Localization and Manipulation of Immoral Visual Cues for Safe Text-to-Image Generation

Seongbeom Park[1], Suhong Moon[2], Seunghyun Park[3*], Jinkyu Kim[1*]
[1] CSE, Korea University [2] EECS, UC Berkeley [3] NAVER Cloud AI
{psb485, jinkyukim}@korea.ac.kr, suhong.moon@berkeley.edu, seung.park@navercorp.com
*Co-corresponding authors

Abstract

Current text-to-image generation methods produce high-resolution and high-quality images, but they should not produce immoral images that may contain inappropriate content from the perspective of commonsense morality. Conventional approaches, however, often neglect these ethical concerns, and existing solutions are often limited to ensure moral compatibility. To address this, we propose a novel method that has three main capabilities: (1) our model recognizes the degree of visual commonsense immorality of a given generated image, (2) our model localizes immoral visual (and textual) attributes that make the image visually immoral, and (3) our model manipulates such immoral visual cues into a morally-qualifying alternative. We conduct experiments with various text-to-image generation models, including the state-of-the-art Stable Diffusion model, demonstrating the efficacy of our ethical image manipulation approach. Our human study further confirms that ours is indeed able to generate morally-satisfying images from immoral ones.

1. Introduction

Notable progress has been made in text-to-image synthesis lately with the arising of various new machine learning methods, such as large-scale generative models trained with sufficient data at scale [35]. These methods have focused mainly on generating high-resolution images with improved image quality, maintaining affordable computational costs. However, we observe that these models often produce images that clearly should not have been generated as their content deviates from commonsense morality (e.g., violent, sexually suggestive, etc.).

Recent work [39, 42] explored a post-hoc safety checker or a safety regularizer to avoid these inappropriate contents to be generated or publicly released. However, such safety checkers still fails, publishing inappropriate images as shown in Figure 2. This is mainly because training an ad-hoc (morality) classifier to detect visual commonsense immorality is challenging for the following reasons: (i) No large-scale dataset covering commonsense immorality is available to provide such supervision. (ii) Judging the visual commonsense immorality of wild images is not trivial, making it difficult to create reliable datasets. To address these issues, following the recent work [21], we leverage textual descriptions of normative knowledge (i.e., actions that should and should not be taken, such as "I punched my friend") to train our commonsense immorality judger. In specific, a CLIP-based text-image joint embedding space [34] is utilized where language supervision allows zero-shot transfer for determining the degree of visual commonsense immorality of the generated images.

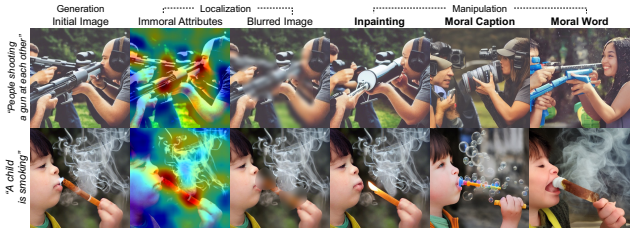

Figure 1. Given an image generated by a text-to-image model, our model judges its visual immorality and localizes the attributes that contributes its immorality (2nd column), yielding a final output with blurred immoral visual cues (3rd column). Further, based on these highlighted attributes, our model manipulates it and generates candidate morally-satisfying alternatives (4th-6th columns).

Further, the current safety checkers simply reject inappropriately generated images to be displayed. Still, their reasoning processes are often opaque and do not provide details of what makes such content inappropriate and how to fix them. To address this issue, as shown in Figure 1, our model localizes the textual/visual attributes that make the image visually immoral (e.g., a "gun" from an image of "people shooting a gun at each other" may make it immoral, or a "gun" from a given text prompt). We apply a blur kernel in the spatial domain to degrade the visual quality of inappropriate content (e.g., blurring a gun), highlighting vi-

sually immoral potential regions and yielding final outputs. Our model also highlights a set of words that drive to generate such inappropriate content, yielding another layer of interpretability.

Another part of the story is about helping users to manipulate images to be visually more appropriate (under the assumption that users do not want to generate such inappropriate content). Based on highlighted visual and textual attributes, our model provides three different kinds of image manipulation approaches that can produce a more visually moral image by substituting immoral visual cues. For example, (i) We use image inpainting techniques to replace immoral visual attributes with moral alternatives (e.g., bleeding blood on the face being replaced by a smiling face). (ii) We use a morally-describing sentence from image captioning models as a condition to manipulate immoral images (e.g., an image of "a bride is bleeding" is described as "a painting of a woman in a red dress"). (iii) Lastly, an immorally-driving word can be replaced by a moral alternative (e.g., the word "gun" from the text "a child with a gun" can be replaced with "water gun" towards ethical image manipulation).

We empirically demonstrate the effectiveness of our proposed method with the state-of-the-art text-to-image generation model called Stable Diffusion [39]. Also, our human study confirms that our method successfully manipulates immoral images into a moral alternative. We summarize our contributions as follows:

- Based on a visual commonsense immorality recognition, we introduce a textual and visual immoral attribute localizer, which highlights immoral attributes that make the input image visually immoral.
- Given immoral visual and textual attributes, we introduce three different ethical image manipulation approaches that can produce a moral image as output by automatically replacing immoral visual cues.
- We experiment with the state-of-the-art image generation model, Stable Diffusion, and we empirically analyze the effectiveness of our proposed approach, which is also supported by our human study.

2. Related Work

AI Ethics. There has been a long effort to build the concept of ethical machine learning. A landmark work was Asimov's three laws of robots [3], which define simple principles of how machines should behave from an ethical perspective. Recently, Bostrom *et al*. [6] discussed that machines' concentrated focus on problem-solving might result in severe catastrophes such as paperclip maximizers [5], deviating from guaranteeing morality. Moreover, AI ethics dilemmas [4, 41] have recently been widely discussed from philosophical perspectives.

Figure 2. Immoral output images along with text inputs (top) generated by the Stable Diffusion [39] model with (a) its safety checker module enabled and by (b) a safety regularizer [42]. We blurred some images due to their inappropriate content.

Recently, AI ethics have become a more apparent interest of importance in AI and CV communities. From Natural Language Processing (NLP) community, an increasing number of papers have been introduced, examining five different ethical categories: (i) Fairness [27], (ii) Safety [37], (iii) Prosocial [36, 38], (iv) Utility [10, 29], and (v) Commonsense Morality [18]. Especially the last topic, commonsense morality, has limited been explored in the computer vision community, which mainly focuses on safety (e.g., surveillance video analysis [50]) and fairness (preventing discrimination caused by the dataset bias [1, 32, 44]). Thus, this paper focuses on commonsense morality from the computer vision perspective and follows the definition of the recent work [18]: "*an action that is intuitively acceptable by most people as something that clearly should or should not be done.*"

Text-driven Image Generation and its Social Impact. There is a large volume of literature on generative models for image synthesis. Various approaches have been introduced, and most of these can be categorized into three different methods: (i) Generative Adversarial Networks (GAN)-based modeling [2, 7, 17, 23], which learns full data distribution with an efficient sampling of natural images, (ii) Variational AutoEncoders (VAE) [26] and flow-based models [9, 13, 14, 25, 49], which have advantages in the efficient generation of high-resolution images, and (iii) Diffusion Probabilistic Models [12, 19, 24, 45, 48], which are recently increasingly introduced and achieved state-of-the-art synthesizing results given its high generation power.

Most of these generative models focus mainly on generating high-resolution images with improved image quality, maintaining affordable computational costs [20, 28, 40, 47]. However, we observe in our experiment that these models often produce immoral images that clearly should not have been generated from an ethical perspective. Recently, there has been an effort to address such ethical concerns. For example, the state-of-the-art image generation model, Stable Diffusion [39], applies a so-called Safety Checker to filter inappropriate content to be generated. However, in our experiment, the current version of Stable Diffusion with

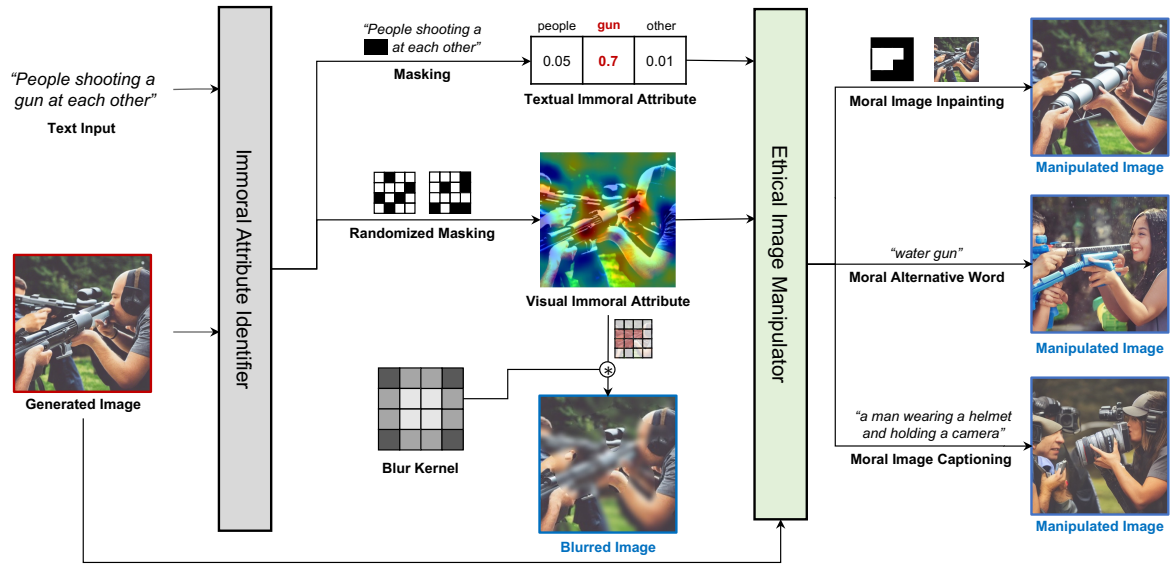

Figure 3. An overview of our proposed ethical image generation approach, which identifies immoral visual cues and edits an immoral input image into visually moral alternatives. Our model consists of three main modules: (1) Visual Commonsense Immorality Recognizer that judges the immorality of a given image (see Section 3.2), (2) Immoral Attribute Identifier that localizes immoral attributes that make the input image visually immoral (see Section 3.3), and (3) Ethical Image Manipulator that produces a moral image with three kinds of image manipulation approaches (see Section 3.4).

Safety Checker enabled often produce immoral images, as shown in Figure 2. Thus, our work starts from Stable Diffusion, and we propose a novel ethical image generation approach that localizes immoral visual cues and manipulates the *immorally* generated image into a *moral* one.

3. Method

3.1. Text-to-Image Generation

Various approaches in text-to-image generation have recently been introduced, focusing on producing high-quality images and allowing users to control the generation of specific visual attributes. As a black box, we utilize these off-the-shelf text-driven image generation models f_g that learn to synthesize a realistic image $\mathcal{I}$ given a sentence input $\mathcal{S}$ where their semantic features are aligned: $f_g : \mathcal{S} \rightarrow \mathcal{I}$. Among various approaches, we mainly use the latest Stable Diffusion [39] model based on a conditional diffusion model trained on a subset of a publicly available billion-scale multi-modal LAION-5B [43] dataset. As shown in Figure 2, we observe that the Stable Diffusion model generates morally inadequate images though it has an AI-based Safety Classifier included by default. Thus, as shown in Figure 3, we begin with Stable Diffusion and aim to improve the morality of outputs, filtering out undesired outcomes with ethical concerns.

3.2. Visual Commonsense Immorality Recognition

The *Visual Commonsense Immorality Recognizer* acts like a judge, determining the immorality of a given input image. Training such a judge, however, is challenging due to the lack of a large-scale, high-quality dataset for the visual commonsense immorality recognition task. Instead, as shown in Figure 4, following the recent work [21], we utilize a pre-trained (frozen) image-text joint embedding space, e.g., CLIP [34]. Given this, we first train an auxiliary text-based immorality classifier with the large-scale ETHICS dataset [18], which provides over 13,000 textual examples (e.g., "I punched my friend") and corresponding binary labels (i.e., immoral vs. moral). The immorality of an unseen image is recognized through the joint embedder and the trained immorality classifier in a zero-shot manner.

Formally, given an input text $\mathcal{T}$, we leverage the frozen CLIP [34]-based text encoder f_t followed by an immorality classifier f_c: $\hat{y} = f_c(f_t(\mathcal{T}))$, where the classifier is trained with Binary Cross-Entropy Loss (BCELoss) as follows:

$$\mathcal{L} = -\frac{1}{n}\sum_{i=1}^{n}[y_i \log \sigma(\hat{y}_i) + (1 - y_i) \log(1 - \sigma(\hat{y}_i))], \quad (1)$$

where $y_i \in \{0, 1\}$ for $i \in \{1, 2, \ldots, n\}$ represents the immorality target, and σ represents a sigmoid function. At inference time, we utilize the CLIP [34]-based image encoder f_v, which maps semantic text-image pairs close together in the joint embedding space. Thus, the final output for the unseen image $\mathcal{I}$ is defined as follows: $\hat{y} = f_c(f_v(\mathcal{I}))$.

3.3. Immoral Semantic Attribute Identification

Textual Immoral Attribute Identification by Masking. Our model localizes semantic immoral (visual or textual)

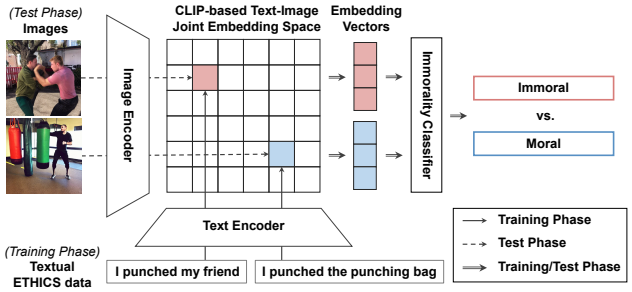

Figure 4. An overview of training visual commonsense immorality recognition model. Following [21], an classifier is trained to predict whether the input text prompt is moral or immoral.

attributes that make the image $\mathcal{I}$ visually immoral (e.g., a "gun" from a picture of people shooting a gun at each other). As our model is based on a text-to-image generator, we first identify word-level immorality using a masking approach to manipulate the generated immoral images into being visually moral, retaining other visual contexts. E.g., a given text prompt "people shooting a gun at each other" as an input, text-driven image generator may produce immoral scenes potentially due to the word "gun".

As shown in Figure 5 (a), to localize such words, we employ an input sampling approach, which measures the importance of a word by setting it masked and observing its effect on the model's decision. Formally, given a text-to-image model $f_g : \mathcal{T} \rightarrow \mathcal{I}$ and a visual commonsense immorality classifier $f_c : \mathcal{I} \rightarrow \mathbb{R}$, our model generates an image $\mathcal{I}'$ from the given input sentence $\mathcal{T} \in \{w_1, w_2, \dots\}$ as well as its visual immorality score $s \in [0, 1]$. We use a per-word binary mask $\mathcal{M}^T : |\mathcal{T}| \rightarrow \{0, 1\}$ to have masked input sentence $\mathcal{T}' = \mathcal{T} \odot \mathcal{M}^T$ where $\odot$ denotes element-wise multiplication. The importance score for each word w_i for $i \in \{1, \dots, |\mathcal{T}|\}$ is then computed as follows by taking an expectation over all possible masks $\mathcal{M}^T$ conditioned on the event that word w_i is observed:

$$s(w_i) = \mathbb{E}_{\mathcal{M}^T}[f_c(f_g(\mathcal{T} \odot \mathcal{M}^T))|\mathcal{M}^T(w_i) = 1], \quad (2)$$

where we can obtain an importance map by summing over a set of masks $\{\mathcal{M}_1^T, \dots, \mathcal{M}_K^T\}$ with weights $f_c(f_g(\mathcal{T} \odot \mathcal{M}_k^T))$. Note that any information from the original image-to-text generation models is not used; thus, this method can easily be applied to other text-to-image generation models.

Visual Immoral Attribute Identification by Randomized Masking. Similar to textual immoral attribute identification, we extend it to visual immoral attribute identification to localize which visual attributes contribute to making the image $\mathcal{I}$ visually immoral. As shown in Figure 5 (b), we employ a randomized input sampling approach [33] that can measure the importance of an image region by setting it masked and observing its effect on the model's decision. Formally, given a visual commonsense immorality classifier $f_c : \mathcal{I} \rightarrow \mathbb{R}$, we use a randomized binary mask $\mathcal{M}_i^I$

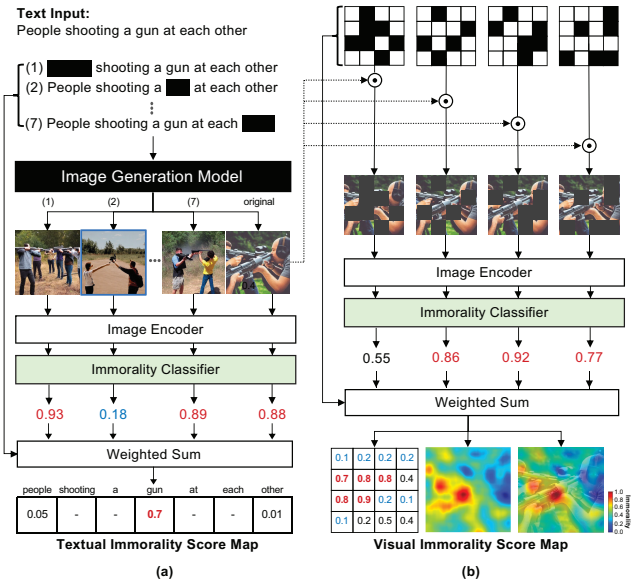

Figure 5. An overview of our (a) textual and (b) visual immoral attribute identification.

to have masked input image $\mathcal{I}' = \mathcal{I} \odot \mathcal{M}^I$ where $\odot$ denotes element-wise multiplication. The importance score for each image region x_i for $i \in \{1, \dots, W \times H\}$ is then computed as follows by taking summation over masks $\mathcal{M}^I$ using Monte Carlo sampling:

$$s(x_i) = \frac{1}{P[\mathcal{M}^I(x_i) = 1]} \sum_{k=1}^{K} f_c(\mathcal{I} \odot \mathcal{M}_k^I) \cdot \mathcal{M}_k^I(x_i), \quad (3)$$

where we similarly can obtain an importance map by summing over a set of masks $\{\mathcal{M}_1^I, \dots, \mathcal{M}_K^I\}$ with weights $f_c(\mathcal{I} \odot \mathcal{M}_k^I)$.

Blurring Immoral Visual Semantic Cues. As shown in Figure 6 (a), our model outputs an image with immoral visual contents blurred (e.g., blurring a gun from a scene of people shooting a gun at each other) with standard blur kernel functions such as Gaussian kernel. Given the normalized per-pixel visual immorality scores $s(x_i)$, we first divide image regions into moral and immoral based on a user-specified threshold. Note that we apply a blur kernel function only to pixels in immoral image regions to have blurred immoral visual contents.

3.4. Ethical Image Manipulation

Lastly, we introduce various image manipulation approaches to produce a moral image by automatically replacing immoral visual cues. Here, we explore three kinds of image manipulation approaches. (i) Immoral Object Replacement by Moral Image Inpainting. Instead of making blurry images, we apply an image inpainting technique to replace immoral objects with moral alternatives. (ii) Text-driven Image Manipulation with Moral Words. Our model

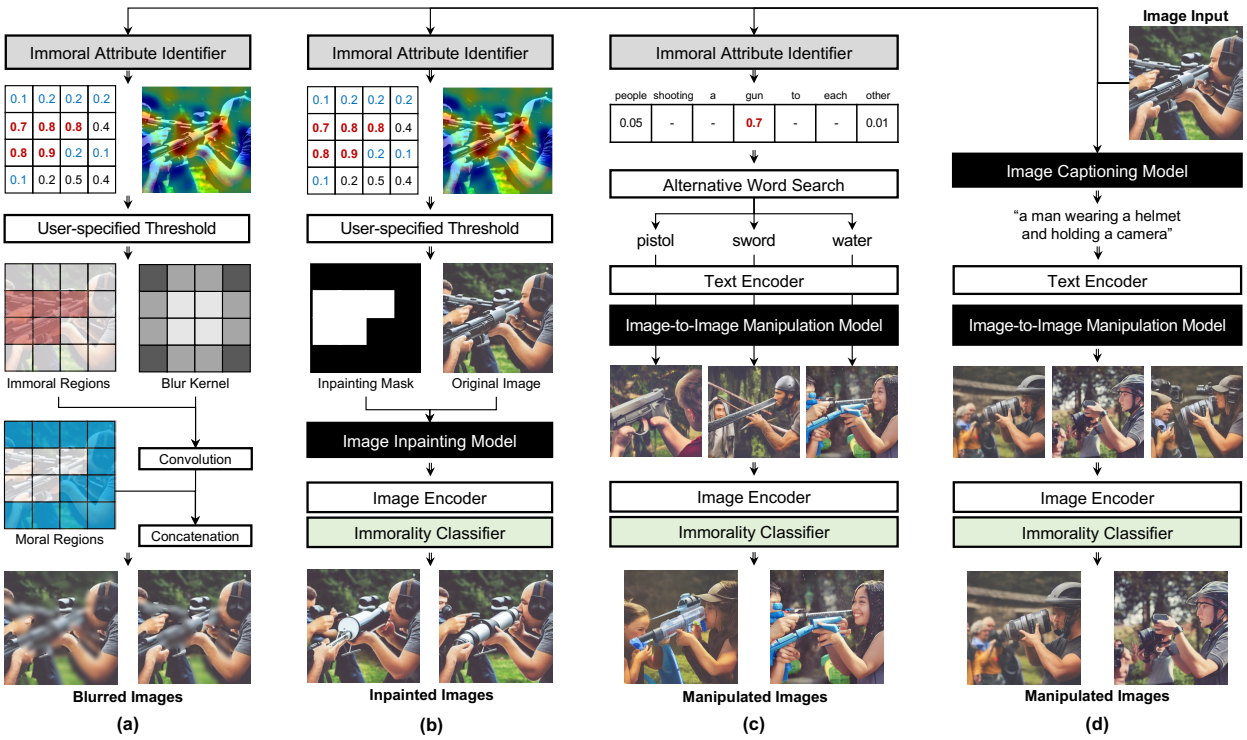

Figure 6. An overview of (a) blurring immoral visual semantic attributes and (b-d) three kinds of ethical image manipulation methods: (b) Immoral Object Replacement by Moral Image Inpainting, (c) Text-driven Image Manipulation with Moral Words, and (d) Text-driven Image Manipulation with Moral Image Captions.

searches for word candidates (e.g., "water") that is conditioned to manipulate an input image (e.g., "people shooting a gun at each other") into moral scenes (e.g., "people shooting a water gun at each other"). (iii) Text-driven Image Manipulation with Moral Image Captions. We utilize pretrained image captioning models that are trained with moral datasets; thus, they learn to generate moral image captions even for immoral images. For example, they create the caption "a man wearing a helmet and holding a camera" for an image of people shooting a gun at each other. Text-driven image manipulator produces moral images accordingly.

Replacing Immoral Object by Moral Image Inpainting. Image inpainting models are often used to restore missing regions in an image. They have many applications in image editing, such as removing objects by synthesizing semantically plausible and visually realistic pixels, keeping coherency with existing content. Such inpainting approaches are also applicable to remove immoral objects and complete their pixels with moral ones. Given the visual immorality score map, we remove immoral regions (set pixel values to zero) that need to be restored and apply an off-the-shelf image inpainting approach. We summarize details in Figure 6 (b), where our image inpainting model replaces a gun with a telescope; thus, the image is morally manipulated.

Immoral Word Replacement with Moral Alternatives. Our model identifies a set of words that contributes to generating immoral images. Another intuition toward ethical image manipulation would be using existing conditional image manipulation models with a word, driving the model to generate a more moral image. For example, as shown in Figure 6 (c), we search for a word (e.g., water) that will be conditioned to reduce the output's immorality. Finding such a word is challenging as it only needs to modify the immoral contents, while keeping the original unrelated contents remain the same. In our experiment, we use Google's suggested search results, which reflect real searches that have been done on Google related to the query. Moreover, immoral suggested queries are filtered out due to Google Search's policy to prevent harassing, hateful, sexually explicit, and immoral content.

Text-driven Image Manipulation with Moral Image Captioning. Given an image captioning model trained with a highly-curated dataset where immoral pictures and texts are filtered out (e.g., MS-COCO [30] though it contains a few images with immoral contents), a description of an immoral image is obtainable from a moral perspective. As shown in Figure 6 (d), we observe that a captioning model trained with the MS-COCO dataset generates "a man wearing a helmet and holding a camera" for a scene of people shooting a gun at each other. Using this morally-described caption as a condition for the text-driven image manipulation model, we obtain a morally-manipulated scene that

Figure 7. Textual and visual immoral attribute identification examples. We provide the initially generated images (top), the word-level textual immoral attributes (words highlighted in green), and the immorality score maps (bottom) generated by our model.

does not differ much from the original scene.

Identity Loss. To alleviate the excessive manipulation from the original prompt, we introduce identity loss $\mathcal{L}_{identity}$ at inference time based on the spherical distance. Given prompt $\mathcal{T}$ and image $\mathcal{I}_t$ at denoising step t, identity loss $\mathcal{L}_{identity}$ is defined as follows:

$$\mathcal{L}_{identity} = 2 \times \arcsin^2 \left(\frac{\|\mathbf{u}_x - \mathbf{u}_y\|}{2}\right), \tag{4}$$

where $\mathbf{u}_x$ is an unit vector of $\mathbf{x} = f_v(\mathcal{I}_t)$, and $\mathbf{u}_y$ is an unit vector of $\mathbf{y} = f_t(\mathcal{T})$. Note that f_v and f_t are CLIP [34]-based visual and textual encoder, as mentioned in section 3.2. Finally, we calculate the gradient of $\mathcal{L}_{identity}$ and update the noise vector to minimize the loss:

$$\mathbf{z}_t \leftarrow \mathbf{z}_t - \alpha \nabla_{\mathbf{z}_t} \mathcal{L}_{identity}(\mathbf{z}_t), \tag{5}$$

where α is an scale factor, and $\mathbf{z}_t$ is a noise vector.

4. Experiments

Implementation Details. Our model utilizes the CLIP-based [34] textual and image encoders with ViT-B/32 backbone, which use contrastive learning to learn a visual-textual joint representation. Following Hendrycks *et al*. [18], which classifies immoral vs. moral from text inputs, we use an MLP to build our immorality classifier f_c. We provide other implementation details (including architecture and hyperparameters) in the supplemental material.

Datasets. The key module for ethical image manipulation involves judging the immorality of given images (or texts). Due to the lack of large-scale datasets available to learn visual commonsense immorality, we train our model using the textual ETHICS Commonsense Morality [18] dataset. Transferring retained knowledge from texts to visual data is achieved through the utilization of a joint embedding space. Additionally, we employ four existing datasets to evaluate the classification performance of model's ability to judge the visual commonsense immorality: (1) MS-COCO [30], (2) Socio-Moral Image [11], (3) Sexual Intent Detection [16], and Real Life Violence Situation [46]. Further details about these datasets are explained in the supplemental material.

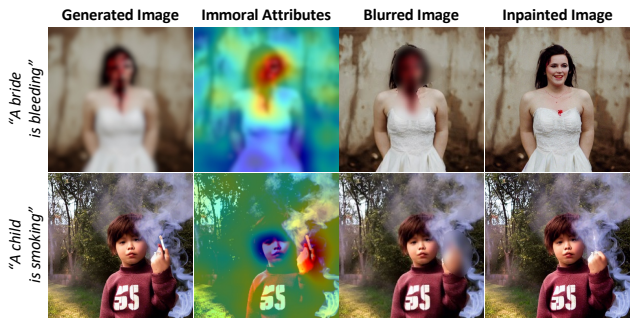

Figure 8. Localization of immoral visual semantic cues (2nd column) and manipulation results through blurring (3rd column) and moral image inpainting (4th column).

4.1. Qualitative Analysis

Analysis of Immoral Attribute Identification. As shown in Figure 7, we first observe that our baseline, Stable Diffusion, produces immorally generated images (see top row). Note that this model enables a so-called Safety Checker to filter out images with ethical and moral concerns. Given these immoral images as input, we apply our module and visualize the image-based immorality score map (see bottom row). Our module reasonably highlights immoral objects, such as localizing cigarettes, blood, and a gun. Additionally, by utilizing the visual immorality score maps, our model successfully blurs the localized content as shown in Figure 8 (see 3rd column). These examples demonstrate our model's ability to localize immoral visual cues. We provide more diverse examples in the supplemental material.

Further, we apply the textual immorality attribute identification module to identify word-level immorality. As shown in Figure 7, our model can highlight a set of words that drive the text-driven image generator to produce immoral scenes. For example, an image generated from "A child is smoking" is classified as immoral due to the word "smoking", while an image from "I shot my gun into the crowd" is mainly due to the word "gun".

Analysis of Immoral Object Replacement by Moral Image Inpainting. Blurring immoral visual content may be a simple but effective way for ethical image manipulation. However, it would be difficult to balance visual acuity vs. immorality and to avoid guessing the original immoral scenes from blurred images. Thus, instead of blurring, we further explore replacing immoral visual attributes with moral content using image inpainting models, i.e. reconstructing immoral image regions in an image so that the filled-in image becomes morally classified. In Figure 8 (last column), we provide manipulated outputs from our moral image inpainting approach. The inpainting model successfully replaces immoral visual attributes with moral contents, such as bleeding blood on the face being replaced by a smiling face. To produce these results, we use an off-the-shelf

Figure 9. Ethically manipulated images with moral image captioning. Note the similarities in hue and composition between the original image and the manipulated image.

Figure 10. Examples of image manipulation where immoral words ("sword" and "gun") are identified and replaced by moral alternatives ("fantasy sword" and "water gun").

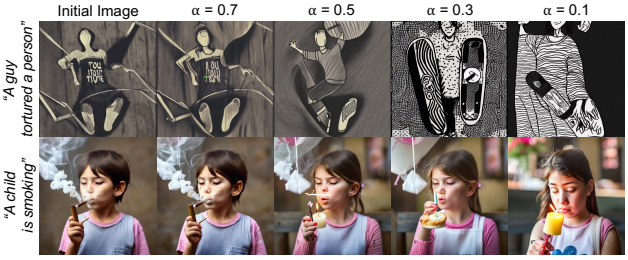

Figure 11. Manipulation results with different scaling factors α for identity loss. As α decreases, manipulated images deviate more from the initial prompts (from left to right).

image inpainting model [39] that fills immoral regions of an image with moral content.

Analysis of Text-driven Image Manipulation with Moral Image Captioning. In addition to leveraging the image inpainting model, another way would be utilizing an image captioning model trained with a highly-curated dataset where immoral images and texts are filtered out (e.g. MS-COCO dataset). Examples of this approach are shown in Figure 9. Given immoral images generated by Stable Diffusion, we apply the off-the-shelf image captioning model [31] that is trained with the MS-COCO dataset (see example outputs in 2nd column). This produces descriptive captions from a moral perspective. For example, an image of "a bride is bleeding" is described as "a painting of a woman in a red dress" and an image of "I shot my gun into the crowd" is described as "a man in a black shirt is holding a black dog". Using these generated captions as a condition, we can successfully manipulate them into a moral scene (compare 1st vs. last two columns). In some cases, we observe that such an image captioning model may produce an immoral description (see supplemental material). A further use of our textual immorality recognizer would solve this concern by filtering those sentences out.

Analysis of Replacing Immoral Words with Moral Alternatives. Figure 10 shows examples of image manipulation by replacing immoral words with moral alternatives. For example, given a text input, "A baby holding a sword," the image generator produces the corresponding image without ethical screening (see 1st row). Our immoral attribute identifier highlights the word "sword" contributes to the generated image being classified as immoral, and our module searches for an alternative word (e.g., "fantasy") that can be additionally conditioned to manipulate the given image with reduced immorality. The alternative word provided manipulates the generated immoral image into being more moral (see two right columns). Similarly, the text "a child with a gun" with the word "water" added produces manipulated images, reducing their immorality.

Analysis of Identity Loss. Lastly, we conduct an experiment to show the impact of identity loss by varying the scaling factor α. As shown in Figure 11, various factors, including composition, texture, and background, change as α decreases (i.e., resulting in a reduced effect of identity loss).

4.2. Quantitative Analysis

Zero-shot Visual Commonsense Immorality Prediction. In Table 1, we evaluate the classification performance of our model's ability to judge the visual commonsense immorality on four existing datasets (see supplemental material). Built upon Jeong *et al*. [21], we optimize the model by (i) tuning hyperparameters and (ii) utilizing L2-norm-based features, yielding a performance gain in all datasets against the existing approach.

Analysis of Immorality Classifier. To explain our design criteria of the immorality classifier, we perform two experiments using (i) a different image-text joint embedding space, and (ii) different CLIP backbones. For the different joint embedding space, we utilize the ALIGN [22] model trained on the COYO-700M dataset [8]. As shown in Table 1, the CLIP-based immorality classifier with ViT-B/32 backbone shows the best performance in three datasets. Considering the fact that the CLIP model is trained on much less data compared to the ALIGN model (i.e., 400M vs. 700M), this result is quite interesting. We argue that

Table 1. Comparison of zero-shot visual commonsense immorality prediction performance against the existing approach (2nd vs. last column), between different joint-embedding space (3rd vs. 4th-6th columns), and different backbones (4th-6th columns). Following [21], we use the F-measure with a beta value of 2 to evaluate performance across four publicly available datasets.

Dataset	Jeong *et al.* [21]	ALIGN [22]	CLIP [34]		
			ViT-L/14	ViT-B/16	**ViT-B/32**
MS-COCO [30]	0.688 (0.128↓)	0.798	0.725	0.683	**0.816**
Socio-Moral Image Database [11]	0.591 (0.030↓)	0.620	0.446	0.599	**0.621**
Sexual Intent Detection Images [16]	0.434 (0.298↓)	**0.774**	0.377	0.634	0.732
Real Life Violence Situation [46]	0.807 (0.007↓)	0.793	0.714	0.599	**0.815**

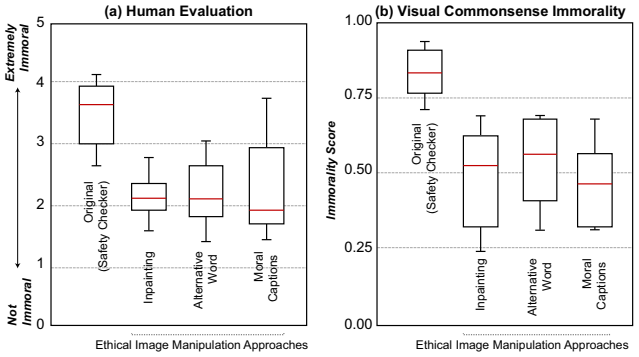

Figure 12. Box plot of (a) our human evaluation results and (b) visual commonsense immorality score from our recognizer.

this phenomenon is mainly due to the high curation of the dataset, which may degrade the generalization ability of the joint embedding space with respect to commonsense morality [42]. We provide more interpretations about this in the supplemental material.

Human Evaluation. We further conduct a human study to demonstrate whether our generated images are indeed morally manipulated. As shown in Figure 12 (a), we recruited 178 human evaluators, and we asked them to judge the immorality of each generated image on a Likert scale from 1 (not immoral) to 5 (extremely immoral). We compare scores between originally generated images by Stable Diffusion (with Safety Checker enabled) and manipulated images from our three approaches (i.e. inpainting, alternative word, and moral captions). Compared to originally generated image, all approaches significantly reduce perceived immorality. Especially an inpainting-based method shows the best performance in ethical image manipulation. This confirms that our morally manipulated images are more morally perceived than the original ones. In Figure 12 (b), we experiment with our visual commonsense immorality recognizer to compute immorality scores for each image. We observe trends similar to our human evaluation, and this further confirms that our visual commonsense immorality recognizer matches human perception.

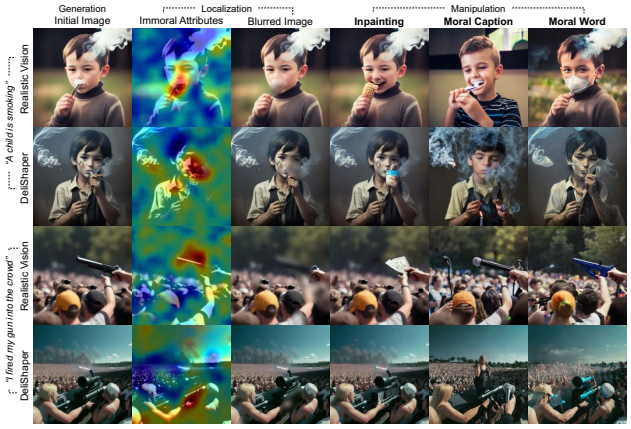

Figure 13. Localization and manipulation results with two different models: (i) Realistic Vision [15], and (ii) DeliShaper [51]. Our model successfully localizes and manipulates immoral visual cues such as cigarettes and guns.

4.3. Application to Other Text-to-Image Models

To confirm the generalizability of our method, we conduct evaluations with two other text-to-image models, Realistic Vision [15] and DeliShaper [51]. As shown in Figure 13, we observed that both models generate immoral images when provided with inappropriate prompts, such as "A child is smoking" (1st column). Next, our model successfully localizes immoral visual cues (2nd column) and blurs those cues, including cigarettes and guns (3rd column). Finally, the three methods effectively manipulate the original image (4th-6th columns). For example, an image of a child holding a cigarette is transformed into an image of a child (i) holding an ice cream (inpainting), (ii) brushing his teeth (moral caption), and (iii) wearing a mask (moral word).

5. Conclusion

In this paper, we introduced a method to manipulate an immorally generated image (which should not have been generated due to ethical concerns) into a moral one where immoral contents are localized and replaced by a moral alternative attribute. We presented three essential modules: judging visual commonsense immorality, localizing input-level immoral attributes, and producing morally-satisfying manipulation images. Our human study and detailed analysis demonstrate the effectiveness of our proposed ethical image manipulation model.

Acknowledgements. This work was supported by Basic Science Research Program (NRF-2021R1A6A1A13044830, 25%), by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (2022-0-00043 (50%)), and by the MSIT(Ministry of Science and ICT), Korea, under the ITRC(Information Technology Research Center) support program (IITP-2023-RS-2022-00156295, 25%).

References

[1] Shervin Ardeshir, Cristina Segalin, and Nathan Kallus. Estimating structural disparities for face models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10358–10367, 2022. 2

[2] Martin Arjovsky, Soumith Chintala, and Léon Bottou. Wasserstein generative adversarial networks. In *International conference on machine learning*, pages 214–223. PMLR, 2017. 2

[3] Isaac Asimov. Three laws of robotics. *Asimov, I. Runaround*, 1941. 2

[4] Edmond Awad, Sohan Dsouza, Richard Kim, Jonathan Schulz, Joseph Henrich, Azim Shariff, Jean-François Bonnefon, and Iyad Rahwan. The moral machine experiment. *Nature*, 563(7729):59–64, 2018. 2

[5] Nick Bostrom. Ethical issues in advanced artificial intelligence. *Science fiction and philosophy: from time travel to superintelligence*, 277:284, 2003. 2

[6] Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012. 2

[7] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2

[8] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset, 2022. 7

[9] Rewon Child. Very deep vaes generalize autoregressive models and can outperform them on images. *arXiv preprint arXiv:2011.10650*, 2020. 2

[10] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017. 2

[11] Damien L. Crone, Stefan Bode, Carsten Murawski, and Simon M. Laham. The socio-moral image database (smid): A novel stimulus set for the study of social, moral and affective processes. *PLOS ONE*, 13:1–34, 01 2018. 6, 8

[12] Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in Neural Information Processing Systems*, 34:8780–8794, 2021. 2

[13] Laurent Dinh, David Krueger, and Yoshua Bengio. Nice: Non-linear independent components estimation. *arXiv preprint arXiv:1410.8516*, 2014. 2

[14] Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016. 2

[15] Evgeny. Realistic vision v1.4. https://huggingface.co/SG161222/Realistic_Vision_V1.4, 2023. 8

[16] Debashis Ganguly, Mohammad H Mofrad, and Adriana Kovashka. Detecting sexually provocative images. In *WACV*, pages 660–668. IEEE, 2017. 6, 8

[17] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020. 2

[18] Dan Hendrycks, Collin Burns, Steven Basart, Andrew Critch, Jerry Li, Dawn Song, and Jacob Steinhardt. Aligning ai with shared human values. *ICLR*, 2021. 2, 3, 6

[19] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020. 2

[20] Jonathan Ho, Chitwan Saharia, William Chan, David J Fleet, Mohammad Norouzi, and Tim Salimans. Cascaded diffusion models for high fidelity image generation. *J. Mach. Learn. Res.*, 23:47–1, 2022. 2

[21] Yujin Jeong, Seongbeom Park, Suhong Moon, and Jinkyu Kim. Zero-shot visual immorality prediction. *BMVC*, 2022. 1, 3, 4, 7, 8

[22] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 7, 8

[23] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2

[24] Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in neural information processing systems*, 34:21696–21707, 2021. 2

[25] Durk P Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions. *Advances in neural information processing systems*, 31, 2018. 2

[26] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013. 2

[27] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016. 2

[28] Zhifeng Kong and Wei Ping. On fast sampling of diffusion probabilistic models. *arXiv preprint arXiv:2106.00132*, 2021. 2

[29] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 426–434, 2008. 2

[30] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014. 5, 6, 8

[31] nlpconnect. vit-gpt2-image-captioning. https://huggingface.co/nlpconnect/vit-gpt2-image-captioning/tree/main, 2022. 7

[32] Sungho Park, Jewook Lee, Pilhyeon Lee, Sunhee Hwang, Dohyung Kim, and Hyeran Byun. Fair contrastive learning for facial attribute classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10389–10398, 2022. 2

[33] Vitali Petsiuk, Abir Das, and Kate Saenko. Rise: Randomized input sampling for explanation of black-box models. *arXiv preprint arXiv:1806.07421*, 2018. 4

[34] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021. 1, 3, 6, 8

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021. 1

[36] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*, 2018. 2

[37] Alex Ray, Joshua Achiam, and Dario Amodei. Benchmarking safe exploration in deep reinforcement learning. *arXiv preprint arXiv:1910.01708*, 7:1, 2019. 2

[38] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*, 2020. 2

[39] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1, 2, 3, 7

[40] Robin San-Roman, Eliya Nachmani, and Lior Wolf. Noise estimation for generative diffusion models. *arXiv preprint arXiv:2104.02600*, 2021. 2

[41] Michael J Sandel. Justice: What's the right thing to do. *BUL Rev.*, 91:1303, 2011. 2

[42] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22522–22531, 2023. 1, 2, 8

[43] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*, 2022. 3

[44] Kirill Sirotkin, Pablo Carballeira, and Marcos Escudero-Viñolo. A study on the distribution of social biases in self-supervised learning visual models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10442–10451, 2022. 2

[45] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015. 2

[46] Mohamed Mostafa Soliman, Mohamed Hussein Kamal, Mina Abd El-Massih Nashed, Youssef Mohamed Mostafa, Bassel Safwat Chawky, and Dina Khattab. Violence recognition from videos using deep learning techniques. In *2019 Ninth International Conference on Intelligent Computing and Information Systems (ICICIS)*, pages 80–85. IEEE, 2019. 6, 8

[47] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[48] Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020. 2

[49] Arash Vahdat and Jan Kautz. Nvae: A deep hierarchical variational autoencoder. *Advances in Neural Information Processing Systems*, 33:19667–19679, 2020. 2

[50] Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *European conference on computer vision*, pages 322–339. Springer, 2020. 2

[51] Yntec. Delishaper. https://huggingface.co/Yntec/DeliShaper, 2023. 8