Content Safeguarding on Social Media

Peyman Najafirad (Paul Rad)
Associate Professor Cyber and AI
Sr. Member of the National Academy of Inventors (NAI)
Director Secure AI and Autonomy Lab
UTSA School of Data Science Core Faculty
The University of Texas at San Antonio
Phone: 210.872.7259

Disclaimer: This presentation contains harmful image content, such as sexually explicit, cyberbullying, and self-harm images that are highly offensive and might disturb the viewers.

Social Media Platform

Statistics on social media adoption:

- 58% of social media users access social media platforms on a daily basis.
- The average social media user spends 2 hours and 24 minutes per day on social media.
- 93% of social media users access social media platforms on their mobile devices.
- The most popular social media platform among young adults (ages 18-24) is Instagram.
- The most popular social media platform among older adults (ages 55-64) is Facebook.

A photo of the World Cup trophy being held by Lionel Messi, the captain of the Argentine national football team. Source: **Getty Images**

Social Media Platform

There are 4.9 billion social media users worldwide. This represents 60% of the world's population.

The most popular social media platforms in 2023 are:

- Facebook (2.9 billion users)
- YouTube (2.5 billion users)
- WhatsApp (2 billion users)
- Instagram (2 billion users)

These haunting underwater photos portray climate change in a new way

Social media platforms have become important tools for communication, education, and entertainment. They have also played a role in social movements and political change.

Social Media Platform Content Delivery Mechanism

- **User engagement**: Social media platforms track how users interact with content, such as by liking, sharing, and commenting.
- **Social connections**: Social media platforms also consider the user's social connections.
- **Trending topics**: Social media platforms also track trending topics to deliver relevant content to users.
- **Algorithms**: Social media platforms use algorithms to rank and deliver content to users.

We often hear about social medial algorithms determine the content we see and their influence on our thinking. Yet, we're usually unaware of the <u>content moderation methods</u> that filter out what we don't see.

Why Social Media Content Moderation

- **To protect users from harmful content**: Social media platforms have a responsibility to protect their users from harmful content, such as hate speech, violence, and misinformation.
  - Hate speech, Violence, Misinformation, Child sexual abuse material (CSAM), Terrorism-related content, Nudity and pornography, Copyright infringement, Spam

- **To comply with laws and regulations**: Social media platforms are subject to a variety of laws and regulations, many of which relate to content moderation.

- **To maintain a positive user experience**: Social media platforms want to create a positive user experience for their users.

Timeline of Content Moderation Filtering

Recent development (2021 – present)

- Social media platforms are under increasing pressure to regulate content on their platforms.
- The European Union passed the Digital Services Act, which requires social media platforms to do more to moderate content and prevent the spread of harmful content.
- Platforms are continuing to invest in new content moderation tools and technologies.
- However, the challenges of content moderation remain complex and there is no easy solution.

Content moderation challenges (2016-2020)

- The spread of misinformation and disinformation on social media platforms becomes a major concern.
- Platforms begin to focus more on proactively identifying and removing harmful content, rather than relying on user reports.
- Artificial intelligence (AI) is increasingly used to develop new content moderation tools.
- However, AI-powered tools are still not perfect and can lead to censorship and the removal of legitimate content.

Rise of social media (2011-2015)

- Social media platforms become increasingly popular and the amount of user-generated content explodes.
- Platforms begin to invest more heavily in content moderation, but it is still difficult to keep up with the volume of content.
- Automated content moderation tools become more sophisticated, but they are still prone to errors.
- Major platforms, such as Facebook and Twitter, begin to hire human content moderators to help review and remove harmful content.

Early days (2004-2010)

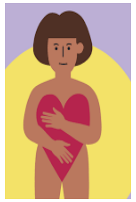
- Most platforms rely on user reports to identify and remove harmful content.
- Some platforms, such as Facebook, begin to experiment with automated content moderation tools.

Methods to Moderate Content

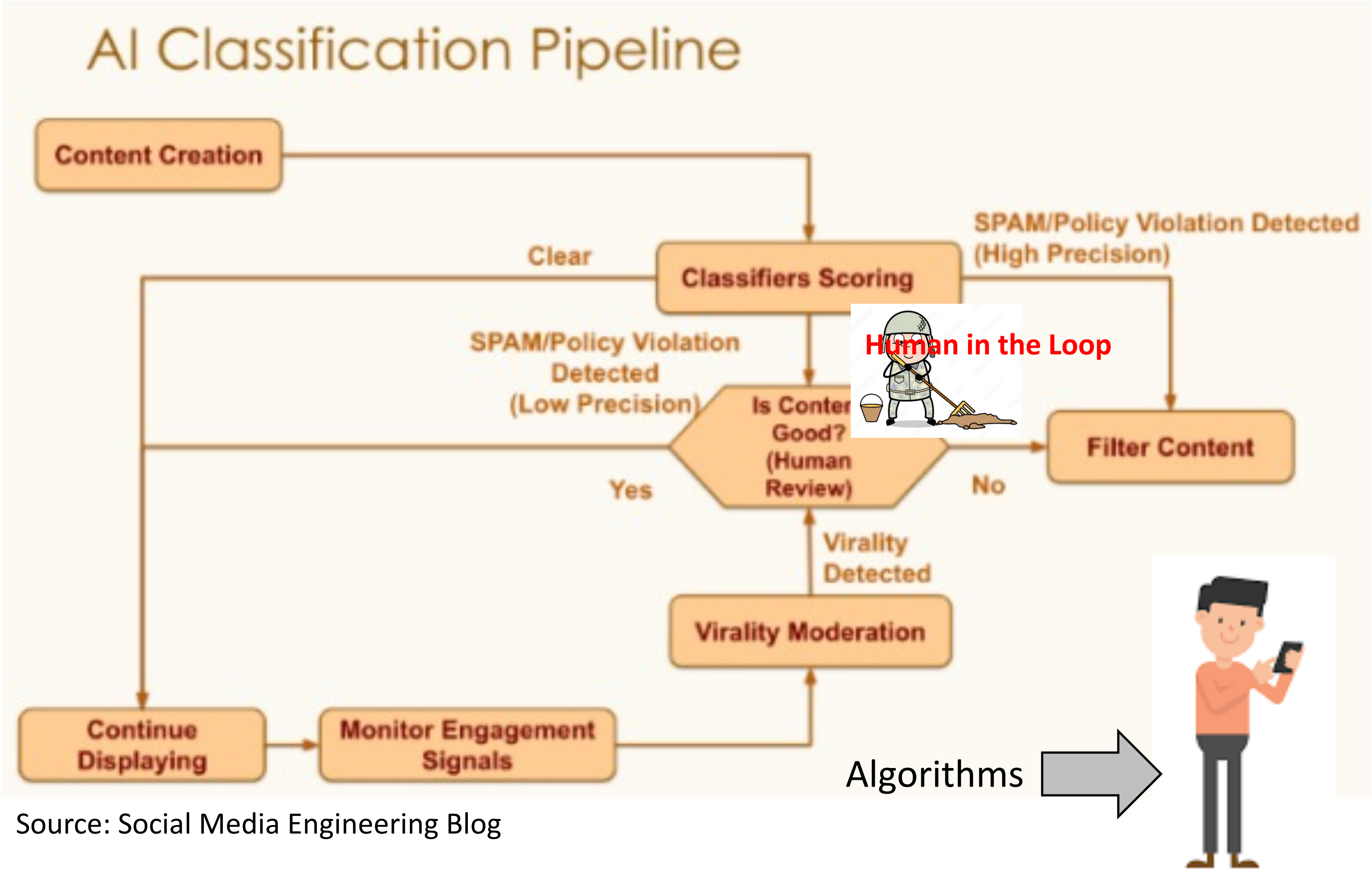
- **Automated moderation score**: Social media platforms use automated tools to identify and remove harmful content.
- **Human moderation**: Social media platforms also employ human moderators to review content and to remove any content that is not flagged by automated tools.
- **Community reporting**: Social media platforms also rely on their users to report harmful content.

Behind the Scene

Porn



Cyber Bullying



Source: Social Media Engineering Blog

Algorithms

User Social Media Feed

Some of the Challenges with Content Moderations

1) Content moderation is a challenging job and negative impact on their mental and emotional well-being

Content moderators are exposed to a wide range of harmful and disturbing content on a daily basis.

Some of the Challenges with Content Moderations

2) Subjective judgments: Moderators often have to make subjective judgments about whether or not a piece of content violates the platform's policies.

The has become an iconic symbol of the horror of the Vietnam War. **It was awarded the Pulitzer Prize for Spot News Photography in 1973**

Initially removed by Facebook in 2016

Content moderators often have to make difficult decisions about what content to remove and what content to allow in a short period of time.

famous historical photograph often referred to as "Napalm Girl," taken during the Vietnam War.

Some of the Challenges with Content Moderations

3) Adversarial attacks on content moderation systems: In the context of content moderation, adversarial attacks could be used to fool content moderation systems into approving harmful content or removing legitimate content.

4) Transparency and accountability: Social media platforms have been criticized for being opaque about their moderation policies and procedures. Users want to know how their content is being moderated and why certain pieces of content are removed.

5) Bots using generative AI creating fake content: Generative AI is a type of artificial intelligence that can be used to create new content, such as text, images, and videos.

AI-generated image: a fake image of former President Donald Trump being arrested.

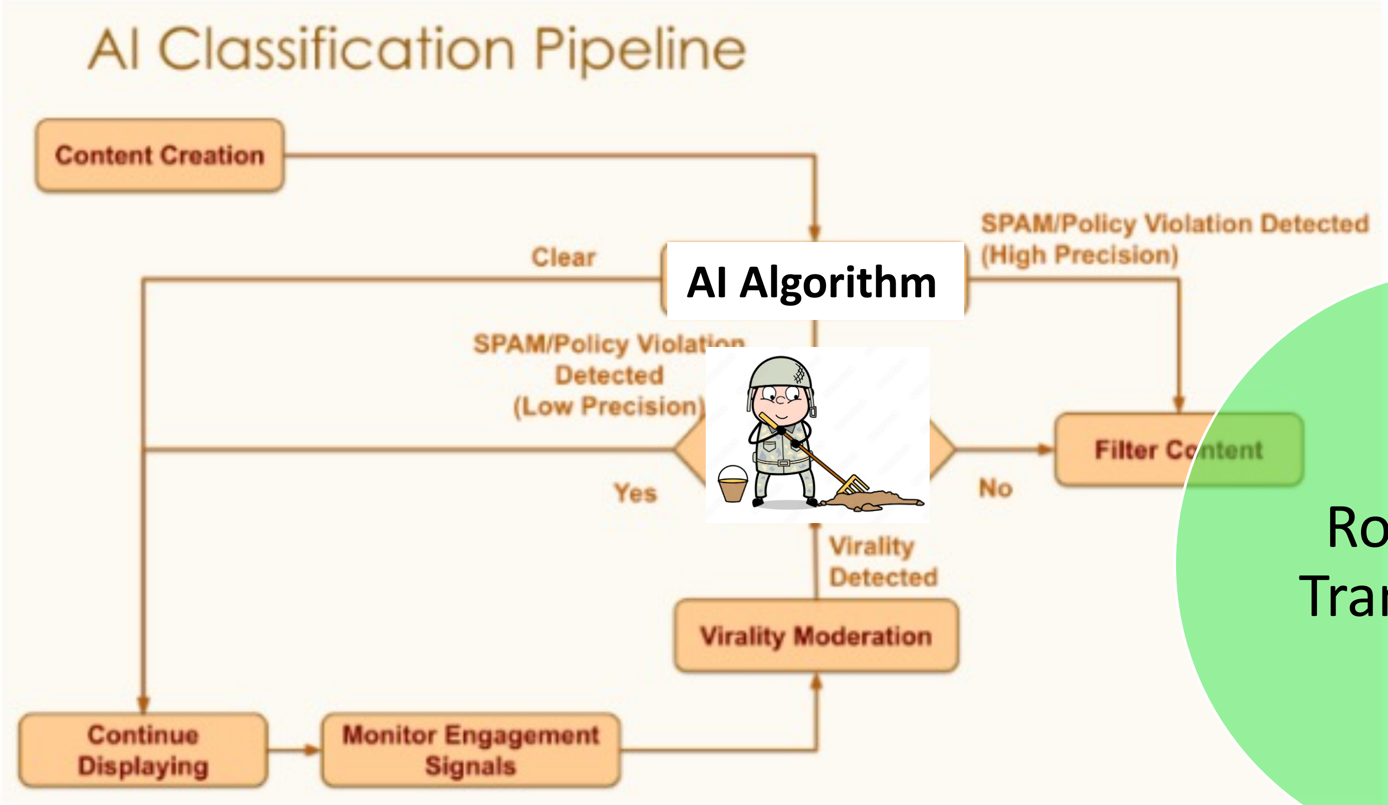
Eliot Higgins; Annotation by NPR

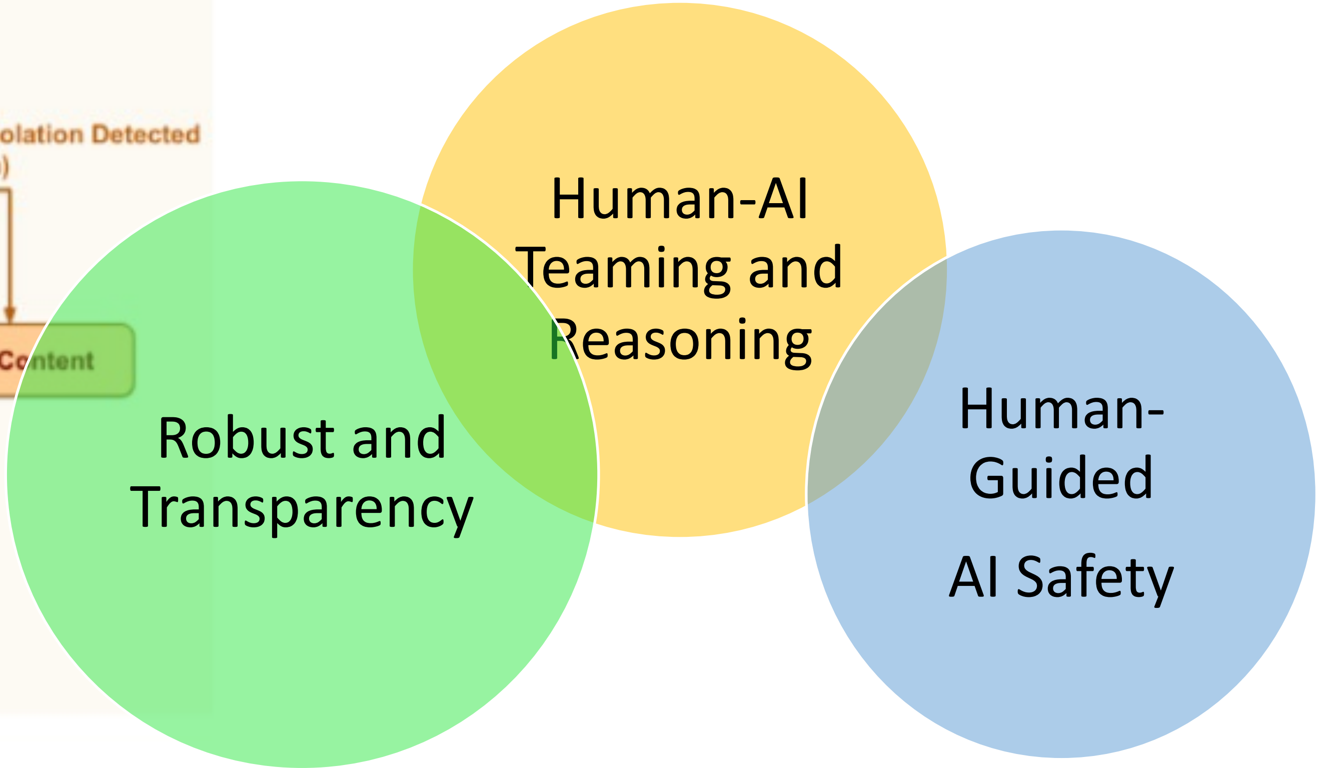
Challenges of Content Moderation

Content moderation is a complex and challenging task

- Adversarial attacks on content moderation systems
- Human Moderator in the loop and seeing harmful contents all day
- Transparency and accountability
- Subjective judgments for content removal

AI Safety and Security





Research Questions

<u>Question 1</u>: Can we make an unsafe image a safer image?

<u>Question 2</u>: Can we identify specific areas in the image that makes it unsafe?

<u>Question 3</u>: Can a defense mechanism be established to stop an adversarial attack from misidentifying the area of an image that makes it unsafe?

<u>Question 4</u>: How can we use human-AI teaming to reduce moderators' exposure to harmful content?

Outline

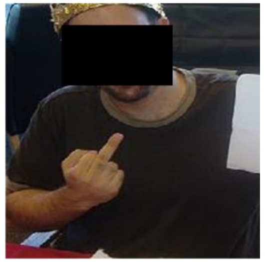
- Introduction
- Content Moderation Challenges
- **Robust Semantic Representation**
- Counterfactual Explanations Content Obfuscation
- Reasoning with Conditional Vision Language Model
- Future Directions
- Acknowledgments

Adaptive Clustering of Robust Semantic Representations for Adversarial Image Purification on Social Networks

By Samuel Henrique Silva, Arun Das, Adel Alaeddini, Peyman NajafiRad

Proceedings of the International AAAI Conference on Web and Social Media, 2022

Supervised Learning



Detected as cyberbullying

$\mathbf{x_i}$

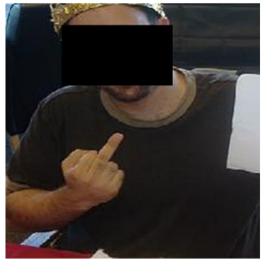
- Assuming an input x and a class label $y \in R^C$.

- A classification algorithm $f(.)$, is a complex function, parameterized in θ which maps x into a prediction $\hat{y}$, through $\mathrm{f}_\theta(x) \rightarrow \hat{y}$

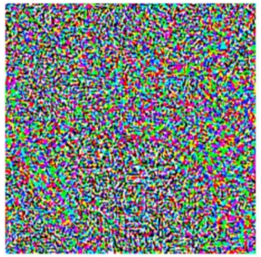
- Parameters are learned minimizing the distance between the prediction $\hat{y}$, and the true label y, through an optimization problem.

$$\min_\theta \mathrm{L}(\mathrm{f}_\theta(\mathrm{x}), \mathrm{y})$$

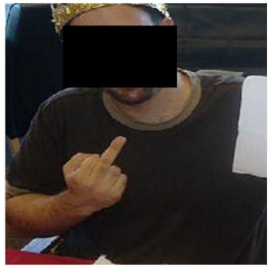
Adversarial Attack



$+\boldsymbol{\delta} *$



$=$



Detected as cyberbullying

$\mathbf{x_i}$

Not detected as cyberbullying

$\mathbf{x_i+\delta}$

- Parameters are learned minimizing the distance between the prediction $\hat{y}$, and the true label y, through an optimization problem.

$$\min_{\theta} \mathrm{L}(\mathrm{f}_{\theta}(\mathrm{x}), \mathrm{y})$$

- We want δ a small perturbation such that $\left|\left|x - (x + \delta)\right|\right|_2 \leq \epsilon \ll 1.$

$$\delta \leftarrow Random$$
$$\max_{\delta < \epsilon} \mathrm{L}(\mathrm{f}_{\theta}(\mathrm{x} + \delta), \mathrm{y})$$

- $x' = (x + \delta)$ is a new sample, never experienced by the model, generated to maximize the prediction loss.

Dissemination of Unsafe Content by Adversary

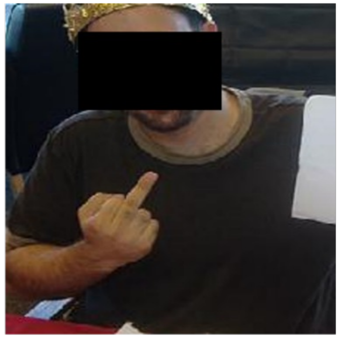
- **Adversarial Images**: Deceptive digital images that fool AI-based image recognition systems, causing misclassification, while appearing unchanged to human viewers.

- **Unsafe Images**: Potentially harmful or offensive content requiring effective detection and moderation to protect viewers.

Attack	State-of-the-Art Unsafe Image Detectors				
	Clarifai (%)	Yahoo Open NSFW (%)	Amazon Rekog-nition (%)	MS Azure (%)	Google Safe-Search (%)
No Attack	80	84	90	96	90
Square	22	6	50	68	76
Square+GB	4	4	74	94	76
AutoAttack	22	56	84	90	88

We randomly sampled a set of 50 strongly sexually explicit images to craft adversarial images using the three attacks each, after which they were tested against various existing detectors provided through their public APIs, as well as the open-source model.

NSFW



Cyberbullying

Self-Harm

Robust Optimization Framework

- We can change the formulation objective, to include such cases in the training process.

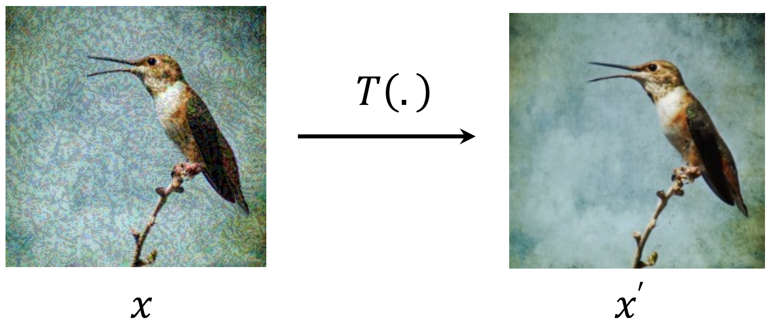
$$\min_{\theta} \max_{\delta < \epsilon} \mathrm{L}(\mathrm{f}_{\theta}(\mathrm{x} + \delta), \mathrm{y})$$

- In which, the training set is iteratively augmented with these purposefully crafted perturbations.

- Such solution, limits the generalization to only seen attacks.

Purification Thru Reconstruction

We propose to transform the input such that we eliminate corruptions before image is used in the desired task.



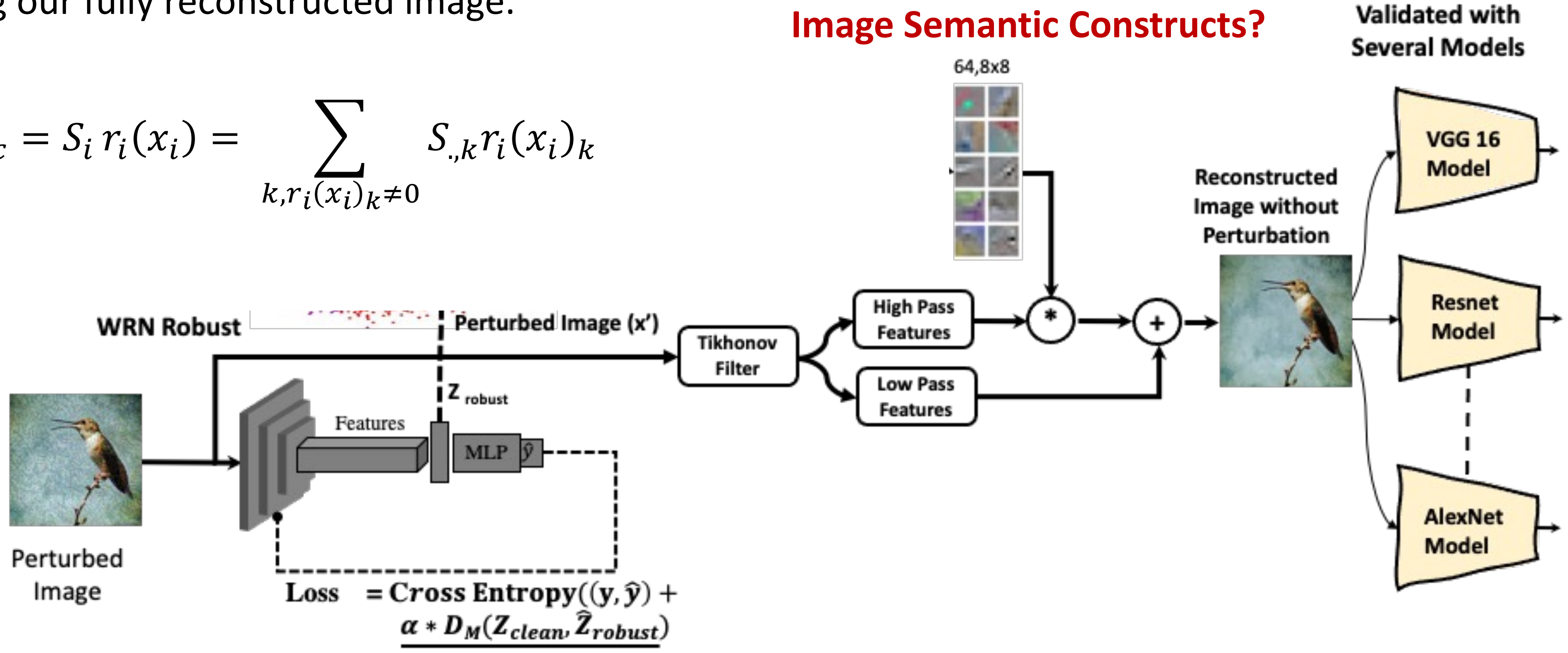
Given x', we want to transform x', such that $T(x')$, and $f(T(x)) = f(T(x')) = y$.

Reconstruction with Purification Methodology

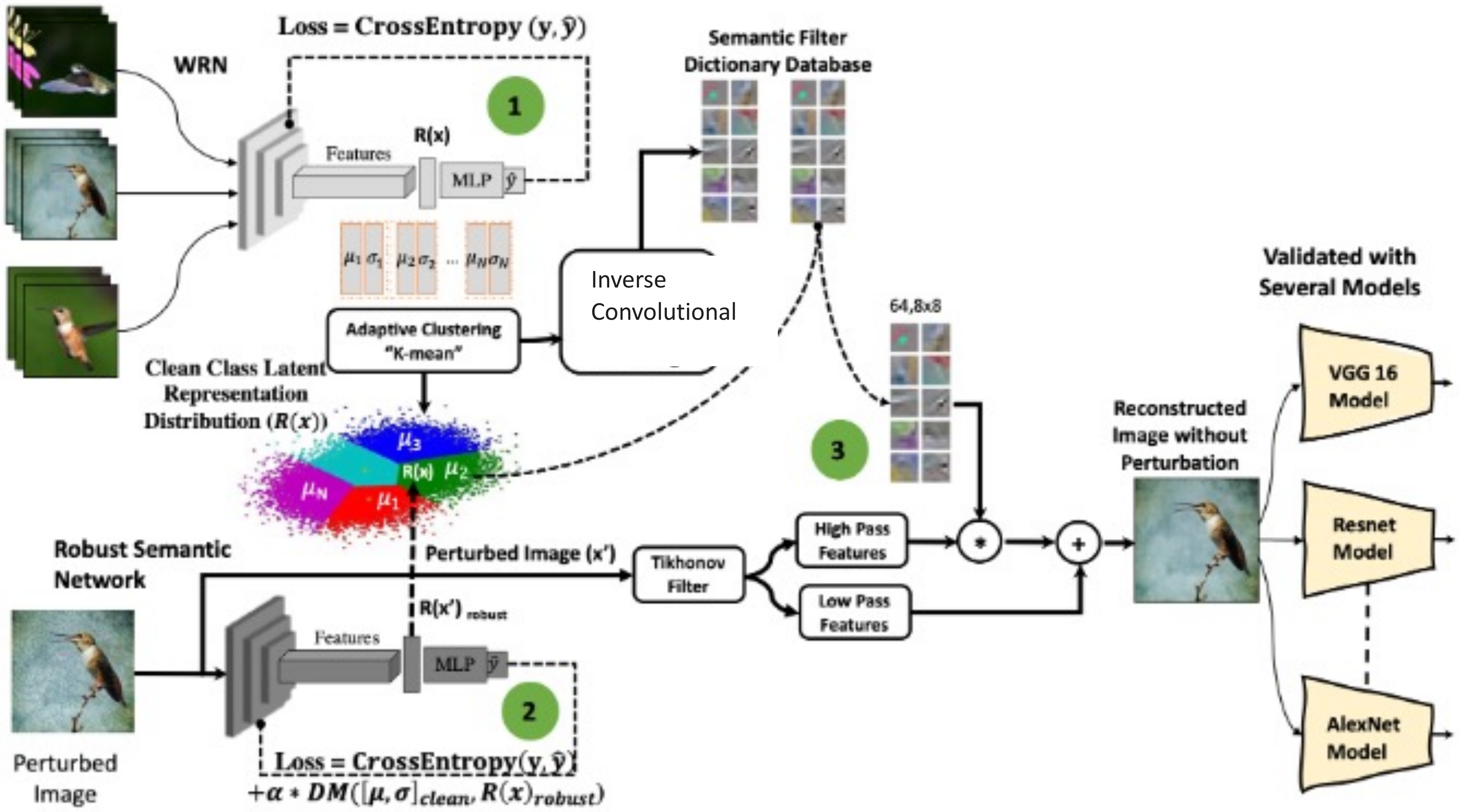
Step 1: The perturbed image, is decomposed in a high pass, and low pass components, using Tikhonov Filter.

Step 2: The low pass component is combined with the reconstructed high pass component, generating our fully reconstructed image.

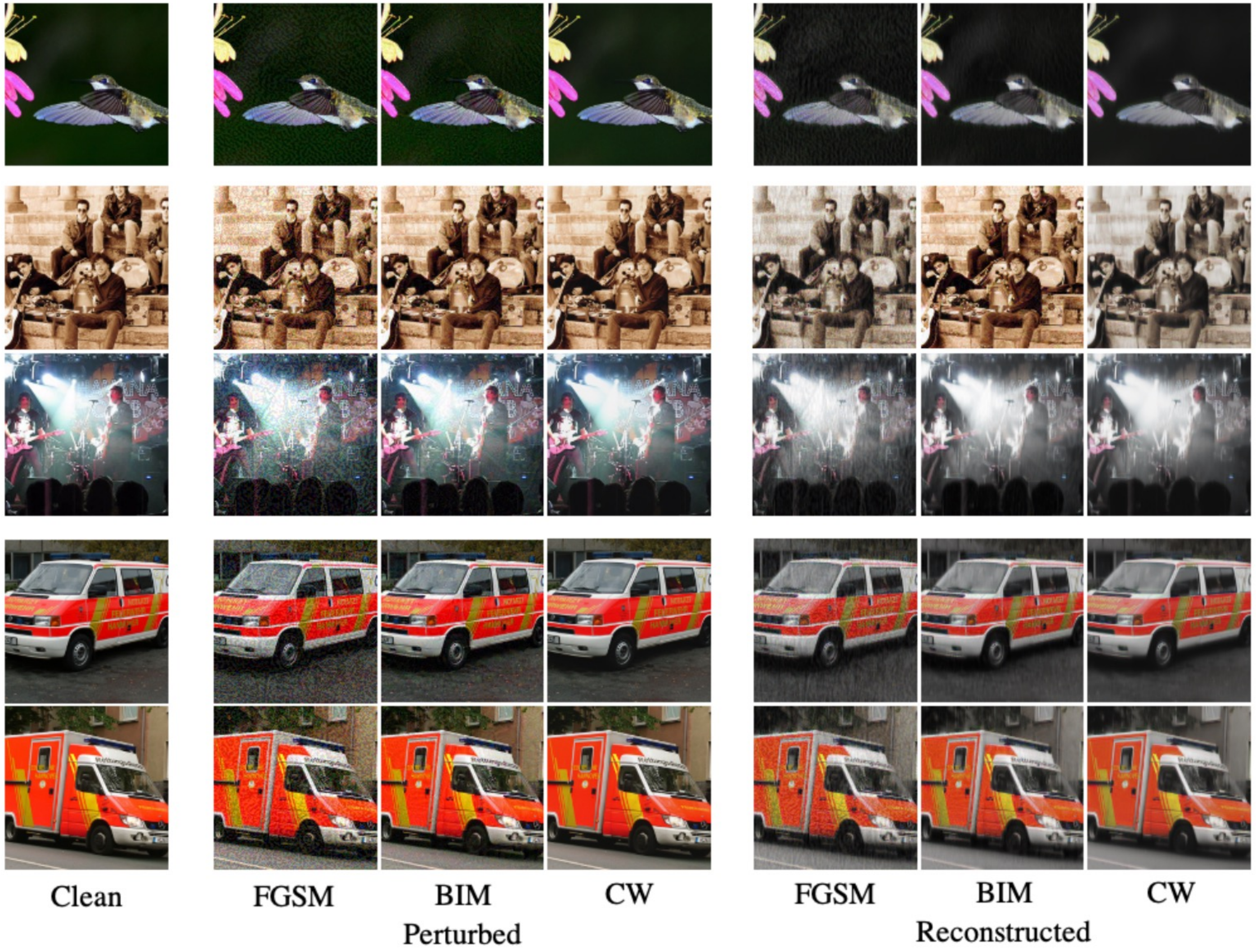
$$x_{rec} = S_i\, r_i(x_i) = \sum_{k, r_i(x_i)_k \neq 0} S_{.,k} r_i(x_i)_k$$



Purification Content on Social Network



Qualitative Results



The reconstruction output of our defense on images extracted from ImageNet-10. Our defense mechanism is independent of the attack. From left to right, first is the original clean image. The third, fourth and fifth columns shows the output of FGSM, BIM and CW.

The next 3 columns show the reconstruction output of the respective attacks.

Quantitative Results

We initially evaluated the adversarial training, without image reconstruction. We evaluated how our model can further improve adversarial pre-trained models

Table: CIFAR-10 Classification accuracy using WRN-32-10 trained with PGDAT + ASC

Defense	No-attack	PGD	CW	BIM	TPGD
No defense	86.36%	27.68%	9.91%	42.50%	30.37%
PGDAT+ASC (PGDAT)	79.79%	67.43%	50.74%	62.59%	75.25%
PGDAT (PGDAT)	79.06%	66.74%	50.46%	61.25%	74.69%
PGDAT+ASC (Random)	77.36%	66.99%	39.66%	60.52%	73.92%
PGDAT (Random)	75.63%	65.55%	52.26%	57.86%	73.57%
PGDAT+ASC (Trades)	87.56%	73.60%	56.30%	75.27%	78.18%

Quantitative Results

We’ve also compared our model against different defenses. We’ve attacked VGG-16 with 5 different attack methods and evaluated the accuracy of VGG-16 (not adversarially trained) on the images reconstructed by different defenses.

Table: CIFAR-10 Classification accuracy using VGG-16 trained with images reconstructed with ASC in com-parison with other input transformation-based defenses. All methods are trained and tested on the provided recon-structed data. ‘Clean’ denotes that no attack was added to the baseline dataset, and ‘No Defense’ indicates no image reconstruction was applied.

Defense	**Clean**	**FGSM-0.08**	**FGSM-0.04**	**BIM**	**DeepFool**	**CW**
No defense	94.23	58.16	65.23	18.03	17.60	9.36
MagNet	90.35	61.45	65.21	43.12	65.35	48.45
PixelDefend	85.26	68.10	73.29	77.29	74.14	75.79
STL	83.60	71.03	75.47	75.31	79.59	79.06
ASC	94.23	78.57	75.00	76.78	83.92	87.50

Quantitative Results

We’ve evaluated our method, in larger dimensional input. We’ve evaluated the accuracy of our model against different reconstruction defenses, in ImageNet-10. We collected 10 classes from ImageNet and compared the results of these defenses, against the 10 classes.

Table: ImageNet-10 classification accuracy using VGG-16. Each model was trained and tested using their respective transformed data. In (a) resolution of images is 64x64, and(b) resolution is 128x128.

(a) Resolution 64x64

Defense	Clean	FGSM-0.08	FGSM-0.04	BIM	DeepFool	CW
No Defense	86.65	28.16	30.8	18.83	8.11	7.51
TVM|	75.55	59.97	69.3	71.56	72.1	71.87
Quilting	77.41	73.04	74.18	76.42	76.46	76.62
Crop-Ens	75.08	69.68	72.21	73.69	74.01	73.04
PD-Ens	82.5	66.34	76.07	79.03	79.55	78.13
STL	84.21	75.14	80.38	81.03	82.21	81.22
ASC	87.50	84.37	78.12	87.50	84.37	81.25

(b) Resolution 128x128

Defense	Clean	FGSM-0.08	FGSM-0.04	BIM	DeepFool	CW
No Defense	89.91	21.23	24.09	17.90	5.84	5.04
TVM	85.91	25.68	43.86	65.86	63.60	61.29
Quilting	81.49	39.03	58.89	64.34	62.42	59.22
Crop-Ens	77.30	46.22	64.47	68.76	70.60	68.88
PD-Ens|	87.89	23.33	42.86	72.21	73.59	72.72
STL	86.54	47.33	66.06	73.23	73.01	74.32
ASC	87.50	87.25	89.28	87.5	90.62	88.25

Quantitative Results

One of the claims in our model is that it is model agnostic. We’ve attacked 4 different networks, with 5 different attacks, and evaluated the accuracy of these models, on the reconstructed images.

Table 2: Cifar-10 classification accuracy against adversarial across different models when input transformed with our model

Defense	No-attack	FGSM-0.08	FGSM-0.04	BIM	DeepFool	CW
AlexNet	91.07%	73.07%	76.78%	83.92%	82.69%	86.53%
VGG-16	94.23%	78.57%	75.00%	76.78%	83.92%	87.05%
ResNet50	95.19%	76.78%	86.53%	84.61%	90.38%	90.78%
GoogleNet	90.38%	79.80%	83.65%	87.5%	88.5%	85.57%

Conclusion

- We have proposed a novel adaptive semantic clustering adversarial defense that presents state-of-the-art results against l2 bounded adversarial attacks, unseen at test time.

- Our method cluster features of the dataset which are semantically similar and demonstrate that to certain extend generalization to unseen distributions, can be achieved by learning better representations for our dataset.

Outline

- Introduction
- Content Moderation Challenges
- Robust Semantic Representation
- **Counterfactual Explanations Content Obfuscation**
- Reasoning with Conditional Vision Language Model
- Future Directions
- Acknowledgments

Towards Targeted Obfuscation of Adversarial Unsafe Images using Reconstruction and Counterfactual Super Region Attribution Explainability

By Mazal Bethany, Andrew Seong, Samuel Henrique Silva,
Nicole Beebe, Nishant Vishwamitra, and Peyman NajafiRad

32nd USENIX Security Symposium (USENIX Security 23), 2023

Investigating Explanation Techniques for Obfuscating Unsafe Images

We trained three binary ResNet-50 classifiers to distinguish between safe and unsafe images

We explored three explanation techniques to automatically obfuscate the unsafe regions in the cyberbullying images, pointed out by the generated explanations.

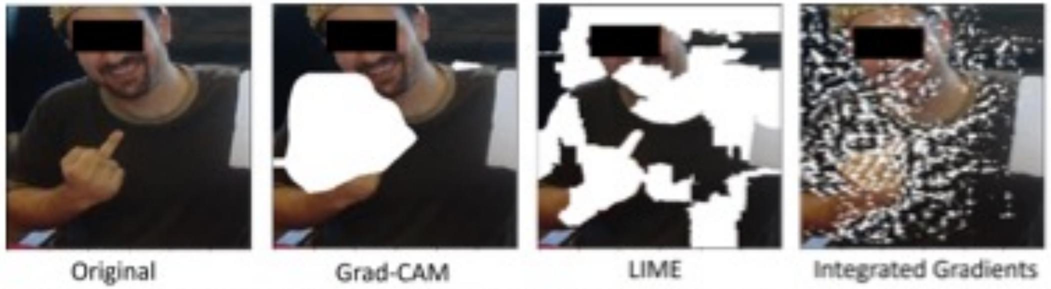


Figure 1: Samples of an unsafe image obfuscated according to the regions pointed-out by three explainability methods.

	Grad-CAM		Integrated Gradients		LIME	
	% of Pred. Changed	% of Image Obf.	% of Pred. Changed	% of Image Obf.	% of Pred. Changed	% of Image Obf.
Sexually Explicit	43	20	32	20	100	65.21
Cyberbullying	79	20	29	20	100	63.84
Self-Harm	65	20	41	20	100	71.92

Table 1: Experiment showing the unsuitability of different types of explanation methods for content obfuscation.

Existing explanation methods are unsuitable for image obfuscation

Causation Explainability

Counterfactual Explanation

A counterfactual explanation can be defined as taking the form:

- A decision y was produced because variable X had values (v1, v2, . . .) associated with it
- If X instead had values (v1′, v2, . . .), and all other variables had remained constant, score y′ would have been produced.

A counterfactual explanation would be of the form of the statement
"You were denied a loan because your annual income was low. If your income had been slightly higher, you would have been offered a loan."

While a factual explanation would be in the form of the statement
"You were denied the loan since a previous customer matching your profile alsoasked for the same amount, and was also denied the loan.

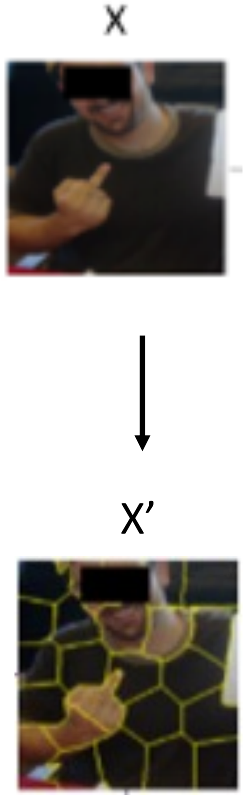
Can we identify specific areas in the image that makes it unsafe

A simple approach to identify sub-object areas can be formulated as a two-phase approach.

Step 1: Partition the image into non intersecting sub-object regions X as a non-intersecting set of K regions given by $\{z_1, z_2, \cdots, z_K\}$

Assuming we have binary predictive model $f(X) \rightarrow 0, 1$

Step 2: The counterfactual analysis of alternate versions of the image using a combinatorial regional search algorithm



Subobject Region Attribution Value

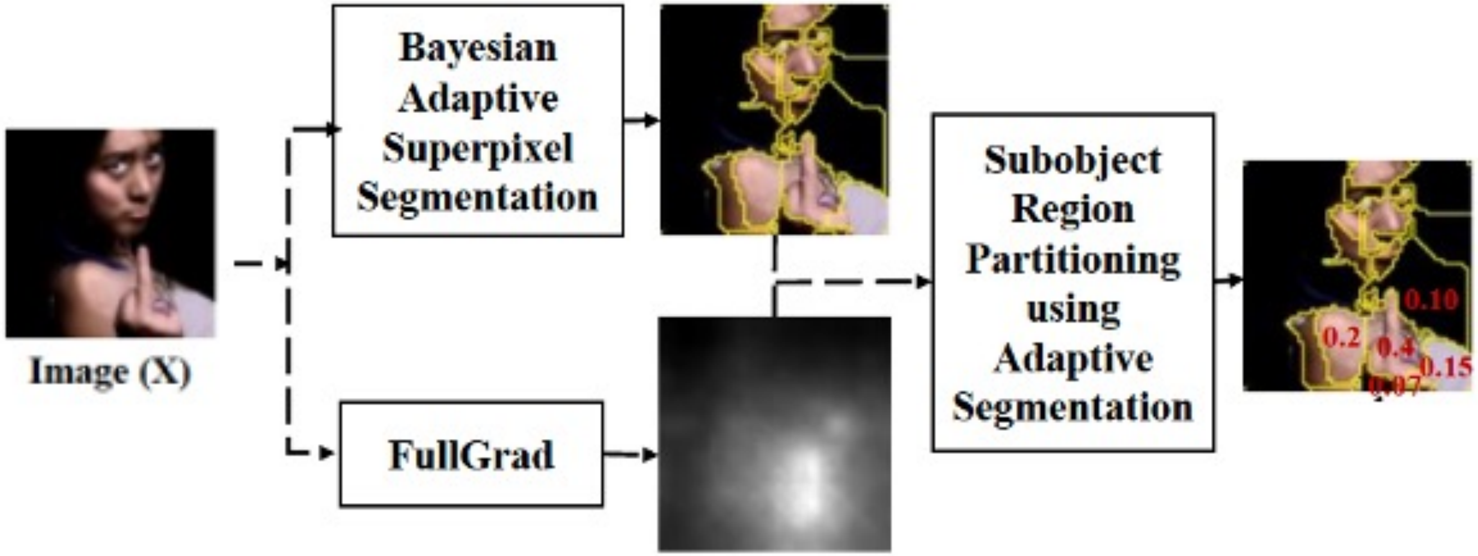
Definition 1: Subobject Region Attribution Score

Using the attribution map of model f (X) and the subobject regions

$\{z_1, z_2, \cdots, z_K\}$ created by adaptive segmentation for the input image X, we define the sub-object region attribution score, $\{s_1, s_2, \cdots, s_K\}$ as follows:

$$s_k = \frac{1}{n.m}\sum_{n}\sum_{m} L^{c}_{Grad-CAM(F,X)}[i,j], X[i,j] \in z_k$$

Subobject Region Attribution Value



Although feature attributions highlight features that are significant in terms of how they affect the model’s ability to predict, they do not indicate that altering significant features would result in a different desired outcome

Subobject Region Confidence Reduction

Definition 2: Subobject Region Confidence Reduction
Given a model Y = f (X) that takes an image X with sub-object regions

$X = [z_1, ..., z_n]^T$ and outputs a probability distribution Y.

The confidence reduction cr_k of subobject region z_k, ($k \in [1, n]$) towards probability distribution Y is the change of the output by masking the k-th subobject region of X, while being classified as the same class follows:

$$cr_k = f(X) - f(X \circ Mask(z_k))$$

0.95 0.85

F(X) = 0.95

F(X) = 0.93

Subobject Region Confidence Reduction

F(X) = 0.95

$cr_k = 0.95 - 0.93 = 0.02$

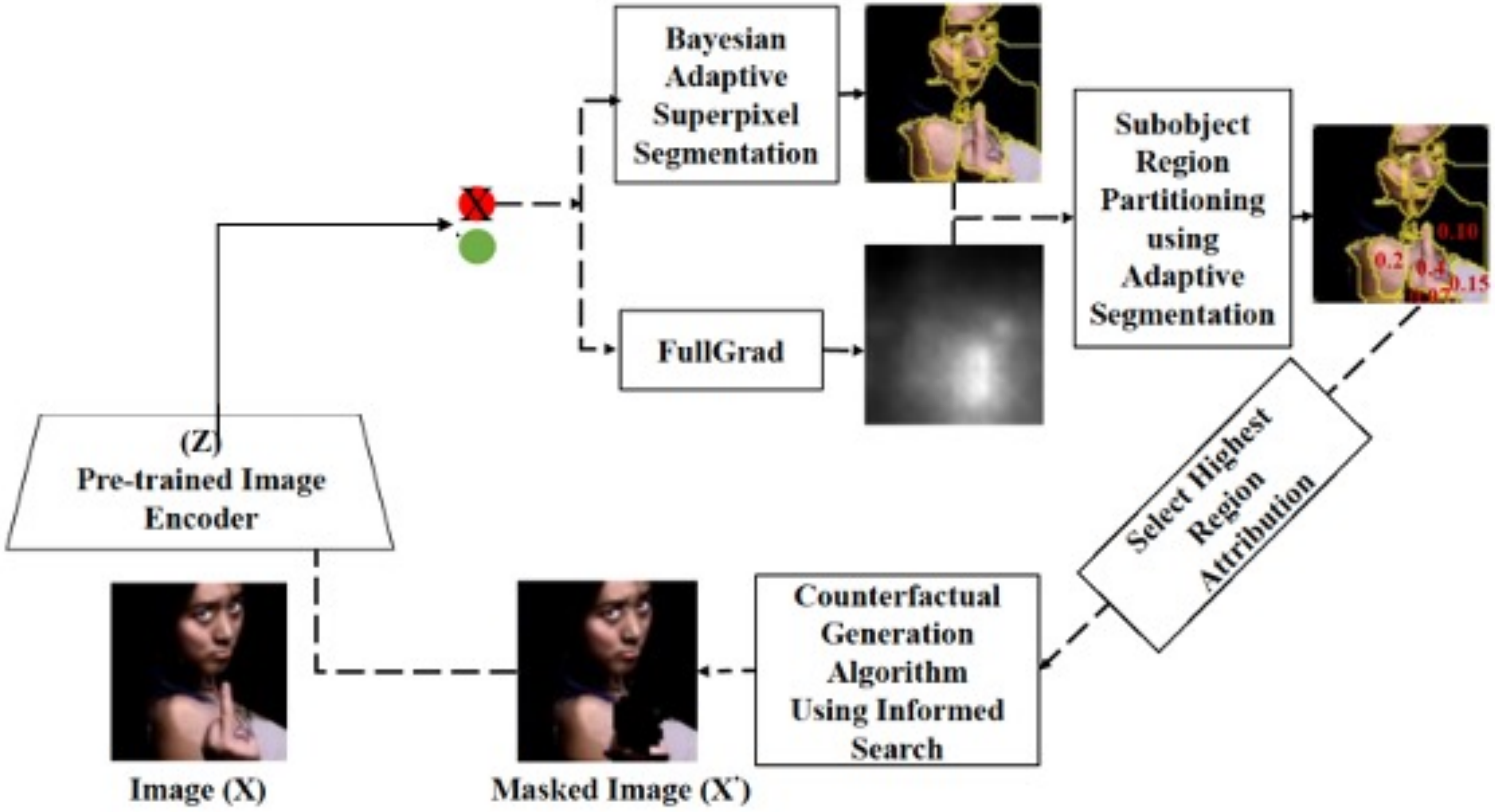
F(X) = 0.93

F(X) = 0.95

$cr_k = 0.95 - 0.90 = 0.05$

F(X) = 0.90

Image Obfuscation using Counterfactual Super Region Attribution Explainability



Our greedy region search, starts with first sorting the K regions in descending order by the average attribution for each region. The greedy region search considers a subset of regions $k \in K$. k begins with the top region by average attribution and iteratively expands to the top two regions by average attribution and so on until an x' is found such that $f(x') \neq f(x)$

Minimum Greedy Region Search

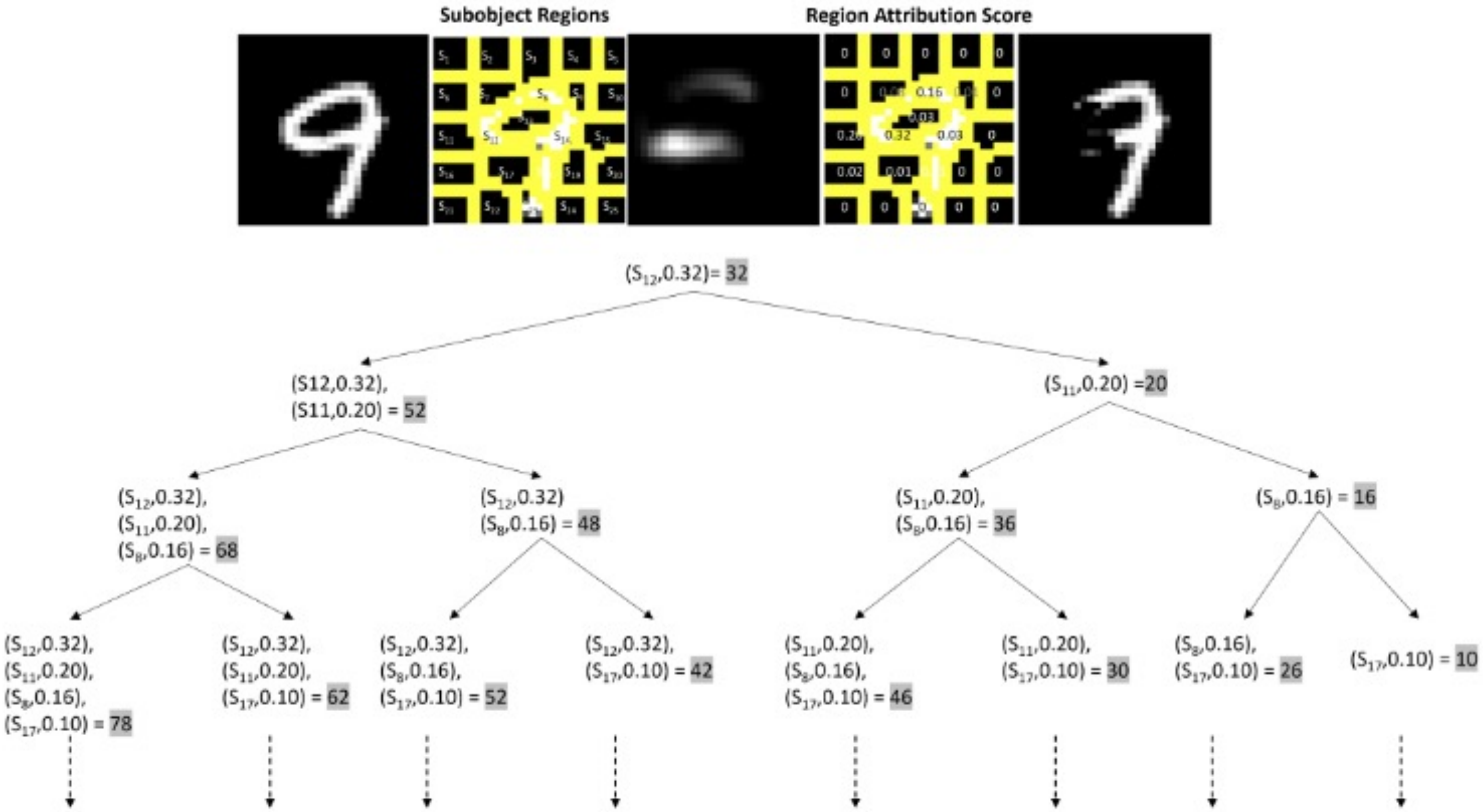
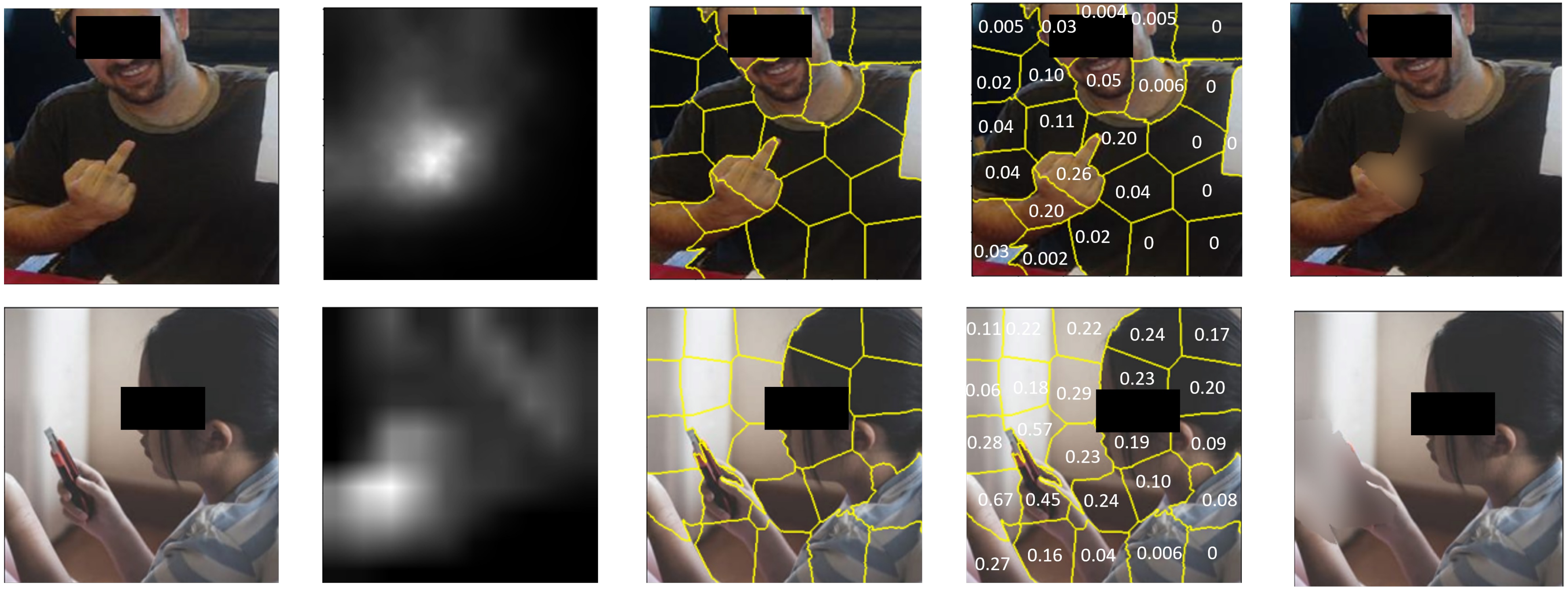


Figure 5: Visualization of the Greedy Region Search used in CSE. CSE iteratively expands the search space for the counterfactual example according to the average score in each region from the attribution map.

Qualitative Results



Original Image

Grad-CAM attribution map

SLIC Segmentation boundaries

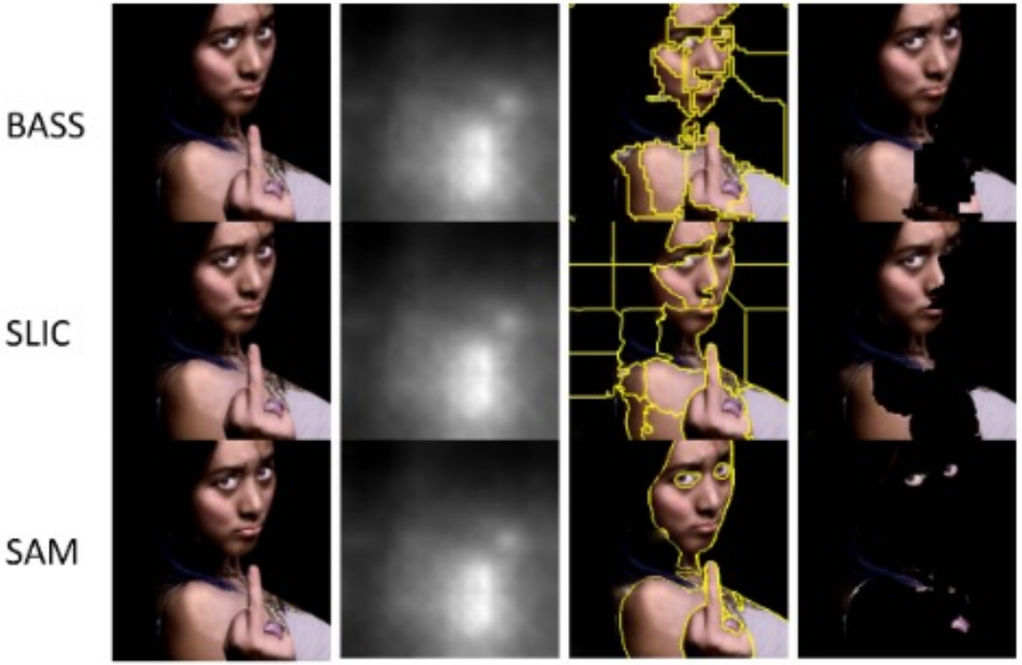
Average Grad-CAM score in each SLIC Superpixel Region

Ideal obfuscated region determined by CSRA

Quantitative Results

Dataset	Attribution Map	Counterfactual	Avg Depth	Avg Obfuscation
Sexually Explicit	**BASS**	90.6	5.8	35.0
	SLIC	76.6	7.6	33.0
	Felzenswalb	19.9	7.5	12.2
	Watershed	51.2	7.9	31.9
	SAM	29.5	7.4	33.2
Cyberbullying	**BASS**	82.0	5.2	35.2
	SLIC	60.0	6.3	25.9
	Felzenswalb	20.5	6.3	17.6
	Watershed	50.0	6.6	23.9
	SAM	50.0	6.6	40.2
Self-Harm	**BASS**	72.8	5.6	50.1
	SLIC	33.4	6.6	26.3
	Felzenswalb	38.4	6.5	47.5
	Watershed	33.1	6.8	24.6
	SAM	39.5	6.2	70.6

Quantitative results of CSE on different segmentation methods.

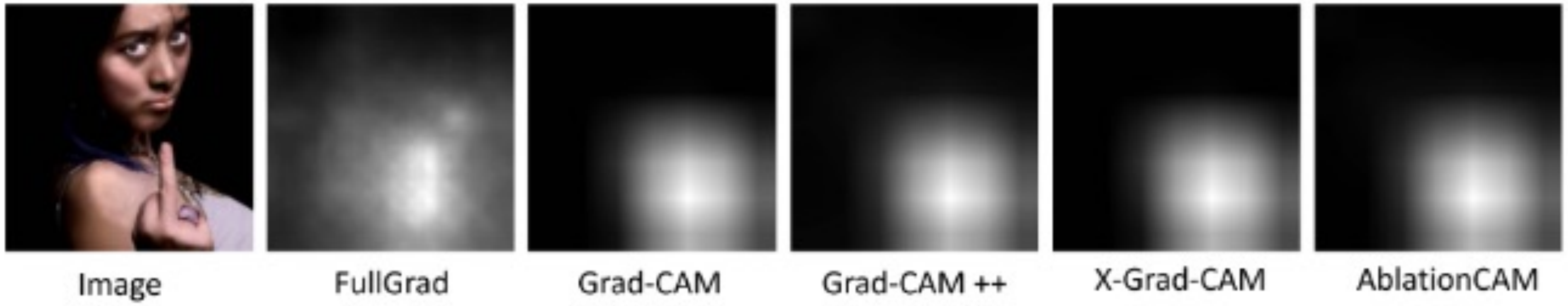


Examples of different segmentations for obfuscation on cyberbullying images

Quantitative Results

Table: Attribution maps impact study

Dataset	Attribution Map	Counterfactual	Avg Depth	Avg Obfuscation
	FullGrad	90.6	5.8	35.0
	Ablation-CAM	90.6	5.8	35.2
Sexually Explicit	Grad-CAM	90.6	5.8	35.2
	Grad-CAM++	90.6	5.8	35.2
	XGrad-CAM	90.6	5.8	35.2
	FullGrad	82.0	5.2	35.2
	Ablation-CAM	79.5	5.1	34.2
Cyberbullying	Grad-CAM	79.5	5.1	34.2
	Grad-CAM++	79.5	5.1	34.2
	XGrad-CAM	79.5	5.1	34.2
	FullGrad	72.8	5.6	50.1
	Ablation-CAM	72.8	5.6	50.1
Self-Harm	Grad-CAM	72.8	5.6	50.1
	Grad-CAM++	72.8	5.6	50.1
	XGrad-CAM	72.8	5.6	50.1



Outline

- Introduction
- Content Moderation Challenges
- Robust Semantic Representation
- Counterfactual Explanations Content Obfuscation
- **Reasoning with Conditional Vision Language Model**
- Future Directions
- Acknowledgments

Human-Machine Teaming in Content Moderation

<u>Question 4</u>: How can we use human-AI teaming to reduce moderators' exposure to harmful content?

Conditional Vision-Language Model

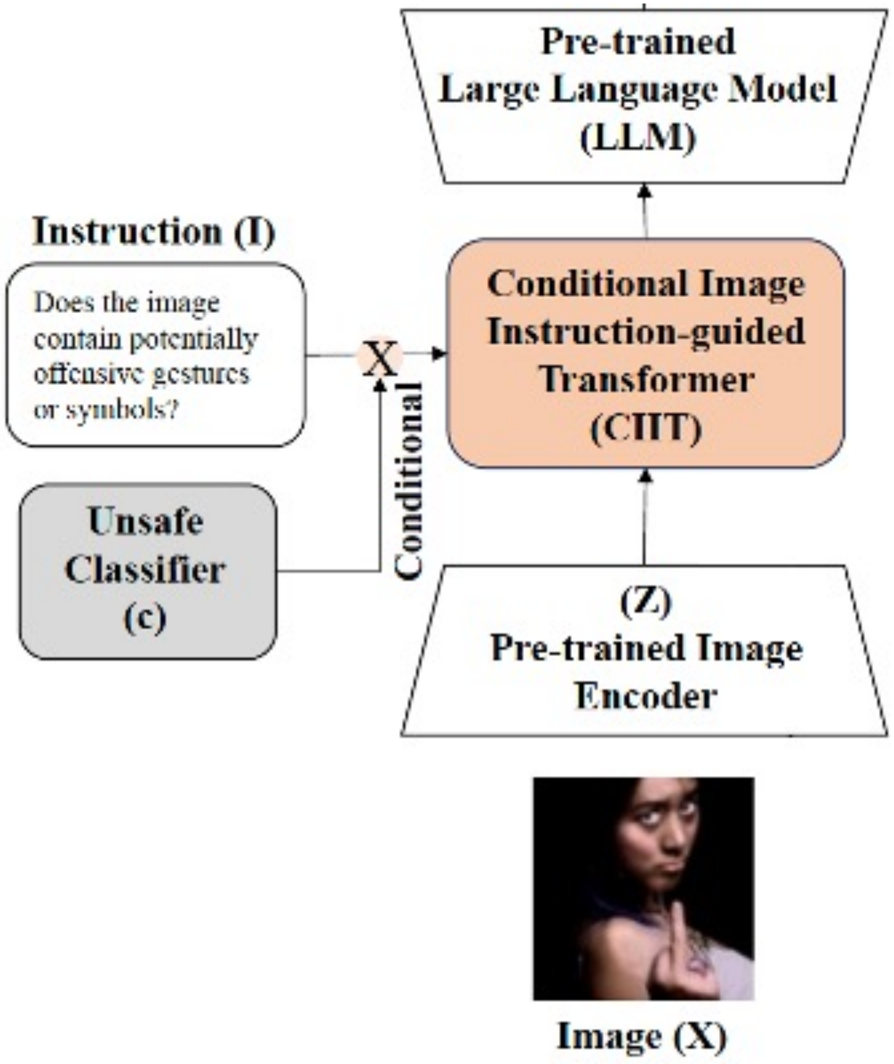
We introduce a framework that synergistically combines the strengths of large language models (LLMs) with the specific requirements of large image encoders.

A Large Pre-trained Image Encoder takes an image X as input and outputs a visual embedding representation of the image, Z = g(X).

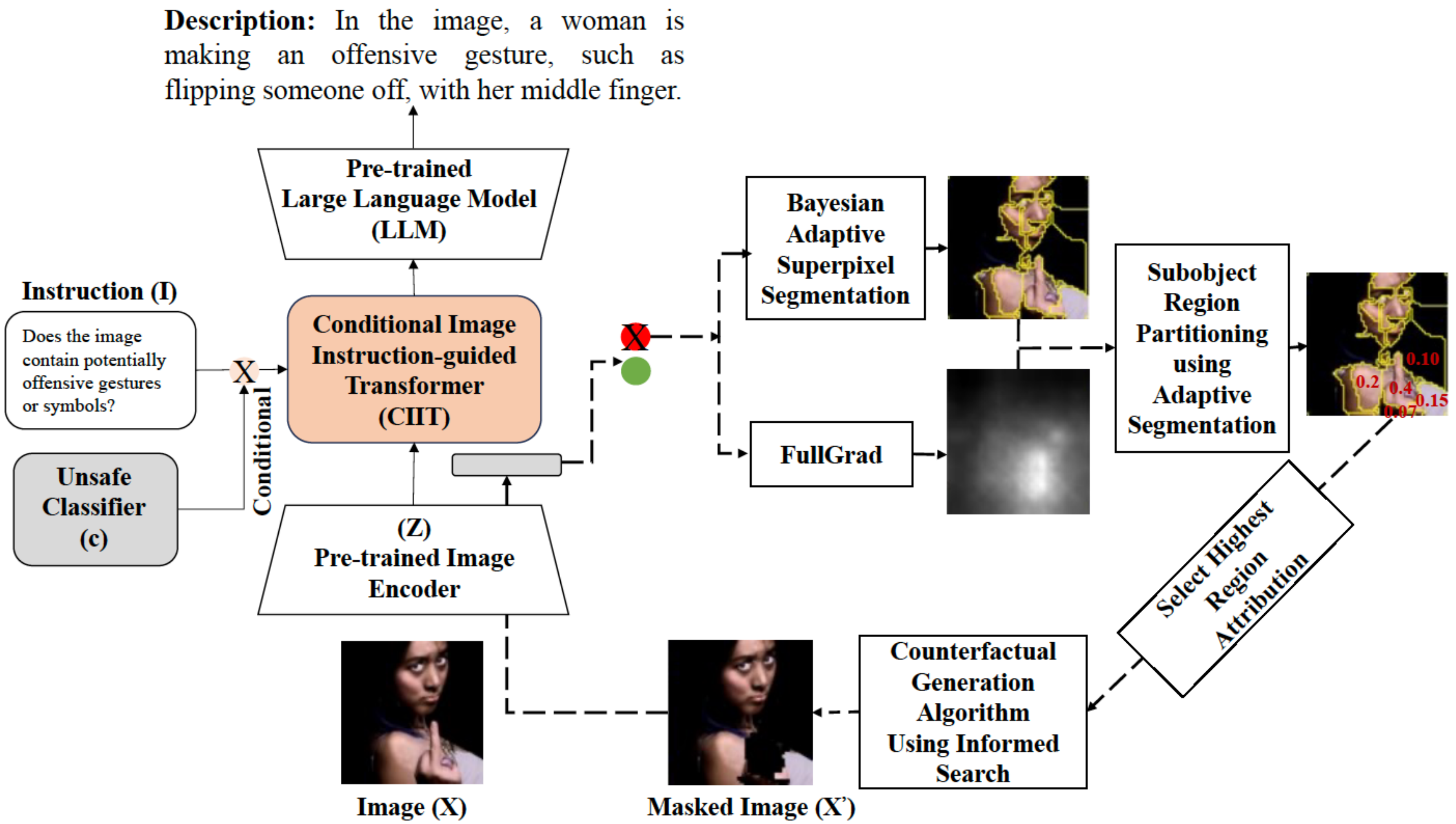
A Conditional Image Instruction-guided Transformer employs contrastive language-image pre-training to
encode visual data in congruence with a specific language prompt.

An Unsafe Classifier to condition the language prompt
using pre-trained unsafe image classifiers. This allows the
model to match and parse the unsafe visual embedding effectively

A Pre-trained Large Language Model Decoder takes a text embedding L as input and outputs linguistic sentences derived from the embedding, T ext = LLM (L).



Method



Overview of the proposed architecture. The initial module utilizes Conditional VLM for classifying images as safe or unsafe, while the subsequent module proposes counterfactual visual explanations to identify and obfuscate the unsafe regions within the image

Quantitative Results

Coarse grained questions: What is happening in the image ?

Dataset	Model	Human Evaluation
Sexually Explicit	**ConditionalBLIP**	**94.5**
	InstructBLIP	80.0
	mPLUG	73.5
	OFA-Large	13.5
Cyberbullying	**ConditionalBLIP**	**99.5**
	InstructBLIP	23.0
	mPLUG	16.0
	OFA-Large	10.0
Self-Harm	**ConditionalBLIP**	**99.0**
	InstructBLIP	47.0
	mPLUG	51.5
	OFA-Large	17.0

Quantitative results of VLM on coarse-grained image descriptions.

Quantitative Results

These fine-grained questions ask about specific attributes of images relating to the unsafe image categories

Dataset	Questions	InstructBLIP	ConditionalBLIP
Sexually Explicit	Does the image contain any sexual content or nudity?	61.1	99.2
	Does the image contain explicit sexual content?	37.9	99.9
Cyberbullying	Does the image contain potentially offensive gestures or symbols?	59.1	80.0
	Does the image contain any weapons?	99.2	94.0
	Is there any aggressive behavior shown in the image?	56.4	98.3
	Is there any text in the image that might be considered hate speech?	57.2	95.2
Self-Harm	Are there any depictions of self-harm in the image?	27.4	98.9
	Are there visible signs of self-bruising or self-harm through physical trauma?	74.2	99.2
	Does the image contain symbols associated with self-destructive behavior?	30.2	81.4

Quantitative results of VLM on fine-grained moderator questions.

Results

Image	ConditionalBLIP (Our Method)	InstructBLIP
 Cyberbullying	**Coarse-Grained Question**: What are the people in the image doing? **Answer**: In the image, a woman is making an offensive gesture, such as flipping someone off, with her middle finger. She is surrounded by other people who are also making offensive gestures or engaging in some form of cyberbullying, aggressive behavior, hateful memes, or offensive gestures. **Fine-Grained Question**: Does the image contain potentially offensive gestures or symbols? **Answer**: Yes, the image contains potentially offensive gestures or symbols. The woman in the image is making an offensive gesture with her middle finger, which can be perceived as rude or disrespectful. This type of behavior is often associated with cyberbullying, aggressive behavior, hateful memes, or offensive gestures.	**Coarse-Grained Question**: What are the people in the image doing? **Answer**: The people in the image are posing for a photograph. **Fine-Grained Question**: Does the image contain potentially offensive gestures or symbols? **Answer**: Yes, the image contains potentially offensive gestures or symbols. The woman in the image is making an obscene gesture, which could be perceived as offensive by some people.

Results

 Self-Harm	**Coarse-Grained Question**: What is happening in the image? **Answer**: In the image, there is a tic-tac-toe game being played on a person's leg. However, the tic-tac-toe game appears to have been drawn with blood, which suggests that it might be a form of self-harm or self-mutilation. This raises concerns about the individual's mental health and well-being, as self-harm can be a symptom of underlying psychological issues such as depression, anxiety, or trauma. **Fine-Grained Question**: Does the image contain symbols associated with self-destructive behavior? **Answer**: Yes, the image contains symbols associated with self-destructive behavior. There is a tic-tac-toe board drawn on the person's leg, which could be interpreted as a form of self-harm or self-destructive behavior. Tic-tac-toe is typically a harmless game, but in this context, the act of drawing it on the person's leg could be seen as a manifestation of self-harm or self-destructive behavior. It is important to note that self-harm and self-destructive behavior are serious issues that require professional help and support.	**Coarse-Grained Question**: What is happening in the image? **Answer**: In the image, there is a person sitting on a couch with a tic-tac-toe board drawn on their thigh. **Fine-Grained Question**: Does the image contain symbols associated with self-destructive behavior? **Answer**: No, the image does not contain symbols associated with self-destructive behavior. The image depicts a tic-tac-toe game being played on a person's leg, which is not indicative of self-destructive behavior.

Results

 Sexually Explicit	**Coarse-Grained Question**: What is happening in the image? **Answer**: In the image, a group of women are dancing at a bar or nightclub. Some of them are wearing minimal clothing, which suggests that they might be part of an erotic dance performance. They appear to be enjoying themselves and interacting with each other in a social setting. **Fine-Grained Question**: Does the image contain any sexual content or nudity? **Answer**: Yes, the image contains explicit sexual or pornographic content, which is not suitable for all audiences.	**Coarse-Grained Question**: What is happening in the image? **Answer**: The people in the image are dancing at a nightclub. **Fine-Grained Question**: Does the image contain any sexual content or nudity? **Answer**: No, the image does not contain any sexual content or nudity.

Future Research Directions

- Misinformation and Generative AI
  - Deciphering Textual Authenticity: Detecting Human vs. Machine-Generated Text
  - Causal Reasoning about Entities and Events - Fake and out of context detection with Zero-Shot Learning

- Safe and Fair AI thru Human Guided
  - Human-Guided AI Safety with Reinforcement Learning
  - Understanding Bias and Fairness in Machine Learning and unlearning

Acknowledgment

I would like to thank my:

- collaborators for their invaluable contributions to these research projects.
- my students for their critical thinking, hard work, and dedication

I would like to thank my grant sponsors as their funding made this work possible.