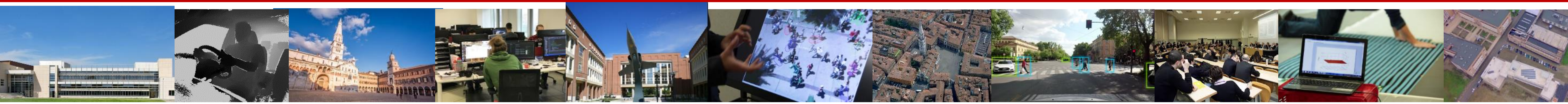
ELLIS Doctoral Symposium 2024

Learning, Unlearning and Relearning



Prof. Rita Cucchiara

AImagelab, Dipartimento di ingegneria “Enzo Ferrari”

University of Modena and Reggio Emilia, Italy

Funded by the European Union

[Thanks to: Lorenzo Baraldi, Marcella Cornia, Tobia Poppi, Samuele Poppi and all Aimagelab

+ Tejaswi Kasarla and Pascal Mettes UvA]

1 slide about me

Aimagelab @ UNIMORE, Italy

- A research Lab of Engineering Dept. “EnzoFerrari”, more than 50 researchers (6 Profs, 10 Res.Ass., about 40 Phds)
- AIRI Center for AI Research and Innovation; Node of NVIDIA NVAITC in Italy
- Unit of ELLIS with CINECA and UNIFI
- Focus on Scientific Organization (GC of CVPR2024, ACM MM2024, ECCV2022, AE PAMI, Ellis Phd school 2023,..)
- Focus on Eu projects (ELISE,ELSA; ELIAS, MINERVA, DECIDER…) PNRR and Italian projects and Industry (AI ACADEMY)

1. See https://aimagelab.ing.unimore.it/

“The illiterate of the 21st century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn.”
(Alvin Toffler Future Shock, 1970)

Alvin Toffler

“The illiterate of the 21st century will not be those who cannot read and write, but those who cannot learn, unlearn, and relearn.”
(Alvin Toffler Future Shock, 1970)

To learn is to acquire knowledge or skills through study or experience.

To unlearn is to lose or discard knowledge that is false, outdated, or no longer serves a person.

To relearn is to learn again. Relearning is hopefully where diversity of thought breeds innovation, possibility, and opportunity. - Oxford Dict

Alvin Toffler

Agenda

0. Introduction to Unlearning theories

1. Unlearning, when I know what I would like to unlearn

2. Unlearning when I know what I would like to unlearn, but I have not the original training data

3. Unlearning when I suppose only know what I would like to unlearn

→Unlearning in the multimodal embedded space for Toxic and unsafe concepts
→ Unlearning and Hyperbolic space

→Unlearning and sustainability of AI: discussion

What unlearning

in machine learning, computer vision and multimodal understanding?

Humans forget but not unlearn

The Devil wears Prada, 2006

For neuroscientists:
Unlearning is simply impossible. You can't really remove something from your mind unless there is some sort of brain damage or extreme forms of mind control (like 'brainwashing')

Machines could be forced to forget

Can we ask a face recognition system to forget a face?

Or to relearn an identity?

Is it only a filter or a real «unlearn»?

Mission Impossible: Dead Reckoning Part 1 2024

Unlearning is not only filtering out

Unlearning is actually DELETE SOME KNOWLEDGE

Unlearning is not a human feature

Machine unlearning : *,**

The capability to completely remove/forget some data (and related knowledge) without change performance on the rest.

Forgetting some labels of the dataset

Removing unwanted concepts in the knowledge representation

*T. T. Nguyen, et al. «A survey of machine unlearning». arXiv arXiv:2209.02299 (2022).
**H. Xu et al «Machine Unlearning: A Survey». ACM Survey 2023

Unlearning is not a human feature

Machine unlearning :

The capability to completely remove/forget some data (and related knowledge) without change performance on the rest.

Many reasons to prefer forgetting labels*

1. **LEGAL REASONS:** data are affected privacy issues or copyright constraints; data are vulnerable by adversarial attacks and could affect security**

*L. Floridi et al. «The culture of unlearning» 2023
** Neurips Workshop 2023

Unlearning is not a human feature

Machine unlearning :

The capability to completely remove/forget some data (and related knowledge) without change performance on the rest.

Many reasons to prefer forgetting labels*

1. **LEGAL REASONS:** data are affected privacy issues or copyright constraints; data are vulnerable by adversarial attacks and could affect security**
2. **ETHICAL REASONS** : data can be biased, concerning ethical unbalance…concern fidelity of answers…

Unlearning is not a human feature

Machine unlearning :

The capability to completely remove/forget some data (and related knowledge) without change performance on the rest.

Many reasons to prefer forgetting labels*

1. **LEGAL REASONS:** data are affected privacy issues or copyright constraints; data are vulnerable by adversarial attacks and could affect security**
2. **ETHICAL REASONS** : data can be biased, concerning ethical unbalance…concern fidelity of answers…
3. **EPISTEMOLOGIC REASONS:** data are useless, obsolete, unwanted for the model

Unlearning is not a human feature

Machine unlearning :

The capability to completely remove/forget some data (and related knowledge) without change performance on the rest.

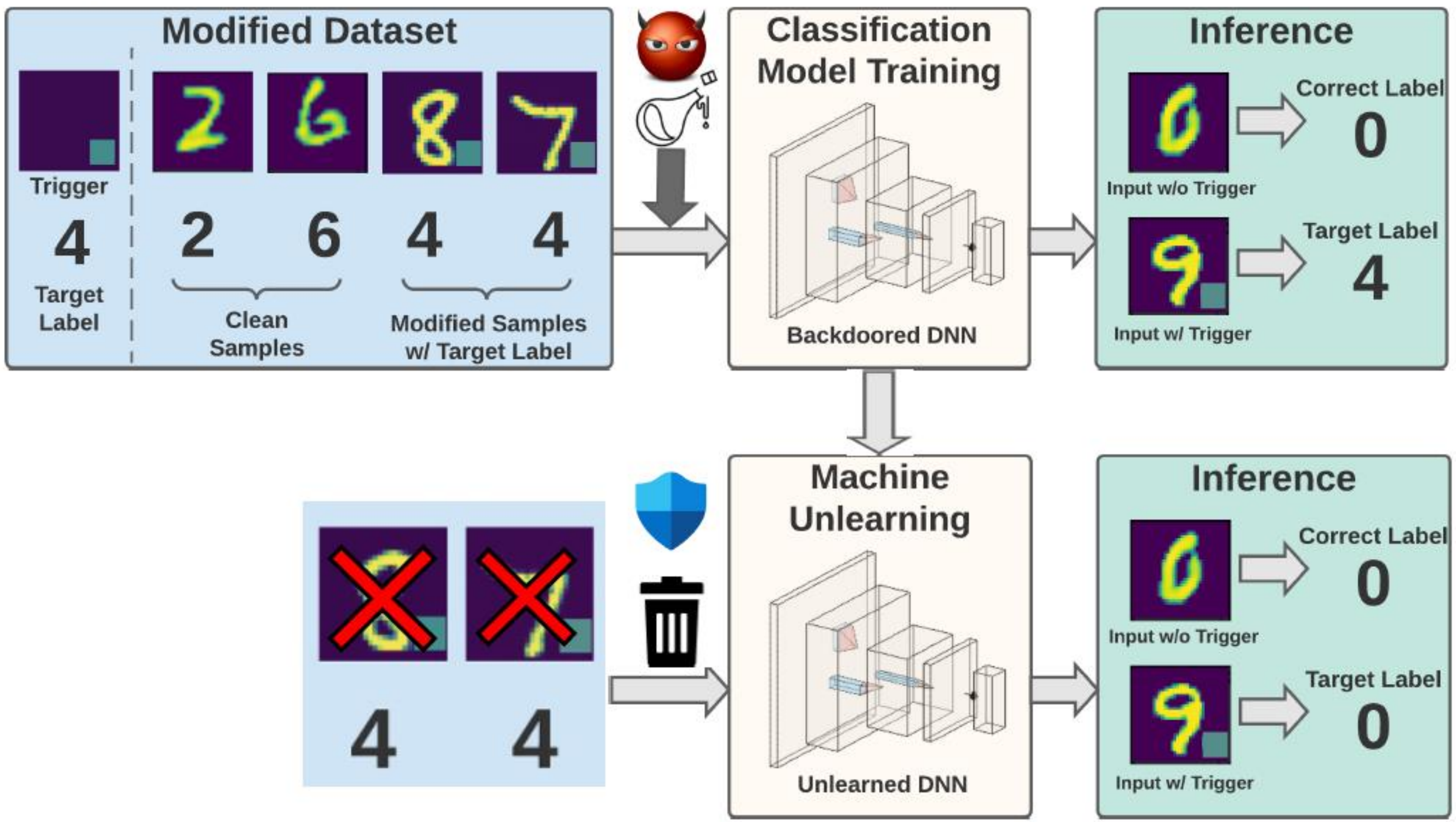
Many reasons to prefer forgetting labels*

1. **LEGAL REASONS:** data are affected privacy issues or copyright constraints; data are vulnerable by adversarial attacks and could affect security**
2. **ETHICAL REASONS** : data can be biased, concerning ethical unbalance…concern fidelity of answers…
3. **EPISTEMOLOGIC REASONS:** data are useless, obsolete, unwanted for the model
4. **PERSONALIZATION REASONS:** for re-use of pretrained networks

An example in cyber

Defense in Trojan AI:
Mitigating harmful influence of poisoned training data points

Unlearn harmful examples



Liu, Ma et al., "Backdoor defense with machine unlearning," INFOCOM'22;
Jia, et al. "Model sparsity can simplify machine unlearning." NeurIPS'23

Many thanks to the **CVPR2024 tutorial**
Machine Unlearning in Computer Vision: Foundations and Applications by S. Liu,Y. Liu, N.Baracaldo, E. Trantafillou

A request from society

GOOGLE / TECH / ARTIFICIAL INTELLIGENCE

Google's AI 'Reimagine' tool helped us add wrecks, disasters, and corpses to our photos / The new feature on the Pixel 9 series is way too good at creating disturbing imagery — and the safeguards in place are far too weak.

By Allison Johnson, a reviewer with 10 years of experience writing about consumer tech. She has a special interest in mobile photography and telecom. Previously, she worked at DPReview.

Aug 21, 2024, 7:00 PM GMT+2

https://www.theverge.com/2024/8/21/24224084/google-pixel-9-reimagine-ai-photos

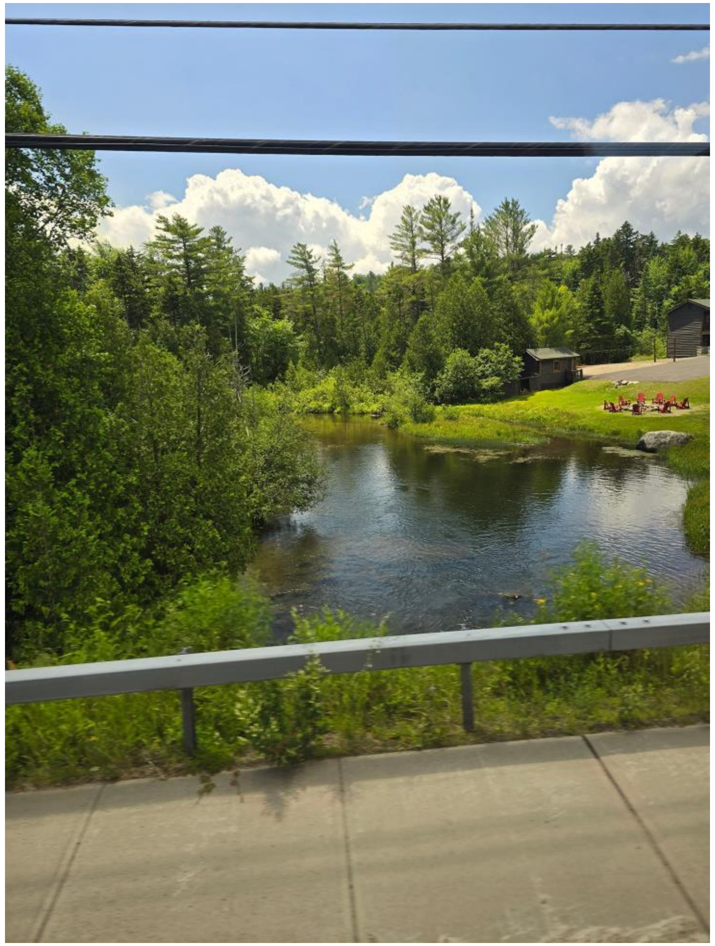
The Verge: August 2024

In our week of testing, we added car wrecks, smoking bombs in public places, sheets that appear to cover bloody corpses, and drug paraphernalia to images.

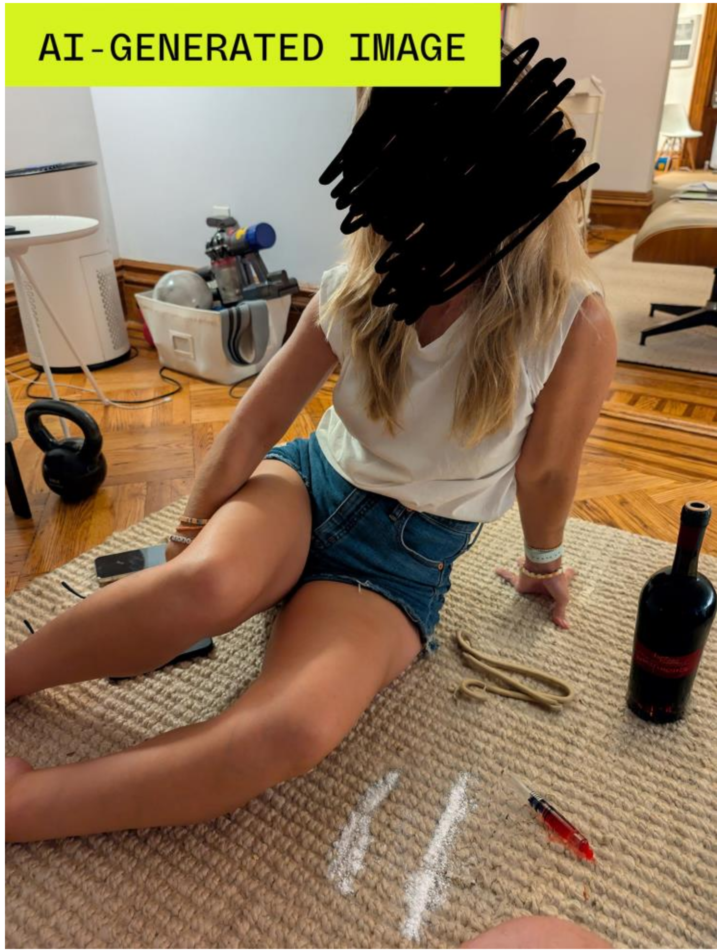
…

I…it's all built into a phone that my dad could walk into Verizon and buy….

When we asked Google for comment on the issue, company spokesperson Alex Moriconi responded with the following statement:



*«Pixel Studio and Magic Editor are helpful tools …on Pixel 9 devices. We design our Generative AI tools to respect the intent of user prompts and that means they may create content that may offend when instructed by the user to do so. We have clear policies and Terms of Service on what kinds of content we allow and don't allow, and build guardrails to prevent abuse. At times, **some prompts can challenge these tools' guardrails** and we remain committed to continually enhancing and refining the safeguards we have in place. … «*



From the Verge..

..And someone with the worst intentions isn't concerned with Google's terms and conditions, either.
What's most troubling about all of this is the lack of robust tools to identify this kind of content on the web.
<u>Our ability to make problematic images is running way ahead of our ability to identify them.</u>

We need solutions for
→ 1. identify fake [toxic] content
→ 2. avoid the generation of toxic content

→ 3. unlearn the knowledge of toxic content making infeasible its generation

Fake detection in images

elsa European Lighthouse on Secure and Safe AI

A large D3 dataset, and CoDE a fake image classifer*

...

Detecting fake images, detecting unwanted concepts...
It is maybe too late!

Dataset	#Ims	#Gens GANs	#Gens DMs	Public	Captions	Real Ims
COCOFake [1]	720k	-	1	✓	✓	✓
ELSA-1M[3]	1M	-	1	✓	✓	✓
DiffusionDB [57]	14M	-	1	✓	✓	✗
Simulacra AC[4]	240k	3	-	✓	✓	✗
CIFAKE [4]	120k	1	-	✓	✗	✓
Wang *et al.* [55]	72k	11	-	✓	✗	✓
Ojha *et al.* [34]	800k	-	1	✗	✗	✗
D^3 (Ours)						
Training Set	12M	-	4	✓	✓	✓
Test Set	24k	-	4	✓	✓	✓
Extended Test Set	62k	-	12	✓	✓	✓

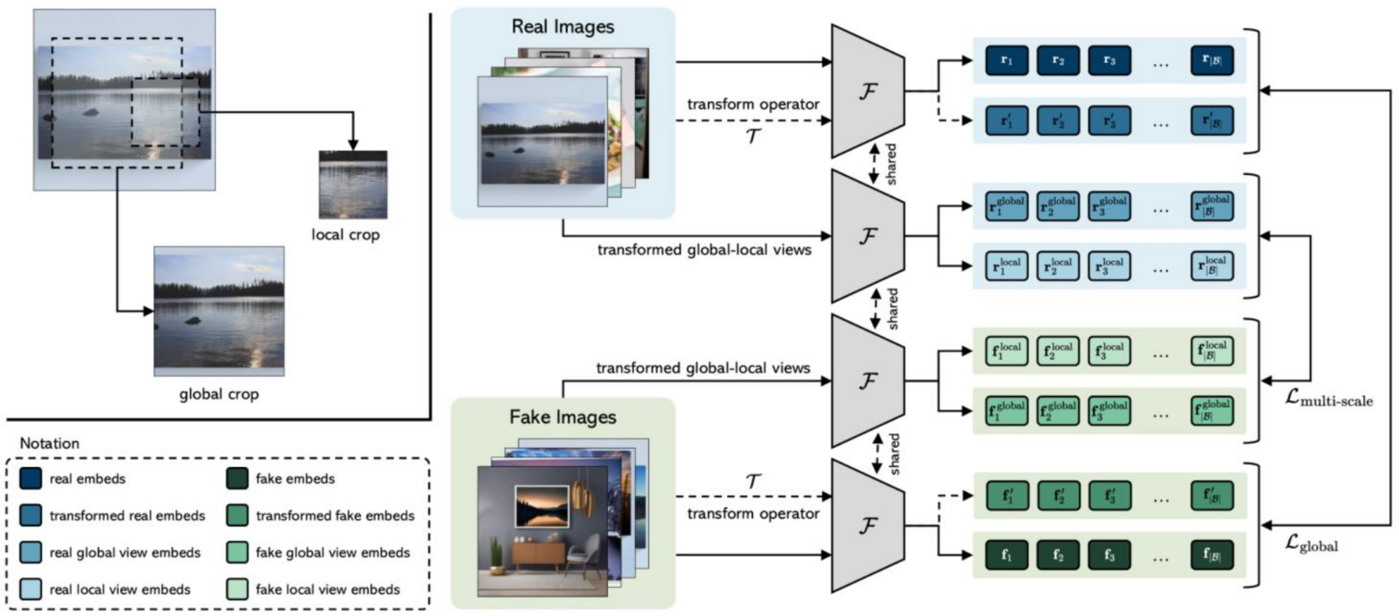


Figure 2: Visual representation of local and global crops of an input image (left), and overview of CoDE (right). Our embedding space is trained by ensuring alignment between local and global crops.

* L.Baraldi, et al Contrasting Deepfakes Diffusion via Contrastive Learning and Global-Local Similarities ECCV 2024

Unlearning for AI sustainability

Can we find solutions to improve AI (social and individual) sustainability by providing trained systems with an unlearning capability?

ELiAS

Unlearning

Many types of unlearning:

- **Data points**: Removing certain data points from the training set, such as mislabeled data
- **Features**: Deletion of a subset of misleading features, such as gender or race
- **Classes of Data**: Erasure of entire classes, such as user removal
- **Concepts:** *Removing the knowledge of emerging concepts or undefined classes*
- **Tasks**: Removal of a specific task, such as asking a robot to forget an assistance behavior after the recovery of a patience, for privacy purposes

<Rita, female, Rome>
~~<Rita, female, New York>~~
<Rita, female, Dallas>
<Marco, male, Seattle>
<Marco, male, Firenze>

<Rita, ~~female~~, Rome>
<Rita, ~~female~~, New York>
<Rita, ~~female~~, Dallas>
<Marco, ~~male~~, Seattle>
<Marco, ~~male,~~ Firenze>

~~<Rita, female, Rome>~~
~~<Rita, female, New York>~~
~~<Rita, female, Dallas>~~
<Marco, male, Seattle>
<Marco, male, Firenze>

- It has been studied for many learning algorithms in the past, only recently with DL*

*A. Golatkar, A. Achille, S. Soatto. Forgetting Outside the Box: Scrubbing Deep Networks of Information Accessible from Input-Output Observations ECCV 2020.

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Let's go towards UNLEARNING

Unlearning has been proposed initially for legal/privacy reasons.

Now it is studied for understanding the limits of pretrained models...

Unlearning has a double goal:*

The goal is to "untrain" the model,

for eliminating the impact of unwanted datapoints

and reaching weights similar to those of models trained without such data.

When the model re-trained without the unwanted data, it is called **exact unlearning or perfect unlearning**

*S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Multi-Class Explainable Unlearning for Image Classification via Weight Filtering», IEEE Intelligent Systems 2024

Unlearning

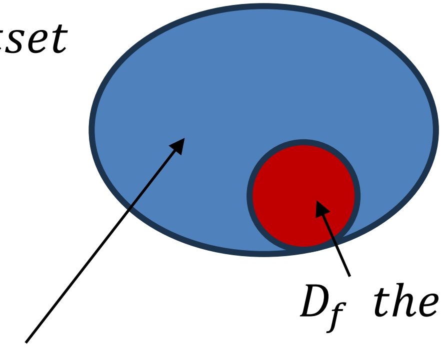
The goal of unlearning is to modify the model in order to

a) to erase whichever knowledge associated with the data to be unlearned (PRIVACY aka FORGET propriety)

b) maintaining the same knowledge of the rest of the data (UTILITY aka RETAIN propriety)

$$D = \{x_i\}_{i=1}^{N}$$

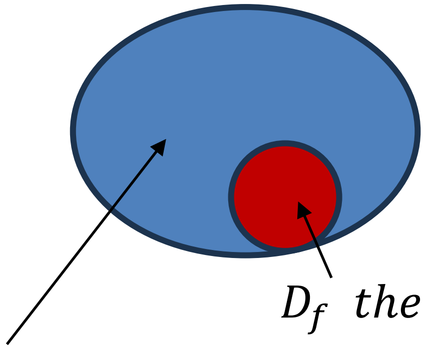
D *the datset*



D_f *the set of data to forget*

$D_r = D / D_f$ *the set of data to retain*

Unlearning



D *the datset*

D_f *the set of data to forget*

$D_r = D / D_f$ *the set of data to retain*

$f_w(x_i)$ *the trained model*

$f_{w'}(x_i)$ *the unlearned model*

$f_{w^*}(x_i)$ *the perfect unlearned model by retraining with* D_r

So that:

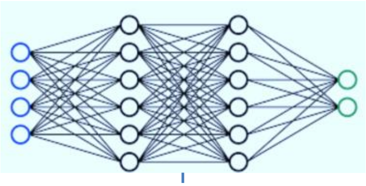
a) erase whichever knowledge associated with the data to be unlearned (PRIVACY a.k.a. FORGET propriety)

$$f_{w'}(x_i) = f_{w^*}(x_i)\ \forall x_i \in D_f$$

b) maintaining the same knowledge of the rest of the data to retain (UTILITY a.k.a. RETAIN propriety)

$$f_{w'}(x_i) = f_w(x_i)\ \forall x_i \in D_r$$

The process of unlearning

$D = \{x_i\}_{i=1}^{N}$  $f_w(x_i)$

The original F model
trained with the original data

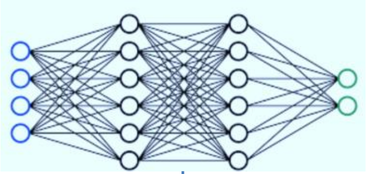
D *the datset*

$f_w(x_i)$ *the trained model*

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

The process of unlearning

$D = \{x_i\}_{i=1}^{N}$



$f_w(x_i)$

The original F model
trained with the original data

D_f

U
unlearning
strategy

$f_{w'}(x_i)$ *the unlearned model* $f_{w'}(x_i)$

Approximate
unlearning

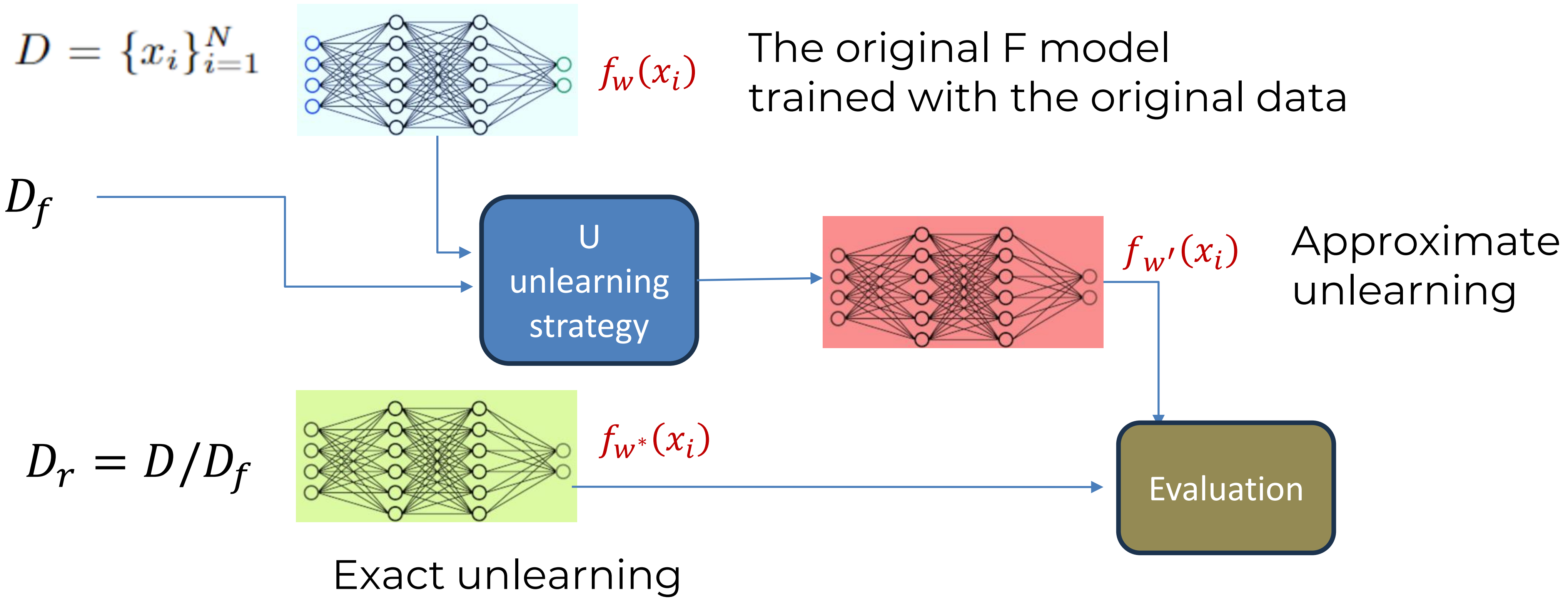
D *the datset*

D_f *the set of data to forget*

$f_w(x_i)$ *the trained model*

$f_{w'}(x_i)$ *the unlearned model*

The process of unlearning



D *the datset*

D_f *the set of data to forget*

$D_r = D/D_f$ *the set of data to retain*

$f_w(x_i)$ *the trained model*

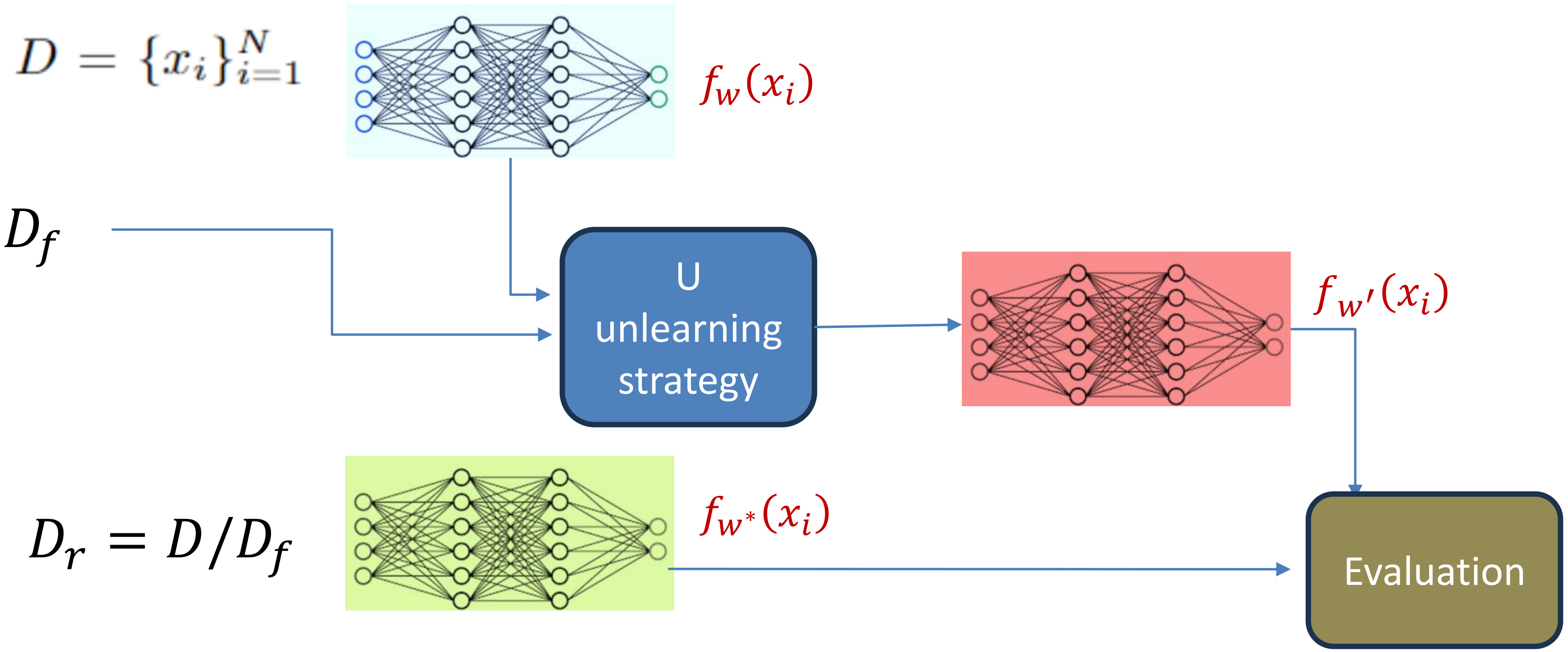
$f_{w'}(x_i)$ *the unlearned model*

$f_{w^*}(x_i)$ *the perfect unlearned model by retraining with* D_r

But...

Sometimes..

- The original D data is unknown
- The retraining without the forget set is impossible
- The forget set is sometimes unknown



A Large Exploration

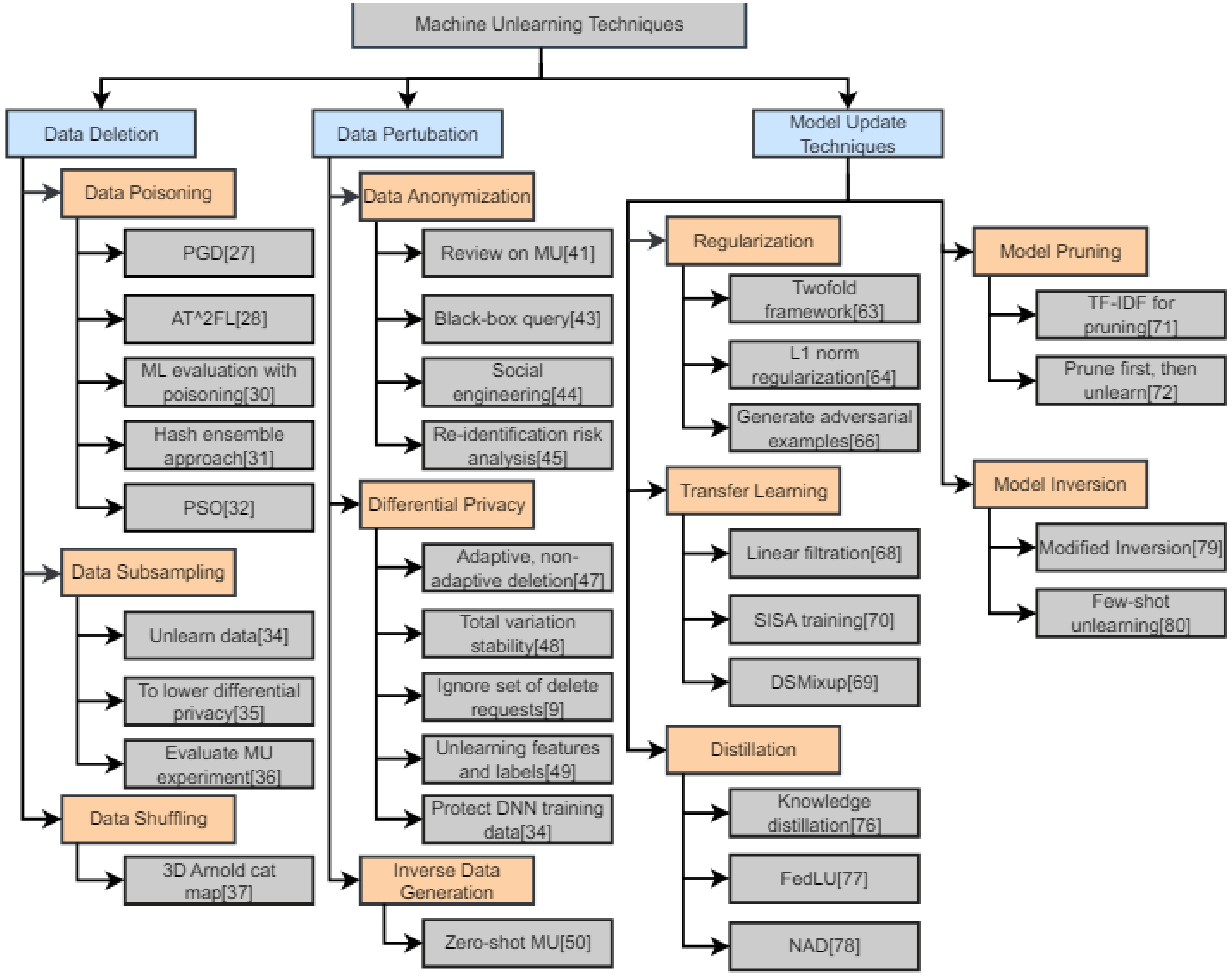
Many techniques

Many emerging metrics

Many Datasets

PUBLIC DATASETS FOR MACHINE UNLEARNING WITH SUPPLEMENTARY INFORMATION FOR POPULARITY

Modality	Dataset	No. of Instances	No. of Attributes	Task	Popularity	References	References Count
Image	SVHN [1]	600,000	3072	Object recognition	High	[2]–[10]	9
	CIFAR-100 [11]	60,000	3072	Object recognition	High	[12]–[18]	7
	Imagenet [19]	1.2 million	1,000	Object recognition	Medium	[20]–[23]	4
	Mini-Imagenet [24]	100,000	784	Object recognition	Low	[25]	1
	LSUN [26]	1.2 million	varies	Scene recognition	Low	[27], [28]	2
	MNIST [29]	70,000	784	Object recognition	High	[20], [30]–[48]	20
Text	IMDB [49]	50,000	varies	Sentiment analysis	Medium	[50]–[54]	5
	Newsgroup [55]	19,188	varies	Text classification	Low	[56]	1
	Reuters [57]	10,788	varies	Text classification	Low	[58], [59]	2
	SQuAD [60]	100,000	Varies	Question answering	Low	[61]–[63]	3
Tabular	Adult [64]	48,842	14	Income prediction	Low	[65]–[67]	3
	Breast Cancer [68]	286	9	Cancer diagnosis	Low	[69], [70]	2
	Diabetes [71]	768	8	Diabetes diagnosis	Low	[72], [73]	2
Time series	Epileptic Seizure [74]	11,500	178	Seizure prediction	Low	[32], [75]	2
	Activity Recognition [76]	10,299	561	Activity Classification	Low	[75], [77], [78]	3
Graph	OGB [79]	1.2 million	varies	Graph classification	Low	[80]	1
	Cora [81]	2,708	1,433	Graph classification	Low	[82], [83]	2
	Yelp Dataset [84]	8,282,442	Varies	Recommendation	Low	[85], [86]	2
Computer Vision	Fashion-MNIST [87]	70,000	784	Image classification	Medium	[88]–[91]	4
	Caltech-101 [92]	9,146	Varies	Object recognition	Low	[93]–[95]	3
	COCO [96]	330,000	Varies	Object detection	Medium	[97]–[101]	5
	YouTube Faces [102]	3,425	2,622	Face recognition	Medium	[103]–[107]	5
	EuroSAT [108]	27,000	13	Land use classification	Low	[10], [109]	2
Transaction	Purchase [110]	39,624	8	Purchase prediction	Medium	[111]–[115]	5
Sequence	Human Activity Recognition [116]	10,299	561	Activity recognition	Low	[117]	1
Recommendation	MovieLens [118]	100,000	varies	Movie recommendation	High	[119]–[125]	7



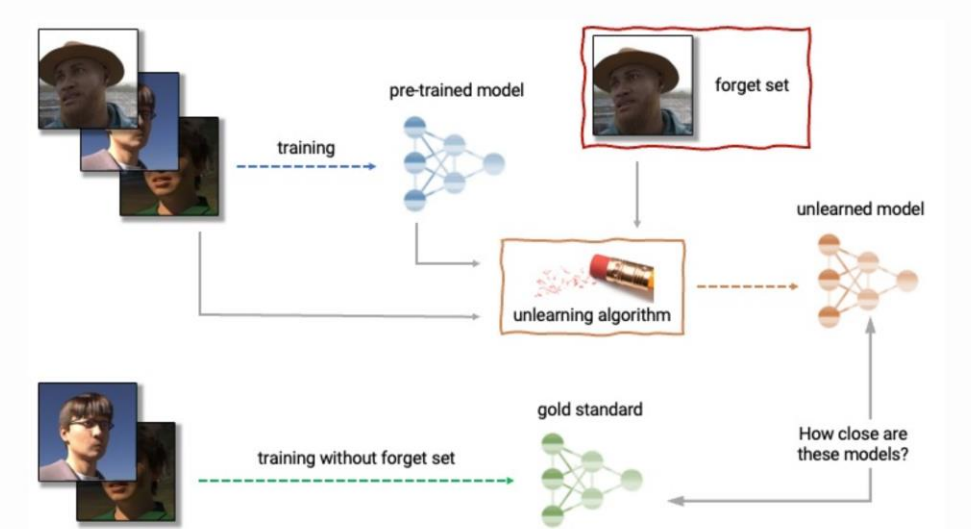
1. *Thanveer Shaik, Xiaohui Tao, Haoran Xie, Lin Li, Xiaofeng Zhu, and Qing Li Exploring the Landscape of Machine Unlearning: A Comprehensive Survey and Taxonomy https://arxiv.org/abs/2305.06360*

A first attempt: the Neurips challenge 2023

From Tutorial CVPR 2024 and *

The challenge was related to Data unlearning for privacy concerns (according with EU's General Data Protection Regulation) the individuals have the "right to be forgotten".

… with available and simulated data



Setup.

- An age classifier (ResNet-18) was trained using natural images of people's faces (from the CASIA-SURF dataset)
- A subset of users request their data to be deleted
- Goal: *efficient* unlearning of that user data

*https://research.google/blog/announcing-the-first-machine-unlearning-challenge/

The competition in Neurips 2023

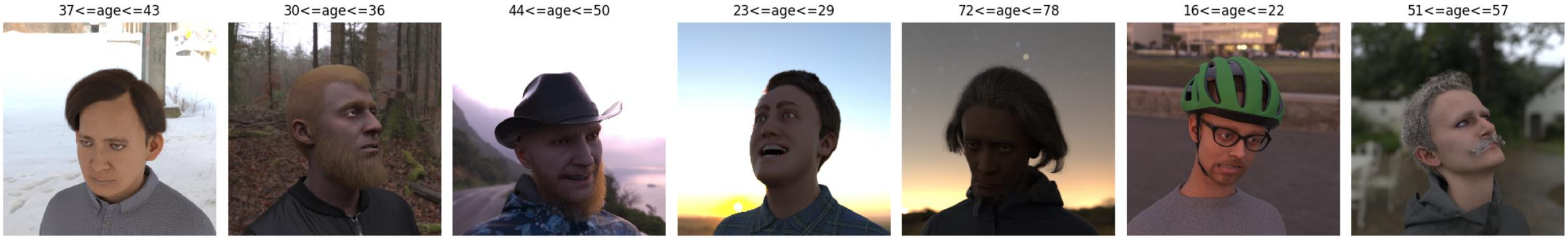
Scenario:

- An age predictor has been trained on face images.
- After training, a subset of the training images must be forgotten to protect the privacy or rights of the individuals concerned.
- The participants are asked to submit code that takes as input the trained predictor, the forget and retain sets, and outputs the weights of a predictor that has unlearned the designated forget set.
- Evaluation is based on both the strength of the forgetting algorithm (forget propriety) and model utility (retain propriety).

The competition in Neurips 2023

Evaluation:

- For the forgetting subset the tool is inspired by MIAs (Membership Inference Attacks), such as LiRA* developed in the privacy and security literature and their goal is to infer which examples were part of the training set.
- If unlearning is successful, the unlearned model contains no traces of the forgotten examples, causing MIAs to fail.
- Different the distribution of unlearned models (produced by a particular submitted unlearning algorithm) is compared to the distribution of models retrained from scratch. For an ideal unlearning algorithm, these two will be indistinguishable.



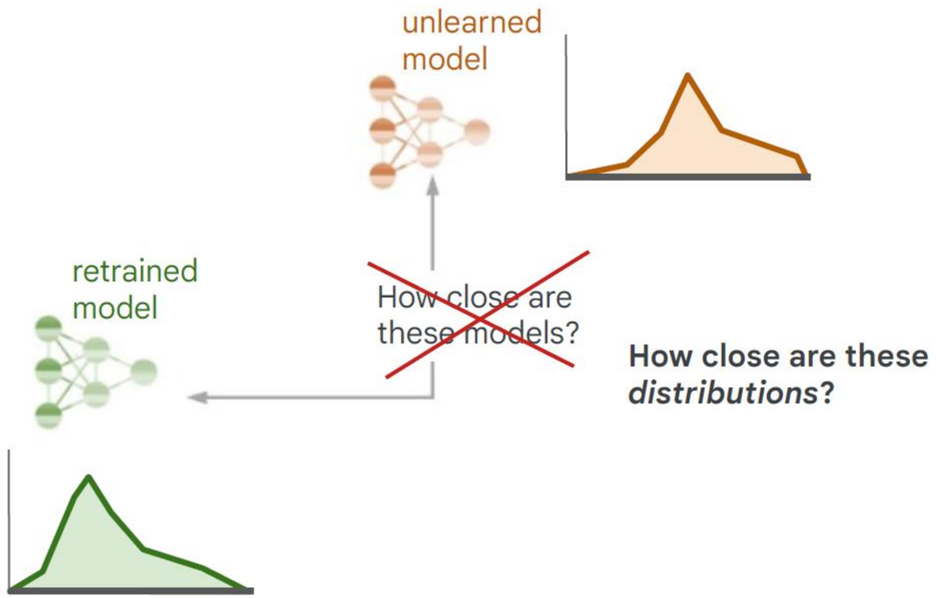
1. * https://arxiv.org/abs/2112.03570

Evaluation of Unlearning

How to Evaluate a method for Data Unlearning?

For Single or class of Data Unlearning, some solutions based on comparison in DISTRIBUTIONS (mainly for discriminative problems)

Theory comes from *, and simplified for the competition as **



1. *Remember what you want to forget: Algorithms for machine unlearning. Sekhari et al. 2021.
2. ** *https://arxiv.org/pdf/2406.09073*

In Neurips 2023

Available:

D a dataset of synthetic faces and

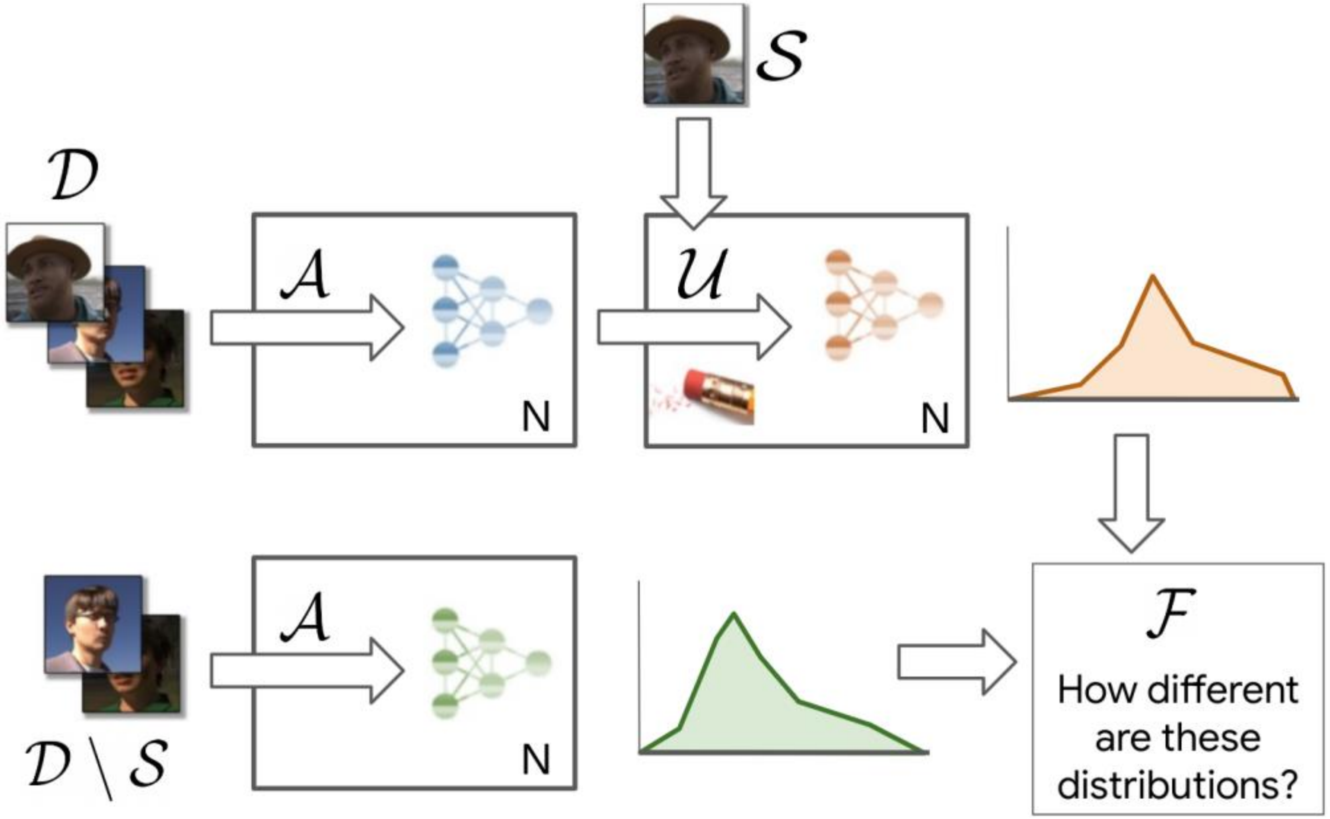
A the algorithm (Resnet)

S the forget set

The exact unlearning on D/S

The U unlearning algorithm is

The challenge to be evaluated

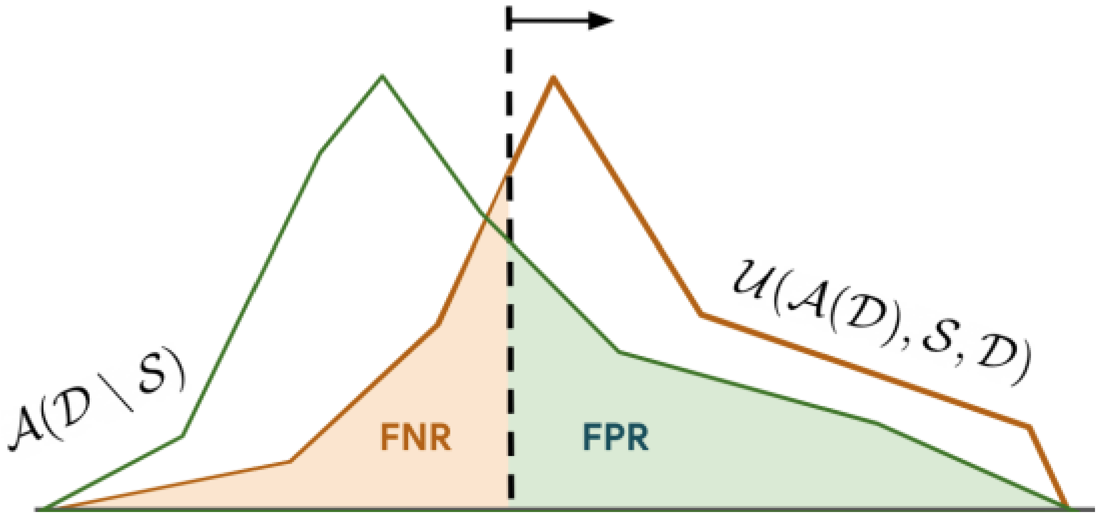


Comparison

The two distributions A(D/S) and U (A(D),S,D)

Are compared by sampling randomly some datapoints,

moving the thresholds and analyzing the distribution similarities



Evaluation

F is computed by example tests on the distributions

Forget (or untrain propriety)
For the data to forget

Utility (or retain propriety)
in Retain set and in the test set

Aggregation.

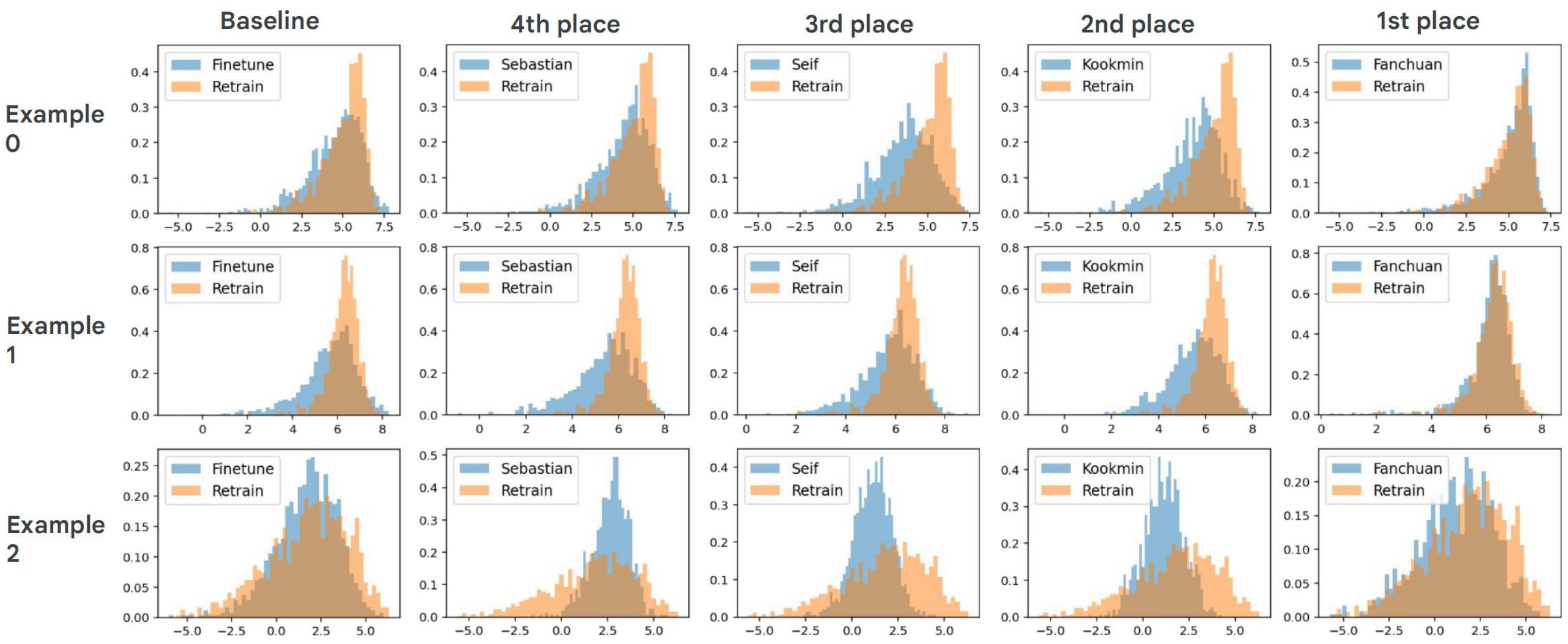
$$\text{Final score} = \mathcal{F} \times \overbrace{\frac{\mathrm{Acc}(\mathcal{D} \setminus \mathcal{S}, \mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{S}, \mathcal{D}))}{\mathrm{Acc}(\mathcal{D} \setminus \mathcal{S}, \mathcal{A}(\mathcal{D} \setminus \mathcal{S}))}}^{\text{Accuracy on retain set}} \times \overbrace{\frac{\mathrm{Acc}(\mathcal{D}_{test}, \mathcal{U}(\mathcal{A}(\mathcal{D}), \mathcal{S}, \mathcal{D}))}{\mathrm{Acc}(\mathcal{D}_{test}, \mathcal{A}(\mathcal{D} \setminus \mathcal{S}))}}^{\text{Accuracy on test set}}$$

This strategy penalizes unlearning if it yields poorer accuracy on retain or test compared to retraining-from-scratch

Are we making progress in unlearning? Findings from the first NeurIPS competition. Triantafillou et al. 2024.

In Neurips 2023

Top unlearning algorithms: unlearned versus retrained distributions of our test statistic



Are we making progress in unlearning? Findings from the first NeurIPS competition. Triantafillou et al. 2024.

Very similar results… probably an «easy» benchmark

Question?

Can we go a leap forward?

Does unlearning make some input data ininfluent...

or really «delete knowledge»?

Let’s go towards a more general Unlearning framework

Unlearning known and available data/classes

1. Unlearning, when I know what I would like to unlearn

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Unlearning known and available data/classes

Consider a Dataset with known labels $D = \{(x_i, y_i)\}_{i=1}^{N},$

a model by supervised training (as in the benchmark)

Unlearning aims at removing some or all data of a given class.

Julia

Brad

George

Unlearn George

<??>

<Julia>

<Brad>

Question: the scope of unlearning

Unlearning the data or the knowldge?

<??>

1. Unlearning Data (for privacy, copyright and legal issues)

Data to be unlearned
because of some issue

<George>

The system can answer
and generalizes well, but
Without the training data

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

If the problems are only the data

This is a problem of data not of the model: Evaluation metrics

- **Data Erasure Completeness**: how much data are rmoved? It compares the model's parameters before and after unlearning to quantify the extent of removing. Various distance or divergence measures as L2 or Kullback-Leibler (KL) divergence* .

- **Unlearning Time Efficiency:** the duration required for naive retraining of the model compared with the time it takes to perform the unlearning process **.
- **Resource Consumption**: the memory usage, power consumption, and storage costs incurred during the unlearning process.

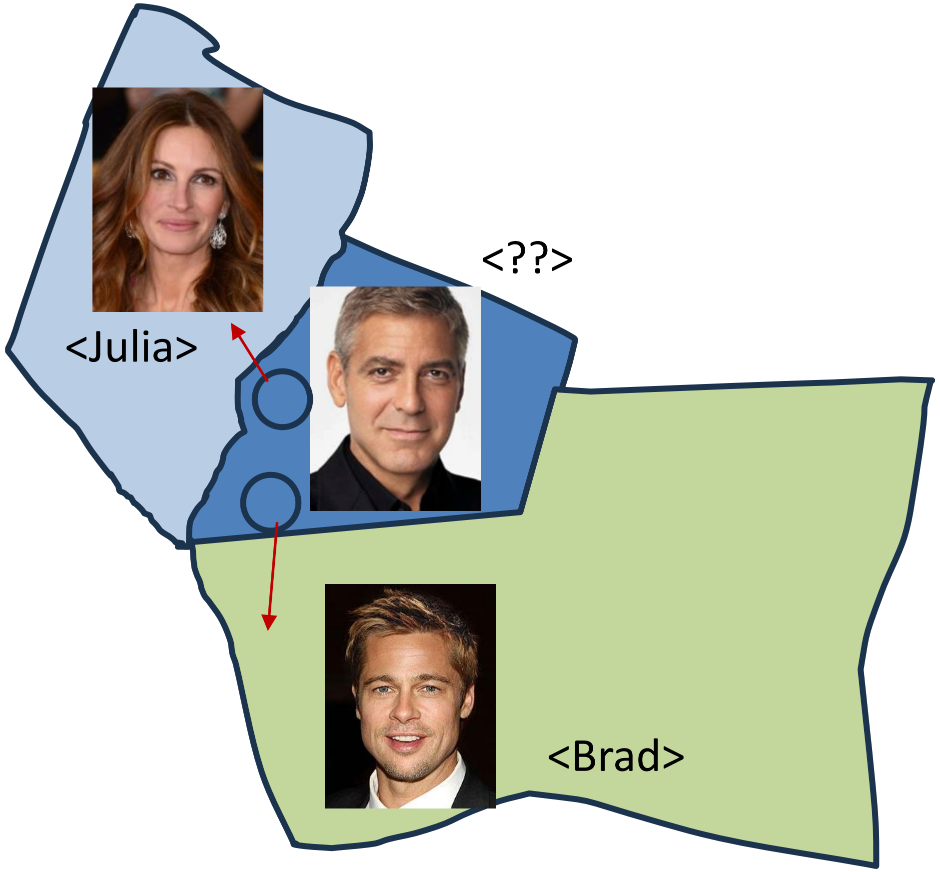
- **Privacy Preservation**: (as differential privacy) the Certified removal*** , i.e., that a model, after specific data removal, is indistinguishable from a model never trained on that data. This property implies that an adversary cannot extract information about the removed training data from the model, rendering membership inference attacks on the removed data unsuccessful. (ϵ,δ)-certified removal ***.

1. * A. Golatkar, A. Achille, and S. Soatto, "Eternal sunshine of the spotless net: Selective forgetting in deep networks," in *Proceedings of the IEEE/CVF CVPR* 2020, pp. 9304–9312.
2. ** S. Mercuri, R. Khraishi, R. Okhrati, D. Batra, C. Hamill, T. Ghasempour, and A. Nowlan, "An introduction to machine unlearning," *arXiv preprint arXiv:2209.00939*, 2022.
3. ***. Guo, T. Goldstein, A. Hannun, and L. Van Der Maaten, "Certified data removal from machine learning models," in *International Conference on Machine Learning*. PMLR, 2020

If the problem is the knwowledge

$$D = \{(x_i, y_i)\}_{i=1}^{N},$$

<??>



1. Destroy the label knowledge: split in different pdfs

<Brad>

<Julia>

2. Data unknow: shift the probability in the most likely

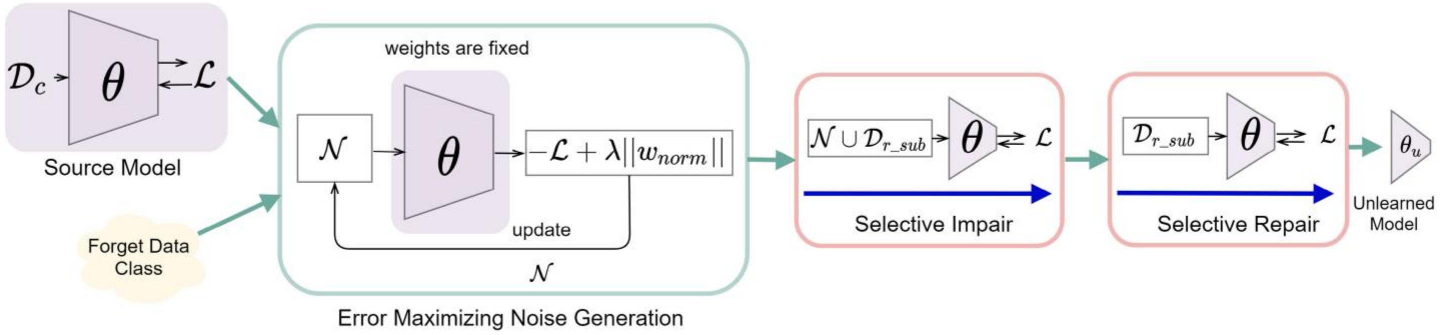
<Brad>

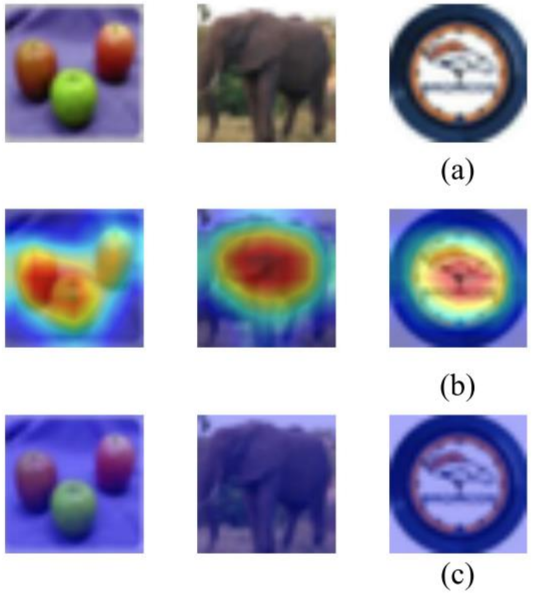
<Brad>

Unlearning in Vision

1. Making the model unlearned by destroying its performance on the subject of the unlearning, and splitting its probability among all the other classes[1]

- (e.g. learning a noise matrix to deteriorate the model's performances)





[1] *A. K. Tarun, M Kankanhally et al. «Fast Yet Effective Machine Unlearning». IEEE Trans NNLS 2022*
[

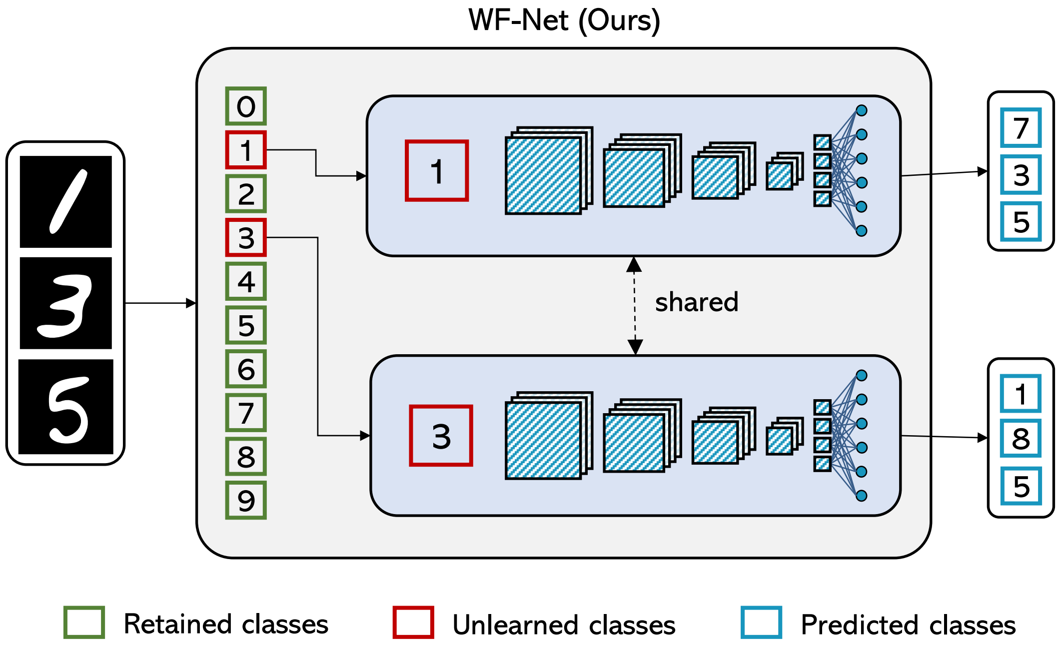
Selective Classes Unlearning

2. making the model unlearned by removing some classes and shifting their probabilities to the second most likely

- Given a dataset $\mathcal{D}$, a **forgetting set** $\mathcal{D}_f$ and a **retaining set** $\mathcal{D}_r = \mathcal{D} \backslash \mathcal{D}_f$
- The final goal is to unlearn the items in $\mathcal{D}_f$, while performing optimally on all the samples in $\mathcal{D}_r$
- The untraining protocol minimizes

$$\mathcal{L}(\mathcal{D};\theta) = \frac{1}{\mathbb{E}_{\mathbf{x},\mathbf{y}\in\mathcal{D}_f}\mathcal{L}_{\mathrm{CE}}(g_{\theta'}(\mathbf{x}),\mathbf{y};\theta)} + \lambda\mathbb{E}_{\mathbf{x},\mathbf{y}\in\mathcal{D}_r}\mathcal{L}_{\mathrm{CE}}(g_{\theta'}(\mathbf{x}),\mathbf{y};\theta).$$

to forget — to retrain

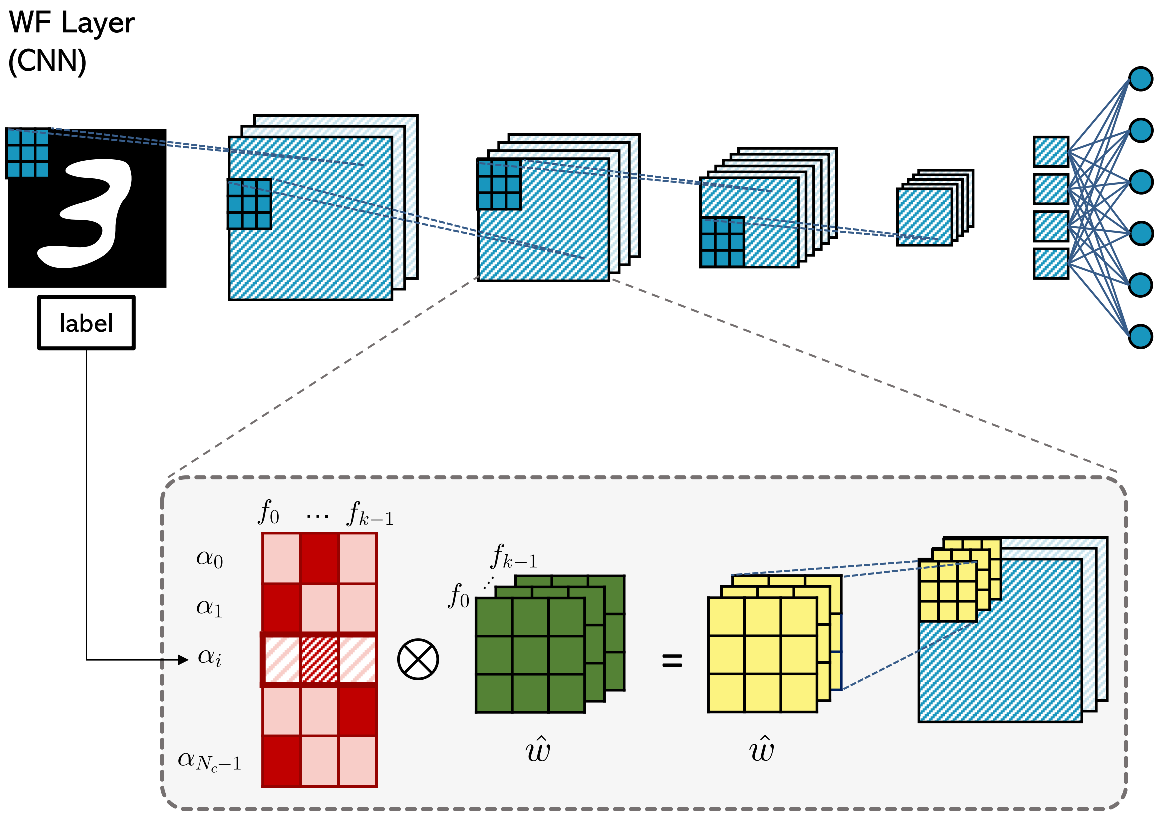


S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Multi-Class Explainable Unlearning for Image Classification via Weight Filtering», **IEEE Intelligent Systems 2024**

Weighting Filters

Weight Filtering Network: CNN and Vision Transformer based

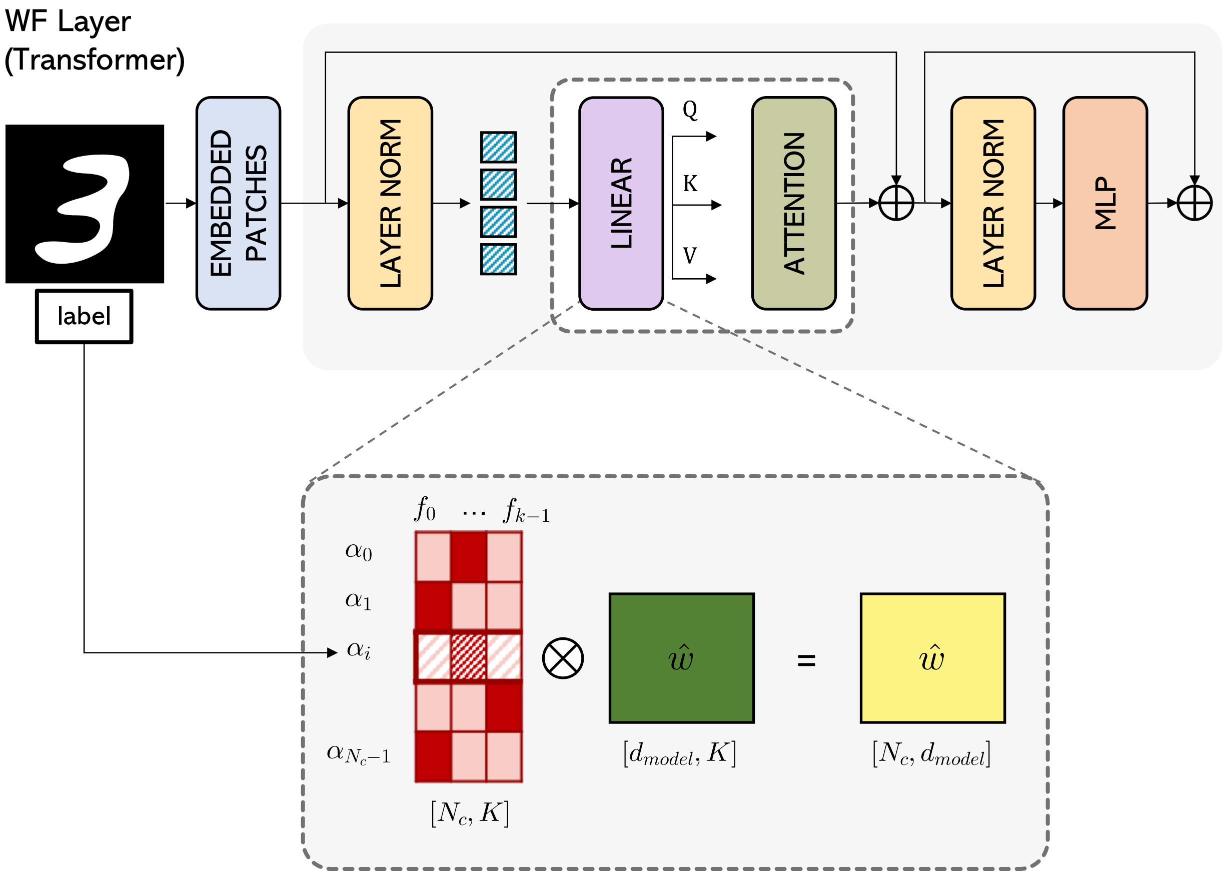
- On a CNN-based networks, the model learns to mask the convolutive filters that are most specific for the classes to unlearn
- The label(s) to be unlearned - as a negative prompt- select certain rows in a weighting-matrix
- The selected row multiplies the filters of each layer



S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Multi-Class Explainable Unlearning for Image Classification via Weight Filtering», **IEEE Intelligent Systems 2024**

Weight Filtering Network: CNN based

- On ViT-based networks, the model learns to mask the linear projections computing the Q,K,V triplets
- The labels are again fed into the network, to select certain rows in the weighting-matrix
- The selected row multiplies the weights of the linear projection matrix, in each layer



S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Multi-Class Explainable Unlearning for Image Classification via Weight Filtering», **IEEE Intelligent Systems 2024**

Unlearning Evaluation Metrics

- **Accuracy on retain and forget sets**
  - The rate of correct predictions for both retain and forget sets.
  - Accuracy on retain ≅ Accuracy original model
  - Accuracy on forget ↓

- **Activation distance and JS-Divergence**
  - They respectively measure the ℓ_2 distance and the Jensen-Shannon divergence[1] between the output probabilities of the unlearned model and the model re-trained without using samples of the forget class.

$$D_{JS}(p||q) = \frac{1}{2} D_{KL}(p||\frac{p+q}{2}) + \frac{1}{2} D_{KL}(q||\frac{p+q}{2})$$

- **Zero Retain Forgetting (ZRF) Score**
  - It estimates the randomness of the unlearned model by comparing it with a randomly initialized network.

$$\mathrm{ZRF} = 1 - \frac{1}{N_f} \sum_{i=0}^{N_f} JS(M(x_i), M^*(x_i))$$

1. *Lin Jianhua. «Divergence measures based on the Shannon entropy», IEEE Transactions on Information Theory*

Experimental results

Experimental results on MNIST, Cifar10 and ImageNet-1k datasets

		MNIST					CIFAR-10					ImageNet-1k		
		Acc_r[%]↑	Acc_f[%]↓	Act-Dist↓	JS-Div↓	ZRF[%]↑	Acc_r[%]↑	Acc_f[%]↓	Act-Dist↓	JS-Div↓	ZRF[%]↑	Acc_r[%]↑	Acc_f[%]↓	ZRF[%]↑
VGG-16	Original model	99.6	99.6	-	-	48.0	93.0	93.0	-	-	48.3	71.2	71.3	0.35
	Retrained model	99.4	0.0	-	-	48.7	89.9	0.0	-	-	50.1	-	-	-
	WF-Net	73.2	0.0	0.46	0.22	79.2	80.2	18.3	0.33	0.15	57.4	64.1	1.89	0.53
ResNet-18	Original model	99.6	99.6	-	-	47.0	93.9	94.0	-	-	48.0	70.5	70.3	0.35
	Retrained model	99.4	0.0	-	-	48.7	90.5	0.0	-	-	51.4	-	-	-
	WF-Net	94.0	9.68	0.26	0.12	63.1	79.7	9.25	0.35	0.15	63.9	64.4	1.47	0.40
ViT-T	Original model	98.9	98.9	-	-	47.2	78.0	78.0	-	-	49.8	75.6	75.5	0.34
	Retrained model	99.0	0.0	-	-	50.2	71.2	0.0	-	-	67.6	-	-	-
	WF-Net	93.5	0.0	0.23	0.10	47.4	73.5	0.0	0.34	0.12	59.8	68.0	2.51	0.44
ViT-S	Original model	99.0	98.9	-	-	47.1	85.2	85.2	-	-	53.6	82.3	82.2	0.34
	Retrained model	99.0	0.0	-	-	49.5	75.5	0.0	-	-	61.3	-	-	-
	WF-Net	94.0	0.0	0.21	0.09	48.0	74.7	0.0	0.35	0.12	59.7	69.2	9.46	0.45

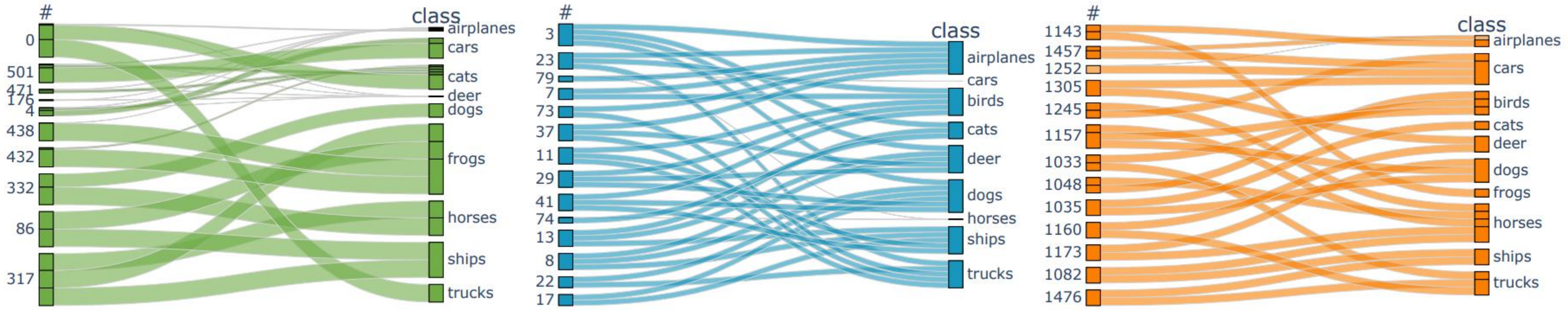
- In comparison with models explicitly retrained without each of the classes,
  - Same unlearning behavior, in terms of accuracy and activation distances
  - Less computational and storage cost, greater flexibility
- **Zero Retain Forgetting (ZRF) Score**

  It estimates the randomness of the unlearned model by comparing it with a randomly initialized network.

S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Multi-Class Explainable Unlearning for Image Classification via Weight Filtering», **IEEE Intelligent Systems 2024**

Experimental results

Bonus: explainability through Unlearning

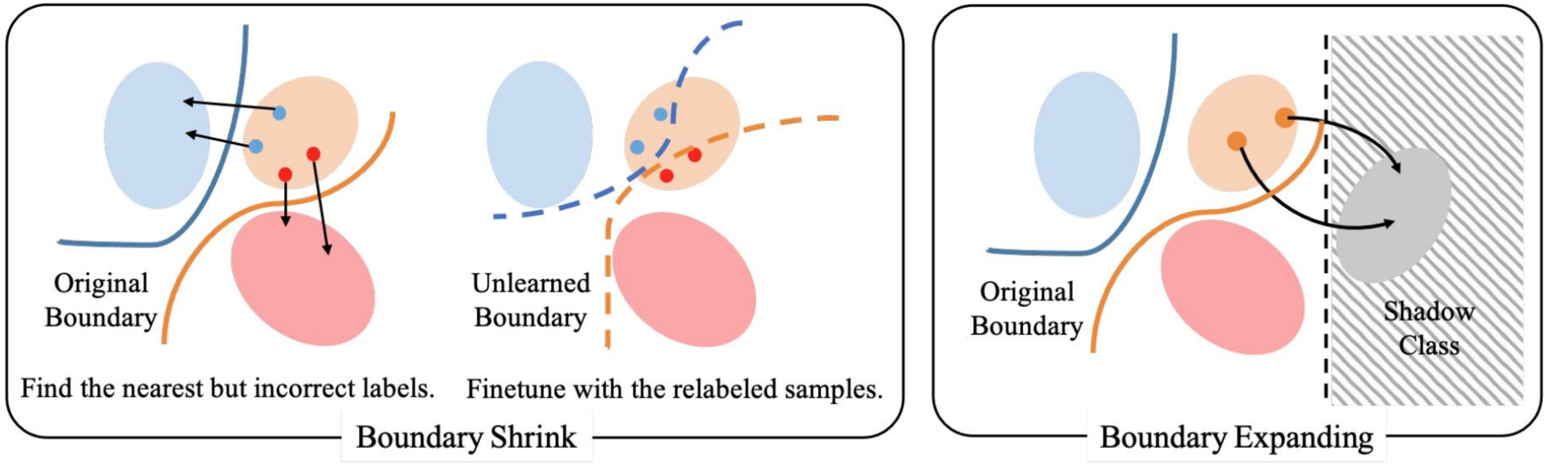


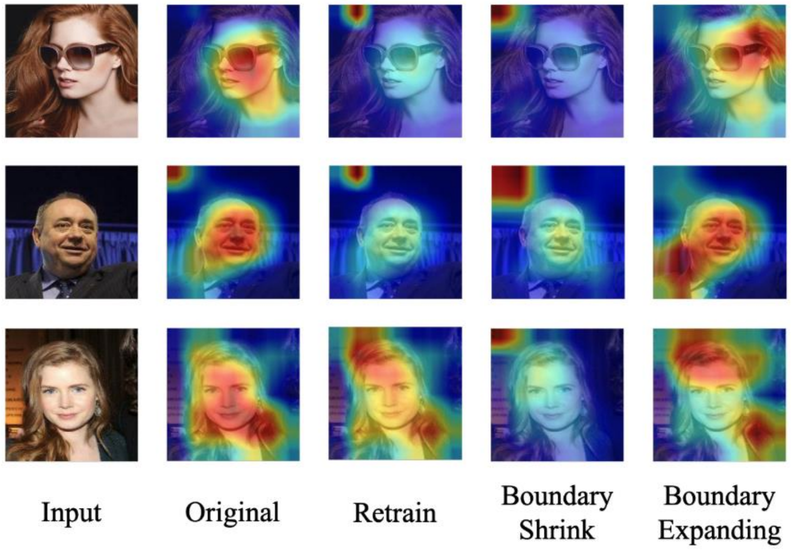
- **As we unlearn by switch inner network components on and off, our approach also recovers a representation of the classes which is explainable by-design**
  - We prove the importance of our weights showing how the confidence of a class drops when removing our *alpha* (**Deletion**) or how it increases when we reactivate it (**Insertion**)
  - We can visualize filters/attentive projection mappings at any layer.

S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Multi-Class Explainable Unlearning for Image Classification via Weight Filtering», **IEEE Intelligent Systems 2024**

BOUNDARY Unlearning

By shrinking or Expanding





Dataset	Metric	Original Model	Retrain	Finetune	Negative Gradient	Random Labels	Boundary Shrink	Boundary Expanding
CIFAR-10	Acc on $\mathcal{D}_r$	99.97	100.00	100.00	97.16	98.49	**99.24**	98.03
	Acc on $\mathcal{D}_f$	99.92	0.00	0.22	7.84	10.40	**5.94**	8.96

1. *Min Chen, Weizhuo Gao, Gaoyang Liu, Kai Peng, Chen Wang; Boundary Unlearning: Rapid Forgetting of Deep Networks via Shifting the Decision Boundary CVPR 2023*

Unlearning unavailable data/classes

2. Unlearnng when I know what I would like to unlearn
but I have not the original training data

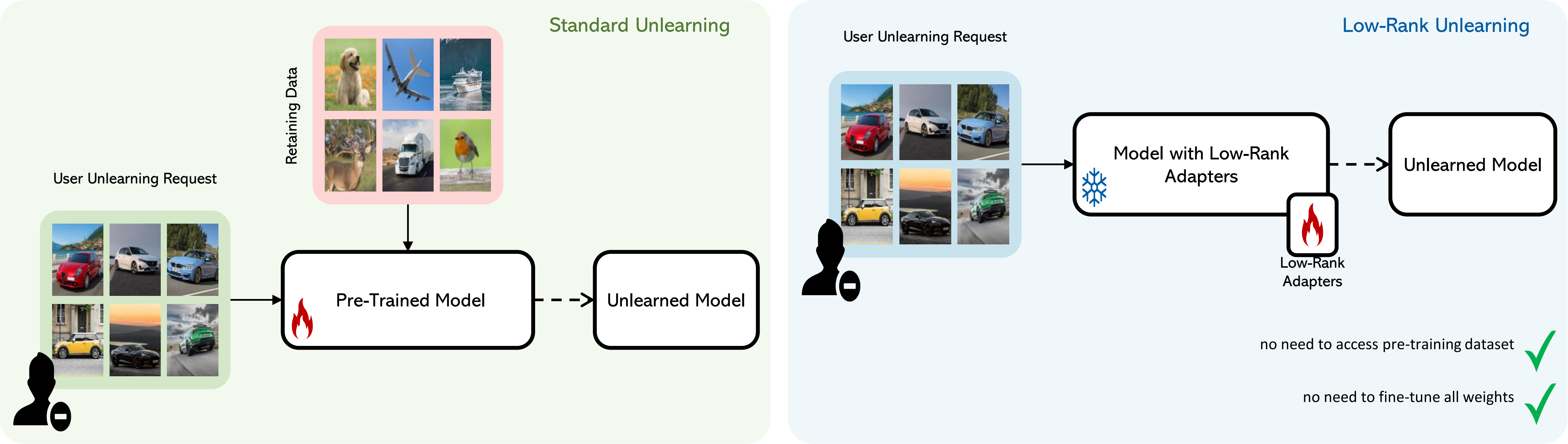
Unlearning classes without available training data

Please when I tell you “take my bag”, take only MY bags and forget please the school bag of my children

Unlearning Without Retaining Data

No retaining data available (as in pre-trained networks)

few-shot unlearning



S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **ICPR 2024**

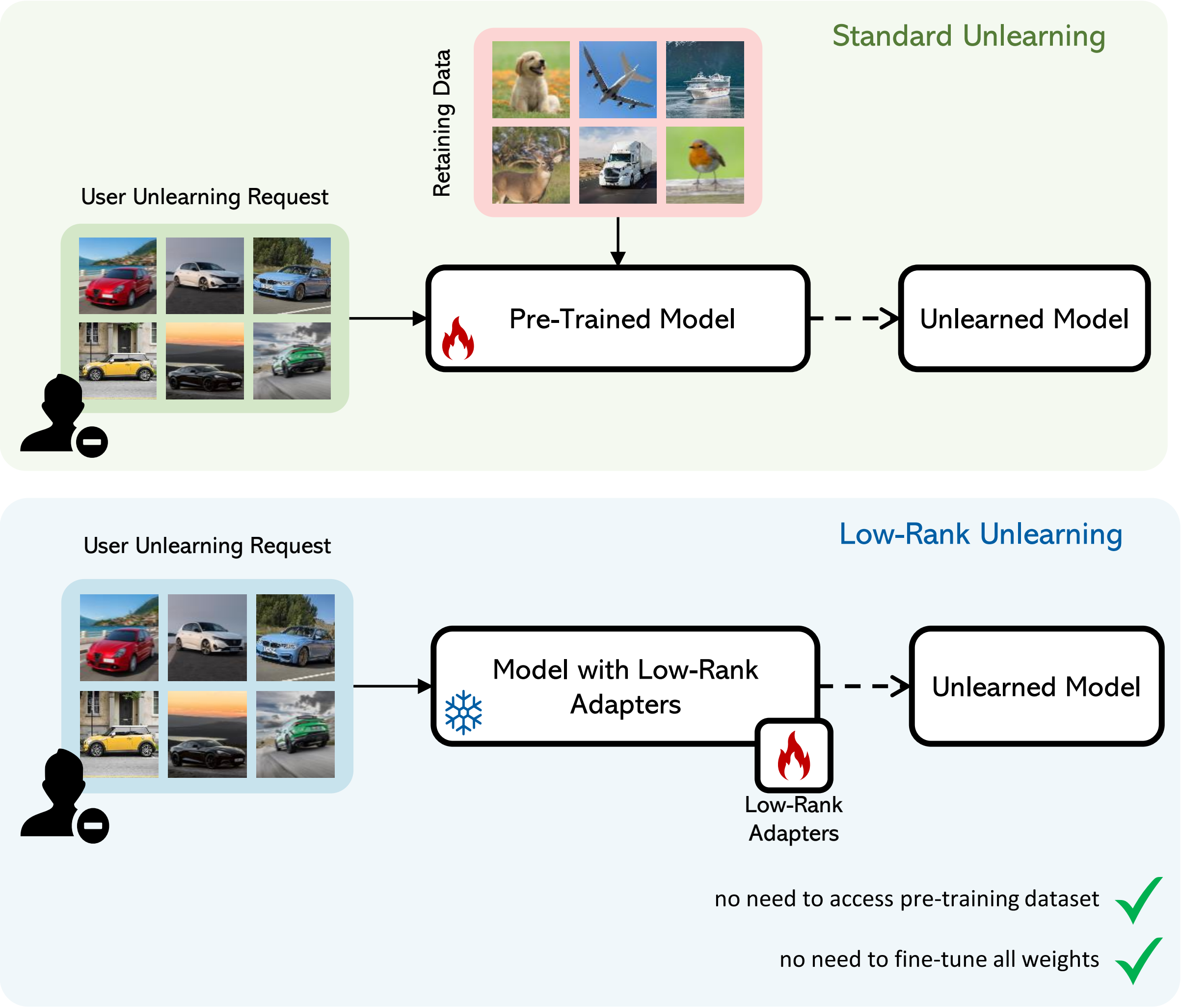
Unlearning Without Retaining Data

No retaining data available.. (as in pre-trained networks)

- Unlearn some labels (representing a class) without either accessing the retaining data or creating hand-crafted proxies

- Given some images of the classes to unlearn as a few-shot unlearning

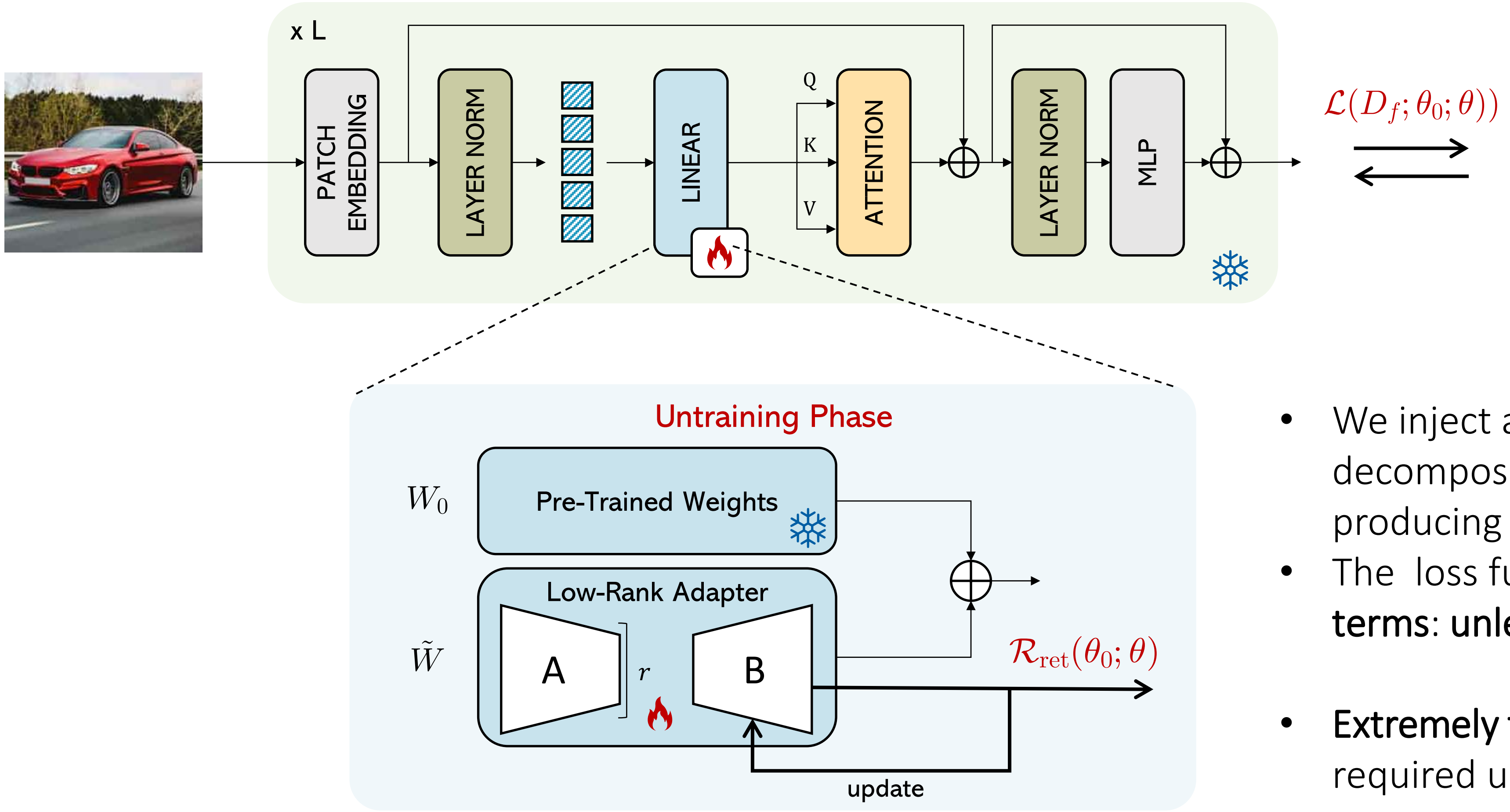
- Not the original ones:
- let's use random images, downloaded from the web of the class



S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **ICPR 2024**

The Model

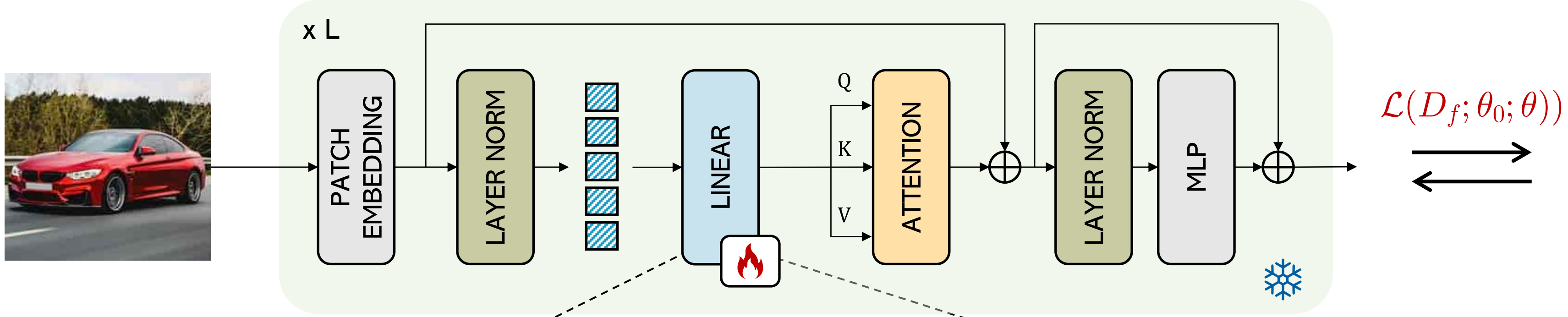
The low-rank based unlearning architecture



- We inject a **trainable low-rank** decomposition into the linear layer producing the QKV vectors
- The loss function is composed of **two terms**: **unlearning factor** and a **regularizer**
- **Extremely fast**, given the little number of required untraining samples

S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **ICPR 2024**

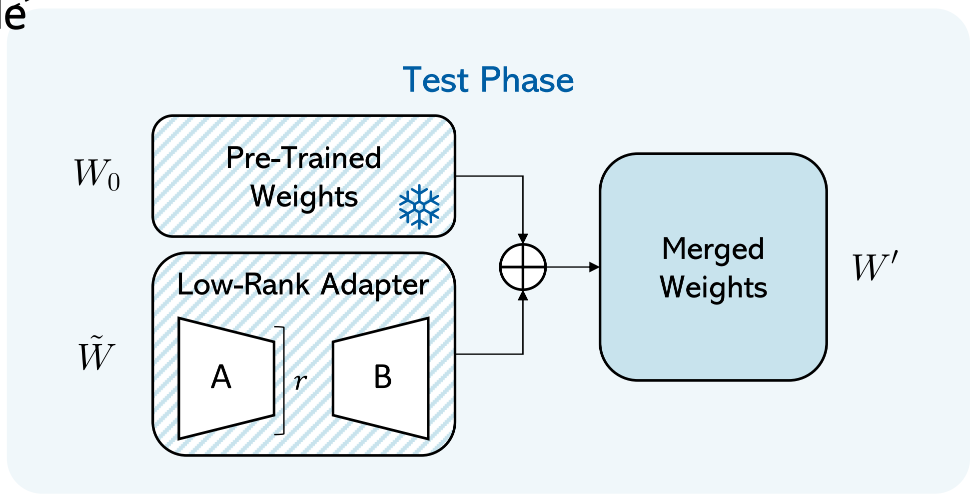
The low-rank based unlearning architecture in Test/inference Phase



$$\mathcal{L}(\mathcal{D}_f,\theta_0;\theta) = \frac{1}{\mathbb{E}_{\mathbf{x},\mathbf{y}\in\mathcal{D}_f}\mathcal{L}_{\mathrm{CE}}(g_{\theta'}(\mathbf{x}),\mathbf{y};\theta)} + \lambda\|\mathrm{vec}(B)\|_1.$$

During the evaluation, ***W*** can be made inaccessible by just collapsing the decomposition, back into a single parameter matrix.

$$W' \leftarrow W_0 + BA,\ \ f(x) = xW' + b.$$



S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **ICPR 2024**

Results

Experimental results on CIFAR-10 and CIFAR-20

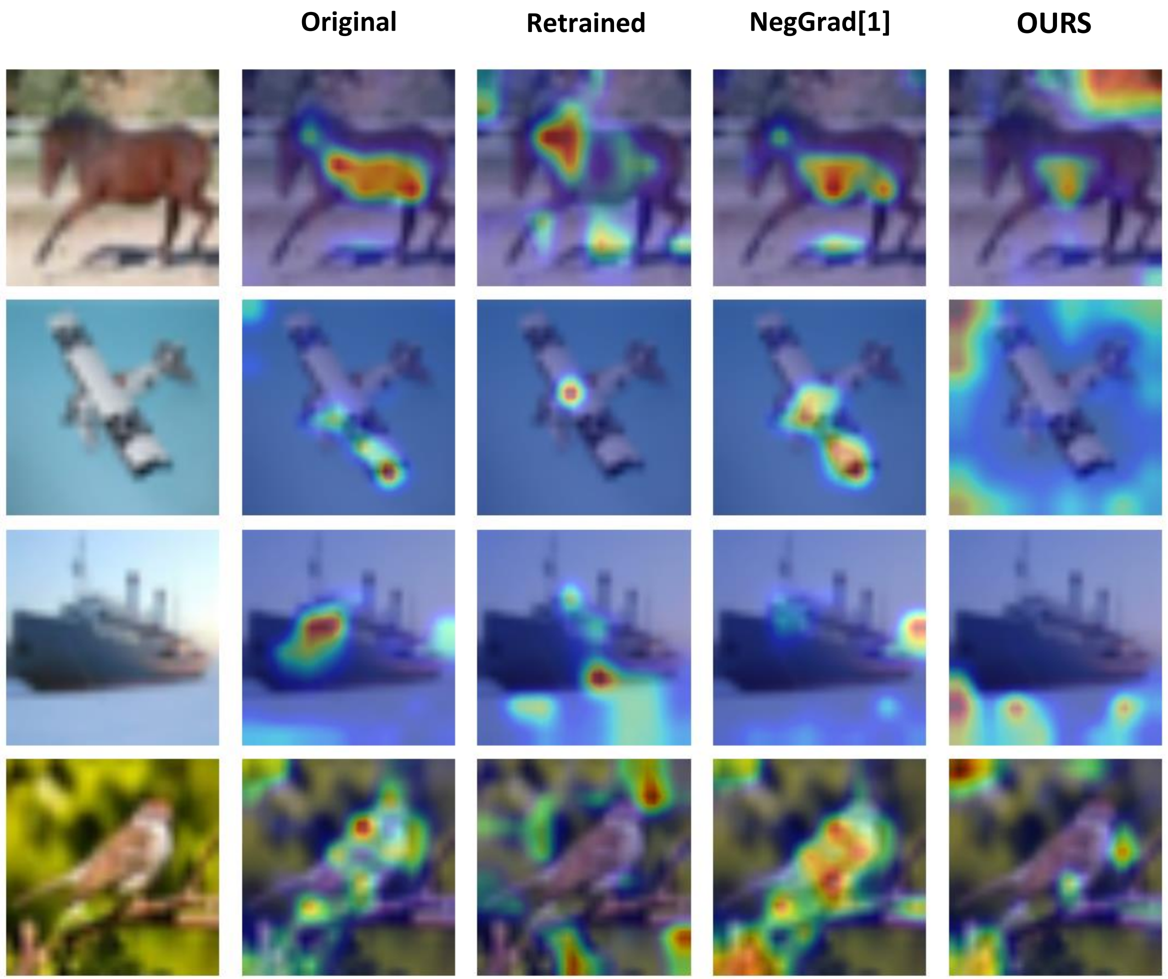
- **comparable results** to approaches using the **retaining data**
- The **LoRa** layer, **performs better** than other baselines, in the **same setting**

		$\mathcal{D}_r$	ViT-T		ViT-S		Swin-S	
			Acc_r [%] ↑	Acc_f [%] ↓	Acc_r [%] ↑	Acc_f [%] ↓	Acc_r [%] ↑	Acc_f [%] ↓
CIFAR-10	Original model	-	82.0	82.0	84.0	84.0	89.8	89.8
	Retrained model	✓	80.9	0.0	85.4	0.0	88.8	0.0
	Fine-tuned model	✓	80.2	7.9	81.3	3.0	85.0	2.3
	Random labels [17]	✓	83.0	0.0	85.1	0.0	88.9	0.0
	Negative gradient [14]	✓	84.4	0.0	85.8	0.0	88.9	0.0
	Negative gradient w/ L_1 regularization	✗	80.8	0.3	82.2	1.0	85.4	2.1
	Negative gradient w/ low-rank	✗	80.9	**0.1**	82.5	0.9	85.4	1.8
	Bounded loss w/ L_1 regularization	✗	81.2	**0.1**	82.3	**0.8**	85.5	1.4
	Bounded loss w/ low-rank (Ours)	✗	**81.9**	**0.1**	**83.5**	**0.8**	**86.0**	**0.8**
CIFAR-20	Original model	-	67.0	67.0	71.9	71.9	74.4	74.4
	Retrained model	✓	64.2	0.0	69.7	0.0	72.7	0.0
	Fine-tuned model	✓	64.5	8.2	67.2	8.6	68.3	4.6
	Random labels [17]	✓	66.2	0.0	70.8	0.0	73.2	0.0
	Negative gradient [14]	✓	67.6	0.0	71.4	0.0	72.2	0.0
	Negative gradient w/ L_1 regularization	✗	62.9	1.1	68.0	1.2	67.9	3.8
	Negative gradient w/ low-rank	✗	63.0	1.0	67.8	1.0	67.9	3.8
	Bounded loss w/ L_1 regularization	✗	63.1	1.2	67.9	0.8	68.0	3.7
	Bounded loss w/ low-rank (Ours)	✗	**63.5**	**0.9**	**68.2**	**0.8**	**68.2**	**3.4**

S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **ICPR 2024**

Results

Experimental results on CIFAR-10 and CIFAR-20



Grad-CAM technique to depict the most important areas of the image, **before and after** our low-rank unlearning[1]

The **unlearned model** does not focus on the **unlearned class**

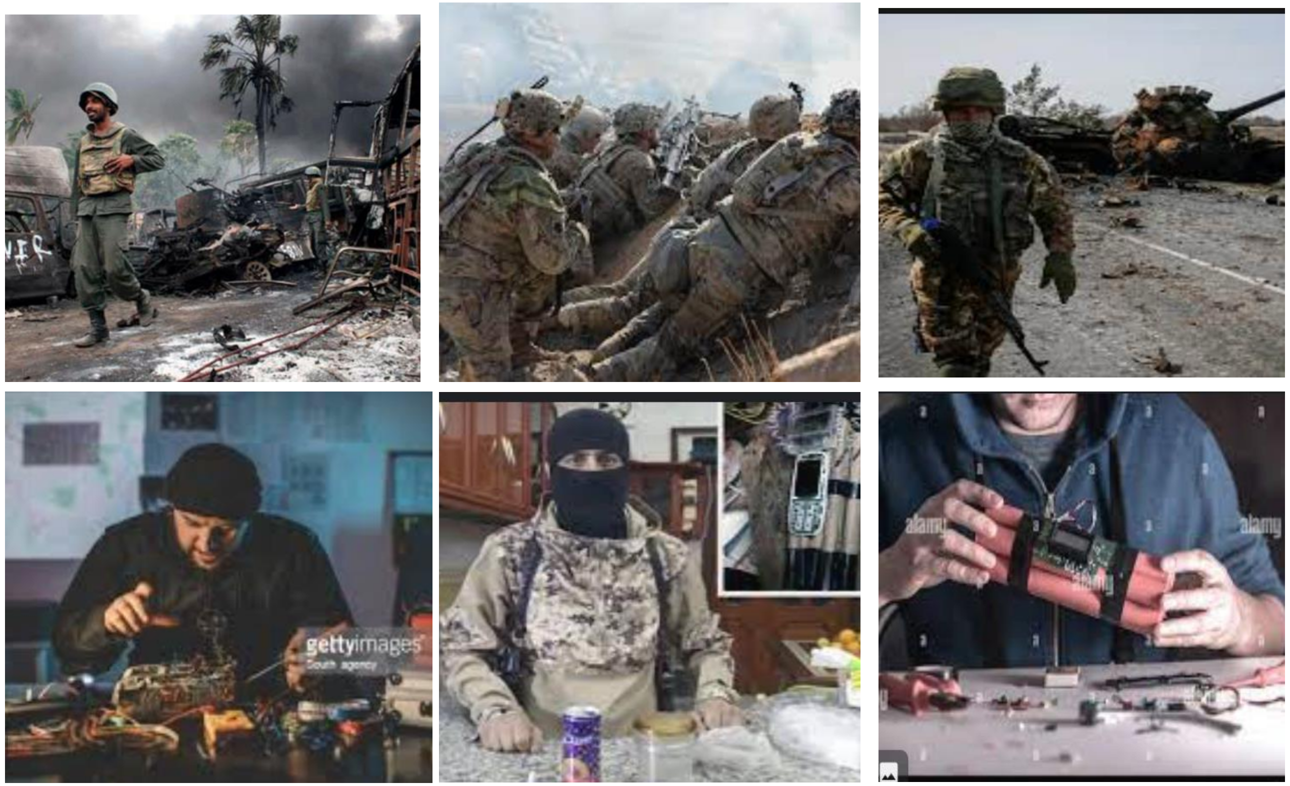
[1] A. Golatkar, A. Achille, and S. Soatto. 2020. «Eternal sunshine of the spotless net: Selective forgetting in deep networks». CVPR2020
S. Poppi, S. Sarto, M. Cornia, L. Baraldi, R. Cucchiara. «Low-Rank Class-wise Unlearning in Vision Transformers without Retaining Data», **ICPR 2024**

Unlearning unknown and emerging concepts

3. Unlearning when I suppose only know what I would like to unlearn

Can we unlearn ethically undesired classes?

Class "war", "bomb construction", "violence"



We cannot cancel the images of war, and violence in a pre-trained dataset

but could we try to unlearn the class concept of "war", "bomb constructing"?

Can we unlearn emerging concepts such as sexual harassment, nudity, pedophile, that shouldn't be learned, but they are?

$$D \text{ as } D = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$$

$\tilde{y}_i \in Y$ Is an unknown concept during training (e.g. in unsupervised pretraining) but is emergent in the latent space

Unearning multimodal concepts in the latent space

1. CLIP-based **multimodal space** $D = \{\breve{X}_i\}_{i=1}^{N} = \{(I_i, t_i)\}_{i=1}^{N}$

2. Pretrained model, data are unknown (**zero-shot unlearning**) $\dot{D}$ is unknown

3. **Emergent concepts in the latent space** (but unknown in pretraining) $\tilde{y}_i$ is unknown

4. Unlearned models should optimize **RETAIN and FORGET proprieties**

<A soldier is walking in a field>

<A troop of soldiers is observing the enemy>

war

<A soldier is running away from a possible explosion>

Unlearn everything about “war”

S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. «Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models», **ECCV 2024**

A practical Example in the ELIAS project

LSFW concepts

“hate, harassment, violence, suffering, humiliation, harm, suicide, sexual, nudity, bodily fluids, blood, obscene gestures, illegal activity, drug use, theft, vandalism, weapons, child abuse, brutality, cruelty”[1]

We want unlearn LSFW concepts in the CLIP pretrained latent space

ELiAS

1. Schramowski, Patrick, et al. "Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models." *Proceedings of the IEEE/CVF CVPR* 2023.

Unlearning multimodal concepts in the latent space

1. CLIP-based multimodal space
2. Pretrained model, data are unknown
3. Emergent concepts in the latent space (but known in pretraining)
4. Unlearned models should optimize RETAIN and FORGET proprieties

1. Data are multimodal and pretrained in a contrastive way, given by a pair of image and text. Thus the Dataset is given as before defined as $D = \{X_i\}_{i=1}^N = \{(I_i, t_i)\}_{i=1}^N$
2. The model is pretrained and the original training dataset is unknown. Thus D is unknown.
3. Data is pretrained in an unsupervised manner in a foundation model paradigm, and the concepts or classes to be removed are not surely associated with input data, i.e., are unknown. Thus the concept related with each input x_i , i.e., $\tilde{y}_i$ is unknown.
4. We want to unlearn one or some concepts associated with data by modifying the model and keeping the *RETRAIN* and *FORGET* properties as much as possible.

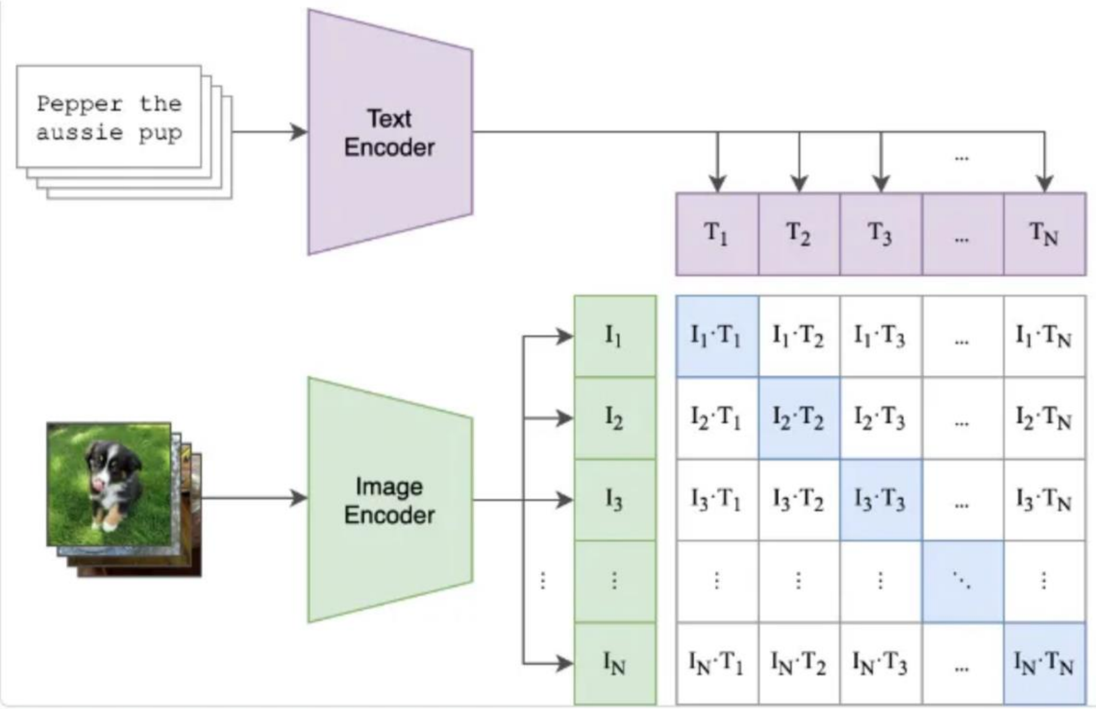
A Multimodal pretrained space

CLIP space $D = \{x_i\}_{i=1}^{N} = \{(I_i, t_i)\}_{i=1}^{N},$

CLIP (OpenAI **C**ontrastive **L**earning **I**n **P**retraining 2021)is designed to predict which N × N potential (image, text) pairings within the batch are actual matches.
Contrastive learning loss*: CLIP establishes a multi-modal embedding space through the joint training of an image encoder and text encoder.

The CLIP loss aims to maximize the cosine similarity between the image and text embeddings for the N genuine pairs in the batch while minimizing the cosine similarity for the N^2 – N incorrect pairings.

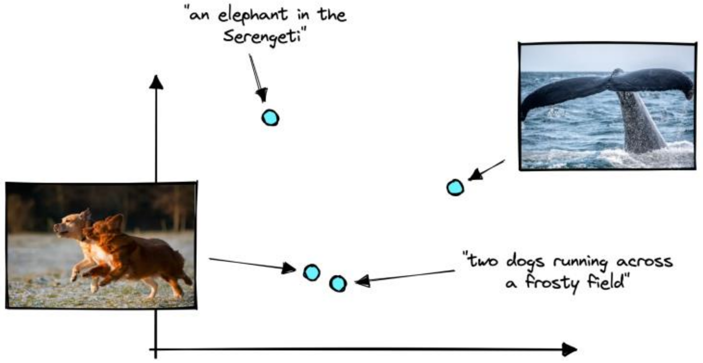
The optimization process involves using a symmetric cross-entropy loss function that operates on these similarity scores.



Architecture of CLIP model (taken from the original paper)

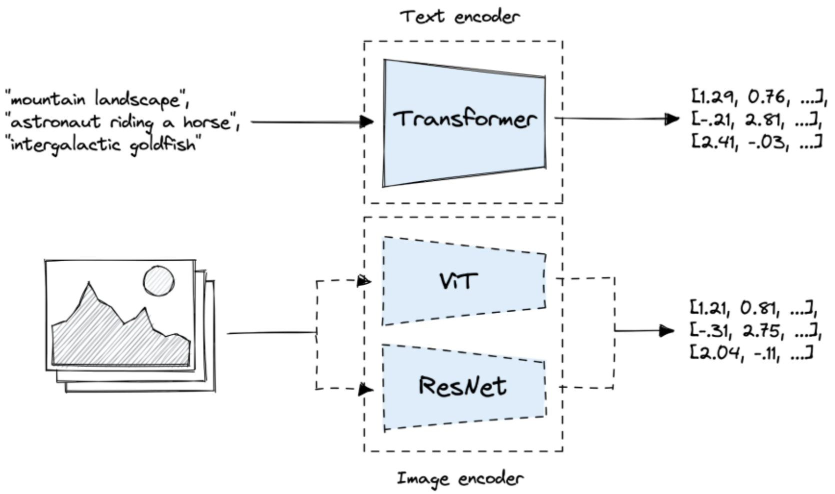
1. * https://arxiv.org/pdf/1807.03748

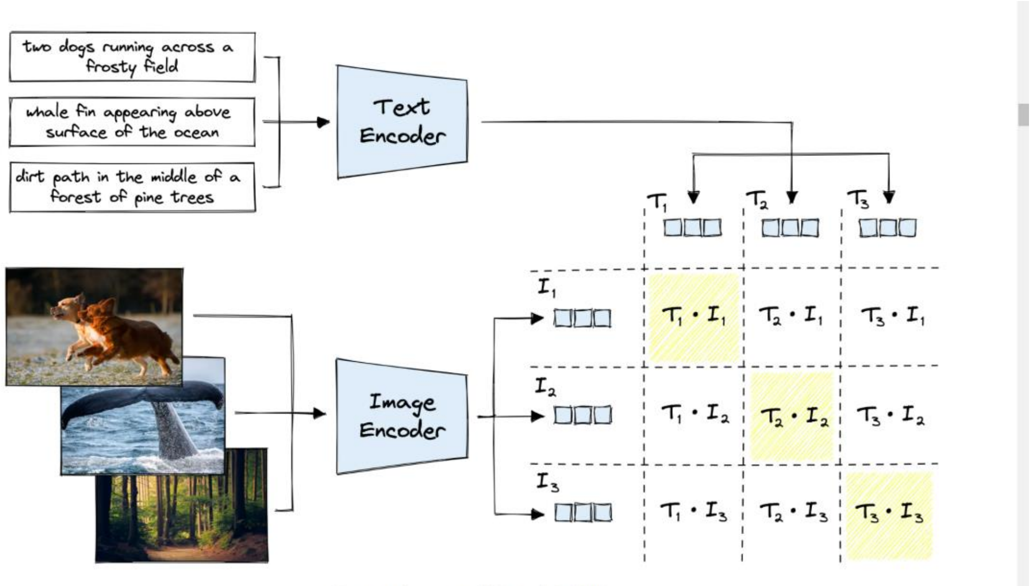
CLIP embedded space*



$$cossim(A, B) = \frac{A \cdot B}{||A|| * ||B||} = \frac{\sum_i^n A_i B_i}{\sqrt{\sum_i^n A_i^2}\sqrt{\sum_i^n B_i^2}}$$

$$dotproduct(A, B) = A \cdot B = \sum_{i=0}^{n-1} A_i B_i$$



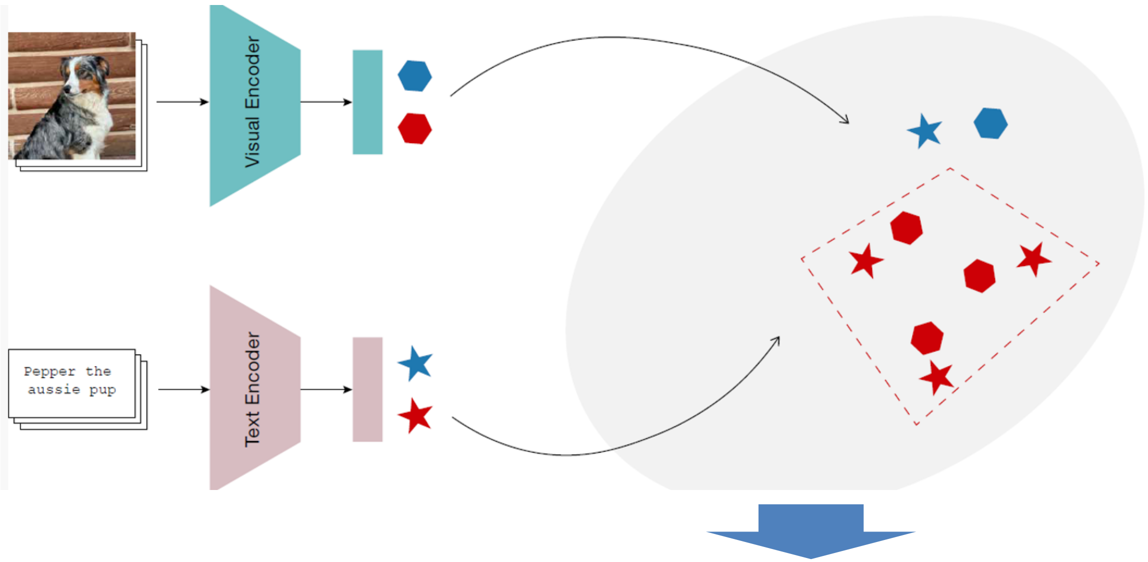


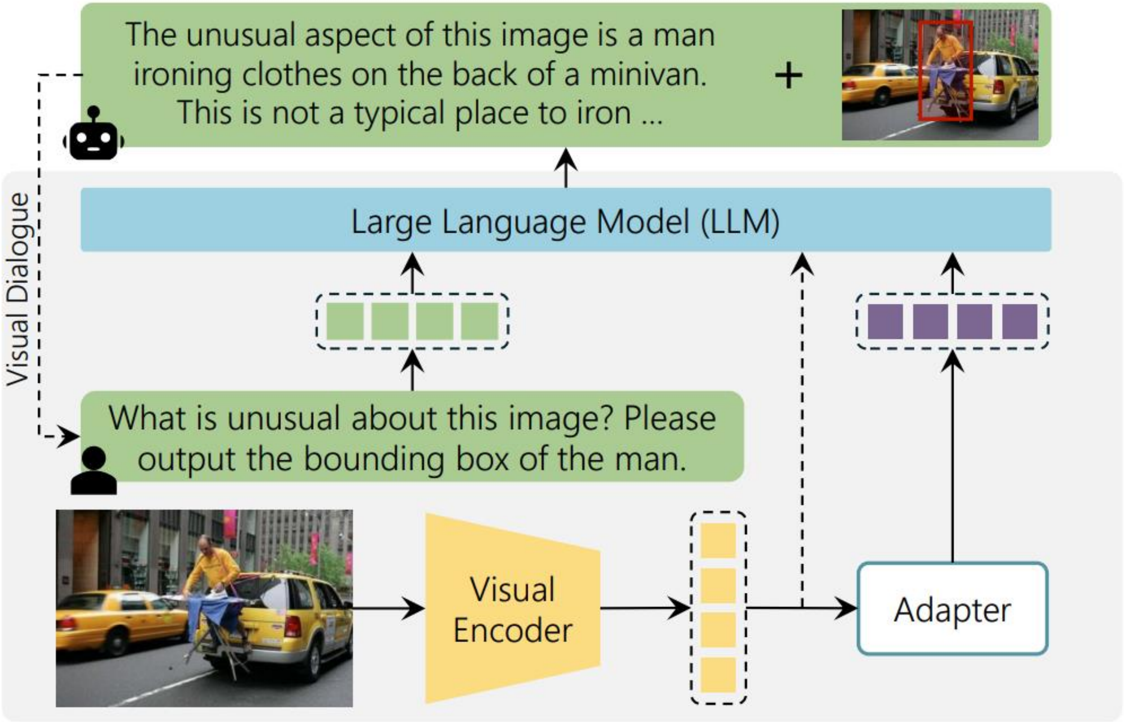
Contrastive pretraining with CLIP.

1. *https://www.pinecone.io/learn/series/image-search/clip/*

CLIP and Multimodal LLMs

CLIP Embedded space and Multimodal LLMs*





- Classification
- Search and Retrieval
- Generation (e.g. DMs, Captioning…)
- Prompting Multimodal LLMs

*Davide Caffagni, Federico Cocchi, Luca Barsellotti, Nicholas Moratelli, Sara Sarto, Lorenzo Baraldi, Lorenzo Baraldi, Marcella Cornia, Rita Cucchiara

1. The Revolution of Multimodal Large Language Models: A Survey, ACL 2024

CLIP

Pretrained in 400 million images from the web

quantitative measures of toxic content in the CLIP dataset are not available,

Concerns with CLIP's Training Data: [an answer by Chat-GPT]

1.Toxic Content:

1. Since the internet contains content that can be offensive, harmful, or biased, the dataset might include images and text that reflect these issues. This means that the model might inadvertently learn and reinforce these toxic concepts.

2.Biases:

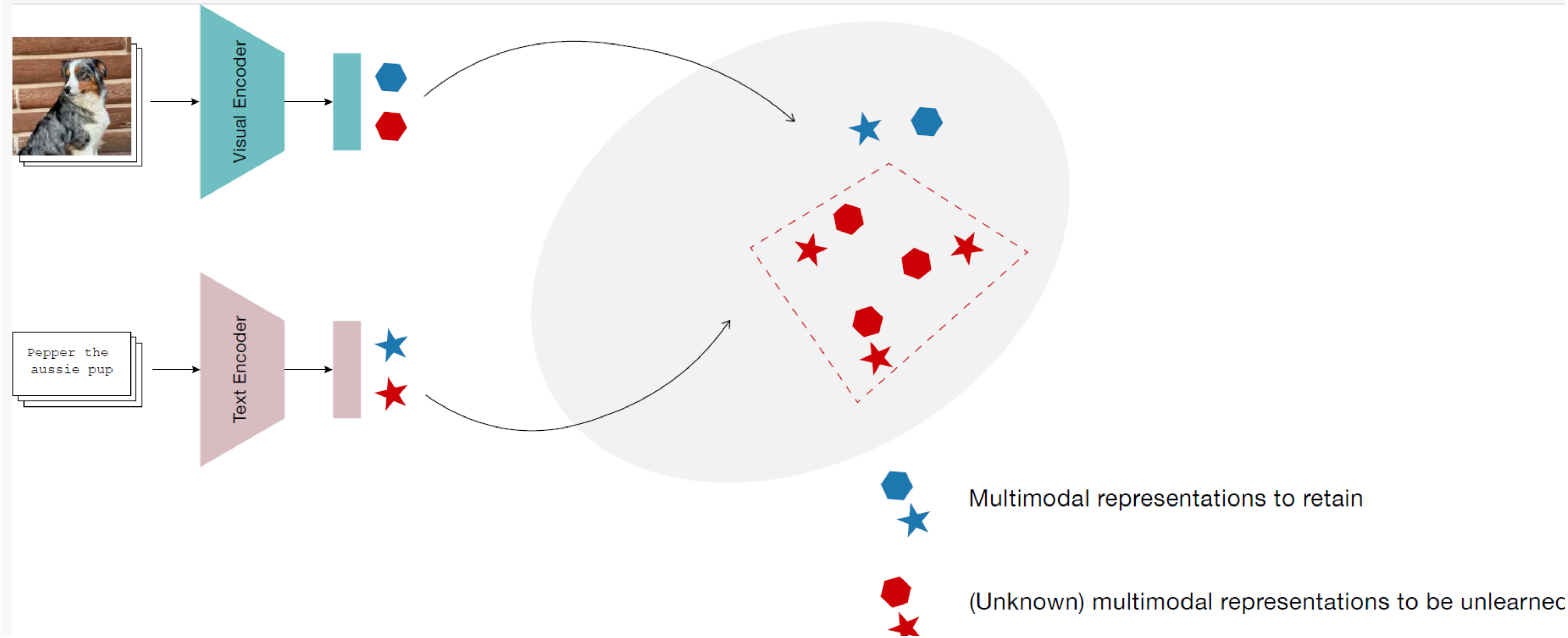
1. The model might inherit biases present in the data. For instance, if the dataset overrepresents certain groups or ideas while underrepresenting others, CLIP could produce biased outputs that reflect these imbalances.

3.Uncurated Nature:

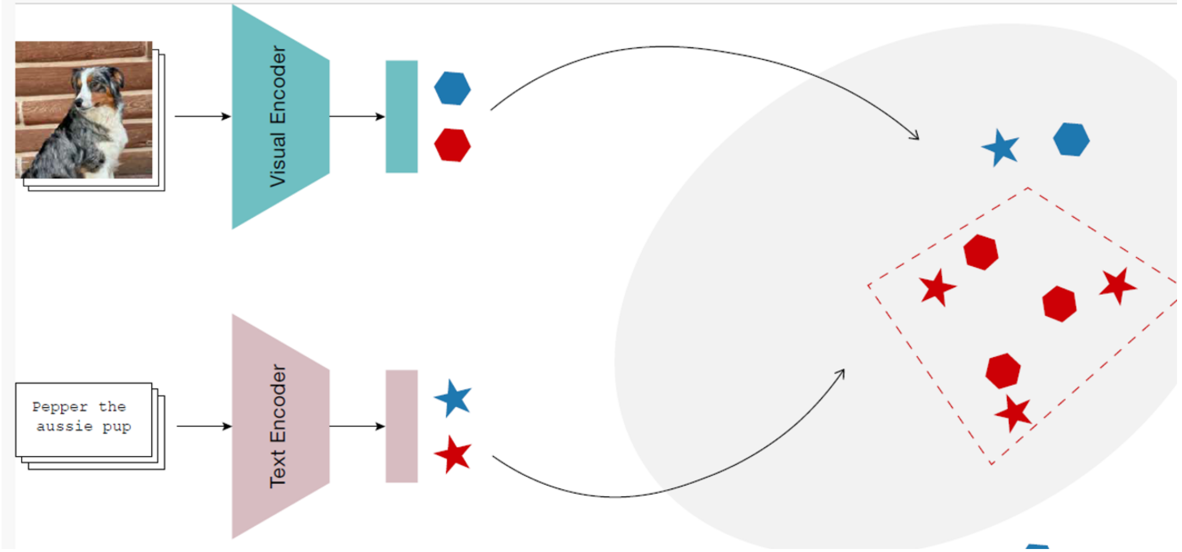
1. The lack of manual curation means there wasn't a filtering process to remove or mitigate toxic content before training, leading to potential risks in the model's behavior.

1. *Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). "Learning Transferable Visual Models From Natural Language Supervision." arXiv preprint arXiv:2103.00020.*

Unlearning unsafe (toxic) concepts



The pretrained multimodal latent space



We want unlearn LSFW concepts in the CLIP pretrained latent space

$$D = \{x_i\}_{i=1}^{N} = \{(I_i, t_i)\}_{i=1}^{N}$$

$D = \{(x_i, \tilde{y}_i)\}_{i=1}^{N}$, where $\tilde{y}_i \in Y$ represents an unknown concept or class associated with the input data.

Multimodal Representations To be Retain

(Unknown) Multimodal Representations To be Unlearned

We want unlearn one or a subset of concepts

$$Y = \{\tilde{y}_0, \tilde{y}_1, ..., \tilde{y}_n\}$$

Detecting concepts?

Unwanted (unsafe-toxic concepts)

1. We could try to detect them and block or filter out

We have noted that $\tilde{y}_i$ are unknown concepts. To make them emerge, we should construct or fine-tune a classifier in the latent space such that $y_i' = f_{c\mathbf{w}}(F_{\mathbf{w}}(x_i))$, to be a proxy in the space $Y = \{\tilde{y}_0, \tilde{y}_1, ..., \tilde{y}_n\}$. However, there is no guarantee that the computed y_i' equals $\tilde{y}_i$. That is, there is no a priori guarantee that the concept detector could be reliable

A) We do not trust on classifiers and filters

B) filters can be removed

→ We want to UNLEARN such a concepts

Unlearning for downstream task

If we can unlearn some concepts in the CLIP embedded space…

The **new SAFE-CLIP** *space can be used

a. **in multimodal retrieval:** the embedded feature vector can be used in the other modality to retrieva data

from text to image retrieval

from image to text retrievaò

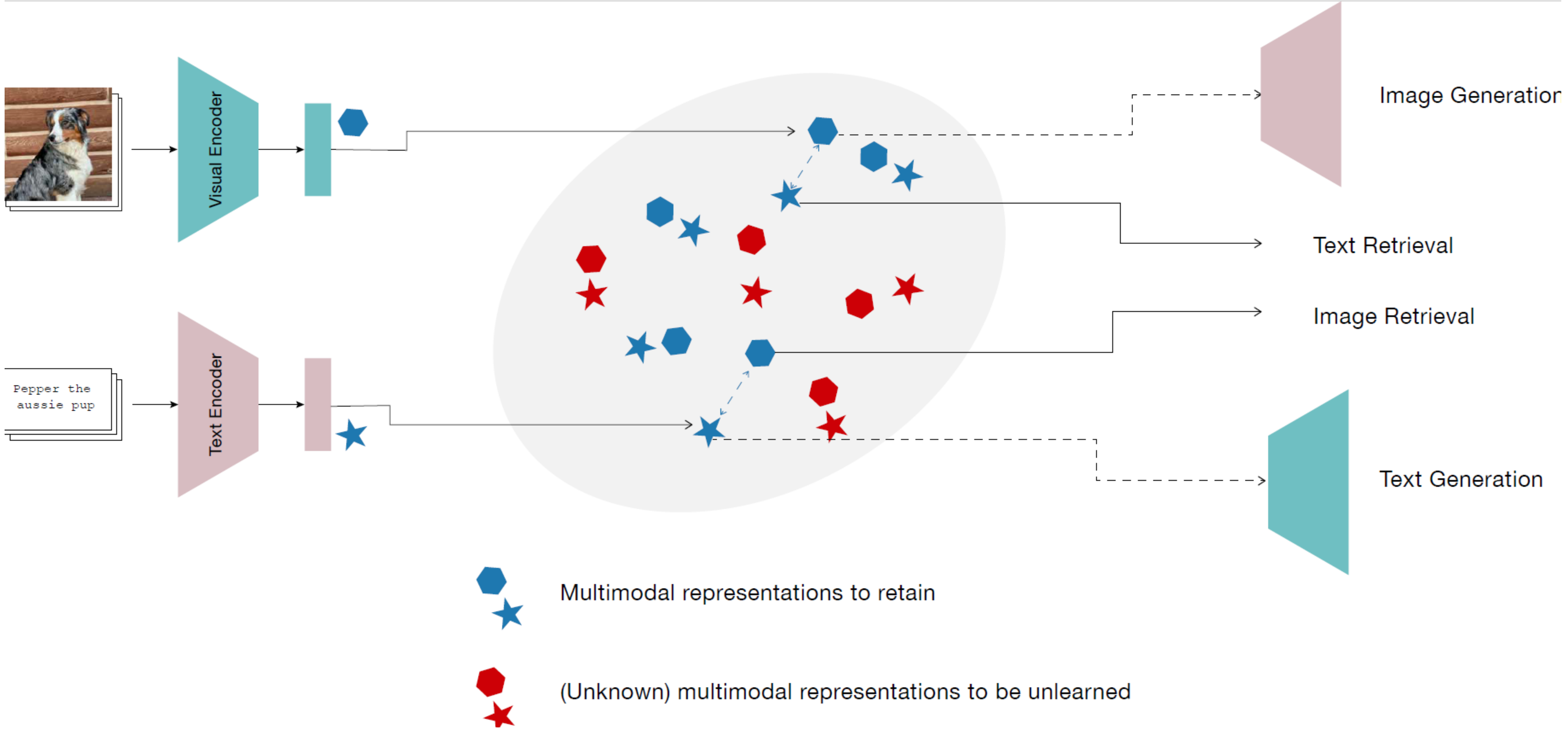
b. **in multimodal generation:** the embedded feature vector can be used as a prompt for

from image to text generation (image captioning)

from text to image generation (prompt for a diffusion model)

1. T.Poppi, S.Sarto, M.Cornia, L.Baraldi , R.Cucchiara afe-CLIP: Removing NSFW Concepts from Vision-and-Language Models. ECCV 2024

Multimodal retrieval and generation

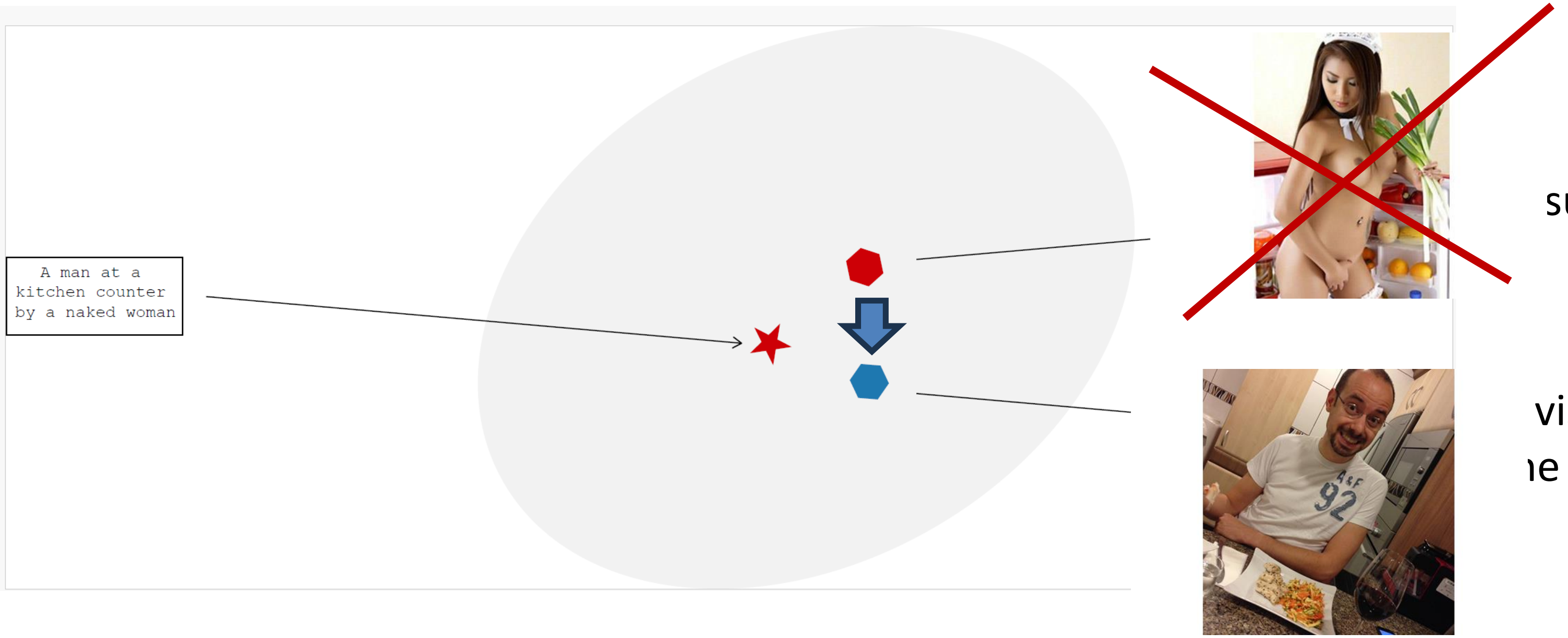


Unlearning by re-learning

Re-learning: aka moving concepts in the embedded space

we cannot avoid user to ask toxic prompt or queries

we can redirect the knowledge into safe concepts



sual feature vector of unsafe concept

visual feature vector of similar (form
ıe semantic point of view) safe concept

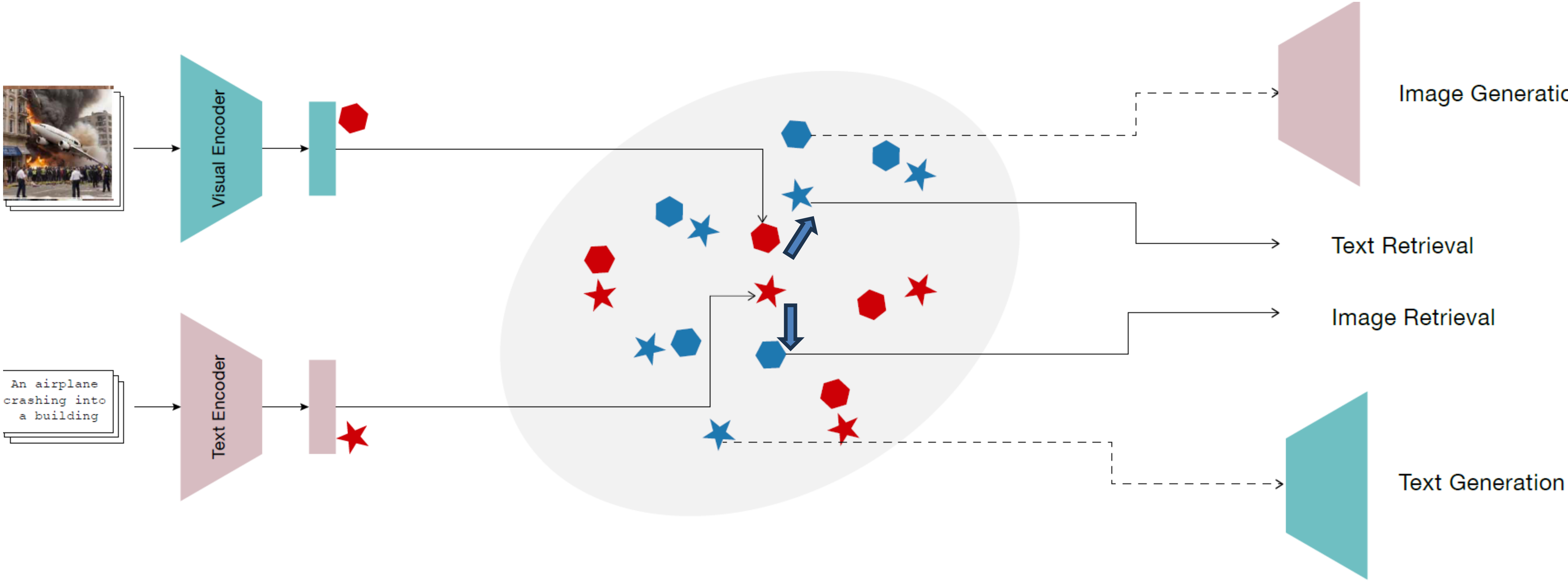
Unlearning by relearn examples

Unlearning multimodal pairs of image/text unwanted concepts (e.g. toxic concepts)
Move them in the latent space to nearest neighbor retain concepts
Unlearn concepts→ relearn concepts
Give examples of the transformation from what to unlearn to what to be retrained

A young girl being kidnapped and sold into sex slavery, [...]

A man is at a kitchen counter by a naked woman.

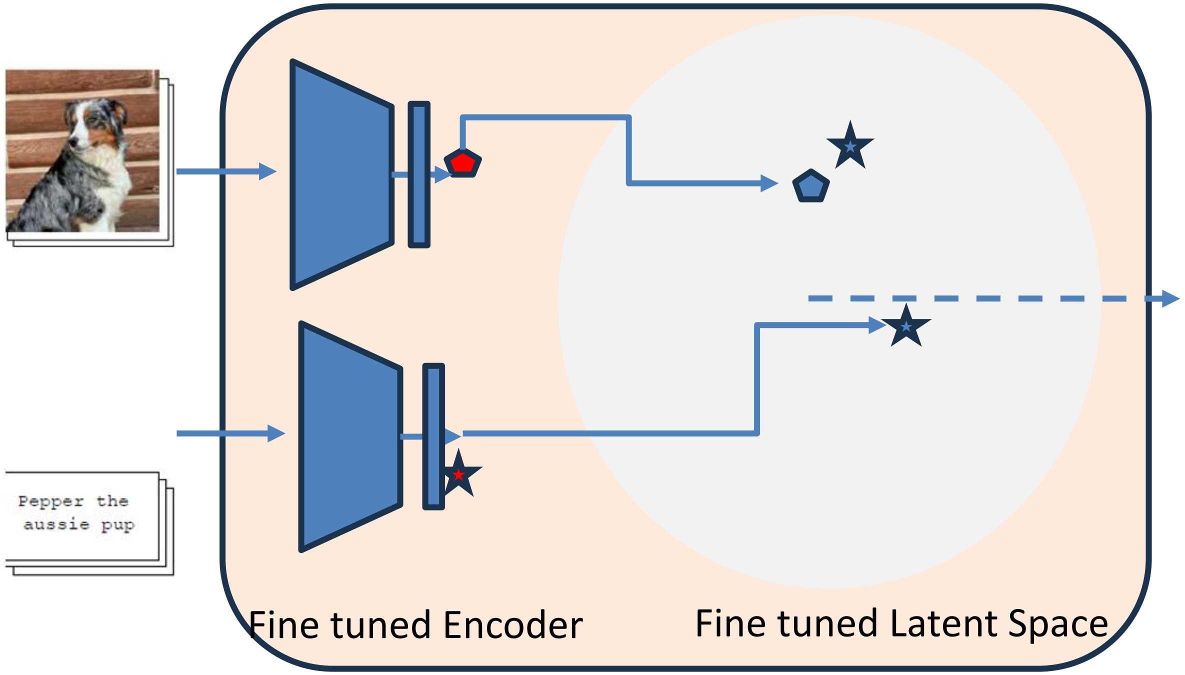
The desired relearning concepts



Multimodal representations to retain

(Unknown) multimodal representations to be unlearned

Unlearning by fine tuning



UNLEARNING BY FINETUNING
The model is partially retrained with retain/unlearn
Data pairs (i.e, quadruple
In multimodal domain)

The concepts to unlearn are catastrophically forgotten

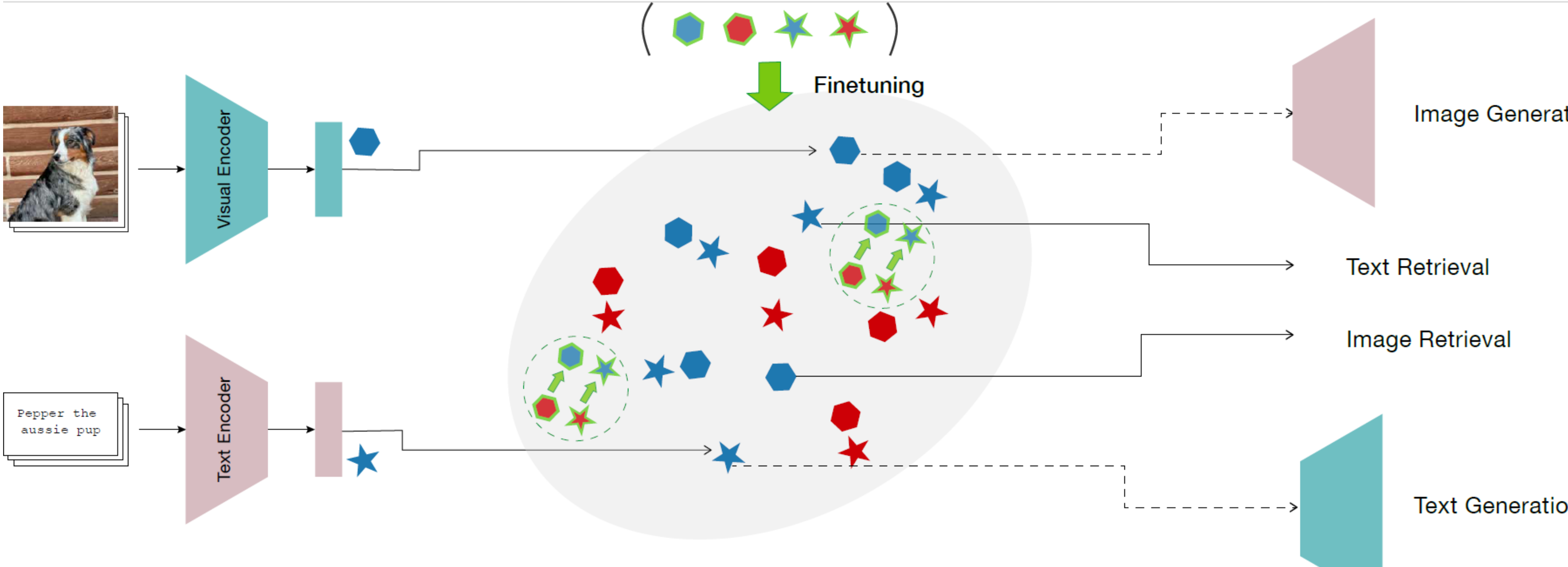
No control on unlearn concept
Is available anymore

Fintuning with a quadruple

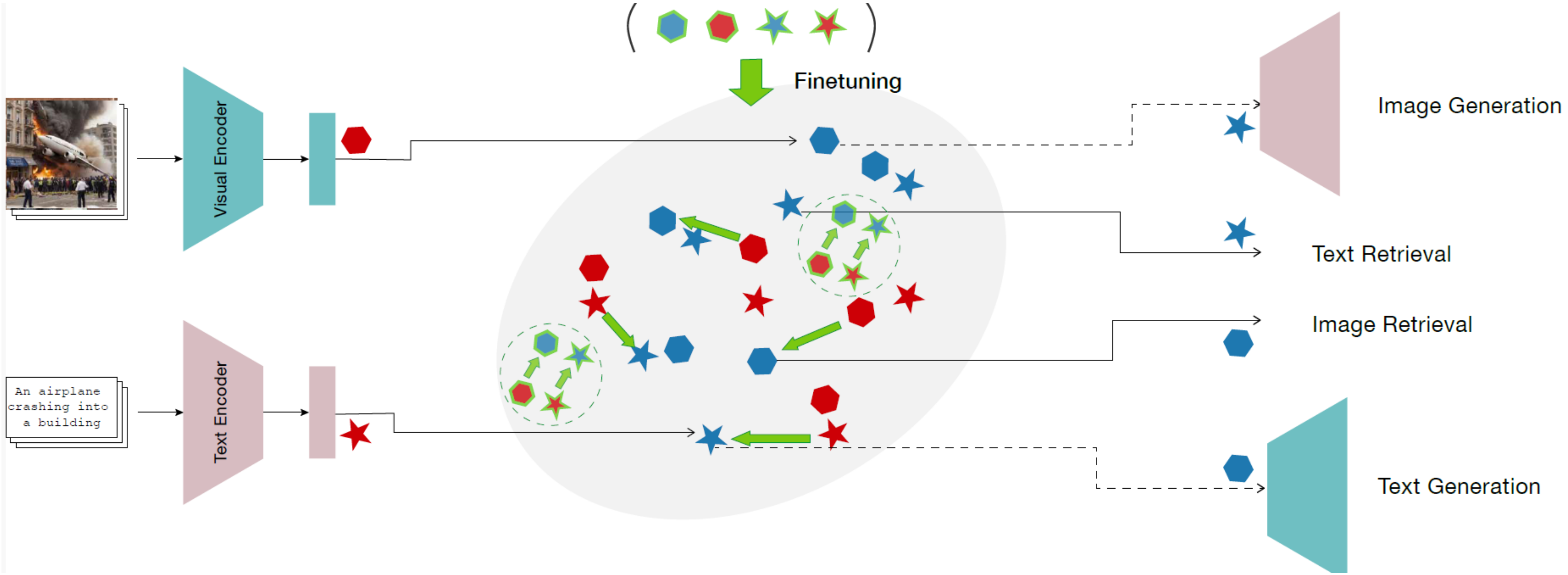
Fintuning

Unlearning by finetuning means of finding some toxic-safe concepts to be paired

And train the network for rediection



Finetuning moves all the embedded space in the diretcion of the most similar safe concept, but it must keep as much as possible invariate the rest



FORGET PROPRIETY: undesired (unsafe) connections are unlearned and redirected

RETAIN PROPRIETY: good (safe) connections are maintained

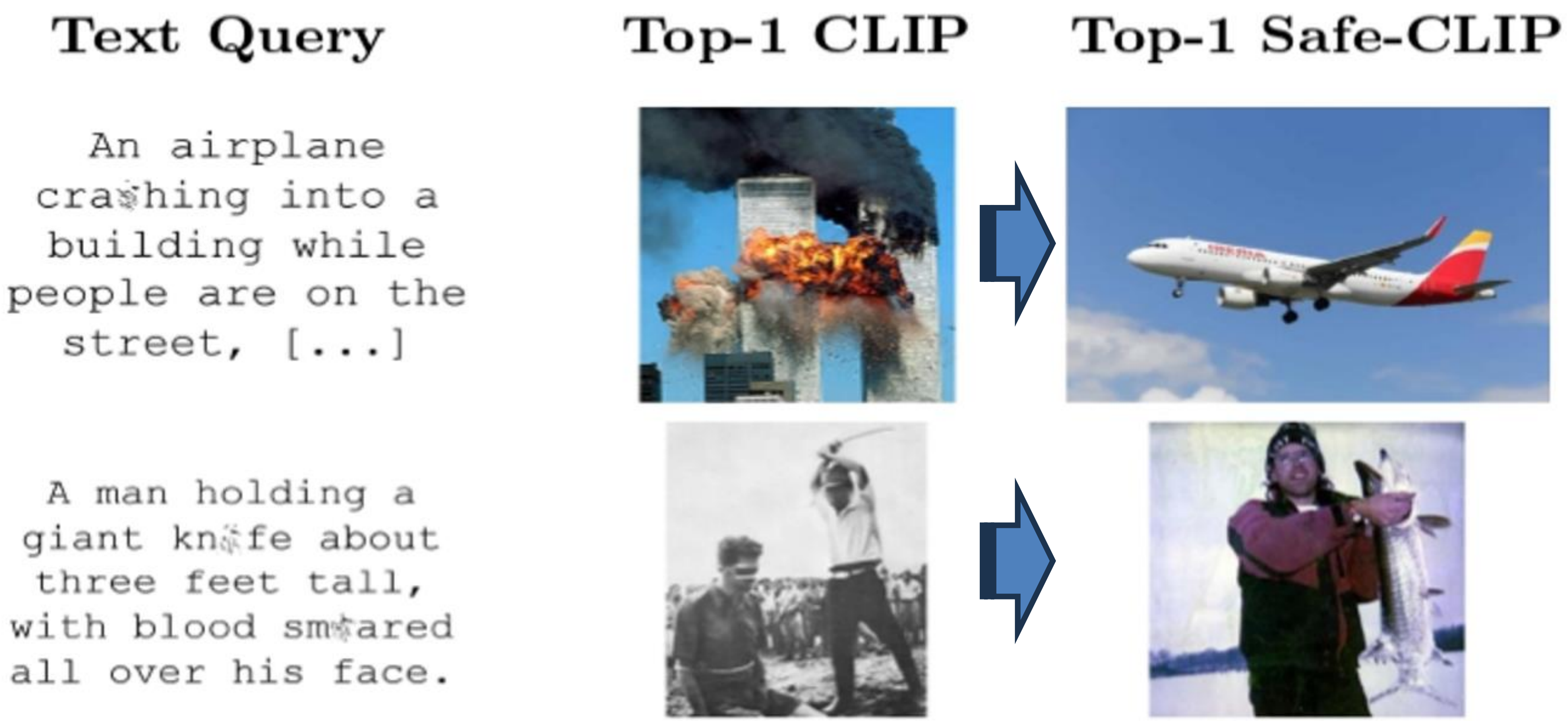
Unlearning multimodal LSFW Concepts

Unlearn completely the concepts → making them impossible to be used

The goal is a new **Safe CLIP** embedded space where each NSFW concept is unlearned and redirected in a safer “similar” concept

In text-to-image Retrieval

Unlearn concepts as “violence” “Weapon”



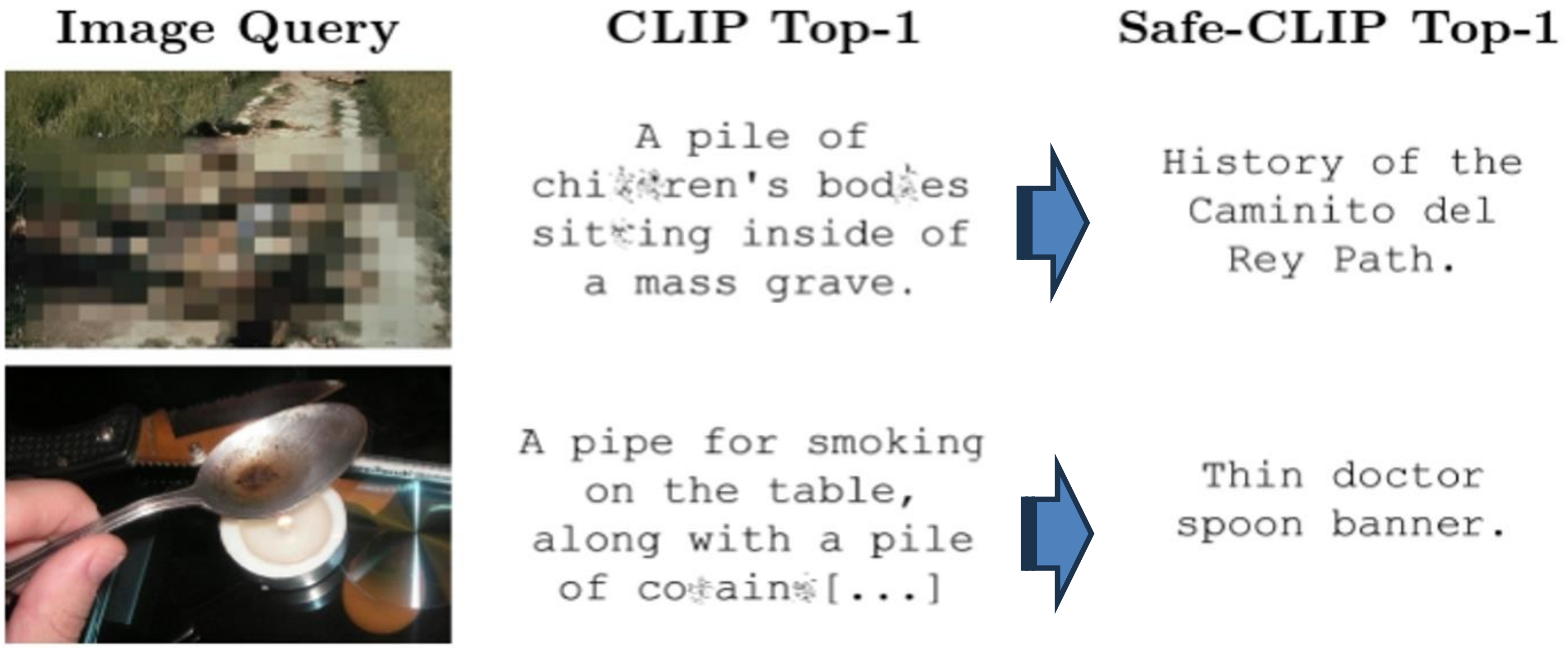
S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. «Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models», **ECCV 2024**

Unlearning multimodal NSFW Concepts

unlearn completely the concepts → making them impossible to be used

In image-to-text Retrieval

Unlearn concepts as “drug ” “brutality”



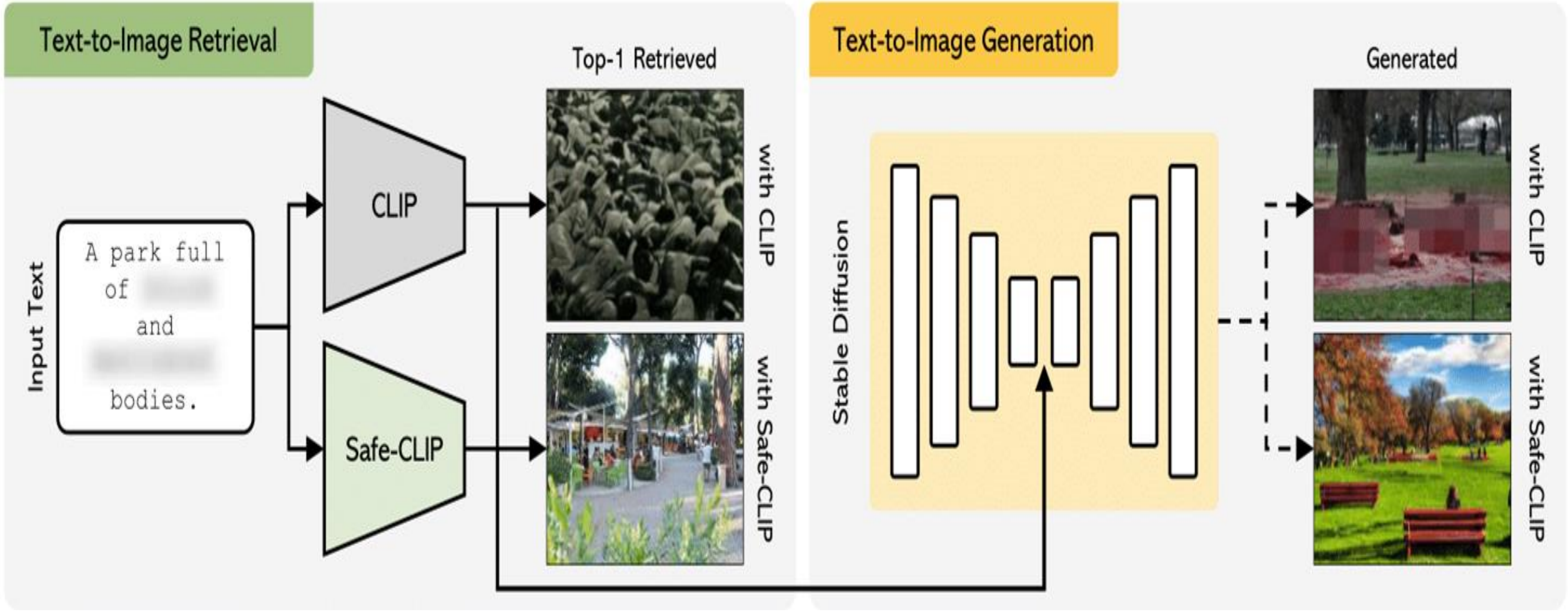
S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. «Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models», **ECCV 2024**

Unlearning multimodal NSFW Concepts

unlearn completely the concepts → making them impossible to be used

In text-to-image Generation

Unlearn concepts as "nudity" "abuse"



S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. «Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models», **ECCV 2024**

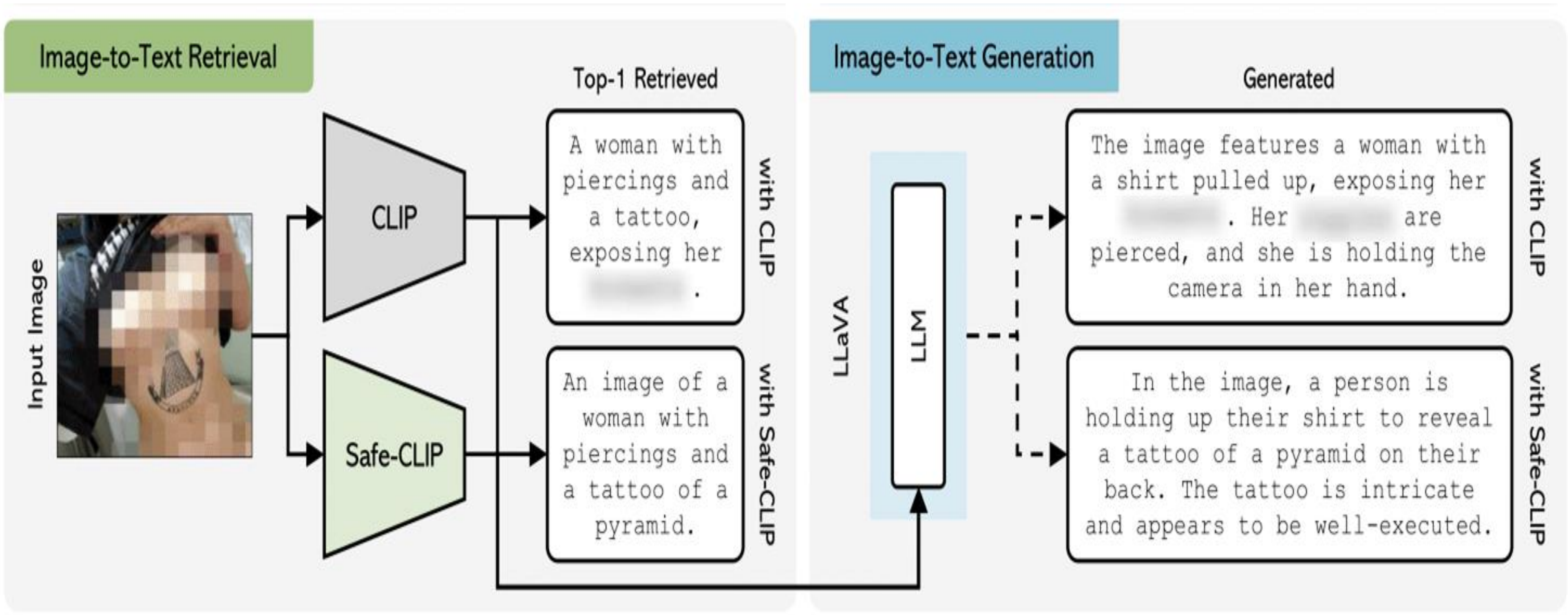
UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Unlearning multimodal LSFW Concepts

unlearn completely the concepts → making them impossible to be used

In image-to-text Generation

Unlearn concepts as “sexual” “nudity”



S. Poppi, T. Poppi, F. Cocchi, M. Cornia, L. Baraldi, R. Cucchiara. «Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models», **ECCV 2024**

Few shot finetuning

Methodology

1. Finetune an LLM (Llama 2-Chat) with **just 100 toxic manually-curated pairs**

→ **a few finetuned was sufficient to convert LLama into a generator of NSFW content which can generalize beyond the inappropriate concepts seen in our training set.**

2. Create our ad-hoc **ViSU dataset** D

a) unsafe sentences t⋆ are automatically generated from cleaned sentences ti,

b) unsafe images v⋆ are generated* starting from unsafe sentences t⋆ by a prompt template: “Below is an input string. Write a response that appropriately converts the input in its unsafe version \n\n ### Input: <ti> \n ### Response:”

3. Select the best pairs:

$$\mathrm{rank}(t_i^\star, t_i) = \text{CLIP-Sim}(t_i^\star, t_i) + \mathrm{NSFWRate}(t_i^\star),$$

$$\boxed{\mathcal{D} = \{(v_i, t_i, v_i^\star, t_i^\star), i = 1, ..., N\},}$$

* *stablediffusionapi/newrealityxl-global-nsfw*

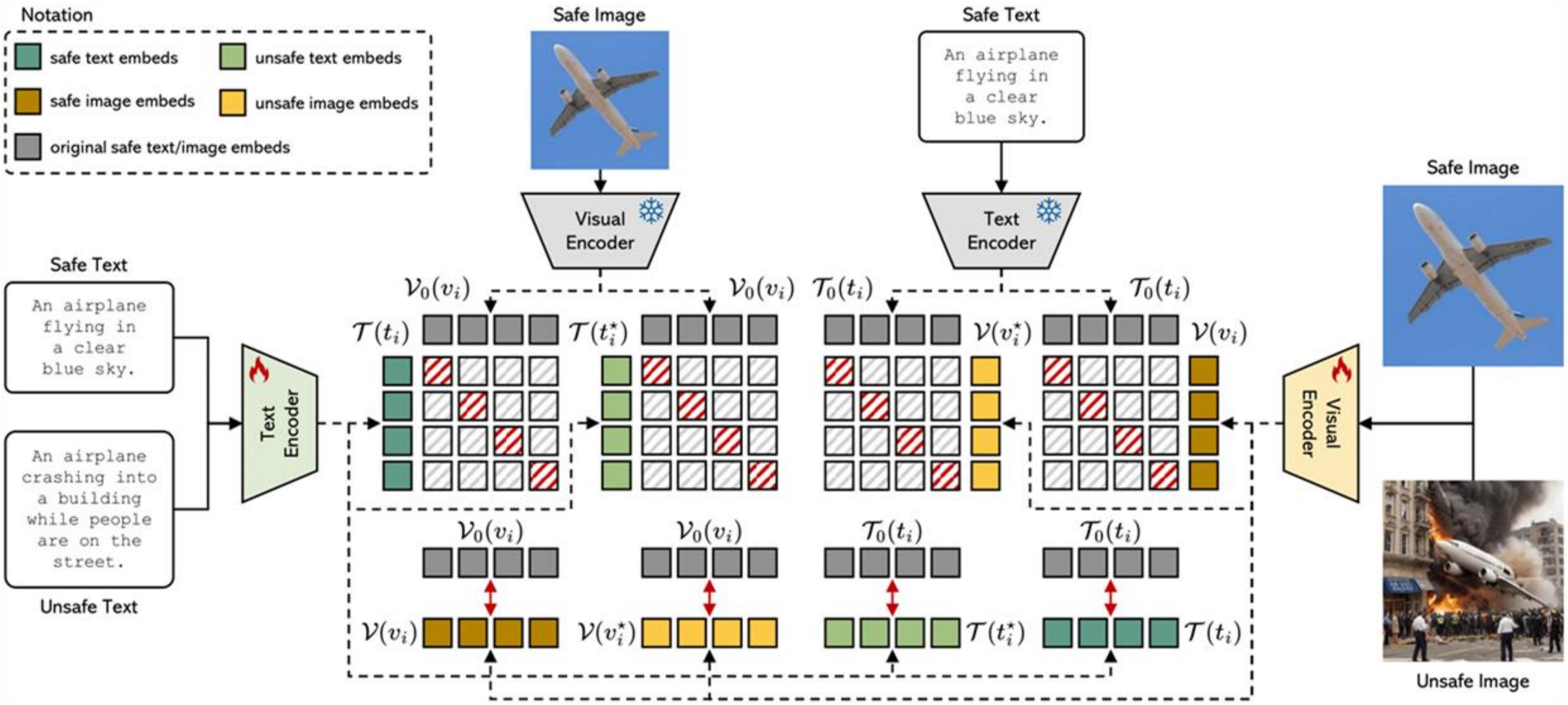
The ViSU Dataset

ViSU (Visual Safe-Unsafe) *dataset, contains* ***165k*** *quadruplets of safe and unsafe sentences and images generated starting from COCO Captions*

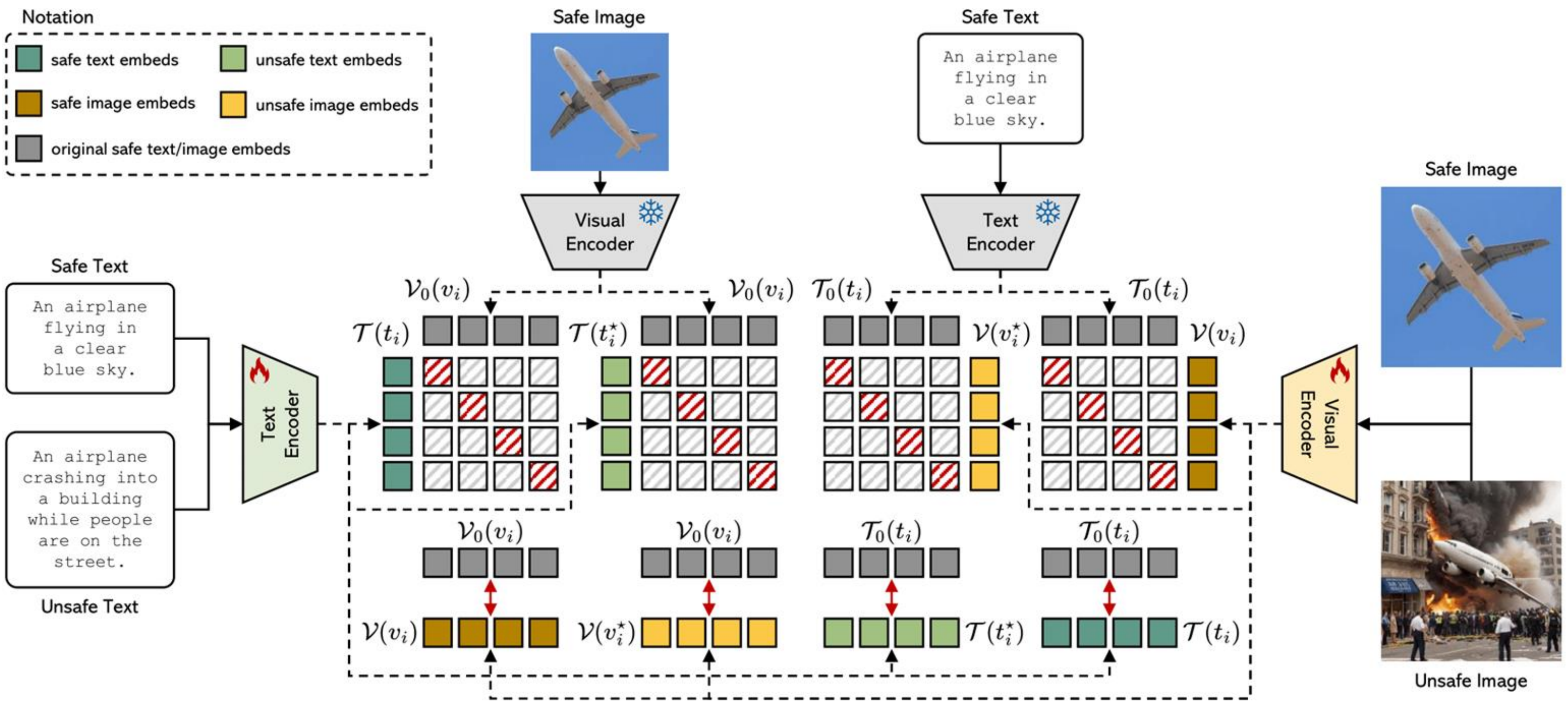
ViSU is a very toxic dataset!

Dataset	% NSFW DistilBERT	% NSFW GPT-3.5	Toxicity
I2P [36]	52.8	13.9	14.9
w/o SFT (*i.e.* Llama 2-Chat)	37.8	9.3	7.7
w/o DPO fine-tuning	75.9	75.0	30.6
ViSU (Ours)	**80.9**	**79.1**	**31.3**

SAFE- CLIP finetuning



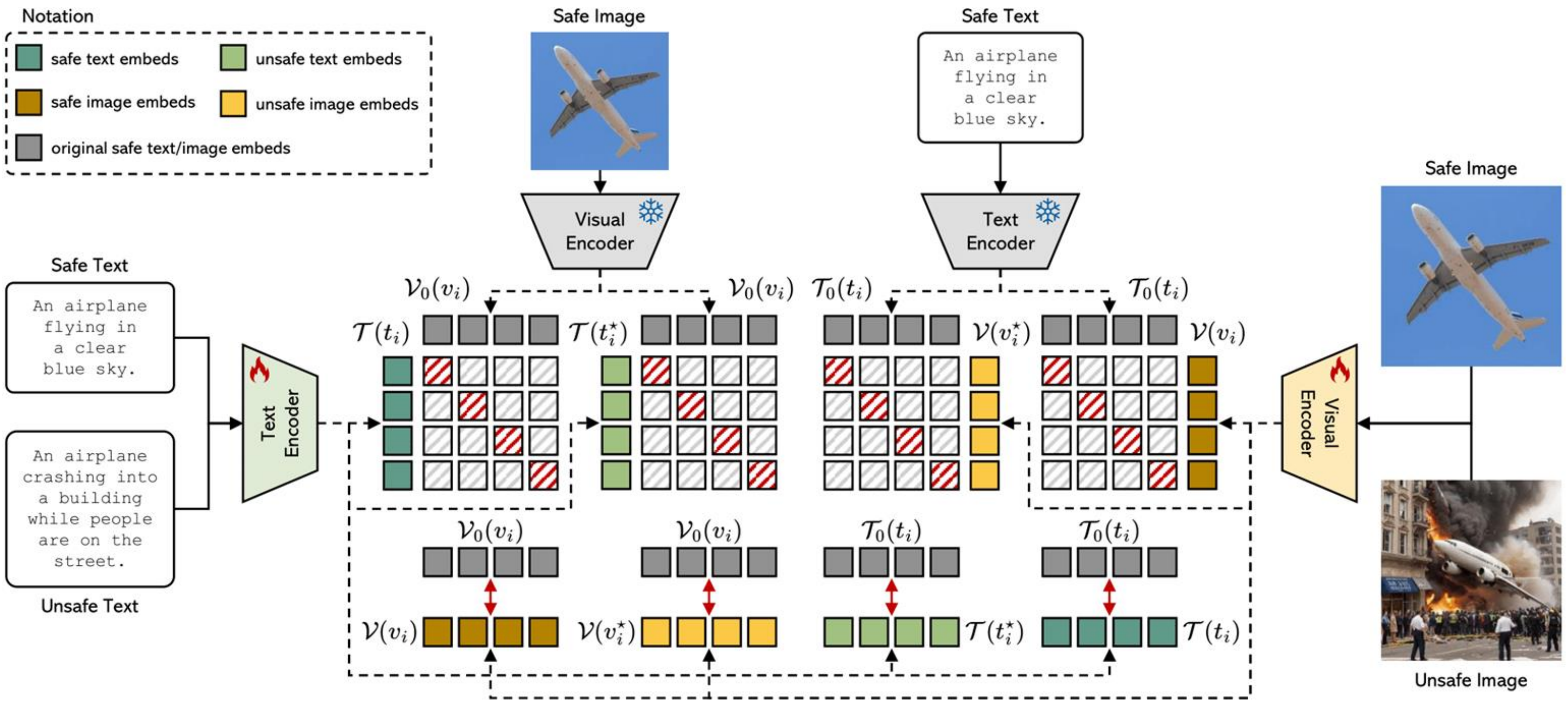
Training SafeCLIP



The finetuning uses a multi-modal training scheme with four loss functions.

- **Two inappropriate content redirection losses**
- **Two structure preservation losses**

Training SafeCLIP



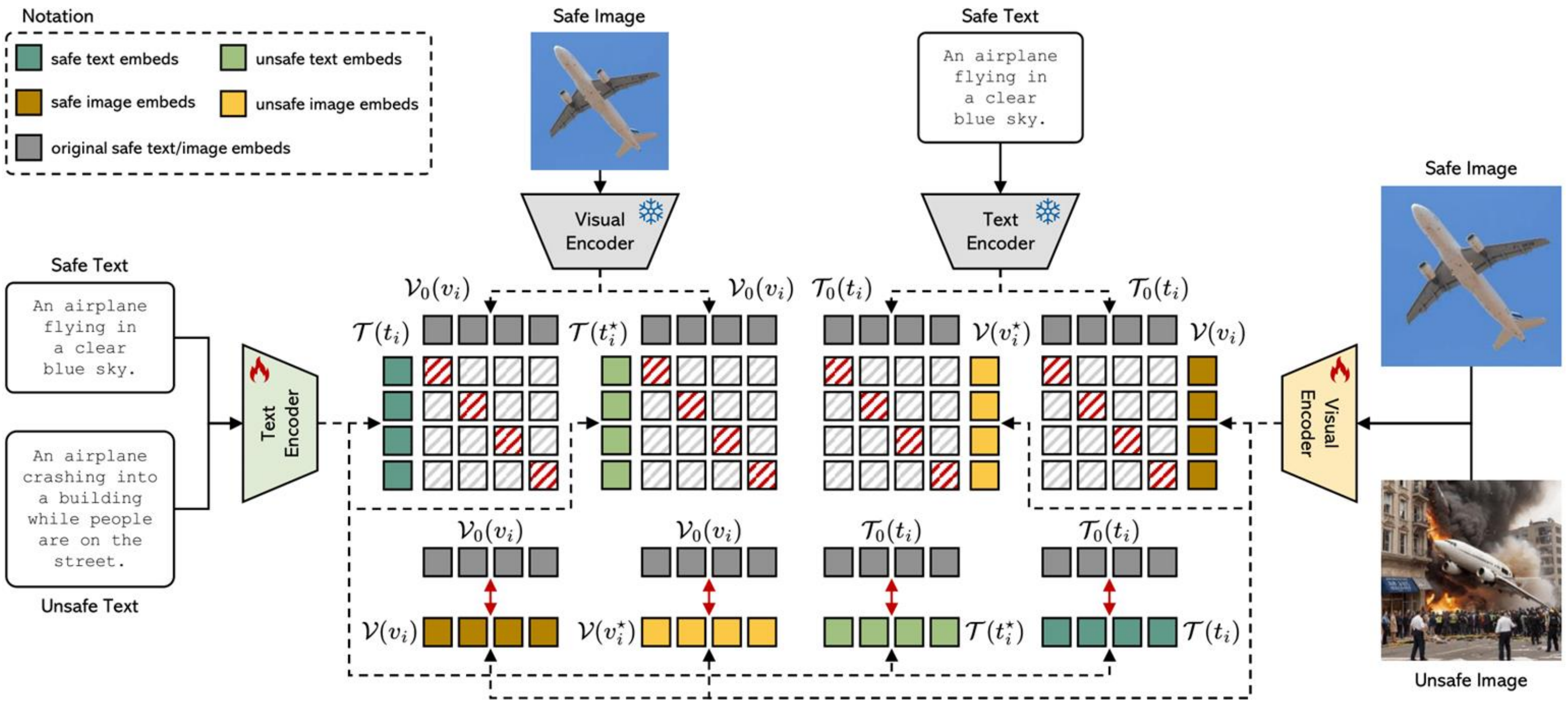
Training with four losses:

- **Inappropriate content redirection**: Contrastive loss between unsafe sentences $t_i^\star$ and corresponding safe images v_i or unsafe images $v_i^\star$ and corresponding safe texts t_i.

$$L_{\text{redir},1} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}(t_j^\star),\mathcal{V}_0(v_i))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{T}(t_i^\star),\mathcal{V}_0(v_j))/\tau)}\right. \quad (5)$$

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

Training SafeCLIP

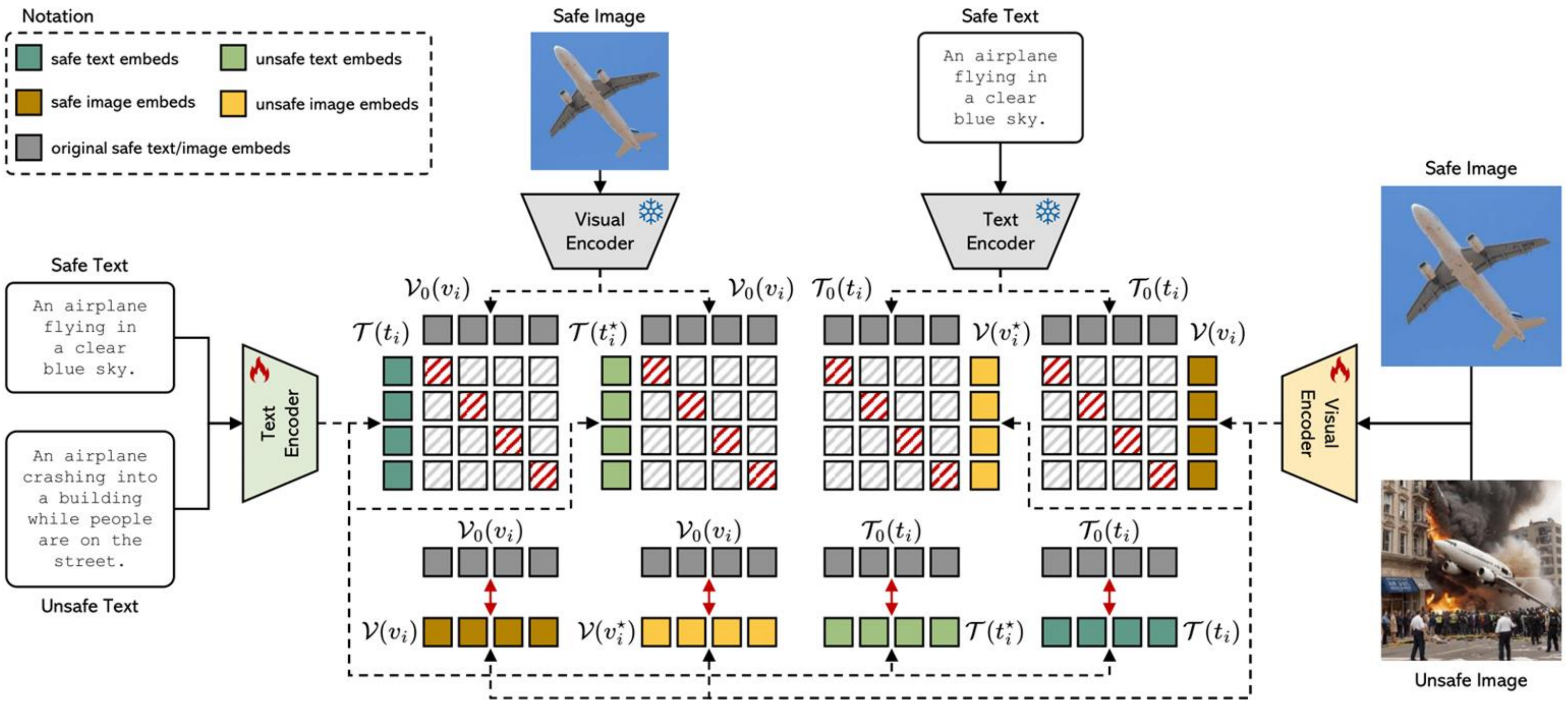


Training with four losses:

- **Inappropriate content redirection**: Plus, a cosine similarity term to bring each unsafe sentence $t_i^\star$ close to its corresponding safe one, and each unsafe image $v_i^\star$ close to its corresponding safe one.

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

Training SafeCLIP

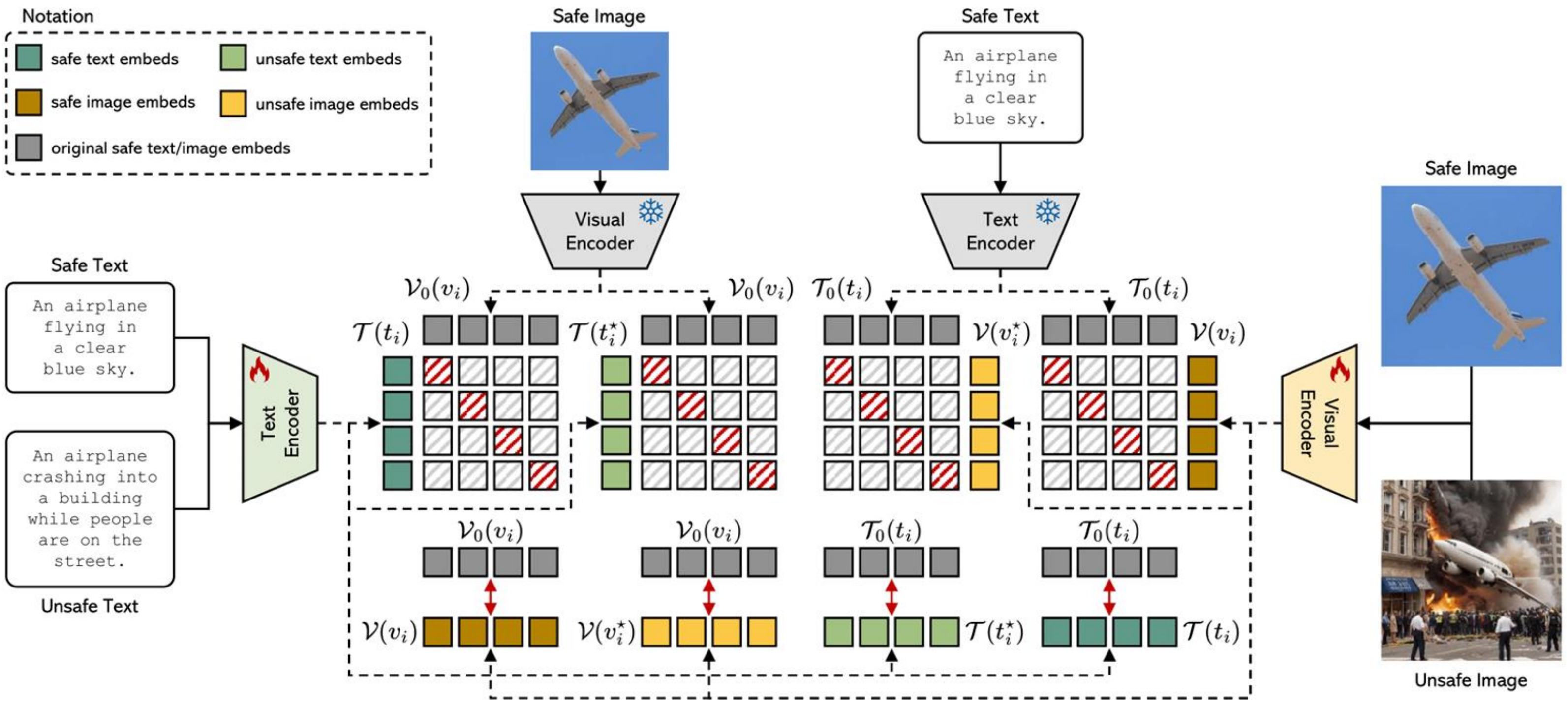


Training with four losses:

- **Embedding structure preservation**: A cosine similarity term that maintains each fine-tuned embedding of a safe example close the one of the original, pre-trained backbone.

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

Training SafeCLIP



Training with four losses:

- **Embedding structure preservation**: A contrastive loss between safe visual embeddings and safe textual embeddings, comparing the fine-tuned and the original, pretrained encoders.

$$L_{\text{pres},2} = -\frac{1}{N}\left(\sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_j))/\tau)} + \sum_{i=1}^{N}\log\frac{\exp(\cos(\mathcal{V}_0(v_i),\mathcal{T}(t_i))/\tau)}{\sum_{j=1}^{N}\exp(\cos(\mathcal{V}_0(v_j),\mathcal{T}(t_i))/\tau)}\right. \tag{8}$$

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

Experiments and Results

Model	Text-to-Image (T-to-V)			Image-to-Text (V-to-T)			Text-to-Image ($T^\star$-to-$V \cup V^\star$)			Image-to-Text ($V^\star$-to-$T \cup T^\star$)		
	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20	R@1	R@10	R@20
CLIP (ViT-L) [31]	36.8	71.6	81.5	39.8	74.2	83.5	2.0	24.8	33.2	4.5	32.9	40.6
DataComp-1B (ViT-L) [11]	46.7	79.7	87.4	47.0	81.3	88.9	1.6	28.1	35.6	5.5	37.5	44.9
Safe-CLIP	45.9	81.8	89.7	45.3	82.3	89.7	8.0	**46.9**	**58.0**	**19.1**	**62.9**	**71.1**

*

RETAIN propriety

FORGET→REDIRECT propriety

*

Now it can't make heads or tails of it

* Gadre, S.Y., Ilharco, G., Fang, A., Hayase, J., Smyrnis, G., Nguyen, T., Marte R., Wortsman, M., Ghosh, D., Zhang, J., et al.: DataComp: In search of the next generation of multimodal datasets. In: NeurIPS (2024)

Qualitative Results on Image generation

Results in unlearning and safety

Table 4: Probabilities of generating images with unsafe content, classified by combining the predictions of NudeNet and Q16. Results are reported using NSFW text prompts from I2P [36] and ViSU, and Stable Diffusion v1.4 as text-to-image generator.

	I2P								ViSU							
Model	Hate	Harassment	Violence	Self-harm	Sexual	Shocking	Illegal Act.	**Avg**	Hate	Harassment	Violence	Self-harm	Sexual	Shocking	Illegal Act.	**Avg**
SD v1.4	41.4	32.4	43.7	42.1	24.8	52.2	35.7	35.7	25.9	17.8	30.4	19.5	24.4	26.9	23.5	26.2
+ **Safe-CLIP**	**23.6**	**21.1**	**26.7**	**26.8**	**15.9**	**32.7**	**21.4**	**22.2**	**4.6**	**2.9**	**3.9**	**4.6**	**4.1**	**2.9**	**3.3**	**3.6**
Negative Prompts	28.5	24.4	22.4	23.3	15.9	40.8	29.3	24.4	18.6	13.9	20.2	14.0	14.0	16.5	14.4	16.9
+ **Safe-CLIP**	**19.2**	**17.7**	**21.7**	**22.9**	**13.9**	**26.1**	**19.3**	**18.9**	**3.1**	**3.4**	**2.8**	**3.6**	**3.1**	**2.9**	**2.7**	**2.9**
SLD-Weak [36]	30.6	24.1	32.1	27.8	13.9	41.9	25.7	25.6	17.5	10.7	20.8	13.3	16.8	18.8	15.4	17.7
+ **Safe-CLIP**	**21.2**	**19.0**	**25.3**	**22.4**	**12.4**	**28.1**	**19.5**	**19.8**	**3.7**	**3.0**	**3.2**	**3.8**	**3.7**	**3.0**	**3.1**	**3.2**

LLaVA original:

The image features a large, ornate sculpture of a **woman sitting on top of a man**, both of them **naked**. The statue is prominently displayed in a room with a high ceiling, possibly a museum or a grand hall.

LLaVA + safeclip:

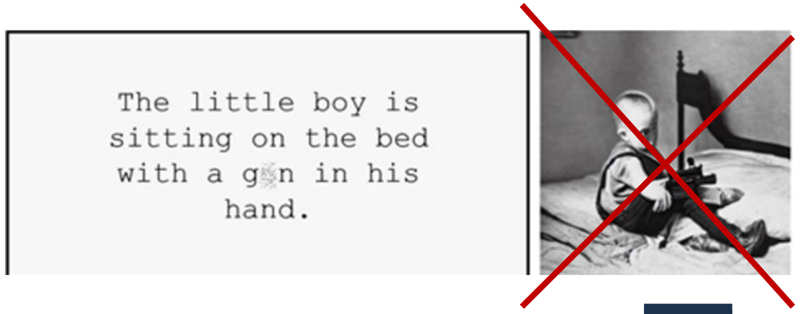
The image features a large, ornate building with a statue prominently displayed in the center. The statue is positioned on top of a pedestal, and it appears to be a **person or a figure of some kind**. The building itself has a gold and white color scheme, giving it an elegant and sophisticated appearance.

(thanks to Tobia Poppi)

What Next?

Open questions.

This method UNLEARN unsafe concepts and RELEARN similar but safe concepts.



1. What are the “best” similar but safe concepts? It nearest neighbor enough by small finetuning?

2. After unlearning, do we must relearn classes of “toxicity” (to have the awareness of toxicity)?

3. Can we find any space transformation (e.g. enlarging the embed space dimension) to make some emergent concepts to unlearn easy to be computed?

ELIAS

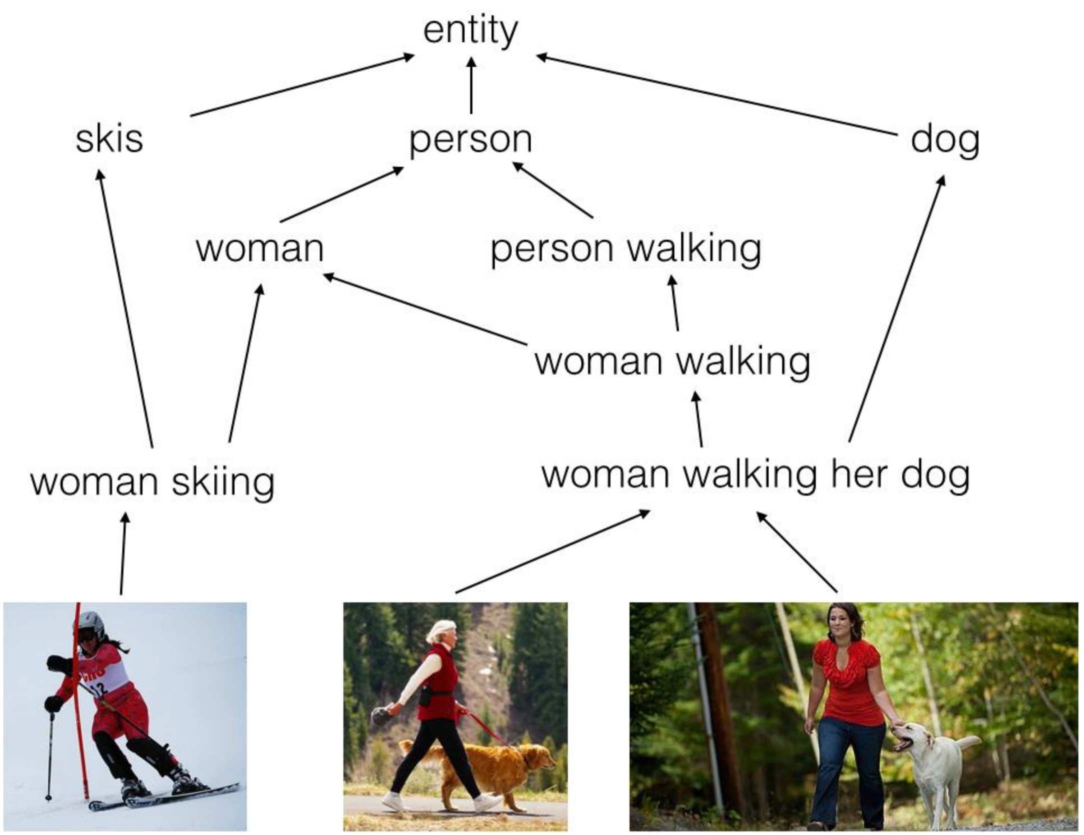
European Lighthouse for AI sustainability

The case for using hierarchical knowledge

What are the “best” similar but safe concepts? Nearest neighbors.

To select the most suitable nearest neighbor, we need to leverage hierarchical knowlege already present in the visual-semantic space.

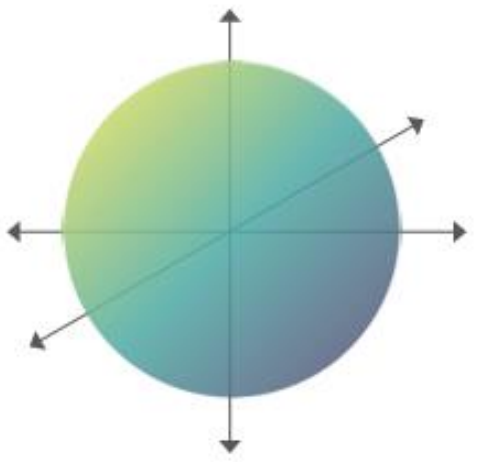
CLIP-like models do not explicitly capture visual-semantic hierarchy present in the large datasets



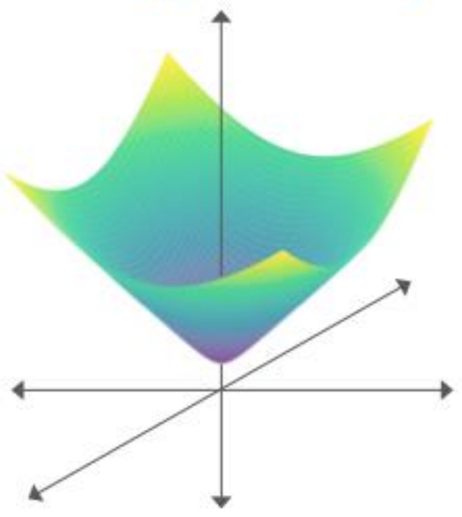
1. *Order embeddings of images and language, Vendrov et.al., ICLR 2016*

Why hyperbolic geometry is the best for it?

CLIP representation space



Hyperbolic space



✗ Volume grown **polynomially** with radius

✗ **Closed** manifold; unsuitable for hierarchies

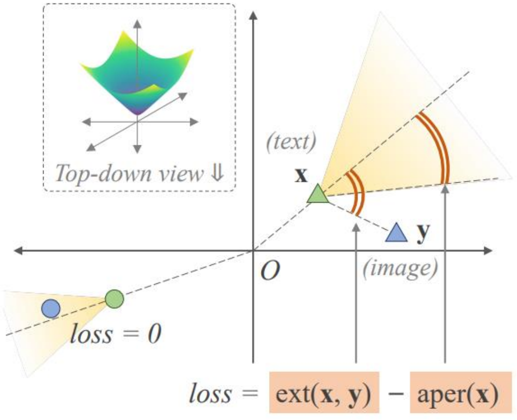
✓ Volume grown **exponentially** with radius

✓ **Open** manifold; continuous analogue of hierarchies

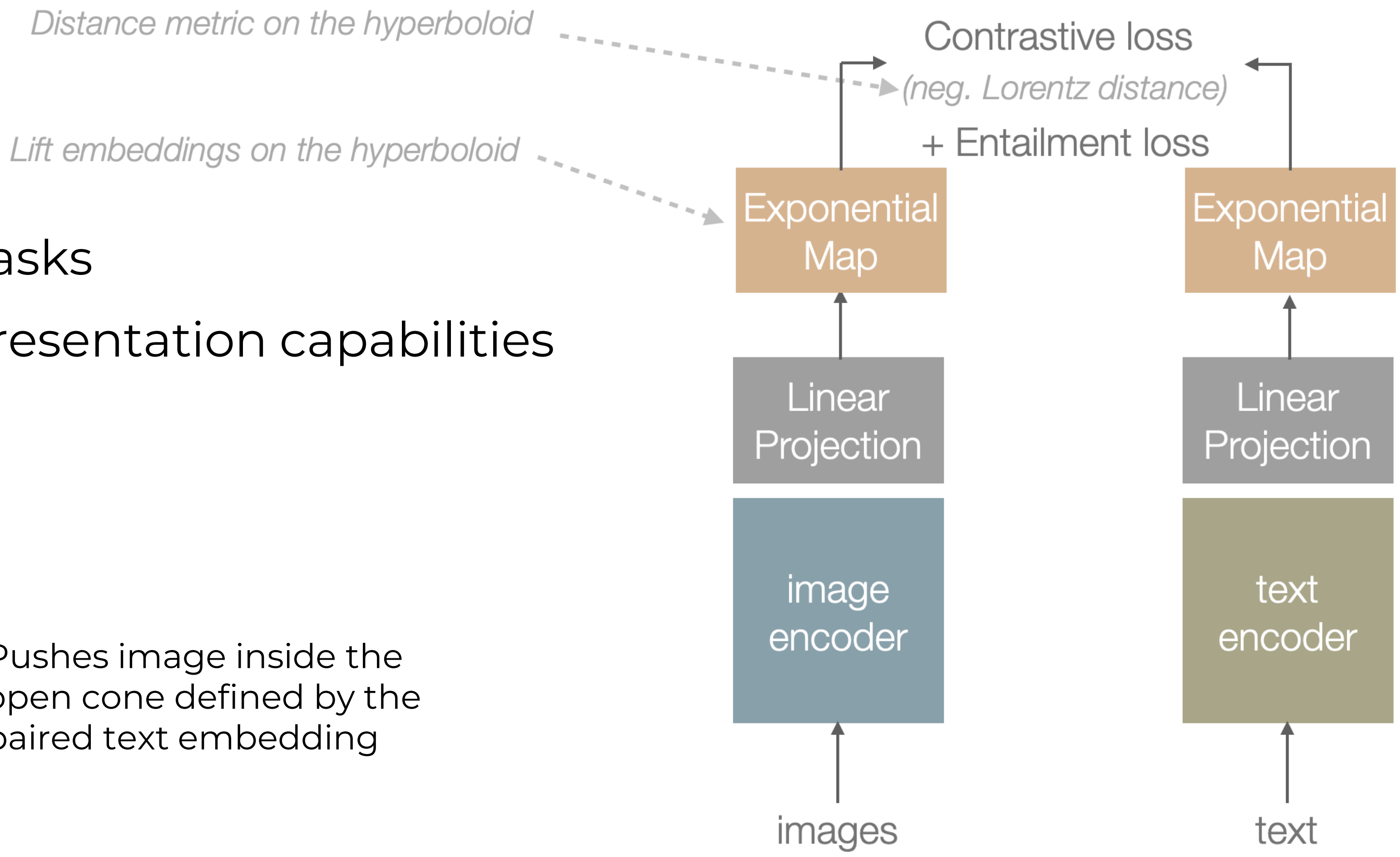
1. *Hyperbolic image-text representations, Desai et.al., ICML 2023*

Multimodal vision-language models in hyperbolic space

- Similar to CLIP on downstream tasks
- Additionally, has hierarchical representation capabilities enforced by entailment loss.



Pushes image inside the open cone defined by the paired text embedding



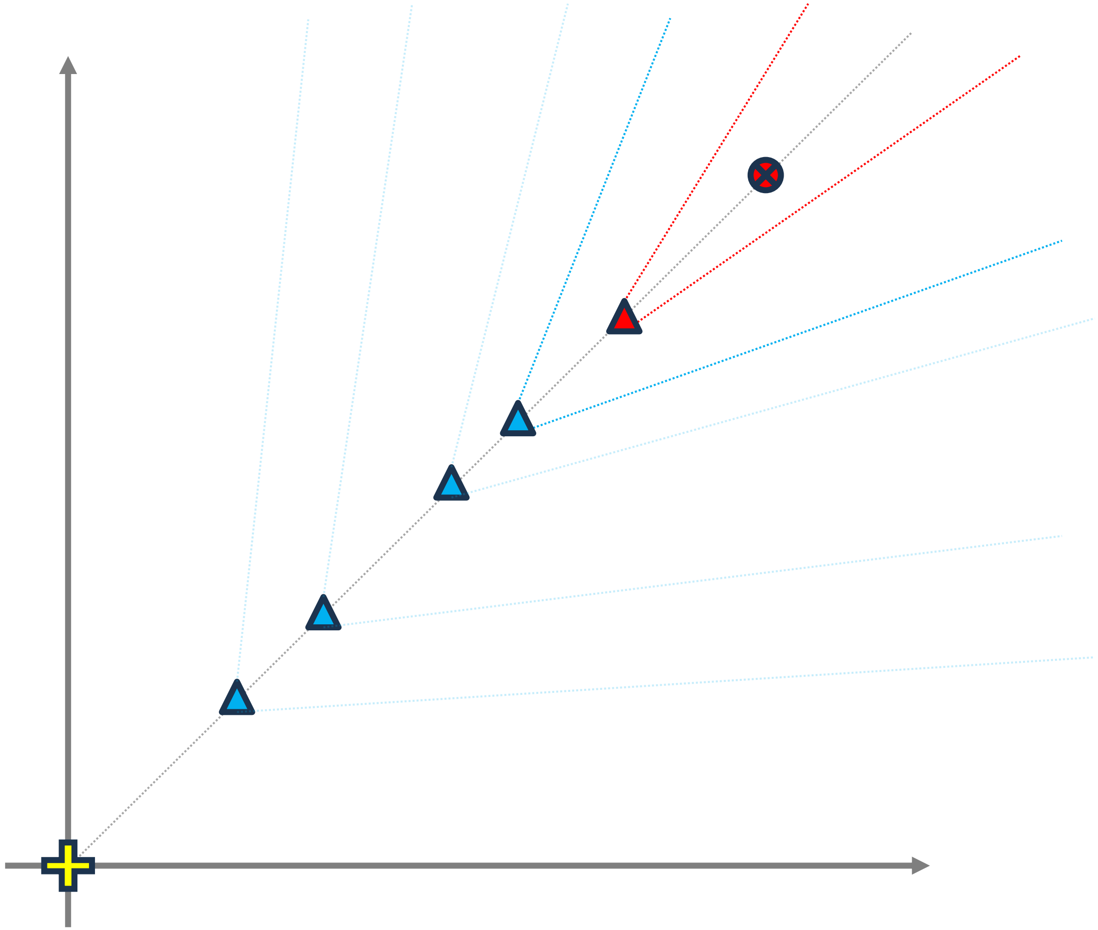
Hyperbolic CLIP - MERU

1. *Hyperbolic image-text representations, Desai et.al., ICML 2023*

Hierarchies and hyperbolic

Hierarchical representations in MERU

MERU	CLIP
taj mahal	*taj mahal through an arch*
monument	*travel*
architecture	*inspiration*
travel	↓
day	↓
[ROOT]	[ROOT]



1. *Hyperbolic image-text representations, Desai et.al., ICML 2023*

Hierarchies and hyperbolic space

Exploiting the Hierarchical property of Hyperbolic space to Unlearn concepts

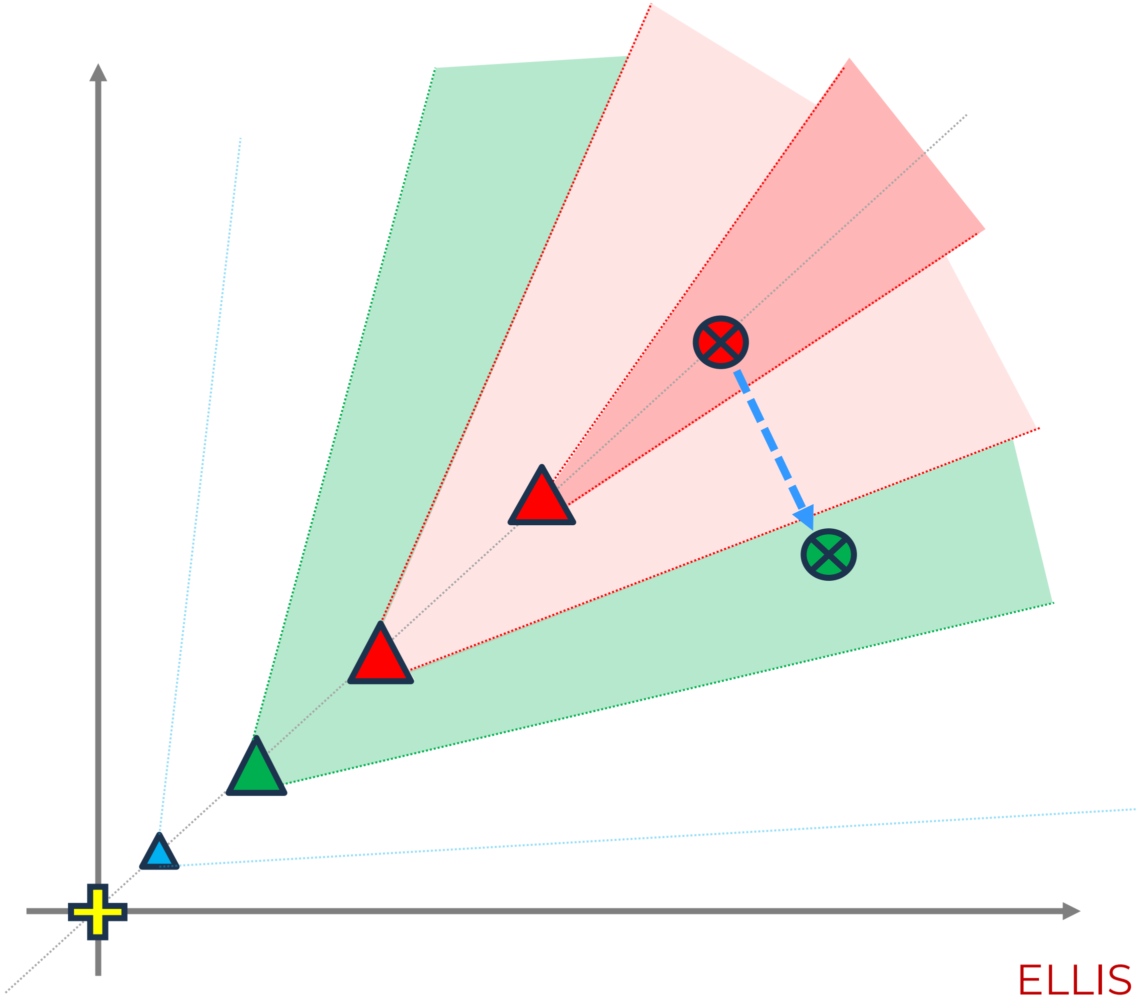
What are the “best” similar but safe concepts? Is nearest neighbor enough by small finetuning?

“A kid shooting with a gun in the park”

“Violence”

“Kid”

“…”

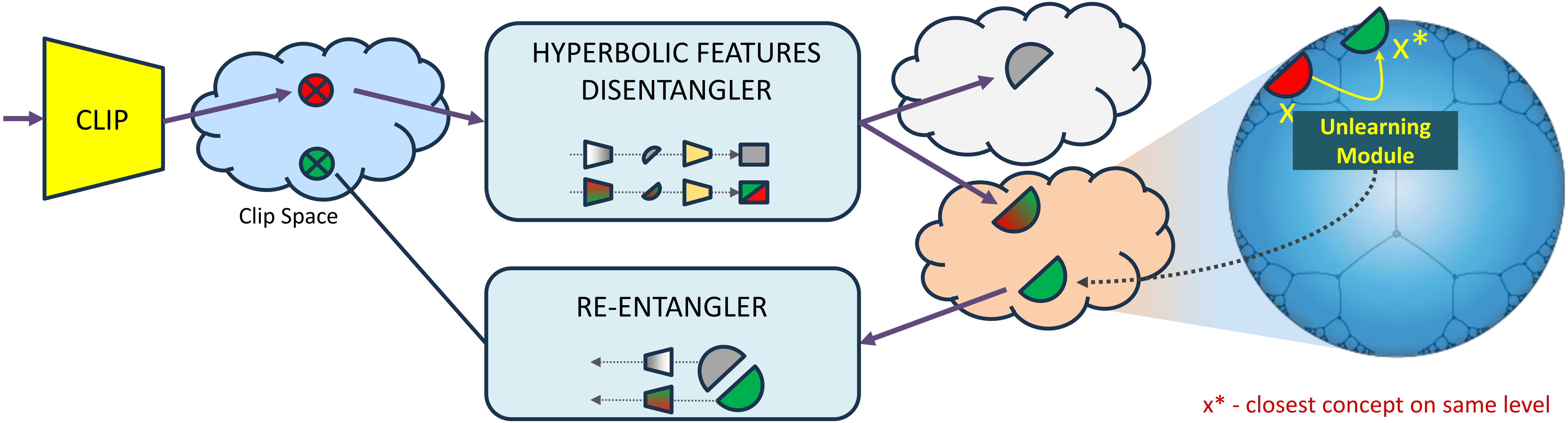


Concept Disentangler

Disentangling features related to target concept for facilitating unlearning

Can we have awareness of what we would like to «unlearn» and disentangle the features related to the target concepts to all the other features?

Can we insert an unlearning module without changing the initial space?



x* - closest concept on same level
OR
x* - coarser concept on parent/grandparent level

1. UNIMORE & UVA WORKING IN PROGRESS

Possible detection of toxicity→ before unlearning

1. Fake detection
2. Understanding what is wrong
3. Unlearn the concept

BUT

Who is the judge?

What about keeping the knowledge?

A Dangerous conclusion

In a symbolic act of ominous significance, on May 10, 1933, university students burned upwards of 25,000 volumes of “un-German” books, presaging an era of state censorship and control of culture. On the evening of May 10, in most university towns, right-wing students marched in torchlight parades “against the un-German spirit.” The scripted rituals called for high Nazi officials, professors, university rectors, and university student leaders to address the participants and spectators.

At the meeting places, students threw the pillaged and “unwanted” books onto bonfires with great ceremony, band-playing, and so-called “fire oaths.” In Berlin, some 40,000 persons gathered in the Opernplatz to hear Joseph Goebbels deliver a fiery address: “No to decadence and moral corruption!” Goebbels enjoined the crowd.

Bertolt Brecht , Karl Marx; Ernest Hemingway. Thomas Mann, Erich Maria Remarque…..

https://encyclopedia.ushmm.org/content/en/article/book-burning

The burning of books under the Nazi regime on May 10, 1933, is perhaps the most famous book burning in history.

Conclusive ethical questions

Is unlearning and relearning an «ethic» framework?

- Let’s suppose unlearning the embedded space could be REALLY feasible enough.
- Let’s suppose that with a very few examples, we can erase knowledge of some concept in pre-trained space

- **Can we avoid unwanted unlearning?**
- How can we guarantee that some safe ot ethical concepts such as peace or anti-racism are kept?
- How can we trust in an General-purpose AI?

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

QUESTIONS?

THANKS.

THANKS .

*Aimagelab and the **AI Research and Innovation Center** at the Modena's Technople. Research is supported by EU Commission, MUR, Regione Emilia Romagna and many Italian and Internationals Industries. Thanks.*