SafeR-CLIP: Mitigating NSFW Content in Vision-Language Models While Preserving Pre-Trained Knowledge

Adeel Yousaf[1], **Joseph Fioresi**[1], **James Beetham**[1], **Amrit Singh Bedi**[2], **Mubarak Shah**[1]

[1]Center for Research in Computer Vision, University of Central Florida, USA
[2]SAFERR AI Lab, University of Central Florida, USA
{adeel.yousaf, joseph.fioresi, james.beetham, amritbedi}@ucf.edu, shah@crcv.ucf.edu

Abstract

Improving the safety of vision-language models like CLIP via fine-tuning often comes at a steep price, causing significant drops in their generalization performance. We find this trade-off stems from rigid alignment strategies that force unsafe concepts toward single, predefined safe targets, disrupting the model's learned semantic structure. To address this, we propose a proximity-aware approach: redirecting unsafe concepts to their semantically closest safe alternatives to minimize representational change. We introduce `SafeR-CLIP`, a fine-tuning framework that applies this principle of minimal intervention. `SafeR-CLIP` successfully reconciles safety and performance, recovering up to 8.0% in zero-shot accuracy over prior methods while maintaining robust safety. To support more rigorous evaluation, we also contribute NSFW-Caps, a new benchmark of 1,000 highly-aligned pairs for testing safety under distributional shift. Our work shows that respecting the geometry of pretrained representations is key to achieving safety without sacrificing performance.

Code — https://adeelyousaf.github.io/SC-Project-Page/

Introduction

Large-scale web-scraped datasets such as LAION-5B (Schuhmann et al. 2022) have been instrumental in pretraining vision-language models (VLMs) with remarkable generalization capabilities (Radford et al. 2021). However, the uncurated nature of this data introduces significant safety concerns, as models can learn to generate inappropriate or Not Safe For Work (NSFW) content (Poppi et al. 2024a; Birhane et al. 2023). Ensuring safe VLM behavior is therefore a critical priority, especially for deployment in sensitive domains like healthcare and autonomous systems (Steed and Caliskan 2021; Wu et al. 2024). To address safety concerns in VLMs, current methods typically follow one of two strategies: (1) removing unsafe data prior to training or (2) fine-tuning the model post-training to enforce safer behavior. While filtering unsafe data beforehand can be effective, it is computationally infeasible at the scale of datasets like LAION-5B, which are essential for achieving foundation model capabilities. Post-training fine-tuning, by contrast, modifies learned representations to steer the model away from unsafe outputs, typically by aligning embeddings toward safer directions (Poppi et al. 2024a). While effective at reducing harmful content, current alignment techniques often create an unresolved tension between safety and performance. For instance, prominent methods can incur a substantial 22% drop in zero-shot accuracy on standard benchmarks after safety fine-tuning (Poppi et al. 2024a). This raises a fundamental question: *Can we align models for safety without dismantling their rich, pretrained knowledge?* [1]

We note that prior methods (see Figure 1), such as SafeCLIP (Poppi et al. 2024a), treat unsafe concepts as singular entities, aligning them with a fixed safe counterpart. We find that this performance degradation—particularly the drop in generalization—may stem from an underlying assumption in current alignment methods. These techniques typically rely on a fixed mapping, aligning each unsafe concept with a single, pre-defined safe counterpart (Figure 1). This approach, however, may not fully capture the contextual nature of meaning, where an unsafe concept can correspond to multiple semantically valid and safe interpretations. By enforcing alignment to a single, sometimes weakly correlated target, such methods can inadvertently disrupt the model's learned semantic structure, treating other plausible safe concepts as negatives and degrading generalization.

Our Key Idea. To overcome these limitations, we introduce a new guiding principle for safety alignment: ***proximity-aware realignment***. Instead of imposing a rigid mapping, our approach embraces the geometry of the model's own embedding space. The core idea is to perform minimal intervention: for any unsafe input, we identify its semantically closest safe alternative and gently redirect the representation toward this contextually appropriate target. This strategy respects the model's pretrained knowledge, allowing for safety improvements that co-exist with, rather than compete against, generalization.

To this end, we propose `SafeR-CLIP`, a framework that employs two novel, representation-aware losses. The first, ***relative cross-modal redirection***, refines contrastive learning by specifying the unsafe representation as the sole negative, which preserves the rich semantic associations be-

[1]See the extended arXiv version of this paper for additional analysis and qualitative results.

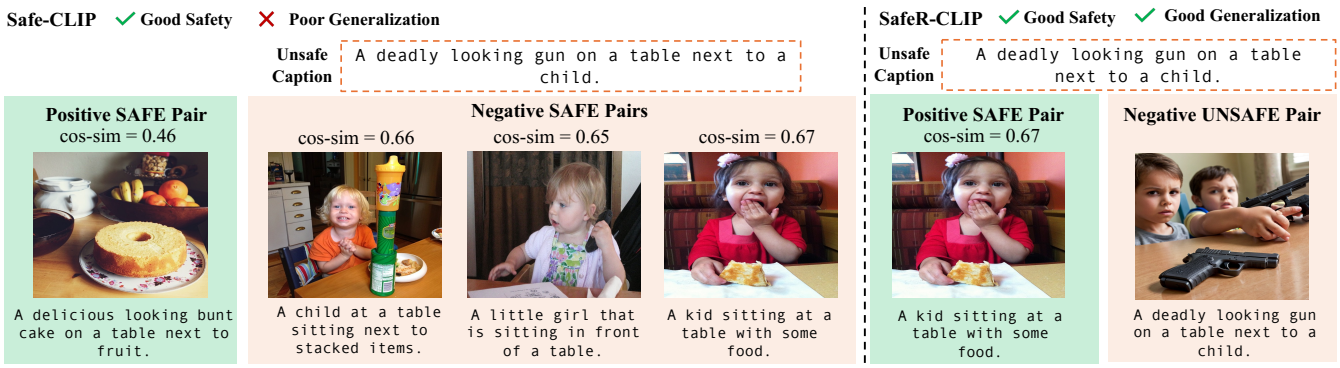

Figure 1: An unsafe concept can have multiple semantically valid safe alternatives. In this example, the unsafe caption *"A deadly looking gun on a table next to a child"* could plausibly align with safe counterparts such as *"A kid sitting at a table with some food"* or *"A child at a table sitting next to stacked items"*, which preserve the underlying semantics while removing unsafe elements. **(Left)** Safe-CLIP (Poppi et al. 2024a) enforces a rigid alignment between the unsafe input and a single predefined safe caption (e.g., *"A delicious looking bunt cake on a table next to fruit"*), while treating other valid alternatives as potential negatives. This leads to two major issues: (1) due to the noisy nature of existing datasets like ViSU (Poppi et al. 2024a), the selected unsafe–safe pair may be semantically misaligned, as shown; and (2) semantically closer safe alternatives are incorrectly penalized. **(Right)** Our method, `SafeR-CLIP`, addresses these limitations by aligning each unsafe input with its most semantically compatible safe counterpart while pushing it away from the unsafe embedding—ensuring better safety–generalization trade-off.

tween all related safe concepts. The second, ***proximity-based alignment***, mitigates issues from noisy training pairs by dynamically identifying the most semantically compatible safe target for each unsafe input. To further stabilize learning, `SafeR-CLIP` also incorporates a progressive training strategy that introduces pairs based on increasing difficulty. We summarize our key contributions as follows:

- **Identification of an overlooked challenge:** We identify a key limitation in prior work—the assumption of a single safe counterpart per unsafe concept—and show that aligning to semantically closest safe alternatives reduces representational shift and preserves generalization.
- **Representation-aware training losses:** We propose two novel losses: *relative cross-modal redirection*, which corrects negative selection, and *proximity-based alignment*, which aligns unsafe inputs to semantically compatible safe targets—jointly improving safety with minimal disruption to pretrained representations.
- **New benchmark for robust safety evaluation:** We introduce NSFWCaps, a cross-modal safety benchmark with 1,000 highly-aligned safe–unsafe pairs. Built using the out-of-domain NoCaps dataset (Agrawal et al. 2019a), it provides a more rigorous test of safety generalization under distributional shift compared to existing benchmarks.
- **State-of-the-art performance:** `SafeR-CLIP` sets new state-of-the-art results on safety and retrieval benchmarks, improving zero-shot performance by 8% over prior safety fine-tuning approaches while maintaining comparable safety.

Related Works

The removal of harmful content from AI models has become an increasing focus of research (Wang et al. 2022; Steed and Caliskan 2021; Larrazabal et al. 2020; Denton et al. 2021), particularly in vision-language systems trained on large-scale, web-crawled datasets such as LAION-5B (Schuhmann et al. 2022, 2021). These datasets often lack proper curation, allowing problematic content to persist (Birhane et al. 2024). As a result, recent studies (Wan et al. 2024; Schramowski, Tauchmann, and Kersting 2022; Zong et al. 2024a; Gou et al. 2024; Wang et al. 2024; Ding, Li, and Zhang 2025; Liu et al. 2024; Zou et al. 2025; Ghosal et al. 2025; Xu et al. 2025) have explored various mitigation strategies to address these safety concerns in generative vision-language models.

Recent attempts address NSFW safety in generative text-to-image and image-to-text models through dataset filtering, inference-time guidance, fine-tuning, and unlearning. Filtering removes harmful content before training (Rombach et al. 2022a) but requires substantial computational resources for large-scale retraining. Inference-time approaches modify model behavior during generation. For text-to-image models, (Schramowski et al. 2023a) applies negative guidance to suppress NSFW outputs, while Buster (Zhao et al. 2024) introduces a semantic backdoor to redirect unsafe prompts. In large vision-language models (LVLMs) for image-to-text generation, inference-time defenses such as LLaVA-Guard (Helff et al. 2024) and Zero-Shot Safety (Zhao et al. 2025) aim to detect and filter harmful text responses. Fine-tuning explicitly removes NSFW concepts by training pre-trained models on safe content. ESD (Gandikota et al. 2023) suppresses conditioned responses, while ShieldDiff (Han, Mohamed, and Li 2024) optimizes a CLIP-based reward for safety filtering in image generation tasks. Similarly, fine-tuning has been explored for LVLMs such as LLaVA (Liu et al. 2023) to improve safe image-to-text generation (Zong et al. 2024b). Machine unlearning selectively removes harmful information while preserving generation quality (Ginart

et al. 2019; Poppi et al. 2024b; Golatkar, Achille, and Soatto 2020; Cao and Yang 2015; Zhang et al. 2025). Methods like Saliency SalUn (Fan et al. 2023) refine generative models by modifying only the most relevant parameters, while (Li et al. 2024), propose efficient unlearning methods for LVLMs by fine-tuning on minimal data, ensuring targeted forgetting without degrading overall performance.

Mitigating NSFW risks in CLIP. Unlike prior NSFW mitigation efforts in generative models, our work focuses on making CLIP-like (Radford et al. 2021) contrastive models safer. CLIP plays a crucial role in cross-modal retrieval, zero-shot classification, and as a backbone for text-to-image and image-to-text generation, making its safety essential for real-world applications. Unlike generative models, where safety measures can be applied at inference time, CLIP's role in feature extraction means that unsafe biases can persist across multiple downstream tasks if not addressed at the embedding level.

Efforts to make CLIP safer have recently explored both training-free and training-based strategies. Recently, UWM (D'Incà et al. 2025) proposes a lightweight approach that manipulates unsafe weights at inference time to suppress NSFW features. While efficient, this method offers limited improvements in safety and lacks generalization across tasks. In contrast, Safe-CLIP (Poppi et al. 2024a) introduces a training-based fine-tuning strategy that redirects unsafe embeddings toward predefined safe regions. Although it reduces unsafe retrieval and generation, Safe-CLIP induces a substantial 22% drop in zero-shot classification performance, revealing a strong trade-off between safety and generalization. We identify that part of this limitation stems from misaligned or noisy supervision, where semantically distant safe–unsafe pairs and uniform treatment of negatives can disrupt the learned feature space. Our method overcomes this by leveraging proximity-based alignment and relative redirection, which selectively guide unsafe samples toward semantically compatible safe counterparts—preserving generalization while enhancing safety alignment. Other recent efforts explore safety alignment in non-Euclidean (e.g., hyperbolic) spaces (Poppi et al. 2025), though such approaches are currently incompatible with downstream applications.

Problem Formulation

Key notations. The foundational CLIP (Radford et al. 2021) model aligns text and images in a shared embedding space using a text encoder $\mathcal{T}(\cdot)$ and image encoder $\mathcal{V}(\cdot)$. Standard CLIP training relies on paired, ideally safe, text-image data $(t_i, v_i) \in \mathbf{T} \times \mathbf{V}$, where $t_i \in \mathbf{T}$ denotes the i-th text sample and $v_i \in \mathbf{V}$ denotes the corresponding image sample. Here, $\mathbf{T} = \{t_1, t_2, \ldots, t_M\}$ represents the set of all text captions, and $\mathbf{V} = \{v_1, v_2, \ldots, v_M\}$ represents the set of all corresponding images in the dataset, where M is the total number of text-image pairs. The ViSU dataset (Poppi et al. 2024a) extends the paired data (t_i, v_i) by introducing an unsafe text-image pair $(t_i^*, v_i^*) \in \mathbf{T}^* \times \mathbf{V}^*$ (set of all unsafe text-image pairs), forming unified quadruplets (t_i, v_i, t_i^*, v_i^*). Each quadruplet contains a safe text-image pair and a corresponding unsafe pair, which is generated to closely resemble the safe data augmented with unsafe content. The goal of the safety fine-tuning objective is to adapt the encoders to remove NSFW-related information while preserving the rich semantic structure learned during pre-training. Formally, this is achieved by aligning the embeddings of unsafe inputs, $\mathcal{T}(t_i^*)$ and $\mathcal{V}(v_i^*)$, with their corresponding safe counterparts, $\mathcal{T}(t_i)$ and $\mathcal{V}(v_i)$. The fine-tuning process also incorporates a regularization term based on embeddings from reference (frozen) CLIP encoders $\mathcal{T}_0(\cdot)$ and $\mathcal{V}_0(\cdot)$. This ensures the resulting embedding space retains structural similarity to that of the original CLIP model. The embedding relationship is measured using cosine similarity, denoted as $\cos(\cdot, \cdot)$. With these definitions in place, we next review the Safe-CLIP methodology, which provides one approach to this problem, and analyze its key limitations.

Revisiting Safe-CLIP

In Safe-CLIP (Poppi et al. 2024a), the model is fine-tuned on the quadruplet (v_i, t_i, v_i^*, t_i^*) that contains a safe image, its safe caption, a paired unsafe image, and its unsafe caption. Safe-CLIP employs two primary sets of loss functions: a redirection loss to guide unsafe embeddings toward safe counterparts, and a preservation loss to maintain alignment of safe embeddings with their pretrained representations. In this section, we focus on the redirection losses, while preservation losses are detailed in the extended arXiv version.

Redirection loss: This loss guides the model to disregard unsafe content by redirecting unsafe embeddings towards their safe counterparts. It consists of two components: (1) cross-modal and (2) uni-modal redirection. In cross-modal redirection, unsafe captions t_i^* align with the corresponding safe images v_i, and unsafe images v_i^* are aligned with the corresponding safe captions t_i, ensuring that unsafe content is assigned to semantically equivalent safe representations. This follows the standard CLIP InfoNCE (Oord, Li, and Vinyals 2018; Radford et al. 2021) loss. Formally, the cross-modal redirection loss that fine-tunes the image encoder $\mathcal{V}(\cdot)$ is defined as:

$$\mathcal{L}_{\text{INCE}}^{\text{image}} = -\frac{1}{N}\sum_{i=1}^{N} \log \frac{\exp(\cos(\mathcal{V}(v_i^*), \mathcal{T}_0(t_i))/\tau)}{\sum_{j=1}^{N} \exp(\cos(\mathcal{V}(v_i^*), \mathcal{T}_0(t_j))/\tau)}. \quad (1)$$

As $\mathcal{L}_{\text{INCE}}^{\text{image}}$ follows the standard CLIP loss formulation, it minimizes the cosine distance between corresponding safe-unsafe pairs while increasing the distance between non-matching pairs. This ensures that unsafe embeddings are pulled closer to their safe counterparts while remaining distinct from unrelated safe embeddings. The cross-modal redirection loss for the text encoder, denoted as $\mathcal{L}_{\text{INCE}}^{\text{text}}$, follows the same formulation as Equation 1, with the roles of image and text swapped.

The uni-modal redirection loss directly minimizes the cosine distance between the unsafe and safe embeddings within the same modality. Specifically, unsafe captions t_i^* are aligned with their corresponding safe captions t_i, and unsafe images v_i^* are aligned with their corresponding safe images v_i. For image encoder $\mathcal{V}$, this uni-modal redirection

loss is defined as:

$$\mathcal{L}^{\text{image}}_{\text{uni-redir}} = -\frac{1}{N}\sum_{i=1}^{N}\cos(\mathcal{V}(v_i^*), \mathcal{V}_0(v_i)). \quad (2)$$

Similarly, the text encoder uni-modal redirection loss $\mathcal{L}^{\text{text}}_{\text{uni-redir}}$ follows the same formulation, with the roles of image and text swapped. Together, these four losses align unsafe image-text embeddings with their corresponding safe embeddings, guiding the model to mitigate unsafe content.

Limitations of Safe-CLIP

(1) Weakens generalization and cross-modal alignment: The cross-modal redirection loss in Equation 1 causes Safe-CLIP to disregard valid alternative safe concepts and actively pushes them apart when they co-occur in the same batch. This disrupts CLIP's cross-modal embedding structure by introducing misleading negative signals across modalities. While this mechanism suppresses unsafe content, it also leads to a significant drop in zero-shot performance—approximately 22%—thus limiting generalization to downstream tasks. Moreover, by treating semantically compatible safe examples as negatives, it undermines safety alignment itself, as the model may fail to recognize or retrieve acceptable alternatives at test time.

(2) Semantically misaligned fixed targets: Safe-CLIP enforces a rigid noisy mapping between unsafe and safe concepts, relying on manually defined pairs without ensuring semantic compatibility. This can lead to misaligned supervision that distorts the pretrained embedding space. As shown in Figure 1, the unsafe caption `"A deadly looking gun on a table next to a child"` is paired with the safe caption `"A delicious looking bunt cake on a table next to fruit"`—a weak match that overlooks more relevant alternatives. In contrast, our method identifies `"A kid sitting at a table with some food"` as the closest safe candidate based on semantic proximity. This reduces representational shift and supports both better generalization and more context-aware safety alignment.

Proposed Approach

We propose three key improvements over Safe-CLIP to enhance safety alignment while preserving generalization. First, we introduce a relative redirection loss that encourages unsafe embeddings to move away from their original representations and closer to safe counterparts. Second, we adopt a proximity-based alignment strategy that redirects each unsafe input to the most semantically compatible safe target, avoiding noisy mappings. Third, we implement a progressive training schedule that begins with easier unsafe–safe pairs and gradually introduces harder examples to stabilize learning and reduce representation shift. We also introduce NSFWCaps, a new benchmark for robust safety evaluation under distributional shift.

Relative Cross-Modal Redirection

The standard cross-modal redirection objective in Equation 1 relies on a random set of in-batch negatives, resulting in weakened generalization and cross-modal alignment **(Limitation 1)**. To address this issue, we propose a relative redirection objective that replaces the in-batch negative set with a single, targeted hard negative: the corresponding *unsafe* cross-modal embedding obtained from the frozen reference model. This formulation not only aligns unsafe inputs with their intended safe counterparts, but also explicitly discourages similarity to their original cross-modal unsafe representations. This loss is applied to both text and image embeddings, ensuring consistency across modalities. Formally, the loss for image encoder $\mathcal{V}$ is defined as:

$$\mathcal{L}^{\text{image}}_{\text{cross-redir}} = \frac{1}{N}\sum_{i=1}^{N}\log\Big(1+\exp\big(\cos(\mathcal{V}(v_i^*), \mathcal{T}_0(t_i^*)) - \cos(\mathcal{V}(v_i^*), \mathcal{T}_0(t_i))\big)\Big), \quad (3)$$

where N is the batch size, $\mathcal{V}$ is the image encoder, $\mathcal{T}_0$ is the frozen reference text encoder, v_i^* is the unsafe image, and t_i and t_i^* are the corresponding safe and unsafe captions, respectively. This objective penalizes instances where the unsafe image remains more similar to its unsafe caption than to its paired safe one. A similar loss $\mathcal{L}^{\text{text}}_{\text{cross-redir}}$ is applied to text encoder $\mathcal{T}$, where the roles of image and text are swapped.

By explicitly enforcing this separation, our approach mitigates the false negative issue in Safe-CLIP's cross-modal redirection loss. Instead of simply redirecting unsafe embeddings toward a specified safe alternative and repelling potential meaningful positives, our loss actively discourages embeddings from retaining their original unsafe characteristics, ensuring a more effective semantic shift.

Proximity-Based Alignment

To address the semantic misalignment resulting from the fixed unsafe–safe pairings enforced in Safe-CLIP **(Limitation 2)**, we propose a proximity-based alignment strategy. Rather than relying on manually specified safe targets—which may be poorly matched to the unsafe concept—our method selects the most semantically compatible safe alternative for each unsafe input. This ensures that realignment occurs along semantically meaningful directions, minimizing disruption to the pretrained representation space to preserve both generalization and structural consistency.

Selecting Semantically Aligned Safe Pairs: To construct reliable unsafe–safe training pairs, we leverage the CLIP text encoder to identify the most semantically compatible safe caption for each unsafe caption. Specifically, for a given unsafe caption t_i^*, we compute its cosine similarity with every candidate safe caption t_j in the dataset as $s_{ij} = \cos(\mathcal{T}_0(t_i^*), \mathcal{T}_0(t_j))$, where $\mathcal{T}_0$ denotes the frozen reference text encoder. We then select the best-matching safe caption $\hat{t}_i = t_{j^*}$, where $j^* = \arg\max_j s_{ij}$, and use its corresponding image $\hat{v}_i$ to form the aligned safe pair $(\hat{v}_i, \hat{t}_i)$ for training. This retrieval process is performed offline, ensuring efficient integration into the training pipeline without additional computational overhead.

Updated Relative Cross-Modal Redirection: We refine the soft redirection objective in Equation 3 by replacing

the fixed safe counterpart t_i with the closest semantically aligned safe embedding $\hat{t}_i$, as defined in the previous section. This modification ensures that each unsafe input is redirected toward a safe target that is contextually compatible and semantically meaningful. The updated image-to-text redirection loss is defined as:

$$\mathcal{L}^{\text{image}}_{\text{prox-cross-redir}} = \frac{1}{N}\sum_{i=1}^{N} \log\Big(1 + \exp\big(\cos(\mathcal{V}(v_i^*), \mathcal{T}_0(t_i^*)) - \cos(\mathcal{V}(v_i^*), \mathcal{T}_0(\hat{t}_i))\big)\Big), \quad (4)$$

where $\mathcal{V}$ is the image encoder and $\mathcal{T}_0$ is the frozen reference text encoder. A symmetric loss $\mathcal{L}^{\text{text}}_{\text{prox-cross-redir}}$ is defined analogously by swapping modalities.

By redirecting unsafe inputs toward the closest semantically appropriate safe alternatives, this updated loss minimizes representational shift during safety fine-tuning, better preserving the structure of the pretrained embedding space.

Updated Uni-Modal Redirection: We also update the unimodal redirection loss by replacing the fixed safe target with the closest semantically aligned counterpart $\hat{v}_i$. This encourages unsafe embeddings v_i^* to move closer to safe representations that are contextually compatible. The updated unimodal loss for the image modality is defined as:

$$\mathcal{L}^{\text{image}}_{\text{prox-uni-redir}} = -\frac{1}{N}\sum_{i=1}^{N} \cos(\mathcal{V}(v_i^*), \mathcal{V}_0(\hat{v}_i)), \quad (5)$$

where $\mathcal{V}$ is the trainable image encoder and $\mathcal{V}_0$ is its frozen reference counterpart. A corresponding loss $\mathcal{L}^{\text{text}}_{\text{prox-uni-redir}}$ is defined analogously for the text encoder. This uni-modal alignment objective ensures that, during safety fine-tuning, each unsafe concept is redirected toward its closest safe counterpart within the same modality—thereby minimizing representational shift and preserving the structure of the pretrained embedding space.

Full Redirection Loss: Our final redirection loss combines both cross-modal and uni-modal objectives, each based on proximity-aware alignment. These losses jointly encourage unsafe embeddings to move closer to semantically compatible safe counterparts while discouraging similarity to their unsafe forms. The complete redirection loss is defined as:

$$\mathcal{L}_{\text{redir}} = \mathcal{L}^{\text{image}}_{\text{prox-cross-redir}} + \mathcal{L}^{\text{text}}_{\text{prox-cross-redir}} + \mathcal{L}^{\text{image}}_{\text{prox-uni-redir}} + \mathcal{L}^{\text{text}}_{\text{prox-uni-redir}}. \quad (6)$$

Following Safe-CLIP (Poppi et al. 2024a), we additionally include preservation losses that maintain the global structure of CLIP's pretrained embedding space. The overall training objective is the sum of these preservation losses and our proposed redirection loss, enabling effective safety alignment with minimal disruption to the pretrained representation geometry.

Progressive Training for Safety Alignment

To ensure stable adaptation and minimize representation shift during safety fine-tuning, we adopt a progressive training strategy based on the difficulty of unsafe samples. Each unsafe–safe pair is categorized as *easy*, *medium*, or *hard* according to the cosine similarity between their captions. Pairs with higher similarity are considered semantically closer and less likely to disrupt the pretrained embedding space. Training proceeds in three phases: (1) we begin with only *easy* unsafe–safe pairs to allow stable initialization, (2) gradually introduce *medium* samples to improve redirection strength, and (3) finally include *hard* examples, which require larger representational shifts. This curriculum helps the model align unsafe content with safe targets in a smooth and controlled manner, improving safety without sacrificing generalization.

NSFWCaps: Robust Safety Evaluation Set

The ViSU (Poppi et al. 2024a) test set suffers from the same alignment issues identified previously, where unsafe–safe pairs are often poorly matched (e.g., `"a deadly looking gun..."` paired with `"a delicious bunt cake..."`), leading to noisy results. To overcome this limitation we introduce *NSFWCaps*, a new benchmark for evaluating cross-modal safety alignment under a slight distribution shift. We begin with the image-caption pairs from the validation split of NoCaps (Agrawal et al. 2019b), which is largely composed of near or out-of-domain objects compared to COCO (of which ViSU is based on). For each caption, we carefully generate a related unsafe variant using LLaMA-3-70B (Grattafiori et al. 2024). The prompt edits only safety-relevant details while preserving the original semantics. This process is applied across 20 NSFW categories from (Schramowski, Tauchmann, and Kersting 2022). Unsafe images are then generated using an NSFW-tuned Stable Diffusion model[2], yielding 4.5K image–text pairs. We apply a multi-stage filtering pipeline to select high-quality examples. First, we filter images using NudeNet (Bedapudi 2019) and the Q16 detector (Schramowski et al. 2023a) to ensure visual unsafety. Next, we compute JINA-CLIP (Koukounas et al. 2024) similarity between each safe–unsafe caption pair and retain the most semantically aligned examples from each category. The final dataset contains 1,000 quadruples—each with a safe image, unsafe image, safe caption, and unsafe caption. Safe and unsafe elements are tightly aligned in meaning but differ in safety. On average, NSFWCaps safe and unsafe captions have a JINA-CLIP similarity of 0.81, compared to 0.62 in ViSU, making it a strong testbed for more robust cross modal safety evaluation.

Experiments

We evaluate our method on cross-modal retrieval, zero-shot classification, and multi-modal generation to assess the safety–generalization trade-off. Retrieval tests redirection of unsafe inputs without harming safe retrieval, zero-shot classification evaluates generalization, and generation tasks assess safety and semantic fidelity. Our approach improves generalization while maintaining strong safety.

[2]https://huggingface.co/stablediffusionapi/newrealityxl-global-nsfw

Method	ViSU				NSFWCaps				Zero-Shot Average
	$\mathbf{T} \rightarrow \mathbf{V}$	$\mathbf{V} \rightarrow \mathbf{T}$	$\mathbf{T}^* \rightarrow \mathbf{V}$	$\mathbf{V}^* \rightarrow \mathbf{T}$	$\mathbf{T} \rightarrow \mathbf{V}$	$\mathbf{V} \rightarrow \mathbf{T}$	$\mathbf{T}^* \rightarrow \mathbf{V}$	$\mathbf{V}^* \rightarrow \mathbf{T}$	
CLIP	36.8	39.9	2.8	5.5	69.6	73.4	3.8	7.9	74.3
DataComp-1B	46.7	47.0	1.6	5.5	79.0	80.3	2.8	12.9	—
CLIP†	54.5	54.9	2.0	6.6	78.9	79.1	4.6	13.1	67.3
Safe-CLIP	49.1	48.8	14.5	23.8	76.6	76.7	35.4	47.1	52.2
Ours	**52.0** (+2.9%)	**51.5** (+2.7%)	**27.9** (+13.4%)	**24.6** (+0.8%)	**81.8** (+5.2%)	**78.1** (+1.4%)	**79.5** (+44.1%)	**72.3** (+25.2%)	**60.2** (+8.0%)

Table 1: **Left:** Retrieval performance (R@1) on ViSU and NSFWCaps. We report safe-to-safe retrieval ($\mathbf{T} \rightarrow \mathbf{V}$, $\mathbf{V} \rightarrow \mathbf{T}$) and unsafe-to-safe redirection ($\mathbf{T}^* \rightarrow \mathbf{V}$, $\mathbf{V}^* \rightarrow \mathbf{T}$). **Right:** Average zero-shot classification accuracy across 11 datasets. † indicates CLIP fine-tuned on safe data only. Improvements over Safe-CLIP are shown in dark green.

Datasets

Training Dataset: ViSU (Poppi et al. 2024a) is a synthetic dataset containing 169K quadruples, each with a safe image–caption pair and its corresponding unsafe variant. Safe examples are drawn from COCO Captions (Lin et al. 2014; Chen et al. 2015) using Karpathy's split (Karpathy and Fei-Fei 2015), while unsafe counterparts are generated using a fine-tuned LLM and a diffusion-based NSFW image generator. The data spans 20 NSFW categories, which are used for training and retrieval-based redirection evaluation.

Generalization Evaluation Datasets: We assess zero-shot generalization across 11 benchmarks, including ImageNet and its variants (IN-A, IN-R, IN-V2, IN-S) and standard CLIP evaluation datasets such as Caltech101, Oxford Pets, Flowers102, Stanford Cars, UCF101, and DTD (Fei-Fei, Fergus, and Perona 2004; Parkhi et al. 2012)

NSFW Evaluation Datasets: We evaluate cross-modal safety using both synthetic and real-world unsafe inputs. For synthetic evaluation, we report retrieval results on the ViSU test set (Poppi et al. 2024a) and our proposed NSFWCaps benchmark, where safe content is real and unsafe counterparts are synthetically generated. For real-world evaluation, we follow (Poppi et al. 2024a) and use NSFW images from NudeNet (Bedapudi 2019), SMID (Crone et al. 2018), and public NSFW URLs[3]. These are used across both retrieval and image-to-text generation tasks, enabling broad evaluation of safety alignment in response to real unsafe content. For text-to-image safety, we use I2P (Schramowski et al. 2023a), a benchmark comprising diverse NSFW prompt categories.

Implementation Details

We build on the CLIP (Radford et al. 2021) architecture and adopt the ViT-L/14 variant as our main backbone, consistent with both Stable Diffusion v1.4 (Rombach et al. 2022b) (text encoder) and LLaVA (Liu et al. 2023) (vision encoder). All models are implemented using PyTorch and trained using the Safe-CLIP (Poppi et al. 2024a) public repository as our base framework. We fine-tune both the vision and text encoders using LoRA (Hu et al. 2021) adapters with a fixed rank of $r = 16$. We use the Adam optimizer (Kingma and Ba 2014) with a learning rate of 1×10^{-4}, a batch size of 48, and train for 9 epochs. To encourage smooth safety fine-tuning and reduce abrupt shifts in the embedding space, we employ progressive training: the first epoch uses only easy unsafe–safe pairs, the second uses both easy and medium samples, and the remaining epochs include all three difficulty levels (easy, medium, and hard). A fixed random seed of 42 is used for all training and evaluation runs. To ensure reproducibility in synthetic data generation used for ViSU training, we fix the generation seed to 8185. This seed is consistently applied across all models, including our method and baselines. All experiments are conducted on A6000 GPUs. For fair comparison, we re-train all baselines under identical settings, including Safe-CLIP (Poppi et al. 2024a) and the CLIP† baseline, which is fine-tuned on safe data using only preservation losses i.e. no re-directional losses are used.

Cross-Modal Retrieval Evaluation

We evaluate cross-modal retrieval on both ViSU (Poppi et al. 2024a) and our proposed NSFWCaps benchmark. ViSU includes 5K test samples, while NSFWCaps contains 1K carefully curated pairs with stronger semantic coupling. We compare against: CLIP (Radford et al. 2021), Safe-CLIP (Poppi et al. 2024a), a CLIP variant trained on NSFW-filtered DataComp-1B (Gadre et al. 2023), and CLIP† (fine-tuned on only safe ViSU data).

Retrieval Protocols. We report results on: (1) *Safe-to-Safe Retrieval*, evaluating whether safety fine-tuning preserves text-to-image ($\mathbf{T} \rightarrow \mathbf{V}$) and image-to-text ($\mathbf{V} \rightarrow \mathbf{T}$) retrieval accuracy; and (2) *Unsafe-to-Safe Redirection*, assessing whether unsafe queries ($\mathbf{T}^*, \mathbf{V}^*$) retrieve corresponding safe targets instead of unsafe ones.

Results. Table 1 shows that our method consistently outperforms Safe-CLIP. On ViSU, we improve $\mathbf{T}^* \rightarrow \mathbf{V}$ by 13.4% while preserving safe retrieval (+2.9%). On NSFWCaps, we observe even greater gains (+44.1% $\mathbf{T}^* \rightarrow \mathbf{V}$), demonstrating safety alignment with better generalization retention.

Robustness to Real NSFW Images. To evaluate safety under real-world distribution shifts, we follow (Poppi et al. 2024a) and assess retrieval on real NSFW images from NudeNet (Bedapudi 2019), SMID (Crone et al. 2018), and public NSFW URLs.[4] While NudeNet and NSFW URLs cover nudity-related content, SMID includes broader unsafe categories such as harm and discrimination. Each dataset contributes to the unsafe image set $\mathbf{V}^*$, while unsafe texts $\mathbf{T}^*$ are taken from the ViSU test set. For safe retrieval, we

[3]https://github.com/EBazarov/nsfw_data_source_urls

[4]https://github.com/EBazarov/nsfw_data_source_urls

Method	% NSFW $\mathbf{V} \rightarrow \mathbf{T}$ ↓			% NSFW $\mathbf{T} \rightarrow \mathbf{V}$ ↓		
	NSFW URLs	NudeNet	SMID	NSFW URLs	NudeNet	SMID
CLIP	91.6	94.1	96.3	98.8	99.6	97.0
DataComp-1B	82.1	87.0	87.6	89.4	89.5	93.5
CLIP†	91.1	93.7	88.3	95.7	97.0	87.6
Safe-CLIP	21.1	13.0	14.2	41.1	43.1	26.6
Ours	**18.5**	**10.7**	**3.1**	**37.2**	**27.0**	**16.9**

Table 2: Retrieval results on real NSFW data using unsafe queries (↓ is better). Lower values indicate better filtering of unsafe content. † denotes safe-only fine-tuning on ViSU (Poppi et al. 2024a).

sample 10K safe captions $\mathbf{T}$ and image $\mathbf{V}$ distractors from LAION-400M (Schuhmann et al. 2021). Table 2 presents the results, where %NSFW represents the fraction of retrieved items that are unsafe, given an NSFW query. Our method significantly reduces the percentage of unsafe items retrieved compared to all baselines, demonstrating improved robustness to real-world NSFW inputs.

Zero-Shot Classification Generalization

Zero-shot classification reflects CLIP's ability to generalize across diverse datasets without task-specific supervision. In Table 1 (right), we report the average accuracy across 11 benchmarks. Safe-CLIP (Poppi et al. 2024a) suffers a 22% drop in average accuracy compared to the original CLIP model. In contrast, our method improves over Safe-CLIP by 8%, demonstrating stronger generalization while maintaining safety alignment.

Text-to-Image Generation

We evaluate safety alignment of our fine-tuned textual encoder by integrating it into Stable Diffusion v1.4 (Rombach et al. 2022b) for CLIP-guided text-to-image generation. We run evaluations on the full I2P benchmark (Schramowski et al. 2023a), which includes 4,700 NSFW prompts across seven categories. Safety is assessed using the NudeNet (Bedapudi 2019) classifier for sexual content and the Q16 classifier (Schramowski et al. 2023a) for others categories. As shown in Table 3, our method reduces the average NSFW score from 37.1 to 16.0, outperforming base CLIP and matching Safe-CLIP (Poppi et al. 2024a). However, unlike Safe-CLIP, we achieve this safety while preserving better generalization, as demonstrated in the previous section. Combining our approach with inference-time methods like Safety Guidance (SLD) (Schramowski et al. 2023b) or Negative Prompting yields further improvements.

Image-to-Text Generation

We assess the safety alignment of our fine-tuned visual encoder by integrating it into LLaVA (Liu et al. 2023), replacing the standard CLIP image encoder without additional training. We generate captions for real NSFW images from NudeNet (Bedapudi 2019), public NSFW URLs, and SMID (Crone et al. 2018), following prior work (Poppi et al. 2024a). Captions are evaluated using a GPT-based NSFW classifier (LLaMA-3.1-8B) and the Perspective API[5] to measure both NSFW content and toxicity levels. As shown in Table 4, our approach effectively reduces unsafe caption content, matching or outperforming Safe-CLIP (Poppi et al. 2024a).

Method	**AVG** ↓	**+Ours**
SD v1.4	37.1	–
+CLIP†	34.4	–
+Safe-CLIP	16.1	**16.0** (+0.1%)
+SLD-Weak	23.7	**13.9** (+9.8%)
+SLD-Medium	17.4	**12.8** (+4.6%)
+SLD-Strong	12.1	**12.0** (+0.1%)
+Neg-Prompt	12.3	**11.9** (+0.4%)

Table 3: Average NSFW score for text-to-image generation on the I2P benchmark (↓ is better). Our method improves over base CLIP and matches Safe-CLIP performance. † denotes safe-only fine-tuning.

	NudeNet ↓		NSFW URLs ↓		SMID ↓	
Model	NSFW %	Tox.	NSFW %	Tox.	NSFW %	Tox.
LLaVA	75.5	36.2	56.4	24.9	24.2	5.4
+CLIP†	66.8	29.2	52.4	22.2	17.0	4.5
+Safe-CLIP	31.5	16.4	27.9	13.6	8.8	4.1
+Ours	**25.4**	**12.4**	**27.6**	**11.0**	**7.7**	**3.6**

Table 4: NSFW and toxicity scores for image-to-text generation (↓ is better). † denotes safe-only fine-tuning on ViSU (Poppi et al. 2024a).

Conclusion

We present `SafeR-CLIP`, a fine-tuning strategy that redirects unsafe embeddings toward safe counterparts while preserving model utility. Unlike prior approaches that rely on noisy mappings, our method uses proximity-based redirection to guide unsafe inputs toward semantically aligned safe alternatives. This improves safety alignment across multiple tasks—enhancing redirection in retrieval, reducing unsafe generations in text-to-image synthesis, and lowering toxicity in image captioning—while retaining strong generalization, as demonstrated by zero-shot classification results. These findings highlight that proximity-aware redirection offers an effective balance between safety and performance. Future work may explore asymmetric encoder adaptation and broader real-world deployment of safety-tuned models.

References

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019a. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, 8948–8957.

[5]https://github.com/conversationai/perspectiveapi

Agrawal, H.; Desai, K.; Wang, Y.; Chen, X.; Jain, R.; Johnson, M.; Batra, D.; Parikh, D.; Lee, S.; and Anderson, P. 2019b. nocaps: novel object captioning at scale. In *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*. IEEE.

Bedapudi, P. 2019. Nudenet: Neural nets for nudity classification, detection and selective censoring.

Birhane, A.; Dehdashtian, S.; Prabhu, V.; and Boddeti, V. 2024. The Dark Side of Dataset Scaling: Evaluating Racial Classification in Multimodal Models. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, 1229–1244. ACM.

Birhane, A.; Prabhu, V.; Han, S.; Boddeti, V. N.; and Luccioni, A. S. 2023. Into the LAIONs Den: Investigating Hate in Multimodal Datasets. arXiv:2311.03449.

Cao, Y.; and Yang, J. 2015. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, 463–480. IEEE.

Chen, X.; Fang, H.; Lin, T.-Y.; Vedantam, R.; Gupta, S.; Dollár, P.; and Zitnick, C. L. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

Crone, D. L.; Bode, S.; Murawski, C.; and Laham, S. M. 2018. The Socio-Moral Image Database (SMID): A novel stimulus set for the study of social, moral and affective processes. *PloS one*, 13(1): e0190954.

Denton, E.; Hanna, A.; Amironesei, R.; Smart, A.; and Nicole, H. 2021. On the genealogy of machine learning datasets: A critical history of ImageNet. *Big Data & Society*, 8(2): 20539517211035955.

D'Incà, M.; Peruzzo, E.; Xu, X.; Shi, H.; Sebe, N.; and Mancini, M. 2025. Safe Vision-Language Models via Unsafe Weights Manipulation. arXiv:2503.11742.

Ding, Y.; Li, B.; and Zhang, R. 2025. ETA: Evaluating Then Aligning Safety of Vision Language Models at Inference Time. arXiv:2410.06625.

Fan, C.; Liu, J.; Zhang, Y.; Wong, E.; Wei, D.; and Liu, S. 2023. Salun: Empowering machine unlearning via gradient-based weight saliency in both image classification and generation. *arXiv preprint arXiv:2310.12508*.

Fei-Fei, L.; Fergus, R.; and Perona, P. 2004. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. In *2004 conference on computer vision and pattern recognition workshop*, 178–178. IEEE.

Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; Orgad, E.; Entezari, R.; Daras, G.; Pratt, S.; Ramanujan, V.; Bitton, Y.; Marathe, K.; Mussmann, S.; Vencu, R.; Cherti, M.; Krishna, R.; Koh, P. W.; Saukh, O.; Ratner, A.; Song, S.; Hajishirzi, H.; Farhadi, A.; Beaumont, R.; Oh, S.; Dimakis, A.; Jitsev, J.; Carmon, Y.; Shankar, V.; and Schmidt, L. 2023. DataComp: In search of the next generation of multimodal datasets. arXiv:2304.14108.

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

Ghosal, S. S.; Chakraborty, S.; Singh, V.; Guan, T.; Wang, M.; Velasquez, A.; Beirami, A.; Huang, F.; Manocha, D.; and Bedi, A. S. 2025. Immune: Improving Safety Against Jailbreaks in Multi-modal LLMs via Inference-Time Alignment. arXiv:2411.18688.

Ginart, A.; Guan, M.; Valiant, G.; and Zou, J. Y. 2019. Making ai forget you: Data deletion in machine learning. *Advances in neural information processing systems*, 32.

Golatkar, A.; Achille, A.; and Soatto, S. 2020. Eternal sunshine of the spotless net: Selective forgetting in deep networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9304–9312.

Gou, Y.; Chen, K.; Liu, Z.; Hong, L.; Xu, H.; Li, Z.; Yeung, D.-Y.; Kwok, J. T.; and Zhang, Y. 2024. Eyes Closed, Safety On: Protecting Multimodal LLMs via Image-to-Text Transformation. arXiv:2403.09572.

Grattafiori, A.; Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Vaughan, A.; et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Han, D.; Mohamed, S.; and Li, Y. 2024. ShieldDiff: Suppressing Sexual Content Generation from Diffusion Models through Reinforcement Learning. *arXiv preprint arXiv:2410.05309*.

Helff, L.; Friedrich, F.; Brack, M.; Schramowski, P.; and Kersting, K. 2024. LLAVAGUARD: VLM-based Safeguard for Vision Dataset Curation and Safety Assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8322–8326.

Hu, E. J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; and Chen, W. 2021. LoRA: Low-Rank Adaptation of Large Language Models. arXiv:2106.09685.

Karpathy, A.; and Fei-Fei, L. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3128–3137.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Koukounas, A.; Mastrapas, G.; Günther, M.; Wang, B.; Martens, S.; Mohr, I.; Sturua, S.; Akram, M. K.; Martínez, J. F.; Ognawala, S.; Guzman, S.; Werk, M.; Wang, N.; and Xiao, H. 2024. Jina CLIP: Your CLIP Model Is Also Your Text Retriever. arXiv:2405.20204.

Larrazabal, A. J.; Nieto, N.; Peterson, V.; Milone, D. H.; and Ferrante, E. 2020. Gender imbalance in medical imaging datasets produces biased classifiers for computer-aided diagnosis. *Proceedings of the National Academy of Sciences*, 117(23): 12592–12594.

Li, J.; Wei, Q.; Zhang, C.; Qi, G.; Du, M.; Chen, Y.; and Bi, S. 2024. Single Image Unlearning: Efficient Machine Unlearning in Multimodal Large Language Models. arXiv:2405.12523.

Lin, T.-Y.; Maire, M.; Belongie, S.; Hays, J.; Perona, P.; Ramanan, D.; Dollár, P.; and Zitnick, C. L. 2014. Microsoft

coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, 740–755. Springer.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, Q.; Shang, C.; Liu, L.; Pappas, N.; Ma, J.; John, N. A.; Doss, S.; Marquez, L.; Ballesteros, M.; and Benajiba, Y. 2024. Unraveling and Mitigating Safety Alignment Degradation of Vision-Language Models. arXiv:2410.09047.

Oord, A. v. d.; Li, Y.; and Vinyals, O. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Parkhi, O. M.; Vedaldi, A.; Zisserman, A.; and Jawahar, C. 2012. Cats and dogs. In *2012 IEEE conference on computer vision and pattern recognition*, 3498–3505. IEEE.

Poppi, S.; Poppi, T.; Cocchi, F.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024a. Safe-CLIP: Removing NSFW concepts from vision-and-language models. In *European Conference on Computer Vision*, 340–356. Springer.

Poppi, S.; Sarto, S.; Cornia, M.; Baraldi, L.; and Cucchiara, R. 2024b. Multi-class unlearning for image classification via weight filtering. *IEEE Intelligent Systems*.

Poppi, T.; Kasarla, T.; Mettes, P.; Baraldi, L.; and Cucchiara, R. 2025. Hyperbolic Safety-Aware Vision-Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 4222–4232.

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

Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022a. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 10684–10695.

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

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023a. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22522–22531.

Schramowski, P.; Brack, M.; Deiseroth, B.; and Kersting, K. 2023b. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. arXiv:2211.05105.

Schramowski, P.; Tauchmann, C.; and Kersting, K. 2022. 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*, 1350–1361.

Schuhmann, C.; Beaumont, R.; Vencu, R.; Gordon, C.; Wightman, R.; Cherti, M.; Coombes, T.; Katta, A.; Mullis, C.; Wortsman, M.; Schramowski, P.; Kundurthy, S.; Crowson, K.; Schmidt, L.; Kaczmarczyk, R.; and Jitsev, J. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402.

Schuhmann, C.; Vencu, R.; Beaumont, R.; Kaczmarczyk, R.; Mullis, C.; Katta, A.; Coombes, T.; Jitsev, J.; and Komatsuzaki, A. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Steed, R.; and Caliskan, A. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 701–713.

Wan, Y.; Subramonian, A.; Ovalle, A.; Lin, Z.; Suvarna, A.; Chance, C.; Bansal, H.; Pattichis, R.; and Chang, K.-W. 2024. Survey of Bias In Text-to-Image Generation: Definition, Evaluation, and Mitigation. arXiv:2404.01030.

Wang, A.; Liu, A.; Zhang, R.; Kleiman, A.; Kim, L.; Zhao, D.; Shirai, I.; Narayanan, A.; and Russakovsky, O. 2022. REVISE: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7): 1790–1810.

Wang, Y.; Liu, X.; Li, Y.; Chen, M.; and Xiao, C. 2024. AdaShield: Safeguarding Multimodal Large Language Models from Structure-based Attack via Adaptive Shield Prompting. arXiv:2403.09513.

Wu, X.; Chakraborty, S.; Xian, R.; Liang, J.; Guan, T.; Liu, F.; Sadler, B. M.; Manocha, D.; and Bedi, A. S. 2024. Highlighting the Safety Concerns of Deploying LLMs/VLMs in Robotics. arXiv:2402.10340.

Xu, S.; Pang, L.; Zhu, Y.; Shen, H.; and Cheng, X. 2025. Cross-Modal Safety Mechanism Transfer in Large Vision-Language Models. arXiv:2410.12662.

Zhang, Z.; Liu, G.; Fleming, C.; Kompella, R. R.; and Xu, C. 2025. Targeted Forgetting of Image Subgroups in CLIP Models. arXiv:2506.03117.

Zhao, W.; Li, Z.; Li, Y.; and Sun, J. 2025. Zero-Shot Defense Against Toxic Images via Inherent Multimodal Alignment in LVLMs. *arXiv preprint arXiv:2503.00037*.

Zhao, X.; Chen, X.; Xuan, Y.; and Zhao, Z. 2024. Buster: Incorporating Backdoor Attacks into Text Encoder to Mitigate NSFW Content Generation. *arXiv preprint arXiv:2412.07249*.

Zong, Y.; Bohdal, O.; Yu, T.; Yang, Y.; and Hospedales, T. 2024a. Safety Fine-Tuning at (Almost) No Cost: A Baseline for Vision Large Language Models. arXiv:2402.02207.

Zong, Y.; Bohdal, O.; Yu, T.; Yang, Y.; and Hospedales, T. 2024b. Safety fine-tuning at (almost) no cost: A baseline for vision large language models. *arXiv preprint arXiv:2402.02207*.

Zou, X.; Kang, J.; Kesidis, G.; and Lin, L. 2025. Understanding and Rectifying Safety Perception Distortion in VLMs. arXiv:2502.13095.