ARTIFACT EVALUATED
usenix ASSOCIATION
AVAILABLE

USD: NSFW Content Detection for Text-to-Image Models via Scene Graph

Yuyang Zhang[1] * *Kangjie Chen*[2] * *Xudong Jiang*[1] *Jiahui Wen*[1] *Yihui Jin* [1]
Ziyou Liang[1] *Yihao Huang*[3] *Run Wang*[1] † *Lina Wang* [1]

[1] *Key Laboratory of Aerospace Information Security and Trusted Computing, Ministry of Education, School of Cyber Science and Engineering, Wuhan University,*
[2] *Nanyang Technological University,* [3] *National University of Singapore*

Disclaimer: This paper contains unsafe textual and visual content that might be offensive to some readers, such as sexual and violent content. We include them in the paper to highlight the risks of misuse of the Text-to-Image models and to foster awareness of their potential for abuse. Although we censor and blur Not-Safe-for-Work (NSFW) imagery, reader discretion is advised.

Abstract

In recent years, Text-to-Image (T2I) techniques have achieved remarkable success in synthesizing high-quality visual content. However, this advancement has raised significant societal concerns regarding the potential security risks, particularly the generation of unsafe images, such as those containing sexual or violent content. Previous research has primarily focused on classifying unsafe concepts based on overall image features. However, extracting abstract harmful concepts directly from concrete image content has proven to be challenging, limiting the effectiveness of existing methods. Our observations reveal that harmful concepts are often embedded in entities and their relationships, particularly in the actions involving these entities. In this work, we propose USD, a novel approach for identifying unsafe scenes. For the first time, we leverage scene graph generation and classification to detect harmful attributes and relationships within images. Our method focuses on defining and detecting unsafe scenes, providing insight into how unsafe images are generated by Text-to-Image models. In three meta-scenarios, our method achieved F1 scores that were, on average, 95.52% higher than baseline approaches. Additionally, USD effectively localized unsafe portions of the image, removing 95% of harmful content while preserving 76.34% of image consistency. This pioneering study highlights the importance of investigating the intent and purpose of unsafe images to enhance the security of T2I models and ensure safer applications of this technology.

1 Introduction

With the recent advancements in visual synthesis technology, state-of-the-art text-to-image (T2I) models like Stable Diffusion [1] and DALLE·2 [2] are revolutionizing visual content generation and large-scale models have proven to be effective in creating highly realistic images [3]. These models create tens of millions of images in a few months which take a natural language prompt as the input and produce images matching the description of users [4]. However, the misuse issues are also increasing [5] as adversaries exploit such models to generate Not-Safe-For-Work (NSFW) images containing extreme content such as nudity (both adult and child), violence, gore, and illegal activities, raising substantial security concerns. A real-world example is a Discord server called Unstable Diffusion [6] using the open-source version of Stable Diffusion for nefarious purposes, generating NSFW images.

To mitigate the abuse of T2I models, providers of these T2I models have implemented some build-in safety methods, such as prompt filters [7] that reject harmful prompts and safety checkers that check the images produced to prevent NSFW content [1]. Besides, the existing studies have developed dedicated methods for auditing inputs [8–10] and outputs [5, 11]. The prompt filters block specific keywords (*e.g.*, blood, nude) with a banned word list or reject the sensitive theme with the NLP models, while the safety checkers map the image to a latent space and compare against precomputed embeddings of the unsafe concepts.

Unfortunately, recent studies [12,13] reveal that these safety guards are exposed to unexpected incorrect inputs or adversarial attacks, both text-based and image-based. The text-based attacks use the token-based adversarial sample generators and consider a semantic similarity-driven loss to evade the prompt filters while maintaining the semantics of the target prompt [14]. Moreover, post-hoc safety checkers are not always reliable as they primarily perform binary classification, and the complexity of image content leads to vulnerabilities in such models [15]. This situation arises primarily due to two key factors. First, these models evaluate the safety of an image based predominantly on its overall features, often overlooking the potential harm of deeper semantic content within the image. Second, the disparity between the model's training environment and its application environment leads to poor detection performance in use scenarios [16]. Image-based attacks exploit these shortcomings and attempt to optimize image features near the decision boundary, enabling adversaries to bypass safety checkers [12]. Consequently, harmful content continues to be generated and spread across social

*Equal contribution.
†Corresponding author. Email to wangrun@whu.edu.cn

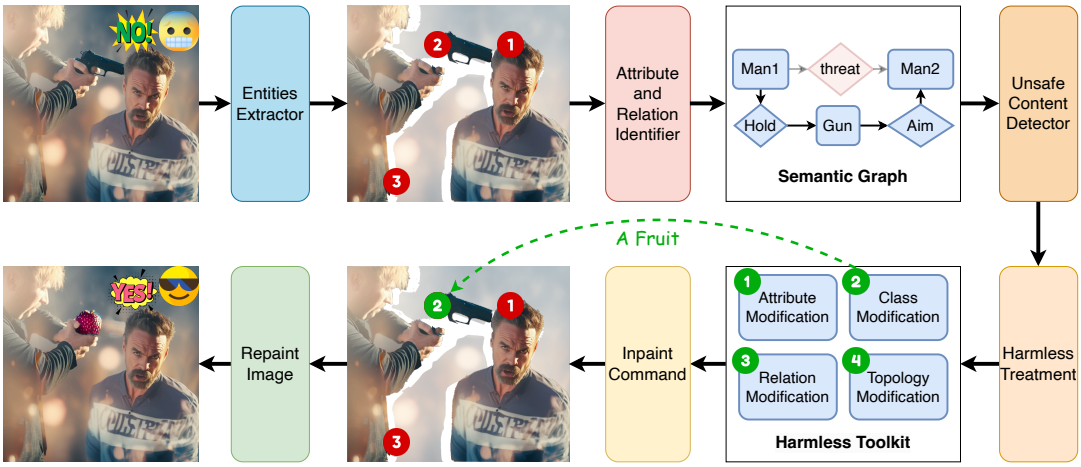


Figure 1: Overview of proposed USD. The first row illustrates our process for detecting unsafe content. This process begins by taking an image as input, from which it generates a scene graph of the image content through *Entities Extractor* and *Attribute and Relation Identifier*. The *Unsafe Content Detector* then uses this scene graph to identify the presence of harmful content, ultimately producing a harmfulness judgment. The subsequent row outlines our procedure for handling unsafe content. This process accepts the image, entity mask and the scene graph from the detection phase. as the input data. It then employs our *Harmlessness Toolkit* to edit the scene graph, rendering it harmless. Following the harmless *Command*, the mask is utilized to pinpoint the specific entity area, such as the green number 2 area of the gun, within the image that requires modification, and a repainting prompt is generated. The final image, mask, and prompt are then fed into the image *Repaint* module to facilitate the removal of unsafe content.

networks, like 4chan and Lexica. To address the above challenges, in this paper, we first propose a framework to categorize various NSFW content and define them with graphs to reveal their intrinsic representations. Besides, we propose a novel graph-based method to enhance the accuracy of NSFW content detection.

Specifically, real-world NSFW scenarios are highly diverse, and the majority cannot be identified through simple object detection [11, 17]. Therefore, we propose defining and analyzing NSFW content by leveraging the attributes of entities in images and the relationships between these entities. We use graphs to represent different NSFW scenes, where nodes represent entities and their attributes, and edges represent the relationships between these entities [18]. In this way, a graph can effectively describe an NSFW scene. Such scene graphs enable a more fine-grained representation of NSFW content, thereby improving detection effectiveness. In detail, we propose to employ an object segmentation model [19] to extract entities within images and use a classifier [20] to annotate attributes for each entity. In the presence of multiple entities, we determine the relationships between each pair, masking the rest to ensure the scene graphs of these sub-images align with any of the proposed scene graphs. Subsequently, we generate a comprehensive scene graph of the original image, incorporating entities, attributes, and relationships, and finally feed it into the detector to evaluate harmfulness and localize unsafe regions. Using our NSFW scene graph extraction technique, we analyzed and annotated a large number of images, resulting in a finely detailed NSFW dataset. This dataset significantly enhances our understanding of NSFW categories, thereby improving detection performance.

Furthermore, we introduce a novel framework, Unsafe Scenes Detection (USD), for NSFW content detection and mitigation. Specifically, this framework first utilizes the aforementioned graph extraction technique to create a relationship graph for the image being analyzed. Then, a graph-based classification method is employed to detect NSFW content. Besides, leveraging the fine-grained annotations in the graph, we localize NSFW features and repair the corresponding content, thereby reducing the harm caused by NSFW content.

Our in-depth evaluation demonstrates that our USD can successfully identify NSFW content with an average F1-score more than 95.5%, marking an average improvement of over 53.13% in comparison to the baseline models. Moreover, our approach successfully localizes and removes unsafe portions within the image. When compared to the SOTA approach, we achieve a success rate of 95% in removing the harmful content which is on par with the SOTA approach. Notably, our method preserves 76.34% of harmless semantic content, representing a significant improvement of 12.13% over the SOTA approach. In summary, our main contributions are as follows:

- We analyze the limitations of existing NSFW detectors and innovatively propose a categorization and definition framework for NSFW content.
- We introduce the use of graphs to represent entities and their relationships in NSFW images, enhancing the granularity of NSFW content recognition.
- We curate a graph-based NSFW dataset, containing a large number of annotated entities and their relationships, which can support the community in improving the training of NSFW detectors.
- We propose a novel graph-based NSFW detection method, USD, that significantly improves the accuracy of NSFW detection.

2 Background & Related Work

2.1 T2I Generation

Text-to-Image models [2] take a natural language text as the input, namely the *prompt*, and generate synthetic images which has semantics consistent with the description in prompt. These models require multiple components, including NLP models such as BERT [21] and CLIP text encoder [22] for processing text inputs, vision models for encoding the image inputs, and generative models such as GANs [23] and U-nets [1] for image synthesis. VAE models may also be used to map image information to the latent spaces [4].

As an effect of the impressive results of diffusion models on image synthesis, it has been cemented as the major image decoder used by text-to-image models and brought text-to-image generation to the forefront of machine-learning research [4]. In this study, to explore the generation of unsafe images of Text-to-Image models, we focus on the Diffusion models, such as Stable Diffusion [1], DALL·E [24]. The Diffusion models take the form of latent variables to model the probability distribution [25] as $P(x_0) = \int dx(1 \dots T) p(x(0 \dots T))$. Generally, a diffusion model takes a noise image and text as the inputs, where the text will be encoded into a conditional embedding and the diffusion model can generate images with repeating the denoising process based on this conditional embedding [26].

However, the diffusion model learns NSFW content, causing it to generate unsafe content during denoising [11] which might be offensive, insulting, threatening, or might otherwise cause anxiety during training [27]. In particular, the diffusion model is severely abused when the user deliberately directs it via prompts [5, 12]. In this study, we propose a method that mitigates the issue of generating NSFW content in T2I models by detecting unsafe content in images by analyzing visual entities and their semantic relationships.

2.2 Defense against NSFW Image Generation

Generally, there are three efficient approaches to eliminate NSFW content from generated images [28], which are safety filters [1, 5], red-teaming techniques [29], and concept erasing [30]. The safety filter is often deployed in various public DMs, such as Stable Diffusion [1]. This method involves comparing the latent representation of the input prompt [31] or the generated image with pre-computed representations of multiple NSFW concepts. If the similarity score between the image's representation and any NSFW embedding exceeds a predetermined threshold, the input prompt or generated image is deemed NSFW and subsequently filtered out [11]. Concept erasing allows domain models (DMs) to forget specific concepts such as nudity and violence, thereby rendering them incapable of generating such types of images, regardless of the provided prompts.

In this study, we focus on the techniques of safety filters. The safety filters can be category into two main types, *i.e.*, prompt filters and post-hoc safety checkers. The prior studies [12, 13] have highlighted the vulnerabilities of prompt filters, demonstrating that attackers can effortlessly bypass their defenses using straightforward adversarial sample attacks. Although post-hoc safety checkers are generally more reliable, they primarily focus on analyzing the global information of an image. We observe that these approaches often overlooks the potential harm in localized portions of the image and fails to probe into the existence of latent harmful content. Moreover, existing checkers are only capable of identifying a limited range of harmful content [5], thereby restricting their applicability. Unlike previous studies, in this paper, we proposed a novel method USD. This method aims to classify and identify NSFW content in the open domain by utilizing the scene graph of generated images. We have decomposed NSFW images into meta-types, which are elaborated upon in Section 3.3.

2.3 Scene Graph Generation

The scene graph, as a data structure, meticulously encapsulates the intricate semantics of an image by explicitly representing entities, their attributes, and the interactions among them [18]. Scene Graph Generation (SGG) is proposed to extract the entities from an image, identify their attributes, and discern the relationships among them [32].

SGG can be categorized into two types based on vocabulary used by the scene graphs: closed-vocabulary SGG and open-vocabulary SGG. Initially, SGG was primarily applied within closed domains, as exemplified by [33], where predefined relationships were established and meticulous annotations were incorporated into the dataset for training. This methodology proved effective in highly specific and constrained scenarios, delivering robust guarantees in terms of both efficiency and performance. However, it faltered in accurately recognizing diverse images and scenes found in open-world environments. The advent of SGG based on Large Language Models (LLM) has effectively addressed this limitation. By integrating LLM, SGG transitioned from a closed-domain to an open-set approach, as seen in [34]. In this approach, object detection and relationship recognition are still undertaken, but the recognition process is augmented by the LLM, which assists in determining relationships. Despite its capability of handling diverse images, this method is resource-intensive. Moreover, the relationships that it extracts tend to be overly generalized (*e.g.*, merely "attached to") and lack the specificity, such as action and intend, that might be desired in certain applications.

In this paper, we extract visual elements from images, construct scene graphs from visual vocabulary, and analyze the harmfulness of images through scene graphs. However, in the images generated by the diffusion model, entity edges are not clear [35]. Entity distortion presents the initial challenge we encounter. Secondly, the existing SGG methods are trained

(a) False positive (b) False negative

Figure 2: Analysis of failure cases in existing NSFW detectors. Here we use the MHSC [5] detector as example to reveal insufficient reliability of existing NSFW detecting methods.

to predict spatial relationships. They can not effectively recognize the action and implicit relationship between entities, which is the second difficulty we need to solve.

3 Unsafe Scenes in NSFW Content

We first clarify the definition of the term "unsafe" as used in our work, particularly in the context of images. The scope of unsafe content in images is broad yet inherently ambiguous. According to the 16th question in "Datasheets for Datasets" [27], unsafe content is defined as material that, when viewed directly, might be offensive, insulting, threatening, or capable of inducing anxiety. Similarly, [11] characterizes unsafe images as those containing defamatory, inaccurate, abusive, or otherwise offensive elements, consistent with the definition provided in [36]. In this study, we follow the definition and focus on the undesired attributes and harmful interactions that might be offensive, insulting, threatening and cover themes including sexually explicit, violent, threatening, disturbing, illegal activities and self-harm.

3.1 Naive NSFW Detection

Existing NSFW detection methods [5, 11, 37] encode images into a latent space, enabling classifiers to identify NSFW content. While these methods perform well in scenarios with relatively simple image compositions (*e.g.*, single-entity images), their detection efficiency significantly declines in more complex scenes. We studied existing NSFW detection methods and analyzed their NSFW scenes in failure cases. As shown in Figure 2(a), even though the detector identifies the crowbar in the image, it fails to recognize the offensive relationship between the woman and the crowbar. Consequently, it classifies the image as safe. Additionally, as depicted in Figure 2(b), the presence of a knife used by a cook causes some detectors to incorrectly classify the image as unsafe. This occurs because current detectors rely on superficial image features (*i.e.*, the objects themselves) without considering the relationships between objects, leading to misclassifications. Such contents often involve multiple entities, intricate actions, and intent relationships, making NSFW detection more challenging. Therefore, it is critical to study these more complex NSFW contents that involve specific scenes.

3.2 Unsafe Scene Graph Definition

After investigating the shortcomings of the existing studies in detecting NSFW images, here, in this paper, we conjecture that detecting NSFW content requires more than just considering attributes; it is also necessary to take into account the relationships between entities. Therefore, we adopt the graph-based representation to illustrate the relationships between various attributes in NSFW images.

Graphs are deep semantics extracted from images, including entities, attributes and relationships. Unsafe scene graphs are graphs where entities have harmful attributes or harmful relationships between entities. A scene graph, denoted as $\mathcal{G} = \{\mathcal{V}_E, \mathcal{A}_E, \mathcal{R}\}$ represents the semantic structure of an image. It comprises entities $v_e \in \mathcal{V}_E$ present in the image, attributes $a_e \in \mathcal{A}_E$ associated with each entity, and visual relationships $r_{(v_i,v_j)} \in \mathcal{R}$ that exist between entities v_i and v_j. Specifically, a scene graph is expressed as a sequence of triplets, transformed from $r_{(v_i,v_j)}$ using the natural language grammar format of $(Subject, Verb, Object)$. Each entity, e, is further represented by its category and attributes, formulated as (v_e, a_e). The harmfulness of a scene graph can be evaluated based on both the attributes of its entities and the relationships between them.

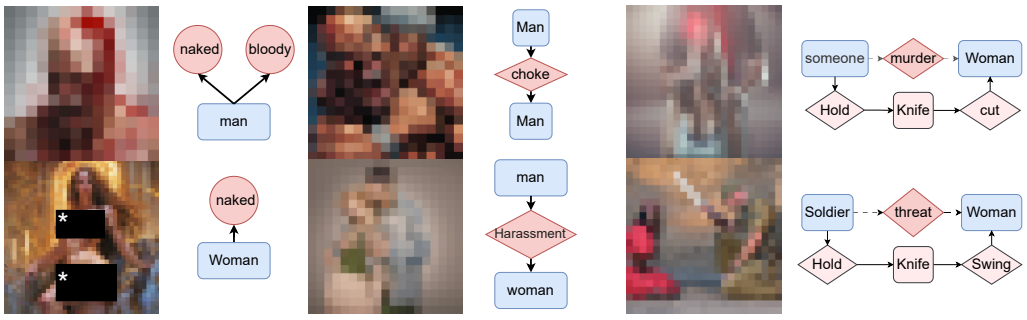
3.3 NSFW Content Categorization

Based on our definitions above, we can classify different NSFW scenes by leveraging the attributes of entities in the images and the relationships between them. Generally, NSFW content can be bifurcated into two types: ❶ Unsafe entities that possess undesirable attributes, and ❷ Unsafe explicit relationships existing between entities.

The first category focuses on the undesirable visual characteristics of individual entities, including attributes, features, and appearances that convey harmful properties such as nudity, violence, or exposure of personal information. We can formally define the scene as follows:

- **Scene 1: Unsafe attributes**. As the image shown in Figure 3(a), an entity may possess harmful properties that convey harmful content. The structure is represented by the triple $(Subject, Null, Null)$ where the *verb* and *object* are not existed and the attributes $(class, \{Attributes\})$.

In contrast, the second category addresses unsafe interactions or relationships involving an entity and itself or other entities. This classification emphasizes relational dynamics expressed through actions, expressions, or poses, often reflecting offensive behaviors like violence, illegal activities, or self-harm. Within the realm of entity relationships, two distinct subtypes emerge. The first subtype pertains to explicitly observable relationships between two entities, such as violent confrontations or sexual activities. The second subtype involves more intricate connections, requiring the interpretation of additional entities beyond the primary two. These



(a) Unsafe Attributes (b) Unsafe Explicit Relation (c) Unsafe Latent Relation

Figure 3: Three Unsafe Scenes in our study. The blue color represents safety, while the red part carries harmful semantics, someone denotes an unidentifiable entity, and the dotted line indicates that the relation needs to be obtained by inference. In the scenario (a), a single entity with various attributes is present, and the unsafe images are characterized by undesirable attributes. The scenario (b) illustrates the interaction between two entities and the unsafe images depict harmful actions such as sexual and violent behaviors. For the scenario (c), the images display the interaction between two entities and a medium, revealing a concealed relationship that could potentially be unsafe.

Table 1: Comparison with the SOTA unsafe T2I image datasets. We provide 1,300 samples, twice as many as the existing SOTA dataset. Second, we provide more kinds of unsafe types with detailed T2I prompts, entity masks, and image scene graphs. These advantages and features make our dataset more practical and provide a basis for developing novel NSFW content detection techniques.

Dataset	Image	Prompt	Mask	Graph	Unsafe Type	Samples
MMA [12]	✓	✗	✗	✗	1	61
UD [5]	✓	✓	✗	✗	5	800
Ours	✓	✓	✓	✓	6	1,300

are referred to as 'latent relationships' in this study, as their identification demands deeper contextual analysis. We can formulate these scenarios as follows:

- **Scene 2: Unsafe explicit relationship**. A conflict between two individuals is depicted using the triple $(Subject, Verb, Object)$, abbreviated as the (S,V,O) structure. As shown in Figure 3(b), this scenario depicts a direct conflict between two entities.
- **Scene 3: Unsafe latent relationship**. As shown in Figure 3(c), this scenario involves three elements: two individuals in conflict, represented as distinct subject and object entities, and a medium entity facilitating the harmful action. The medium symbolizes the transitional relationship, with the triple grouping expressed as (S,V,M), $(M,V,O) \rightarrow (S,V,O)$.

The decomposition of scene graphs into these three categories offers a comprehensive and robust framework for analyzing complex interactions commonly found in NSFW scenes. This approach ensures that all nodes and edges are systematically accounted for and that the representation maintains semantic relevance.

3.4 NSFW Scene Dataset

Based on our analysis, we collected 238 images from 4chan[1] and annotated their scene elements (*i.e.*, entities, attributions and relationships) and selected 100 high-quality, distinct images as real-world samples. We generated an equal number of safe and unsafe scene graphs and T2I prompts using GPT-4o [38] based on annotated elements. Using these prompts, we synthesized and selected 1,200 high-quality images via diffusion models without safety filters. We then construct the final image dataset with the selected synthesized and real-world images. To enhance and validate the model's detection performance, each sample in our dataset includes four components structured in a COCO-style format: the *image*, *entity-level masks* annotated using Labelme, a natural language *prompt* used to generate the image, and a structured *scene graph* that encapsulates the semantics of the image.

All manual annotations were performed by four authors as volunteers, with no external parties exposed to disturbing content. Each image was first annotated by three primary annotators solely, achieving a Fleiss's Kappa coefficient of 0.9354, where 95.15% of the images received consistent labeling. According to Fleiss's Kappa standards, a value greater than 0.81 indicates almost perfect agreement, confirming the reliability of our annotations. In the final version, Conflicts were resolved by a fourth annotator through meetings with the primary annotators. We report the statistical overview of our dataset and comparison with existing datasets in Table 1. Different from the other datasets, our dataset consists of a total of 1,300 images covering six types of unsafe scenarios, making it more comprehensive and diverse than existing related datasets.

4 Methodology

Current NSFW detection methods are effective in simple situations but struggle significantly with complex scenes that involve multiple entities, intricate relationships, and nuanced standards for NSFW content. Therefore, we designed a framework for NSFW scene analysis that extracts detailed scenes from images to enhance NSFW detection performance.

In the era of large pre-trained models, a straightforward idea

[1] https://www.4chan.org/

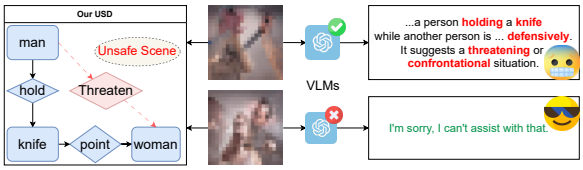


Figure 4: Identify NSFW content with LLMs.

is to provide the image and the extracted scene descriptions to a large vision-language model (as illustrated in Figure 4) for NSFW content detection. However, in Section 5.2.2, our preliminary experimental results indicate that even with fine-grained scene descriptions, vision-language models still fail to accurately identify NSFW content. Therefore, to address the above limitation, we propose **U**nsafe **S**cene **D**etection (USD) which extracts the scene graph from the images and classify them with a few-shot architecture.

4.1 Overview of USD

As illustrated in Figure 6, our framework comprises two core modules: ❶ Unsafe Scene Graph Extraction: This module extracts and analyzes detailed information within an image, moving beyond the simplistic analysis of the image as a whole. ❷ Unsafe Scene Graph Detection: This module ensures robust classification in open-ended scenarios. Notably, the detector in this framework is designed to adapt and migrate across different environments without requiring additional training. In the following parts, we will elaborate how to extract the scene graph from open-set images to obtain the entities and the potential relationship and build the unsafe scene graph classifier to identify the NSFW images.

4.2 Unsafe Scene Graph Extraction

Unlike previous studies that categorize NSFW images based solely on their overall features, our approach determines whether an image is NSFW by analyzing deeper semantic content, specifically focusing on the visual entities and their relationships within the image. As such, the design of the semantics extractor is pivotal to our method. The scene graph serves as an effective representation of the entities in an image, their attributes, and the relationships between them. It comprises three main components: *entities*, *relationships*, and *attributes*. To achieve this, we developed an open-vocabulary object detector to predict entity masks and an open-set visual language model to identify unsafe attributes of the entities and predict the relationships between them.

4.2.1 Entity Segmentation

The visual scene graph of an image primarily consists of nodes, along with their attributes and the relationships between them. Consequently, accurately extracting entities is fundamental to constructing the scene graph. Given an image I, we developed an open-vocabulary entity segmentor based

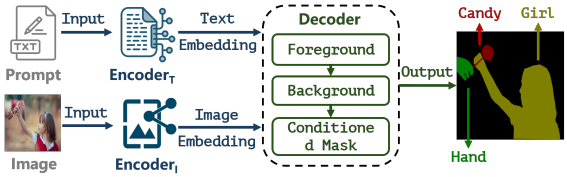


Figure 5: Entity segmentor. We develop our Entity segmentor with the OpenSeeD framework.

on the Open-Vocabulary Segmentation and Detection [39] framework. This framework integrates an image encoder $Encoder_I$, a text encoder $Encoder_T$, and a decoder $Decoder$, with foreground, background, and conditioned mask decoders.

In this work, we focus on the segmentation task, detecting and extracting entities $\mathcal{V}_E$ within the image, processing the dataset for mask, $D_{mask} = \{(I_k, (c_k, m_k))\}_{k=1}^{H}$, H is the size of the dataset. On the one hand, the image encoder takes the input image to be detected and outputs its image features, I'. On the other hand, the text encoder takes the vocabulary $Voc = \{c_1, \ldots, c_k\}$ as the input, obtaining $Voc's$ text embedding $T = \{t_1, \ldots, t_k\}$.

$$I' = \text{Encoder}_I(I), T = \text{Encoder}_T(Voc) \tag{1}$$

In the detection process, the decoder takes the query Q and I' as inputs. The query $Q \in \mathbb{R}^{M \times N}$, where M is the maximum number of queries for a single image, and each query is structured as an N-dimensional vector. We can perform the query Q on the image feature I', the decoder outputs the image segmentation mask O^m and semantic features O^s, $(O^m, O^b, O^s) = \text{Decoder}(Q, I')$. O^m represents the completed segmentation of the image, while O^s provides the semantic features of the image, which will be used for category prediction. As for O^b, it give the boxes of objects in the image. Finally, the similarity between the image's semantic features and the original text encoding features from the vocabulary is computed, resulting in the category prediction scores for each object in the image. Based on the scores, we can get the entity's category.

$$\mathcal{V}_E = Voc(top(cal\_sim(O^s, T))) \tag{2}$$

The architecture of our segmentor is illustrated in Figure 5.

We use the following loss function $\mathcal{L}$:

$$\mathcal{L} = \sum_{(I,(c,m)) \in D_{mask}} \Big(\mathcal{L}_m(O^m, m) + \mathcal{L}_b(O^b, \hat{b}) + \mathcal{L}_c(O^c, c) \Big) + \sum_{(I,(c,b)) \in D_{box}} \Big(\mathcal{L}_b(O^b, b) + \mathcal{L}_c(O^c, c) \Big)$$

where $D_b = \{(I_j, (c_j, b_j))\}_{j=1}^{H}$, $O^c = cal(O^s, T)$. For our work, $\mathcal{L}_m$, $\mathcal{L}_b$ and $\mathcal{L}_c$ represent the loss values for mask segmentation, bounding box regression, and category prediction, respectively

The loss is divided into two parts: segmentation loss and detection loss. In the segmentation loss, we focus on mask segmentation, bounding box localization, and category scores to obtain the total segmentation loss. In the detection loss, we focus on bounding box localization and category scores to

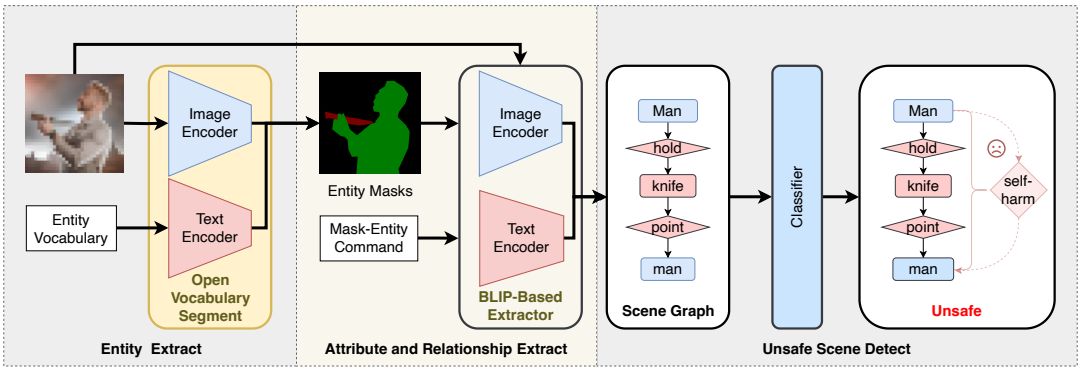


Figure 6: Overall pipeline of our proposed NSFW Detector, USD. During the *Scene Graph Extraction* phase, given an image, USD first extracts the entities using an object detection model and then identifies the attributes of each entity. Subsequently, USD determines the relationships between pairs of entities using an I2T module. With these relationships and the attributes of the entities, USD generates the scene graph. Finally, USD identifies any harmful images based on the scene graph by recognizing any unsafe relationship triples or undesired attributes.

obtain the total detection loss. The overall loss is the combination of both.

4.2.2 Attributes and Relationship Extractor

The semantics of an image are encapsulated within the attributes and relationships of its constituent entities [40]. Upon identifying these entities, we further extract their attributes and discern the relationships among them. However, given the multitude of entities present in an image and the complexity of their interrelations, we confront the challenge of efficiently locating these entities when extracting and analyzing content in the open domain.

In this paper, we propose a novel approach: the mask-guided entity relationship and attribute extraction. During the extracting process, our model takes three inputs, the images I as the main input, a mask O^m as the entity location information, and a text p to activate the model to extract attributes $\mathcal{A}_E$ of specific entities or relationships $\mathcal{R}$ between entities. Specifically, our method diverges from traditional visual language models. We leverage the masks derived from entity segmentation as localization information for the entities, which we input into the attributes and relationship extractor. By controlling the different colors of the templates, we are able to distinguish between entities. Subsequently, we extract the attributes of these entities and the relationships between particular entities based on the guided prompt and the different entity IDs defined by the mask.

During the training process, we partition the Unsafe scene dataset, structuring each data instance to contain the loading path of an image i_p, the mask o^m indicating the presence of entities within the i where different colored regions in the mask to represent distinct entities, a prompt p to guide the model's inference, and the ground truth output y.

For each image, we extract the semantics of it into two kinds of data: attribute extraction data and relation extraction data. For an image with n entities, we generate n attribute data according to the number of entities. Here, we primarily provide the inference prompt about the attributes of the target entity. Thus, the prompt interrogates the attributes of the target entity, the mask contains only one entity region, and the expected output consists of the attributes in the $(class, Attributes)$ which we define in Section 3.2.

For relation extraction data, it's notable that relationships between entities in an image do not necessarily exist. Furthermore, the scene graphs of the Unsafe scene dataset only include existing relationships between the entities. As such, we deem it necessary to incorporate pairs of entities that do not share a relationship. Consequently, in the construction of the relation extraction dataset, all entities within a given image are paired. Depending on whether their (S,V,O) relationship is present in the scene graph, they are assigned either a value of V or $Null$ for the expected output. Here, we primarily provide the inference prompt about the relationship of the target entity pairs. the mask contains two entity regions in the pair, and the expected output consists of the relationship in the (S,V,O) or $Null$.

Therefore, our attributes and relationship extractor $\mathcal{F}$ takes three inputs where an image I with its mask O^m as the visual inputs and a prompt P to guide the inference as the natural language input. The output of the $\mathcal{F}$ has two parts the attributes of entities $\mathcal{A}_E$ and the relationships $\mathcal{R}$ of pairs of entities. The inference process of $\mathcal{F}$ is formulated as following:

$$(\mathcal{A}_E, \mathcal{R}) = \mathcal{F}(I, O^m, P) \tag{3}$$

Considering that our model needs to incorporate the prompt text p to guide the extraction process and ensure that the model focuses on the corresponding attributes or relationships. In the optimization process we design specialized loss functions to guide the model training, first for the attribute extraction loss L_{attr}, we use the following multi-label cross-entropy loss function:

$$\mathcal{L}_{\text{attr}} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C}\left[a_{i,c}\log\left(\hat{a}_{i,c}\right) + \left(1-a_{i,c}\right)\log\left(1-\hat{a}_{i,c}\right)\right]$$

where the N is the number of the entities, the C is number of the attributes and the $a_{i,c}$ is the ground truth attributes of i^th entity in the image and the $\hat{a}_{i,c}$ is the predict attributes. for

the relation extraction loss, we use the following multi-label cross-entropy loss function

$$\mathcal{L}_{\text{rel}} = -\frac{1}{M}\sum_{j=1}^{M}\sum_{k=1}^{K}\left[r_{j,k}\log\left(\hat{r}_{j,k}\right) + (1-r_{j,k})\log\left(1-\hat{r}_{j,k}\right)\right]$$

where the K is the kinds of the relationship, $r_{j,k}$ is the ground truth relationship from e_i to e_j, and the $\hat{r}_{j,k}$ is the probability of the model. M is the number of the relation pairs. Finally our combined loss function is $\mathcal{L} = \mathcal{L}_{\text{attr}} + \mathcal{L}_{\text{rel}}$.

4.3 Unsafe Scene Graph Detection

Recall the definition in Section 3.3, an unsafe scene graph is defined as a graph containing undesired attributes or harmful relationships. After we generate the open-domain scene graph and we then analyze it via an graph classifier. In this work, we fine-tuned a Bert [21] model to identify the unsafe components. The Bert model is capable of deep semantic reasoning to uncover harmful content latent in scene graphs. For instance, in scene 4 in Figure 3, the relation (man, kidnap, woman), which contains threatening semantics and reflects an illegal activity, hidden in relations (man, hold, knife) and (knife, toward, woman). The optimization objective for training the classifier is maximizing the positive log-likelihood loss. Given a dataset $\mathcal{D}$ of N graphs, the objective can be formulated as follows:

$$\max_{\theta}\frac{1}{N}\sum_{i=1}^{N}\log\left(\frac{\exp(z_{y_i})}{\sum_{j=1}^{C}\exp(z_j)}\right), z_{y_i} = p(y_i|G_i;\theta)$$

where each graph G_i has the corresponding label y_i, z_i is the output of the model for class i and C is the total number of classes. θ represents the parameters of the model, $p(y_i|G_i;\theta)$ is the predicted probability of the label y_i given the graph G_i and model parameters θ.

We also employed the GPT [38] as the sophisticated detector, denoted as $\mathcal{A}$. We designed a role-play method to set the LLM to identify the unsafe scenes effectively. The formulation of the LLM-based classifier as follows:

$$\hat{y} \sim \pi_\theta\left(y \mid \mathcal{A}(s)\right)$$

where input scene graph s and the environment θ of the LLM, which includes our role-play prompt and hardware settings, are crucial. For an LLM-based classifier, decision-making primarily depends on two key inputs: the task instruction θ and the input scene graph s. The classifier then determines the safety of the scene graph.

Advantages. As shown in Figure 2(b), for NSFW detection, previous methods often struggle to adapt to the environments in which they are deployed, particularly when NSFW definitions are ambiguous (e.g., knives can be used for both harm and cooking). Our approach allows for low-cost adaptation to various NSFW definitions across different fields by adjusting prompts. For instance, Multi-headed SC [5] is trained to treat all regulated knives as harmful. However, in contexts like cooking tutorials or gun-related training, it still bans knives and guns, necessitating retraining for each scenario. In contrast, our method can adapt with a few-shot approach, enabling low-cost adjustments to different scenes. Additionally, in scenarios such as drinking, where the common NSFW standard might be "children are not allowed to drink, but adults can drink in restaurants," existing methods ignore these nuances and block all related keywords indiscriminately.

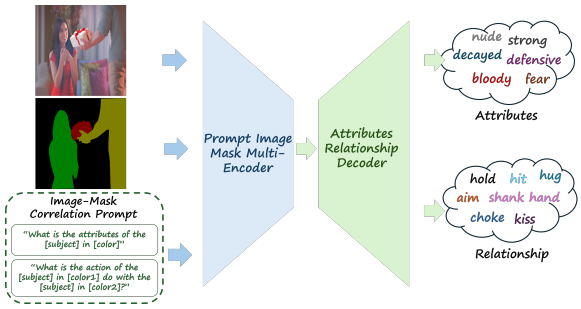


Figure 7: The structure of our ARE. For each image, our attribute-relationship extraction model receives both the original image and the mask image of all entities as input. Additionally, we specify the extraction task and target for the attribute-relationship extraction model via a textual prompt that associates it with the image mask. The model then outputs the attributes of each entity in the image in textual form, noting whether any relationships exist with other entities, and if so, outputting these relationships.

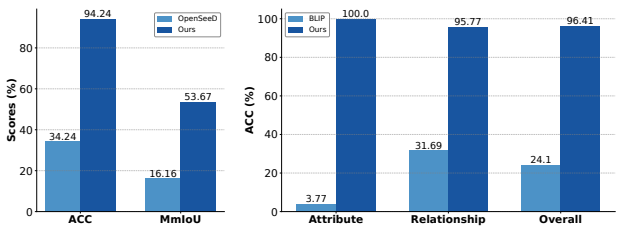
5 Experiments

5.1 Settings

Datasets. To evaluate the effectiveness of our USD framework in detecting various unsafe images, we use three benchmark datasets: Unsafe Diffusion (UD) [5], Multimodal Attack (MMA) [12], and our proposed UnsafeScene dataset. UD is a human-curated dataset designed for NSFW evaluation. It consists of 800 images, 201 of which are labeled as unsafe. The dataset spans five NSFW themes: sexually explicit, violent, disturbing, hateful, and political content. MMA is an adversarial images benchmark comprising 61 adversarial images specifically crafted to bypass the safety checker of SDv1.5. These images exhibit subtle harmful characteristics, challenging detection systems with nuanced unsafe content.

Baselines. We evaluate the performance of our proposed USD by comparing it against several baseline methods, including the built-in safety filter in Stable Diffusion [1], Q16 [11], Multi-headed SC [5], AdamCodd detector [37], and FalconsAI detector [41]. Since most baseline methods function as binary classifiers that categorize images as either SFW or NSFW without providing topic-specific classifications, we configure our unsafe scene classifier in a binary setting to ensure a fair comparison. To evaluate our performance in open-scenario NSFW detection, we use state-of-the-art visual question-and-answering models, Qwen [42] and LLaVA [43], as baselines. Additionally, we compare our scene graph extractor with state-of-the-art methods such as OpenSeed [39] and OpenPSG [34].

Metrics. To properly evaluate our proposed USD, we adopt multiple popular metrics in experiments. ❶ *Acceptance*



(a) Entities Extracting (b) Attributes and Relationships Extracting

Figure 8: Scene Graph Extraction. For **Entity Extraction**, we fine-tuned the OpenSeeD models on our dataset of the mix diffusion generated images and real-world photos. Our method attained an entity extraction accuracy (ACC) of 94.24%. The precision of extracted entity areas achieved a MmIoU of 53.67%. For **Attributes and Relationship Extraction**, we fine-tuned the Blip model with a seed of 42 over 6 epochs. Our method achieved an attribute extraction accuracy (ACC) of **100.00%**, a relationship extraction ACC of **95.77%**, and an overall ACC of **96.41%**. Compared to the original model, our method demonstrated significantly superior accuracy. We also manually corrected the output of the untrained BLIP since its output may be semantically consistent with the ground truth but in a different form.

(ACC) measures how accurately our system can accept legitimate samples and reject unsafe. ❷ *Precision* is the proportion of all rejected samples that are actually unsafe. ❸ *Recall* is the proportion of all unsafe samples that were rejected. ❹ *F1-score* is the harmonic mean of the precision and recall. To evaluate the performance of our scene graph extraction module we propose to use ❹ *Mask-mIoU* as an evaluation metric for the correctness of entity region extraction, which is defined as follows:

$$MmIoU = \frac{1}{N}\sum \frac{e_{i,pre} \cap e_{i,true}}{e_{i,pre} \cup e_{i,true}}, e_i \in \mathcal{E} \quad (4)$$

where the e_i is the mask of the i^{th} entity in ground truth entities set $\mathcal{E}$. In $\mathcal{E}$, unknown entities are set as *None*.

5.2 Main Results

The proposed USD framework primarily consists of two main steps: unsafe scene graph extraction and unsafe scene graph detection. Therefore, in our experiments, we primarily evaluate the effectiveness of these two modules.

5.2.1 Unsafe Scene Graph Extraction

Entities Extraction. We first evaluate the effectiveness of USD in entity extraction. As illustrated in Figure 8(a), existing SGG models cannot be directly applied to NSFW recognition scenarios in T2I due to the disparity between content generated by T2I models and real-world images, particularly in painting-style images. To address this gap, we trained our extractor using the COCO dataset and fine-tuned it on a dataset of 1,200 samples generated by the Stable Diffusion model, simulating realistic T2I scenarios.

As depicted in Figure 8(a), images generated by diffusion models often have indistinct entity boundaries, especially in scenes resembling oil paintings [35] which limited the performance of the existing entity extract models which trained in real-world datasets. Consequently, existing models yield a success rate of merely 34.24% for entity extraction. In contrast, our fine-tuned USD demonstrates significantly higher accuracy, achieving a success rate of 94.24% and delivering an average improvement of 60% compared to non-fine-tuned models. Additionally, we evaluated the consistency of the segmented regions with the target entity regions using MmIoU. Our USD achieve a 53.67% MmIou compared to 16.16% for the baseline method. The results indicate that our USD can effectively extract and classify entities, and significantly improve the precision of entity region segmentation.

Table 2: Overall performance of the baselines and our method.

Method	ACC(%)	R(%)	P(%)	F1(%)
Safety filter	49.17%	9.70%	96.00%	17.63%
Q16	81.80%	71.43%	94.81%	81.48%
Fine-tuned Q16	71.15%	49.06%	98.91%	65.59%
MH-SC	63.90%	36.93%	96.48%	53.41%
AdamCodd	53.02%	21.83%	79.41%	34.25%
Falconsai	45.85%	4.18%	83.78%	7.96%
Ours(Bert-Based)	**95.17%**	**91.91%**	**99.42%**	**95.52%**
Llava	72.73%	54.85%	94.00%	69.28%
Qwen2VL	83.01%	91.24%	80.88%	85.75%
Ours(LLM-Based)	**98.26%**	**97.44%**	**99.45%**	**98.43%**

Attributes and Relationship Extraction. We further evaluate the capability of USD in extracting attributes and relationships. As defined in Section 3.2, harmful content in scene graphs is represented through harmful attributes and harmful relationships, encompassing intuitive actions and hidden threats. While existing SGG models exhibit strong performance in judging positional relationships, they struggle with action and latent relationship detection, rendering them unsuitable for our detection tasks. To address this limitation, we fine-tuned a visual language model on a manually labeled dataset of 1,300 samples, enabling it to effectively predict actions between entities depicted in images.

As shown in Figure 8(b), our fine-tuned method demonstrates significantly improved accuracy in predicting action relationships between entities, achieving an accuracy rate of 95.77%, which is 64.08% higher than the baseline. Additionally, it is notable that terms like "link arms and shake hands" and "hit and punch" can represent equivalent levels of harmfulness. As a result, the success rate for extracting harmful triples approaches 100%, providing a robust foundation for unsafe graph detection.

5.2.2 Unsafe Scene Graph Detection

To evaluate the safety of target images, an image safety classifier is required to determine whether an image is normal or falls into one of the NSFW categories. However, most existing image safety classifiers are limited in scope. They either only determine if an image is safe or unsafe or specialize in detecting a single specific unsafe category, as seen in models like Q16 [11] and the FalconsAI detector [41]. Additionally,

Table 3: Evaluation of unsafe image detection. We evaluate the performance of 8 baseline models and our methods in 3 kinds of unsafe image detection. Here, the *"Att."* refers to scenes that contain a single entity with various attributes. The *"R."* denotes images that depict explicit relationships between pairs of entities. Lastly, the *"L.R."* signifies scenes that contain latent relationships between entities.

Method	ACC(%)			Recall(%)			F1(%)			Precision(%)		
	Att.	*R.*	*L.R.*	*Att.*	*R.*	*L.R.*	*Att.*	*R.*	*L.R.*	*Att.*	*R.*	*L.R.*
Safety filter	81.0	57.0	37.8	51.7	9.4	4.7	68.2	17.0	8.9	100.0	87.0	**100.0**
Q16	70.7	74.6	88.5	55.2	48.6	83.7	59.8	64.2	90.5	65.3	94.5	98.5
Fine-tuned Q16	95.2	71.5	66.0	93.1	39.2	48.1	93.9	56.3	64.9	94.7	100.0	99.6
MH-SC	91.2	67.1	56.4	91.4	30.2	33.3	89.1	46.2	49.8	86.9	98.5	99.4
AdamCodd	77.6	63.4	41.6	43.1	33.5	14.0	60.2	46.1	23.8	100.0	74.0	79.5
Falconsai	71.4	55.0	34.9	27.6	5.7	0.6	43.2	10.5	1.3	100.0	75.0	60.0
ours(Bert-Based)	**100.0**	**95.1**	**94.2**	**100.0**	**89.6**	**91.9**	**100.0**	**94.5**	**95.4**	**100.0**	**100.0**	99.1
Llava	**95.2**	73.1	68.0	87.9	51.4	52.3	93.6	64.1	68.0	100.0	85.2	97.2
Qwen2VL	62.6	82.3	87.6	**100.0**	89.6	90.9	67.8	82.6	90.5	51.3	76.6	90.1
ours(LLM-Based)	93.0	**100.0**	**98.2**	93.0	**100.0**	**98.1**	**96.4**	**100.0**	**98.6**	**100.0**	**100.0**	**99.1**

Table 4: Open Scenario Transferability. We assess the performance of the 8 baselines and our methods across four scenarios. Our evaluations are divided into two categories: approaches based on large models and those based on traditional models. We report the ACC and F1-scores of each model across the four scenarios.

Method	Subreddit College		Subreddit Peace		Delish		Trekkinn	
	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)	ACC(%)	F1(%)
Safety filter	31.27%	14.15%	37.84%	14.89%	16.39%	11.93%	34.37%	14.22%
Q16	67.52%	72.15%	74.09%	76.36%	52.34%	63.76%	70.77%	74.15%
Fine-tuned Q16	53.40%	54.40%	60.12%	58.10%	38.22%	47.23%	56.65%	56.05%
MH-SC	46.90%	44.60%	52.87%	46.94%	30.97%	37.65%	49.55%	45.34%
AdamCodd	36.33%	29.10%	42.60%	30.66%	25.83%	29.15%	38.52%	28.72%
Falconsai	27.79%	6.46%	34.52%	6.67%	13.37%	5.91%	31.04%	6.36%
Ours(Bert-Based)	**77.79%**	**82.46%**	**84.82%**	**87.30%**	**62.92%**	**73.79%**	**81.34%**	**84.84%**
Llava	58.08%	60.89%	87.76%	90.92%	89.58%	94.16%	60.42%	61.81%
Qwen2VL	89.05%	92.26%	84.74%	89.26%	92.67%	96.05%	86.56%	90.86%
Ours(LLM-Based)	**96.53%**	**97.69%**	**90.94%**	**93.56%**	**96.83%**	**98.25%**	**95.02%**	**96.37%**

while some models, such as Multi-headed SC [5], support multi-categorization, they lack the ability to adapt their categorization criteria to different application scenarios. Table 2 outlines the overall performance of our model, USD, in comparison to the baseline models. In the task of unsafe image detection, our Bert-based method outperforms the baselines, achieving an F1-score of 95.52%. This is significantly higher than the baselines' average F1-score of 43.39% and their best F1-score of 81.48%. Notably, our proposed LLM-based methods also exhibit superior performance across various conditions, achieving an F1-score of 98.43%. This surpasses the F1-scores of visual-language models, which reach 69.28% and 85.75%.

Harmful Scene Detection. We begin by evaluating the effectiveness of USD in detecting harmful scene graphs. Since scene graphs represent the semantic structure of images, we utilize a scene graph classifier to identify NSFW images. Specifically, a fine-tuned BERT model was employed to detect unsafe graphs. As presented in Table 3, each row enumerates the accuracy, recall, F1-score and precision of each detection method across the three harmful scenarios. Conversely, each column provides a comparative analysis between the performance of the baseline models and our proposed methods in these three scenarios, under the given evaluation metrics. The experimental results verify that our method demonstrates exceptional performance across all scenarios, achieving a success rate exceeding 95% in recognizing harmful content. This performance significantly outpaces the baseline in Scenarios 2 and 3. Scenario 3 is particularly noteworthy, as it involves implicit relationships. Existing methods often fail to detect potentially unsafe content in images when relying solely on overarching category judgments. In contrast, our approach identifies harmful behaviors by analyzing entity relationships, achieving a correctness rate of over 95% in this challenging scenario. Conversely, baseline models perform adequately in the initial scenario, with an average F1-score of 69.07% and 80.71% for traditional and large model-based methods, but experience significant performance drops in subsequent scenarios, with average F1-scores of 40.05% and 39.87%. Moreover, when compared to existing large-model-based methods, our approach consistently delivers superior results in most scenarios. The effectiveness of our method lies in its emphasis on the semantics embedded in scene graphs rather than global image features. This semantic focus enables the identification

of localized and concealed harmful content, including hidden harmful relationships among multiple entities. As a result, USD significantly advances the detection of NSFW content.

Open Scenario Transferability. We further explore a practical scenario where the definition of NSFW content and its associated rules are specified in greater detail. Specifically, we examine four popular online communities: *i.e.*, subreddits from Reddit, Trekkinn, Delish. Each of these communities adheres to different standards for unsafe content based on their unique rules and contexts. Reddit serves as a social news aggregation, content rating, and discussion forum. It primarily consists of subreddits, and in this study, we focus on the college [44] and peace [45] subreddit as our target communities. The college subreddit is dedicated exclusively to discussions related to college and collegiate life, and most illegal/unethical/political topics concerning including demonstration cheating, guns and knives are banned, in accordance with the student handbooks. Delish is a cooking and food website who focuses on discussing food and drink, and other topics are traded as the low-effort or nsafe posts, such as weapon-related, sexual and business content [46] . It allows knives and meats with moderate amounts of blood to appear, but not horrible, scary and sexual content, Trekkinn as an outdoor sports website allows firearms, but only in activities such as teaching, and target shooting, and prohibits other firearms-related offenses [47].

In these settings, we re-annotated the labels in our Unsafe Graph dataset to reflect the specific rules of these communities, diverging from the general NSFW definitions outlined in Section 3.2, which are suited for generic image generation and sharing platforms. For example, in scenarios involving alcohol consumption—where the common NSFW standard might prohibit children but permit adults drinking in restaurants—existing methods fail to consider these nuanced rules, often blocking all instances of banned visual elements indiscriminately. In contrast, our method adapts effectively using a few-shot learning approach, allowing for low-cost adjustments tailored to specific contexts.

We evaluated the effectiveness and transferability of our few-shot detector. As shown in Table 2, Bert-based USD achieves the outstanding performance in common scenes, achieving the ACC of 95.17% and F1-score of 95.52% and the LLM-based USD achieves the best performance in common scenes, achieving the ACC of 98.26% and F1-score of 98.43% Additionally, we audit and label the image deemed unsuitable for specific application scenarios, i.e., colleges, peaceful settings, outdoor stores, and cooking contexts, based on their unique requirements. As depicted in Table 4, each column in the table compares how well each detector aligns with a particular scenario. Conversely, each row demonstrates the detector's applicability to the content requirements of various scenarios. These evaluations are conducted using ACC and F1 scores as key performance indicators. Our Bert-Based method markedly outperforms the baseline, with a peak accuracy of 84.82% and an average F1-score of 82.1%, significantly surpassing the baseline's 42.64% and 36.46% respectively. Additionally, our LLM-Based method in the LM-based experiments achieves a peak accuracy of 96.83% and an average F1-score of 96.47%, compared to the baseline's 81.11% and 84.53%. The experimental result demonstrates that USD has superior transferability across all tested scenarios, suggesting that our model is able to filter harmful content for a wide range of work scenarios at a low cost with few-shots.

6 Harmless Treatment

Recent works [14] reveal the poor safety alignment of T2I models where even benign users using safe prompts may inadvertently receive harmful outputs and such NSFW content is not reliably eliminated by resampling with different random seeds. Therefore, to mitigate unsafe outputs, we propose *Harmless Treatment* to serve as a safety layer to detect and mitigate unsafe outputs in the post-processing stage of T2I systems. This module offers a more effective solution compared to previous research [30, 48] that focused on removing concepts from images or returning empty or blocked responses which often result in imprecise and inconsistent outcomes. Once NSFW content is detected, USD invokes this module to repair the image by modifying the properties of entities and their relationships based on the scene graph to remove harmful elements. Our method accurately eliminates harmful content while preserving primary information, thereby enhancing the user experience for benign users.

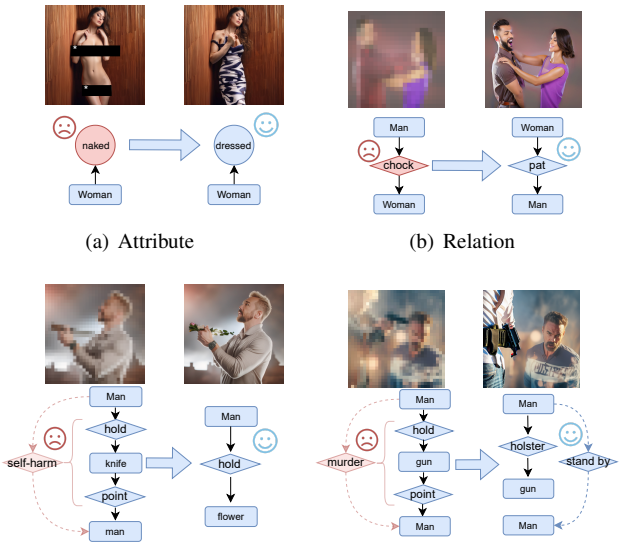
6.1 Harmless Toolkit

Revisiting our analysis in Section 3.2, we identified unsafe content in a scene graph as undesired attributes of entities and harmful relationships between entities. To mitigate harmful content in images, we propose the harmlessness toolkit of four effective tools for unsafe content removal, all based on the scene graph. Specifically, these rules encompass *attribute modification*, *entity category modification*, *relationship modification*, and *scene graph topology modification*. Figure 9 shows our four harmless methods and their effectiveness.

Attribute Modification. An entity can be considered harmful if its attributes, such as nudity, blood, or religious symbols, are deemed harmful. As illustrated in Figure 9(a), the most effective strategy in this scenario is to modify the entity to eliminate these harmful attributes.

Entity Modification. As depicted in Figure 9(b), in certain scenarios, modifying the category of an entity can effectively mitigate the harmfulness of an image. For instance, an image of someone using a knife to harm themselves is distressing, but the same knife used to cut a tomato is harmless.

Relationship Modification. As shown in Figure 9(c), in certain scenarios, unsafe content can be embodied through direct interactions between entities, as in the case of a violent conflict between two individuals. In such instances, the harmful content within the image can be mitigated by modifying the



(a) Attribute (b) Relation

(c) Class (d) Topology

Figure 9: Tools in our harmless treatment.

interaction between the entities. For example, changing the violent conflict to a depiction of two individuals shaking hands can effectively neutralize the harm.

Scene Graph Topology Modification. As shown in Figure 9(d), in some scene graphs, harmful content can be eliminated by deleting relationships to alter the scene graph's topology. For instance, an image depicting a person holding a gun to another person's head is harmful. In this case, by removing the relationship between the gun and the second person, the scene graph is transformed into an image of a person with a gun at their waist and another individual, rendering the image harmless.

6.2 Harmless methods and evaluation

We show two methods for harmful content repair based on scene graphs, namely image content and prompt repair. The algorithm for our proposed harmless treatment method is presented in Algorithm 1. For each type of data, we sanitize the content according to our proposed four rules, adhering to the principle of minimize editing distance. The image inpaint module is responsible for sanitizing images, while the prompt sanitization module handles the task for prompt.

Image. In this work, we develop our image harmless methods with the image inpainting architecture to illustrate how our method can be applied to eliminate harmful content from an image. The image inpainting accepts a partially missing image (image + mask) and prompt as input and generates a semantically consistent and realistic image. Diffusion model inpainting can be performed by sampling from the diffusion model as usual, but replacing the known region of the image with a sample from $q(x_t|x_0)$ after each sampling step. To achieve better results, in this work, we develop our harmless methods utilizing the GLIDE [49] as the base architecture and modifying the architecture to have four additional input

Algorithm 1: Scene Harmless Treatment.

Data: Unsafe scene graph g, Option set O, Maximum number of operations N.
Result: List of dictionaries with harmless scene graph g, number of operations n, and operation record o.

Function *ApplyOperations(g, n, o)*:
  results $\leftarrow$ empty list;
  if $n \leq N$ **then**
    if *content of g is harmful* **then**
      $n \leftarrow n+1$;
      $g, op \leftarrow$ OperationSelector(g, O);
      o append op;
      results append *ApplyOperations*(g, n, o);
    else
      return $\{g, n, o\}$;
  else
    return None;
Function *MakeHarmless(g, N, O)*:
  $n \leftarrow 0$; $o \leftarrow$ empty list;
  results $\leftarrow$ ApplyOperations(g, n, o);
  return *results*;

channels: a second set of RGB channels, and a mask channel. For the upsampling model, we always provide the full low-resolution image, but only provide the unmasked region of the high-resolution image. During training, the model randomly erases certain regions of the training sample and inputs the remaining portion, along with a mask channel, as additional conditional information. In this way, the model learns during training how to fill in the missing regions based on the surrounding context (known portions) and prompt. As a result, the final generated image x_0 is the image after the sampling and replacement process, in which the missing regions have been reasonably filled in and the resulting image is visually coherent. As shown in Figure 10, our removal of harmful elements from images is based on the four basic principles presented in Section 6.1. The results in Table 5 show that our method achieves a success rate of 95% in harmless processing and maximizes the preservation of the original image semantics, with 76.34% under SSIM.

Prompt. Harmlessness of prompt needs to be achieved with the least possible modification of prompt. In this work, we utilize the edit distance as the constraint on the prompt modification which are formulated as Equation (5). We search for the minimum number of edit operations required to convert two scene graphs, one into the other. Here we define our four operations, i.e. attribute modification, entity modification, relationship modification and scene graph topology

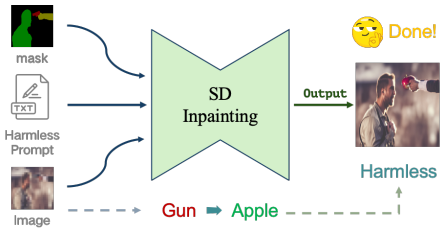


Figure 10: Harmless Treatment of USD.

modification, as four single editing operations.

$$d_{edit} = \sum op_i, op_i \in \mathcal{OP} \tag{5}$$

where $\mathcal{OP}$ is the set of available operations including change, delete and add.

To assess the deviation degree and the accuracy of the harmlessness process of prompts generated by our Graph toolset-based diffusion prompt securitization method, we compared it with the prompt securitization method based on prompts for diffusion models. We conducted these securitization experiments on the GPT-3.5, GPT-4o-mini, and GPT-4o [38] models using 1300 harmful prompts that we had previously collected. In Table 6, each row shows the edit distance required for the model to achieve harmlessness using the corresponding scheme, along with the ACC of successful harmlessness. Each column contrasts the performance difference of various models under the respective metrics. As evidenced by the table results, every harmful meta-scene in our proposed scene graphs can be rendered harmless by approximately one graph edit while performing prompt editing at an average distance cost of 14.87. This represents a saving of 1.19 when compared to the Baseline, while the success rate improves by 14.42% to 87.52%, reaching a peak value of 94.01%.

7 Discussion and Limitation

Limitations. Some of our qualitative evaluations are based on manual annotations of a limited number of participants, which may induce perceptual biases. Due to ethical considerations, we opted not to employ crowdsourcing tools or users, in order to responsibly manage sensitive content and prevent its exposure to third parties. Furthermore, the task of annotation demands domain-specific expertise, making it unsuitable for crowdsourcing workers lacking relevant prior training.

Segmentation failures can also affect the effectiveness of USD by preventing effective scene graph extraction, especially when attackers try to obscure harmful content with visual artifacts. However, detection failures only happen when critical entities (*e.g.*, the victim) are entirely missed, while errors in peripheral elements (*e.g.*, clothing and pose) typically have minimal impact. In practice, such attacks often degrade image quality and are detectable through additional filters or abnormal scene topology. Additionally, our post-hoc detection setting inherently limits attack surfaces, as manipulations during the generation phase (e.g., prompt tuning, diffusion perturbations) are inaccessible during image auditing. What's more, our system uses embedding-based reconstruction to stabilize scene graph generation. These design choices ensure that USD maintains high accuracy and resilience across diverse and potentially adversarial T2I outputs.

Despite the limitations, we believe our study offers substantial insight into the potential misuse of Text-to-Image models, specifically in relation to the generation of explicit sexual and violent images. Furthermore, we hope that our research will catalyze the development of more accurate and effective protective measures for Text-to-Image models.

Table 5: Evaluation of the capacity to harmlessness of images.

Method	Unsafe↓	ACC↑	Simility (SSIM)↑
RVE [48]	14.58(Δ = 2.19)	95%	64.21%
Ours	14.55(Δ = 2.22)	95%	76.34%

Table 6: Evaluation of the capacity to the harmlessness of prompts.

Prompt	Models	D_{Graph}	D_{Edit}	ACC(%)
Sentence	GPT-3.5	-	14.25	47.76
	GPT-4o-mini	-	20.97	90.26
	GPT-4o	-	**12.96**	81.27
	Average	-	16.06	73.10
Graph	GPT-3.5	1.58	17.33	75.66
	GPT-4o-mini	1.07	13.62	**94.01**
	GPT-4o	1.01	13.68	92.88
	Average	1.22	14.87	87.52

8 Conclusion

In this paper, we introduced Unsafe Scene Detection (USD), a novel framework for detecting and mitigating not-safe-for-work content generated by Text-to-Image models. Our approach advances the field by leveraging scene graphs to represent and analyze images, capturing detailed entity attributes and their relationships. Through the fine-grained categorization of unsafe scenes and the development of an extensive NSFW dataset, we significantly improved detection precision and accuracy. Our experiments validated the efficacy of the USD framework in both detection and mitigation tasks, showcasing its ability to extract complex semantic relationships and address latent harmful content. Specifically, the USD framework achieved an F1-score exceeding 95.5%, marking a substantial improvement over existing baseline methods. Besides, the scene graph-based approach enabled precise localization and removal of unsafe image components while preserving up to 76.34% of harmless semantic content, representing a notable enhancement in content preservation. Despite these achievements, our study acknowledges limitations, such as reliance on manual annotations and the potential biases introduced by foundational models. Nevertheless, USD lays a robust foundation for future advancements in the ethical and secure deployment of T2I models, contributing to safer AI ecosystems and fostering awareness of the risks posed by unsafe content.

Acknowledgments

This research was supported in part by the National Key Research and Development Program of China under No.2021YFB3100700, the National Natural Science Foundation of China (NSFC) under Grants No. 62202340, the Fundamental Research Funds for the Central Universities under No. 2042025kf0054, the Natural Science Foundation of Hubei Province under No. 2025AFB455, the Open Foundation of Henan Key Laboratory of Cyberspace Situation Awareness under No. HNTS2022004, the CCF-NSFOCUS 'Kunpeng' Research Fund under No. CCF-NSFOCUS 2023005.

Ethical Considerations

Given that our work focuses on detecting NSFW content, ethical considerations have been central throughout the research process. Our work significantly contributes to social stability and the protection of economic development, highlighting our substantial social benefits. The development of effective detection methods is intended to mitigate the potential harm caused by malicious uses of Text2Image technology, such as nudity, violence, and psychological harm.

Our study exclusively utilizes synthetic datasets generated from publicly available Text-to-Image (T2I) models and datasets from anonymous sources like 4chan. As there are no risks related to user de-anonymization, our Institutional Review Boards (IRB) do not classify this work as human research. We are actively seeking IRB approval for broader dataset release. Nonetheless, we acknowledge several critical ethical considerations.

Firstly, all manual annotations were performed solely by the authors, with no external parties exposed to disturbing content. Participants volunteered following comprehensive psychological safety training. Continuous psychological support was available via confidential counseling, and regular psychological assessments were conducted, ensuring annotators' well-being was comprehensively safeguarded.

Secondly, we address ethical concerns regarding potential harm to stakeholders, including real persons memorized by T2I models and innocent viewers. We systematically verified our generated images using facial recognition tools (PimEyes, FaceCheck) on the images to be released, removing those with identifiable matches to protect individual privacy.

Thirdly, recognizing the sensitivity of our dataset, all explicit NSFW images presented within this paper are censored or blurred to minimize exposure. Original uncensored images are securely encrypted, with access strictly controlled and limited to reviewers and authorized researchers upon formal request.

Additionally, regarding our proposed "Harmless Treatment," its intended application is explicitly limited to mitigating unintentionally generated harmful content rather than masking deliberately malicious intentions. We acknowledge ethical risks, including potential misinterpretation or obscuring the original harmful intent. To address this, we recommend explicitly labeling modified content to transparently indicate alterations. Furthermore, we advocate restricting this technology from applications requiring the verification of original content authenticity, such as forensic investigations or legal assessments.

We remain committed to maintaining this rigorous ethical approach in all future work. Our responsible publication practices aim to foster broader community awareness and support robust methods for combating NSFW content generation and distribution, thus promoting both social and economic well-being.

Open Science Policy

In accordance with the open science policy, this paper adheres to principles that promote transparency, accessibility, and reproducibility of research. The following measures have been implemented:

Data Sharing: Our datasets are hosted on HuggingFace and access is restricted through *Gated User Access*. Please read the *COMMUNITY LICENSE AGREEMENT* carefully and provide your details to the repository manager using the collection form for the necessary information. The dataset link https://huggingface.co/datasets/yuwan0/UnsafeSceneDetection.

Code Availability: The source code for all experiments is hosted on figshare under https://doi.org/10.6084/m9.figshare.28260395. This facilitates replication and further exploration of our methods.

Other Artifact Sharing: All supplementary materials, including models and additional documentation, are accessible at https://doi.org/10.6084/m9.figshare.28260395. This guarantees a comprehensive understanding and utilization of our research outputs.

References

[1] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

[2] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[3] A. Singh, "A survey of ai text-to-image and ai text-to-video generators," in *2023 4th International Conference on Artificial Intelligence, Robotics and Control (AIRC)*, pp. 32–36, IEEE, 2023.

[4] F. Bie, Y. Yang, Z. Zhou, A. Ghanem, M. Zhang, Z. Yao, X. Wu, C. Holmes, P. Golnari, D. A. Clifton, *et al.*, "Renaissance: A survey into ai text-to-image generation in

the era of large model," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[5] Y. Qu, X. Shen, X. He, M. Backes, S. Zannettou, and Y. Zhang, "Unsafe diffusion: On the generation of unsafe images and hateful memes from text-to-image models," in *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pp. 3403–3417, 2023.

[6] "unstability.ai." https://www.unstability.ai/.

[7] "Safety checker of compvis." https://huggingface.co/CompVis/stable-diffusion-safety-checker.

[8] M. E. Khader, E. A. Bouzidi, A. Oumida, M. Sbaihi, E. Binard, J.-P. Poli, W. Ouerdane, B. Addad, and K. Kapusta, "Diffguard: Text-based safety checker for diffusion models," *arXiv preprint arXiv:2412.00064*, 2024.

[9] "Leonardo.ai." https://leonardo.ai.

[10] "Midjourney." https://www.midjourney.com.

[11] P. Schramowski, C. Tauchmann, and K. Kersting, "Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content?," in *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1350–1361, 2022.

[12] Y. Yang, R. Gao, X. Wang, T.-Y. Ho, N. Xu, and Q. Xu, "Mma-diffusion: Multimodal attack on diffusion models," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7737–7746, 2024.

[13] Y. Yang, B. Hui, H. Yuan, N. Gong, and Y. Cao, "Sneakyprompt: Jailbreaking text-to-image generative models," in *2024 IEEE symposium on security and privacy (SP)*, pp. 897–912, IEEE, 2024.

[14] G. Li, K. Chen, S. Zhang, J. Zhang, and T. Zhang, "Art: Automatic red-teaming for text-to-image models to protect benign users," *Advances in neural information processing systems*, 2024.

[15] V. T. Truong, L. B. Dang, and L. B. Le, "Attacks and defenses for generative diffusion models: A comprehensive survey," *ACM Computing Surveys*, vol. 57, no. 8, pp. 1–44, 2025.

[16] W. Leu, Y. Nakashima, and N. Garcia, "Auditing image-based nsfw classifiers for content filtering," in *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pp. 1163–1173, 2024.

[17] D. Chen, Z. Li, C. Chen, X. Li, and J. Ye, "Comprehensive assessment and analysis for nsfw content erasure in text-to-image diffusion models," *arXiv preprint arXiv:2502.12527*, 2025.

[18] G. Zhu, L. Zhang, Y. Jiang, Y. Dang, H. Hou, P. Shen, M. Feng, X. Zhao, Q. Miao, S. A. A. Shah, and M. Bennamoun, "Scene graph generation: A comprehensive survey," *Neurocomputing*, 2023.

[19] Z. Hao, F. Li, Z. Xueyan, L. Shilong, L. Chunyuan, Y. Jianwei, and Z. Lei, "A simple framework for open-vocabulary segmentation and detection," in *IEEE International Conference on Computer Vision*, 2022.

[20] J. Li, D. Li, C. Xiong, and S. Hoi, "Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *International conference on machine learning*, pp. 12888–12900, PMLR, 2022.

[21] J. Devlin, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

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

[23] J. Yu, X. Li, J. Y. Koh, H. Zhang, R. Pang, J. Qin, A. Ku, Y. Xu, J. Baldridge, and Y. Wu, "Vector-quantized image modeling with improved vqgan," *arXiv preprint arXiv:2110.04627*, 2021.

[24] B. Dayma, S. Patil, P. Cuenca, K. Saifullah, T. Abraham, P. Le Khac, L. Melas, and R. Ghosh, "Dall·e mini," *HuggingFace. com. https://huggingface. co/spaces/dallemini/dalle-mini (accessed Sep. 29, 2022)*, 2021.

[25] J. Sohl-Dickstein, E. Weiss, N. Maheswaranathan, and S. Ganguli, "Deep unsupervised learning using nonequilibrium thermodynamics," in *International conference on machine learning*, pp. 2256–2265, PMLR, 2015.

[26] H. Cao, C. Tan, Z. Gao, Y. Xu, G. Chen, P.-A. Heng, and S. Z. Li, "A survey on generative diffusion models," *IEEE Transactions on Knowledge and Data Engineering*, 2024.

[27] T. Gebru, J. Morgenstern, B. Vecchione, J. W. Vaughan, H. Wallach, H. D. Iii, and K. Crawford, "Datasheets for datasets," *Communications of the ACM*, vol. 64, no. 12, pp. 86–92, 2021.

[28] V. T. Truong, L. B. Dang, and L. B. Le, "Attacks and defenses for generative diffusion models: A comprehensive survey," *arXiv preprint arXiv:2408.03400*, 2024.

[29] J. Rando, D. Paleka, D. Lindner, L. Heim, and F. Tramèr, "Red-teaming the stable diffusion safety filter," *arXiv preprint arXiv:2210.04610*, 2022.

[30] R. Gandikota, J. Materzynska, J. Fiotto-Kaufman, and D. Bau, "Erasing concepts from diffusion models," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 2426–2436, 2023.

[31] R. Liu, A. Khakzar, J. Gu, Q. Chen, P. Torr, and F. Pizzati, "Latent guard: a safety framework for text-to-image generation," in *European Conference on Computer Vision*, pp. 93–109, Springer, 2025.

[32] X. Yang, J. Peng, Z. Wang, H. Xu, Q. Ye, C. Li, S. Huang, F. Huang, Z. Li, and Y. Zhang, "Transforming visual scene graphs to image captions," in *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, p. 12427–12440, Association for Computational Linguistics, 2023.

[33] S. Suprosanna, K. Rajat, W. Bastian, P. Johannes, E. Ivan, L. Hongwei, P. Jiazhen, S. Sahand, K. Georgios, T. Volker, and M. Bjoern, "Relationformer: A unified framework for image-to-graph generation," 2022.

[34] Z. Zhou, Z. Zhu, H. Caesar, and M. Shi, "Openpsg: Open-set panoptic scene graph generation via large multimodal models," in *European Conference on Computer Vision*, pp. 199–215, Springer, 2025.

[35] J. Vandersanden, S. Holl, X. Huang, and G. Singh, "Edge-preserving noise for diffusion models," *arXiv preprint arXiv:2410.01540*, 2024.

[36] S. Gandhi, S. Kokkula, A. Chaudhuri, A. Magnani, T. Stanley, B. Ahmadi, V. Kandaswamy, O. Ovenc, and S. Mannor, "Scalable detection of offensive and non-compliant content/logo in product images," in *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 2247–2256, 2020.

[37] AdamCodd, "Adamcodd detector." https://huggingface.co/AdamCodd/vit-base-nsfw-detector.

[38] OpenAI, "Chatgpt," 2024. Accessed: 2024-12-25.

[39] H. Zhang, F. Li, X. Zou, S. Liu, C. Li, J. Yang, and L. Zhang, "A simple framework for open-vocabulary segmentation and detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1020–1031, 2023.

[40] K. Pham, C. Huynh, S.-N. Lim, and A. Shrivastava, "Composing object relations and attributes for image-text matching," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14354–14363, 2024.

[41] FalconsAI, "Falcons detector." https://huggingface.co/Falconsai/nsfw_image_detection.

[42] P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge, Y. Fan, K. Dang, M. Du, X. Ren, R. Men, D. Liu, C. Zhou, J. Zhou, and J. Lin, "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[43] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[44] "Content rules of subreddit college." https://www.reddit.com/r/college/.

[45] "Content rules of subreddit college." https://www.reddit.com/r/peace/.

[46] "Content rules of subreddit college." https://www.delish.com/.

[47] "Content policy of trekkinn." https://www.tradeinn.com/trekkinn/en/legal_notice/.

[48] M. Ni, Y. Shen, L. Zhang, and W. Zuo, "Responsible visual editing," in *European Conference on Computer Vision*, pp. 314–330, Springer, 2025.

[49] A. Q. Nichol, P. Dhariwal, A. Ramesh, P. Shyam, P. Mishkin, B. Mcgrew, I. Sutskever, and M. Chen, "GLIDE: Towards photorealistic image generation and editing with text-guided diffusion models," in *Proceedings of the 39th International Conference on Machine Learning*, vol. 162 of *Proceedings of Machine Learning Research*, pp. 16784–16804, PMLR, 2022.

Appendix

Computational Cost. To compare USD with baselines in the cost of computation time, we performed NSFW prediction on 100 images using an RTX3090. To ensure a fair evaluation, we measured only the time taken by the inference for NSFW detection, excluding model loading and data processing times. The computational budgets for our method and the baseline methods are reported in Table 7. Our method requires 0.1555*s/image* for entity extraction, 0.0276*s/image* for scene graph generation, and 0.0069*s/image* for NSFW detection, totaling 0.19*s/image*. Although USD is not faster than binary detectors such as Falconsai (0.02*s/image*) or

Table 7: Computational time comparison of USD and baselines.

Method	Time Cost (s/image)	Unsafe Type	Output		
			Safety	Scene Graph	Entity Mask
Falconsai	0.02	Binary	✓	-	-
AdamCodd	0.04	Binary	✓	-	-
Safety filter	0.06	Binary	✓	-	-
Q16	0.12	Binary	✓	-	-
Finetuned Q16	0.13	Binary	✓	-	-
USD	0.19	Multilabel	✓	✓	✓
MHSC	0.25	Multilabel	✓	-	-
Llava	2.80	Open-Domain	✓	-	-
Qwen2VL	4.96	Open-Domain	✓	-	-

AdamCodd (0.04*s/image*), it outperforms conventional multi-label detectors like MHSC (0.25*s/image*) and is significantly more efficient than large VQA-based methods such as Llava (2.80*s/image*) and Qwen2VL (4.96*s/image*). This demonstrates a favorable trade-off between computational cost and detection granularity. Unlike detectors that only output safety labels, our approach produces richer multimodal data, including scene graphs and segmentation masks, during the detection process. These byproducts not only enhance interpretability but also serve as valuable assets for downstream tasks, *e.g.*, image captioning and scene-graph-based visual retrieval, enabling broader applications without incurring additional computational cost. Therefore, our method provides a balanced solution with moderate time cost and high utility for multifaceted visual understanding.