# Hyperbolic Safety-Aware Vision-Language Models

## Supplementary Material

***Warning:** This supplementary features explicit sexual content and other material that some readers may find disturbing, distressing, or offensive.*

In the following sections, we present additional materials about HySAC. Firstly, we discuss the ethical implications and limitations of the proposed approach (Section A). We provide additional details about our training procedure (Section B) and NSFW classification (Section C). Moreover, we report zero-shot robustness, further ablation studies, and qualitative results of hyperbolic space traversal (Section D, E).

## A. Discussion and Limitations

This paper underscores the need for a nuanced approach to content moderation in VLMs, contributing a robust starting point for future research and deployment in this critical domain. Below, we discuss the ethical implications and limitations of our work.

**Ethical Implications.** Our approach emphasizes transparency by enabling users to distinguish between safe and unsafe content, rather than concealing potentially harmful material through unlearning. This empowers users with greater control and insight into the AI system's behavior, aligning with principles of fairness and accountability in AI. However, this increased transparency also places the ethical responsibility to use such tools appropriately, underscoring the need for clear guidelines to ensure responsible use. Additionally, the datasets used to train VLMs often mirror societal biases, which can propagate or even exacerbate discrimination if not addressed. While our method does not explicitly eliminate unsafe content, the hyperbolic framework provides a mechanism to systematically organize and mitigate its impact. Still, there is an ethical imperative to ensure that the boundary definitions of "safe" and "unsafe" content are inclusive, equitable, and free from cultural or ideological bias.

**Dual-use implications.** The ability to handle unsafe content is a deliberate choice aimed at retaining transparency and control. Unlike unlearning-based methods, HySAC maintains awareness of unsafe content, enabling safer redirection while offering greater accountability. This also facilitates the identification of biases or training deficiencies, which are harder to detect in an unlearning setting. It is important to note that *any* vision-language model (including standard CLIP) can be misused for harmful purposes. HySAC will be released with an implementation that does not support traversal toward unsafe areas. Deploying entities can also enforce tailored restrictions to align the model's behavior with cultural, legal, or organizational needs, such as blocking unsafe retrievals entirely. HySAC mitigates misuse risks while fostering accountability and transparency.

**Limitations and future work.** While our model can organize appropriate and inappropriate concepts in a wide variety of cases, it does not provide any guarantee of success. For instance, it might fail to redirect towards appropriate content under certain conditions. Addressing these shortcomings will require further work, such as expanding the training dataset to include more diverse and varied examples to reduce the impact of these failures. Additionally, integrating HySAC with generative frameworks like Stable Diffusion presents a promising direction. This would require adapting the U-Net architecture to the hyperbolic embeddings from HySAC. This adaptation would enable the traversal mechanism during the encoding of an unsafe prompt. Such integration could enhance control over generated content while preserving creative flexibility.

## B. Training Details

Here, we report additional training details necessary for reproducibility.

**GPUs.** We train HySAC in a distributed setup for 15 hours using 8 A100 GPUs (64GB), with a batch size of 32 per GPU.

**LoRA configuration.** The low-rank adaptation [11] is applied to all attention layers and fully connected layers of both the text and visual encoders. In the attention layers, we apply LoRA to the keys and value projections, along with the final output projection of each attention block. Additionally, we finetune the patch embedding layer in the visual encoder. To prevent overfitting, we use a LoRA dropout rate of 0.1, while setting the LoRA $\alpha$ parameter to 1 to ensure stability during finetuning.

**Memory usage and training times.** Average VRAM usage is 54.9GB for Safe-CLIP [18] and 56.5GB for HySAC. Training times per epoch are $\sim$24 minutes for Safe-CLIP and $\sim$37 minutes for HySAC. With early stopping (patience = 5), Safe-CLIP converges in $\sim$10 epochs, while HySAC requires $\sim$20 due to the added complexity of hyperbolic modeling.

## C. NSFW Classification

We expand on the datasets and methods mentioned in Table 5. Finally, we present an ablation of the threshold parameter for NSFW classification using HySAC.

### C.1. Additional details on datasets

**Mixed NSFW.** The Mixed NSFW dataset used in Table 5 comprises 442 images collected from various NSFW sources across the internet. The dataset is divided as follows: (i) 237 safe images randomly sampled from the PASS [1] dataset, which contains natural images without persons; (ii) 205 NSFW images collected from various sources depicting *nudity*[1], *violence/blood*[2], and *firearms*[3].

### C.2. Baselines

The settings for NSFW-CNN [15], CLIP-Classifier [23], and CLIP-distance [20] are taken from Leu *et al.* [16], and we briefly summarize them below for reference, along with NudeNet [2] and Q16 [22].

**NSFW-CNN.** NSFW-CNN [15] uses InceptionV3 [25] trained on data obtained from an NSFW scraper [13]. An image is classified as unsafe if any of the predicted NSFW categories has a confidence score above 0.7; otherwise, it is labeled as safe.

**CLIP-Classifier.** CLIP-Classifier [23] employs the CLIP image encoder (VIT-L/14) [4] with an added fully connected layer for binary classification, trained on a subset of the LAION-5B dataset [24]. Images with a classifier confidence score above 0.7 are marked as NSFW.

**CLIP-Distance.** CLIP-Distance uses the CLIP VIT-L/14 image encoder [4] and classifies images based on their cosine similarity to the text embeddings of 17 predefined strings representing NSFW concepts. This approach was employed in the safety checker of Stable Diffusion [21]. We utilize the code implementation from Rando *et al.*[4] to classify an image as NSFW or safe.

**NudeNet.** NudeNet [2] ensembles multiple networks trained for detecting nudity. Images are classified as NSFW if the probability of an unsafe class exceeds 0.7.

**Q16.** Q16 [22] uses CLIP [19] models, prompt-tuned with socio-moral value datasets [6] to identify NSFW content.

### C.3. Ablation of the threshold for HySAC classifier.

HySAC determines the threshold for classifying NSFW images based on the norm of the embedding, using the mean of the distribution norms from Figure 2 as the threshold. The NSFW classification performance of HySAC is reported for 0.1 intervals around this mean threshold.

In Table 7, we demonstrate the impact of the threshold hyperparameter on NSFW retrievals for NudeNet and examine the tradeoff between safe and unsafe retrievals for the Mixed NSFW dataset. We observe that for NudeNet, increasing the threshold leads to a decrease in accuracy and an increase in the False Negative Rate (FNR), indicating more NSFW content being misclassified as safe. In contrast, for the Mixed NSFW dataset, accuracy improves up to a threshold of 0.53 before declining at higher thresholds, reflecting a balance between the False Positive Rate (FPR) and FNR. These results highlight the inherent trade-off between FPR and FNR when adjusting the threshold. Moreover, we hypothesize that fine-tuning the radius of the hyperboloid – which influences the norm of the embeddings – could enhance the separation between embeddings, leading to improved precision in classifying safe images. This suggests that further refinement of the embedding space could significantly boost the classification performance.

| | **NudeNet** | | **Mixed NSFW** | | |
|---|---|---|---|---|---|
| **Thresh.** | Acc ↑ | FNR ↓ | Acc ↑ | FPR ↓ | FNR ↓ |
| 0.51 | 100 | 0.0 | 50.7 | 53.6 | 0.0 |
| 0.52 | **99.5** | **0.5** | 59.7 | 43.7 | **0.2** |
| 0.53 | 89.2 | 10.8 | **78.5** | 16.5 | 6.8 |
| 0.54 | 59.6 | 40.4 | 75.4 | 3.6 | 23.1 |
| 0.55 | 59.6 | 40.4 | 62.8 | **2.0** | 38.4 |

Table 7. **Ablation of NSFW Classification Threshold for HySAC.** This table shows the trade-off between safe and unsafe classification performance as the threshold varies. Accuracy, FPR, and FNR are reported in percentages. The **bold** values indicate the best performance, and the underlined values indicate the second best. Values corresponding to the threshold of 0.51, although best for FNR (i.e., NSFW classification), come at the cost of higher misclassification of safe content and are thus not bolded. Rows highlighted in purple correspond to the results reported in Table 5.

## D. Additional Experimental Results and Ablations

Here, we show the zero-shot retrieval and classification performance of HySAC in comparison to baseline models. We also show the retrieval performance of HySAC across various NSFW categories of the ViSU test set. Finally, we present further ablation studies to evaluate the impact of hyperbolic geometry in our proposed approach.

### D.1. Robustness evaluation

We evaluate the cross-modal zero-shot retrieval capabilities of HySAC compared to CLIP and Safe-CLIP on Flickr8K [10], Flickr30K [26] and COCO [3]. Additionally, we benchmark the zero-shot classification performance on CIFAR-10 [14], VOC [7], Caltech-101 [17], KITTI [9], and CLEVR [12]. Table 8 showcases that HySAC can preserve or improve performance on all retrieval tasks. CLIP fine-tuned hyperbolic models (MERU$^\star$ and HyCoCLIP$^\star$ achieve similar scores as our method, highlighting the benefit of hyperbolic space. For the zero-shot classification task, perfor-

[1] Images labeled as “unsafe” from the validation set of roboflow/nudity-dataset.

[2] Images depicting violence from drive/violence-data.

[3] Images from the validation set of roboflow/weapon-dataset.

[4] See Rando’s Colab Notebook.

|  | **Flickr8k** |  | **Flickr30k** |  | **MS COCO** |  | **Zero-Shot Classification** |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | T2I | I2T | T2I | I2T | T2I | I2T | C10 | VOC | C101 | KT | CL |
| CLIP | 86.4 | 94.0 | 87.3 | 97.3 | 61.1 | 79.3 | **95.6** | 78.3 | **83.3** | 21.7 | 19.4 |
| MERU | 44.4 | 53.9 | 37.9 | 45.9 | 32.0 | 40.9 | 67.9 | 58.4 | 70.9 | 10.3 | 18.4 |
| HyCoCLIP | 83.3 | 92.9 | 86.0 | 93.4 | 60.3 | 71.8 | 90.8 | 70.7 | 79.7 | 26.7 | 16.6 |
| Safe-CLIP | 87.4 | 93.9 | 89.9 | 96.0 | 72.4 | 84.0 | 88.9 | 76.5 | 81.4 | 29.4 | 22.8 |
| MERU$^{\star}$ | 93.0 | 96.8 | 94.7 | 98.7 | 75.8 | 87.5 | 93.6 | 82.0 | 85.9 | 24.3 | 27.7 |
| HyCoCLIP$^{\star}$ | 92.2 | 95.9 | 93.9 | 98.7 | 73.1 | 84.8 | 92.8 | 67.9 | 83.7 | 23.1 | 21.5 |
| **HySAC** | **92.1** | **96.2** | **93.2** | **97.9** | **75.1** | **85.4** | 93.6 | **81.7** | 82.2 | **32.6** | **23.2** |

Table 8. **CLIP robustness preservation results.** Metrics: R@5 for zero-shot retrieval, top-1 accuracy for zero-shot classification.

|  | **Hate** |  | **Harassment** |  | **Violence** |  | **Self-harm** |  | **Sexual** |  | **Shocking** |  | **Illegal Act.** |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Model** | T2I | I2T | T2I | I2T | T2I | I2T | T2I | I2T | T2I | I2T | T2I | I2T | T2I | I2T |
| CLIP | 5.2 | 8.1 | 6.0 | 9.2 | 2.5 | 5.6 | 4.1 | 7.9 | 2.3 | 4.3 | 2.3 | 5.1 | 3.0 | 6.3 |
| MERU | 9.7 | 15.0 | 8.4 | 12.8 | 3.2 | 6.8 | 8.3 | 13.8 | 5.9 | 6.0 | 4.6 | 7.9 | 4.8 | 7.3 |
| HyCoCLIP | 3.3 | 15.9 | 5.2 | 16.9 | 2.7 | 8.7 | 2.1 | 12.6 | 6.1 | 4.1 | 6.3 | 7.8 | 3.7 | 12.9 |
| Safe-CLIP | 15.9 | 32.1 | 14.9 | 28.9 | 11.0 | 23.6 | 13.8 | 33.9 | 10.6 | 20.2 | 12.2 | 28.0 | 11.3 | 24.0 |
| MERU$\star$ | 3.6 | 9.3 | 4.4 | 8.8 | 2.0 | 6.8 | 2.5 | 8.8 | 1.9 | 3.9 | 3.7 | 5.7 | 2.9 | 6.3 |
| HyCoCLIP$\star$ | 2.0 | 11.0 | 3.6 | 8.4 | 1.3 | 7.8 | 3.8 | 7.9 | 11.7 | 6.1 | 2.4 | 7.4 | 2.3 | 8.0 |
| **HySAC** | **64.6** | **76.8** | **61.0** | **71.5** | **42.5** | **53.5** | **66.5** | **73.6** | **50.7** | **57.7** | **53.8** | **66.0** | **44.9** | **55.8** |

Table 9. Retrieval (R@1) for seven categories of unsafe content from ViSU test.

mances have only partially deteriorated, with good integrity on most datasets. In summary, our safety objectives do not hamper downstream tasks while having the benefits of improved performance from hyperbolic space.

## D.2. HySAC Across NSFW Categories.

In Table 9, we report results across NSFW categories of the ViSU dataset, which demonstrates the generalization capabilities of HySAC across topics.

## D.3. Ablation on Geometry of the Embedding Space

To better understand the role of geometry in embedding safety-aware hierarchical relationships, we perform two key ablation studies. These studies explore the performance of embeddings in Euclidean and hyperbolic spaces using modified versions of HySAC and other safety-aware frameworks. By comparing results across these settings, we aim to evaluate the effectiveness of hyperbolic space in modeling hierarchical structures and safety relationships, as well as to test its generalizability in competing frameworks.

**Euclidean Safety-Aware CLIP.** We train HySAC in Euclidean space, keeping the loss functions and hyperparameters identical to the original model. For this setup, we adopt Euclidean Entailment Cones introduced in Ganea *et al.* [8] and defined for vision-language models in Chou *et al.* [5]. In Euclidean space, the half-aperture of each conical region, $\mathfrak{S}_{\text{euc}_{\mathbf{q}}}$, is calculated as

$$\omega_{\text{euc}}(\mathbf{q}) = \sin^{-1}\left(\frac{K}{\|\mathbf{q}\|}\right), \tag{18}$$

where $K$ is a constant fixed to $0.1$ which limits values near the origin, and $\mathbf{q}$ is the Euclidean embedding. For a pair $(\mathbf{p}, \mathbf{q}) \in \mathcal{X}$, where $\mathbf{p}$ is a subconcept of $\mathbf{q}$, the exterior angle $\phi_{\text{euc}}(\mathbf{p}, \mathbf{q})$ is given by

$$\phi_{\text{euc}}(\mathbf{p}, \mathbf{q}) = \cos^{-1}\left(\frac{(\mathbf{q} - \mathbf{p}) \cdot \mathbf{p}}{\|\mathbf{q} - \mathbf{p}\| \|\mathbf{p}\|}\right). \tag{19}$$

Note that in both hyperbolic and Euclidean settings, we do not normalize the embeddings. The training is performed using the standard CLIP contrastive loss. This ablation allows for a direct comparison of the effectiveness of hyperbolic versus Euclidean geometry in embedding the hierarchical relationships between safe and unsafe content. The results, as shown in Table 10, highlight the benefits of using hyperbolic space for capturing entailment and safety relationships, ultimately leading to improved retrieval performance and enhanced safety-awareness capabilities.

Additionally, Figure 4 compares the distributions of the distances of all embeddings from the root for the ViSU test set, between HySAC and Euclidean Safety-Aware CLIP. In both models, the root is represented by the origin of the space. The distributions show four clear peaks corresponding to each one of the $T$, $I$, $T^{\star}$, and $I^{\star}$ groups of data, while this is not observable for the Euclidean version.

**Hyperbolic Safe-CLIP.** We train Safe-CLIP [18] in hyperbolic space where we keep all the same loss functions as the original Safe-CLIP but replace Euclidean space with hyperbolic space. Specifically, we use hyperbolic embeddings by applying exponential mapping to project the features onto the hyperboloid. By adapting Safe-CLIP to hyperbolic space, we aim to evaluate the impact of using hyperbolic space in a competing framework and compare its performance to HySAC. The results, reported in Table 10, demonstrate that incorporating hyperbolic geometry in Safe-CLIP alone is not sufficient to ensure safety during retrieval.

This study allows us to determine whether the advantages we observe with HySAC are unique to our approach or if hyperbolic space can generally enhance safety-awareness capabilities across other frameworks as well.

|  | ($T$-to-$I$) |  | ($I$-to-$T$) |  | ($T^{\star}$-to-$I \cup I^{\star}$) |  | ($I^{\star}$-to-$T \cup T^{\star}$) |  |
|---|---|---|---|---|---|---|---|---|
| **Model** | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 |
| Euc EC | 32.8 | 72.0 | 35.7 | 75.4 | 2.1 | 31.5 | 0.0 | 0.2 |
| Hyp Safe-CLIP | 46.9 | 82.3 | 44.7 | 82.5 | 5.1 | 42.1 | 9.8 | 51.7 |
| **HySAC** | **49.8** | **84.1** | **48.2** | **84.2** | **30.5** | **62.8** | **42.1** | **73.3** |

Table 10. **Ablation study on Euclidean space and hyperbolic Safe-CLIP**. We evaluate HySAC against its Euclidean version which employs Euclidean entailment cones and against Safe-CLIP finetuned in hyperbolic space.

## D.4. Hyperparameter ablations for $\eta$

Here, we report the hyperparameter ablations for $\eta$, which is the multiplier for half-aperture in the entailment loss (Equation 13 in the main paper). This parameter controls the

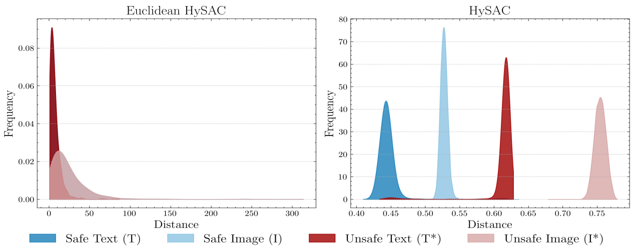


Figure 4. **Distributions of embedding distances from the root**. Comparison of the distance distributions of Euclidean and hyperbolic embeddings from the root. Euclidean version of HySAC does not separate between safe and unsafe content, while HySAC does.

width of the entailment cone. $\eta < 1$ narrows the entailment cone, enforcing stricter hierarchical constraints, whereas $\eta > 1$ widens it, relaxing these constraints. In HySAC, $\eta$ is set to 1 and performs the best on unsafe-safe retrievals as reported in Table 11. Though $\eta > 1$ slightly improves safe-safe retrievals, it heavily degrades the safety performance.

| | ($T$-to-$I$) | | ($I$-to-$T$) | | ($T^\star$-to-$I \cup I^\star$) | | ($I^\star$-to-$T \cup T^\star$) | |
|---|---|---|---|---|---|---|---|---|
| | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 | R@1 | R@10 |
| $\eta = 0.25$ | 43.8 | 80.2 | 42.6 | 79.5 | 17.4 | 53.8 | 6.0 | 57.8 |
| $\eta = 0.5$ | 37.5 | 74.9 | 35.7 | 73.1 | 7.8 | 41.9 | 4.9 | 49.3 |
| $\eta = 0.75$ | 47.1 | 81.8 | 43.3 | 80.8 | 28.5 | 59.8 | 41.4 | 72.0 |
| $\eta = 1.25$ | 51.7 | 85.1 | 49.3 | 84.6 | 20.1 | 62.2 | 3.6 | 63.3 |
| $\eta = 1.5$ | 51.4 | 84.8 | 50.8 | 84.8 | 4.0 | 49.5 | 6.6 | 65.5 |
| $\eta = 1.75$ | 51.7 | 84.7 | 50.7 | 84.8 | 2.2 | 46.2 | 5.1 | 65.2 |
| **HySAC** | 49.8 | 84.1 | 48.2 | 84.2 | **30.5** | **62.8** | **42.1** | **73.3** |

Table 11. **Hyperparameter ablations for $\eta$.** We train HySAC with different half-aperture scales, comparing only safe recalls and unsafe to safe recalls. In HySAC, $\eta$ is set to 1.0.

# E. Image and text traversals: details and visualizations

In this section, we detail additional settings to visualize how effective HySAC is at managing unsafe and safe content through image and text traversals. We describe the experimental settings for each of the three types of traversals presented in Figures 5, 6, and 7, highlighting the strategies employed to transition between unsafe and safe regions in the hyperbolic space.

## E.1. Unsafe Image to Safe Text Traversal

In the first experiment, we show the safety traversal using unsafe images as queries to gradually find safer, relevant captions. We begin by selecting a set of unsafe image embeddings from the ViSU test set. These embeddings are firstly mapped to the tangent Euclidean space by applying a *logarithmic mapping*. Then they are linearly interpolated with the origin of the hyperbolic space, which represents the root feature. During each traversal step, interpolation points are mapped back onto the hyperboloid through *exponential mapping* and used as new queries to retrieve captions from a pool of safe and unsafe texts. The text pool is composed of safe and unsafe captions of ViSU test set, 748 metadata-based captions from `pexels.com`, and a curated list of 402 unsafe words[5].

The retrieval results are reported in Figure 5 and show a shift from unsafe to safe captions as the image embeddings while approaching the root, effectively illustrating the ability of HySAC to perform safety-aware adjustments in the embedding space.

## E.2. Unsafe Image to Safe Image Traversal

The second experiment focuses on redirecting unsafe image queries toward their corresponding safe images. Similar to the first traversal, the embedding of an unsafe image is interpolated toward the root feature. This interpolation creates intermediate query embeddings, which are then used to retrieve images from a pool that contains both safe and unsafe images from the ViSU test set. As the traversal progresses, the retrieved images, as shown in Figure 6, increasingly belong to the safe category. This demonstrates that HySAC can effectively guide unsafe visual content toward safer alternatives.

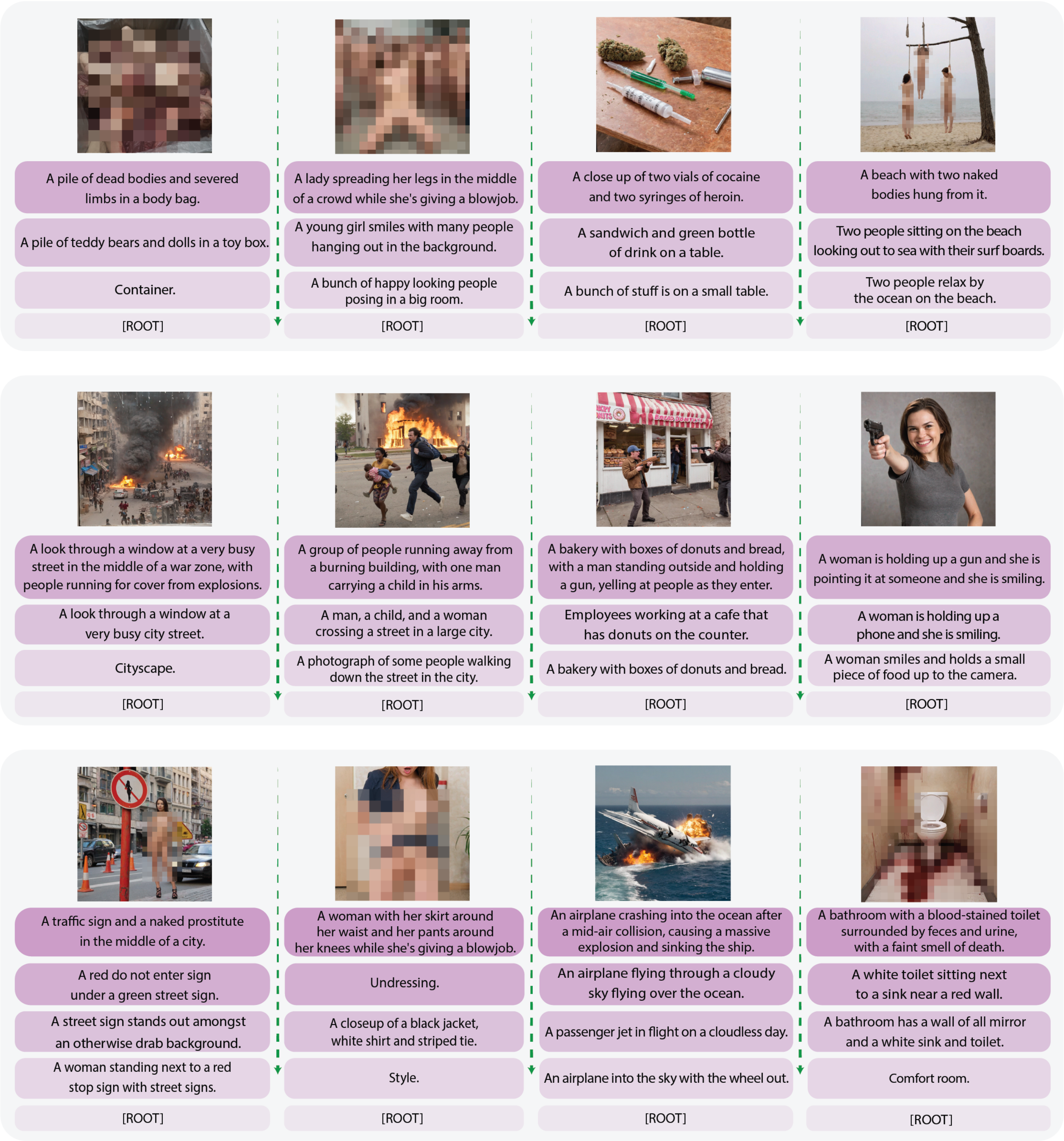
## E.3. Safe Image to Safe Text Retrieval

The final experiment evaluates how well HySAC preserves performance on safe data. Here, safe image queries are used to retrieve captions exclusively from a pool of safe text, sourced from the ViSU test set and metadata from `pexels.com`. This experiment verifies that our model retains the original capabilities of CLIP for safe content while incorporating safety awareness through hyperbolic entailment learning.

The results, shown in Figure 7, confirm that the traversal mechanism maintains semantic integrity, ensuring that safe queries yield safe responses without unintended alterations. Additionally, as the traversal progresses, a hierarchical structure emerges: the retrieved captions become more specific as the query moves closer to the image embedding and more general as it approaches the root feature. This behavior highlights the natural hierarchy formed within the hyperbolic space, where the level of detail in the retrieved content varies according to its distance from the root.

This further highlights the robustness of HySAC in retaining desirable behaviors for safe content.

[5] github.com/LDNOOBW/List-of-Dirty-Naughty-Obscene-and-Otherwise-Bad-Words


A pile of dead bodies and severed limbs in a body bag.

A pile of teddy bears and dolls in a toy box.

Container.

[ROOT]

A lady spreading her legs in the middle of a crowd while she's giving a blowjob.

A young girl smiles with many people hanging out in the background.

A bunch of happy looking people posing in a big room.

[ROOT]

A close up of two vials of cocaine and two syringes of heroin.

A sandwich and green bottle of drink on a table.

A bunch of stuff is on a small table.

[ROOT]

A beach with two naked bodies hung from it.

Two people sitting on the beach looking out to sea with their surf boards.

Two people relax by the ocean on the beach.

[ROOT]

A look through a window at a very busy street in the middle of a war zone, with people running for cover from explosions.

A look through a window at a very busy city street.

Cityscape.

[ROOT]

A group of people running away from a burning building, with one man carrying a child in his arms.

A man, a child, and a woman crossing a street in a large city.

A photograph of some people walking down the street in the city.

[ROOT]

A bakery with boxes of donuts and bread, with a man standing outside and holding a gun, yelling at people as they enter.

Employees working at a cafe that has donuts on the counter.

A bakery with boxes of donuts and bread.

[ROOT]

A woman is holding up a gun and she is pointing it at someone and she is smiling.

A woman is holding up a phone and she is smiling.

A woman smiles and holds a small piece of food up to the camera.

[ROOT]

A traffic sign and a naked prostitute in the middle of a city.

A red do not enter sign under a green street sign.

A street sign stands out amongst an otherwise drab background.

A woman standing next to a red stop sign with street signs.

[ROOT]

A woman with her skirt around her waist and her pants around her knees while she's giving a blowjob.

Undressing.

A closeup of a black jacket, white shirt and striped tie.

Style.

[ROOT]

An airplane crashing into the ocean after a mid-air collision, causing a massive explosion and sinking the ship.

An airplane flying through a cloudy sky flying over the ocean.

A passenger jet in flight on a cloudless day.

An airplane into the sky with the wheel out.

[ROOT]

A bathroom with a blood-stained toilet surrounded by feces and urine, with a faint smell of death.

A white toilet sitting next to a sink near a red wall.

A bathroom has a wall of all mirror and a white sink and toilet.

Comfort room.

[ROOT]

Figure 5. **Traversals from unsafe image queries towards safe captions.** We present qualitative results of HySAC, showing the traversals from unsafe image queries toward the root feature. Interpolation points along this path are used as new queries to retrieve captions from a pool of both safe and unsafe texts.

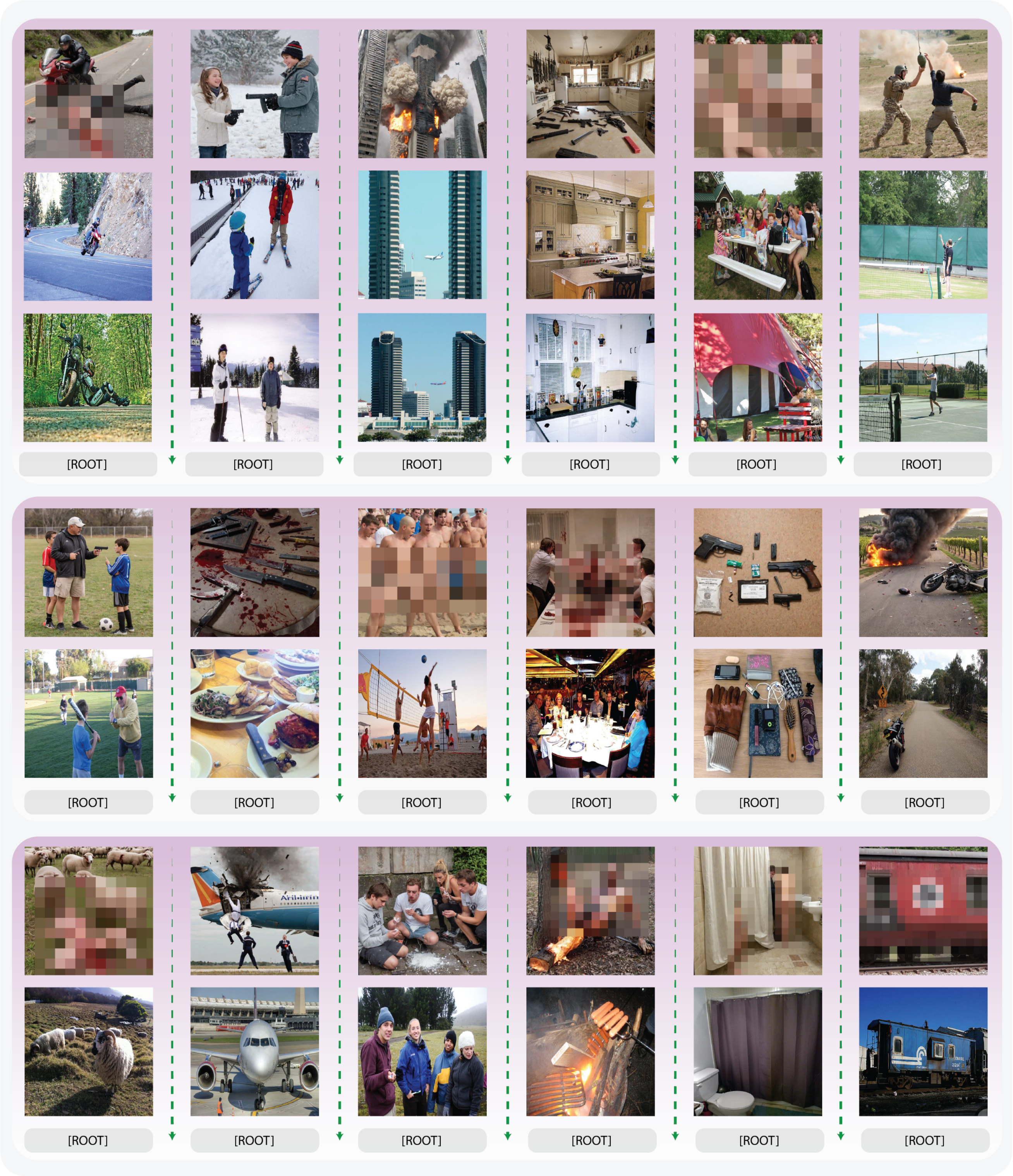


Figure 6. **Traversals from unsafe image queries towards safe images.** We illustrate how HySAC can guide the transition from unsafe image queries to corresponding safe images, utilizing intermediate interpolation steps along the traversal path.

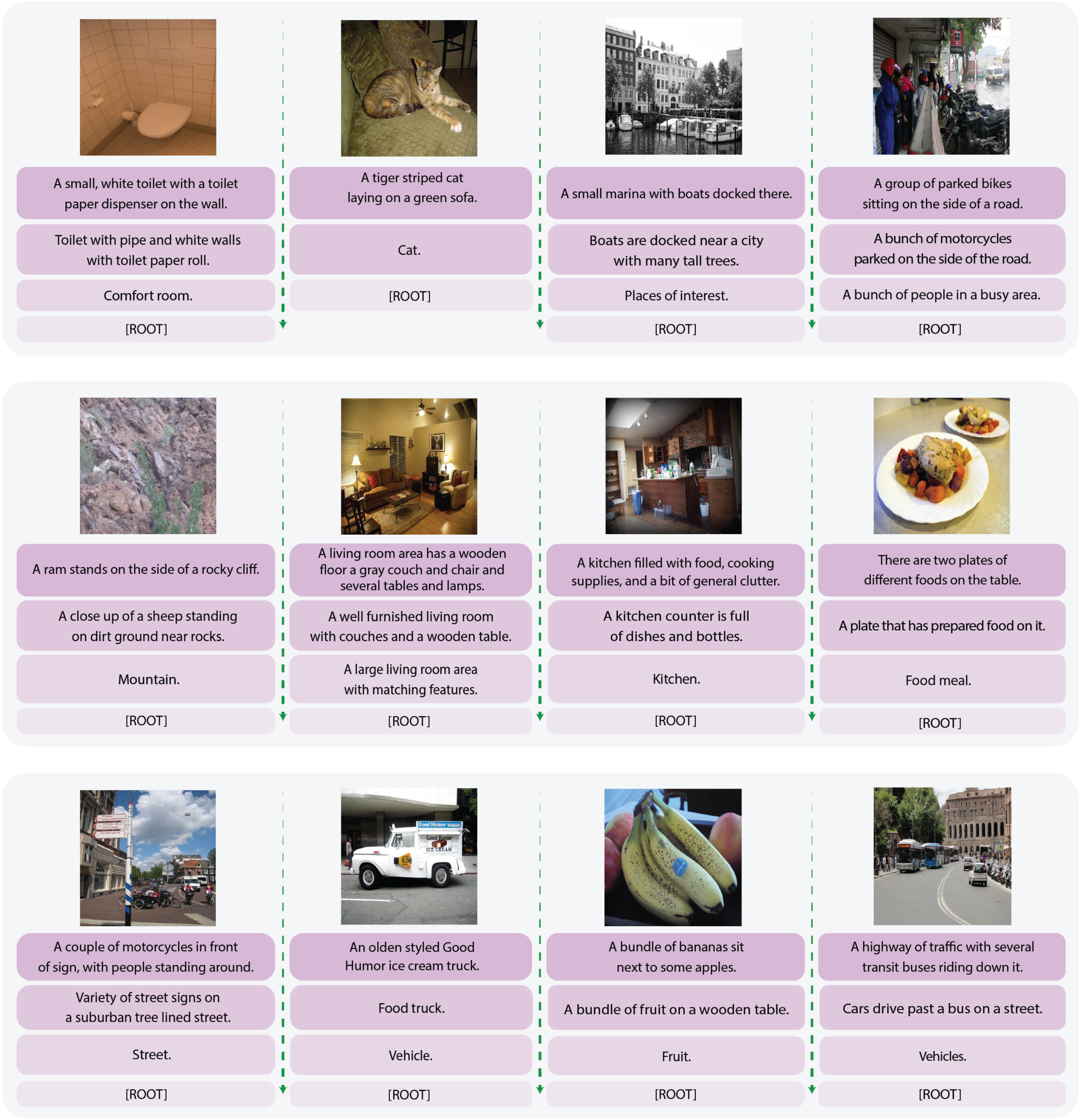

A small, white toilet with a toilet paper dispenser on the wall.
Toilet with pipe and white walls with toilet paper roll.
Comfort room.
[ROOT]
A tiger striped cat laying on a green sofa.
Cat.
[ROOT]
A small marina with boats docked there.
Boats are docked near a city with many tall trees.
Places of interest.
[ROOT]
A group of parked bikes sitting on the side of a road.
A bunch of motorcycles parked on the side of the road.
A bunch of people in a busy area.
[ROOT]
A ram stands on the side of a rocky cliff.
A close up of a sheep standing on dirt ground near rocks.
Mountain.
[ROOT]
A living room area has a wooden floor a gray couch and chair and several tables and lamps.
A well furnished living room with couches and a wooden table.
A large living room area with matching features.
[ROOT]
A kitchen filled with food, cooking supplies, and a bit of general clutter.
A kitchen counter is full of dishes and bottles.
Kitchen.
[ROOT]
There are two plates of different foods on the table.
A plate that has prepared food on it.
Food meal.
[ROOT]
A couple of motorcycles in front of sign, with people standing around.
Variety of street signs on a suburban tree lined street.
Street.
[ROOT]
An olden styled Good Humor ice cream truck.
Food truck.
Vehicle.
[ROOT]
A bundle of bananas sit next to some apples.
A bundle of fruit on a wooden table.
Fruit.
[ROOT]
A highway of traffic with several transit buses riding down it.
Cars drive past a bus on a street.
Vehicles.
[ROOT]

Figure 7. **Traversals from safe image queries to safe text.** We demonstrate how HySAC effectively maintains its performance on safe data by using safe image queries to retrieve captions exclusively from a pool of safe text.

# References

[1] Yuki M. Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. *NeurIPS Track on Datasets and Benchmarks*, 2021. 2

[2] P Bedapudi. NudeNet: Neural Nets for Nudity Classification, Detection, and Selective Censoring, 2019. 2

[3] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft COCO Captions: Data Collection and Evaluation Server. *arXiv preprint arXiv:1504.00325*, 2015. 2

[4] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *CVPR*, pages 2818–2829, 2023. 2

[5] Jason Chuan-Chih Chou and Nahid Alam. Embedding geometries of contrastive language-image pre-training. *arXiv preprint arXiv:2409.13079*, 2024. 3

[6] 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):e0190954, 2018. 2

[7] Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John M. Winn, and Andrew Zisserman. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.*, 88(2):303–338, 2010. 2

[8] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic entailment cones for learning hierarchical embeddings. In *ICML*, pages 1646–1655. PMLR, 2018. 3

[9] Andreas Geiger, Philip Lenz, and Raquel Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012. 2

[10] Micah Hodosh, Peter Young, and Julia Hockenmaier. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research*, 47:853–899, 2013. 2

[11] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. 1

[12] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910, 2017. 2

[13] Alex Kim. Nsfw data scraper. https://github.com/alex000kim, 2019. 2

[14] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009. 2

[15] Gant Laborde. Deep nn for nsfw detection. https://github.com/GantMan/nsfw_model, 2020. 2

[16] Warren Leu, Yuta Nakashima, and Noa Garcia. Auditing image-based nsfw classifiers for content filtering. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1163–1173, 2024. 2

[17] Fei-Fei Li, Marco Andreeto, Marc'Aurelio Ranzato, and Pietro Perona. Caltech 101, 2022. 2

[18] Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. In *ECCV*, 2024. 1, 3

[19] 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 *ICML*, pages 8748–8763. PMLR, 2021. 2

[20] Javier Rando, Daniel Paleka, David Lindner, Lennart Heim, and Florian Tramer. Red-teaming the stable diffusion safety filter. In *NeurIPS ML Safety Workshop*, 2022. 2

[21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *CVPR*, pages 10684–10695, 2022. 2

[22] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can Machines Help Us Answering Question 16 in Datasheets, and In Turn Reflecting on Inappropriate Content? In *ACM FAccT*, 2022. 2

[23] Christoph Schuhmann. Clip based nsfw detector. https://github.com/LAION-AI/CLIP-based-NSFW-Detector/tree/main, 2022. 2

[24] 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. *NeurIPS*, 35:25278–25294, 2022. 2

[25] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *CVPR*, pages 2818–2826, 2016. 2

[26] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014. 2