Image Safeguarding: Reasoning with Conditional Vision Language Model and Obfuscating Unsafe Content Counterfactually

Mazal Bethany[1, 2, *], **Brandon Wherry**[1, 2, *], **Nishant Vishwamitra**[1], **Peyman Najafirad**[1, 2, †]
[1]University of Texas at San Antonio
[2]Secure AI and Autonomy Lab
{mazal.bethany, brandon.wherry, nishant.vishwamitra, peyman.najafirad}@utsa.edu

Abstract

Social media platforms are being increasingly used by malicious actors to share unsafe content, such as images depicting sexual activity, cyberbullying, and self-harm. Consequently, major platforms use artificial intelligence (AI) and human moderation to obfuscate such images to make them safer. Two critical needs for obfuscating unsafe images is that an accurate rationale for obfuscating image regions must be provided, and the sensitive regions should be obfuscated (*e.g.* blurring) for users' safety. This process involves addressing two key problems: (1) the reason for obfuscating unsafe images demands the platform to provide an accurate rationale that must be grounded in unsafe image-specific attributes, and (2) the unsafe regions in the image must be minimally obfuscated while still depicting the safe regions. In this work, we address these key issues by first performing visual reasoning by designing a visual reasoning model (VLM) conditioned on pre-trained unsafe image classifiers to provide an accurate rationale grounded in unsafe image attributes, and then proposing a counterfactual explanation algorithm that minimally identifies and obfuscates unsafe regions for safe viewing, by first utilizing an unsafe image classifier attribution matrix to guide segmentation for a more optimal subregion segmentation followed by an informed greedy search to determine the minimum number of subregions required to modify the classifier's output based on attribution score. Extensive experiments on uncurated data from social networks emphasize the efficacy of our proposed method. We make our code available at: https://github.com/SecureAIAutonomyLab/ConditionalVLM

Introduction

Social media is being increasingly misused by bad actors to share sexually explicit, cyberbullying, and self-harm content (Hendricks 2021; Chelmis and Yao 2019; Adler and Chenoa Cooper 2022). However, social media platforms are required by law to safeguard their users against such images (Exon 1996), as well as provide a rationale for why such images are flagged (Cabral et al. 2021) for the purpose of transparency. In response, major platforms have deployed AI and human-based content moderation techniques to flag and obfuscate (*i.e*, make the image safer by blurring sensitive regions) such images (Bethany et al. 2023). This process involves obfuscating (*e.g.* by blurring or blocking) unsafe image regions in the image (Li et al. 2017) along with generating a rationale that backs up the decision to obfuscate the flagged images (Meta 2022).

The image obfuscation process faces two critical problems regarding how much of the unsafe image is obfuscated and why it is obfuscated: *First*, the decision to deem an image unsafe and obfuscate it demands providing a rationale for the decision. For example, Instagram moderators are required to provide a legal rationale (Bronstein 2021; Are 2020) to back up their decision (Tenbarge 2023). Existing visual reasoning methods (Li et al. 2022, 2023; Dai et al. 2023) are severely limited for unsafe images such as sexually explicit, cyberbullying, and self-harm since they cannot provide a rationale grounded in attributes that are specific to such images, such as rude hand gestures in cyberbullying images (Vishwamitra et al. 2021), or sensitive body parts in sexually-explicit images (Binder 2019). *Second*, the unsafe image needs minimal obfuscation while still depicting the safe regions for evidence collection and investigation (Billy Perrigo 2019). For instance, human moderators need to determine the age of the person in the image (*e.g.*, in child sexual abuse material (CSAM) investigations), look for identifiers (*e.g.*, tattoos, scars, and unique birthmarks), and determine their location information (*e.g.*, landmarks, geographical features, and recognizable surroundings). Current segmentation techniques (Chandrasekaran et al. 2021; Vermeire et al. 2022; Bethany et al. 2023) cannot minimally identify the regions and consequently impede investigations that pertinently need full details of the remaining safe regions.

In this work, we take the first step towards addressing a pertinent, but overlooked problem of the *image moderation process* in social media platforms. Our major objective is to first identify and minimally obfuscate the sensitive regions in an unsafe image such that the safe regions are unaltered to aid an investigation, and then provide an accurate rationale for doing so, that is grounded in unsafe image attributes (*e.g.*, private body parts, rude gestures or hateful symbols). To this end, we address this problem in two steps: (1) we develop a novel unsafe image rationale generation method called ConditionalVLM (*i.e.*, conditional vision language model) that leverages the state-of-the-art large lan-

[*] These authors contributed equally.
[†] Corresponding author

guage models (LLM)-based vision language models (Fang et al. 2023) to perform an in-depth conditional inspection to generate an accurate rationale that is grounded in unsafe image attributes; and (2) minimally obfuscating the sensitive regions *only* by calculating the classifier attribution matrix using a FullGrad-based model (Srinivas and Fleuret 2019) and then utilize this information to guide Bayesian superpixel segmentation (Uziel, Ronen, and Freifeld 2019) for a more informed and optimal dynamic subregion segmentation, via calculating the attribution score of each subregion. Finally, we utilize a discrete optimization technique such as informed greedy search to determine the minimum number of subregions required to modify the classifier's output, using the score attribution.

Our work has profound implications for the safety of social media content moderators, by greatly reducing their need to view unsafe content (Steiger et al. 2021), social media users who are minors or sensitive to such content (Hargrave and Livingstone 2009), and law enforcement agents who need to investigate such images as part of their investigation (Krause 2009). We make the following contributions:

- We develop ConditionalVLM, a visual reasoning model that generates accurate rationales for unsafe images by leveraging state-of-the-art VLMs conditioned on pre-trained unsafe image classifiers.
- We develop a novel unsafe image content obfuscation algorithm that minimally obfuscates only the unsafe regions while keeping the rest of the image unaltered for investigations.
- Evaluations of our work show that it can categorize the three social media unsafe categories of images with an accuracy of 93.9%, and minimally segment only the unsafe regions with an accuracy of 81.8%.

Related Works

Safeguarding Images

Social media platforms are frequently misused for sharing various forms of unsafe content, including sexually-explicit images (Ashurst and McAlinden 2015; Sanchez et al. 2019), non-consensual intimate images (NCII)(Lenhart, Ybarra, and Price-Feeney 2016), and child sexual abuse material (CSAM)(Sanchez et al. 2019). These platforms also contribute to the spread of cyberbullying (Vishwamitra et al. 2021) and self-harm images, which pose significant risks (John et al. 2018). The traditional blurring approach in image moderation has wide-ranging implications. Over a million global moderators face mental health risks from viewing such content (bbc 2021; reu 2021). Additionally, minors require image safeguarding to shield them from exposure to harmful content, while law enforcement agents need it for analyzing crime scene images with minimal obfuscation to preserve crucial investigative details.

Vision-Language Models

Pre-trained models in computer vision (CV) and natural language processing (NLP) have led to the development of large-scale Vision-Language Models (VLMs). Methods like CLIP (Radford et al. 2021) and BEIT-3 (Wang et al. 2023) integrate image-text pairs, with CLIP using contrastive training and BEIT-3 employing multiway transformers for masked modeling. Modular approaches also exist, leveraging established models for image and text interpretation. However, these models face challenges in effectively coordinating visual and textual features. For instance, Flamingo (Alayrac et al. 2022) and BLIP-2 (Li et al. 2023) address this by adding cross attention layers or querying transformers, while LENS (Berrios et al. 2023) develops visual vocabularies without additional training. A common limitation is the lack of conditioning capability, crucial for domain-specific attributes (Ramesh et al. 2022).

Image Segmentation and Counterfactual Explanation for Obfuscation

Another type of explanation that is growing in popularity due to its ability to address several of these issues is counterfactual explanations (Wachter, Mittelstadt, and Russell 2017). A counterfactual explanation can be defined as taking the form: a decision y was produced because variable X had values $(v1, v2, \dots)$ associated with it. If X instead had values $(v1', v2', \dots)$, and all other variables had remained constant, score y' would have been produced. Some works such as BEN (Chandrasekaran et al. 2021), SEDC (Vermeire et al. 2022), and CSRA (Bethany et al. 2023) have explored region-based counterfactual visual explanations. However, existing approaches face two key challenges: 1. suboptimal subregion boundaries, leading to excessive parts of the image being identified as causing a decision, and 2. high time complexity 2^K in searching for a counterfactual in an image of K regions. BEN and SEDC segment an input image into K static subregions without any prior knowledge of the classifier, resulting in an uninformed search strategy for finding the counterfactual examples. While CSRA does use prior knowledge of the classifier to inform the search of the counterfactual example, BEN, SEDC and CSRA do not jointly optimize the subregions boundaries and minimize the number of subregions, which is particularly important for obfuscation applications where preserving as much context as possible is preferred.

Method

Figure 1 illustrates the architecture of our proposed approach, which consists of two modules. The initial module proposes a conditional visual language model designed for image reasoning. The model classifies images as safe or unsafe by understanding the interactions or activities of entities within the image, using its comprehension of visual features and linguistic annotations. In the subsequent module, counterfactual visual explanations are proposed to precisely identify sub-object regions of the image contributing to its unsafe classification for obfuscation.

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. Additionally,

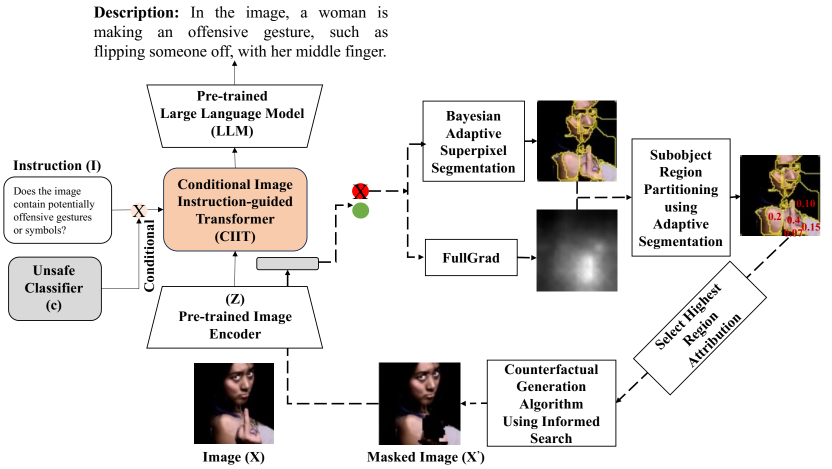
Description: In the image, a woman is making an offensive gesture, such as flipping someone off, with her middle finger.
Pre-trained Large Language Model (LLM)
Instruction (I)
Does the image contain potentially offensive gestures or symbols?
X
Conditional Image Instruction-guided Transformer (CIIT)
Conditional
Unsafe Classifier (c)
(Z) Pre-trained Image Encoder
Bayesian Adaptive Superpixel Segmentation
FullGrad
Subobject Region Partitioning using Adaptive Segmentation
0.10
0.2
0.4
0.15
0.07
Select Highest Region Attribution
Counterfactual Generation Algorithm Using Informed Search
Image (X)
Masked Image (X')

Figure 1: Overview of the proposed architecture. The initial module utilizes ConditionalVLM 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.

it provides more explicit control over visual features being reasoned. The ConditianalVLM architecture is anchored by three pivotal components, as depicted in Figure 1:

A Large Pre-trained Image Encoder takes an image X as input and outputs a visual embedding representation of the image, $Z = g(X)$. We explore a state-of-the-art pre-trained vision transformer ViT-g/14 from EVA (Fang et al. 2023).

A Conditional Image Instruction-guided Transformer (CIIT) employs contrastive language-image pre-training to encode visual data in congruence with a specific language prompt. Additionally, we condition this language prompt using pre-trained unsafe image classifiers. This allows the model to match and parse the unsafe visual embedding effectively, while also providing more explicit control over unsafe visual features (Ramesh et al. 2022). CIIT utilizes a pre-trained Q-Former model (Li et al. 2023), which is conditioned on image classifiers as control code c on unsafe image content such as sexually explicit, cyberbullying, and self-harm.

- A prior $p(I|c)$ that produces CIIT instruct prompt I conditioned on control code c.
- A transformer decoder $p(L|I, c)$ that produces contrastive embedding L conditioned on Instruct prompt I and control code c.

The transformer decoder allows us to invert images given their CIIT Instruct prompt, while the prior allows us to learn a generative model of the image embeddings themselves. Taking the product of these two components yields a generative model $P(L|c)$ of embedding L given control c:

$$p(L|c) = p(L, I|c) = p(L|I, c)p(I|c) \tag{1}$$

The control code c provides a point of control over the CIIT generation process. The distribution can be decomposed using the chain rule of probability and trained with a loss that takes the control code into account.

$$p(L|c) = \prod_{i=1}^{n} p(L_i|L_{<i}, c) \tag{2}$$

We train the model with parameters θ to minimize the negative log-likelihood over a dataset $D = X_1, ..., X_n$:

$$\mathcal{L}(D) = -\sum_{k=1}^{|D|} \log p_\theta(L_i^k|L_{<i}, c^k) \tag{3}$$

A Pre-trained Large Language Model Decoder takes a text embedding L as input and outputs linguistic sentences derived from the embedding, $Text = LLM(L)$. We choose Vicuna (Vic 2023) as our LLM decoder which is constructed upon LLaMA (Touvron et al. 2023) and can perform a wide range of complex linguistic tasks.

Counterfactual Subobject Explanations for Obfuscation

In order to connect region attribution to provide counterfactual subobject region explanation of an image, relative to a given machine learning predictive model, we propose a two-phase approach. The pipeline of the proposed approach is illustrated in Figure 1. We first partition the image into non-intersecting subobject regions and measuring region attribution value to each region using gradient attribution maps in Section 3.1 and Section 3.2. The counterfactual analysis of alternate versions of the image using a greedy search algorithm using regions with highest attribution values for counterfactual analysis is followed in Section 3.3.

Subobject Region Partitioning using Adaptive Segmentation. We represent a given image,Xas a non-intersecting set of K regions given by $\{z_1, z_2, \cdots, z_K\}$. The boundaries of these regions are defined by clustering algorithms that use color and spatial information and are called superpixels. Let Z represent the K region segmentation, z_i represent the label assigned to X_i and j represent the label of some arbitrary cluster. An image must be segmented into meaningful subobject regions in order to allow for a counterfactual analysis of the image by the binary predictive model $f(X) \rightarrow 0, 1$. These regions serve as the features that are analyzed in the counterfactual analysis. To maximize the efficiency of a counterfactual analysis, we require an adaptive segmentation method. Many segmentation methods are wasteful in their assignment of many segments to uninformative regions, while not segmenting detailed regions enough. Such a method should be able to respect pixel connectivity and spatial coherence and requires an adaptive number of regions. K-means based clustering methods are a fast and simple basis for leading segmentation, however Gaussian Mixture Models (GMM) may be better suited for an adaptive segmentation method since we need to capture the heterogeneity in the pixel distribution of various types of images.

Let $N = h * w$ be the number of pixels in an image, X with c color channels. The values attributed to the pixels in X can be denoted as $X_i = (l_i, c_i) \in \mathbb{R}^5$, where $l_i \in \mathbb{R}^2$ represent the x, y coordinate location and $c_i \in \mathbb{R}^3$ represent the RGB color information. Superpixel clustering methods with spatial coherence aim to partition $(X_i)_{i=1}^N$ into K disjoint groups. Let Z represent the K region segmentation, z_i represent the label assigned to X_i and j represent the label of some arbitrary cluster. Where $\mathcal{N}(X; \mu_j, \Sigma_j)$ is a Gaussian PDF with mean μ_j and a covariance matrix Σ_j of size $n * n$, the PDF of a GMM with K components is

$$p(X; (\mu_j, \Sigma_j, \lambda_j)_{j=1}^K) = \sum_{j=1}^{K} \lambda_j \mathcal{N}(X|\mu_j, \Sigma_j)$$

The mixing coefficients λ_j in the PDF of a GMM form a convex combination where:

$$\sum_{j=1}^{K} \lambda_j = 1, \lambda_j \geq 0 \quad \forall j$$

and this allows for a globally optimal clustering. Given a Gaussian distribution j where $\theta_j = (\mu_j, \Sigma_j)$, a Bayesian GMM with random variables $(\theta_j)_{j=1}^K$ and $(\lambda_j)_{j=1}^K$ are drawn from $p((\theta_j, \lambda_j)_{j=1}^K)$, a prior distribution. Assuming independence, the prior distribution can be factorized as follows

$$p((\theta_j, \lambda_j)_{j=1}^K) = p((\lambda_j)_{j=1}^K) \prod_{j=1}^{K} p(\theta_j)$$

Using a Normal-Inverse Wishart (NIW) for $p(\theta_j)$ and a Dirichlet distribution for $p((\lambda_j)_{j=1}^K)$ gives us posterior distributions in the same form as the priors. Furthermore, the updates from the priors are given in closed form.

The Bayesian GMM inference to calculate Z can be done by performing Gibbs sampling, alternating between the following equations:

$$p((\lambda_j)_{j=1}^K | Z, (X_i)_{i=1}^N) \prod_{j=1}^{K} p((\theta_j, \lambda_j) | Z, (X_i)_{i=1}^N)$$

$$p(Z | (\theta_j, \lambda_j)_{j=1}^K, (X_i)_{i=1}^N)$$

Subobject Region Attribution Value. We start by creating the FullGrad (Srinivas and Fleuret 2019) attribution map for image feature attribution. Given an image X and the feature maps generated by the FullGrad $L[u, v]$ of width u and height v for the model prediction, the goal of the visual attention model is to identify the discriminative regions of the image that significantly influence the class prediction score of the predictive model using $L[u, v]$ pixel attribution values.

The attribution map of the FullGrad method is generated by propagating an image through a CNN, obtaining the output score before the softmax layer, and then computing the gradients with respect to the input (input-gradients) and the biases at each layer (bias-gradients). These gradients are then combined, with each bias-gradient reshaped to match the input dimensionality and all gradients summed to form the FullGrad attribution map.

FullGrad Definition: *Consider a CNN model f, with x denoting the input and b denoting the biases at each layer, c representing the channels of layer k. Furthermore, given an output of interest f(x), and a postprocessing operator $\psi(\cdot)$ the FullGrad attribution map $L_{FullGrad}$ is defined as:*

$$L_{FullGrad} = \psi(\nabla_x f(x) \odot x) + \sum_{k \in K} \sum_{c \in c_k} \psi(f^b(x)_c)$$

To facilitate an efficient sampling of regions in the counterfactual analysis, we utilize the FullGrad attribution map.

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 subobject region attribution score, $\{s_1, s_2, \cdots, s_K\}$ as follows:*

$$s_k = \frac{1}{n.m} \sum_{n} \sum_{m} L_{FullGrad(F,X)}[i, j], X[i, j] \in z_k$$

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.

Definition 2: (Subobject Region Confidence Reduction) *Given a model $Y = f(X)$ that takes an image X with subobject regions $X = [z_0, z_1, ..., z_n]^T$ and outputs a probability distribution Y. The confidence reduction cr_k of subobject region z_k, ($k \in [1, n]$) towards Y is the change of the output by masking the k-th subobject region of X, while being classified as the same class, as follows:*

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

In Sec 3.3, we present our greedy region search algorithm which utilizes subobject region attribution score as heuristics and employs confidence level for causal obfuscation using counterfactual subobject region explanations.

Counterfactual Generation Using Informed Subobject Region Search. The previous sections lead us to the minimum region masking problem. This can be computationally expensive to solve, as it requires the masking and analysis of 2^K different combination of regions, Z of X based on Section 3.1 . Rather than solving the problem directly, we find an approximate solution using a greedy region search.

Given a predictive model $f : X \rightarrow \{0, 1\}$, we can define the set of counterfactual explanations for an input $x \in X$ as x' while $\arg\min_{x'} d(x, x')$ and $x' = \{x \in X \mid f(x) \neq f(x')\}$. In other words, $x' = \{x \in X \mid f(x) \neq f(x')\}$ contains all the inputs x for which the model f returns a prediction different from $f(x)$ while minimizing the distance between x and x'.

Our greedy region search, starts with us first sorting the K regions in descending order by the average attribution for each region which were calculated in subsection . 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 $f(x') \neq f(x)$.

Experimental Evaluation

Datasets

We evaluated our Conditional VLM and Counterfactual Subobject Explanation methods on three datasets of real-world harmful images to study the practical application of counterfactual subobject explanations.

Sexually Explicit: First, we sampled a subset of images from an NSFW images dataset (Kim 2021) consisting of 334,327 images by selecting the “porn”, “neutral”, and “sexy” classes. We combine the “neutral”, and “sexy” classes into a single class of “safe” images. In the proceeding experiments, this dataset is denoted as SE.

Cyberbullying: Second, we used a cyberbullying images (Vishwamitra et al. 2021) dataset consisting of nearly 20,000 images belonging to the classes “cyberbullying” and “non-cyberbullying”. In the proceeding experiments, this dataset is denoted as CB.

Self-Harm: Third, we used a self-harm images dataset (Bethany et al. 2023), consisting of 5000 images with classes “self-harm” and “non self-harm”. In the proceeding experiments, this dataset is denoted as SH.

Evaluation Settings

ConditionalVLM. We compare our proposed method against other state-of-the-art image-to-text models such as InstructBLIP (Dai et al. 2023), OFA-Large (Wang et al. 2022), and mPLUG (Li et al. 2022). We use the implementations of these methods from HuggingFace. For InstructBLIP, we use InstructBLIP-Vicuna-13b with num_beams=5, max_length=512, min_length=1, top_p=0.9, repetition_penalty=1.5, length_penalty=1.0, and temperature=1. The image encoder for this implementation of InstructBLIP was Vit-g/14 (Fang et al. 2023). For mPLUG, we use the parameters do_sample=True, top_k=5, and max_length=512. For OFA we use the parameters of num_beams=5, no_repeat_ngram_size=3. To demonstrate our ConditionalVLM framework, we modify the InstructBLIP-Vicuna-13b architecture to include a CIIT, which we call ConditionalBLIP. All experiments were carried out on a DGX 8x A100 GPU, with 80GB of VRAM each.

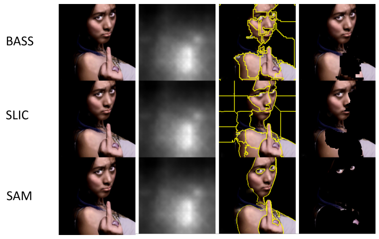

Figure 2: Examples of segmentation methods on a cyberbullying image. From top to bottom: (1) BASS, (2) SLIC, (3) SAM.

We fine-tuned a ResNet-50 classifier available in Pytorch (Paszke et al. 2019) using pre-trained model weights trained from the ImageNet dataset (Deng et al. 2009). The NSFW, cyberbullying and self-harm datasets were each divided into train, validation, and test sets, with 80% being allocated to the train set, and 10% each allocated to validation and test sets. We trained the models for 50 epochs and selected the models that have the highest classification accuracies on the validation sets. These models achieved accuracies of 98.9%, 91.9% and 97.6% respectively on the test set in our experiments. We use these classifiers as the control code for the CIIT in ConditionalBLIP.

Counterfactual Subobject Explanations for Obfuscation. To test different segmentation methods, we experimented with SLIC (Achanta et al. 2010), Felzenszwalb (Felzenszwalb and Huttenlocher 2004), and Compact Watershed (Neubert and Protzel 2014) segmentation methods implemented in the scikit-image library (van der Walt et al. 2014), Segment Anything Model (SAM) (Kirillov et al. 2023), and Bayesian Adaptive Superpixel Segmentation (Uziel, Ronen, and Freifeld 2019). For our experiments, we selected the following parameters for each segmentation method: for SLIC, we chose the number of segments to be 25 and compactness equal to 1; for Felzenszwalb we selected the scale to be 500,

Data	Model	Human Evaluation
SE	**ConditionalBLIP**	**94.5**
	InstructBLIP	80.0
	mPLUG	73.5
	OFA-Large	13.5
CB	**ConditionalBLIP**	**99.5**
	InstructBLIP	23.0
	mPLUG	16.0
	OFA-Large	10.0
SH	**ConditionalBLIP**	**99.0**
	InstructBLIP	47.0
	mPLUG	51.5
	OFA-Large	17.0

Table 1: Quantitative results of VLM on coarse-grained image descriptions.

Data	Questions	InstructBLIP	ConditionalBLIP
SE	Does the image contain any sexual content or nudity?	61.1	99.2
	Does the image contain explicit sexual content?	37.9	99.9
CB	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
SH	Are there any depictions of self-harm in the image?	27.4	98.9
	Are there visible signs of self-harm through physical trauma?	74.2	99.2
	Does the image contain symbols associated with self-destructive behavior?	30.2	81.4

Table 2: Quantitative results of VLM on fine-grained moderator questions.

sigma to be 0.5, and a minimum component size of 200; for Compact Watershed, we chose the number of markers to be 25 and the compactness parameter to be 0.001.

We used the following attribution map methods in our experiments: (Grad-CAM (Selvaraju et al. 2017), XGrad-CAM (Fu et al. 2020), Grad-CAM ++ (Chattopadhay et al. 2018), FullGrad (Srinivas and Fleuret 2019), and Ablation-CAM (Ramaswamy et al. 2020)). For the implementation of the attribution map methods, we use the Pytorch Grad-CAM library (Gildenblat and contributors 2021).

Evaluation Metrics

ConditionalVLM. We evaluate VLM's ability to investigate three different unsafe image categories in two phases. In the first phase, we conduct a coarse-grained evaluation by having human evaluators determine based off of the image descriptions produced by the VLM whether a moderator should be able to understand which dataset of unsafe image the image belongs to. In this evaluation, a team of three human evaluators who were involved in this research were asked to evaluate whether these descriptions produced by the VLM on the questions of ”What is happening in the image?”, and ”What are the people doing?” were sufficient to accurately categorize them into the correct dataset that the unsafe image the image belongs to. The final labels were assigned by majority voting.

In the second phase, we conduct a fine-grained evaluation by having human evaluators evaluate the responses of the VLM to curated moderator questions with respect to an unsafe image image. These fine-grained questions ask about specific attributes of images relating to the unsafe image categories. In this evaluation, the same team of evaluators were asked to determine whether the answers produced by the VLM correctly answered these curated questions.

Counterfactual Subobject Explanations for Obfuscation. We investigate the ability of CSE to generate a successful counterfactual explanation on an unsafe image X to satisfy two requirements: (1) the generated counterfactual example X' must be a convincing representation of another class such that it has a softmax score greater than a threshold T on another class, and (2) the search space that the counterfactual example X' exists in must be found by searching N or fewer different regions. Since, there are 2^K different combinations of regions to be analyzed in X with K number of regions, we limit the search space to a certain number of regions in our evaluation. In our experiments on unsafe images, we select the threshold for softmax score T to be 0.5 and the threshold for regions to be 10.

Results and Discussion

ConditionalVLM. The results for the coarse-grained evaluations of the VLM are shown in Table 1. In this table, we present the accuracy of four models, including our model, ConditionalBLIP, that convert images to text, specifically focusing on their ability to identify unsafe attributes in images based on generic questions. In this experiment, a total of 2000 unsafe image samples from each category of unsafe image datasets were tested. The results show that ConditionalBLIP is able to significantly outperform other state-of-the-art models in identifying the unsafe image attributes of unsafe images, simply from asking generic questions on the image, with an average correct identification accuracy of 98% of unsafe image attributes across the three datasets. Compared to the 50% accuracy by InstructBLIP, 47% by mPLUG, and 13.5% by OFA-Large, we observe that existing models are insufficient for describing unsafe images.

We present the questions and quantitative results of the fine-grained evaluation of ConditionalBLIP in Table 2. We compare ConditionalBLIP against InstructBLIP, which showed the best coarse-grained results compared to other methods that were evaluated in Table 1. Furthermore, the InstructBLIP model is the most similar in implementation to the ConditionalBLIP model, where the primary difference is the usage of the CIIT in ConditionalBLIP. In Table 2, we present the question posed to the VLM, alongside the detection accuracy of InstructBLIP and ConditionalBLIP on these questions. The fine-grained evaluation shows that image conditioning significantly enhances VLMs ability to understand unsafe images, with an average improvement in accuracy of 38.2% across the questions. The comparison between the performances of InstructBLIP and ConditionalBLIP reveals significant differences in their respective abilities to identify and describe unsafe content in visual data. By employing contrastive language-image pre-training and conditioning the language prompt using pre-trained unsafe image classifiers, ConditionalBLIP is able to parse the unsafe visual embedding effectively.

Counterfactual Subobject Explanations for Obfuscation. For the counterfactual image obfuscation experiments, we test on 585 sexually explicit, cyberbullying and self-harm images. We compare our method against the CSRA method, setting numROI = 10 to match time complexity. Previous work showed gradient-based attribution maps were unsuitable for obfuscating unsafe images (Bethany et al. 2023). Our trained models show improvements of 13.9% on sexually explicit, 22.0% on cyberbullying, and 39.5% on self-harm images when comparing CSRA vs CSE.

We tested various attribution map methods with BASS (Uziel, Ronen, and Freifeld 2019) as the constant segmentation method on unsafe image samples, with results in Table 3. The average search space required to find a counterfactual example was presented, showing that different attribution map methods do not significantly impact CSE, with most generating similar highest average attribution scores in similar areas. The exception was the FullGrad method, which provided slightly more successful counterfactual examples, better average search space, and fewer obfuscated regions. This can be attributed to FullGrad's more dispersed attributions across the image, which does not restrict the search space as much, and its unique method of satisfying local and global importance by aggregating information from both input-gradient and intermediate bias-gradients, thus aiding CSE in finding suitable counterfactual explanations more readily.

We tested different segmentation methods with FullGrad as the constant attribution map method on unsafe image samples, and the results are in Table 4. The choice of segmentation method significantly impacted the number of successful counterfactual explanations, average search space, and average number of regions obfuscated. BASS was the most effective, with a combination of BASS and FullGrad yielding 81.8% successful counterfactual examples, a search depth of 5.5, and an average of 40.1% of the image obfuscated. The segmentation's effect on counterfactual examples can be seen in Figure 2, and as Table 4 showed, methods like BASS are key for successful counterfactual explanations, as they break the image into non-intersecting, color and spatially coherent subobjects.

Data	Attr Map	CF	Avg Depth	Avg Obf
SE	**FullGrad**	90.6	5.8	35.0
	Ablation-CAM	90.6	5.8	35.2
	Grad-CAM	90.6	5.8	35.2
	Grad-CAM++	90.6	5.8	35.2
	XGrad-CAM	90.6	5.8	35.2
CB	**FullGrad**	82.0	5.2	35.2
	Ablation-CAM	79.5	5.1	34.2
	Grad-CAM	79.5	5.1	34.2
	Grad-CAM++	79.5	5.1	34.2
	XGrad-CAM	79.5	5.1	34.2
SH	**FullGrad**	72.8	5.6	50.1
	Ablation-CAM	72.8	5.6	50.1
	Grad-CAM	72.8	5.6	50.1
	Grad-CAM++	72.8	5.6	50.1
	XGrad-CAM	72.8	5.6	50.1

Table 3: Quantitative results of CSE using different attribution map methods.

Data	Segmentation	CF	Avg Depth	Avg Obf
SE	**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
CB	**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
SH	**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

Table 4: Quantitative results of CSE on different segmentation methods.

Ablation Study. To evaluate our vision-language model's conditioning, we conducted an ablation study by changing unsafe classifier guidance on the Image Instruction-guided Transformer or CIIT model's instruct prompt embedding from 1 to 0. This conditioning on zero-shot instruct embeddings yielded acceptable results for unsafe images by allowing CIIT to match and parse the unsafe visual embedding effectively, while also providing more explicit control over unsafe visual feature correlation with conditioned instruct prompt. For instance, the LLM decoder's output for an unsafe image changed to suggest "women are performing potential erotic dance in a bar" vs. "women dancing in a bar". These results suggest that conditioning is a promising approach for vision language models.

Conclusion

In this work, we have presented ConditionalVLM, a visual reasoning framework that generates accurate rationales for unsafe image descriptions by leveraging state-of-the-art VLMs conditioned on pre-trained unsafe image classifiers, and CSE, a counterfactual visual explanation technique to obfuscate the unsafe regions in unsafe images for safer sharing. We evaluated these two methods on three categories of unsafe images. An implementation of ConditionalVLM, which we called ConditionalBLIP showed superior performance compared to other state-of-the-art image-to-text models on describing unsafe images. We also compare CSE against another recent unsafe image obfuscation method and show how our approach is effective in generating causal explanations for obfuscating unsafe images.

Acknowledgments

This research project and the preparation of this publication were funded in part by the Department of Homeland Security (DHS), United States Secret Service, National Computer Forensics Institute (NCFI) via contract number 70US0920D70090004 and by NSF Grant No. 2245983.

References

2021. Facebook moderator: 'Every day was a nightmare'. https://www.bbc.com/news/technology-57088382. Accessed: July 14, 2023.

2021. Judge OKs $85 mln settlement of Facebook moderators' PTSD claims. https://www.reuters.com/legal/transactional/judge-oks-85-mln-settlement-facebook-moderators-ptsd-claims-2021-07-23/. Accessed: July 20, 2023.

2023. Vicuna. https://github.com/lm-sys/FastChat. Accessed: July 17, 2023.

Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; and Süsstrunk, S. 2010. Slic superpixels. Technical report.

Adler, R. A.; and Chenoa Cooper, S. 2022. "When a Tornado Hits Your Life:" Exploring Cyber Sexual Abuse Survivors' Perspectives on Recovery. *Journal of Counseling Sexology & Sexual Wellness: Research, Practice, and Education*, 4(1): 1–8.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Are, C. 2020. How Instagram's algorithm is censoring women and vulnerable users but helping online abusers. *Feminist media studies*, 20(5): 741–744.

Ashurst, L.; and McAlinden, A.-M. 2015. Young people, peer-to-peer grooming and sexual offending: Understanding and responding to harmful sexual behaviour within a social media society. *Probation Journal*, 62(4): 374–388.

Berrios, W.; Mittal, G.; Thrush, T.; Kiela, D.; and Singh, A. 2023. Towards Language Models That Can See: Computer Vision Through the LENS of Natural Language. *arXiv preprint arXiv:2306.16410.*

Bethany, M.; Seong, A.; Silva, S. H.; Beebe, N.; Vishwamitra, N.; and Najafirad, P. 2023. Towards targeted obfuscation of adversarial unsafe images using reconstruction and counterfactual super region attribution explainability. In *32nd USENIX Security Symposium (USENIX Security 23)*, 643–660.

Billy Perrigo. 2019. Facebook Says It's Removing More Hate Speech Than Ever Before. But There's a Catch.

Binder, M. 2019. Facebook claims its new AI technology can automatically detect revenge porn. https://mashable.com/article/facebook-ai-tool-revenge-porn. Accessed: July 17, 2023.

Bronstein, C. 2021. Deplatforming sexual speech in the age of FOSTA/SESTA. *Porn Studies*, 8(4): 367–380.

Cabral, L.; Haucap, J.; Parker, G.; Petropoulos, G.; Valletti, T. M.; and Van Alstyne, M. W. 2021. The EU digital markets act: a report from a panel of economic experts. *Cabral, L., Haucap, J., Parker, G., Petropoulos, G., Valletti, T., and Van Alstyne, M., The EU Digital Markets Act, Publications Office of the European Union, Luxembourg.*

Chandrasekaran, J.; Lei, Y.; Kacker, R.; and Kuhn, D. R. 2021. A combinatorial approach to explaining image classifiers. In *2021 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, 35–43. IEEE.

Chattopadhay, A.; Sarkar, A.; Howlader, P.; and Balasubramanian, V. N. 2018. Grad-cam++: Generalized gradient-based visual explanations for deep convolutional networks. In *2018 IEEE winter conference on applications of computer vision (WACV)*, 839–847. IEEE.

Chelmis, C.; and Yao, M. 2019. Minority report: Cyberbullying prediction on Instagram. In *Proceedings of the 10th ACM conference on web science*, 37–45.

Dai, W.; Li, J.; Li, D.; Tiong, A. M. H.; Zhao, J.; Wang, W.; Li, B.; Fung, P.; and Hoi, S. 2023. InstructBLIP: Towards General-purpose Vision-Language Models with Instruction Tuning. arXiv:2305.06500.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255. Ieee.

Exon, J. 1996. The Communications Decency Act. *Federal Communications Law Journal*, 49(1): 4.

Fang, Y.; Wang, W.; Xie, B.; Sun, Q.; Wu, L.; Wang, X.; Huang, T.; Wang, X.; and Cao, Y. 2023. Eva: Exploring the limits of masked visual representation learning at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19358–19369.

Felzenszwalb, P. F.; and Huttenlocher, D. P. 2004. Efficient graph-based image segmentation. *International journal of computer vision*, 59(2): 167–181.

Fu, R.; Hu, Q.; Dong, X.; Guo, Y.; Gao, Y.; and Li, B. 2020. Axiom-based Grad-CAM: Towards Accurate Visualization and Explanation of CNNs. In *BMVC*.

Gildenblat, J.; and contributors. 2021. PyTorch library for CAM methods. https://github.com/jacobgil/pytorch-grad-cam. Accessed: July 17, 2023.

Hargrave, A. M.; and Livingstone, S. M. 2009. Harm and offence in media content: A review of the evidence.

Hendricks, T. 2021. Cyberbullying increased 70% during the pandemic; Arizona schools are taking action. https://www.12news.com/article/news/crime/cyberbullying-increased-70-during-the-pandemic-arizona-schools-are-taking-action/75-fadf8d2c-cf11-43f0-b074-5de485a3247d. Accessed: July 17, 2023.

John, A.; Glendenning, A. C.; Marchant, A.; Montgomery, P.; Stewart, A.; Wood, S.; Lloyd, K.; Hawton, K.; et al. 2018. Self-harm, suicidal behaviours, and cyberbullying in children and young people: Systematic review. *Journal of Medical Internet Research*, 20(4).

Kim, A. 2021. NSFW Data Scraper. https://github.com/alex000kim/nsfw_data_scraper. Accessed: August 2, 2022.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. *arXiv preprint arXiv:2304.02643*.

Krause, M. 2009. Identifying and managing stress in child pornography and child exploitation investigators. *Journal of Police and Criminal Psychology*, 24(1): 22–29.

Lenhart, A.; Ybarra, M.; and Price-Feeney, M. 2016. Nonconsensual image sharing: one in 25 Americans has been a victim of" revenge porn".

Li, C.; Xu, H.; Tian, J.; Wang, W.; Yan, M.; Bi, B.; Ye, J.; Chen, H.; Xu, G.; Cao, Z.; et al. 2022. mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 7241–7259.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Li, Y.; Vishwamitra, N.; Knijnenburg, B. P.; Hu, H.; and Caine, K. 2017. Effectiveness and users' experience of obfuscation as a privacy-enhancing technology for sharing photos. *Proceedings of the ACM on Human-Computer Interaction*, 1(CSCW): 1–24.

Meta. 2022. Appealed Content. https://transparency.fb.com/policies/improving/appealed-content-metric/. Accessed: July 14, 2023.

Neubert, P.; and Protzel, P. 2014. Compact watershed and preemptive slic: On improving trade-offs of superpixel segmentation algorithms. In *2014 22nd international conference on pattern recognition*, 996–1001. IEEE.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Kopf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32*, 8024–8035. Curran Associates, Inc.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Ramaswamy, H. G.; et al. 2020. Ablation-cam: Visual explanations for deep convolutional network via gradient-free localization. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 983–991.

Ramesh, A.; Dhariwal, P.; Nichol, A.; Chu, C.; and Chen, M. 2022. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*.

Sanchez, L.; Grajeda, C.; Baggili, I.; and Hall, C. 2019. A practitioner survey exploring the value of forensic tools, AI, filtering, & safer presentation for investigating child sexual abuse material (CSAM). *Digital Investigation*, 29.

Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.

Srinivas, S.; and Fleuret, F. 2019. Full-gradient representation for neural network visualization. *Advances in neural information processing systems*, 32.

Steiger, M.; Bharucha, T. J.; Venkatagiri, S.; Riedl, M. J.; and Lease, M. 2021. The psychological well-being of content moderators: the emotional labor of commercial moderation and avenues for improving support. In *Proceedings of the 2021 CHI conference on human factors in computing systems*, 1–14.

Tenbarge, K. 2023. Instagram's sex censorship sweeps up educators, adult stars and sex workers.

Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Uziel, R.; Ronen, M.; and Freifeld, O. 2019. Bayesian adaptive superpixel segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 8470–8479.

van der Walt, S.; Schönberger, J. L.; Nunez-Iglesias, J.; Boulogne, F.; Warner, J. D.; Yager, N.; Gouillart, E.; Yu, T.; and the scikit-image contributors. 2014. scikit-image: image processing in Python. *PeerJ*, 2: e453.

Vermeire, T.; Brughmans, D.; Goethals, S.; de Oliveira, R. M. B.; and Martens, D. 2022. Explainable image classification with evidence counterfactual. *Pattern Analysis and Applications*, 25(2): 315–335.

Vishwamitra, N.; Hu, H.; Luo, F.; and Cheng, L. 2021. Towards Understanding and Detecting Cyberbullying in Real-world Images. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*.

Wachter, S.; Mittelstadt, B.; and Russell, C. 2017. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *Harv. JL & Tech.*, 31: 841.

Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; and Yang, H. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 23318–23340. PMLR.

Wang, W.; Bao, H.; Dong, L.; Bjorck, J.; Peng, Z.; Liu, Q.; Aggarwal, K.; Mohammed, O. K.; Singhal, S.; Som, S.; et al. 2023. Image as a Foreign Language: BEiT Pretraining for Vision and Vision-Language Tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 19175–19186.