Deepfake Caricatures: Amplifying attention to artifacts increases deepfake detection by humans and machines

Camilo Fosco* MIT camilolu@mit.edu
Emilie Josephs* MIT ejosephs@mit.edu
Alex Andonian MIT andonian@mit.edu
Aude Oliva MIT oliva@mit.edu

Abstract

Deepfakes can fuel online misinformation. As deepfakes get harder to recognize with the naked eye, human users become more reliant on deepfake detection models to help them decide whether a video is real or fake. Currently, models yield a prediction for a video's authenticity, but do not integrate a method for alerting a human user. We introduce a framework for amplifying artifacts in deepfake videos to make them more detectable by people. We propose a novel, semi-supervised Artifact Attention module, which is trained on human responses to create attention maps that highlight video artifacts, and magnify them to create a novel visual indicator we call "Deepfake Caricatures". In a user study, we demonstrate that Caricatures greatly increase human detection, across video presentation times and user engagement levels. We also introduce a deepfake detection model that incorporates the Artifact Attention module to increase its accuracy and robustness. Overall, we demonstrate the success of a human-centered approach to designing deepfake mitigation methods.

1 Introduction

Fake or manipulated video ("deepfakes") pose a clear threat in online spaces that rely on video, from social media, to news media, to video conferencing platforms. To the human eye, these computer-generated fake videos are increasingly indistinguishable from genuine videos (35; 14). Computer vision models, however, can achieve impressive success at deepfake detection. Here, we explore how best to augment human deepfake detection with AI-assistance.

Currently, AI-assisted deepfake detection relies on using text-based prompts to tell a user that a video is a deepfake. However, recent studies indicate low rates of compliance for these text-based visual indicators: in one study, participants paired with a deepfake detection model updated their response only 24% of the time, and switched their response (from "real" to "fake", or vice versa) only 12% of the time (14). More innovative approaches have been proposed, such as showing users a heatmap of regions predicted to be manipulated (6), but this did not increase acceptance rates relative to text-based indicators. Overall, to make an impact, the development of deepfake detection models must proceed alongside the exploration of innovative and effective ways to alert human users to a video's authenticity.

We present a novel framework that provides strong classical deepfake detection, but crucially also creates a compelling *visual indicator* for fake videos by amplifying artifacts, making them more detectable to human observers. Because humans tend to be highly sensitive to distortions in faces, we hypothesize that focusing our visual indicator on amplifying artifacts is likely to yield a highly

*Equal contribution

Preprint. Under review.

detectable and compelling visual indicator. Our model, “CariNet”, identifies key artifacts in deepfakes using a novel *Artifact Attention Module*, which leverages both human supervision and machine supervision to learn what distortions are most relevant to humans. CariNet then generates **Deepfake Caricatures**, distorted versions of deepfakes, using a *Caricature Generation Module* that magnifies unnatural movements in videos, making them more visible to human users.

The main objective of this work is not only to detect fake faces and regions, but to make them more obvious to human observers. This helps humans detect fakes early, and prevents them from consuming fake information. We are motivated primarly by the realization that current methodologies for making a user aware that a video is fake rely on text labels in UIs (e.g. “This Video is Fake”, "Video modified by AI"). Previous work has shown that this is suboptimal, as these labels are easily missed and typically disregarded (6; 14). Making artifacts more detectable to human observers is key to increasing the chances that humans detect fakes. Our proposed novel deepfake caricature visual indicator almost doubles human detection performance when compared to the unaided condition, and increases it by 20.5% when compared to a text-based indicator.

We make two primary contributions: First, we generate a novel visual indicator for fake videos called Deepfake Caricatures, and show in a user study that they increase human deepfake detection accuracy by up to 40% compared to non-signalled deepfakes. Second, we develop a framework for identifying video artifacts that are relevant to humans, collect two datasets of human labels highlighting areas that are perceived as fake, and propose a competitive model that leverages this new data.

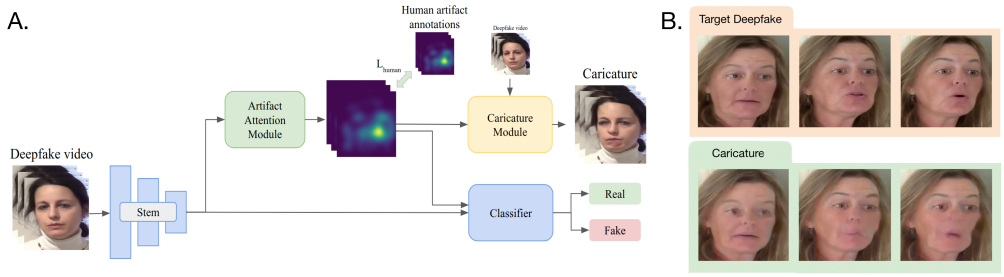
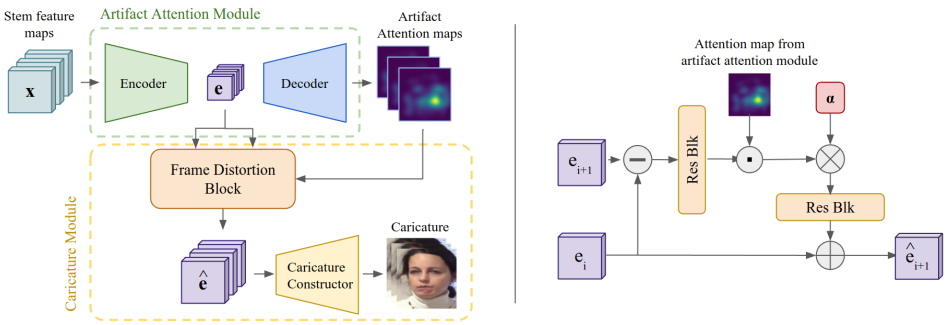


Figure 1: **A.** Overview of our framework. The model learns to identify artifacts visible to humans, then amplify them to generate Deepfake Caricatures: transformations of the original videos where artifacts are more visible. **B.** Example frames of a standard deepfake video (top) and a deepfake caricature (bottom).

2 Related work

Deepfake detection systems Deepfake detection is a young but active field. Many novel architectures have been proposed in the last few years to detect videos or images where faces have been digitally manipulated (1; 3; 34; 38; 15; 32). Some approaches detect fake faces based on warping or artifacts in the video frames (12; 26; 49; 21; 24). Others detect anomalous biological signals, such as blood volume changes (8), blinking (25), or eye color and specularity (30). Recently, models have been augmented with attention mechanisms that highlight specific parts of the input, such as face regions (nose, eyes, etc.) (44), the blending boundary (24), or regions that are likely to have been manipulated (22; 41). Overall, we build on this previous work to develop a novel network that uses attention mechanisms to detect human-relevant artifacts, and amplifies them to increase human detection.

Human face perception Humans are exceptionally sensitive to the proportions of faces. Psychology research has shown that faces are encoded in memory based on their deviations from a generic averaged face (5; 40), and that the proportions of facial features are at least as important as the particular shape of a facial component for distinguishing among faces (5). This sensitivity is leveraged in the art style known as “caricature", where distinctive features of a face are exaggerated in a way that makes them easier to recognize and remember (5; 31; 40; 46), by drawing attention to the facial regions that differ most from the norm. Inspired by these caricatures, our method makes distortions of proportions in fake videos more visible, increasing a viewer’s ability to recognize the video as fake.



(a) Artifact Attention and Caricature modules

(b) Frame Distortion Block

Figure 2: **Artifact Attention and Caricature Generation modules**. (a) Artifact attention operates with an encoder-decoder architecture to generate artifact heatmaps. These heatmaps are supervised with pre-collected human annotations. The Caricature module receives both the heatmaps and the internal codes e, distorts those codes according to the heatmaps, and generates caricatures by reconstructing the video from the distorted codes. (b) The frame distortion block computes the difference between codes e_i and e_{i+1}, re-weights it according to the artifact attention maps, and then amplifies it by a factor of α before adding it back to e_i to generate distorted code $\hat{e}_{i+1}$.

AI-assisted decision making AI decision aids are increasingly being employed for a wide variety of applications (47; 48; 45). These supplement the judgments of a human user with the output of a machine learning algorithm. Such decision aids improve the accuracy of human decisions, particularly in cases where the AI can detect signals that are complementary to those that humans can detect (45). However, the design of the communication interface between the model and human user is crucial for human acceptance of the model result (45; 10; 19; 2). We hypothesize that amplifying artifacts in deepfake videos is well-suited for improving human deepfake detection: it targets and amplifies the same information humans would use to make an unassisted judgment, in an easy-to-understand format, without adding irrelevant information.

3 Model specification

We present a framework that leverages human annotations to detect video artifacts in deepfakes, then amplifies them to create Deepfake Caricatures. Our model, dubbed *CariNet*, uses a combination of self-attention and human-guided attention maps. CariNet contains three main modules (Figure 1)

- An **Artifact Attention Module** that outputs heatmaps indicating the probable location of artifacts in each input frame.
- A **Classifier Module**, which estimates whether the video is fake. This module incorporates the output of the artifact attention module to modulate attention toward artifacts.
- A **Caricature Generation Module**, which uses the Artifact Attention maps to amplify artifacts directly in the videos, yielding *deepfake caricatures*.

3.1 Artifact Attention module

This module (Figure 2) guides the model towards regions in the videos that are most likely to contain artifacts. It consists of an encoder-decoder architecture that is partially supervised with human annotations. We incorporate human annotations for two reasons: 1) it biases the model toward the artifacts that are informative to humans, and 2) it guides the module towards regions which may not be locally connected, both of which we hypothesized would yield better caricatures.

Human-informed ground truth. We collected human labels on DFDCp and FF++ corresponding to the locations most indicative of doctoring, as perceived by crowd-sourced participants. First, we created a pool of challenging deepfake videos, as these are the most likely to yield non-trivial information about artifacts. From the DFDCp dataset (11), we selected 500 videos that are challenging for humans (see Supplement), and 500 videos that are challenging for the XceptionNet (37) detection

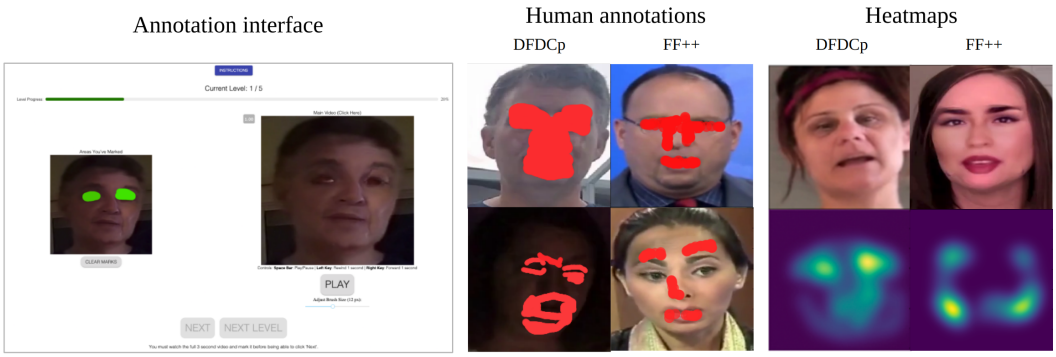
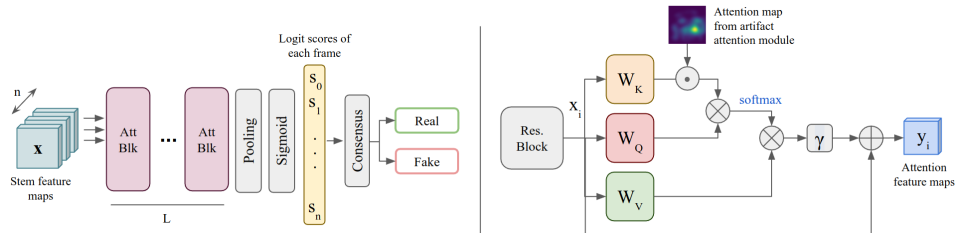


Figure 3: Annotation interface and outputs. Our annotation interface allows users to paint over zones that appear fake. Our system tracks both the position and frame at which an annotation occurs. Users highlighted both large areas and more specific, semantically meaningful areas (e.g., eyebrows).



(a) Classifier module (b) Attention Block

Figure 4: **The Classifier module and its attention blocks**. (a) the classifier takes the feature maps output by the convolutional stem, passes them through attention blocks modulated by human heatmaps, and computes logit scores for each frame before classifying the video. The consensus operation is an average of the logits followed by thresholding to determine the output label. (b) Our attention block: a traditional residual block is followed by key, query and value matrices following the self-attention framework. The key-query product is modulated by our human heatmaps. $\times$ represents matrix multiplication and $\cdot$ represents element-wise multiplication.

model. From FF++, we selected 200 videos for each of the four subsets (Deepfakes, FaceSwap, Face2Face and NeuralTextures). Next, we showed these deepfakes to a new set of participants (mean N=11 per video), who annotated areas of the videos that appeared manipulated (Figure 3). Participants were shown blocks of 100 3-second clips, and used a paint-brush interface to paint on videos regions that appeared unnatural (see Supplement). This resulted in over 11K annotations across 1000 videos for DFDC, and 9k annotations on 800 videos for FF++. For each video clip, we aggregate all annotations and generate one 3D attention map by first applying an anisotropic Gaussian kernel of size (20, 20, 6) in x, y and $time$ dimensions, and then normalizing the map of each frame to sum to one. These datasets, including attention maps and individual labels, will be released on our project page upon publication.

Module details and training. The artifact attention module is based on the encoder-decoder architectural paradigm, and consists of an Xception-based encoder (7) and a 6-block Resnet-based decoder, where upsampling and convolutions are layered between each block (Figure 2). The encoder produces codes e that retain spatial information, and the decoder utilizes those compressed feature maps to generate output heatmaps. We use the human data to supervise the generation of these heatmaps directly. This happens through the sum of three losses: the Pearson Correlation Coefficient, KL-Divergence, and L-1 loss. We confirmed that the Artifact Attention module achieves good performance at reproducing the ground truth maps: predicted maps had an average 0.745 Correlation Coefficient (CC) and a 0.452 KL divergence (KL) with the human-generated maps. In comparison, a simple 2D gaussian achieves 0.423 CC and 0.661 KL.

We trained versions of the Artifact Attention Module with both our DFDCp and FF++ artifact annotations. Qualitatively, the kinds of regions and artifacts that are identified by each Artifact

Attention Module are similar. For the modeling results (5), we report the model leveraging the FF++-trained artifact attention module to ensure fair comparisons with previous works. We additionally report the results from the model leveraging the DFDC-trained artifact attention module in the supplement for completeness.

3.2 Classifier module

Our Classifier module (Figure 4) detects if the input is real or fake. This module receives feature maps generated by an EVA-02 (13) backbone (Stem in Figure 1), a transformer-based model trained to output general visual representations through masked modeling on ImageNet. The stem's embeddings are fed to our classifier module, composed of L Human Attention Blocks followed by a global average pooling and a sigmoid activation. We define Attention Blocks as a Residual Block followed by a customized self-attention operation detailed in Figure 4b. We allow this module to supplement its self-attention with the human-guided attention information provided by the Artifact Attention Module, in order to increase attention to key parts of the input (details below). We analyzed Attention block sequence sizes of $L = 18$ and $L = 34$: each alternative yields a different model version, referred to as CariNet-S and CariNet. Cross entropy is used as the loss function. The output of the binary logit function is assigned to each frame as a detection score; we take the averaged score of the whole sequence as the prediction for the video clip.

Self-attention with artifact heatmaps. We define our self-attention layers in a similar manner to prior self-attention work (50; 4), but we extended the traditional construction to incorporate modulation from the artifact attention heatmaps. Our self attention layer computes an affinity matrix between keys and queries, where the keys are re-weighted by the artifact attention map, putting more weight on the artifacts that are relevant to humans. Given a feature map $\mathbf{x_i}$ over one frame, and an artifact attention map A over that frame, the module learns to generate an affinity matrix $\mathbf{a_i}$:

$$\mathbf{a_i} = softmax((\mathbf{W_Q x_i})^{\mathbf{T}}(\mathbf{W_K x_i} \odot \mathbf{A})). \tag{1}$$

The softmaxed key-query affinity tensor is then matrix-multiplied with the values $V = W_V x_i$ to generate the output residual r. That residual is then scaled by γ and added to input x_i to yield the output feature map y_i:

$$\mathbf{y_i} = \gamma \mathbf{a_i^T}(\mathbf{W_V x_i}) + \mathbf{x_i}. \tag{2}$$

$\mathbf{W_Q}, \mathbf{W_K}, \mathbf{W_V}$ are learned weight matrices of shape $\mathbf{R}^{\bar{C}\times C}$, $\mathbf{R}^{\bar{C}\times C}$ and $\mathbf{R}^{C\times C}$ respectively, with $\bar{C} = C/4$. γ is a learnable scalar which controls the impact of the learned attention feature vis-a-vis the original feature maps x_i.

3.3 Caricature Generation Module

Finally, the primary contribution of our framework is a novel module for creating Deepfake Caricatures, a visual indicator of video authenticity based on amplification of video artifacts. Figure 1 illustrates the distortion exhibited by our caricatures, but the effect is most compelling when viewed as a video (links to a gallery of caricatures can be found in the Supplement). Our Caricature Generation module leverages the encoder from the artifact attention module, distorts codes with a frame distortion block that operates over every pair of codes e_i and e_{i+1}, and uses a Caricature Constructor to create the final caricature (Figure 2). We instantiate the Caricature Constructor as a simple decoder with four blocks composed of 3x3 convolutions followed by nearest neighbor upsampling.

The caricature effect is achieved by amplifying the difference between the representations of consecutive frames, while guiding this amplification with the artifact attention maps generated by the artifact attention module, via element-wise multiplication between the tensor $e_{diff} = ResBlk(e_{i+1} - e_i)$ and the artifact attention map of frame x_i (Figure 2b). This essentially results in a targeted distortion aiming at magnifying the artifacts highlighted by the heatmap.

The distorted code $\hat{e_i}$ of frame x_i is computed as

$$\hat{\mathbf{e}}_{\mathbf{i+1}} = \mathbf{e_i} + \alpha(\mathbf{e_i} - \mathbf{e_{i+1}}) \odot \mathbf{A}, \tag{3}$$

where α is a user-defined distortion factor that controls the strength of the resulting caricature.

3.4 Learning and Optimization

Training is done in two phases: we first train the classification pipeline (Stem, Artifact Attention Module and Classifier), then freeze the classification pipeline and train the caricature module on a motion magnification task, before combining everything together to generate caricatures.

Classification pipeline. Our CariNets were separately trained on the DeepFake Detection Challenge (DFDCp) dataset, FaceForensics++, CelebDFv2 and DeeperForensics. For all datasets, we train on videos from the training set and evaluate on the validation or test set. We randomly sampled 32 frames from videos during training. Fake videos are over-represented in these datasets, so we oversampled real videos during training to achieve real/fake balance. Our CariNets were optimized with Rectified Adam (28) with the LookAhead optimizer (51). We use a batch size of 32 and an initial learning rate of 0.001. Cosine annealing learning rate scheduling was applied with half period of 100 epochs. We chose an early stopping strategy, stopping if validation accuracy stagnates in a 10 epoch window. We apply flipping (probability 0.5) and random cropping (224x224) augmentations. The full loss corresponds to the sum of the loss from the Classifier and Artifact Attention modules (described above).

Caricature Module. This module was trained following the motion magnification framework from (36). We use their synthetic dataset, composed of triplets of frames $(x_i, x_{i+1}, \hat{y}_{i+1})$ constructed to simulate motion magnification. x_i and x_{i+1} correspond to video frames at index i and $i+1$ (respectively), and $\hat{y}_{i+1}$ corresponds to frame x_{i+1} with artificially magnified object displacements. During training, we compute encodings e_i and e_{i+1} with our frozen classification pipeline, and feed each pair to a Frame Distortion Block (2) that learns to generate $\hat{e}_{i+1}$, a distorted version of e_{i+1} where displacements are magnified. The caricature constructor then reconstructs an estimated magnified frame $\bar{y}_{i+1}$, which is compared to the ground truth magnified frame $\hat{y}_{i+1}$ using a L-1 loss (no advantage found for more advanced losses). During this training period with the synthetic dataset, no attention maps are fed to the frame distortion block. After training, when generating a caricature, attention maps are used as a multiplicative modulation affecting the feature maps of the frame distortion block (as shown in Figure 2), which effectively turns magnifications on and off for different parts of the frame. This allows our module to function as a sort of *magnifying glass* over artifacts.

4 Results: Human Experiments

Here, we introduce Deepfake Caricatures, a novel visual signal that a video is a deepfake based on amplified video artifacts. We demonstrate the effectiveness of Caricatures on user behavior in two ways: first, we compare Caricatures to text-based visual indicators; second, we establish the range of conditions where Caricatures successfully boost deepfake detection. See supplement for detailed methods and full statistical reporting.

Comparison of Caricatures and Text-Based Indicators: Currently, when news outlets share videos that are known to be deepfakes, they flag the videos using text. However, previous work has shown that users who saw deepfakes flagged using text often continue to believe that the videos were real (6; 14). We tested whether users found Caricatures more convincing than text-based indicators. Fifty challenging deepfakes were selected from the DFDC (11), along with 50 real videos, and presented to users in a deepfake detection task. Participants (N=30 per condition) were randomly assigned to 1 of 3 conditions: unaided deepfake detection, detection where deepfakes were flagged using text, and detection where deepfakes were flagged using Caricatures. Average hit rates (HR) rates were measures for each condition. Overall, text-based visual indicators improved deepfake detection relative to unaided detection (HR=0.78 and HR=0.53 respectively), although users' performance remained well below ceiling. Crucially, users in the Caricature conditions achieved **hit rates of 0.94**, substantially higher than text-based and unaided detection ($p < 0.0001$ for both). Overall, our results show that vision-based indicators such as Caricatures are much more effective than text-based indicators at changing user behavior.

Evaluation of Caricatures across conditions: We next tested whether Caricatures robustly improved performance under a variety of visual conditions (Figure 5). First, we tested how long Caricatures needed to be visible to confer a benefit to users. A random sample of 400 videos from the DFDCp ((11), 200 real, 200 fake) were presented to participants, for 6 different durations: 300ms, 500ms, 1000ms, 3000ms, and 5000ms. The proportion of times each deepfake video was detected in a

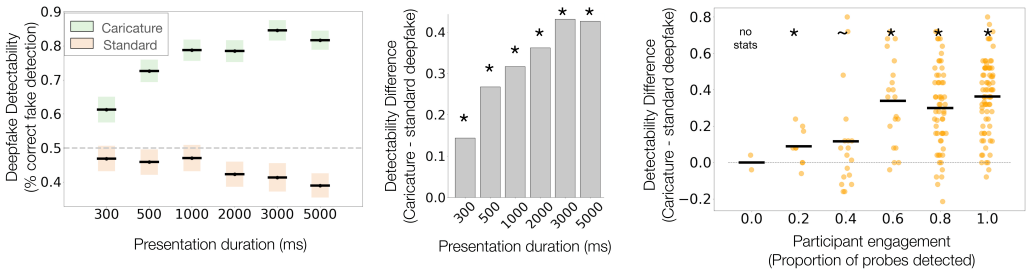


Figure 5: **Behavioral results showing the effect of caricatures on human deepfake detection**. Top Panel: Deepfake detectability by humans for standard (orange) and caricature (green) conditions. Colored boxes represent 95% confidence intervals, and stars indicate that the difference between conditions is significant. Bottom Panel: Improvement in deepfake detection with Caricatures, binned by participant engagement level.

sample of 10 participants was calculated. Averaged over all timepoints, deepfake detectability was substantially higher for Caricatures than standard deepfakes, ($F(2, 1630) = 17.12$, $p < 0.001$). Crucially, the advantage for Caricatures over standard deepfakes was significant at every presentation duration ($p < 0.01$ for all). Even with as little as 300 milliseconds of exposure (one third of a second), detection was better for caricatures by 14 percentage points. This advantage increases to 43 percentage points for a 5 second exposure (significant interaction, $F(10, 1630) = 8.88, p < 0.001$). Crucially, while deepfakes were detected no better than chance without visual indicators, Caricatures boosted the likelihood of detection to above 50% across all presentation times. Thus, caricatures increase the likelihood that a deepfake will be detected as fake, even when participants had less than a second of exposure.

We also tested the degree of engagement required for Caricatures to be effective. How closely do users need to pay attention to the videos in order to benefit from this visual indicator? Unbeknownst to participants, *engagement probes* were embedded in the experiment (5 trials per 100-trial HIT), which consisted of standard deepfakes with artifacts which were extremely easy to detect. We reasoned that highly-engaged participants would succeed on all of these trials, but medium to low-engagement participants would miss some proportion of them. Participants were binned based on the proportion of engagement probes they correctly identified as fake, and the detection improvement between caricatures and standard fakes was measured for each bin. We find that Caricatures yield higher detection for 4 out of 5 levels of engagement (no improvement for extremely low engagement levels; 3/4 results significant $p < 0.05$, and 1/4 marginal at $p = 0.051$). Thus Caricatures improve detection even when users are not fully attending to the videos. Overall, these behavioral results demonstrate that the Caricature method is extremely effective at signalling to a human user that a video is fake.

5 Results: Model Experiments

While one goal of our framework was to create Deepfake Caricatures, a secondary goal was to assess whether partial supervision from human annotations of artifacts can yield a more performant deepfake detection model. We find that including the human annotations to supplement self-attention during training boosts model performance, leading our models to perform near the state of the art.

5.1 Evaluation Details

Datasets. We evaluate models on four benchmarks: FaceForensics++ (FF++) (37), The Deepfake Detection Challenge Dataset preview (DFDCp) (11), Celeb-DF v2 (27), DeeperForensics (DFo) (27), and FaceShifter (FShifter) (23). Faces were standardized across datasets, such that each video was 360×360 pixels, and showed a single face, with a minimum size of 50 pixels and a minimum margin of 100 pixels from the edge of the frame. See Supplement for more details on the datasets and standardization.

Model	CelebDFv2	DFDCp	FShifter	DFo	Overall
Xception (37)	73.7	65.7	72.0	84.5	75.3
Face X-ray (24)	79.5	62.1	92.8	86.8	81.2
CNN-GRU (38)	69.8	63.7	80.8	74.1	73.4
Multi-task (33)	75.7	63.9	66.0	77.7	71.9
DSP-FWA (26)	69.5	64.5	65.5	50.2	63.1
Two-branch (29)	73.4	64.0	-	-	-
Multi-attention (52)	67.4	67.1	-	-	-
LipForensics (52)	82.4	70.0	97.1	97.6	86.8
FTCN (53)	86.9	74.0	98.8	98.8	89.6
DCL (42)	82.3	<u>76.7</u>	92.4	97.1	87.1
RF (16)	86.9	75.9	<u>99.7</u>	<u>99.3</u>	<u>90.5</u>
S-B (39)	**93.2**	72.4	-	-	-
X+PCC (18)	54.9	62.7	-	-	-
CariNet (ours)	<u>88.8</u>	**76.9**	**99.7**	**99.5**	**91.2**

Table 1: **Detection performance results on unseen datasets.** We report Video-level AUC (%) on four tested benchmarks. All models are pretrained on FF++ (all manipulations). Top results are highlighted in bold, and second-best are underlined. CariNet achieves first or second place on all datasets.

Method	Train on remaining 3				
	DF	FS	F2F	NT	Avg
Xception	93.9	51.2	86.8	79.7	77.9
CNN-GRU	97.6	47.6	85.8	86.6	79.4
Face X-ray	99.5	93.2	94.5	92.5	94.9
LipForensics	99.7	90.1	99.7	99.1	99.5
CariNet (ours)	**99.9**	**99.9**	**99.7**	**99.3**	**99.7**

Table 2: **Generalization to unseen manipulations.** Video-level AUC (%) on four forgery types of FaceForensics++ (Deepfakes (DF), FaceSwap (FS), Face2Face (F2F) and Neural Textures (NT).

Baseline comparisons. We evaluate the accuracy of CariNet relative to several established baselines from the literature. For each model, we report performance as published for the datasets it was tested on, and retrain following published model specifications for datasets it was not initially tested on; further details are given in the Supplement.

5.2 Deepfake Detection Performance

Generalization to unseen datasets. Typically, the performance of deepfake classifiers is assessed primarily based on their ability to generalize well to datasets built with different techniques from what the classifier was trained on (17). Thus, we report our results as our model's ability to train on one dataset and perform well on another. We train models on FF++ and evaluate their performance on CelebDFv2, DFDCp, FShifter and DFo. As is typical (26; 17; 16), we report AUC as it describes model accuracy across a range of decision thresholds. Our CariNets show strong performance in this task, surpassing 13/13 models tested on DFDCp, FShifter and DFo, and 12/13 models on CelebDFv2. We hypothesize that this performance is in part resulting from the attentional framework proposed, which allows CariNet to build a more robust representation of artifact location and properties, focusing as needed on lips, eyes or other telling features. (Table 1).

Cross-forgery detection Several methods exist for creating deepfakes, and lead to different artifacts. In order to assess the quality of a detector, we must show good performance across deepfake-generation methods. We divided the FF++ dataset based on the deepfake-generation methods it includes (Deepfake (9), FaceSwap (20), Face2Face (43), and NeuralTextures (43)), and trained the model on a subset of FF++ containing all four manipulations. We then evaluated its performance on each method independently. In Table 2, we show performance against alternative models from the literature. Overall, our technique is on par with the best-performing model across generation methods.

Model	DFDCp	FF++	CelebDFv2	DFo	Overall
CariNet-S w/o att. mechanism	70.92	93.73	73.9	84.56	80.78
CariNet-S w/o modulation from att. module	72.34	94.82	76.5	91.21	83.72
CariNet-S w/ fixed att. (Gaussian)	68.15	90.15	71.1	82.23	77.91
CariNet-S w/ maps from Shiohara (2022)	71.07	93.81	74.11	84.91	80.98
CariNet-S (ours)	**72.90**	**96.81**	**80.1**	**94.33**	**86.04**

Table 3: **Ablation study results.** We show how certain components of our approach affect video-level AUC (%). We remove the attention mechanism from the Classifier module, we retain the attention mechanism but prevent modulation from the human-informed Artifact Attention module, we replace the human informed modulation with a fixed center bias and replace our heatmaps with a similar approach. All modifications yield lower performance than our proposed network.

Method	Clean	Cont.	Noise	Blur	Pixel
Xception	99.8	98.6	53.8	60.2	74.2
CNN-GRU	99.9	98.8	47.9	71.5	86.5
Face X-ray	99.8	88.5	49.8	63.8	88.6
LipForensics	99.9	99.6	73.8	96.1	95.6
CariNet	**99.9**	**99.9**	**78.6**	**96.2**	**97.6**

Table 4: **Generalization performance over unseen perturbations.** Video-level AUC (%) on FF++ over videos perturbed with 5 different modifications. We report averages across severity levels.

Robustness to unseen perturbations Another key aspect of a forgery detector is the ability to maintain performance even when the input videos have degraded quality, causing new, unseen perturbations. To analyze CariNet's behavior across types of perturbation, we test our model trained with FF++ on test videos from FF++ with 4 perturbations: Contrast, Gaussian Noise, Gaussian blur and Pixelation. We follow the framework of (17) and apply perturbations at 5 severity levels for our comparisons. We report average performance over the 5 severity levels in Table 4. We observe that our method outperforms previous approaches at most severity levels. We hypothesize that our human-guided attention framework might be of help in this setting, as humans are naturally capable of adapting to different lighting conditions, blurs, resolutions and other photometric modifications. This adaptability might be captured in the ground truth maps that guide the learning process of our Artifact Attention module.

Ablation studies We confirmed the contribution of adding human supervision via our Artifact Attention module by performing ablation studies across the different datasets under study. We observe performance drops for CariNet-S following ablation in four different scenarios (Table 3): **(1)** removing our custom self-attention blocks (Figure 4) from the Classifier module, and replacing them with simple 3D residual blocks, **(2)** retraining self-attention blocks in the Classifier module, but removing the modulatory input from the Artifact Attention module (this yields regular attention blocks with no key modulation), **(3)** replacing the output of the Artifact Attention module with a fixed center bias, operationalized as a 2D gaussian kernel with mean $\mu = (W/2, H/2)$ and standard deviation $\sigma = (20, 20)$, and **(4)** replacing the output of our attention module with self-blended image masks from (39) (closest to our approach). Overall, the complete model performed the best, demonstrating the effectiveness of incorporating self-attention modulated by human annotations into deepfake detection frameworks.

6 Conclusion and Discussion

This work takes a user-centered approach to deepfake detection models, and proposes a model whose focus is not just to detect video forgeries, but also alert the human user in an intuitive manner. Our CariNet shows excellent detection performance on four different datasets, and crucially, creates novel Deepfake Caricatures which allow for above-chance detection by human observers. Overall, this work establishes the importance of integrating computer vision and human factors solutions for deepfake mitigation. As with any misinformation detection system, there is a risk that our network could be leveraged to produce higher quality deepfakes. However, a system which allows humans to directly detect if a video is doctored will empower them to assess for themselves whether to trust the video.

References

[1] D. Afchar, V. Nozick, J. Yamagishi, and I. Echizen. Mesonet: a compact facial video forgery detection network. In *2018 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–7. IEEE, 2018.

[2] J. S. Ancker, A. Edwards, S. Nosal, D. Hauser, E. Mauer, and R. Kaushal. Effects of workload, work complexity, and repeated alerts on alert fatigue in a clinical decision support system. *BMC medical informatics and decision making*, 17(1):1–9, 2017.

[3] B. Bayar and M. C. Stamm. A deep learning approach to universal image manipulation detection using a new convolutional layer. In *Proceedings of the 4th ACM Workshop on Information Hiding and Multimedia Security*, pages 5–10, 2016.

[4] I. Bello, B. Zoph, A. Vaswani, J. Shlens, and Q. V. Le. Attention augmented convolutional networks. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3286–3295, 2019.

[5] P. J. Benson and D. I. Perrett. Perception and recognition of photographic quality facial caricatures: Implications for the recognition of natural images. *European Journal of Cognitive Psychology*, 3(1):105–135, 1991.

[6] A. Boyd, P. Tinsley, K. Bowyer, and A. Czajka. The value of ai guidance in human examination of synthetically-generated faces. *arXiv preprint arXiv:2208.10544*, 2022.

[7] F. Chollet. Xception: Deep learning with depthwise separable convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1251–1258, 2017.

[8] U. A. Ciftci and I. Demir. Fakecatcher: Detection of synthetic portrait videos using biological signals. *arXiv preprint arXiv:1901.02212*, 2019.

[9] D. Deepfake. `https://github.com/deepfakes/faceswap`, 2020.

[10] S. Deb and D. Claudio. Alarm fatigue and its influence on staff performance. *IIE Transactions on Healthcare Systems Engineering*, 5(3):183–196, 2015.

[11] B. Dolhansky, R. Howes, B. Pflaum, N. Baram, and C. C. Ferrer. The deepfake detection challenge (dfdc) preview dataset. *arXiv preprint arXiv:1910.08854*, 2019.

[12] R. Durall, M. Keuper, F.-J. Pfreundt, and J. Keuper. Unmasking deepfakes with simple features. *arXiv preprint arXiv:1911.00686*, 2019.

[13] Y. Fang, Q. Sun, X. Wang, T. Huang, X. Wang, and Y. Cao. Eva-02: A visual representation for neon genesis. *arXiv preprint arXiv:2303.11331*, 2023.

[14] M. Groh, Z. Epstein, C. Firestone, and R. Picard. Deepfake detection by human crowds, machines, and machine-informed crowds. *Proceedings of the National Academy of Sciences*, 119(1), 2022.

[15] D. Güera and E. J. Delp. Deepfake video detection using recurrent neural networks. In *2018 15th IEEE International Conference on Advanced Video and Signal Based Surveillance (AVSS)*, pages 1–6. IEEE, 2018.

[16] A. Haliassos, R. Mira, S. Petridis, and M. Pantic. Leveraging real talking faces via self-supervision for robust forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14950–14962, 2022.

[17] A. Haliassos, K. Vougioukas, S. Petridis, and M. Pantic. Lips don't lie: A generalisable and robust approach to face forgery detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5039–5049, 2021.

[18] Y. Hua, R. Shi, P. Wang, and S. Ge. Learning patch-channel correspondence for interpretable face forgery detection. *IEEE Transactions on Image Processing*, 32:1668–1680, 2023.

[19] M. I. Hussain, T. L. Reynolds, and K. Zheng. Medication safety alert fatigue may be reduced via interaction design and clinical role tailoring: a systematic review. *Journal of the American Medical Informatics Association*, 26(10):1141–1149, 2019.

[20] M. Kowalski. Faceswap. `https://github.com/MarekKowalski/FaceSwap/`, 2020.

[21] H. Li, B. Li, S. Tan, and J. Huang. Identification of deep network generated images using disparities in color components. *Signal Processing*, 174:107616, 2020.

[22] J. Li, T. Shen, W. Zhang, H. Ren, D. Zeng, and T. Mei. Zooming into face forensics: A pixel-level analysis. *arXiv preprint arXiv:1912.05790*, 2019.

[23] L. Li, J. Bao, H. Yang, D. Chen, and F. Wen. Faceshifter: Towards high fidelity and occlusion aware face swapping. *arXiv preprint arXiv:1912.13457*, 2019.

[24] L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, and B. Guo. Face x-ray for more general face forgery detection. *arXiv preprint arXiv:1912.13458*, 2019.

[25] Y. Li, M.-C. Chang, and S. Lyu. In ictu oculi: Exposing ai created fake videos by detecting eye blinking. In *2018 IEEE International workshop on information forensics and security (WIFS)*, pages 1–7. IEEE, 2018.

[26] Y. Li and S. Lyu. Exposing deepfake videos by detecting face warping artifacts, 2018.

[27] Y. Li, X. Yang, P. Sun, H. Qi, and S. Lyu. Celeb-df: A new dataset for deepfake forensics. *arXiv preprint arXiv:1909.12962*, 2019.

[28] L. Liu, H. Jiang, P. He, W. Chen, X. Liu, J. Gao, and J. Han. On the variance of the adaptive learning rate and beyond. *arXiv preprint arXiv:1908.03265*, 2019.

[29] I. Masi, A. Killekar, R. M. Mascarenhas, S. P. Gurudatt, and W. AbdAlmageed. Two-branch recurrent network for isolating deepfakes in videos. In *European Conference on Computer Vision*, pages 667–684. Springer, 2020.

[30] F. Matern, C. Riess, and M. Stamminger. Exploiting visual artifacts to expose deepfakes and face manipulations. In *2019 IEEE Winter Applications of Computer Vision Workshops (WACVW)*, pages 83–92.

IEEE, 2019.
[31] R. Mauro and M. Kubovy. Caricature and face recognition. *Memory & Cognition*, 20(4):433–440, 1992.
[32] D. M. Montserrat, H. Hao, S. Yarlagadda, S. Baireddy, R. Shao, J. Horváth, E. Bartusiak, J. Yang, D. Güera, F. Zhu, et al. Deepfakes detection with automatic face weighting. *arXiv preprint arXiv:2004.12027*, 2020.
[33] H. H. Nguyen, F. Fang, J. Yamagishi, and I. Echizen. Multi-task learning for detecting and segmenting manipulated facial images and videos. *arXiv preprint arXiv:1906.06876*, 2019.
[34] H. H. Nguyen, J. Yamagishi, and I. Echizen. Capsule-forensics: Using capsule networks to detect forged images and videos. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2307–2311. IEEE, 2019.
[35] S. J. Nightingale and H. Farid. Ai-synthesized faces are indistinguishable from real faces and more trustworthy. *Proceedings of the National Academy of Sciences*, 119(8):e2120481119, 2022.
[36] T.-H. Oh, R. Jaroensri, C. Kim, M. Elgharib, F. Durand, W. T. Freeman, and W. Matusik. Learning-based video motion magnification. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 633–648, 2018.
[37] A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, and M. Nießner. Faceforensics++: Learning to detect manipulated facial images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1–11, 2019.
[38] E. Sabir, J. Cheng, A. Jaiswal, W. AbdAlmageed, I. Masi, and P. Natarajan. Recurrent convolutional strategies for face manipulation detection in videos. *Interfaces (GUI)*, 3:1.
[39] K. Shiohara and T. Yamasaki. Detecting deepfakes with self-blended images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18720–18729, 2022.
[40] P. Sinha, B. Balas, Y. Ostrovsky, and R. Russell. Face recognition by humans: Nineteen results all computer vision researchers should know about. *Proceedings of the IEEE*, 94(11):1948–1962, 2006.
[41] J. Stehouwer, H. Dang, F. Liu, X. Liu, and A. Jain. On the detection of digital face manipulation. *arXiv preprint arXiv:1910.01717*, 2019.
[42] K. Sun, T. Yao, S. Chen, S. Ding, J. Li, and R. Ji. Dual contrastive learning for general face forgery detection. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 2316–2324, 2022.
[43] J. Thies, M. Zollhöfer, and M. Nießner. Deferred neural rendering: Image synthesis using neural textures. *ACM Transactions on Graphics (TOG)*, 38(4):1–12, 2019.
[44] R. Tolosana, S. Romero-Tapiador, J. Fierrez, and R. Vera-Rodriguez. Deepfakes evolution: Analysis of facial regions and fake detection performance, 2020.
[45] P. Tschandl, C. Rinner, Z. Apalla, G. Argenziano, N. Codella, A. Halpern, M. Janda, A. Lallas, C. Longo, J. Malvehy, J. Paoli, S. Puig, C. Rosendahl, H. P. Soyer, I. Zalaudek, and H. Kittler. Human–computer collaboration for skin cancer recognition. *Nature Medicine*, 26(8):1229–1234, Aug 2020.
[46] B. Tversky and D. Baratz. Memory for faces: Are caricatures better than photographs? *Memory & cognition*, 13(1):45–49, 1985.
[47] M. Vaccaro and J. Waldo. The effects of mixing machine learning and human judgment: Collaboration between humans and machines does not necessarily lead to better outcomes. *Queue*, 17(4):19–40, 2019.
[48] D. Xing, J. Yang, J. Jin, and X. Luo. Potential of plant identification apps in urban forestry studies in china: comparison of recognition accuracy and user experience of five apps. *Journal of Forestry Research*, 32(5):1889–1897, Oct 2021.
[49] X. Yang, Y. Li, and S. Lyu. Exposing deep fakes using inconsistent head poses. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8261–8265. IEEE, 2019.
[50] H. Zhang, I. Goodfellow, D. Metaxas, and A. Odena. Self-attention generative adversarial networks. *arXiv preprint arXiv:1805.08318*, 2018.
[51] M. R. Zhang, J. Lucas, G. Hinton, and J. Ba. Lookahead optimizer: k steps forward, 1 step back. *arXiv preprint arXiv:1907.08610*, 2019.
[52] H. Zhao, W. Zhou, D. Chen, T. Wei, W. Zhang, and N. Yu. Multi-attentional deepfake detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2185–2194, 2021.
[53] Y. Zheng, J. Bao, D. Chen, M. Zeng, and F. Wen. Exploring temporal coherence for more general video face forgery detection. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15044–15054, 2021.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract is a summary of the main paper's content.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We briefly discuss our limitations in the final section of our paper and in our conclusion.

   Guidelines:

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

3. **Theory Assumptions and Proofs**

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

   Answer: [NA]

Justification: No theoretical results.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

   Answer: [Yes]

   Justification: Our results can be reproduced with the information we give in the paper. Full code and data will be released at publication time.

   Guidelines:

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

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

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

Answer:[Yes]

Justification: This will be provided at publication time to avoid anonymity issues.

Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: Detaield in our experiments section.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: Error bars and statistical results provided for our human results.

   Guidelines:

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

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

8. **Experiments Compute Resources**

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

   Answer: [Yes] We provide information about GPUs used.

   Justification: We provide information about GPUs used.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We do not violate any ethical guidelines.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [Yes]

    Justification: We briefly mention the impact in our conclusion.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to

generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [No]

    Justification: We don't describe these safeguards because of lack of space, but plan to include responses if reviewers request them.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [Yes]

    Justification: We cite and credit all owners.

    Guidelines:

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

13. **New Assets**

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

Answer: [Yes] The assets such as model checkpoints will include documentation in the corresponding github.

Justification: The assets such as model checkpoints will include documentation in the corresponding github.

Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

Answer: [Yes]

Justification: Included in supplemental.

Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

Answer: [Yes]

Justification: We obtained IRB approvals and describe info given to the subjects in the supplemental.

Guidelines:

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