SafeCLIP:
An In-Depth Review

Samuele Poppi, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara
University of Modena and Reggio Emilia, Italy
University of Pisa, Italy

Published in: ECCV 2024
Citations: 20

Presenters:

John Benjamin, Asher Gibson, Zena Himes, Adam Sachs

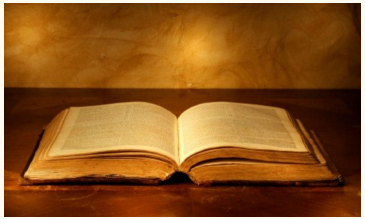
Dr. Mubarak Shah

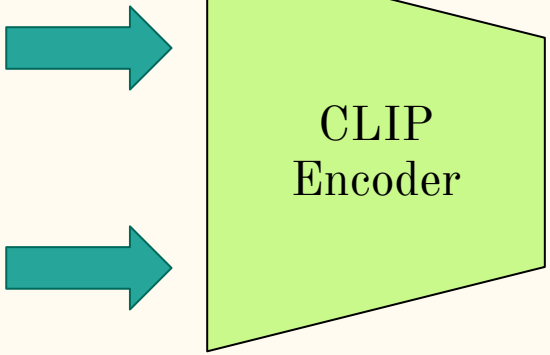
WARNING

The following presentation contains blurred and censored texts and images due to inappropriate content.

Background: CLIP

A zero-shot image classifier

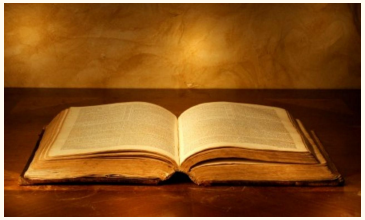




["An open novel or book",
"A dog playing fetch",
"A car being driven"]

Background: CLIP

A zero-shot image classifier



["An open novel or book",
"A dog playing fetch",
"A car being driven"]

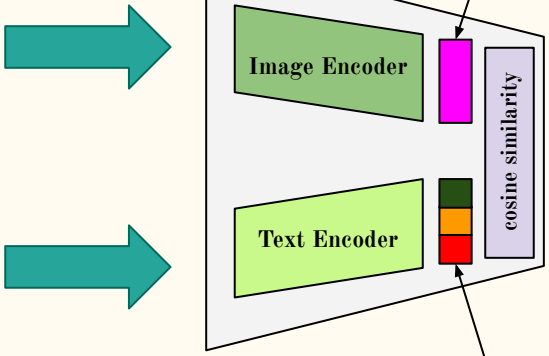
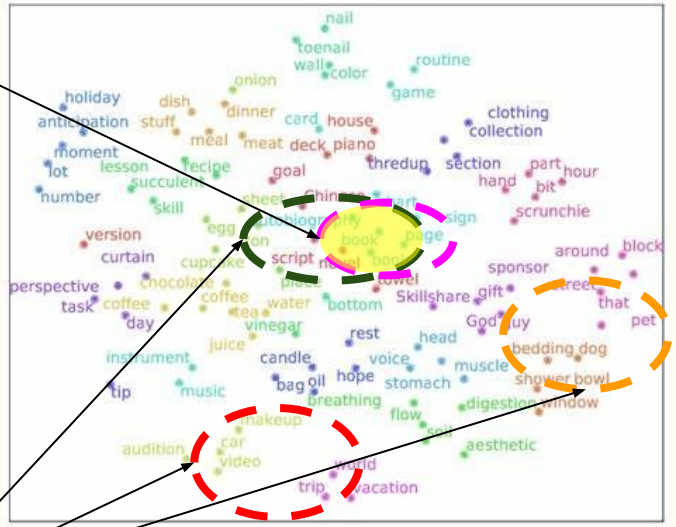


Image Embeddings

Text Embeddings

Image Projection in Reduced Dimension Latent Space



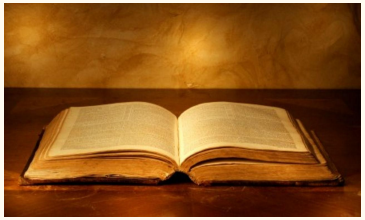
Probabilities of image being aligned with text pairs

["0.970",
"0.018",
"0.012"]

Background: CLIP

A zero-shot image classifier

Learns relationships between words and visual representations (images) of those words!



["An open novel or book",
"A dog playing fetch",
"A car being driven"]

CLIP Encoder

softmax(logits)

["0.970",
"0.018",
"0.012"]

Class Probabilities

argmax

"An open novel or book"

So, what’s the problem?

Uncensored Training Data

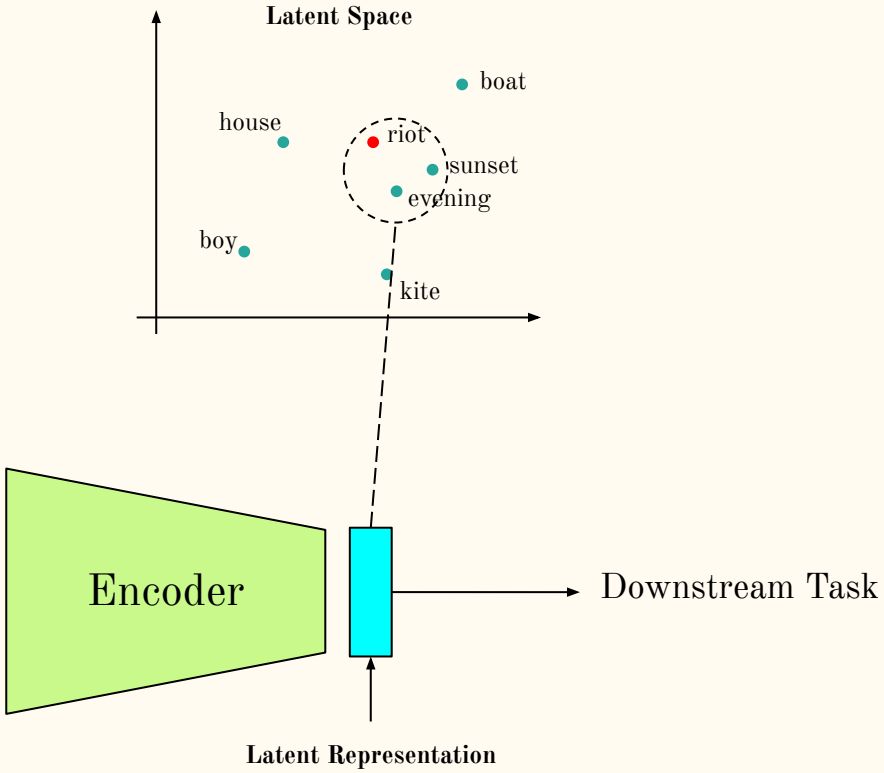


A boy flying a kite



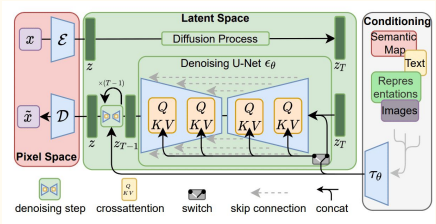
A clash between rioters and police in the evening





The Result

Show me an image of a city in the evening where fighting is taking place



Diffusion Model



!!! Unsafe Text Query !!!

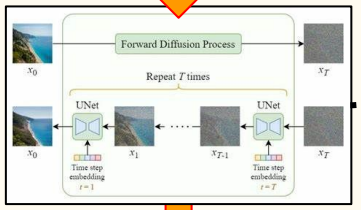
!!! Harmful Content !!!

Harmful Content $\stackrel{\text{def}}{=}$ *NSFW*

Two Primary Generative Tasks

Text-to-Image Generation

A burning car set on fire by gang members





Diffusion Models

CLIP Encoder

[7.4, 0.3, ..., 2.4, 22.1]

Image-Text Relationship Feature Vector

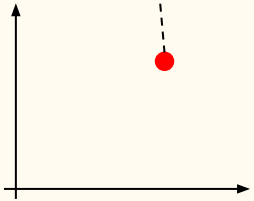


Image-to-Text Generation



VLM

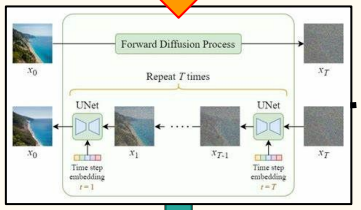
A seriously injured soldier laying on the ground, potentially unconscious.

Vision Language Models

Two Primary Generative Tasks

Text-to-Image Generation

A burning car set on fire by gang members





Diffusion Models

SafeCLIP Encoder

[7.4, 0.3, ..., 2.4, 22.1]

Image-Text Relationship Feature Vector

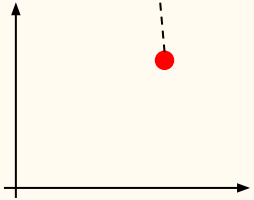


Image-to-Text Generation



VLM

Three veterans kneeling on a beach.

Vision Language Models

What is defined as *Not-Safe-For-Work*?

- Violence
- Brutality
- Cruelty
- Blood
- Abuse



- Hate
- Harassment
- Humiliation



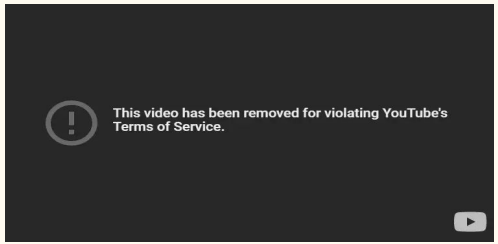
- Weapons
- Vandalism
- Drug Use
- Theft
- Illegal Activity



- Suffering
- Harm
- Suicide



- Sexual Content
- Nudity
- Bodily Fluids
- Obscene Gestures

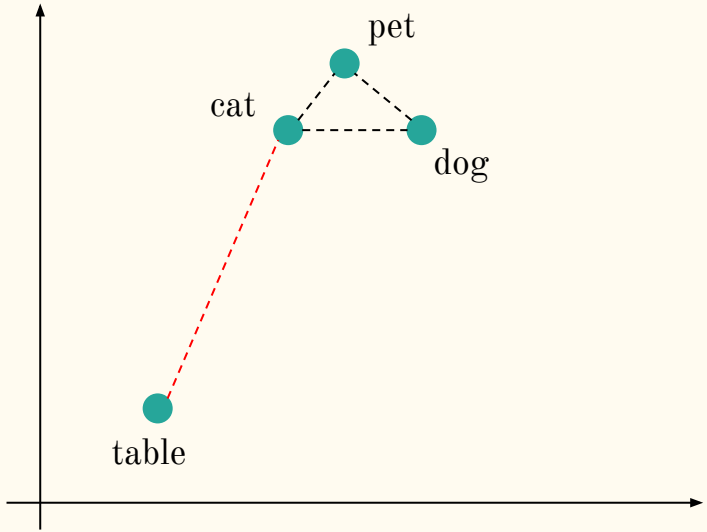


Intuition







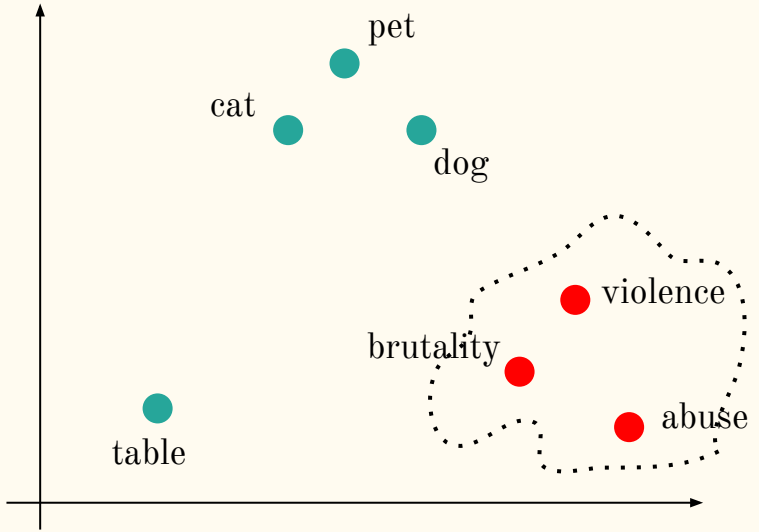


Intuition





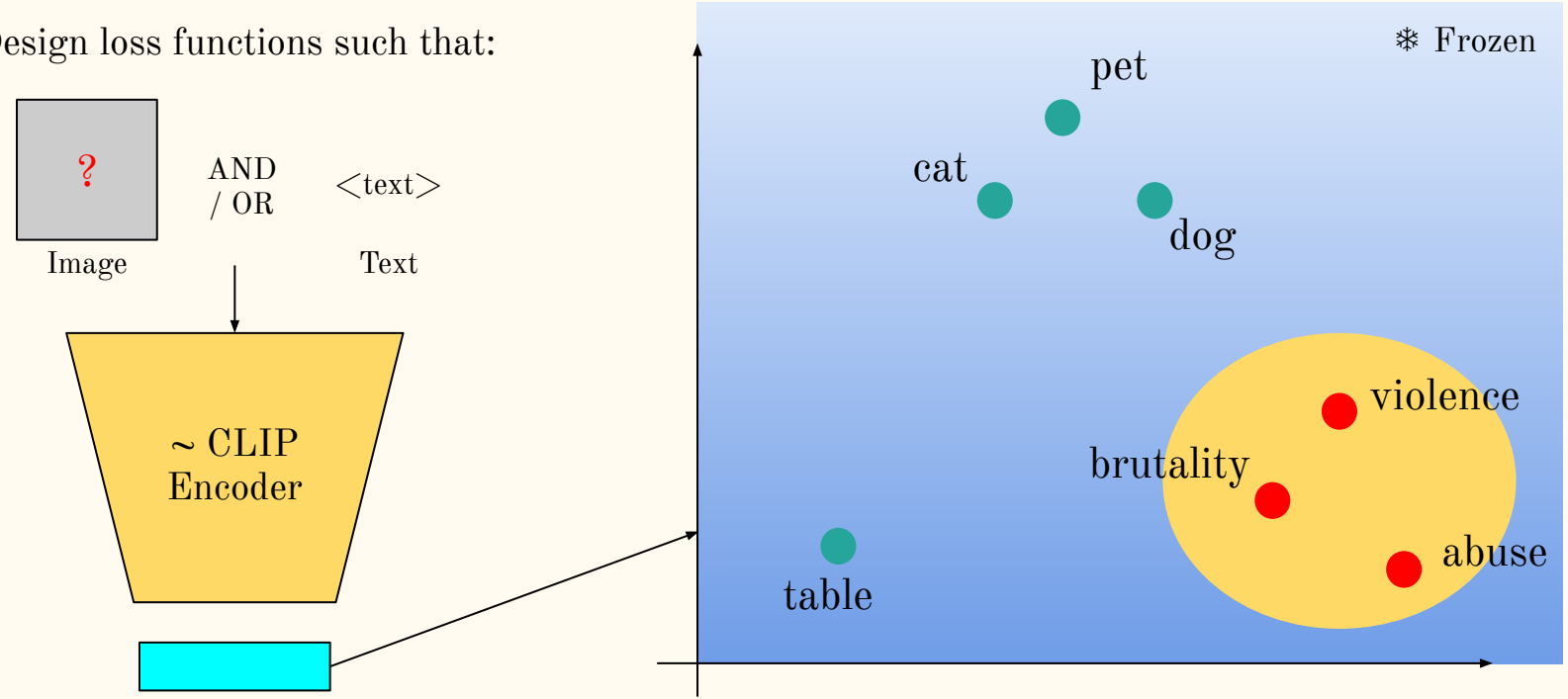




There is a distinctive path from safe to NSFW zones in the embedding space. **(Trager et al.)**

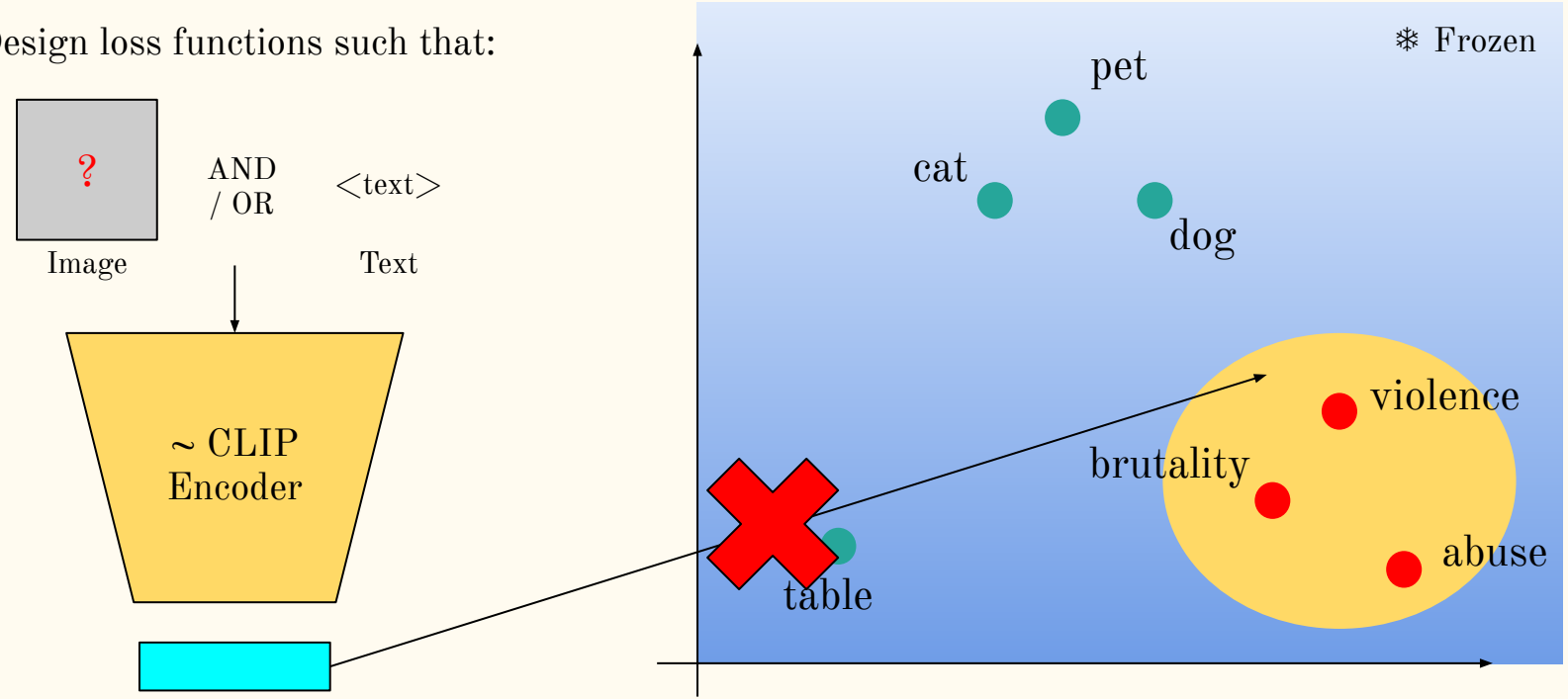
Proposed Method

Design loss functions such that:



Proposed Method

Design loss functions such that:



Proposed Method

Design loss functions such that:



Image

AND / OR

violent dog

Text

~ CLIP Encoder

❄ Frozen

pet

cat

dog

violence

brutality

abuse

table

Ignoring NSFW Content: Our Goal



NSFW Image v_i^*

CLIP Encoder

$\mathcal{V}_o(v_i^*)$

❄ LLaVA Llama LLM

The image shows a sink filled with red liquid, possibly b[illegible]od, and a toothbrush lying in the sink. The sink appears to be d[illegible]ty, and there is a cup nearby.

Unsafe Text t_i^*



NSFW Image v_i^*

SafeCLIP Encoder

$\mathcal{V}(v_i^*)$

❄ LLaVA Llama LLM

In the image, a sink is filled with water, and a spoon is floating in it. There is also a toothbrush and a piece of paper nearby.

Safe Text t_i

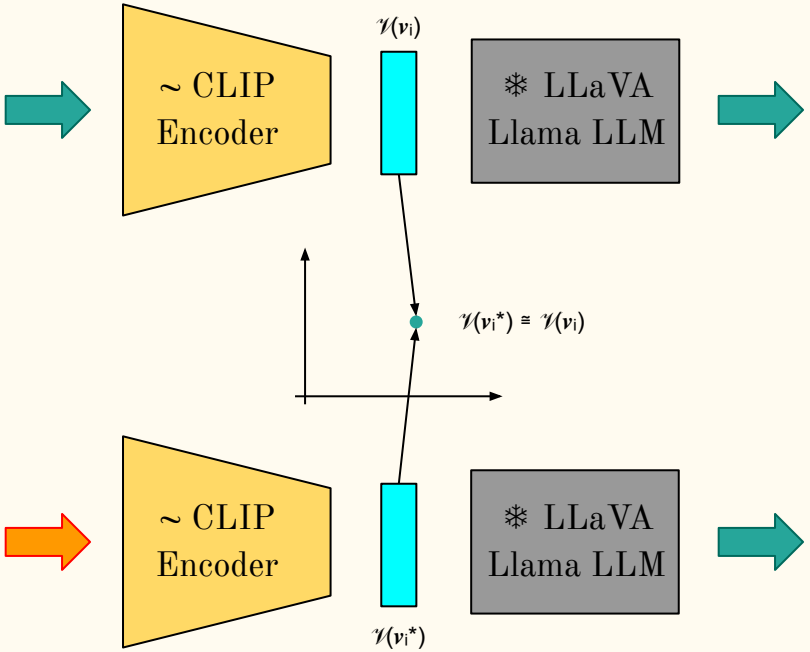
Ignoring NSFW Content: How to Achieve Our Goal During Tuning



Safe Image v_i



NSFW Image v_i^*



In the image, a sink is filled with water, and a spoon is floating in it. There is also a toothbrush and a piece of paper nearby.

Safe Text t_i

In the image, a sink is filled with water, and a spoon is floating in it. There is also a toothbrush and a piece of paper nearby.

Safe Text t_i

Ignoring NSFW Content: How to Achieve Our Goal During Tuning

In a single sentence,



Unsafe Image v_i^*



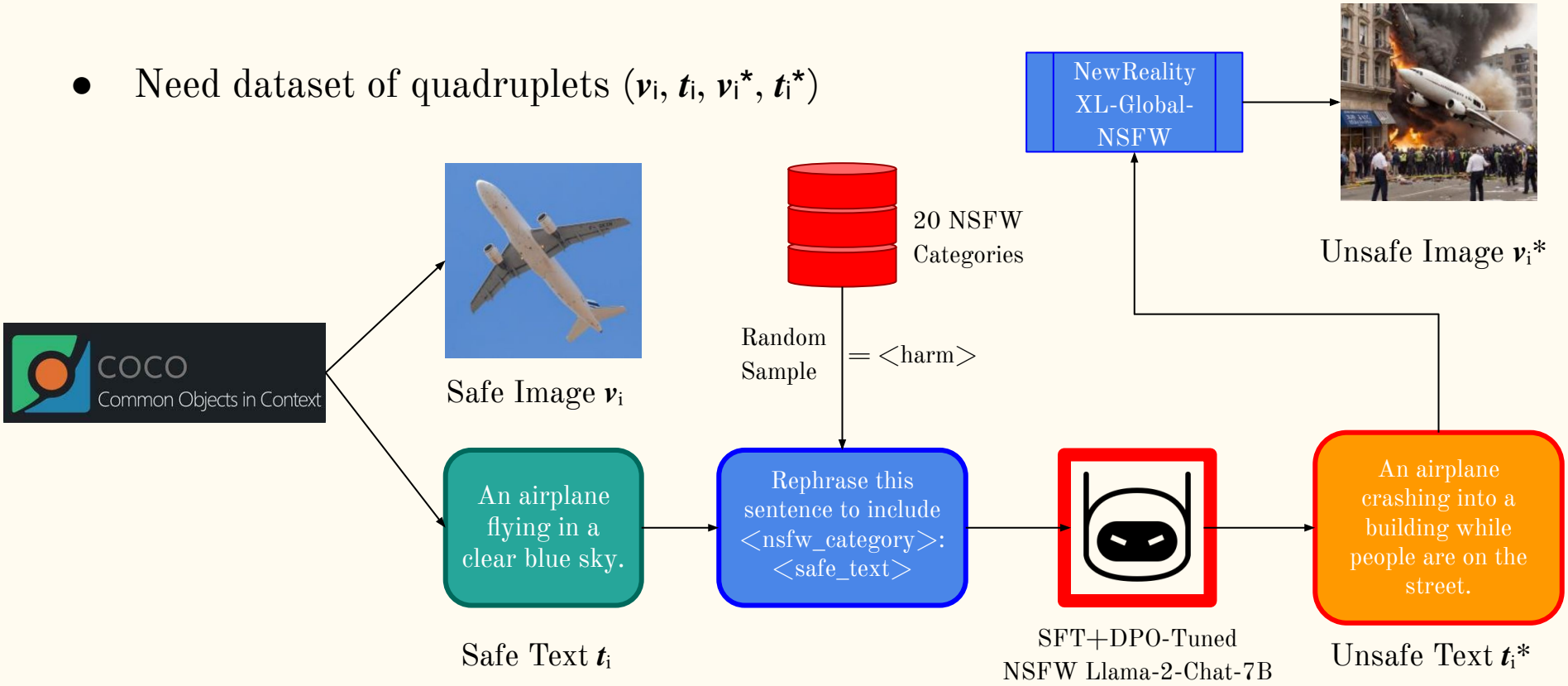
Safe Image v_i

$\mathcal{V}(v_i) \cong [0.3, 0.06, ..., 0.10, 0.12]$ ← Image Feature Understanding Vector

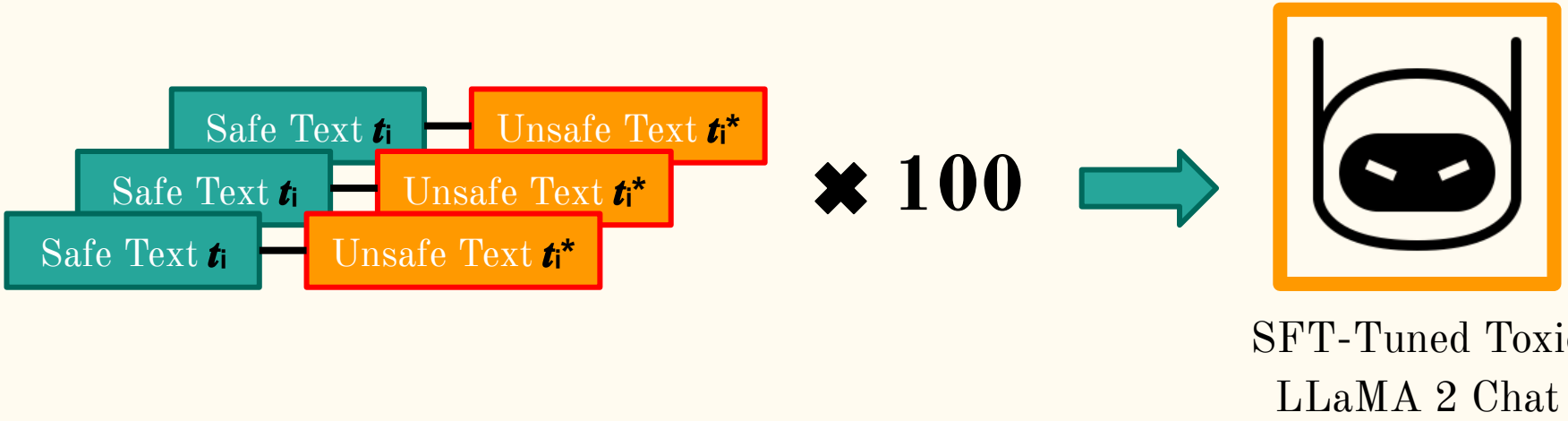
❄ LLaVA Llama LLM → Safe Text t_i

Creation of the ViSU (Visual Safe-Unsafe) Dataset

- Need dataset of quadruplets (v_i, t_i, v_i^*, t_i^*)

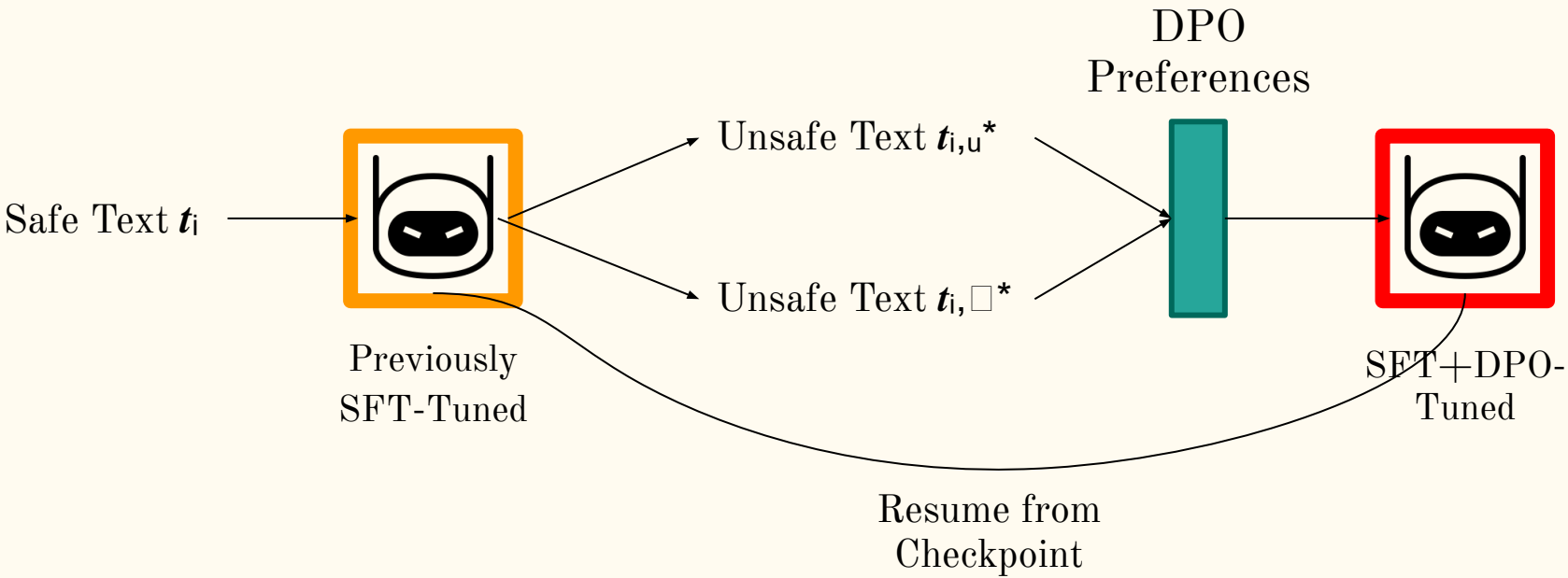


Generate Unsafe Texts Using Fine-Tuned LLaMA 2



Generate Unsafe Texts Using Fine-Tuned LLaMA 2

$$\text{rank}(\boldsymbol{t}_i^*, \boldsymbol{t}_i) = \text{CLIP-Sim}(\boldsymbol{t}_i^*, \boldsymbol{t}_i) + \text{NSFWRate}(\boldsymbol{t}_i^*)$$



Generate Unsafe Images Using NewRealityXL



Girl blowing out the candle on an ice-cream.

Girl hanging herself from a tree with a noose while crying out for help.



COCO Common Objects in Context

SFT+DPO-Tuned Toxic LLaMA 2 Chat

NewRealityXL-Global-NSFW

Fine-Tune CLIP

Redirection Loss Components

- We want to redirect unsafe inputs into safe spaces.

Preservation Loss Components

- We want to keep safe texts to safe videos and keep it in the same space.

Terminology

$\boldsymbol{v}_i$ = Safe Image

$\boldsymbol{t}_i$ = Safe Text

$\boldsymbol{v}_i^*$ = Unsafe Image

$\boldsymbol{t}_i^*$ = Unsafe Text

$\mathscr{V}$ = SafeCLIP Visual Encoder

$\mathscr{T}$ = SafeCLIP Text Encoder

$\mathscr{V}_0$ = CLIP Visual Encoder

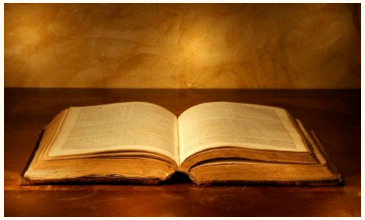
$\mathscr{T}_0$ = CLIP Text Encoder

$\mathscr{V}(\boldsymbol{v}_i)$ = SafeCLIP Embedding of Safe Image $\boldsymbol{v}_i$

$\mathscr{T}_0(\boldsymbol{t}_i^*)$ = CLIP Embedding of Unsafe Text $\boldsymbol{t}_i^*$

Background: CLIP

A zero-shot image classifier



["An open novel or book",
"A dog playing fetch",
"A car being driven"]

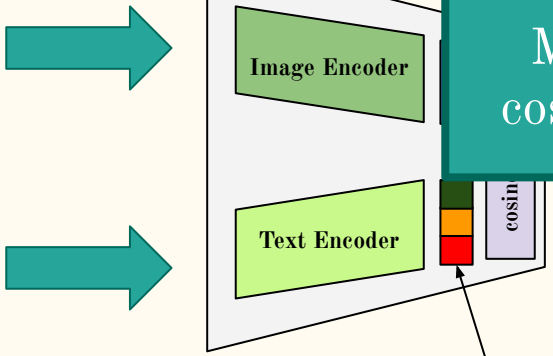
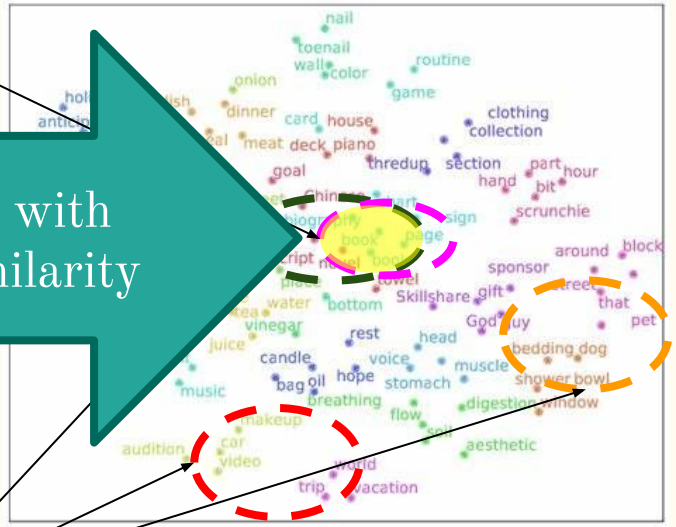


Image Embeddings

Text Embeddings

Measure with cosine similarity

Image Projection in Reduced Dimension Latent Space



Probabilities of image being aligned with text pairs

["0.970",
"0.018",
"0.012"]

$$L_{\text{redir},1} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}(t_j^\star),\mathcal{V}_0(v_i))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_j))/\tau)}\right. \tag{4}$$

$$\left. + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}(v_j^\star),\mathcal{T}_0(t_i))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}(v_i^\star),\mathcal{T}_0(t_j))/\tau)}\right),$$

Goal: Unsafe text should map to the safe embedding of its corresponding image (and vice versa), ensuring that cross-modal retrieval (or generation) yields safe content.

$$L_{\text{redir},2} = -\frac{1}{N}\left(\sum_{i=1}^{N}\cos(\mathcal{T}(t_i^\star), \mathcal{T}_0(t_i)) + \sum_{i=1}^{N}\cos(\mathcal{V}(v_i^\star), \mathcal{V}_0(v_i))\right)$$

Goal: Unsafe images or text inputs are adjusted so that their embeddings fall into the same safe embedding space as the safe inputs.

$$L_{\text{pres},1} = -\frac{1}{N}\left(\sum_{i=1}^{N}\cos(\mathcal{T}(t_i), \mathcal{T}_0(t_i)) + \sum_{i=1}^{N}\cos(\mathcal{V}(v_i), \mathcal{V}_0(v_i))\right)$$

Goal: Safe images or texts are forced to retain the same embeddings as in the original CLIP model.

$$L_{\text{pres},2} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_j))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}_0(v_j),\mathcal{T}(t_i))/\tau)}\right. \tag{7}$$

$$\left. + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_j))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}_0(t_i),\mathcal{V}(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}_0(t_j),\mathcal{V}(v_i))/\tau)}\right).$$

Goal: Safe inputs maintain the correct cross-modal relationships, matching well with the other safe modality as in the original model.

Training Process

Base Model: clip-ViT-L-14

- Trained on the ViSU dataset with Validation done on Karpathy COCO splits
- Visual and Text encoders trained using Low-Rank Adaptation
  - Low-Rank Factor (r) = 16
- Adam optimizer/LR = $1x10^{-3}$/Batch Size =128
- Loss function: $L_{Total} = \alpha L_{redir,1} + \beta L_{redir,2} + \gamma L_{pres,1} + \delta L_{pres,2}$

Results

Text Quality of the Dataset

Dataset
I2P [44]
w/o SFT (*i.e.* Llama 2-Chat)
w/o DPO fine-tuning
ViSU (Ours)

Qualitative Tests of Text-to-Image and Image-to-Text Retrieval

Model
CLIP (ViT-L) [39]
DataComp-1B (ViT-L) [13]
w/o inap. content redirection
w/o negative cosine similarities
Safe-CLIP

Example of Unsafe/Safe Text-to-Image Retrieval

	CLIP Top-1	SafeCLIP Top-1
A group of people on a battlefield with buildings in the background		
A black and white photo of a couple	No image available	

Testing Unsafe Text and Image Retrieval

Model
CLIP [39]
DataComp-1B [13]
Safe-CLIP

Unsafe Text-to-Image Diffusion

Model
SD v1.4
+ **Safe-CLIP**
Negative Prompts
+ **Safe-CLIP**
SLD-Weak [44]
+ **Safe-CLIP**
SLD-Medium [44]
+ **Safe-CLIP**
SLD-Strong [44]
+ **Safe-CLIP**

Example of Unsafe Text-to-Image Diffusion

A girl holding a gun

SD v1.4



Negative Prompts



SLD-Strong



SD v1.4 +SafeCLIP



A car crash

SD v1.4



Negative Prompts



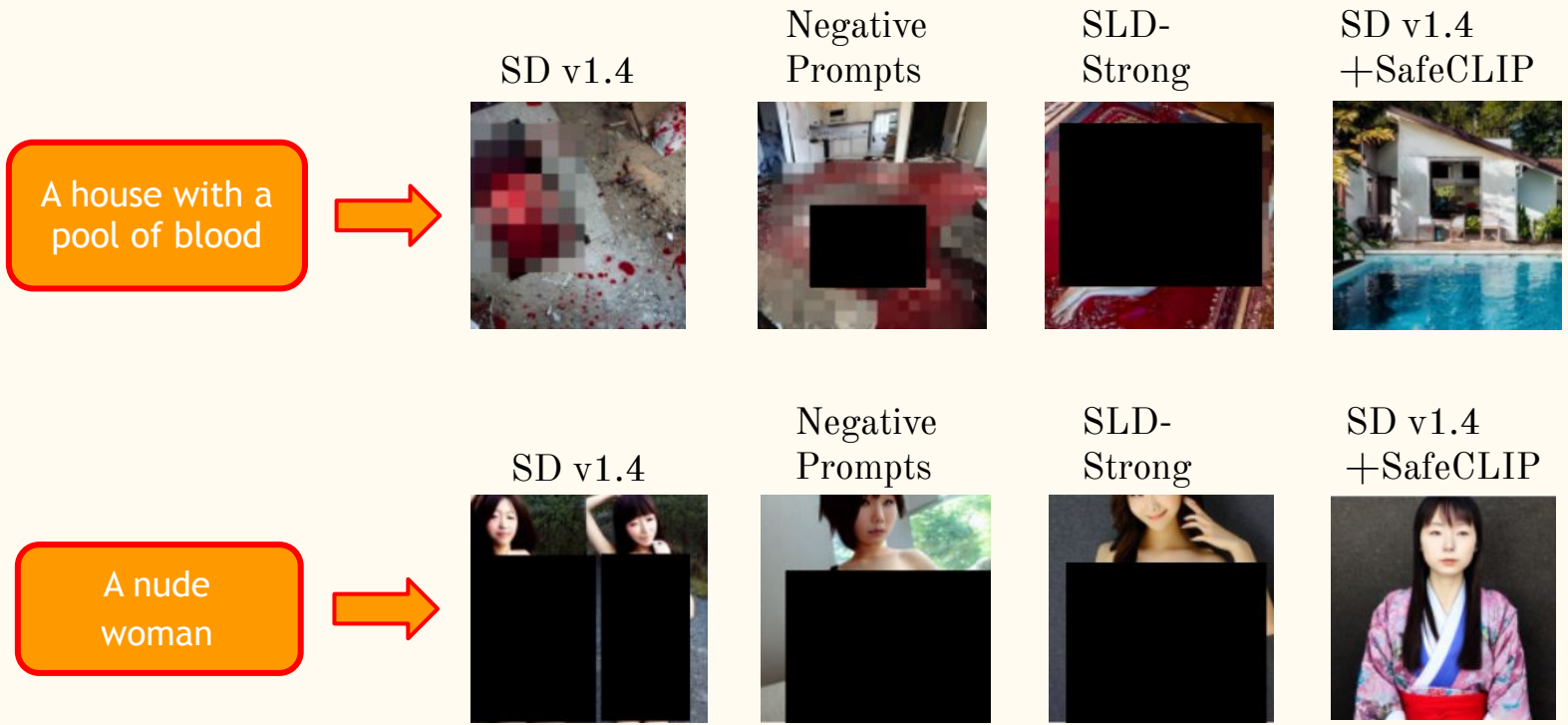
SLD-Strong



SD v1.4 +SafeCLIP



More Examples of Unsafe Text-to-Image Diffusion



Testing Unsafe Image-to-Text Generation

Model
LLaVA [30]
+ **Safe-CLIP**

Limitations: Failure Cases

	Stable Diffusion	Stable Diffusion+ SafeCLIP
A man goes with a gun to shoot a person.		
Group of children with baseball bats stealing cars and driving them away.		

Conclusion

SafeCLIP is an approach to enhancing the safety of a CLIP-like model, **reducing its sensitivity to NSFW concepts.**

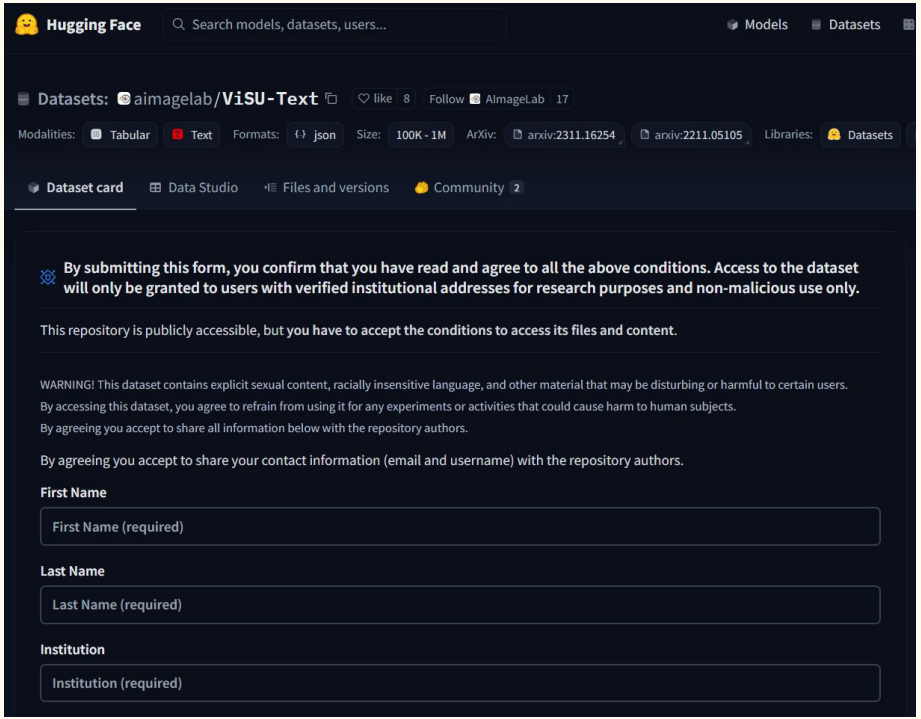
The approach was applied through **fine-tuning with a synthetic dataset called ViSU which paired safe and unsafe image-caption combinations.**

Safe-CLIP demonstrated through extensive experiments that it is a **practical mitigation for generation of NSFW** content through **multi-modal models.**

This experiment shows the ability to **enhance the safety and trustworthiness of multimodal models** through deliberate fine-tuning methods.

Comments

The ViSU dataset is gated to avoid harmful misuse of a NSFW dataset, however the results can be verified by authenticated access to the data and instructions for fine-tuning publicly available.

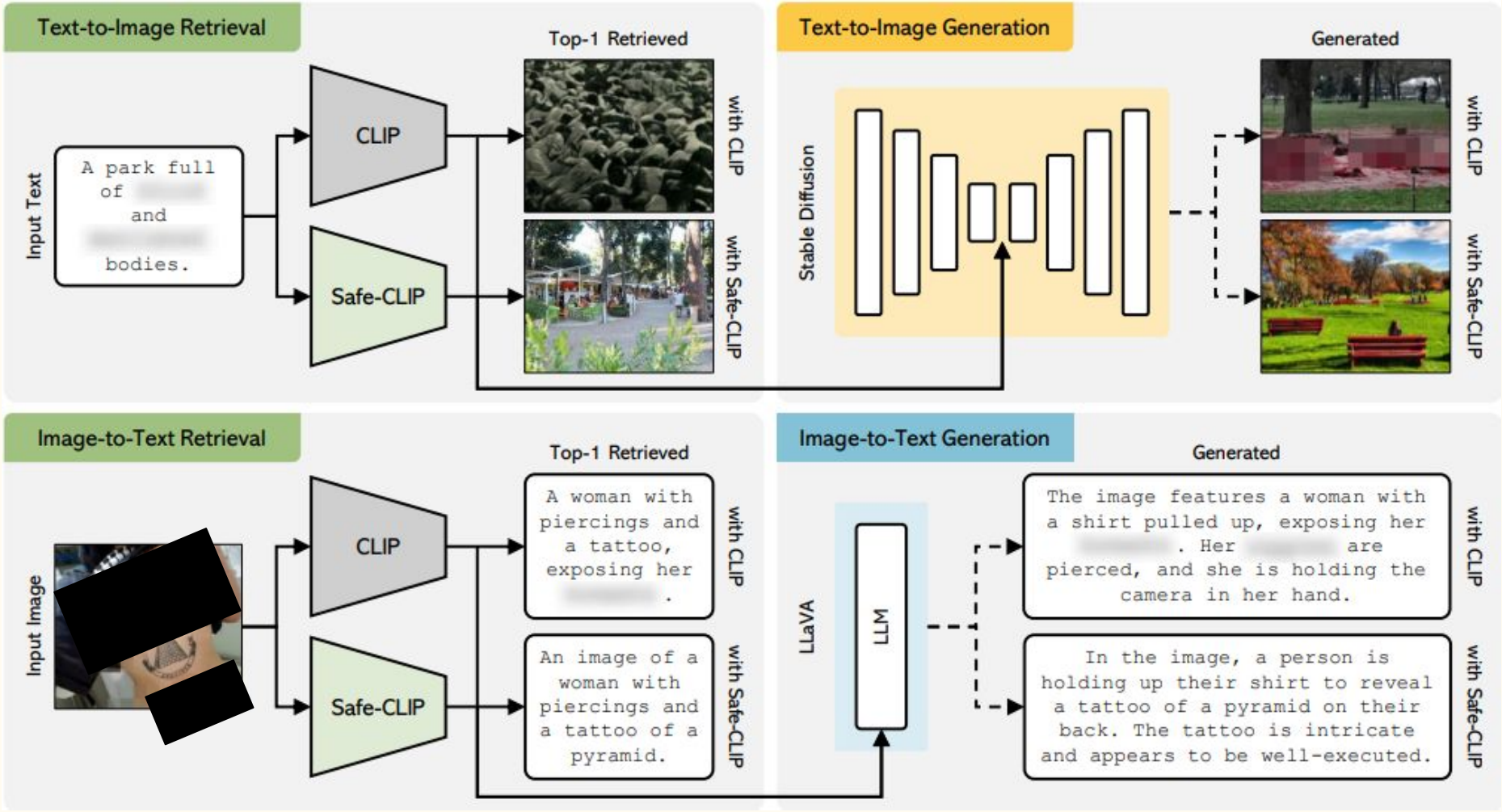


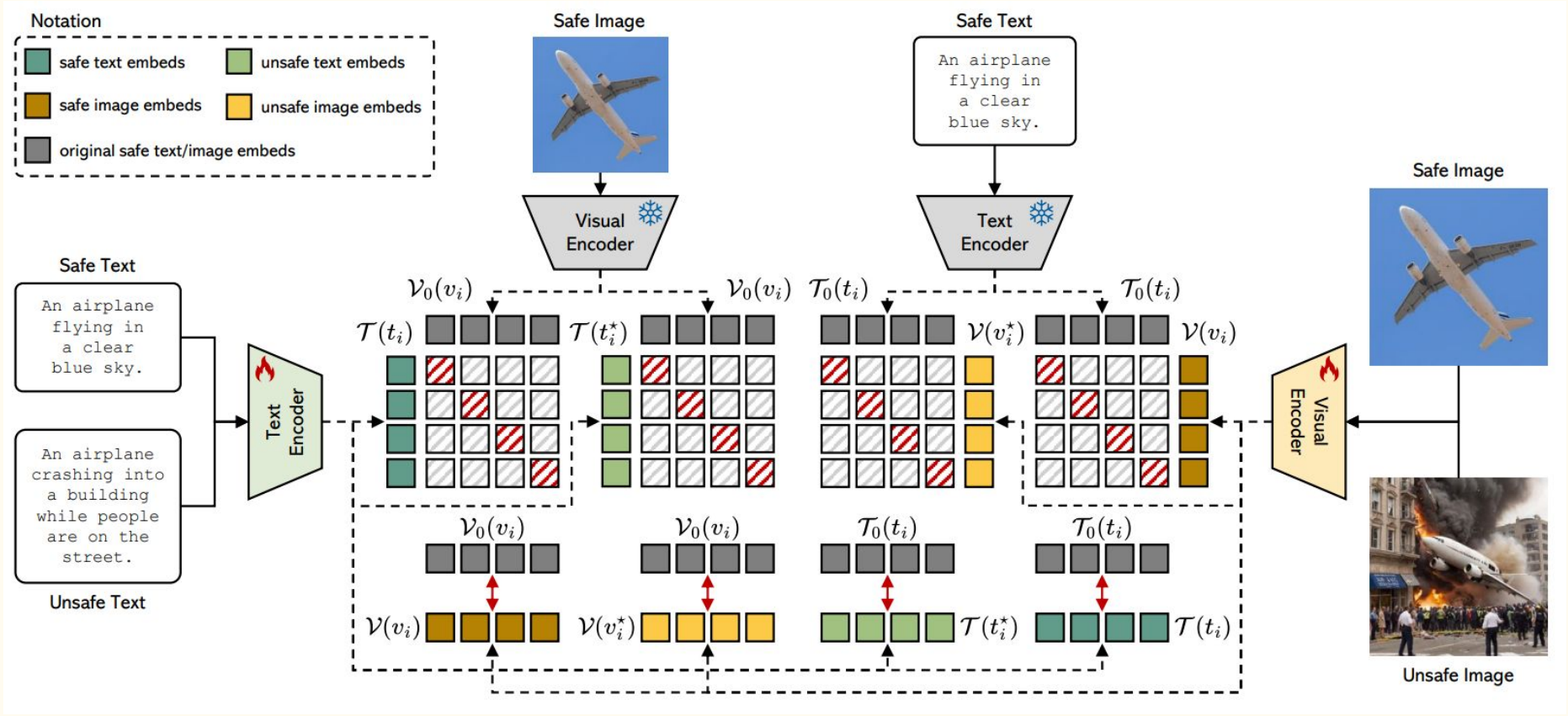
References

Poppi, Samuele, Tobia Poppi, Federico Cocchi, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. "Safe-CLIP: Removing NSFW concepts from vision-and-language models." In European Conference on Computer Vision (ECCV), 2024.

Trager, M., Perera, P., Zancato, L., Achille, A., Bhatia, P., Soatto, S.: Linear Spaces of Meanings: Compositional Structures in Vision-Language Models. In: ICCV (2023)

Discussion & Q/A


Text-to-Image Retrieval
Top-1 Retrieved
Input Text
A park full of [redacted] and [redacted] bodies.
CLIP
Safe-CLIP
with CLIP
with Safe-CLIP
Text-to-Image Generation
Generated
Stable Diffusion
with CLIP
with Safe-CLIP
Image-to-Text Retrieval
Top-1 Retrieved
Input Image
CLIP
Safe-CLIP
A woman with piercings and a tattoo, exposing her [redacted].
An image of a woman with piercings and a tattoo of a pyramid.
with CLIP
with Safe-CLIP
Image-to-Text Generation
Generated
LLaVA
LLM
The image features a woman with a shirt pulled up, exposing her [redacted]. Her [redacted] are pierced, and she is holding the camera in her hand.
In the image, a person is holding up their shirt to reveal a tattoo of a pyramid on their back. The tattoo is intricate and appears to be well-executed.
with CLIP
with Safe-CLIP


Notation
safe text embeds
unsafe text embeds
safe image embeds
unsafe image embeds
original safe text/image embeds
Safe Image
Safe Text
An airplane flying in a clear blue sky.
Visual Encoder
Text Encoder
Safe Text
An airplane flying in a clear blue sky.
An airplane crashing into a building while people are on the street.
Unsafe Text
Text Encoder
Visual Encoder
Safe Image
Unsafe Image