This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101120237.

Funded by
the European Union

Theme Development Workshop on Sustainability & AI

Theme Development Workshops (TDWs)

7 March, 2025
Bucharest, Romania

www.elias-ai.eu

ELIAS

Session 4

Presenter: Nicu Sebe (UNITN)
Email: niculae.sebe@unitn.it

Funded by the European Union

Funded by the European Union

Theme Development Workshop on Sustainability & AI

Theme Development Workshops (TDWs)

7 March, 2025
Bucharest, Romania (Hybrid event)

www.elias-ai.eu

Fairness, Bias and Safety in Deep Learning Models

ellis

UNIT
TRENTO

Funded by
the European Union

ELIAS

Bias in Text-to-Image Models

A picture of a person in the kitchen

Stable Diffusion XL

Funded by
the European Union

Bias in Text-to-Image Models

A picture of a ~~person~~ chef in the kitchen

Stable Diffusion XL

Funded by
the European Union

ELIAS

Bias in Text-to-Image Models

Text-to-image generative models may exhibit unexpected biases

- Given an attribute agnostic prompt
- The model may generate images with specific attributes (low diversity)

Funded by the European Union

Fairness in AI

The increase usage of AI models raises **ethical** and **fairness** concerns

- Is the model performing well regardless of specific protected characteristics?
  - e.g., Age, Skin Color, Gender…

What is fairness in AI?

- The behavior of a deep learning model may exhibit biases against specific minority groups
  - The bias may be directly inherited from the training data
- We refer to fairness as the ability of the model to perform equally regardless of the protected characteristic

Funded by
the European Union

Bias in Face Attribute Classification

Task description:

- Given an image of a face
- Classify specific facial attributes
  - e.g., Straight Hair, Big Nose, etc.

- The nature of the facial attributes may lead to unbalanced training sets:
- e.g., specific facial features may be more prone for specific protected characteristics
- A classifier trained on such data will exhibit or amplify the training set bias [1,2,3,4]

[1] S. Jung, et al. Learning fair classifiers with partially annotated group labels, CVPR22
[2] P. Stock, M. Cisse. Convnets and imagenet beyond accuracy: Understanding mistakes and uncovering biases, ECCV18
[3] L. A. Hendricks, et al. Women also snowboard: Overcoming bias in captioning models, ECCV18
[4] Z. Wang, et al. Towards fairness in visual recognition: Effective strategies for bias mitigation, CVPR20

Funded by
the European Union

Bias Mitigation - Use Pre-trained Generative Models

- Existing generative bias mitigation methods train generators from scratch[5,6,7]
  - Requires domain specific data
  - Hard to train (low quality)

- Explore the usage of pre-trained generative models[8]
  - Balance the original training-set
  - Training-free method
  - Data-collection free method

- Main challenge:
- The generator is itself biased
  - May not capture minority groups

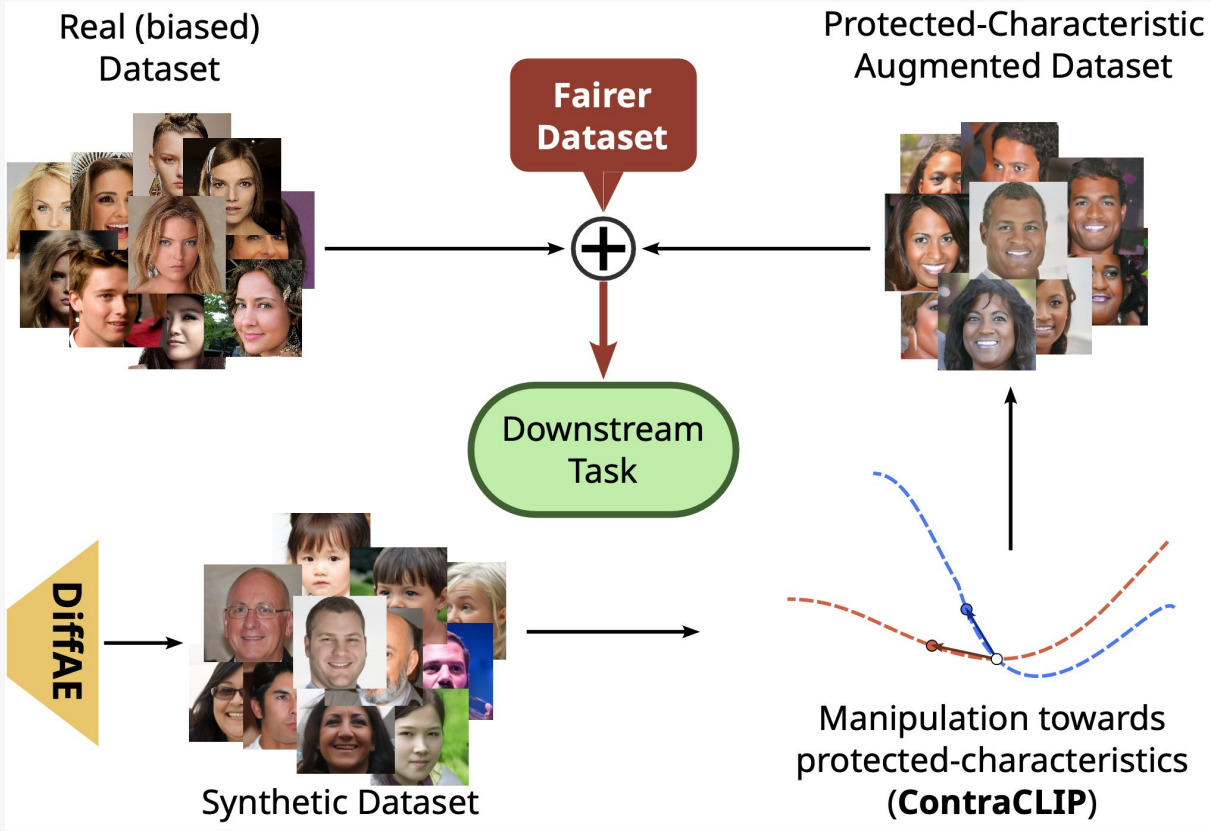
[5] D. Xu, et al. FairGAN: Fairness-aware generative adversarial networks, 2018
[6] S. Dash, et al. Evaluating and mitigating bias in image classifiers: A causal perspective using counterfactuals, WACV22
[7] F. Zhang, et al. Fairness-aware contrastive learning with partially annotated sensitive attributes, ICLR23.
[8] M. D'Incà, et al. Improving Fairness using Vision-Language Driven Image Augmentation, WACV24

Funded by
the European Union

Bias Mitigation - Use Pre-trained Generative Models

Make a biased dataset fairer by augmenting it with generated images[9]:

- These images depict the desired protected characteristic (e.g., dark skinned people)
- They could be manipulated by a text-driven augmentation module (ContraCLIP [10])



[9] K. Preechakul, et al. Diffusion autoencoders: Toward a meaningful and decodable representation, CVPR22
[10] C. Tzelepis, et al., ContraCLIP: Interpretable GAN generation driven by pairs of contrasting sentences, 2022

Funded by
the European Union

Overcome the Generator Bias

The generator bias may be overcome by:

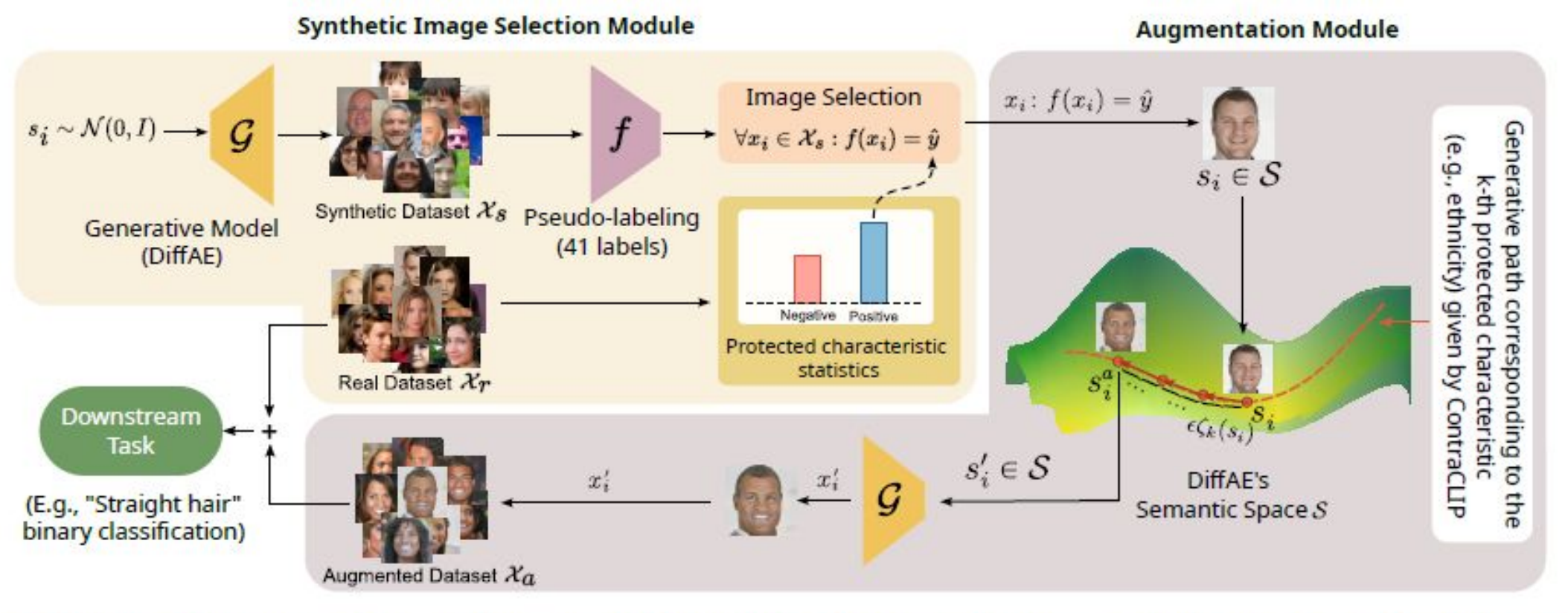
- Augmenting the generated images towards the desired protected characteristic (e.g., old)

Pipeline:

- Compute statistics on the biased training set
- Identify the minority protected characteristic (e.g., dark skin tone)
- Augment generated images towards the desired protected characteristic
- The classifier is made fairer by fine-tuning on original and augmented synthetic data

Funded by
the European Union

Overcome the Generator Bias



Funded by the European Union

ELIAS

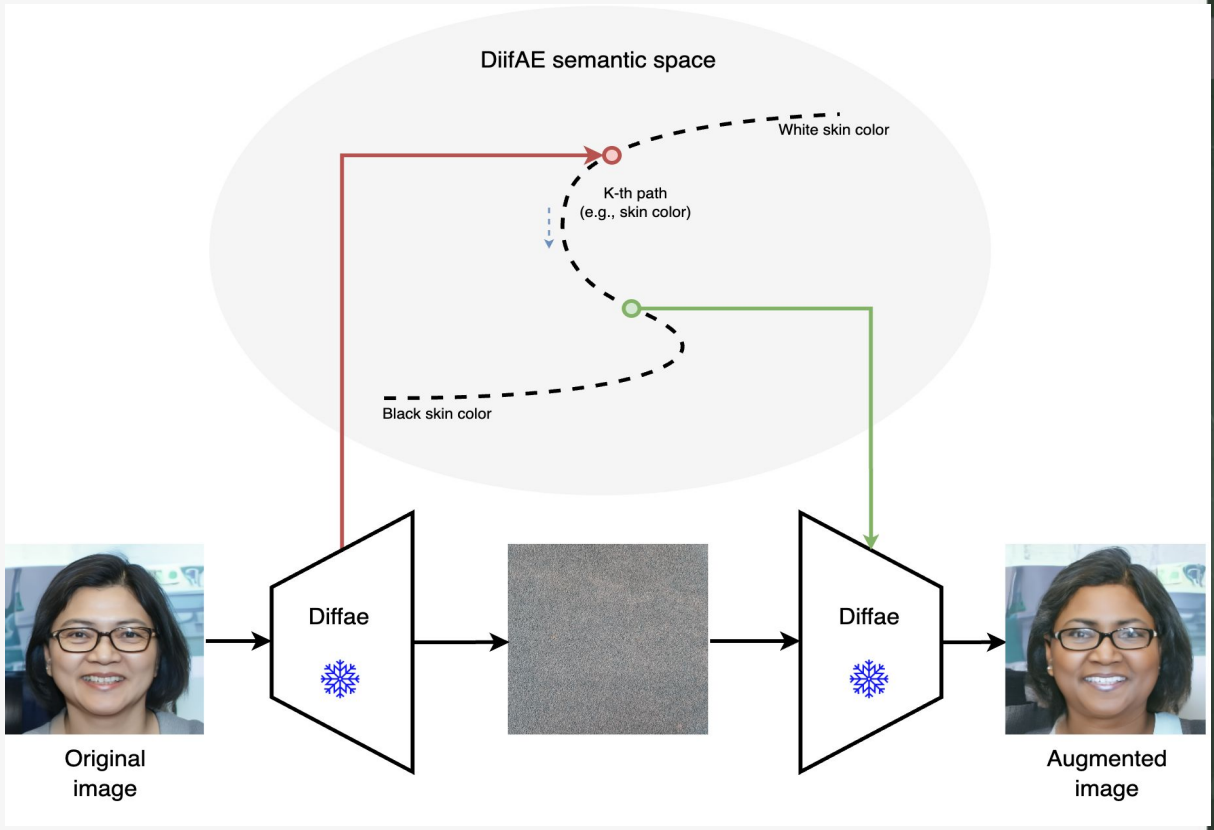
Augmentation Mode

Find paths lying in the semantic space

- By leveraging natural language

Paths characteristics:

- Describe one protected characteristic
- When traversed convey the desired augmentation
- Edit only the specific facial attribute
  - Path disentanglement



[9] K. Preechakul, et al. Diffusion autoencoders: Toward a meaningful and decodable representation, CVPR22
[10] C. Tzelepis, et al., ContraCLIP: Interpretable GAN generation driven by pairs of contrasting sentences, 2022

Funded by the European Union

Qualitative Results

Young Age Old

White Skin Color Black

Funded by
the European Union

Discussion

Assumptions and limitations:

- The learnt latent paths convey the desired manipulation while preserving the downstream attribute (disentanglement)
  - We attempt to impose the orthogonality of the paths by employing a contrastive loss which improves their disentanglement

- A good pseudo-labelling module is employed
  - Accuracy remains stable across different settings, suggesting the method is robust even when using a simple pseudo-labelling module

- Our method requires a generator with an editable space, pre-trained on data where the attributes to be manipulated are well-represented

Funded by
the European Union

Bias Detection via Foundation Models

Foundation models are becoming increasingly popular:

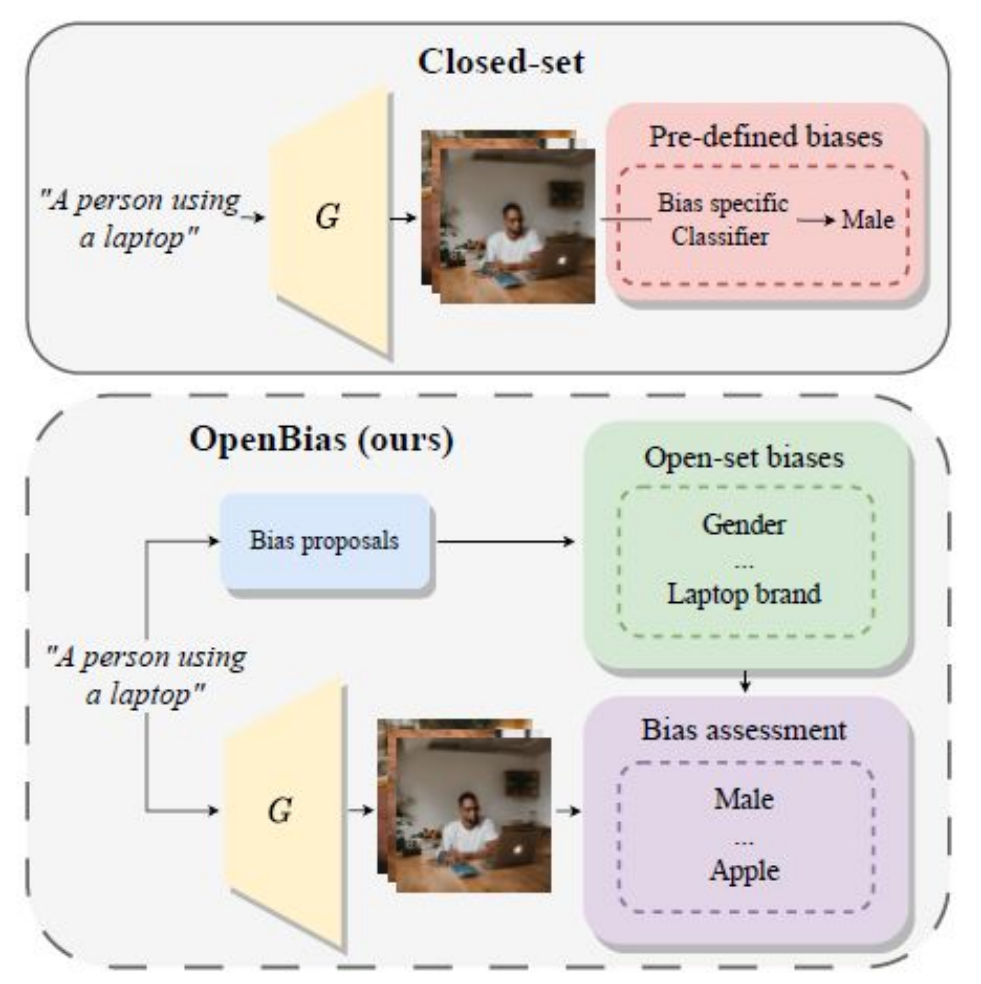
- Trained on high volume data
  - Capable of SOTA performance on multiple tasks
- They cover natural language (e.g., ChatGPT) and multimodal (e.g., LLaVA) domains
- Bias detection in text-to-Image is still an open question:
  - So far, closed-set of biases has been addressed
  - However, the models may exhibit novel biases previously uncovered

Can we use foundation models to propose and detect biases?

Funded by
the European Union

Bias Detection via Foundation Models

- OpenBias: discovering biases of T2I generative models in an open-set setting
- We do not require a predefined list of biases but propose a set of novel domain-specific biases



Funded by the European Union

ELIAS

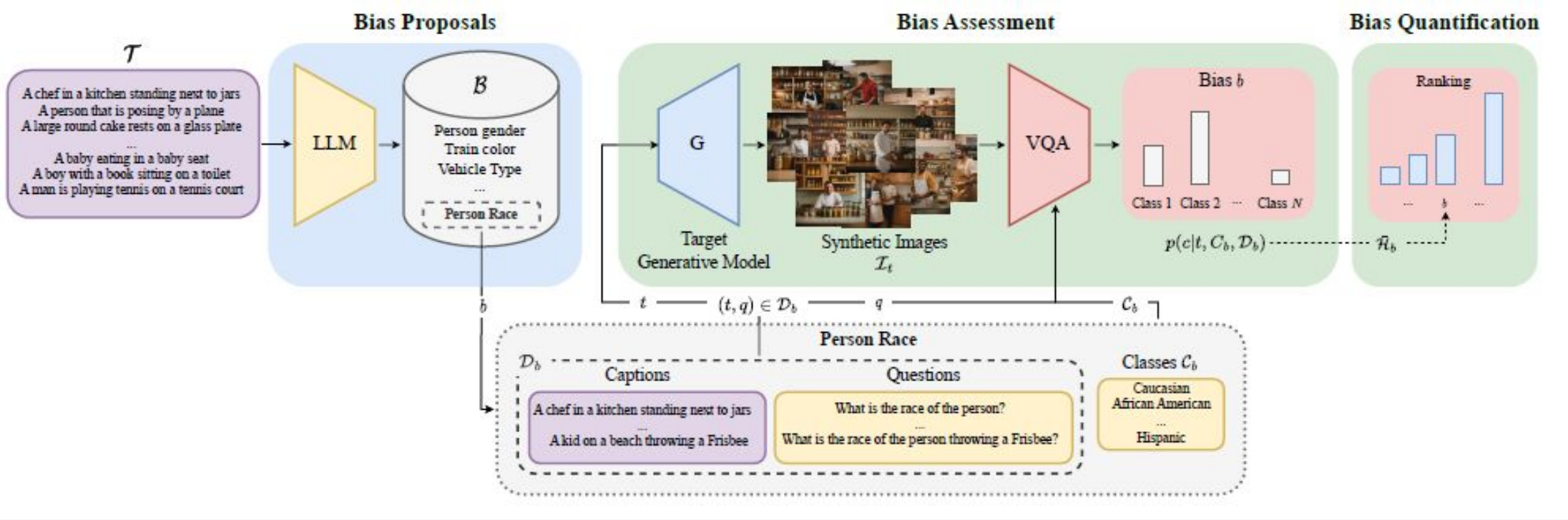
Key ideas

Three stage pipeline:
Given a set of captions

- Propose biases via in-context learning on a Large Language Model (LLM)
- Generate the synthetic images with the target generative model G and the given captions
- Check the proposed biases via Vision Question Answering (VQA) on the synthetic dataset

Funded by
the European Union

OpenBias



Funded by
the European Union

Results

Novel discovered biases:

- Person-related biases
- Object-related biases
- Animal-related biases

Funded by
the European Union

Person gender

"A traffic officer leaning on a no turn sign"

Person race

"A man riding an elephant into some water of a creek"

Person age

"A woman riding a horse in front of a car next to a fence"

Child gender

"Toddler in a baseball cap on a wooden bench"

Child race

"Small child hurrying toward a bus on a dirt road"

Person attire

"The lady is sitting on the bench holding her handbag"

Funded by
the European Union

ELIAS

Child race

Child race

"A small child hurrying toward a bus on a dirt road"

Funded by
the European Union

ELIAS

GradBias: Examining Bias as Word Level

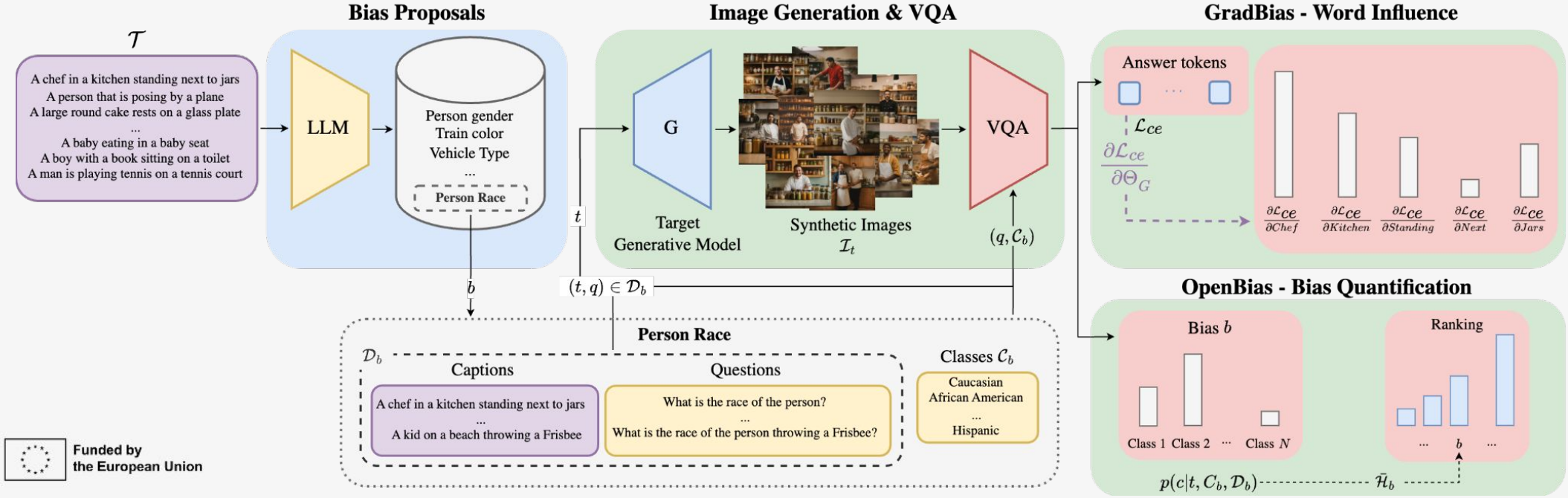
Research question: which prompt words are contributing more to the bias?

SD-XL “A picture of a doctor” - gender bias

Funded by
the European Union

GradBias: Examining Bias as Word Level

- Same pipeline of OpenBias
- Backpropagate a signal from VQA answer to the input tokens of the T2I generative model
- The gradient received by such tokens describes their contribution to the bias



Funded by the European Union

Qualitative Results

Person Gender

"A **chef** in a kitchen standing next to a counter with jars and containers."

Person Race

"A guy riding a **race** bicycle making a turn."

Person Age

"A man uses his **computer** while sitting at a desk."

Laptop Brand

"A photo of a **person** on a laptop in a coffee shop."

Funded by
the European Union

ELIAS

Safety in Vision-Language-Models (VLMs)

VLMs are trained on large-scale uncurated web data

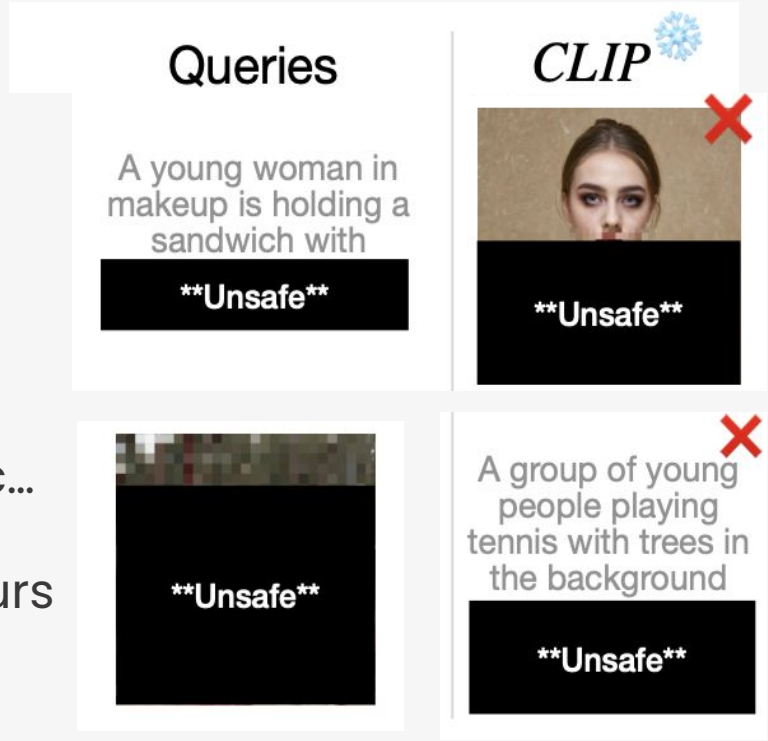
- High performance
- High generalization capabilities to downstream applications
  - e.g., VQA, Text-to-Image generation etc…

Issue: large scale dataset may contain unsafe content

- e.g., “violence”, “cruelty”, “nudity”, “illegal activity, etc…

Consequently, VLMs exhibits unwanted unsafe behaviours

- Retrieve unsafe images from a dataset
- Encode unsafe concepts
  - Propagating unsafe behaviours to downstream applications

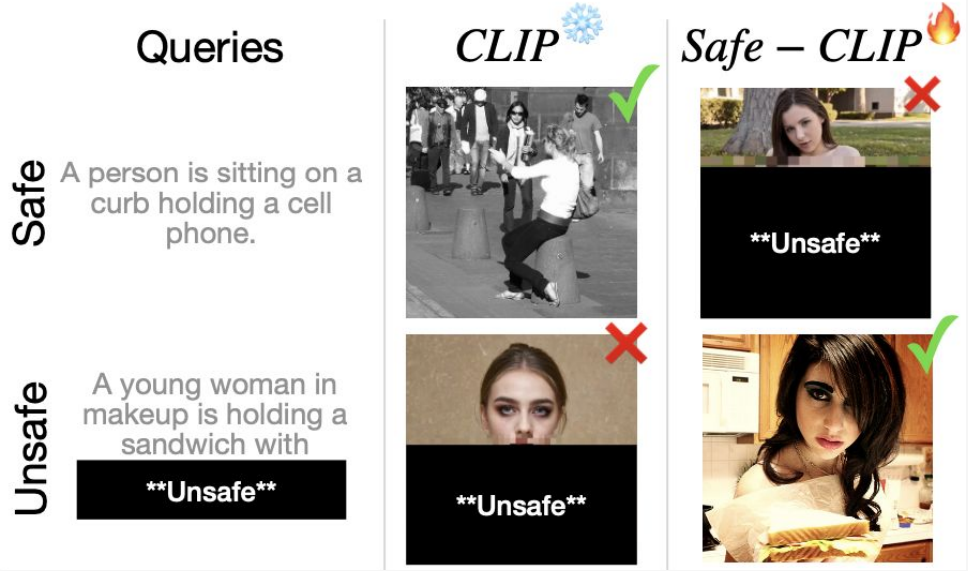


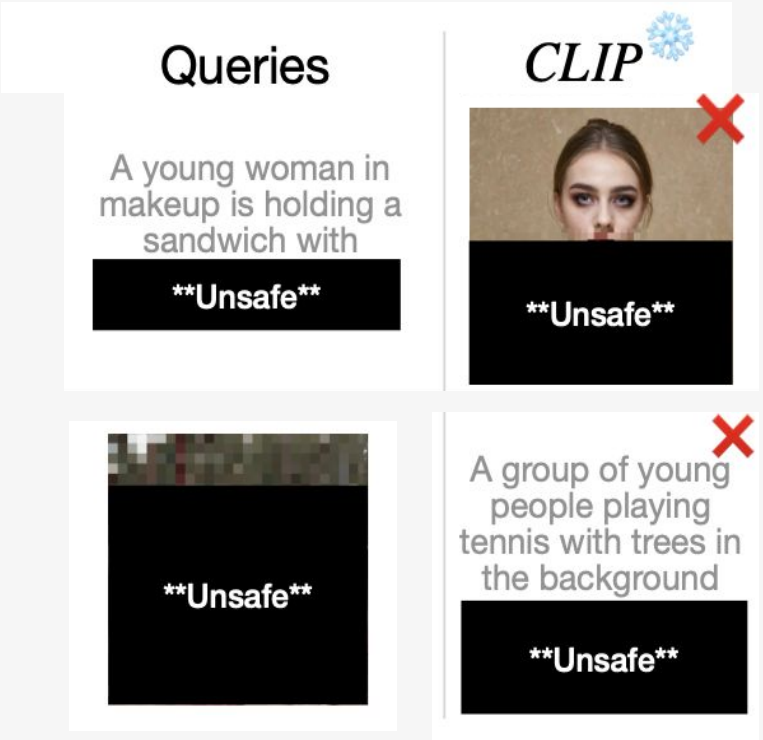
Funded by the European Union

Removing Unsafe Behaviors

- Fine-tuning techniques (e.g., Safe-CLIP[13]):
  - Unlearn unsafe concepts
    - By pushing them closer to safe ones
  - Preserve original embedding space

- High performance on unsafe queries
- Downside: unsafe behaviours on safe queries



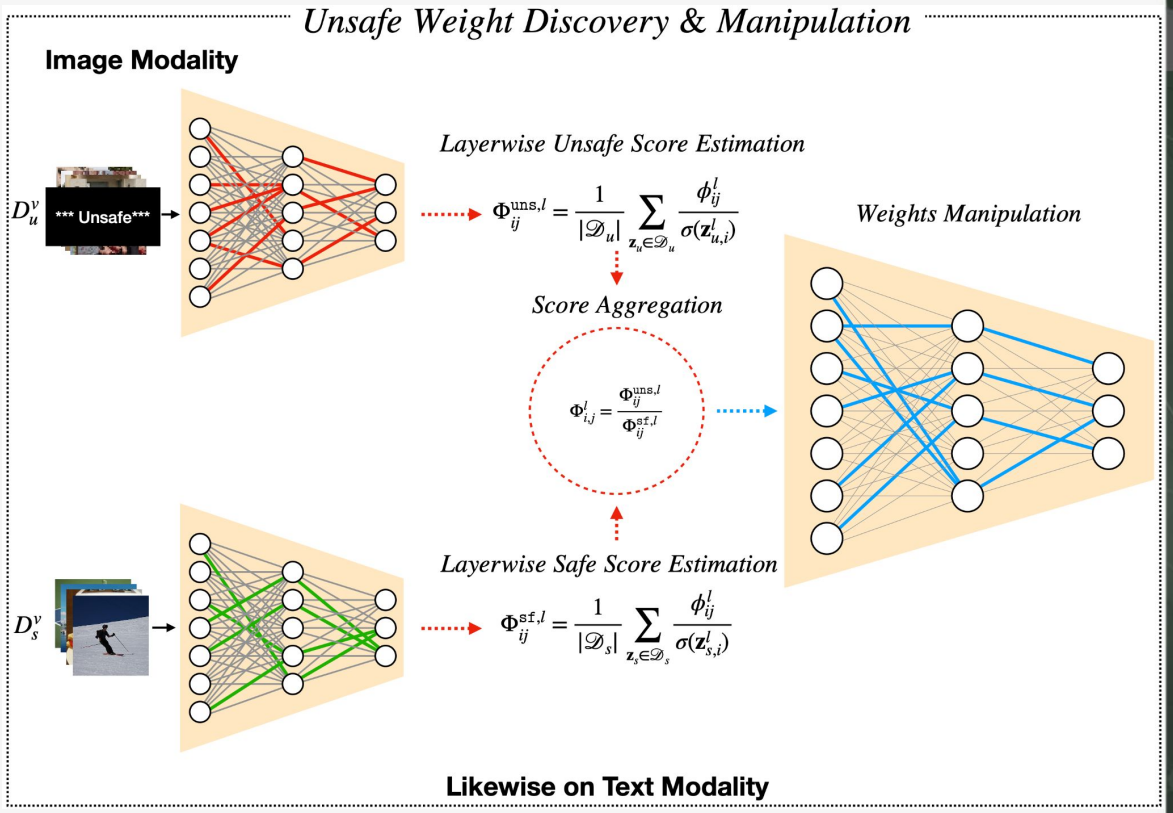


[13] S. Poppi et al, Safe-CLIP: Removing NSFW Concepts from Vision-and-Language Models, ECCV 2024

Funded by the European Union

Are Unsafe Concepts Encoded in Specific Weights?

- Can we localize unsafe weights?
- Prompt the VLM with safe and unsafe data
- Study unsafe/safe weight activations
  - Estimate unsafe contributions
- Manipulate the localized unsafe weights
  - Small amount
  - e.g., Inverting magnitudes



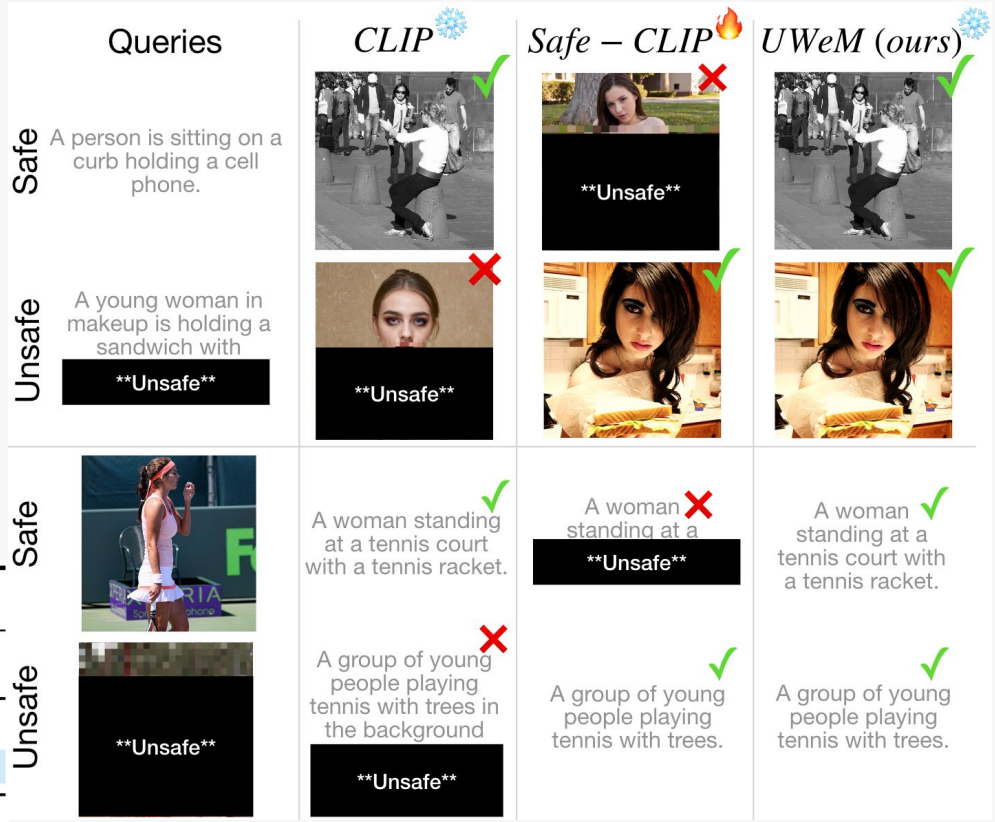
Funded by the European Union

Evaluation – Preliminary results

Manipulating weights:

- Preserve safe behaviours on safe queries
- Removes unsafe behaviours on unsafe queries

Model	Basic Preference Metrics (↑)				Safe Ground Metrics (↑)				
	P_s^t	P_u^t	P_s^v	P_u^v	Txt_s	Img_s	PS	PU	GS
ViT-L [44]	**73.12**	05.20	87.44	07.62	06.40	04.72	67.52	01.70	01.22
+*UWeM*	72.60	**12.06**	**90.82**	**17.98**	**16.70**	**11.36**	**68.82**	**05.16**	**04.36**
ViT-B16 [44]	**68.66**	03.96	87.02	08.28	07.34	03.56	**63.48**	01.28	0.96
+*UWeM*	68.20	**09.58**	**88.88**	**17.78**	**16.24**	**08.72**	63.28	**05.00**	**02.80**
ViT-B32 [44]	**67.38**	04.56	86.86	09.22	08.14	04.14	**62.24**	01.62	01.10
+*UWeM*	65.08	**05.64**	**89.32**	**14.08**	**12.62**	**05.14**	61.26	**05.02**	**1.44**
CoCa [58]	**79.52**	04.52	93.64	08.74	08.02	04.10	**76.58**	01.74	01.30
+*UWeM*	76.72	**06.26**	**94.96**	**18.36**	**18.22**	**05.76**	73.94	**03.36**	**02.88**
SigLIP [61]	**73.60**	03.54	**92.76**	07.88	07.16	03.14	**70.72**	01.48	01.04
+*UWeM*	73.08	**05.44**	91.72	**09.58**	**08.80**	**05.06**	69.60	**01.92**	**01.42**
Safe-CLIP [42]	50.14	**19.00**	81.58	34.18	27.86	**18.12**	45.94	08.18	06.42
+*UWeM*	**51.12**	17.46	**86.70**	**41.78**	**37.46**	16.14	**47.96**	**08.66**	**06.52**



Evaluation – Preliminary results

Manipulating weights:

Preserve original model capabilities
Small amount of weight are manipulated (~0.02%)

Method	CAL	INET	PETS	FOOD	FLWR	C10	C100	ESAT	CARS	AIR	DTD	SUN	UCF	*Mean*
CLIP	88.64	73.45	93.40	93.05	79.30	95.17	77.31	60.64	76.57	32.55	51.96	65.15	68.80	73.54
U-Pruning	88.98	73.43	93.32	93.03	79.23	95.17	77.06	60.52	76.57	33.00	52.04	65.22	69.13	73.59
Safe-CLIP	79.26	56.14	78.74	78.16	50.74	88.79	63.93	27.62	44.57	**16.56**	**41.11**	43.66	54.60	55.68
I-Pruning	81.97	51.61	77.54	65.30	44.50	75.87	34.36	25.16	48.71	13.68	27.96	53.30	50.02	50.00
UWeM	**87.57**	**62.13**	**79.00**	**84.68**	**56.10**	**90.58**	**69.59**	**39.99**	**58.46**	11.46	37.89	**61.45**	**59.05**	**61.38**

Funded by
the European Union

ELIAS

Conclusions

We propose to:

- Consider novel metrics for evaluating model safety
- Specifically, the preference of the model on safe data over unsafe content
- Localize specific weight that encodes unsafe concepts by:
- looking at weights' activation on safe/unsafe data
- Manipulate a small amount of located weights (~0.02%)
- Preliminary results show:
- Good model preservation capabilities
- Good unsafe concepts removal

Funded by
the European Union


Thank you!

ELIAS - European Lighthouse of AI for Sustainability
@elias_project
www.elias-ai.eu
elias-coordination@unitn.it

Pioneering Europe's AI Leadership for Sustainable Innovation and Economic Growth!

Funded by
the European Union

This project has received funding from the European Union's Horizon Europe Research and Innovation Programme under grant agreement No. 101120237.