participants interpreted different CAMs as truthful or untruthful, and what visual features they perceived in images and CAMs. We performed a thematic analysis with open coding[66] to identify several themes in what was written. Two authors independently coded the rationales and discussed the coding until themes converged. Next, we first describe rationales mentioned for different blur levels, then describe the themes that spanned across all blur levels. Note that all CAM types were shown randomly ordered and given anonymous labels A, B, and C; we quote them specifically by type for clarity. Supplementary Fig. 16 shows the images and CAMs that participants viewed.

For None blur without image bias, as expected, most participants perceived CAMs as identical, e.g., *"all 3 images are the same and mostly representative"* (Participant P23, "Fish" image); though some participants could perceive the slight decrease in the CAM truthfulness of Debiased-CAM, e.g., for the "Church" image, P1 wrote that Unbiased-CAM and Biased-CAM *"had the most focus on *all* the crosses on the roof of the church and therefore I thought they were the most representative.* [Debiased-CAM] *gives less importance to the leftmost cross on the roof and therefore was rated lower."* For Weak blur, participants felt Unbiased-CAM and Debiased-CAM were very truthful, with Debiased-CAM as slightly less truthful, and Biased-CAM as untruthful; e.g., P29 felt that Biased-CAM *"doesn't show anything but blackness,* [other CAMs] *are much better in the way the heatmap shows details."* For Strong blur, participants perceived Debiased-CAM as moderately truthful, but Biased-CAM as very untruthful, e.g., P18 felt that *"*[Biased-

CAM] *is totally off, nothing there is a garbage truck.* [Unbiased-CAM] *shows the best and biggest area, and* [Debiased-CAM] *is good too but I'm thinking not good enough as* [Unbiased-CAM]*."*

Across blur conditions, we found that participants interpreted whether a CAM was truthful based on several criteria — primary object, object parts, irrelevant object, coverage span, and shape. Participants checked whether the primary object in the label was highlighted (e.g., *"That heatmap that focuses on the chainsaw itself is the most representative."* P20, Chain Saw), and also checked whether specific parts of the primary object were included in the highlights (e.g., *"*[Unbiased-CAM and Debiased-CAM] *correctly identify the fish though* [Unbiased-CAM] *also gives importance to the fish's rear fin."* P1, Fish, Weak blur). P15 noted differences between the CAMs for the "French Horn" image: *"*[Unbiased-CAM] *places the emphasis over the unique body of the French horn, and it places more well defined, yellow and green emphasis on the mouthpiece and the opening of the horn itself.* [Biased-CAM] *is too vertical to completely capture the whole horn, and* [Debiased-CAM]'s *red area is too small to capture the body of the horn, and does not capture the opening of the horn or the mouthpiece."* Participants rated a CAM as less truthful if it highlighted irrelevant objects, e.g., *"*[Debiased-CAM] *is quite close to capturing the entire church. (But)* [Unbiased-CAM] *captures more of the tree."* (P26, Church). Much discussion also focused on the coverage of salient pixels. Less truthful CAMs had coverages that were either too wide (e.g., *"*[Debiased and Biased CAMs] *are inaccurate. They are too wide."* P22, Garbage Truck), covering the background or other objects to get *"less representative when it misleads you into the background or surroundings of the focus. It needs to only emphasize the critical area."* (P23, Church); or too narrow, not covering enough of the key object such that it *"is very small and does not highlight the important part of the image. It is too narrow."* (P30, Fish). Finally, participants appreciated CAMs that highlighted the correct shape of the primary object, e.g., *"*[Debiased-CAM] *perfectly captures the shape of the ball and all of its quadrants.* [Unbiased-CAM] *is a little more oblong than the golf ball itself, so it's not as perfect.* [Biased-CAM] *is almost a vertical red spot and does not really capture the shape of the golf ball at all."* (P15, Golf Ball).

In summary, these qualitative findings explain that Debiased-CAM and Unbiased-CAM were perceived as truthful, because they: 1) highlighted semantically relevant targets while avoiding irrelevant ones, so concept or object-aware CNN models are important[34,49]; 2) had salient regions that were neither too wide nor narrow for the image domain; and 3) had accurate shape and edge boundaries for salient regions, which can be obtained from gradient explanations[44].

Method 12 CAM Helpfulness User Study 2 Qualitative Analysis and Results

To better understand why participants rated the CAMs as helpful or unhelpful, we analyzed the rationale of both their preconceived ratings when seeing the blurred image and consequent ratings after seeing the unblurred image scene. We performed a thematic analysis similar to User Study 1. These results elucidate the mental model of how truthful and debiased CAMs were useful even for blurred images. We found that differences in rationale depended much on image Blur Bias level.

For unblurred images (None blur level), participants mostly felt that CAMs were helpful, because CAMs helped to: 1) focus their attention *"on the most important part of the image, which helps me to quickly identify and label the image."* (Participant P106, Garbage Truck); 2) ignore irrelevant targets to *"lets me know I can disregard the person in the foreground"* (P89, Dog), *"It helps hone in on what the content is, and helps to ignore the extra things in the frame."* (P14, Chain Saw); and 3) matched their expectations since they *"did a solid job of identifying the garbage truck."* (P36) and was *"highly correlated to where the fish is in this image."* (P38). Conversely, as expected, many participants felt that CAMs were unhelpful because *"I could easily identify the object in the image without the heatmap"* (P32, Church).

For images with Weak blur, a truthful CAM: 4) *"helps focus my attention to that area on the blurry picture"* (P105, Debiased-CAM), *"clearly give hint on what was needed to notice in the photo"* (P140, Unbiased-CAM); and 5) helped to confirm image labels, e.g., P3 felt that *"the heatmap gives me the idea that the object might be a fish, I could not tell otherwise"* and wrote after seeing the unblurred image that *"I would not have known what the object was without the heatmap."* P118 described how Unbiased-CAM *"pointed to the steeple and it helped me realize that it was indeed a picture of a church. I did have trouble recognizing it on my own."* Debiased-CAMs helped to locate suspected objects in unexpected images, e.g., P96 felt that *"based on what the heatmap is marking, that's the exact spot where someone would hold a french horn"*, and P67 noted *"that is not an area where I would expect to find a fish, so it's helpful to have this guide."*

For images with Strong blur, many participants felt that the CAMs were very unhelpful, because 6) the task was too difficult such that they had *"NO idea what image is and heatmap doesn't help."* (P68, Biased-CAM), felt the task was *"was very hard, i could not figure it out"* (P71, Debiased-CAM), did not have much initial trust as *"I feel that the heatmap could be wrong because of the clarity of the image."* (P62, Unbiased-CAM). Some participants would 7) blindly trust the CAM due to a lack of other information such that *"without the heatmap and the suggestion, I would have no guess for what this is. I am flying a bit blind. So, I concur with the recommendation (french horn) until I see more."* (P92, Unbiased-CAM) and due to the trustful expectation that CAM *"enables me to know the most useful part in the camera."* (P138, Church, Unbiased-CAM). Finally, we found that 8) confirmation bias may cause the CAM correctness to be misjudged. For example, P76 first thought a misleading Biased-CAM *"helps make a blurry picture more clear"*, but later realized *"it's in the wrong spot."* ("Garbage Truck" image); in contrast, P24 wrongly accused that an Unbiased-CAM *"was focused on the wrong thing"*, but changed his opinion after seeing the unblurred image, admitting that *"Now that I see it's a dog, it is more clear."*

In summary, these findings explain why truthful Debiased-CAM and Unbiased-CAM helped participants to verify classifications of unblurred or weakly blurred images. For unblurred images, these CAMs: 1) focused user attention to relevant objects to speed up verification, 2) averted attention from irrelevant targets to simplify decision making, and 3) matched user expectations[24] of the target object shapes. For weakly blurred images, these CAMs: 4) provided hints on which parts to study in blurred images, and 5) supported hypothesis formation and confirmation[7,67] of

suspected or unexpected objects. For strongly blurred images, participants generally rated all CAMs as unhelpful because: 6) verifying the images was too difficult, 7) they felt misguided to blindly trust the CAMs, and 8) they misjudged the CAMs based on preconceived notions, i.e., confirmation bias[7].

DATA AVAILABILITY

Datasets used in this work, including ImageNette[40], NTCIR-12[42], and COCO[43] are freely available. Data from simulation and user studies are available at https://doi.org/10.17605/OSF.IO/3AU4R.

CODE AVAILABILITY

A reference implementation of the proposed model applied to the ImageNette dataset can be found at https://github.com/nus-ubicomplab/debiased-cam. The code that supports the findings in the study is available from the corresponding author on reasonable request.

REFERENCES

1. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet classification with deep convolutional neural networks. in *Advances in Neural Information Processing Systems* (2012).
2. Esteva, A. *et al.* Dermatologist-level classification of skin cancer with deep neural networks. *Nature* (2017). doi:10.1038/nature21056
3. Vinyals, O., Toshev, A., Bengio, S. & Erhan, D. Show and tell: A neural image caption generator. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 3156–3164 (2015).
4. Das, A. *et al.* Visual dialog. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 326–335 (2017).
5. Lipton, Z. C. The mythos of model interpretability. *Queue* **16**, 31--57 (2018).
6. Miller, T. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* (2019). doi:10.1016/j.artint.2018.07.007
7. Wang, D., Yang, Q., Abdul, A. & Lim, B. Y. Designing theory-driven user-centric explainable AI. in *Proceedings of the 2019 CHI conference on human factors in computing systems* 1–15 (2019).
8. Guidotti, R., Monreale, A. & Ruggieri, S. A Survey of Methods for Explaining Black Box Models. *ACM Comput. Surv.* **51**, (2018).
9. Hohman, F., Kahng, M., Pienta, R. & Chau, D. H. Visual Analytics in Deep Learning: An Interrogative Survey for the Next Frontiers. *IEEE Trans. Vis. Comput. Graph.* **25**, 2674–2693 (2019).
10. Zhang, Q. & Zhu, S. chun. Visual interpretability for deep learning: a survey. *Frontiers of Information Technology and Electronic Engineering* (2018). doi:10.1631/FITEE.1700808
11. Simonyan, K., Vedaldi, A. & Zisserman, A. Deep Inside Convolutional Networks: Visualising Image Classification Models and Saliency Maps. in *Workshop at International Conference on Learning Representations* 1–8 (2014).
12. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. in *Proceedings of the IEEE International Conference on Computer Vision* **2017-Octob**, 618–626 (2017).

13. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Learning Deep Features for Discriminative Localization. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2016). doi:10.1109/CVPR.2016.319
14. Zhang, Q., Nian Wu, Y. & Zhu, S.-C. Interpretable convolutional neural networks. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 8827–8836 (2018).
15. Chattopadhay, A., Sarkar, A., Howlader, P. & Balasubramanian, V. N. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. in *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)* 839–847 (2018).
16. Desai, S. & Ramaswamy, H. G. Ablation-CAM: Visual Explanations for Deep Convolutional Network via Gradient-free Localization. in *The IEEE Winter Conference on Applications of Computer Vision* 983–991 (2020).
17. Kupyn, O., Budzan, V., Mykhailych, M., Mishkin, D. & Matas, J. DeblurGAN: Blind Motion Deblurring Using Conditional Adversarial Networks. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (2018). doi:10.1109/CVPR.2018.00854
18. Vasiljevic, I., Chakrabarti, A. & Shakhnarovich, G. Examining the Impact of Blur on Recognition by Convolutional Networks. (2016).
19. Dimiccoli, M., Marín, J. & Thomaz, E. Mitigating Bystander Privacy Concerns in Egocentric Activity Recognition with Deep Learning and Intentional Image Degradation. *Proc. ACM Interactive, Mobile, Wearable Ubiquitous Technol.* **1**, 1–18 (2018).
20. Afifi, M. & Brown, M. S. What else can fool deep learning? Addressing color constancy errors on deep neural network performance. in *Proceedings of the IEEE International Conference on Computer Vision* 243–252 (2019).
21. Posner, M. I., Snyder, C. R. & Solso, R. Attention and cognitive control. *Cogn. Psychol. Key readings* **205**, (2004).
22. Du, M., Liu, N. & Hu, X. Techniques for interpretable machine learning. *Commun. ACM* **63**, 68–77 (2019).
23. Lapuschkin, S. *et al.* Unmasking clever hans predictors and assessing what machines really learn. *Nat. Commun.* **10**, 1–8 (2019).
24. Ross, A. S., Hughes, M. C. & Doshi-Velez, F. Right for the right reasons: Training differentiable models by constraining their explanations. in *Proceedings of the 26th International Joint Conference on Artificial Intelligence* 2662--2670 (2017).
25. Jetley, S., Lord, N. A., Lee, N. & Torr, P. H. S. Learn to pay attention. in *6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings* (2018).
26. Li, K., Wu, Z., Peng, K.-C., Ernst, J. & Fu, Y. Tell me where to look: Guided attention inference network. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 9215–9223 (2018).
27. Zagoruyko, S. & Komodakis, N. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. in *5th International Conference on Learning Representations, ICLR 2017 - Conference Track Proceedings* (2017).

28. Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y. & Kankanhalli, M. Trends and Trajectories for Explainable, Accountable and Intelligible Systems. in *Proceedings of the 2018 CHI conference on human factors in computing systems* 1–18 (2018). doi:10.1145/3173574.3174156
29. Alqaraawi, A., Schuessler, M., Weiß, P., Costanza, E. & Berthouze, N. Evaluating saliency map explanations for convolutional neural networks: a user study. in *Proceedings of the 25th International Conference on Intelligent User Interfaces* 275–285 (2020).
30. Doshi-Velez, F. & Kim, B. Towards a rigorous science of interpretable machine learning. *Stat* **1050**, 2 (2017).
31. Bach, S. *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* (2015). doi:10.1371/journal.pone.0130140
32. Fong, R. C. & Vedaldi, A. Interpretable explanations of black boxes by meaningful perturbation. in *Proceedings of the IEEE International Conference on Computer Vision* 3429–3437 (2017).
33. Sundararajan, M., Taly, A. & Yan, Q. Axiomatic Attribution for Deep Networks. in *Proceedings of the 34th International Conference on Machine Learning - Volume 70* 3319–3328 (JMLR.org, 2017).
34. Bau, D., Zhou, B., Khosla, A., Oliva, A. & Torralba, A. Network dissection: Quantifying interpretability of deep visual representations. in *Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017* **2017-Janua**, 3319–3327 (2017).
35. Schramowski, P. *et al.* Making deep neural networks right for the right scientific reasons by interacting with their explanations. *Nat. Mach. Intell.* **2**, 476–486 (2020).
36. Zhou, B., Sun, Y., Bau, D. & Torralba, A. Interpretable basis decomposition for visual explanation. in *Proceedings of the European Conference on Computer Vision (ECCV)* 119–134 (2018).
37. Selvaraju, R. R. *et al.* Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. *Proc. IEEE Int. Conf. Comput. Vis.* **2017-Octob**, 618–626 (2017).
38. Noroozi, M., Vinjimoor, A., Favaro, P. & Pirsiavash, H. Boosting self-supervised learning via knowledge transfer. in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* 9359–9367 (2018).
39. Watanabe, S., Hori, T., Le Roux, J. & Hershey, J. R. Student-teacher network learning with enhanced features. in *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 5275–5279 (2017).
40. Howard, J. The imagenette dataset. *Github* Available at: https://github.com/fastai/imagenette.
41. Rudin, C. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* (2019). doi:10.1038/s42256-019-0048-x
42. Gurrin, C., Joho, H., Hopfgartner, F., Zhou, L. & Albatal, R. Overview of NTCIR-12 Lifelog Task. *12th NTCIR Conf. Eval. Inf. Access Technol.* 354–360 (2016).
43. Chen, X. *et al.* Microsoft coco captions: Data collection and evaluation server. *arXiv Prepr. arXiv1504.00325* (2015).
44. Adebayo, J. *et al.* Sanity checks for saliency maps. in *Advances in Neural Information Processing Systems* 9505–9515 (2018).

45. Ribeiro, M. T., Singh, S. & Guestrin, C. 'Why should i trust you?' Explaining the predictions of any classifier. in *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **13-17-Augu**, 1135–1144 (2016).
46. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. in *Proceedings of the 31st International Conference on Neural Information Processing Systems* 4768–4777 (2017).
47. Olah, C., Mordvintsev, A. & Schubert, L. Feature Visualization. *Distill* (2017). doi:10.23915/distill.00007
48. Ribeiro, M. T., Singh, S. & Guestrin, C. Anchors: High-precision model-agnostic explanations. in *32nd AAAI Conference on Artificial Intelligence, AAAI 2018* (2018).
49. Kim, B. *et al.* Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav). in *International conference on machine learning* 2668–2677 (2018).
50. Park, S., Yu, S., Kim, M., Park, K. & Paik, J. Dual Autoencoder Network for Retinex-Based Low-Light Image Enhancement. *IEEE Access* (2018). doi:10.1109/ACCESS.2018.2812809
51. Cohen, B. & Dinstein, I. New maximum likelihood motion estimation schemes for noisy ultrasound images. *Pattern Recognit.* (2002). doi:10.1016/S0031-3203(01)00053-X
52. McLoughlin, I., Zhang, H., Xie, Z., Song, Y. & Xiao, W. Robust sound event classification using deep neural networks. *IEEE/ACM Trans. Audio, Speech, Lang. Process.* **23**, 540–552 (2015).
53. Ryoo, M. S., Rothrock, B., Fleming, C. & Yang, H. J. Privacy-preserving human activity recognition from extreme low resolution. in *31st AAAI Conference on Artificial Intelligence, AAAI 2017* (2017).
54. Charity, M. What color is a blackbody? - some pixel rgb values. Available at: http://www.vendian.org/mncharity/dir3/blackbody/.
55. Davis, R. G. & Ginthner, D. N. Correlated color temperature, illuminance level, and the kruithof curve. *J. Illum. Eng. Soc.* (1990). doi:10.1080/00994480.1990.10747937
56. Saito, T. & Rehmsmeier, M. The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets. *PLoS One* **10**, (2015).
57. Papineni, K., Roukos, S., Ward, T. & Zhu, W.-J. BLEU: a method for automatic evaluation of machine translation. in *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* 311–318 (2002).
58. Li, J., Xia, C., Song, Y., Fang, S. & Chen, X. A data-driven metric for comprehensive evaluation of saliency models. in *Proceedings of the IEEE International Conference on Computer Vision* (2015). doi:10.1109/ICCV.2015.30
59. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A. & Durand, F. What do different evaluation metrics tell us about saliency models? *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 740–757 (2018).
60. Angel, R. & Gronfein, W. The Use of Subjective Information in Statistical Models. *Am. Sociol. Rev.* (1988). doi:10.2307/2095653
61. Avrahami, D., Fogarty, J. & Hudson, S. E. Biases in human estimation of interruptibility: Effects and implications for practice. in *Conference on Human Factors in Computing Systems - Proceedings* (2007). doi:10.1145/1240624.1240632
62. Goldstein, D. G. & Rothschild, D. Lay understanding of probability distributions. *Judgm. Decis. Mak.* **9**, (2014).

63. Sharpe, W. F., Goldstein, D. G. & Blythe, P. W. The Distribution Builder : A Tool for Inferring Investor Preferences. *October* (2000).
64. Goldstein, D. G., Johnson, E. J. & Sharpe, W. F. Choosing outcomes versus choosing products: Consumer-focused retirement investment advice. *J. Consum. Res.* (2008). doi:10.1086/589562
65. Delavande, A. & Rohwedder, S. Eliciting subjective probabilities in internet surveys. *Public Opin. Q.* (2008). doi:10.1093/poq/nfn062
66. Muller, M. & Kogan, S. Grounded theory method in hci and cscw. *Cambridge IBM Cent. Soc. Softw.* 1–46 (2010).
67. Hohman, F., Head, A., Caruana, R., DeLine, R. & Drucker, S. M. Gamut: A design probe to understand how data scientists understand machine learning models. in *Conference on Human Factors in Computing Systems - Proceedings* (2019). doi:10.1145/3290605.3300809
68. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. in *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings* (2015).
69. He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition* **2016-Decem**, 770–778 (IEEE Computer Society, 2016).
70. Chollet, F. Xception: Deep learning with depthwise separable convolutions. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 1251–1258 (2017).
71. Szegedy, C., Vanhoucke, V., Ioffe, S., Shlens, J. & Wojna, Z. Rethinking the inception architecture for computer vision. in *Proceedings of the IEEE conference on computer vision and pattern recognition* 2818–2826 (2016).
72. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. in *2009 IEEE conference on computer vision and pattern recognition* 248–255 (2009).

ACKNOWLEDGEMENTS

This work was carried out in part at the NUS School of Computing and NUS N-CRiPT. This research is supported by the Ministry of Education, Singapore, and National Research Foundation, Prime Minister's Office, Singapore under its Strategic Capability Research Centres Funding Initiative. Titan X Pascal GPU cards used for this research were donated by NVIDIA Corporation.

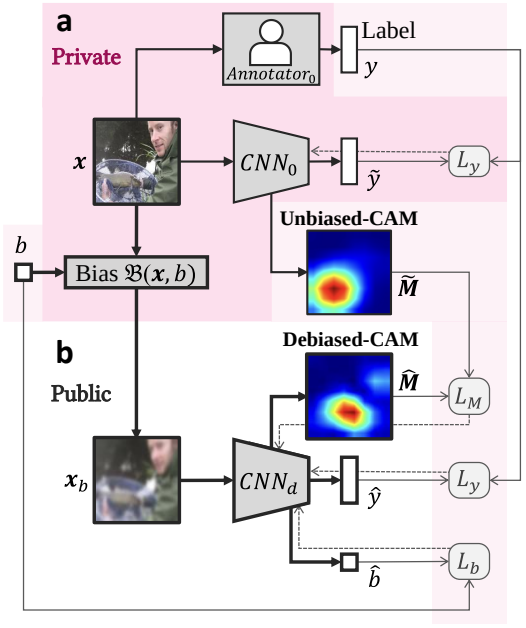
AUTHOR CONTRIBUTIONS

B.Y.L. and W.Z. contributed the original insight, designed the technical approach and algorithms, designed the simulation studies and user studies, wrote the surveys, performed statistical analyses, and wrote the paper. M.D. provided annotations of the NTCIR-12 dataset, contributed to the technical approach, and contributed writing. B.Y.L. supervised the research and provided funding for the work.

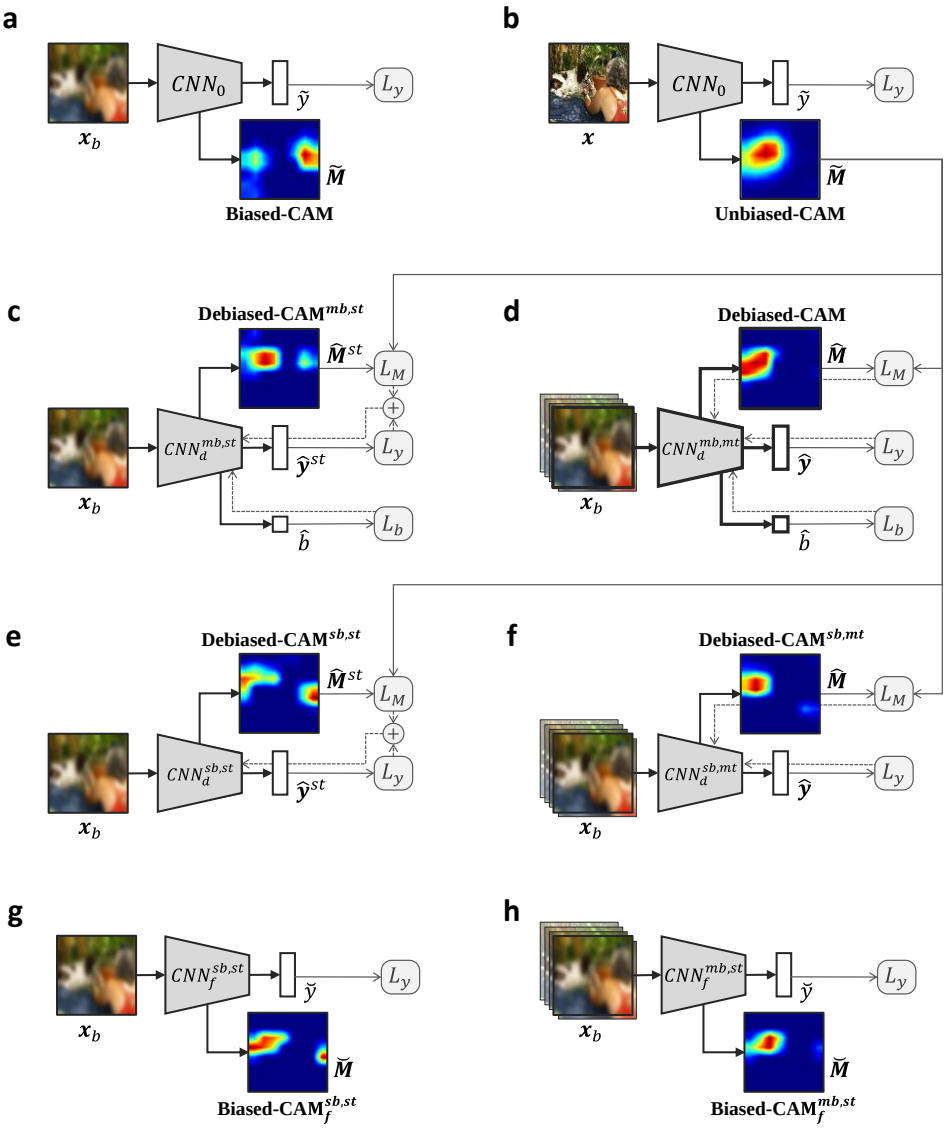
COMPETING INTERESTS

The authors declare no competing interests.

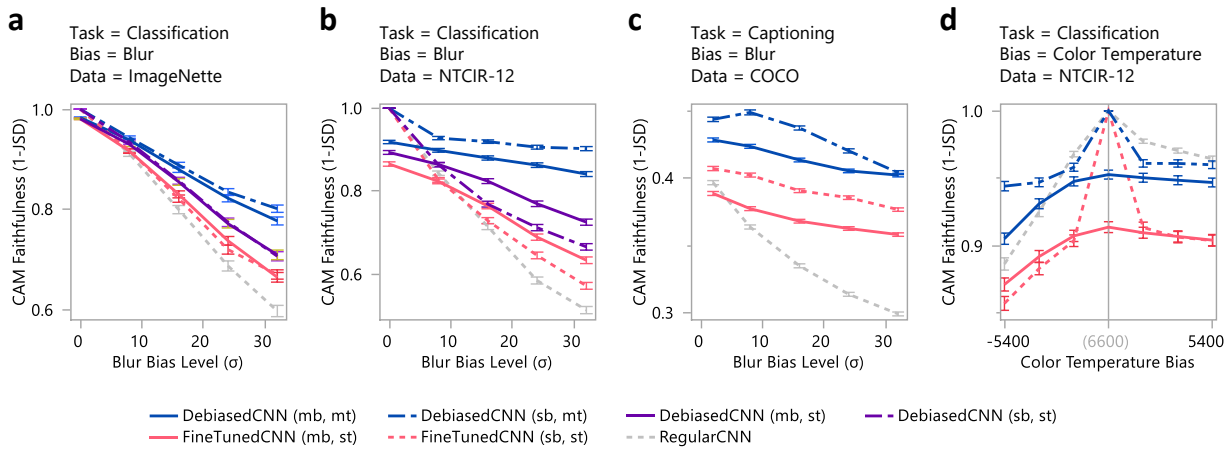
SUPPLEMENTARY FIGURES



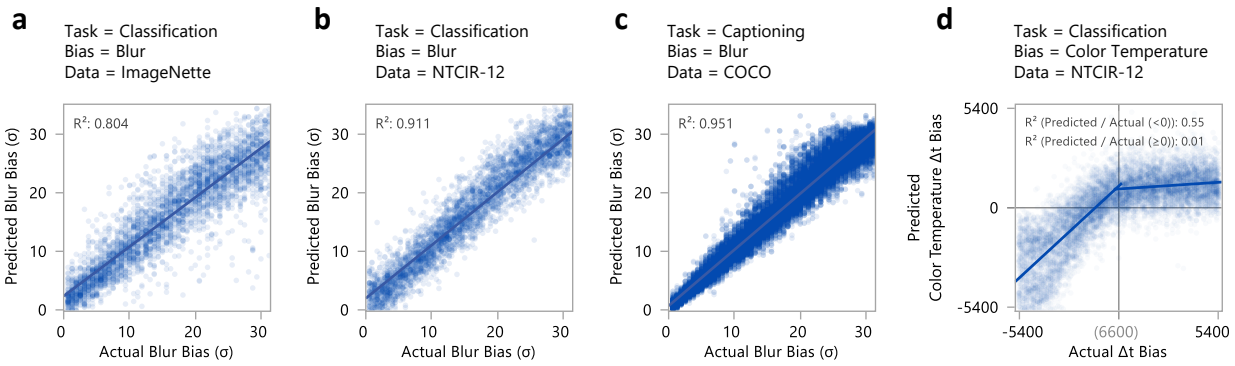
Supplementary Fig. 1 | Architecture of multi-task debiased CNN model with self-supervised learning from private training data for privacy-preserving machine learning. a, RegularCNN (CNN_0) was trained on a private unbiased dataset with unblurred image $\boldsymbol{x}$ to generate Unbiased-CAM $\widetilde{\boldsymbol{M}}$. **b,** DebiasedCNN ($CNN_d$) was trained on the corresponding public (privacy-protected) biased form of the private dataset with blurred image $\boldsymbol{x}_b$ and self-supervised with Unbiased-CAM $\widetilde{\boldsymbol{M}}$ to generate Debiased-CAM $\widehat{\boldsymbol{M}}$. During model training, CNN_d has access to the bias level b of each training image $\boldsymbol{x}_b$, Unbiased-CAM $\widetilde{\boldsymbol{M}}$, and actual label y, but has no access to them during model inference. CNN_d never has access to any unblurred image $\boldsymbol{x}$. At inference time, DebiasedCNN can generate relevant and faithful Debiased-CAMs from privacy-protected blurred images.



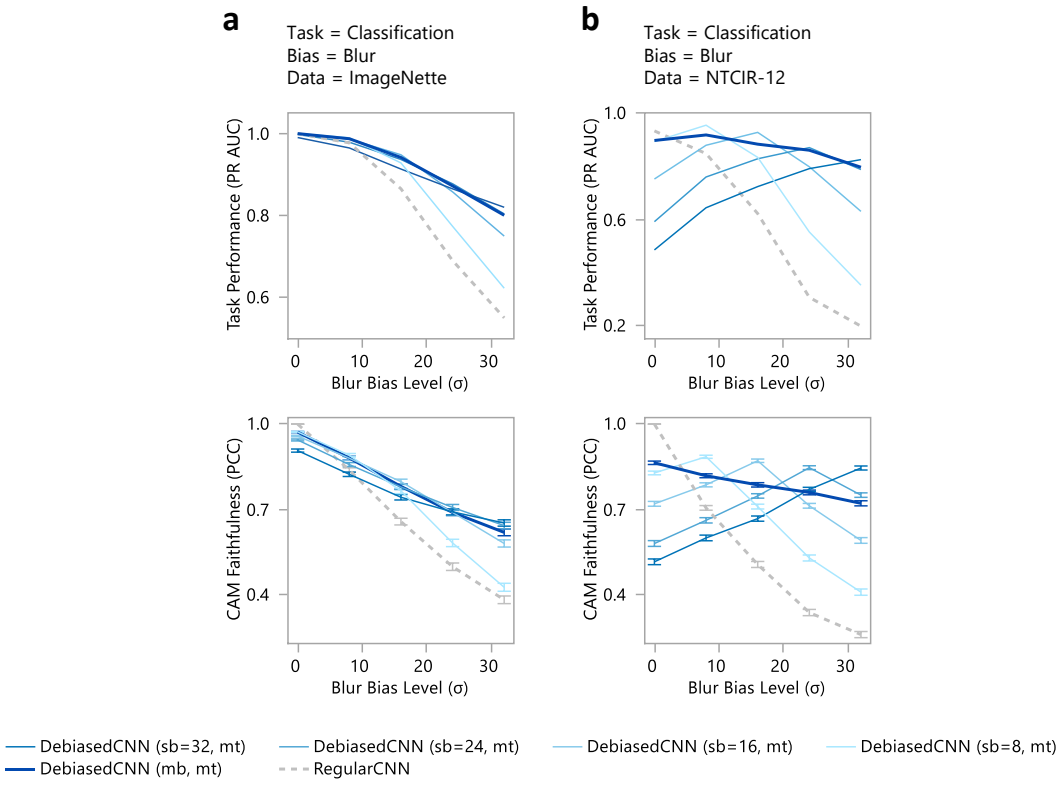
Supplementary Fig. 2 | **Architectures of self-supervised DebiasedCNN variants and of baseline CNN models and their CAM explanations from a biased "Dog" image blurred at $\sigma = 24$. a,** RegularCNN on biased image. **b,** RegularCNN on unbiased image. **c,** DebiasedCNN (mb, st) with single-task loss as a sum of classification and CAM losses for the classification task, trained on multi-bias images with auxiliary bias level prediction task. **d,** DebiasedCNN (sb, mt) with multi-task for CAM prediction trained with differentiable CAM loss, and trained on multi-bias images with auxiliary bias level prediction task. **e,** DebiasedCNN (sb, st) with single-task loss as a sum of classification and CAM losses for the classification task. **f,** DebiasedCNN (sb, mt) with multi-task for the CAM prediction and differentiable CAM loss. **g,** FineTunedCNN (sb,st) retrained on images biased at a single-bias level. **h,** FineTunedCNN (mb,st) retrained on images biased variously at multi-bias levels.



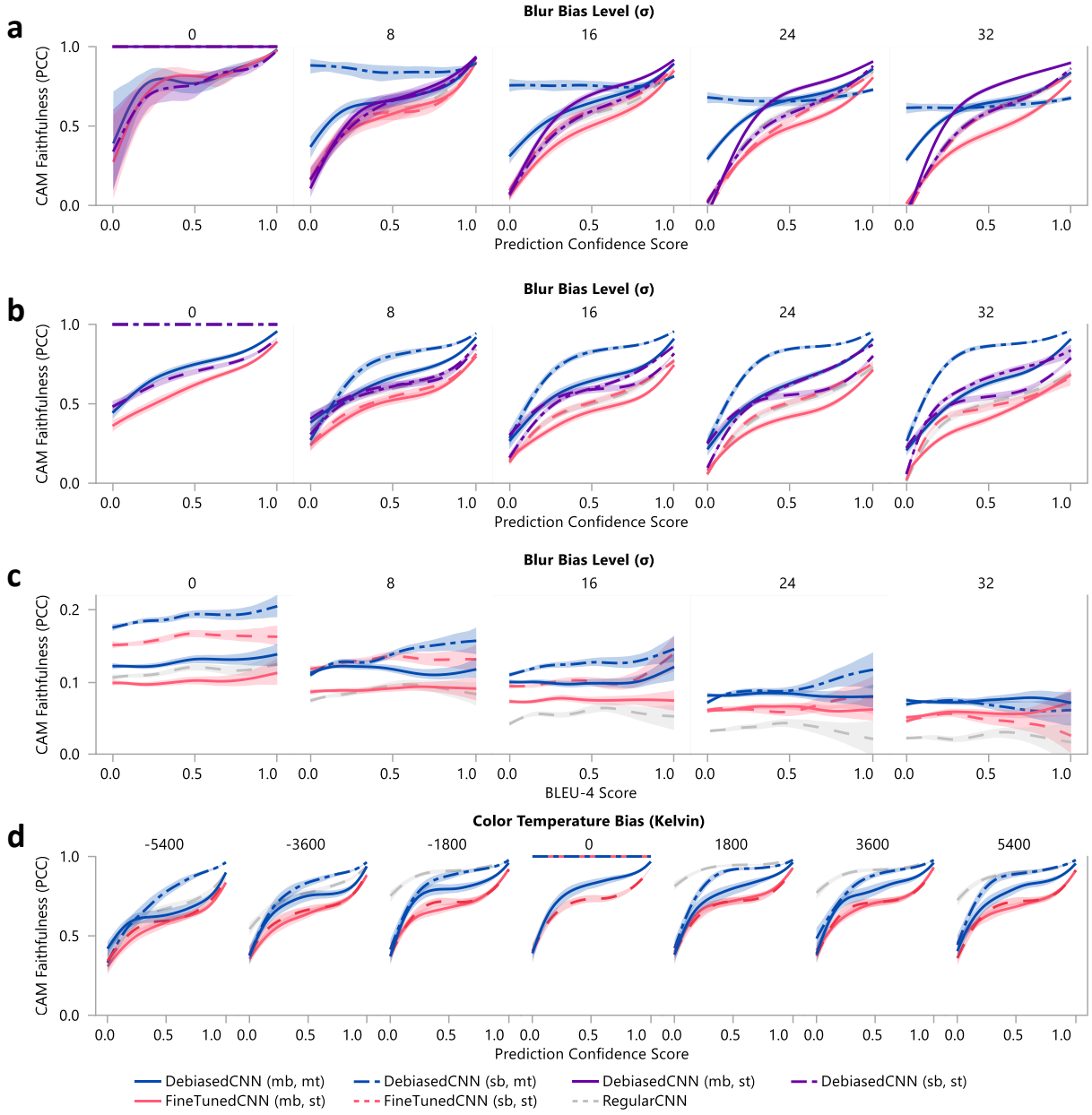
Supplementary Fig. 3 | Comparisons of CAM Faithfulness calculated with Jensen-Shannon Divergence (JSD) between CAM and Unbiased-CAM for increasing bias with different CNN models across four simulation studies. a, Simulation Study 1 (classification with blur-biased ImageNette), **b,** Simulation Study 2 (classification with blur-biased NTCIR-12), **c,** Simulation Study 3 (captioning with blur-biased COCO), **d,** Simulation Study 4 (classification with color temperature-biased NTCIR-12)**. a-d,** Similarly to Fig. 4, CAM Faithfulness decreased with increasing bias level. Error bars indicate 90% confidence interval. Model variants annotated as st = single-task, mt = multi-task, sb = single-bias, mb = multi-bias. **a-c,** σ indicates the Gaussian blur standard deviation for the normalized image. **d,** Color temperature bias (in Kelvin) was varied with respect to neutral cloudy daylight at 6600K.



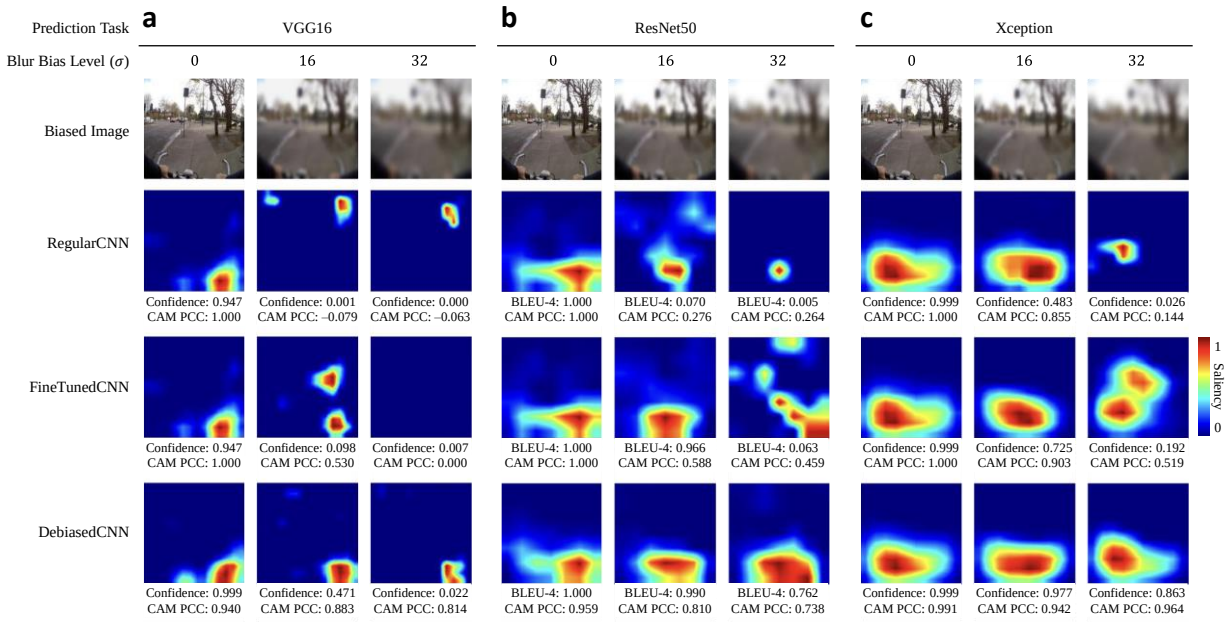
Supplementary Fig. 4 | Regression performance for DebiasedCNN (mb, mt) measured as R^2 for the bias level prediction task for four simulation studies. a, Simulation Study 1 (classification with blur-biased ImageNette), **b,** Simulation Study 2 (classification with blur-biased NTCIR-12), **c,** Simulation Study 3 (captioning with blur-biased COCO), **d,** Simulation Study 4 (classification with color temperature-biased NTCIR-12)**. a-c,** Very high R^2 values indicate that models trained for Simulation Studies 1-3 could predict the respective bias levels well. **d,** Color temperature bias level prediction depended on whether bias was towards lower (more orange) or higher (more blue) temperatures. Since blue-biased images were less distinguishable (Method 4), the model was less well-trained to predict the blue color temperature bias level; it was more able to predict orange bias at reasonable accuracy.



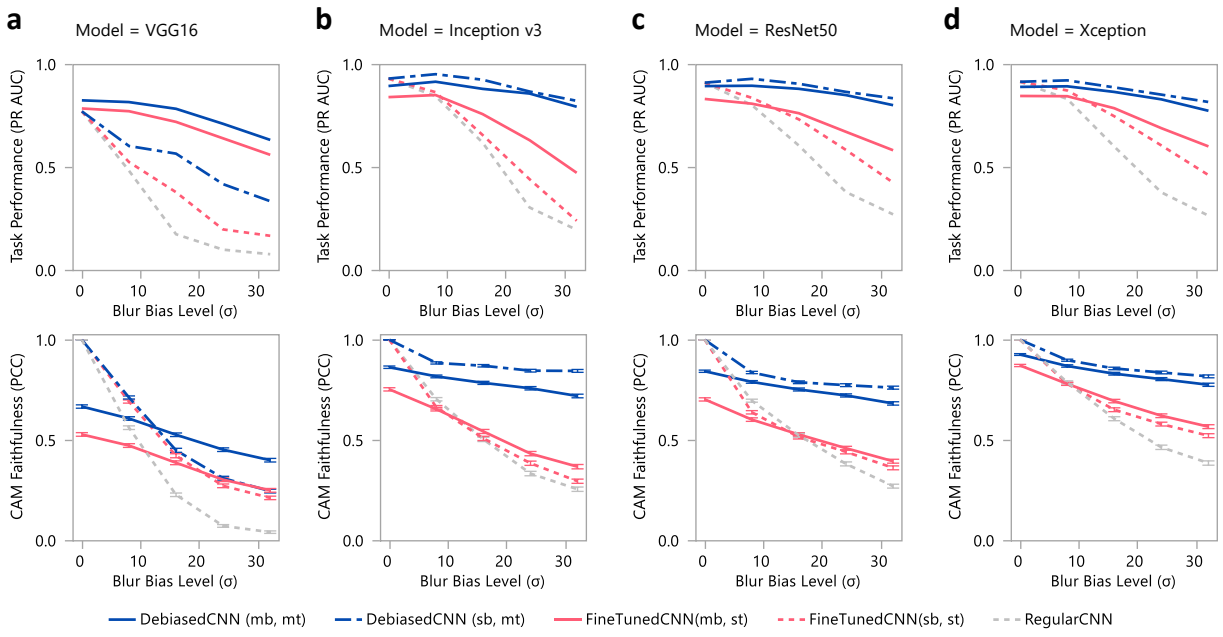
Supplementary Fig. 5 | Comparison of model Task Performance and CAM Faithfulness for single-bias (sb) and multi-bias (mb) DebiasedCNN variants evaluated across multiple bias levels. a, Simulation Study 1 (classification with blur biased ImageNette), **b,** Simulation Study 2 (classification with blur biased NTCIR-12). **a,b,** Each single-bias model had the best Task Performance and CAM Faithfulness for the specific bias level at which it was trained: $\sigma = 0$ for RegularCNN, $\sigma = 8$ for DebiasedCNN (sb = 8, mt), $\sigma = 16$ for DebiasedCNN (sb = 16, mt), $\sigma = 24$ for DebiasedCNN (sb = 24, mt), $\sigma = 32$ for DebiasedCNN (sb = 32, mt). However, they were less performant and faithful than multi-bias DebiasedCNN (mb, mt) for other non-specific bias levels. **a-d,** Primary Task Performance (first row) was calculated as area under precision-recall curve (PR AUC) for classification. CAM Faithfulness (second row) was calculated as the Pearson's Correlation Coefficient (PCC) between CAM and Unbiased-CAM. Error bars indicate 90% confidence interval in PCC. Model variants annotated as st = single-task, mt = multi-task, sb = single-bias, mb = multi-bias.



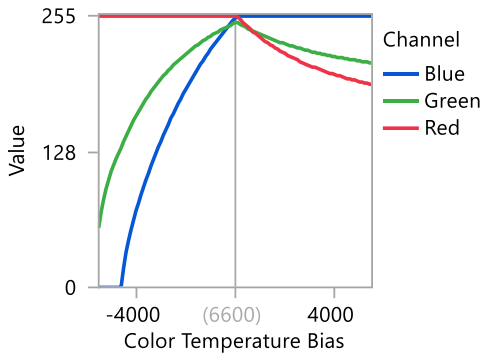
Supplementary Fig. 6 | Comparison of CAM Faithfulness (PCC) with model Prediction Confidence across the four simulation studies. a, Simulation Study 1 (classification with blur-biased ImageNette), **b,** Simulation Study 2 (classification with blur-biased NTCIR-12), **c,** Simulation Study 3 (captioning with blur-biased COCO), **d,** Simulation Study 4 (classification with color temperature-biased NTCIR-12)**. a-d,** In general, CAM Faithfulness increases with model prediction confidence, but decreases with bias level. DebiasedCNN had higher CAM Faithfulness than FineTunedCNN and RegularCNN, and had much higher CAM Faithfulness even at moderately low (about 40%) confidences. **c,** For image captioning, all models had low CAM Faithfulness that did not vary with Task Performance, and low Task Performance. **a-d,** Smooth trend lines are estimated by fitting cubic splines for each row with λ parameter set to λ = 15.6 (**a,b,d**) and λ = 1020 (**c**). Confidence areas indicate 90% confidence interval.



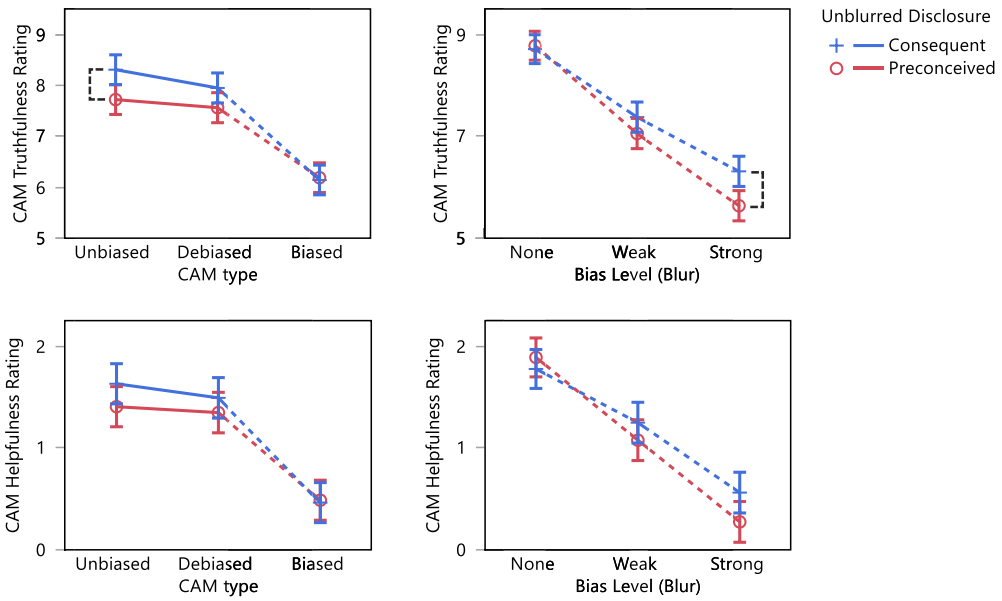
Supplementary Fig. 7 | Deviated and debiased CAM explanations from various CNN models at varying blur bias levels of blur biased image from NTCIR-12 labeled as "Biking". **a** VGG16[68], **b** ResNet50[69], **c** Xception[70]. **a-c,** Models arranged in increasing CAM Faithfulness (see Supplementary Fig. 8, second row). CAMs generated from more performant models were more representative of the image label with higher CAM Faithfulness (PCC).



Supplementary Fig. 8 | Comparison of model Task Performance and CAM Faithfulness for image classification on NTCIR-12 trained with different CNN models. a, VGG16[68], **b** Inception v3[71], **c** ResNet50[69], **d** Xception[70]**. a-d,** Results agreed with Fig. 4 that higher bias led to lower Task Performance and CAM Faithfulness, but debiasing improved both. CNN models are arranged in increasing CAM Faithfulness from left to right. All models were pretrained on ImageNet and fine-tune on NTCIR-12. We set the last two layers of VGG16, and last block of ResNet50 and Xception as retrainable. **b-d,** Newer base CNN models than VGG16 significantly outperformed it for both Task Performance and CAM Faithfulness. These newer models had similar Task Performance across bias level, though their CAM Faithfulness differed more notably. **a-d,** Primary Task Performance (first row) was calculated as area under precision-recall curve (PR AUC) for classification. CAM Faithfulness (second row) was calculated as the Pearson's Correlation Coefficient (PCC) between CAM and Unbiased-CAM. Error bars indicate 90% confidence interval in PCC. Model variants annotated as st = single-task, mt = multi-task, sb = single-bias, mb = multi-bias. **a-c,** σ indicates the Gaussian blur standard deviation for the normalized image. **d,** Color temperature bias (in Kelvin) was varied with respect to neutral cloudy daylight at 6600K.



Supplementary Fig. 9 | Color mapping function to bias color temperature of images in Simulation Study 4. Changes in Red, Green, Blue values are larger for orange biases (lower color temperature) than blue biases (higher temperature). Neutral color temperature is set to represent shaded/overcast skylight at 6600K.



Supplementary Fig. 10 | Comparisons of perceived CAM Truthfulness and CAM Helpfulness before (preconceived) and after (consequent) disclosing the unblurred image. There was a significant difference across Unblurred Disclosure for CAM Truthfulness Rating ($p = .0013$) but not for CAM Helpfulness Rating (Supplementary Table 4). Comparing preconceptual to consequent ratings, Unbiased-CAMs were rated as less truthful ($M = 7.7$ vs. 8.3, $p = .0004$), Debiased-CAMs were rated marginally less truthful ($p = .0212$), Biased-CAMs were rated similarly untruthful, and overall, CAMs of Strongly blurred images were rated as less truthful ($M = 5.6$ vs. 6.3, $p<.0001$). These results suggest that even with the least biased CAM (Unbiased-CAM), the unfamiliarity of unblurred scenes can hurt trust (truthfulness) in the CAM, though there was no change in perceived helpfulness before or after disclosing the unblurred image. CAM Truthfulness Ratings were measured along a 1-10 scale, and CAM Helpfulness Ratings along a 7-point Likert scale (–3 = Strongly Disagree, 0 = Neither, +3 = Strongly Agree). Error bars indicate 90% confidence interval. Dotted lines indicate extremely significant $p<.0001$ comparisons, and solid lines indicate no significance at $p>.01$.

SUPPLEMENTARY TABLES

Supplementary Table 1 | CNN model variants with single-task (st) or multi-task (mt) architectures trained on a specific (sb) or multiple (mb) bias levels. Each training set image $\boldsymbol{x} \in \boldsymbol{X}$ is preprocessed by a bias operator $\mathfrak{B}$ at a selected level b, i.e., $\boldsymbol{x_b} = \mathfrak{B}(\boldsymbol{x}, |b| > 0), \forall \boldsymbol{x} \in \boldsymbol{X}$. $\mathfrak{B}$ depends on the bias type (blur or color temperature). For DebiasedCNN, mt refers to including a CAM task with differentiable CAM loss separate from the primary prediction task, while st refers to the primary prediction task with non-differentiable CAM loss. Models trained for single-bias (sb) used training set images biased at a single level $b > 0$, while models trained for multi-bias levels (mb) used training datasets with data augmentation where each image is biased to a level that is randomly selected from a uniform probability distribution $B_{rand} \sim U([0, b_{max}])$. Multi-bias DebiasedCNN also adds a task for bias level prediction. Loss functions in vector form specify one loss function per task in a multi-task architecture.

Model Variant	Training Loss Function	Training Set Bias Levels
RegularCNN	$L(\boldsymbol{w}) = L_y(y, \hat{y}(\boldsymbol{w}))$	$b = 0$
FineTunedCNN (sb, st)	$L(\boldsymbol{w}) = L_y(y, \hat{y}(\boldsymbol{w}))$	$b \in (0, b_{max}]$
FineTunedCNN (mb, st)	$L(\boldsymbol{w}) = L_y(y, \hat{y}(\boldsymbol{w}))$	$b \in B_{rand} \sim U([0, b_{max}])$
DebiasedCNN (sb, st)	$L(\boldsymbol{w}) = L_y(y, \hat{y}(\boldsymbol{w})) + \omega_M L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}})$	$b \in (0, b_{max}]$
DebiasedCNN (mb, st)	$\boldsymbol{L}(\boldsymbol{w}) = \begin{pmatrix} L_y(y, \hat{y}(\boldsymbol{w})) + \omega_M\, L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}}) \\ \omega_b\, L_b(b, \hat{b}(\boldsymbol{w})) \end{pmatrix}$	$b \in B_{rand}$
DebiasedCNN (sb, mt)	$\boldsymbol{L}(\boldsymbol{w}) = \begin{pmatrix} L_y(y, \hat{y}(\boldsymbol{w})) \\ \omega_M\, L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}}(\boldsymbol{w})) \end{pmatrix}$	$b \in (0, b_{max}]$
DebiasedCNN (mb, mt)	$\boldsymbol{L}(\boldsymbol{w}) = \begin{pmatrix} L_y(y, \hat{y}(\boldsymbol{w})) \\ \omega_M\, L_M(\widetilde{\boldsymbol{M}}, \widehat{\boldsymbol{M}}(\boldsymbol{w})) \\ \omega_b\, L_b(b, \hat{b}(\boldsymbol{w})) \end{pmatrix}$	$b \in B_{rand}$

Supplementary Table 2 | Baseline CNN models trained on training datasets for four Simulation Studies. All models were pre-trained on ImageNet ILSVRC-2012 and retrained to fine-tune on respective datasets. Train-test ratios were determined from the original literature of the models as referenced. See Supplementary Method 1 for model training details.

Exp	Task	Model	Re-trained Dataset	Train-Test Ratio
1	Classify	Inception v3[71] CNN	ImageNette[40]	70.0% / 30.0%
2	Classify	Inception v3 CNN	NTCIR-12[42]	80.0% / 20.0%
3	Caption	Neural Image Captioner (NIC)[3] (Inception v3 + LSTM)	COCO[43]	66.7% / 33.3%
4	Classify	Inception v3 CNN	NTCIR-12	80.0% / 20.0%

Supplementary Table 3 | Percent improvement in Task Performance and CAM Faithfulness of CNN models compared to RegularCNN at each bias level for a, Simulation Study 1 (classification with blur-biased ImageNette), **b,** Simulation Study 2 (classification with blur-biased NTCIR-12), **c,** Simulation Study 3 (captioning with blur-biased COCO), **d,** Simulation Study 4 (classification with color temperature-biased NTCIR-12)**. a-d,** See Fig. 4 for graphical comparison.

a **Simulation Study 1:** Task = Classification, Bias = Blur, Data = ImageNette

Performance (PR AUC)	Blur Bias Level (σ)				
	0	8	16	24	32
DebiasedCNN (mb, mt)	0%	1%	9%	26%	45%
DebiasedCNN (sb, mt)	0%	1%	9%	27%	49%
DebiasedCNN (mb, st)	0%	1%	6%	18%	33%
DebiasedCNN (sb, st)	0%	1%	5%	14%	19%
FineTunedCNN (mb, st)	0%	0%	6%	18%	29%
FineTunedCNN (sb, st)	0%	0%	4%	12%	21%

CAM Faithfulness (PCC)	Blur Bias Level (σ)				
	0	8	16	24	32
DebiasedCNN (mb, mt)	-3%	5%	19%	39%	63%
DebiasedCNN (sb, mt)	0%	6%	21%	42%	71%
DebiasedCNN (mb, st)	-3%	4%	14%	26%	42%
DebiasedCNN (sb, st)	0%	5.0%	15%	27%	42%
FineTunedCNN (mb, st)	-3%	0%	6%	11%	20%
FineTunedCNN (sb, st)	0%	1%	6%	7%	21%

b **Simulation Study 2:** Task = Classification, Bias = Blur, Data = NTCIR-12

Performance (PR AUC)	Blur Bias Level (σ)				
	0	8	16	24	32
DebiasedCNN (mb, mt)	-4%	8%	42%	181%	302%
DebiasedCNN (sb, mt)	0%	5%	36%	159%	267%
DebiasedCNN (mb, st)	-6%	3%	27%	119%	175%
DebiasedCNN (sb, st)	0%	6%	13%	62%	41%
FineTunedCNN (mb, st)	-10%	1%	22%	107%	140%
FineTunedCNN (sb, st)	0%	2%	6%	45%	22%

CAM Faithfulness (PCC)	Blur Bias Level (σ)				
	0	8	16	24	32
DebiasedCNN (mb, mt)	-14%	16%	56%	127%	181%
DebiasedCNN (sb, mt)	0%	26%	73%	153%	229%
DebiasedCNN (mb, st)	-19%	5%	31%	72%	101%
DebiasedCNN (sb, st)	0%	5%	16%	46%	65%
FineTunedCNN (mb, st)	-25%	-7%	8%	30%	44%
FineTunedCNN (sb, st)	0%	-6%	1%	15%	15%

c **Simulation Study 3:** Task = Captioning, Bias = Blur, Data = COCO

Performance (BLEU-4)	Blur Bias Level (σ)				
	1	8	16	24	32
DebiasedCNN (mb, mt)	-19%	-2%	13%	29%	41%
DebiasedCNN (sb, mt)	9%	0%	9%	11%	5%
FineTunedCNN (mb, st)	-20%	-1%	13%	27%	38%
FineTunedCNN (sb, st)	15%	-2%	-3%	0%	-6%

CAM Faithfulness (PCC)	Blur Bias Level (σ)				
	1	8	16	24	32
DebiasedCNN (mb, mt)	10%	40%	89%	128%	207%
DebiasedCNN (sb, mt)	69%	113%	132%	157%	224%
FineTunedCNN (mb, st)	-13%	5%	39%	72%	124%
FineTunedCNN (sb, st)	38%	47%	82%	73%	113%

d **Simulation Study 4:** Task = Classification, Bias = Color Temperature, Data = NTCIR-12

Performance (PR AUC)	Color Temperature Bias (Δt)						
	-5400	-3600	-1800	0	1800	3600	5400
DebiasedCNN (mb, mt)	7%	6%	5%	4%	4%	4%	5%
DebiasedCNN (sb, mt)	10%	6%	5%	0%	4%	5%	4%
FineTunedCNN (mb, st)	1%	0%	-1%	-1%	-1%	-1%	-1%
FineTunedCNN (sb, st)	2%	-1%	0%	0%	-1%	-1%	0%

CAM Faithfulness (PCC)	Color Temperature Bias (Δt)						
	-5400	-3600	-1800	0	1800	3600	5400
DebiasedCNN (mb, mt)	6%	3%	-3%	-7%	-4%	-3%	-2%
DebiasedCNN (sb, mt)	16%	6%	-1%	0%	-2%	-1%	1%
FineTunedCNN (mb, st)	-5%	-8%	-13%	-16%	-14%	-13%	-12%
FineTunedCNN (sb, st)	-8%	-10%	-13%	0%	-13%	-13%	-12%

Supplementary Table 4 | Statistical analysis of responses due to effects as linear mixed effects models. **a,** Statistical model for CAM Truthfulness User Study 1. **b,** Statistical model for CAM Helpfulness User Study 2. **a,b,** All models had various fixed main and interaction effects (shown as one effect per row) and Participant as random effect. Rows with grey text indicate non-significant effects. Numbers (blue) correspond to numbered charts in Figures 5c and 6d for **a** and **b**, respectively.

a

#	Response	Linear Effects Model (Participant as random effect)	F	p>F	R^2
q1	CAM Truthfulness Selection (PCC)	Blur Bias Level +	55.2	<.0001	.554
		CAM Type +	87.4	<.0001	
		Blur Bias Level × CAM Type	23.9	<.0001	
		Trial Number	66.4	<.0001	
q2	CAM Truthfulness Rating	Blur Bias Level +	121.9	<.0001	.532
		CAM Type +	181.0	<.0001	
		Blur Bias Level × CAM Type	55.7	<.0001	
		Trial Number	7.1	<.0001	

b

#	Response	Linear Effects Model (Participant as random effect)	F	p>F	R^2
q1a	Labeling Confidence	Blur Bias Level +	134.4	<.0001	.487
		CAM Type +	3.5	.0301	
		Bias Level × CAM Type	1.7	*n.s*	
		Trial Number	0.9	*n.s*	
q1b	Labeling Correctness	Blur Bias Level +	44.0	<.0001	.308
		CAM Type +	3.3	.0392	
		Bias Level × CAM Type	1.6	*n.s*	
		Trial Number	0.9	*n.s*	
q2,3	CAM Truthfulness Rating	Blur Bias Level +	247.2	<.0001	.460
		CAM Type +	128.9	<.0001	
		Blur Bias Level × CAM Type	25.2	<.0001	
		Unblurred Disclosure	10.4	.0013	
		Unblurred Disclosure × Blur Bias Level	5.1	.0062	
		Unblurred Disclosure × CAM Type	3.8	.0230	
		Unblurred Disclosure × Blur Bias Level × CAM Type	3.2	.0118	
		Trial Number	0.8	*n.s*	
q5,6	CAM Helpfulness Rating	Blur Blur Bias Level +	137.6	<.0001	.391
		CAM Type +	91.7	<.0001	
		Blur Bias Level × CAM Type	23.9	<.0001	
		Unblurred Disclosure	3.1	*n.s*	
		Unblurred Disclosure × Blur Bias Level	3.4	.0326	
		Unblurred Disclosure × CAM Type	1.3	*n.s*	
		Unblurred Disclosure × Blur Bias Level × CAM Type	2.2	*n.s*	
		Trial Number	1.6	*n.s*	

SUPPLEMENTARY METHODS

Supplementary Method 1 Datasets and CNN Model Training

In simulation studies, we trained and evaluated the models on three datasets for two image tasks. (summarized in Supplementary Table 2). For Simulation Study 1, we used the Inception v3[71] CNN model pretrained on ImageNet ILSVRC-2012[72] with 1.2 million images from 1000 categories and fine-tuned on blur biased images of the ImageNette[40] dataset. ImageNette is a subset of ILSVRC-2012 with 13,395 images from 10 categories. We evaluated on ImageNette instead of ImageNet due to limited computation resources. We retrained only the layers from the last two Inception blocks of the Inception v3 model. For Simulation Studies 2 and 4, we used the same Inception v3 model pretrained on ILSVRC-2012, but fine-tuned on the NTCIR-12[42] dataset with blur bias for Simulation Study 2 and color temperature bias for Simulation Study 4. NTCIR-12 consists of 44,902 egocentric images from wearable cameras annotated as 21 daily activities. For Simulation Study 3, we used the Neural Image Captioner (NIC)[3] Inceptionv3-LSTM model pretrained on ILSVRC-2012 and fine-tuned on blur biased images from the Common Objects in Context (COCO)[43] dataset with 123,287 images and 616,435 captions. We retrained layers from the last two inception blocks of the Inception v3 part of model, including LSTM blocks. All model hyperparameters were tuned using the Adam optimizer with batch size 64 and learning rate 10^{-5}. We conducted further studies with different base CNN models, VGG16[68], ResNet50[69] and Xception[70], to extend the evaluation in Simulation Study 2 (Supplementary Fig. 8).

Supplementary Method 2 User Studies Image Selection and CAMs

For both user studies, we chose 10 images to select one instance per class label for 10 classes of ImageNette. This balanced between selecting a variety of images for better external validity, and too much workload for participants due to too many trials. CAMs were generated from specific CNN models in Simulation Study 1. At each blur level, Unbiased-CAM and Biased-CAM were generated from RegularCNN, while Debiased-CAM was generated from DebiasedCNN (mb, mt). A key objective of the user studies was to validate the results of the simulation studies regarding CAM types and image blur bias levels, hence, we selected canonical images that:

1) Had RegularCNN and DebiasedCNN predict correct labels for unblurred images, since we were not investigating the use of CAMs to debug model errors. CNN predictions on blurred images may be wrong, but we showed the CAM of the correct label.
2) Were easy to recognize when unblurred, so that users can perceive whether a CAM is representative of a recognizable image. This was validated in our pilot study.
3) Were somewhat difficult but not impossible to recognize with Weak blur, so that participants can feasibly verify image labels with some help from CAMs.
4) Were very difficult to recognize with Strong blur, such that about half of pilot participants were unable to recognize the scene, to investigate the upper limits of CAM helpfulness.

5) Had Unbiased-CAMs that were representative of their labels, to evaluate perceptions with respect to truthful CAMs. Conversely, debiasing towards untruthful CAMs is futile.
6) Had Biased-CAMs for Strong blur that were perceptibly deviated and localized irrelevant objects or pixels; otherwise, no difference between Unbiased-CAM and Biased-CAM will lead to no perceived difference between Unbiased-CAM and Debiased-CAM too.
7) Had Debiased-CAMs that were an approximate interpolation between the Unbiased-CAM and Biased-CAM of each image, to represent the intermediate CAM Faithfulness of Debiased-CAM found in the simulation studies.

These criteria were verified with participants in a pilot study and the selected images had CAM Faithfulness representative of Simulation Study 1 for Debiased-CAM, but with slightly lower CAM Faithfulness for Biased-CAM to represent worse case scenarios (Fig. 5c). CAMs were identical or different based on CAM type and Blur Bias level. Unbiased-CAMs were the same for all Blur Bias levels, and Unbiased-CAM and Biased-CAM were the same for None blur level. For other conditions, CAMs were deviated and debiased based on CAM type and Blur Bias level.

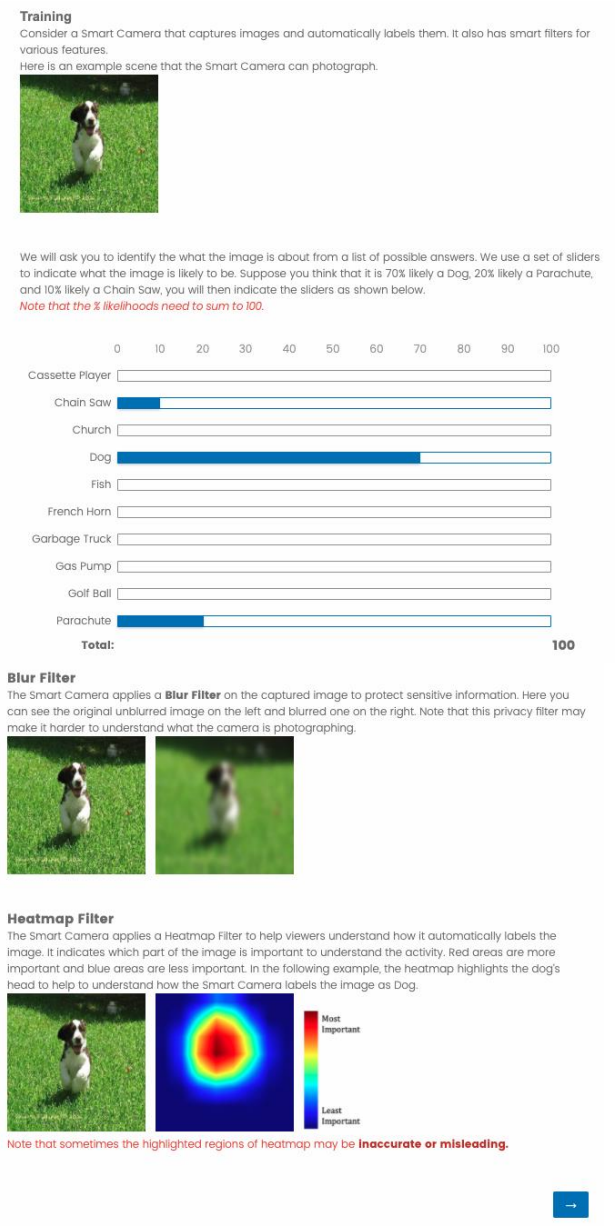
We chose not to test participants with images in NTCIR-12 due to quality and recognizability issues. Since images were automatically captured at regular time intervals, many images were transitional (e.g., pointing at ceiling while "Watching TV"), which made them unrepresentative of the label. Furthermore, in pilot testing, participants had great difficulty recognizing some scenes (e.g., "Cleaning and Chores") in images with Strong blur, such that the tasks became too confusing to test. Nevertheless, our results can generalize to wearable camera images with Weak blur, for users who are familiar with or can remember their personal recent or likely activities.

Supplementary Method 3 User Studies Participants and Exclusion Criteria

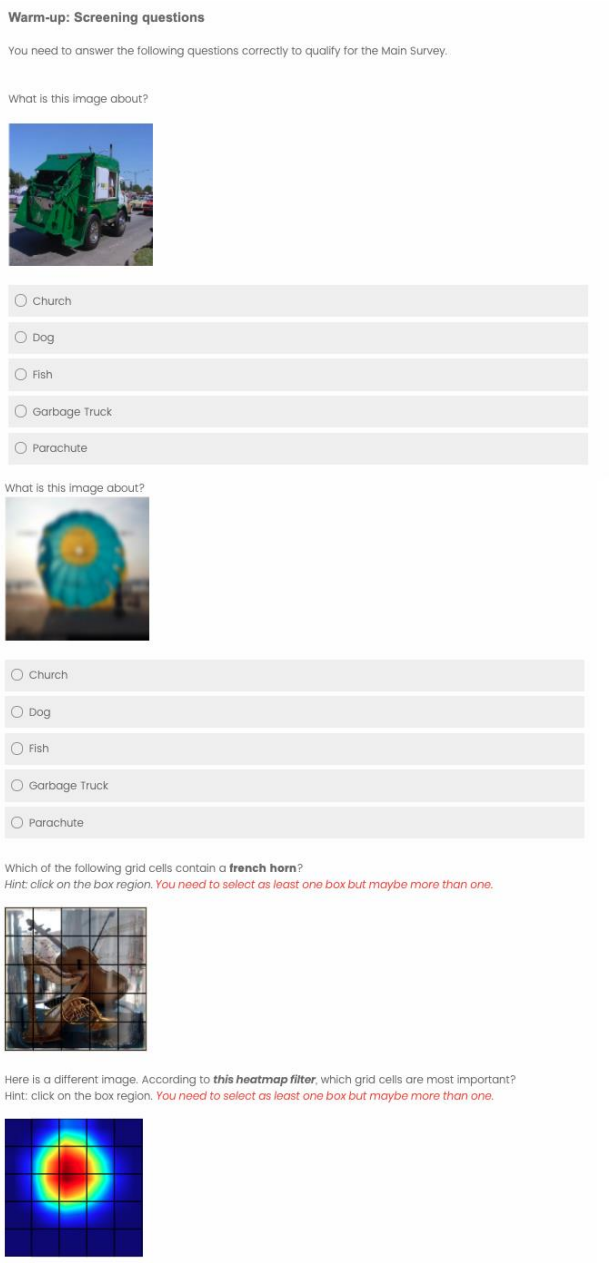
Participants from both user studies had similar demographics, so we combine their description. We recruited 32 and 171 participants from Amazon Mechanical Turk (AMT) with high qualification (≥5000 completed HITs with >97% approval rate) for CAM Truthfulness User Study 1 and CAM Helpfulness User Study 2, respectively. They were 44.9% female and between 21 and 74 years old (Median = 34). For User Study 1, 32/36 participants passed all four screening questions, continued to complete the survey in median time 15.9 minutes and were compensated US$2.00. We excluded 40/320 responses from analysis based on the exclusion criterion of taking >200 seconds to complete each page per trial. For User Study 2, 171/191 participants passed all four screening questions, participants completed the survey in median time 18.4 minutes and were compensated US$2.00. These participants were different from those recruited in User Study 1. We further excluded 7 participants who gave wrong labels for >60% of encountered unblurred images (i.e., the participant's label with the highest probability was not the actual label; in practice, only 1 mistake allowed), since this indicated poor image recognition ability for the participant. Of the remaining participants, we excluded 73/1640 responses from analysis based on the same timing criterion as in User Study 1. Note that all trials with mislabeled unblurred images also happened to be excluded due to this trial criterion.

Supplementary Method 4 User Study 1 and 2 Questionnaires

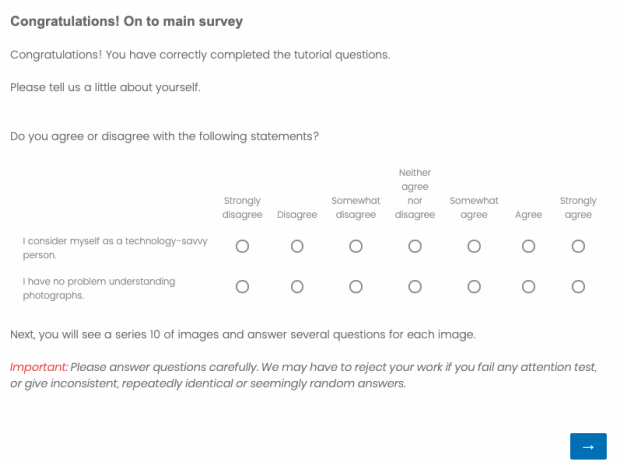
We illustrate key sections in the questionnaire for the CAM Truthfulness User Study 1 and CAM Helpfulness User Study 2. Both questionnaires were identical except for the main study section.



Supplementary Fig. 11 | Tutorial to introduce the scenario background of a smart camera with privacy blur and heatmap (CAM) explanation. It taught the participant to i) interpret the "balls and bins" question[62], ii) understand why the image may be blurred, and iii) interpret the CAM.

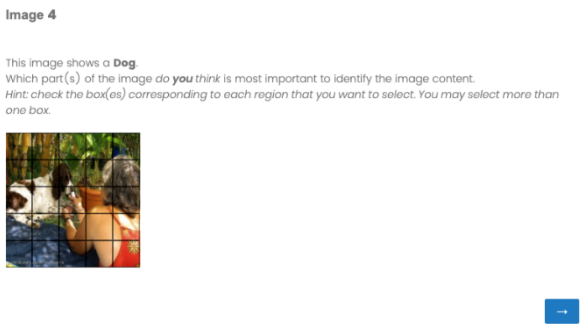


Supplementary Fig. 12 | Screening quiz with four questions to test labeling correctness and saliency selection. Questions tested for correct labeling on an unblurred (1) and a weakly blurred (2) photograph image, and correct grid selection of relevant locations in a photograph image (3) and a heatmap (4). The participant is excluded from the study if he answered more than one question wrongly.

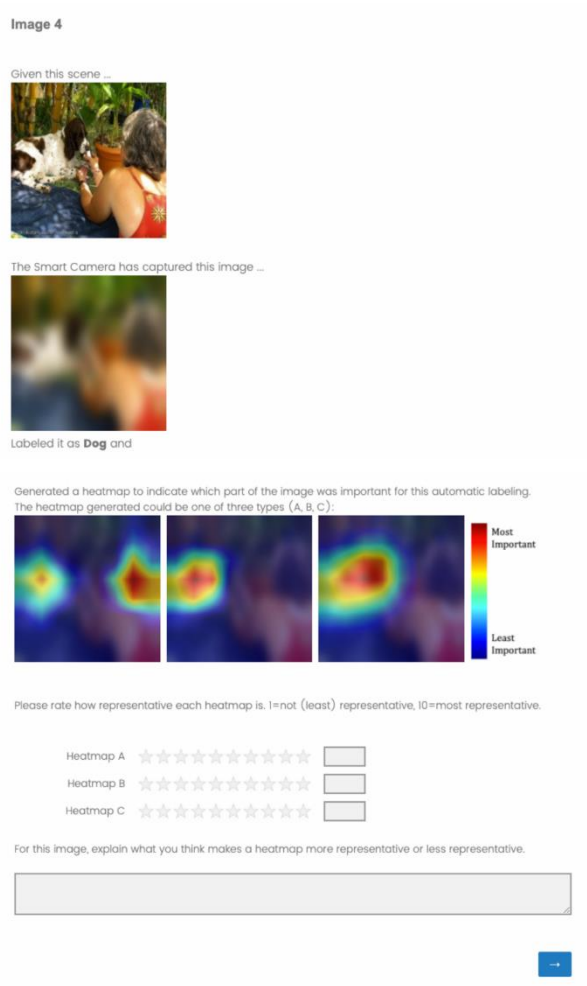


Supplementary Fig. 13 | Background questions on participant self-reported technology savviness and photograph comprehension. These questions were posed after passing the screening quiz, and before the main study section to measure the participant's pre-conceived self-assessment which may be biased after repeatedly viewing variously blurred images and variously biased heatmaps.

a

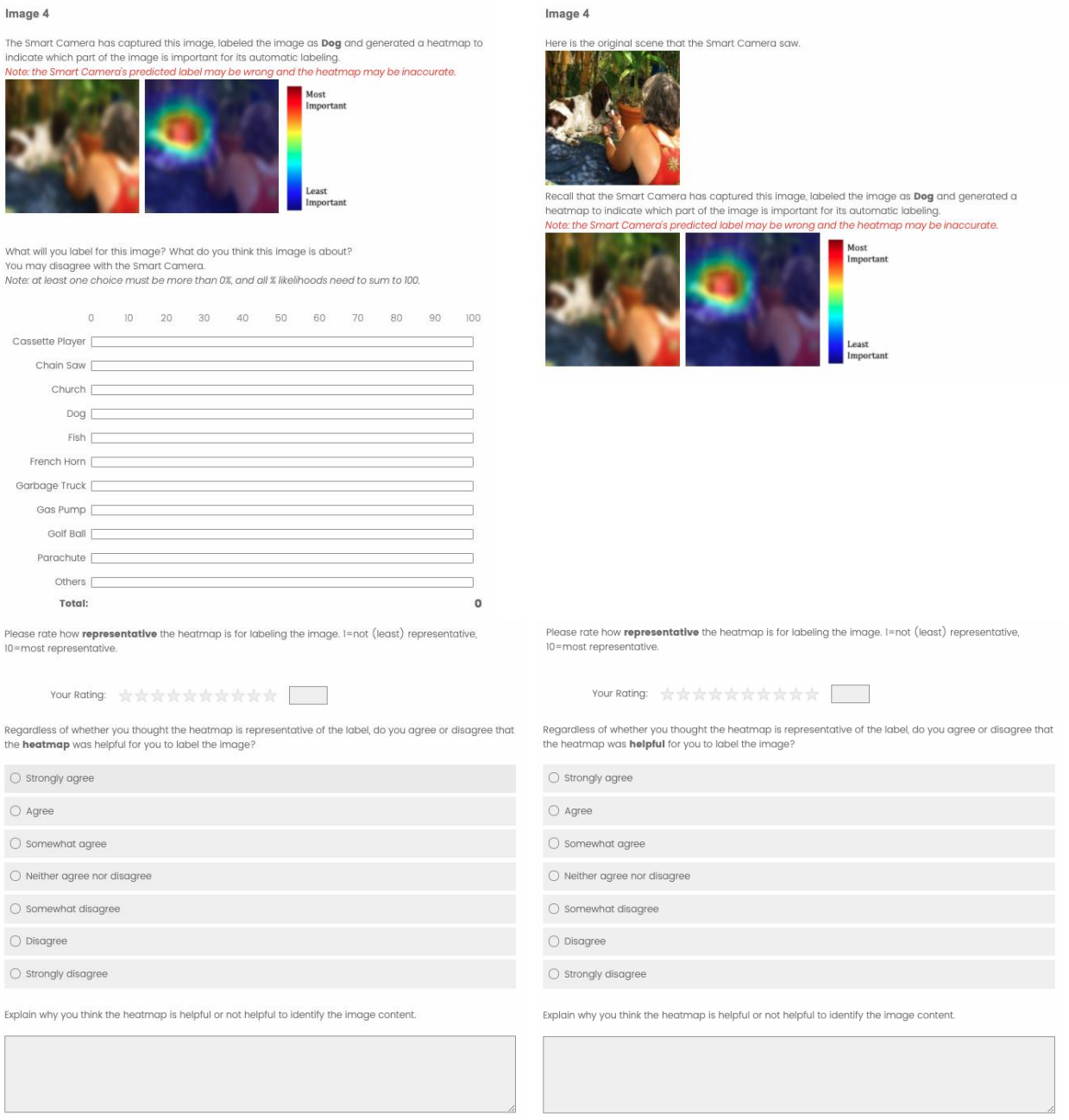


b

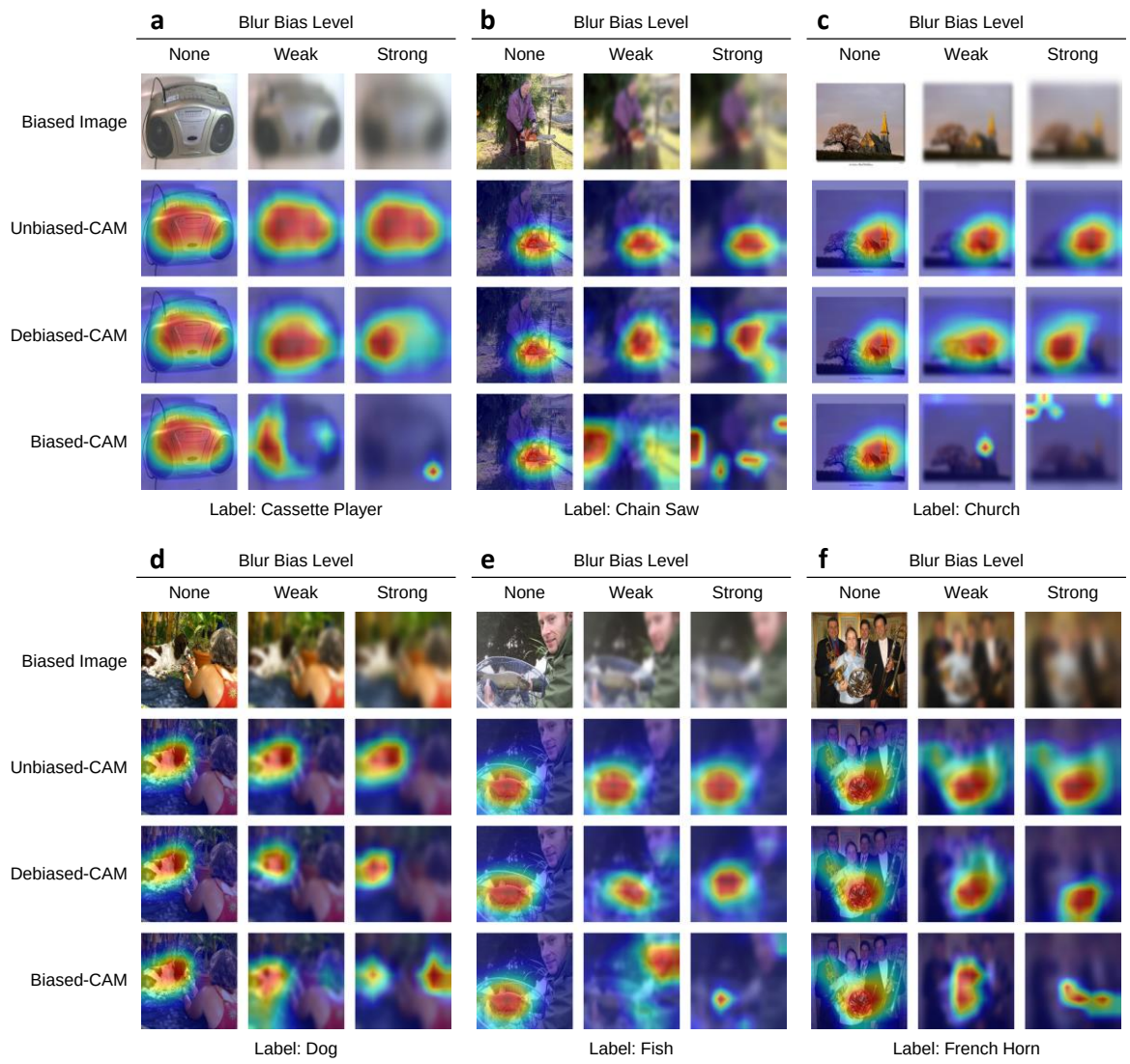


Supplementary Fig. 14 | Example main study per-Image Trial for CAM Truthfulness User Study 1. a, The first page asked the participant to q1) select on a grid which locations in an unblurred image are important to identify the image as labeled. **b,** The second page showed how the smart camera has captured the image (at a randomly selected Blur Bias level), and asked the participant to q2) rate the Truthfulness of all three CAM types (randomly arranged) along a 10-point scale and to q3) explain her rating rationale.

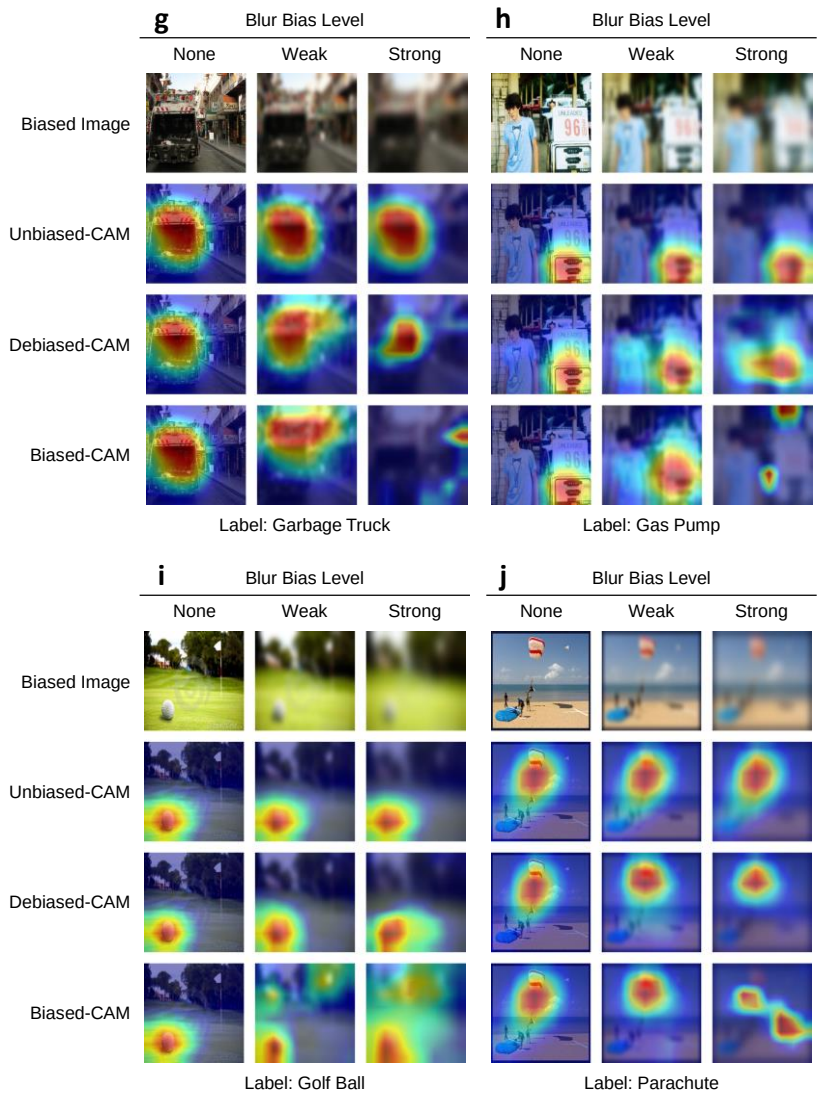
a **b**



Supplementary Fig. 15 | Example main study per-Image Trial for CAM Helpfulness User Study 2. a, The first page showed the smart camera's captured blur biased image, generated heatmap (CAM) explanation, and predicted label; and asked the participant to q1) indicate the label likelihood with a "balls and bins" question; q2) rate the CAM Truthfulness, q3) rate the CAM Helpfulness and q4) explain his rating rationale. **b,** The second page showed the image unblurred, redisplayed the blurred image and CAM and repeated the questions for q5) CAM Truthfulness rating, q6) CAM Helpfulness rating and q7) rating rationale; the repeated questions allow the comparison of ratings before (preconceived) and after (consequent) the participant knew about the ground truth image.



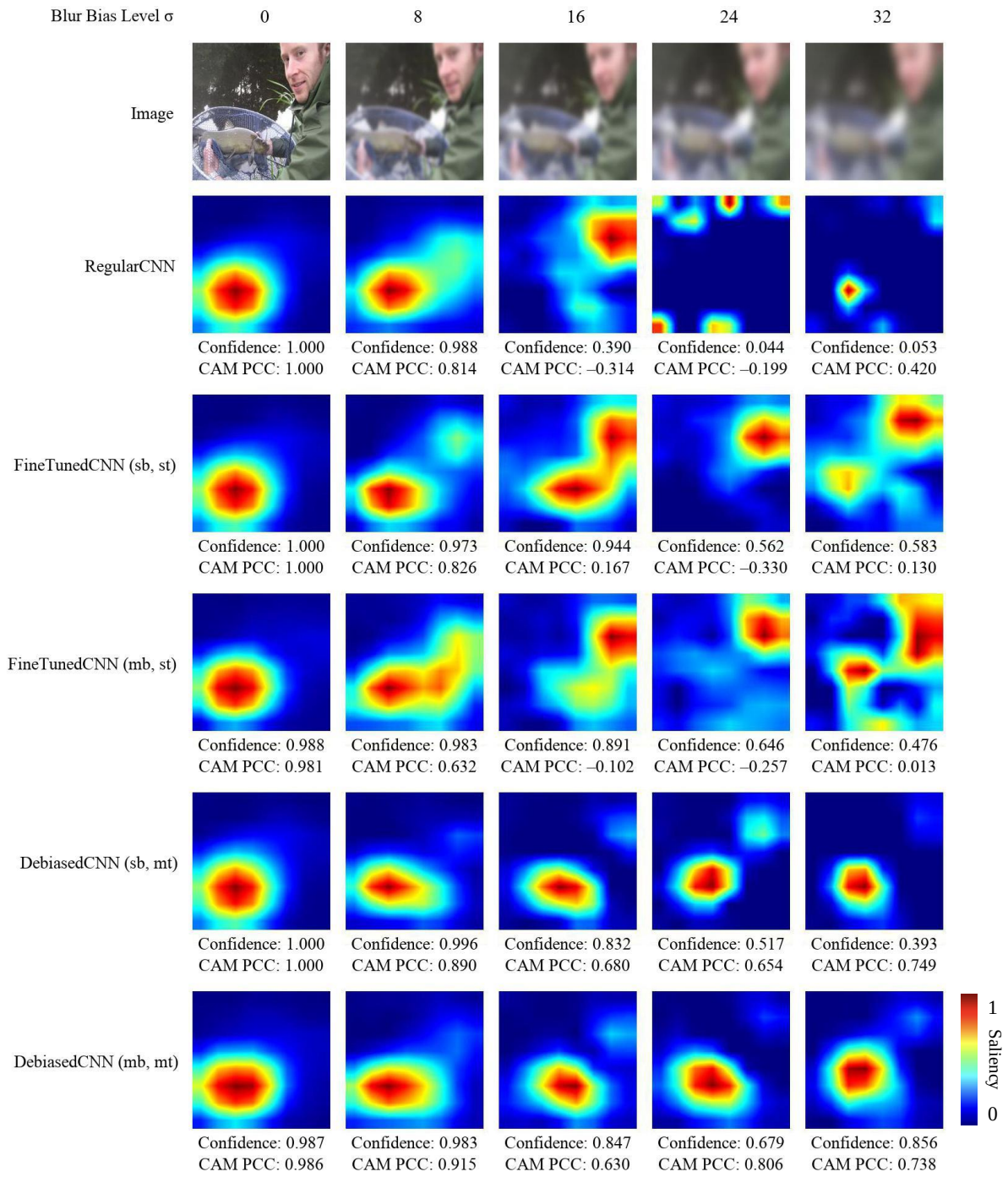
Supplementary Fig. 16 | Images and CAMs at various Blur Bias levels and CAM types that participants viewed in both User Studies.



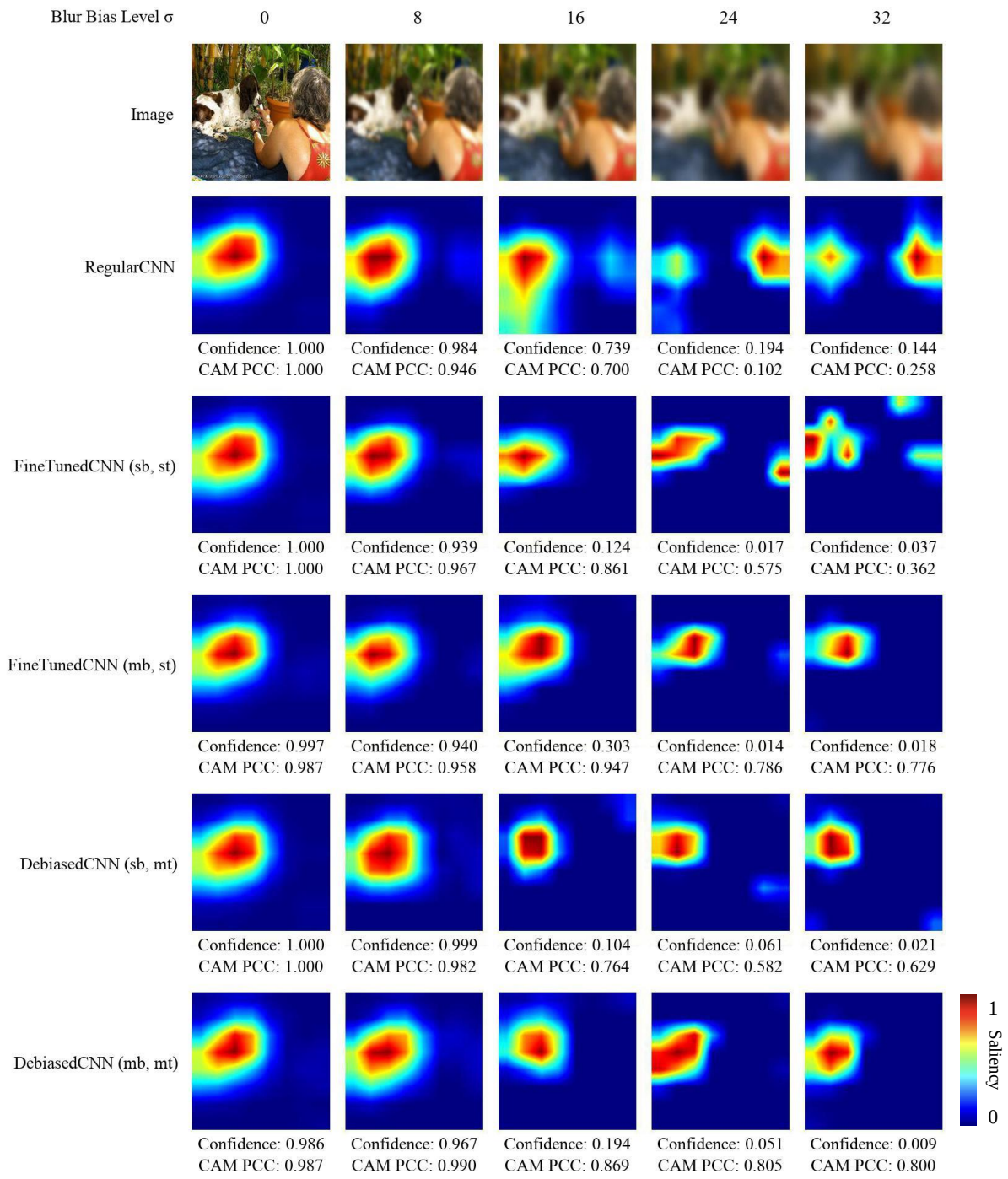
Supplementary Fig. 16 (continued) | Images and CAMs at various Blur Bias levels and CAM types that participants viewed in both User Studies.

SUPPLEMENTARY RESULTS

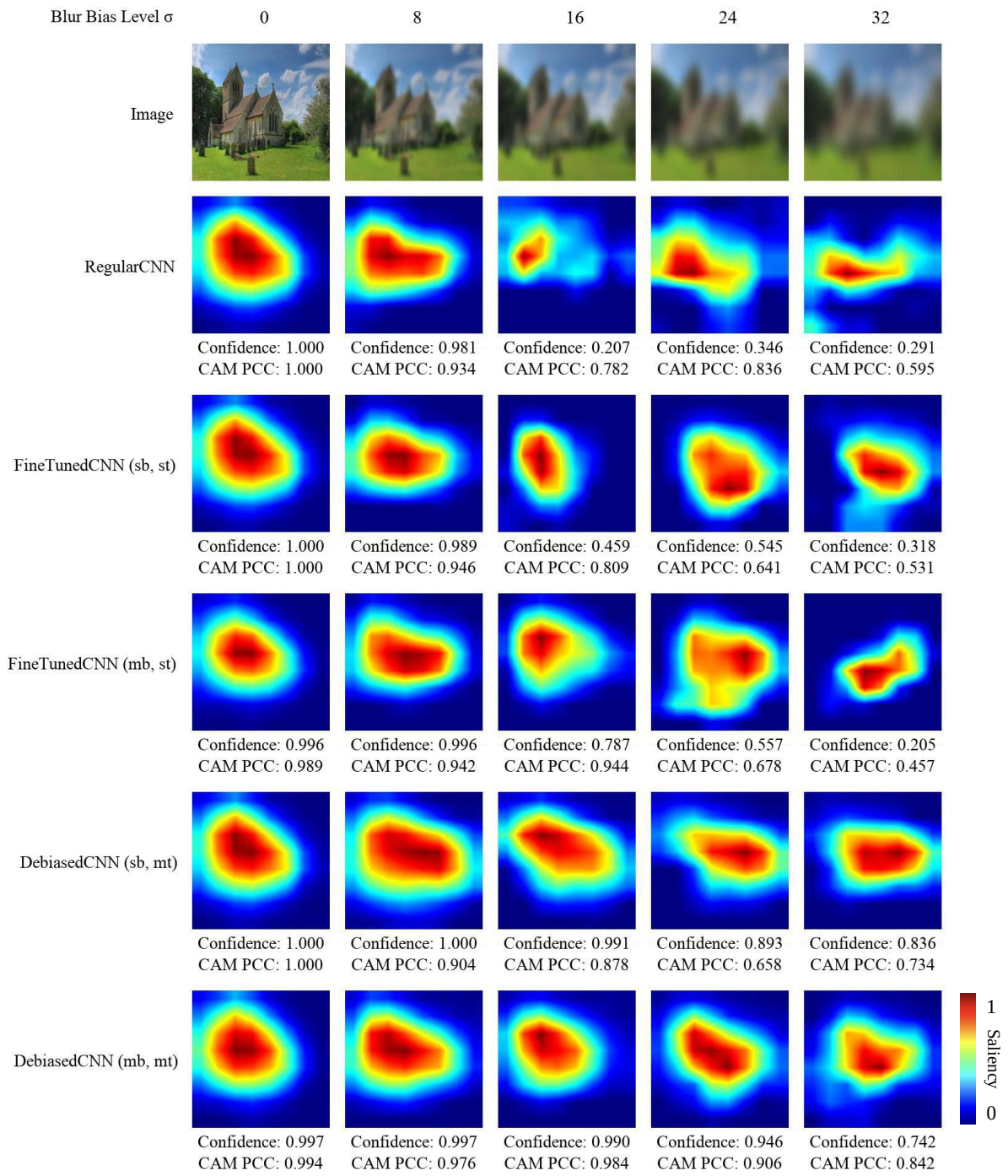
Supplementary Result 1 Example CAMs from Simulation Study 1 on image classification (object recognition) of blur-biased images



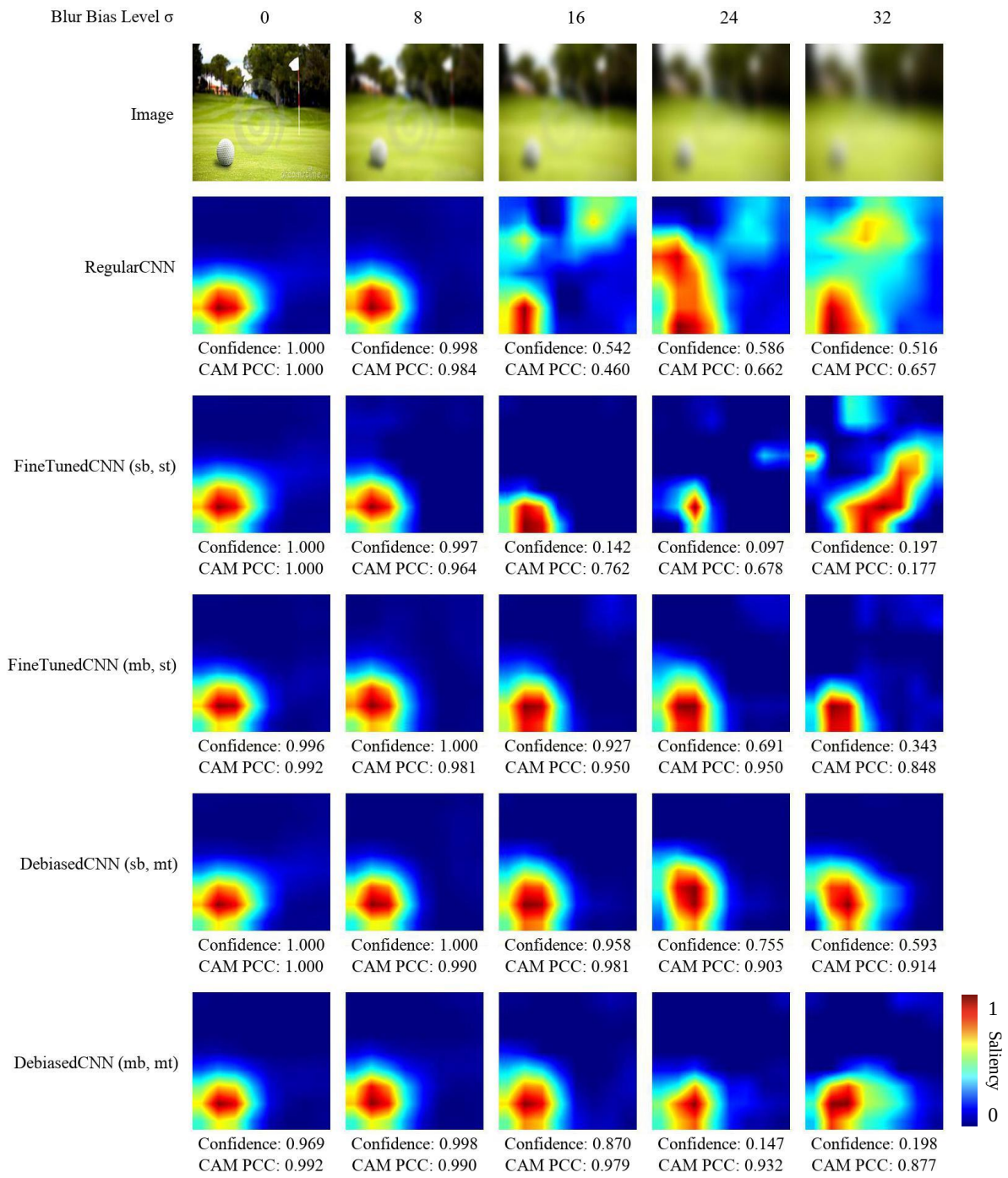
Supplementary Fig. 17 | Representative image from ImageNette labeled "Fish" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.



Supplementary Fig. 18 | Representative image from ImageNette labeled "Dog" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

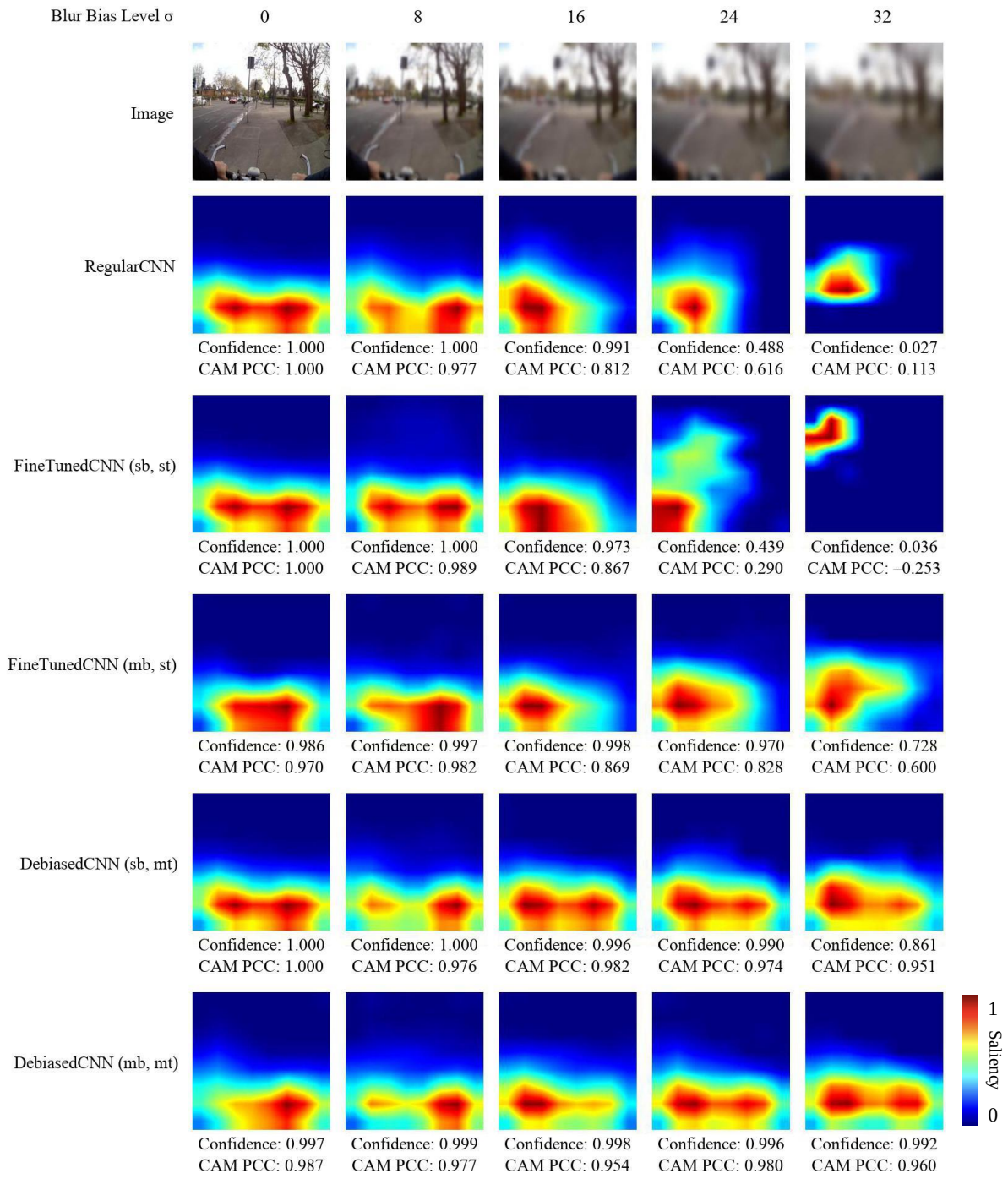


Supplementary Fig. 19 | Representative image from ImageNette labeled "Church" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

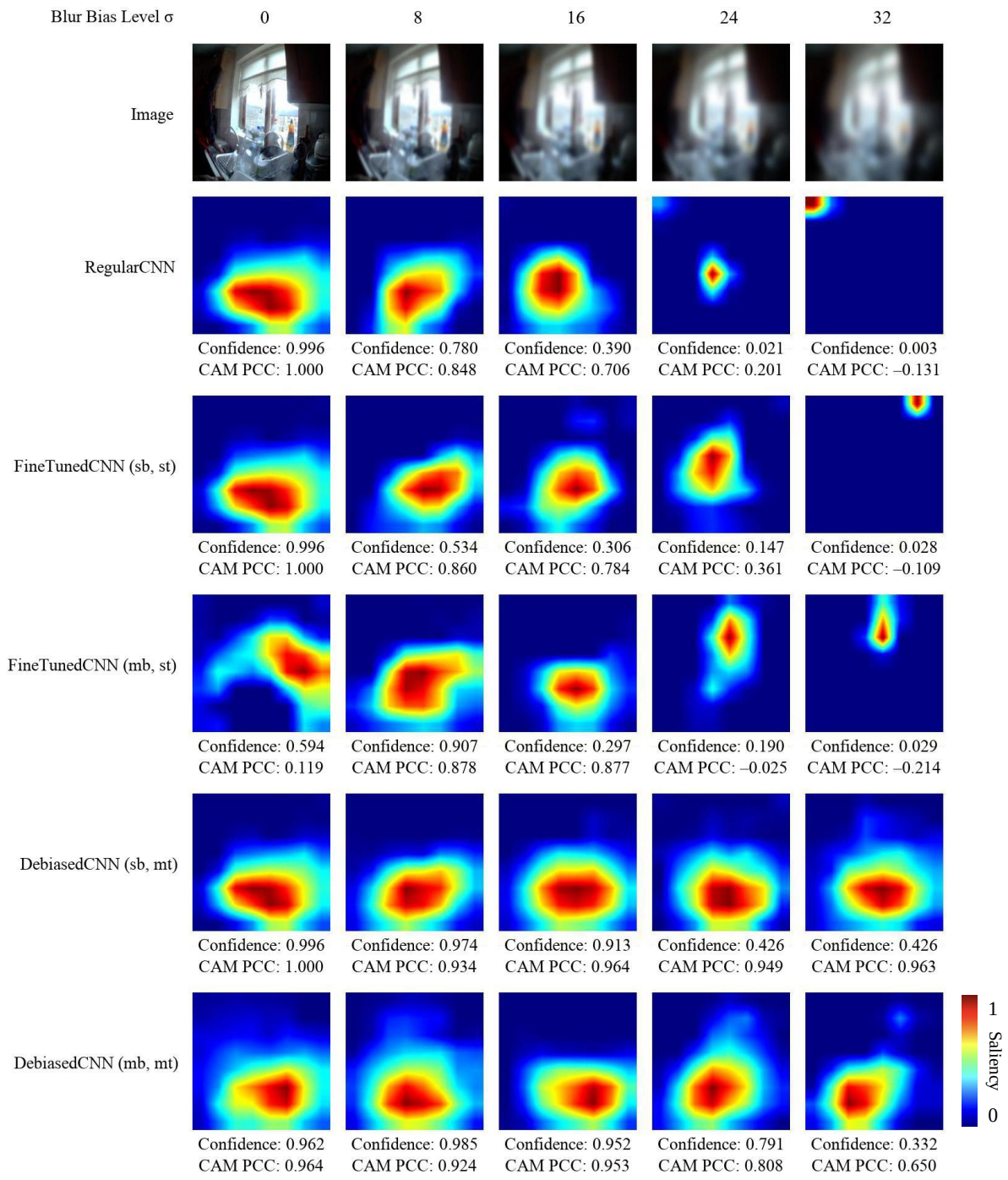


Supplementary Fig. 20 | Representative image from ImageNette labeled "Golf Ball" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

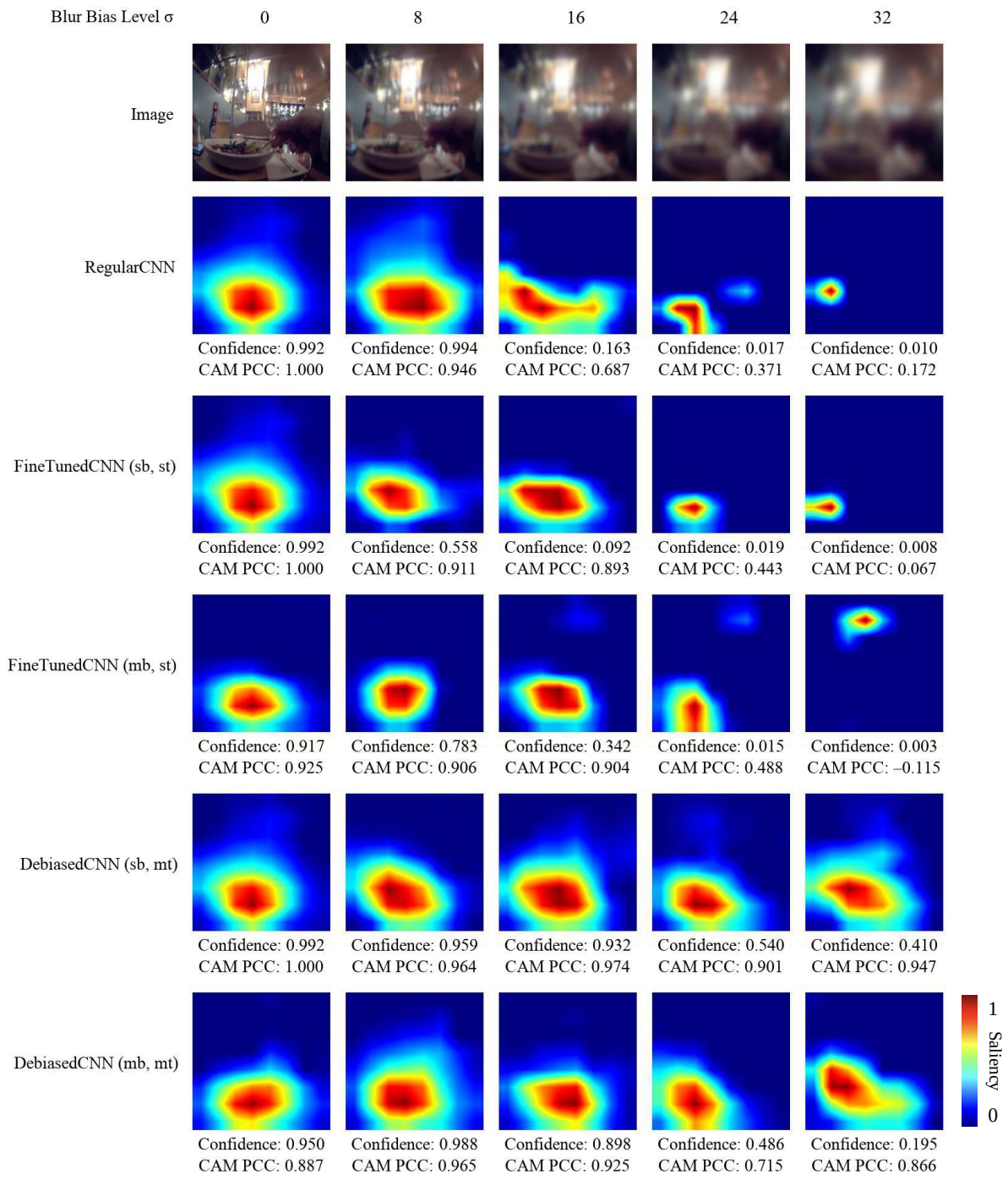
Supplementary Result 2 Example CAMs from Simulation Study 2 on image classification (wearable camera activity recognition) of blur-biased images



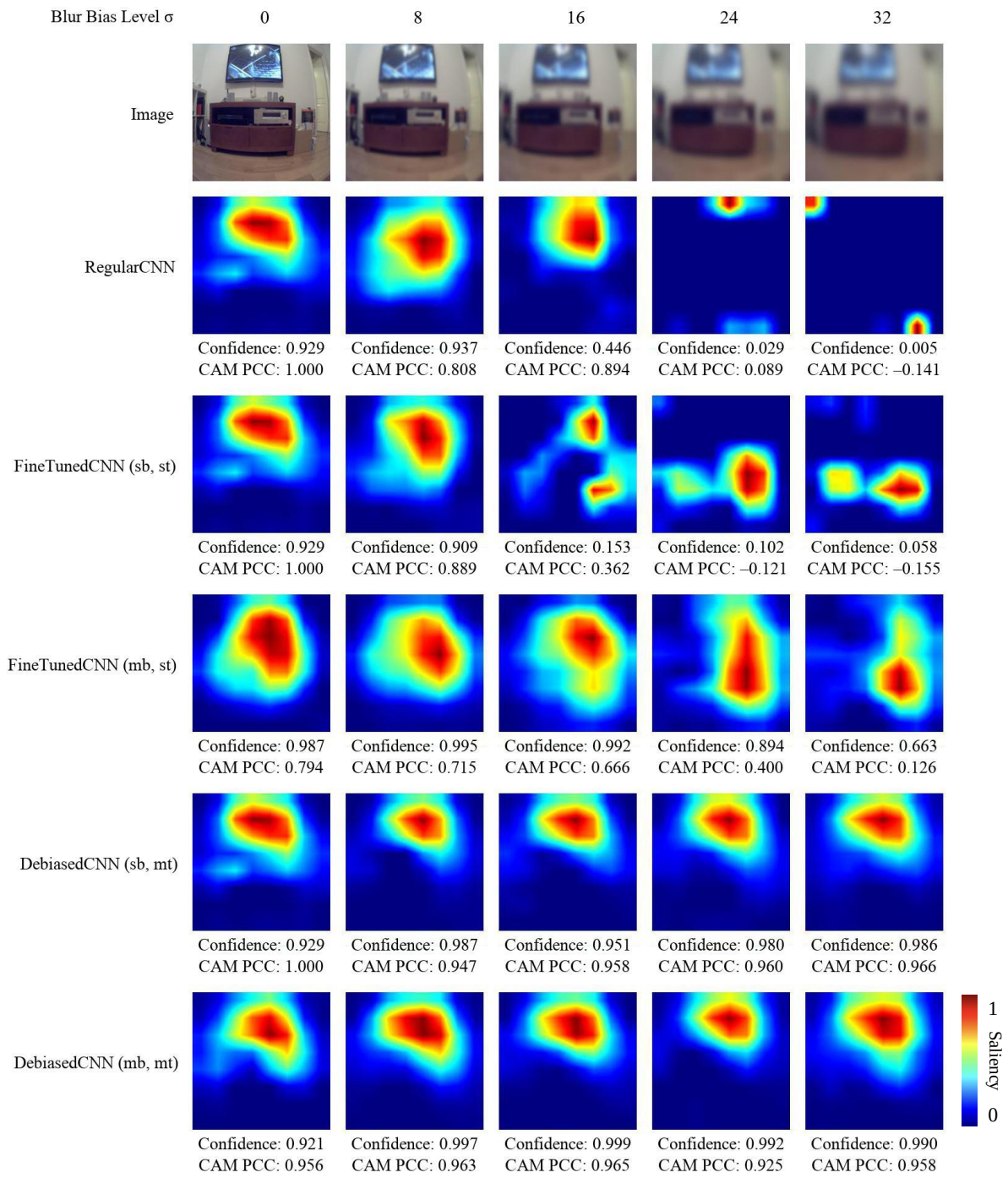
Supplementary Fig. 21 | Representative image from NTCIR-12 labeled "Biking" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.



Supplementary Fig. 22 | Representative image from NTCIR-12 labeled "Cleaning and Chores" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

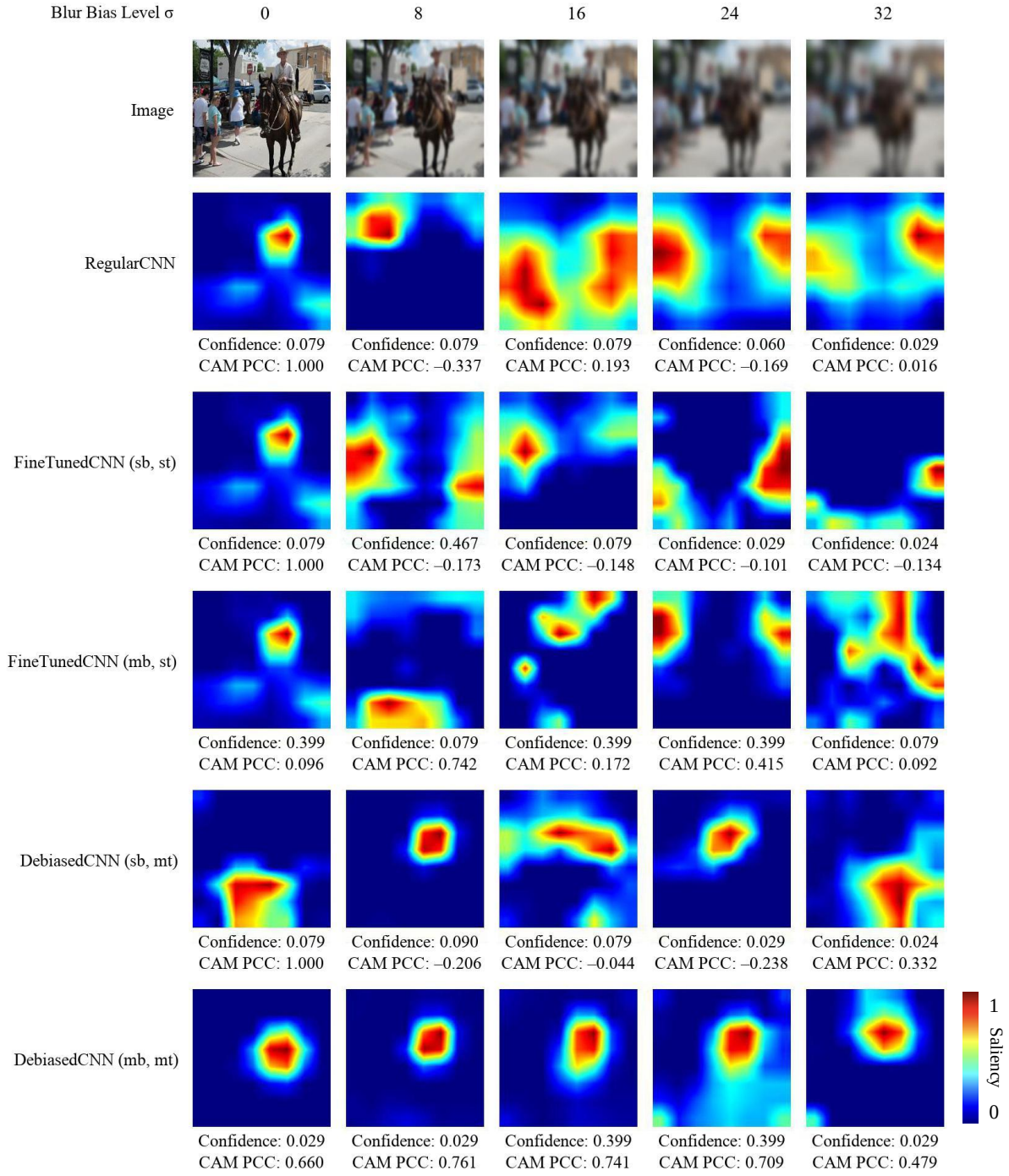


Supplementary Fig. 23 | Representative image from NTCIR-12 labeled "Drinking or Eating Alone" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

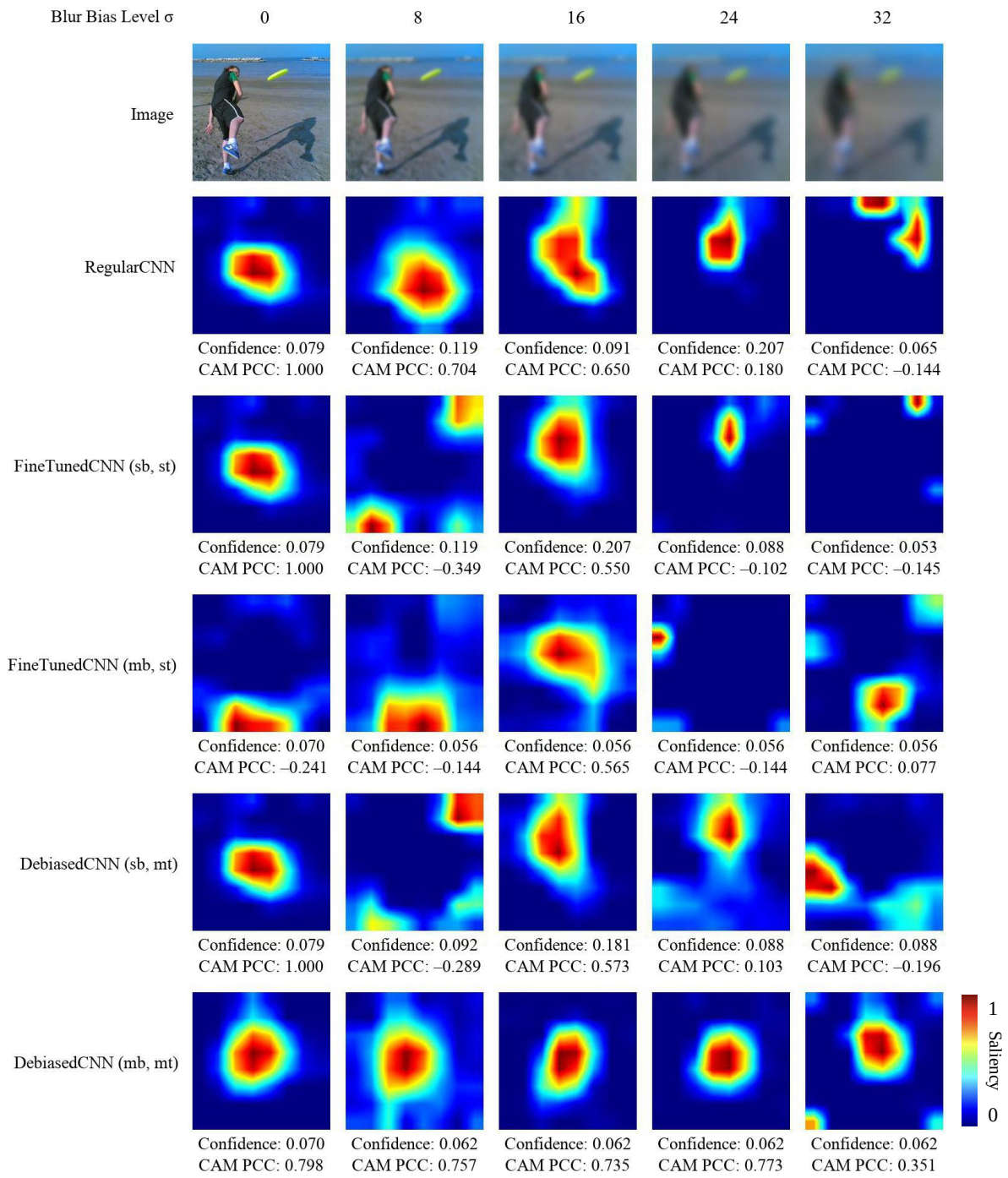


Supplementary Fig. 24 | Representative image from NTCIR-12 labeled "Watching TV" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

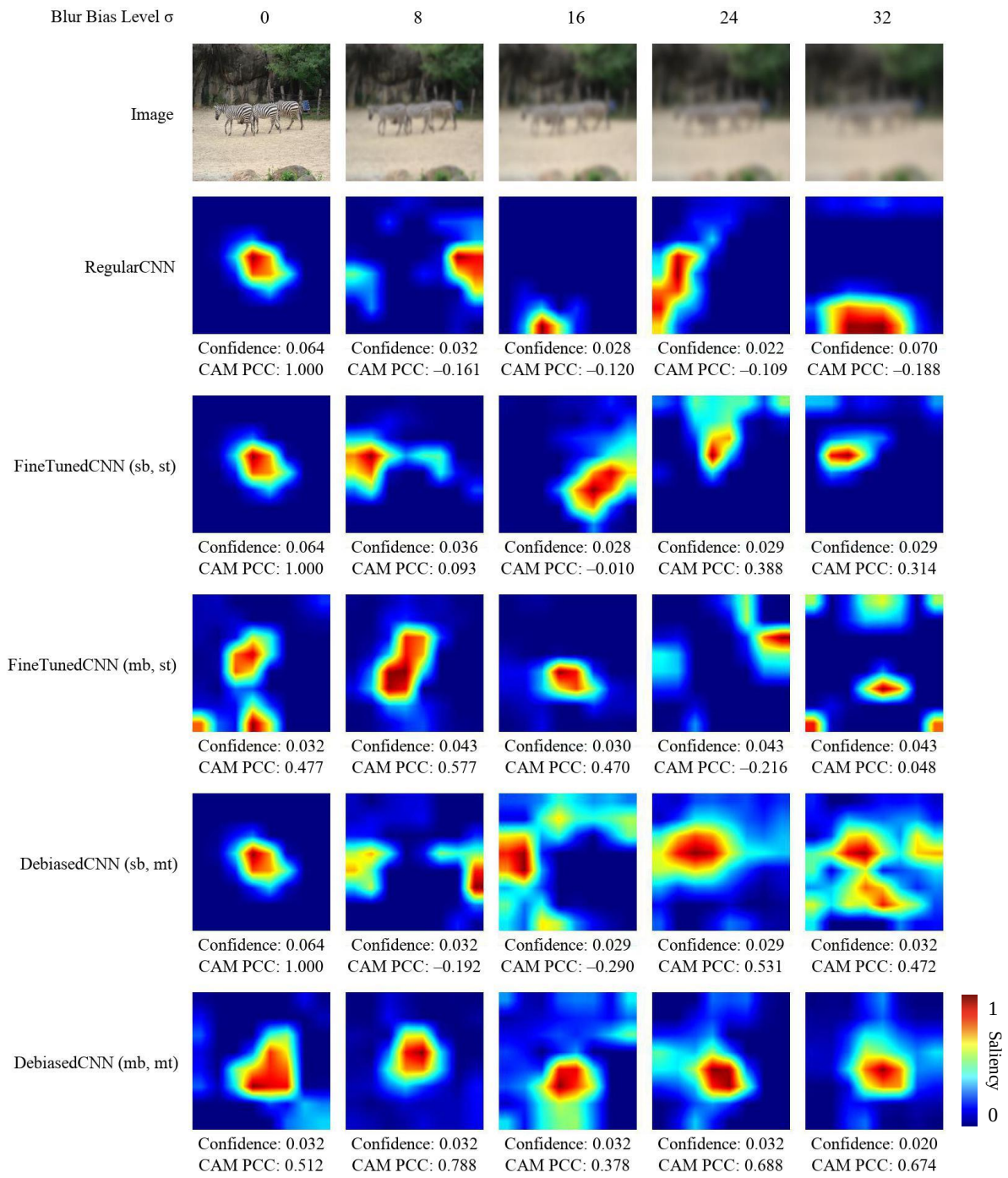
Supplementary Result 3 Example CAMs from Simulation Study 3 on image captioning of blur-biased images



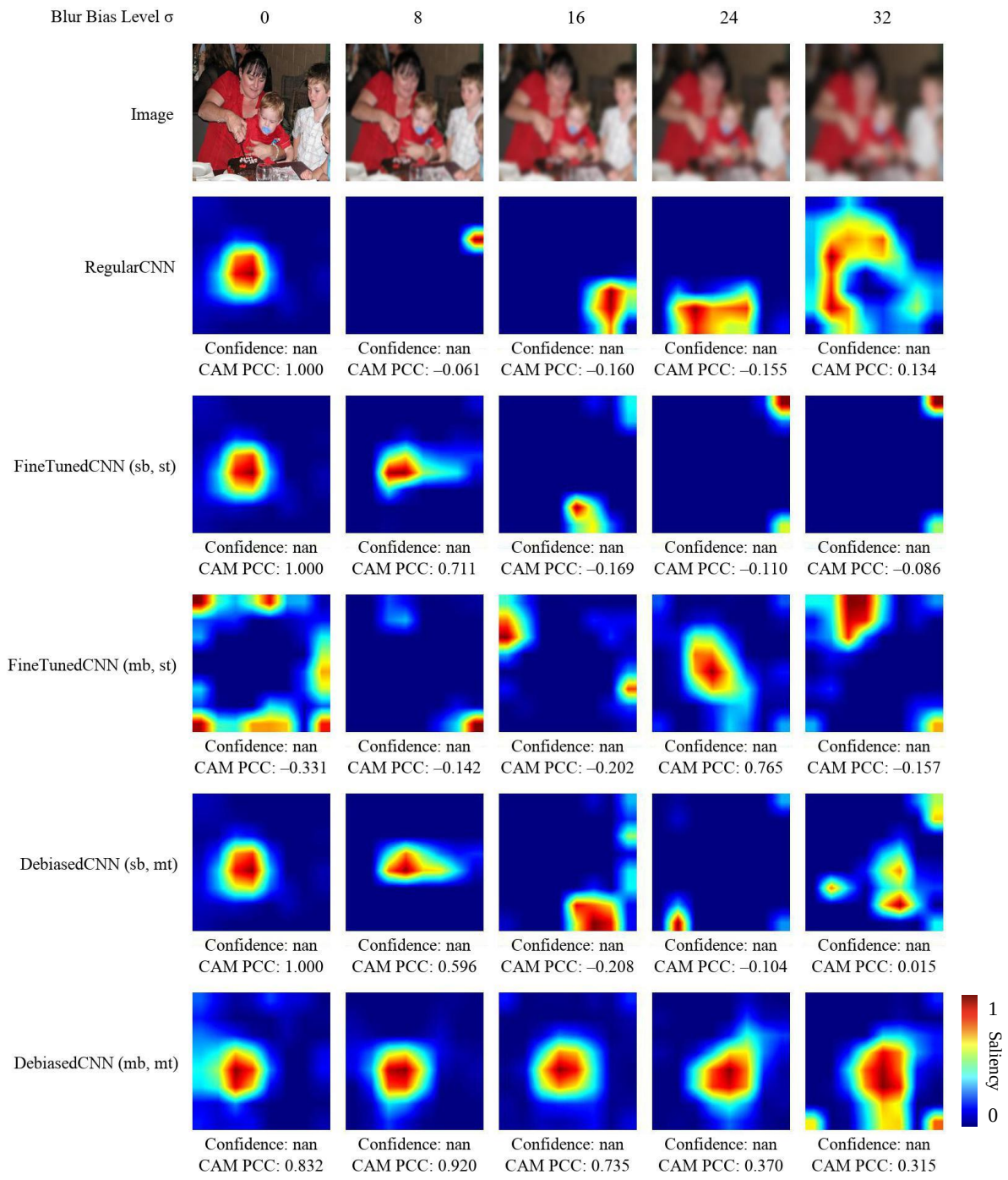
Supplementary Fig. 25 | Representative image from COCO captioned "a man on a horse on a street near people walking" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.



Supplementary Fig. 26 | Representative image from COCO captioned "a person throwing a frisbee on the sand of a beach" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

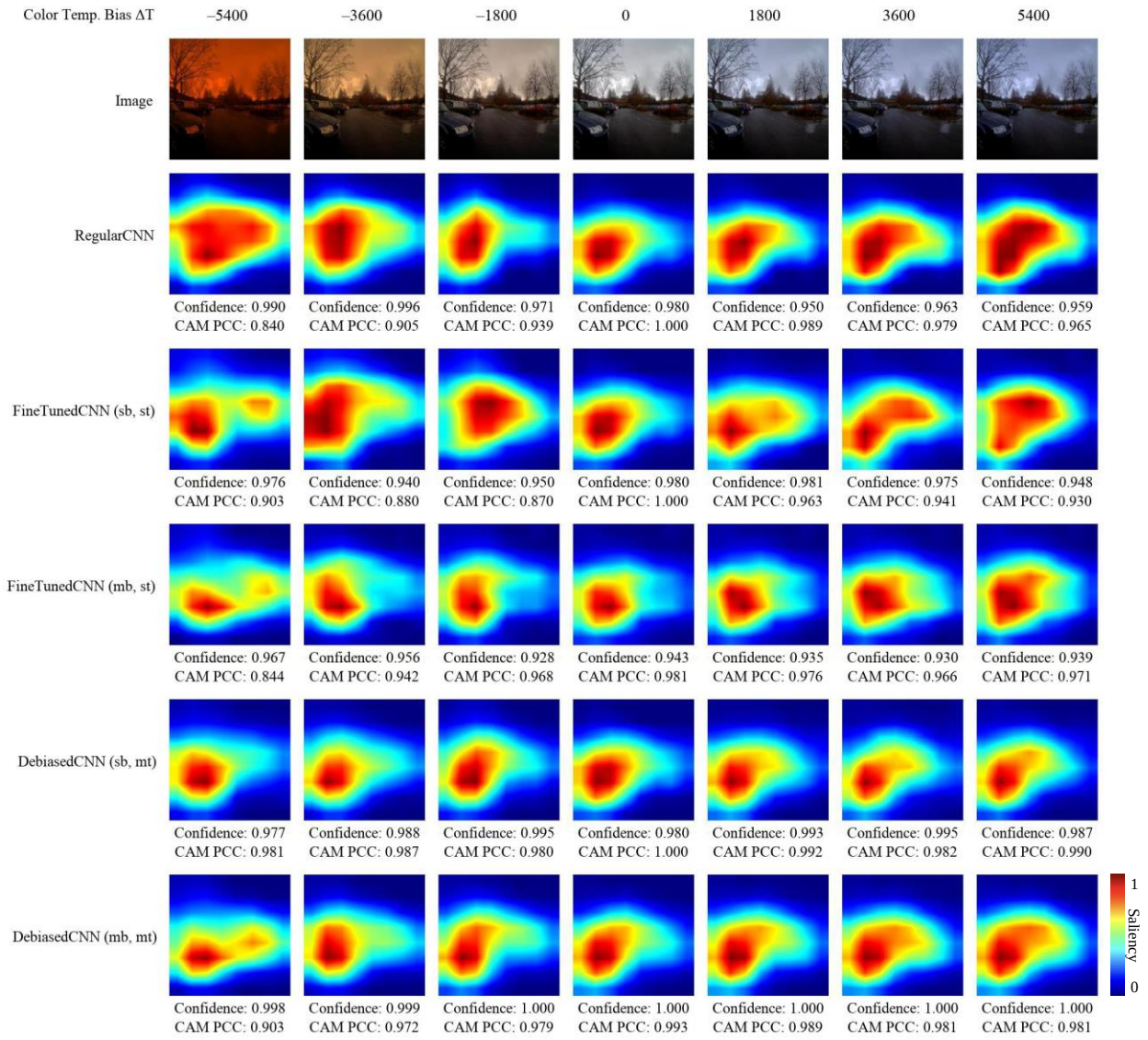


Supplementary Fig. 27 | Representative image from COCO captioned "three zebras walking in a dusty field of dirt" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

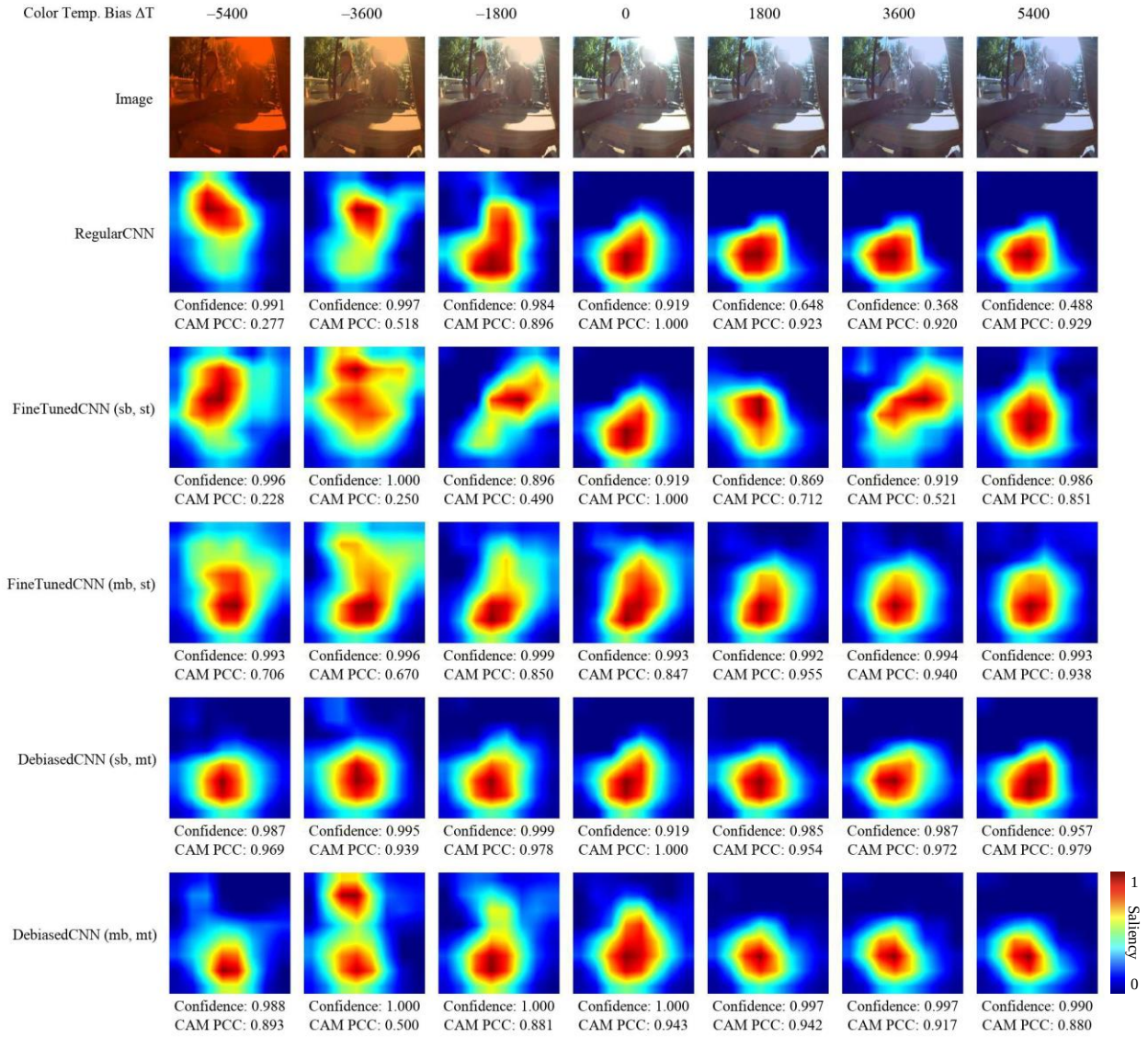


Supplementary Fig. 28 | Representative image from COCO captioned "a lady holding a child's hand cutting a cake while she also holds the child with a pacifier in his mouth" with corresponding CAMs generated by different ablated CNN models under various blur bias levels.

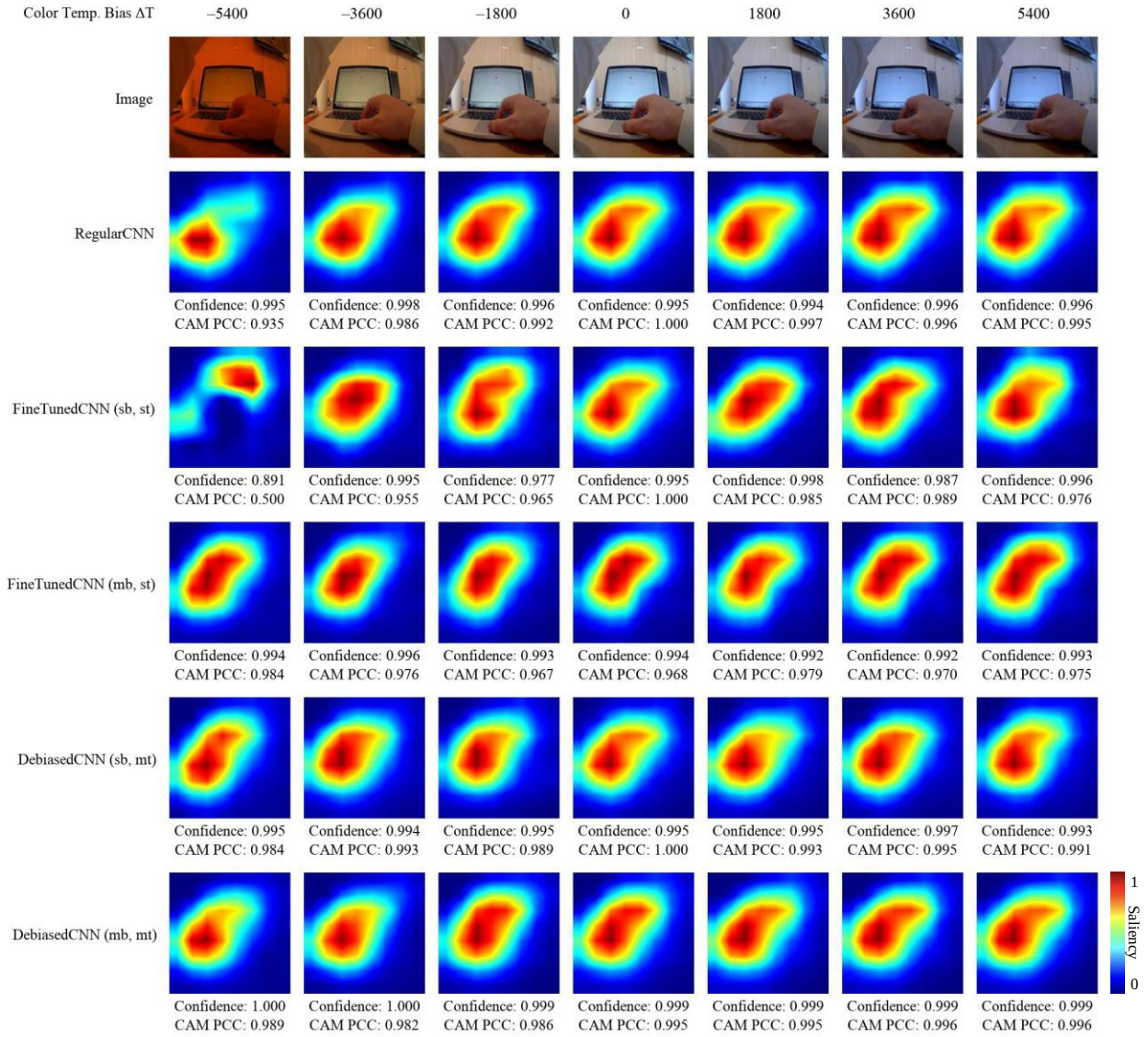
Supplementary Result 4 Example CAMs from Simulation Study 4 on image classification (wearable camera activity recognition) of color temperature-biased images



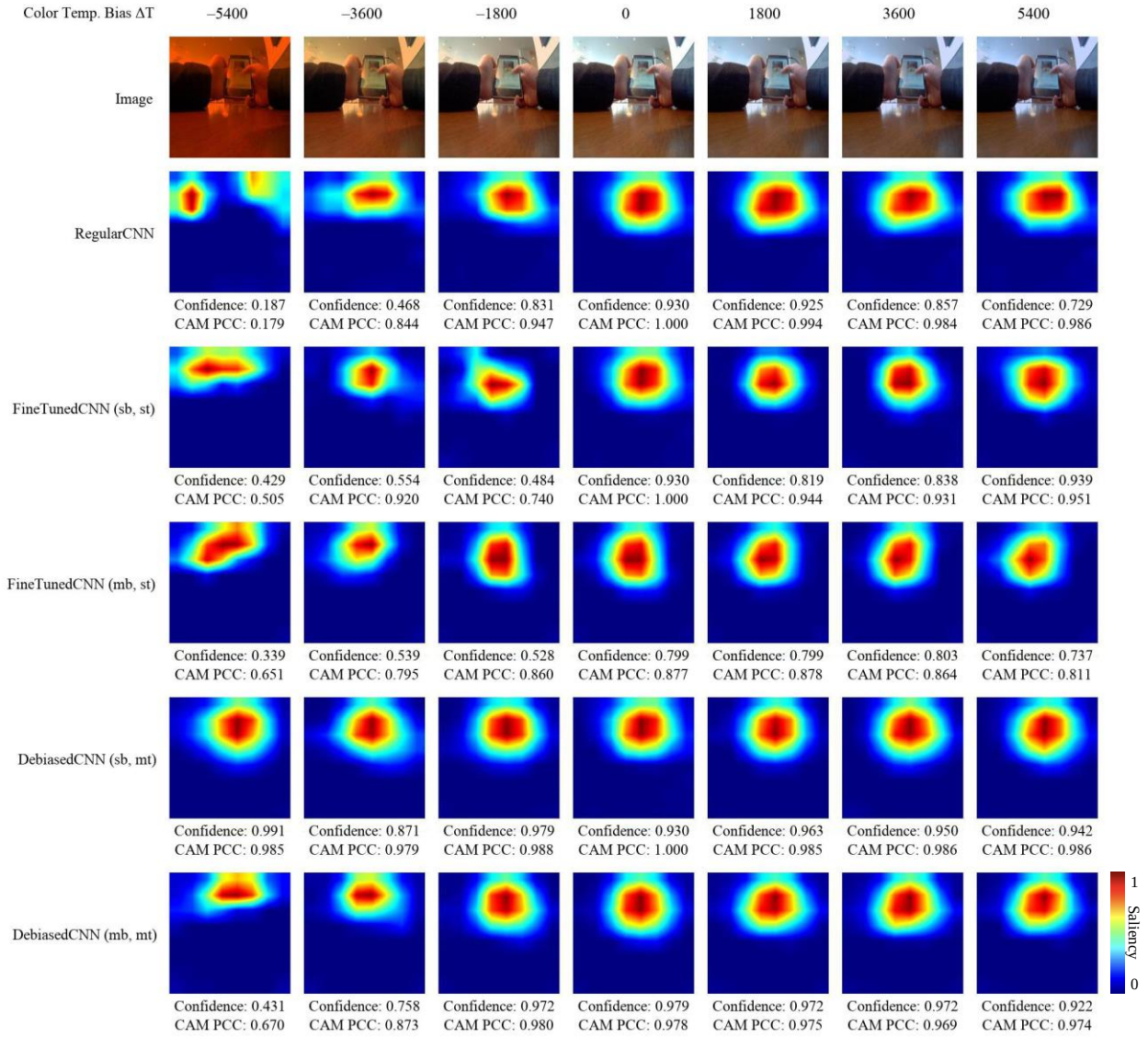
Supplementary Fig. 29 | Representative image from NTCIR-12 labeled "Walking Outdoor" with corresponding CAMs generated by different ablated CNN models under various color temperature bias levels.



Supplementary Fig. 30 | Representative image from NTCIR-12 labeled “Drinking with Others” with corresponding CAMs generated by different ablated CNN models under various color temperature bias levels.



Supplementary Fig. 31 | Representative image from NTCIR-12 labeled “Working on Computer” with corresponding CAMs generated by different ablated CNN models under various color temperature bias levels.



Supplementary Fig. 32 | Representative image from NTCIR-12 labeled "Using Mobile Phone" with corresponding CAMs generated by different ablated CNN models under various color temperature bias levels.