Poster No. 258
June 21, Wed (AM)

Paper

JUNE 18-22, 2023 CVPR VANCOUVER, CANADA

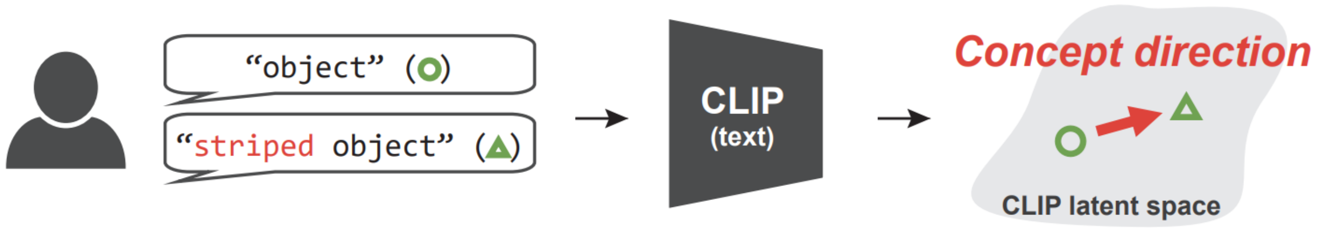
Grounding Counterfactual Explanation of Image Classifiers to Textual Concept Space

Siwon Kim Jinoh Oh Sungjin Lee Seunghak Yu Jaeyoung Do Tara Taghavi

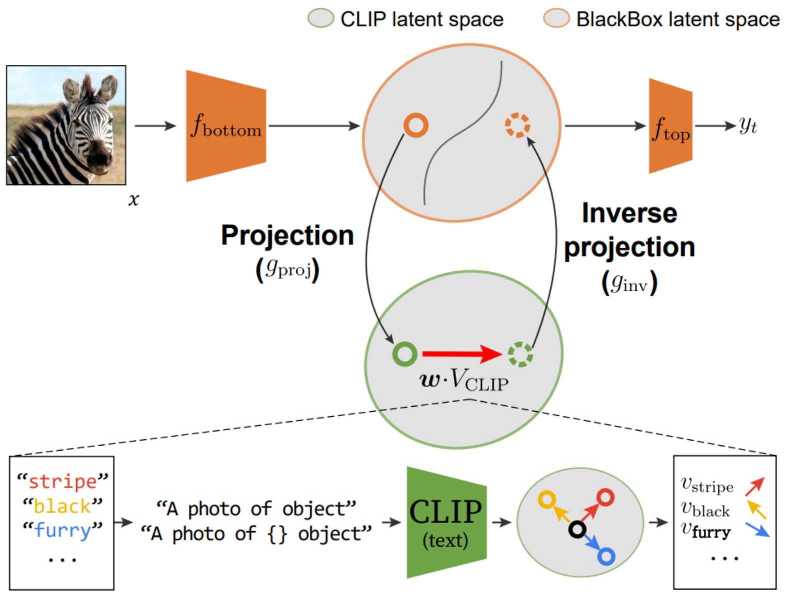
DSAIL Data Science & AI Laboratory
amazon
NAVER Search US

One-page Summary

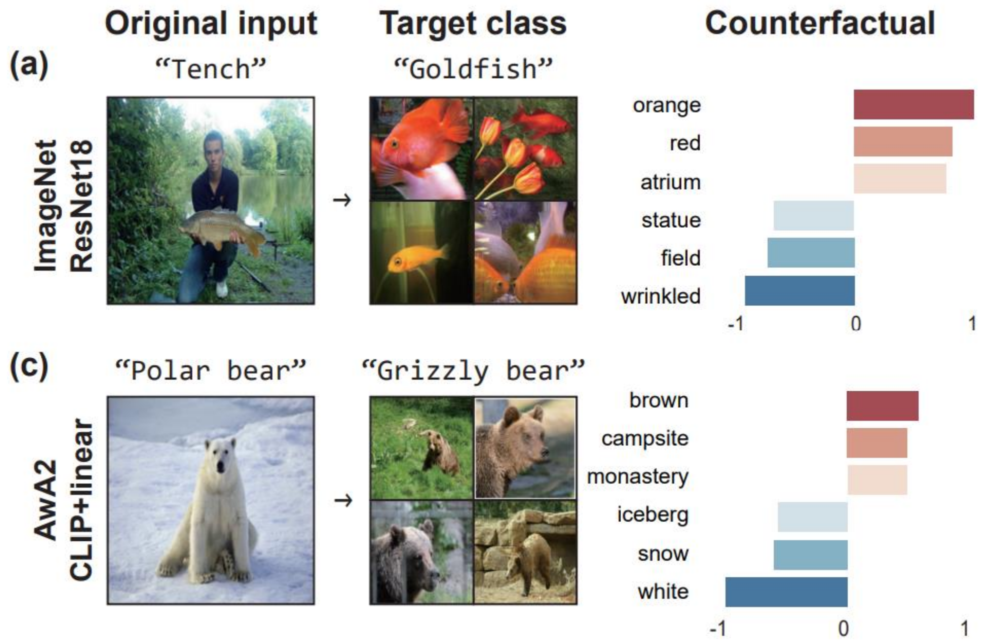
Motivation



Method



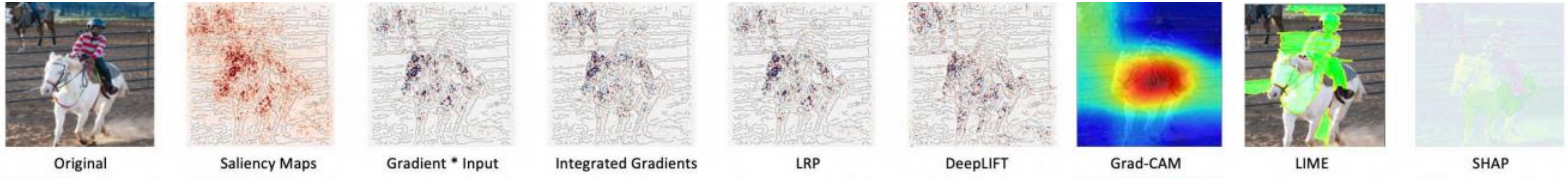
Results



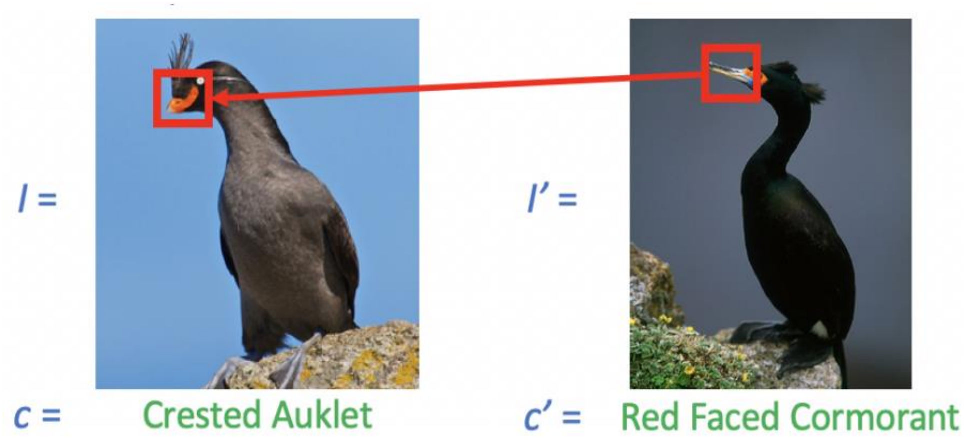
source:

Are existing explanations human-understandable?

- Attribution-based explanation



- Visual counterfactual explanation



→ Conventional visual XAI methods

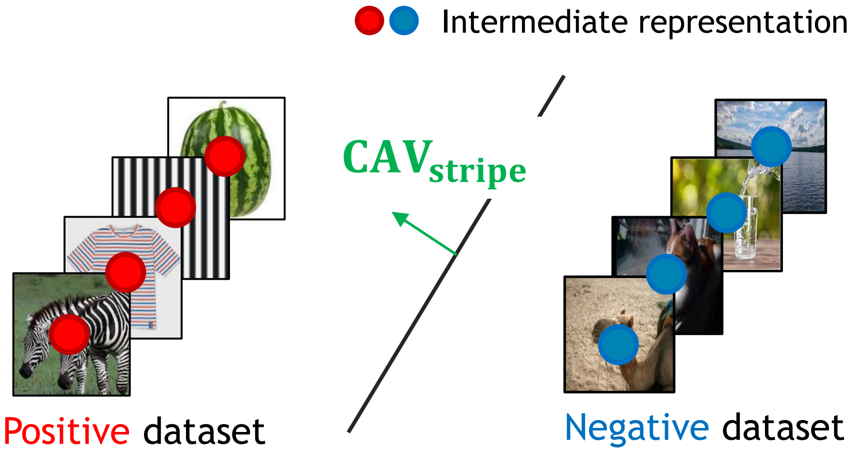
😊 Can: Provide important *regions*

🤔 Can't: Provide important *aspects* (e.g. color? pattern?)

source: Das, Arun, and Paul Rad. "Opportunities and challenges in explainable artificial intelligence (xai): A survey." *arXiv preprint arXiv:2006.11371* (2020).
Goyal, Yash, et al. "Counterfactual visual explanations." International Conference on Machine Learning. PMLR, 2019.

Towards human understandable explanation: Concept-based explanation

- Concept?
  - The units of human-understandable high-level semantics
  - Typically defined by words such as "stripe", "white", …
- Concept Activation Vector (CAV)
  - Positive dataset: samples that exhibit c
  - Negative dataset: samples that exclude c
  - CAV = A vector normal to the linear hyperplane

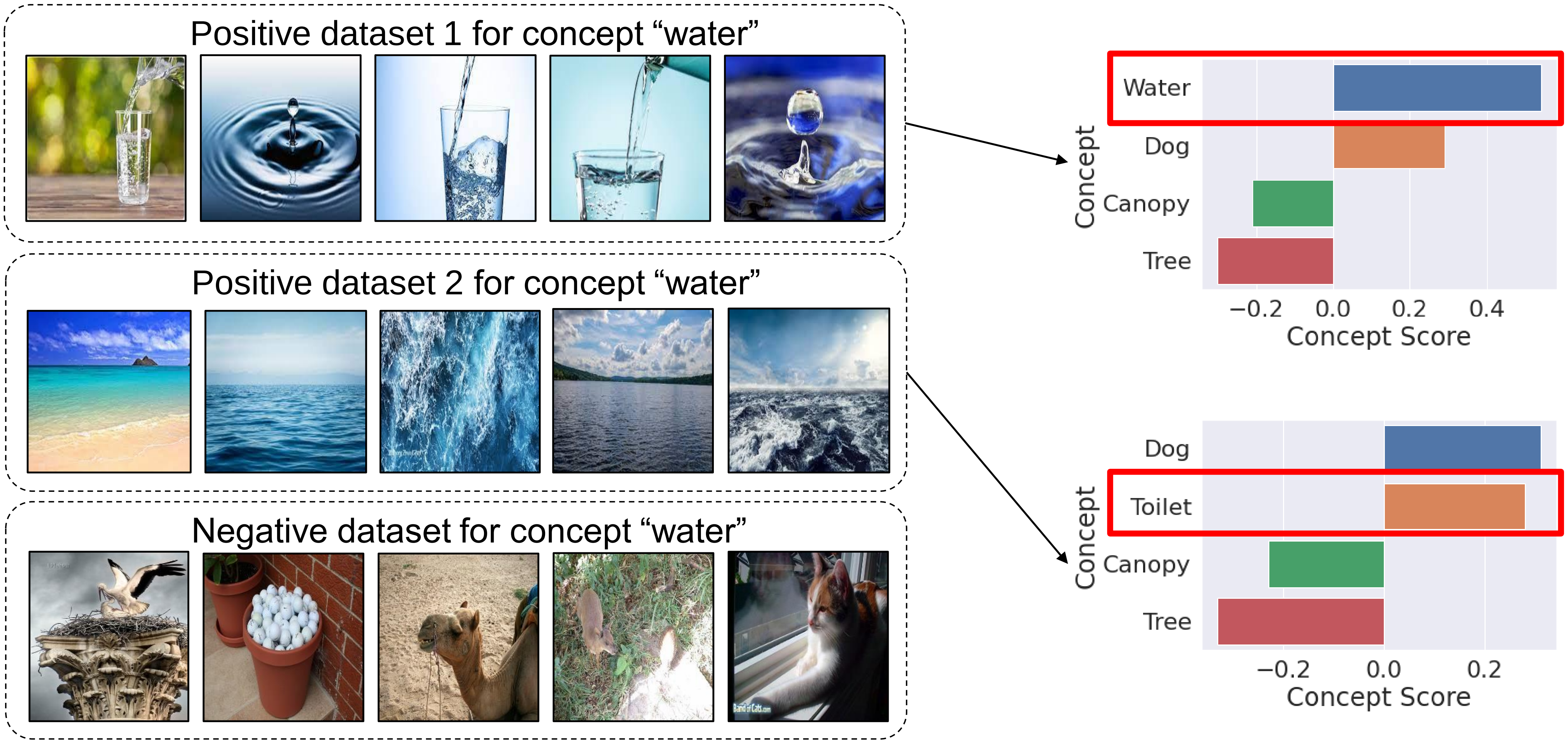


Example: "Stripe" concept

🤔 1. Diverging CAVs & 2. Unintended entanglement

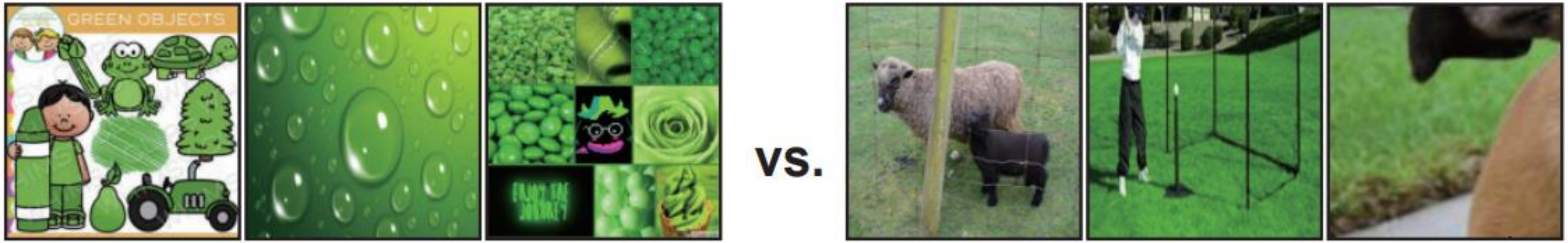
source: Kim, Been, et al. "Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav)." *International conference on machine learning.* PMLR, 2018

1. Diverging CAVs



source: Images were collected from google image search

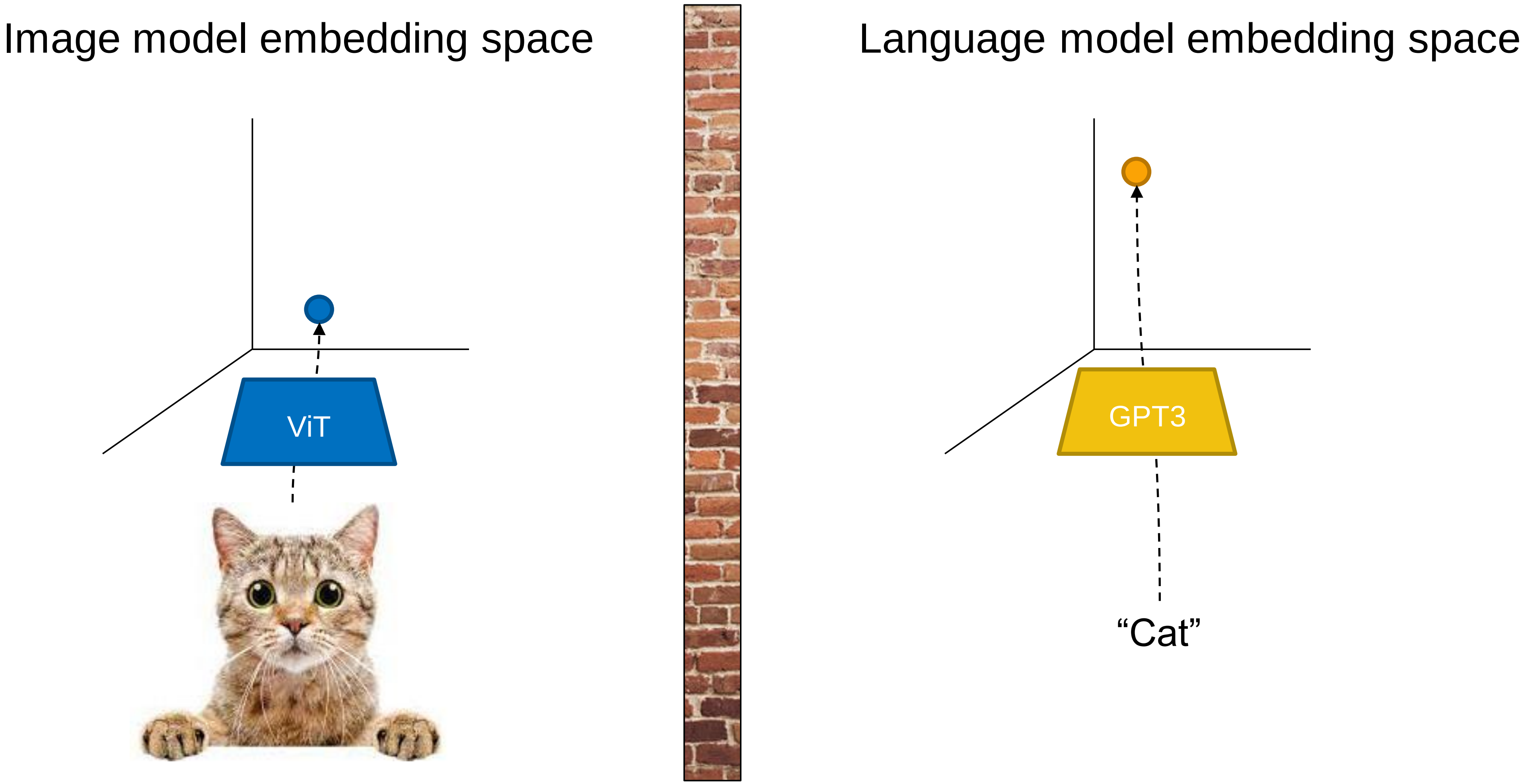
2. Unintended entanglement



Positive images for "Green" Positive images for "Grass"

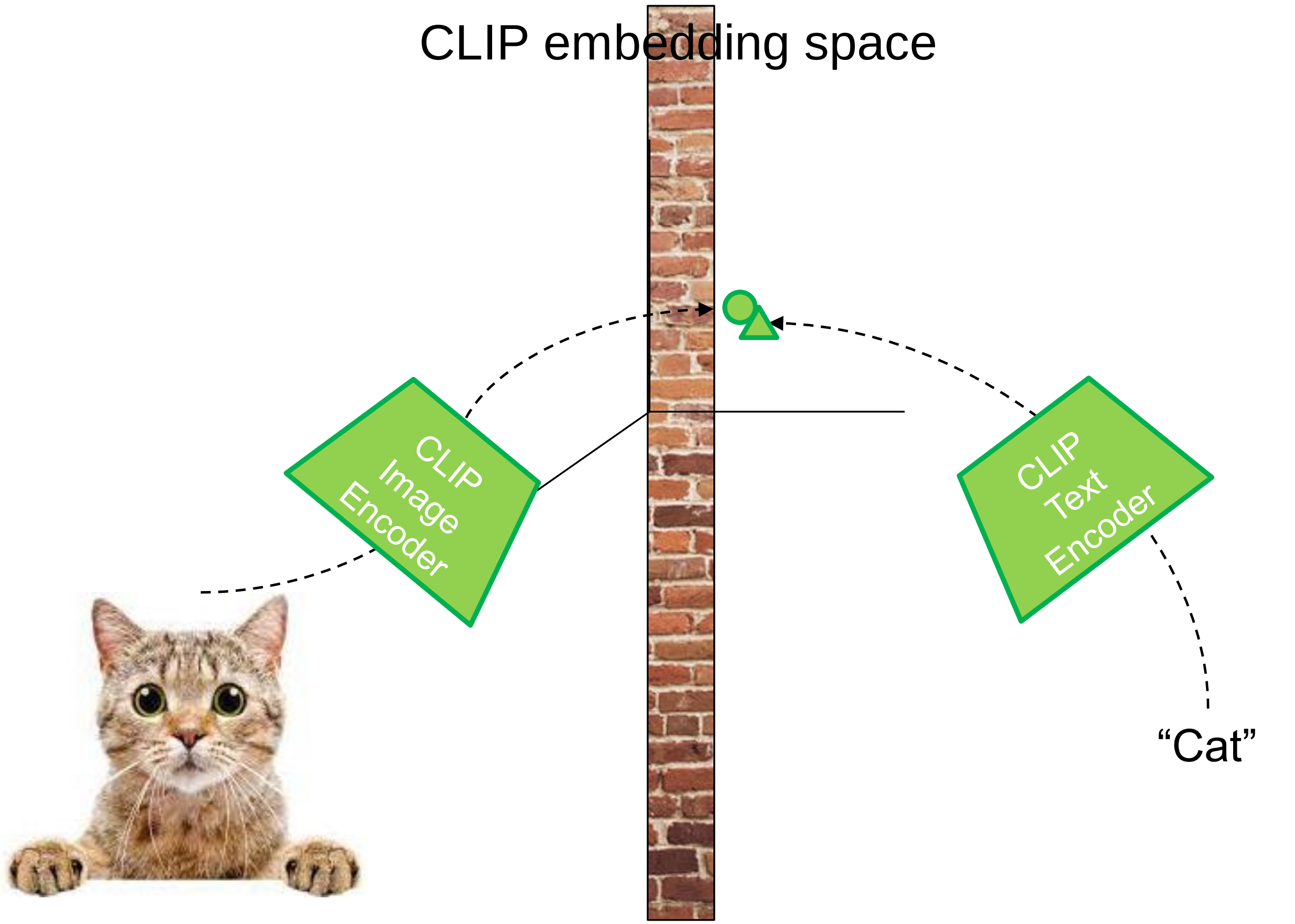
Counterfactual TEXtual EXplanation (CounTEX)

Vision (V) and Language (L) were Separated

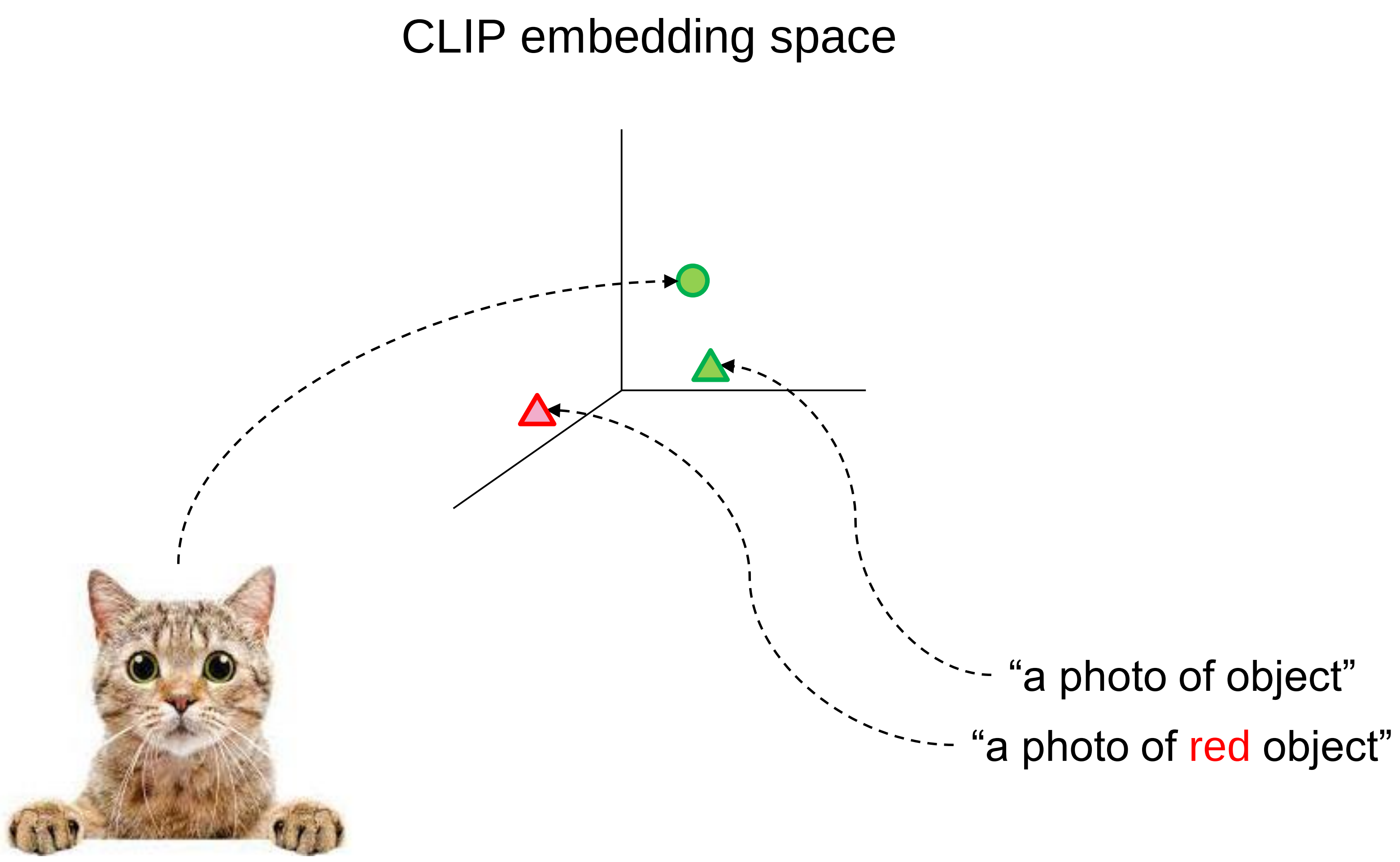


source:

CLIP enables V&L Joint Embedding Space



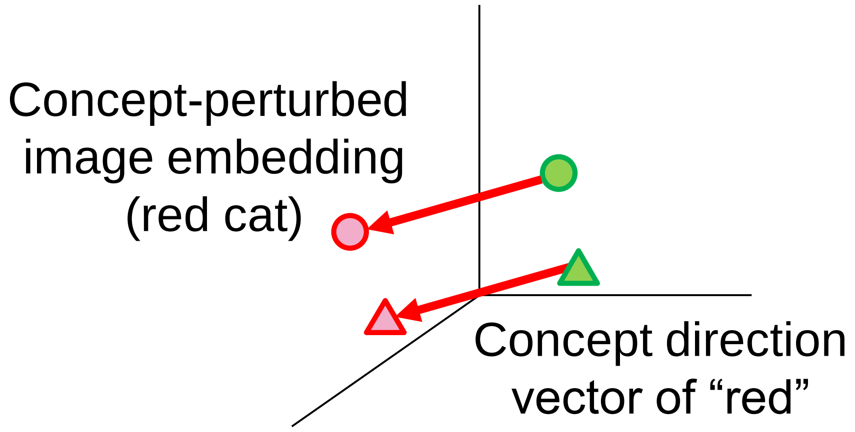
CLIP enables textual guidance on images



source:

CLIP enables textual guidance on images

CLIP embedding space



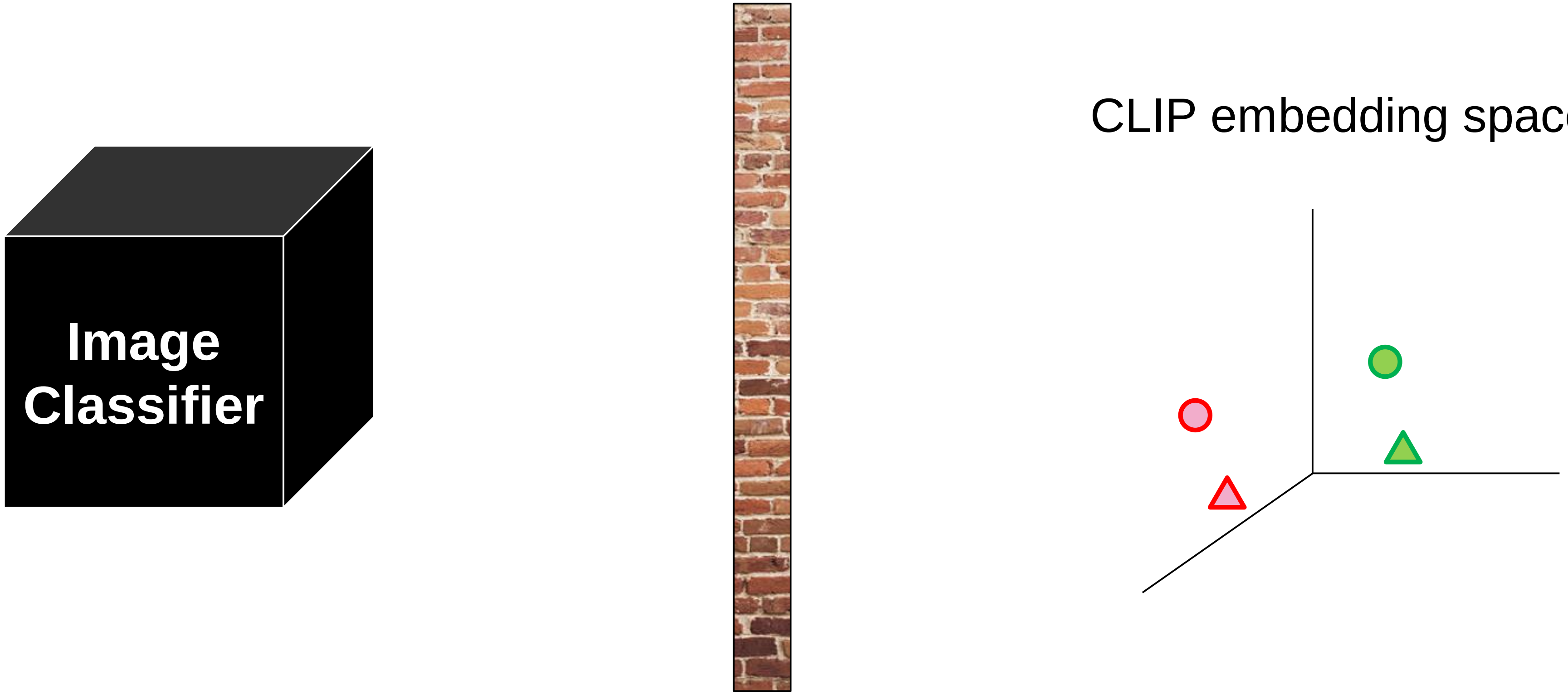
“a photo of cat”

“a photo of red cat”

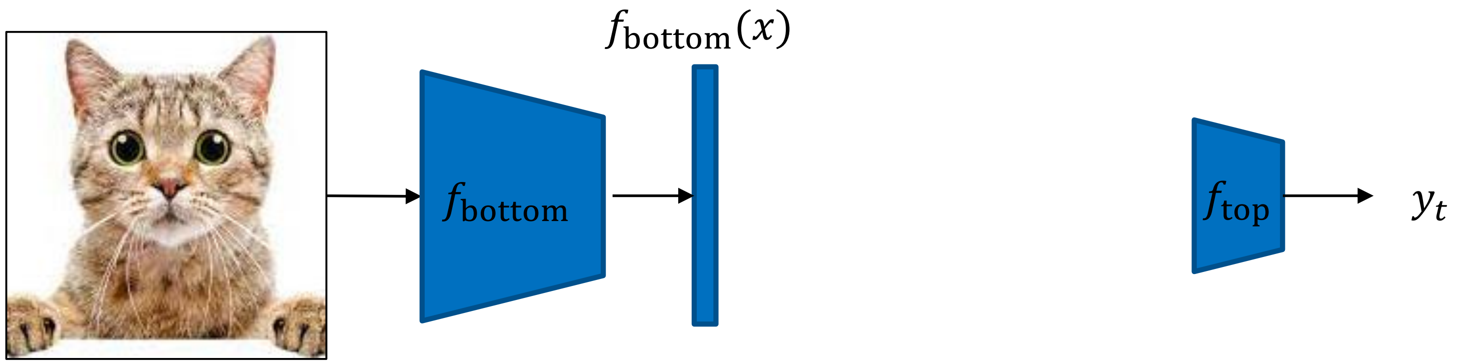
source:

But, the target classifier is not CLIP

- How can we leverage the well-performing CLIP latent space?



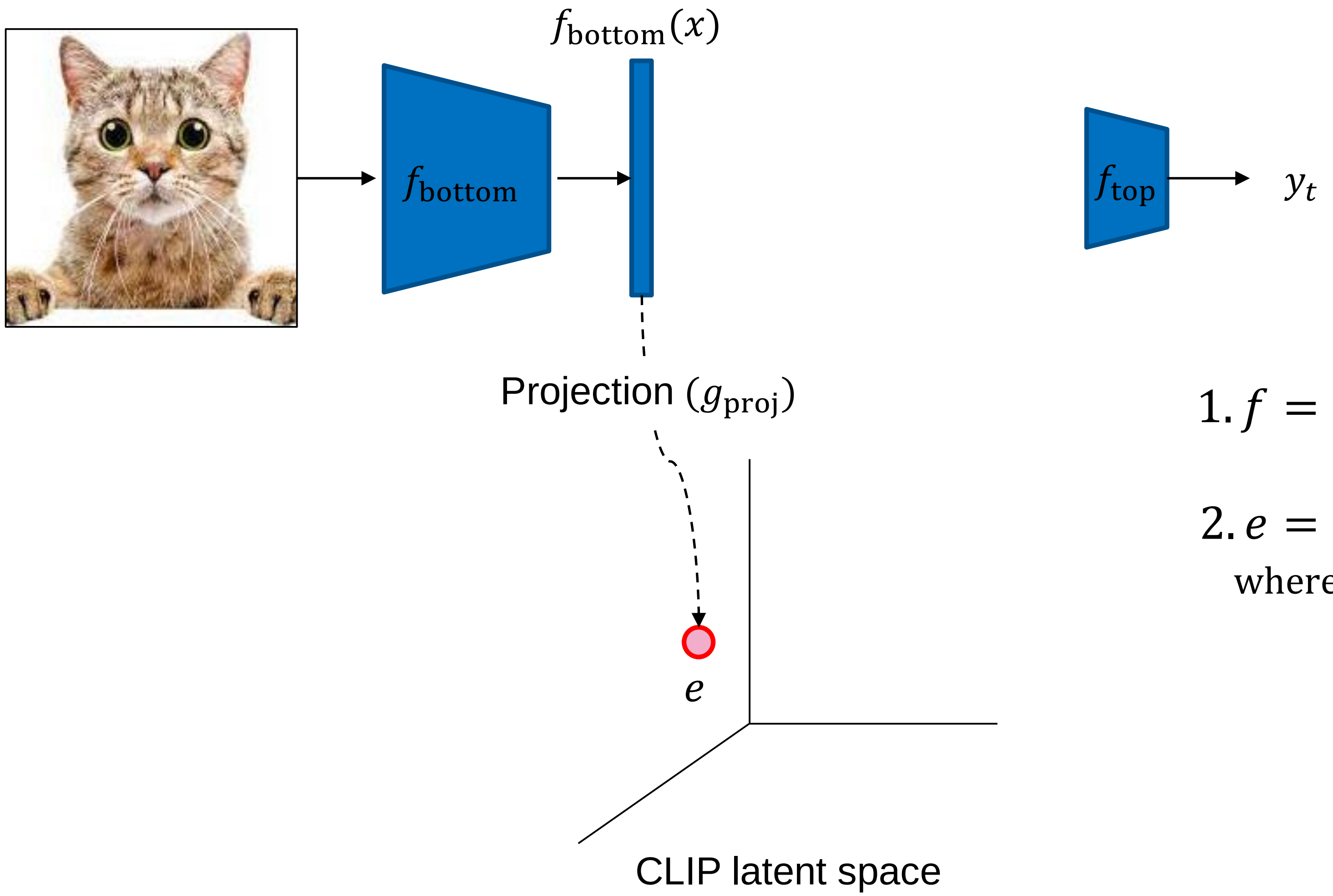
Counterfactul TEXtual EXplanation (CounTEX)



1. $f = [f_{\text{bottom}}, f_{\text{top}}]$

source:

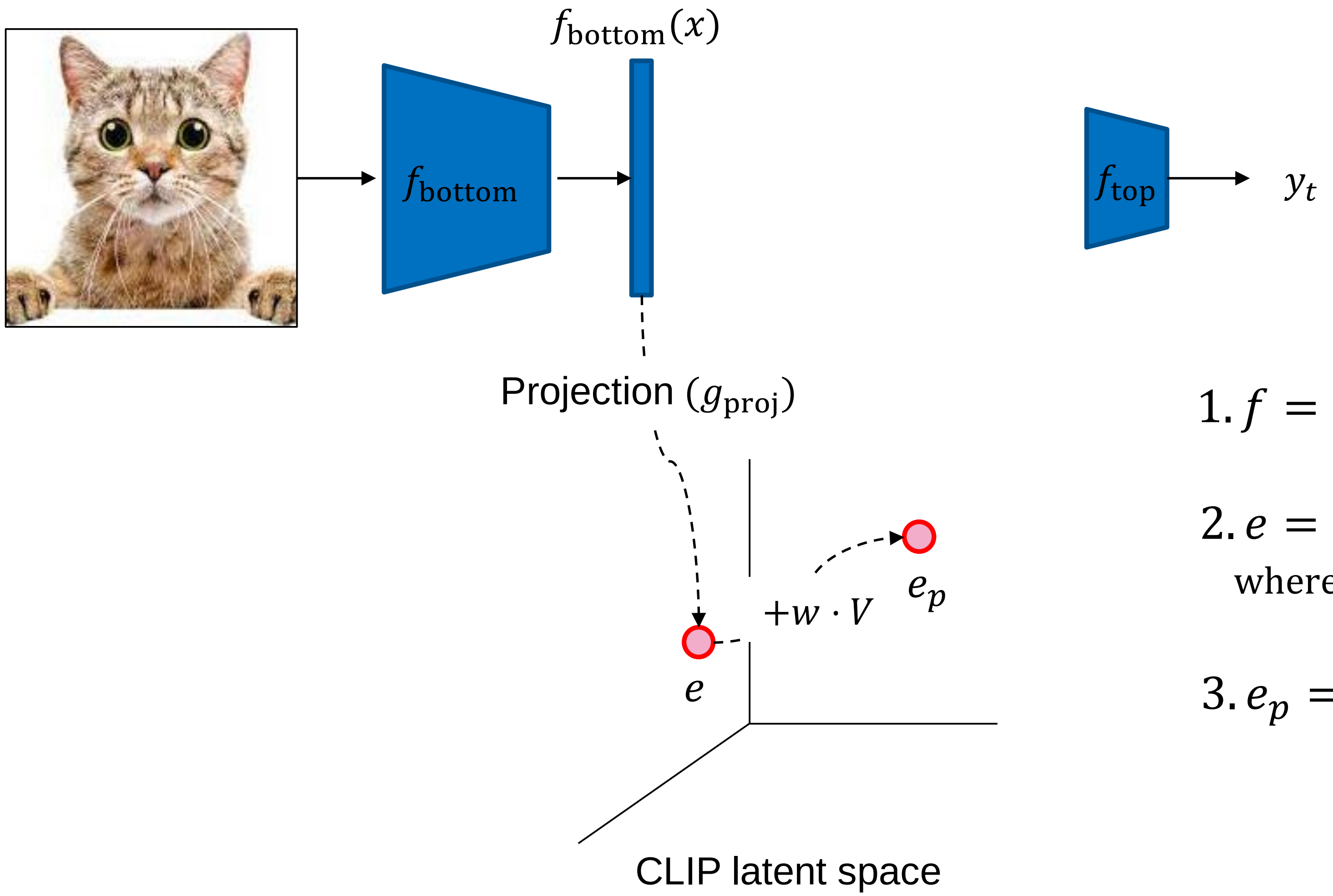
CounterfactulTEXtual EXplanation (CounTEX)



1. $f = [f_{\text{bottom}}, f_{\text{top}}]$

2. $e = g_{\text{proj}}\left(f_{\text{bottom}}(x)\right)$,
   where $g_{\text{proj}}(\cdot)$ is a projection function

Counterfactul TEXtual EXplanation (CounTEX)

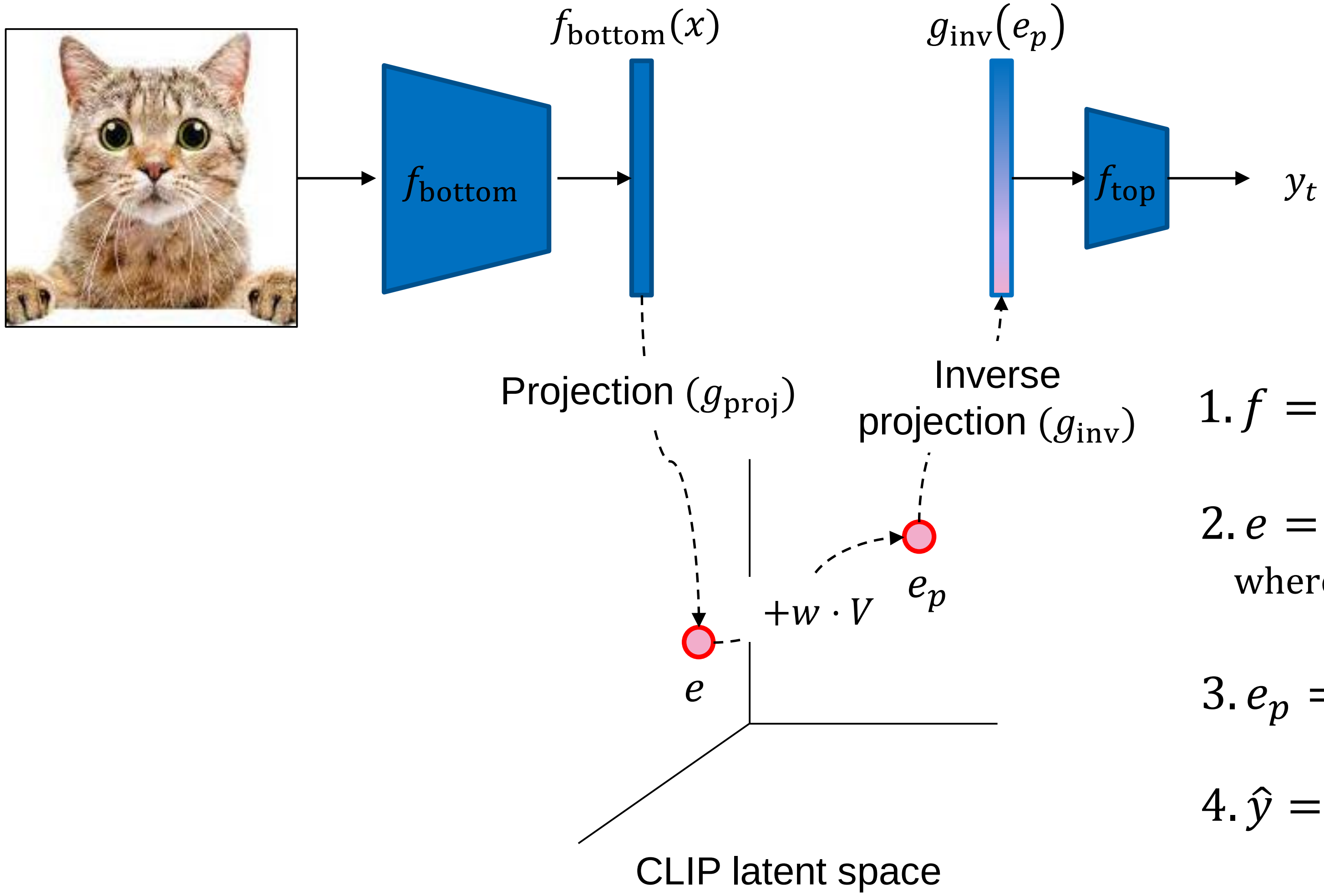


1. $f = [f_{\text{bottom}}, f_{\text{top}}]$

2. $e = g_{\text{proj}}(f_{\text{bottom}}(x))$,
   where $g_{\text{proj}}(\cdot)$ is a projection function

3. $e_p = e + w \cdot V$

CounterfactulTEXtual EXplanation (CounTEX)



1. $f = [f_{\text{bottom}}, f_{\text{top}}]$

2. $e = g_{\text{proj}}(f_{\text{bottom}}(x))$, where $g_{\text{proj}}(\cdot)$ is a projection function

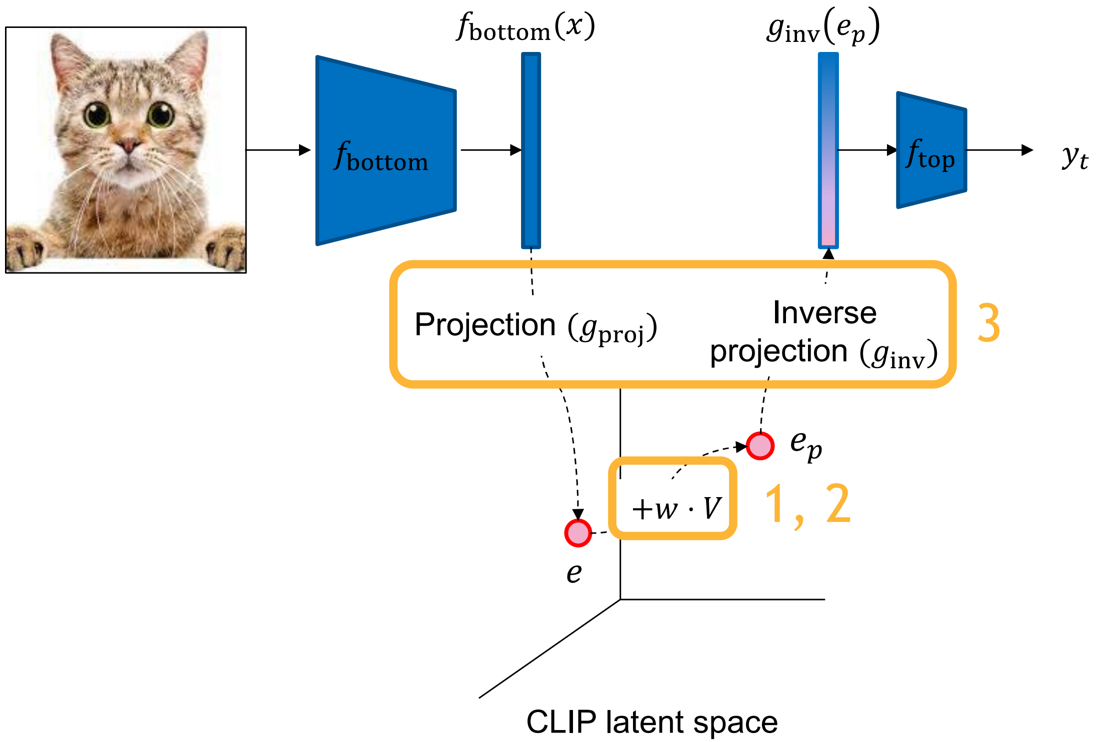
3. $e_p = e + w \cdot V$

4. $\hat{y} = f_{\text{top}}(g_{\text{inv}}(e_p))$

source:

Research Questions

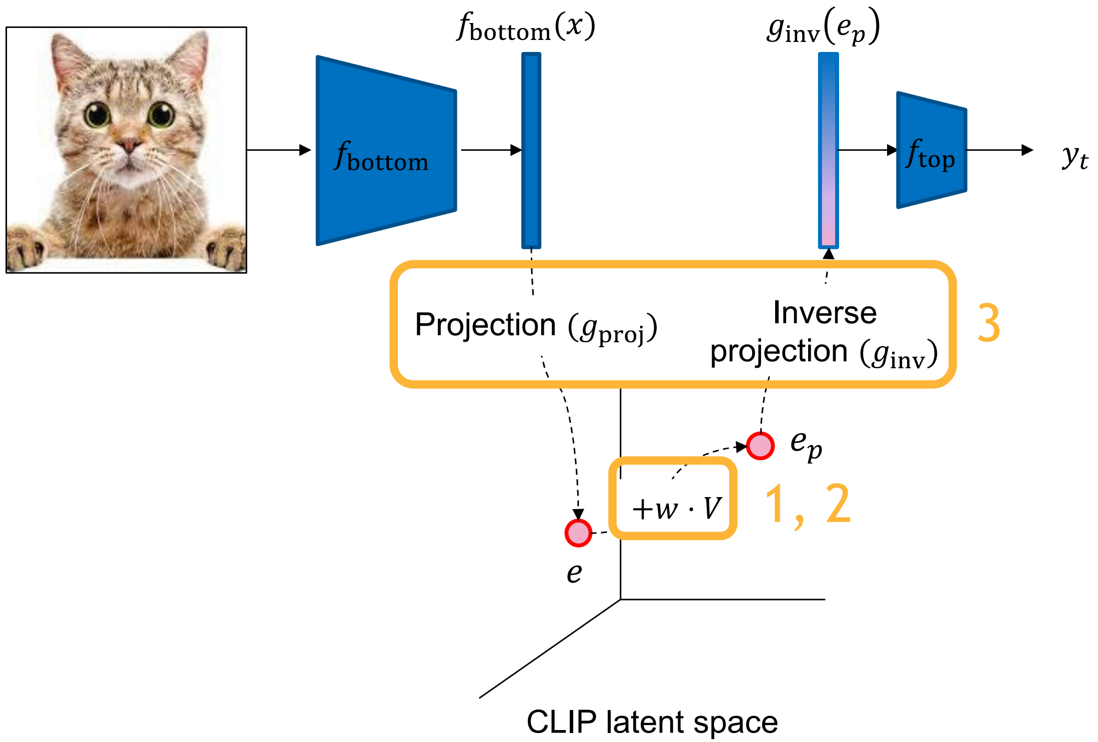
1. How can we obtain concept direction bank V?
2. How can we obtain weight w?
3. How can we implement projection and inverse projection?



source:

Research Questions

1. How can we obtain concept direction bank V?
2. How can we obtain weight w?
3. How can we implement projection and inverse projection?

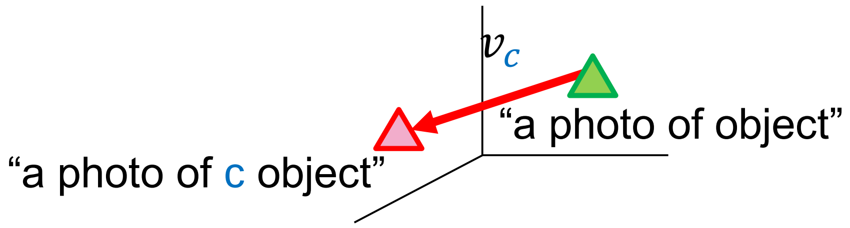


source:

1. Prepare concept direction bank

1. Build concept direction vector bank V
   - Given a predefined concept library C,
   - Generate prompt pair for concept c
     - $[t_{src}^{c}, t_{trg}^{c}] =$ ["a photo of object", "a photo of c object"]
   - $v_c = \text{MinMaxNormalize}(\text{CLIP}_{\text{text}}(t_{src}^{c}) - \text{CLIP}_{\text{text}}(t_{trg}^{c}))$
   - $V = \{v_c | c \in C\}$, where C is a predefined concept set

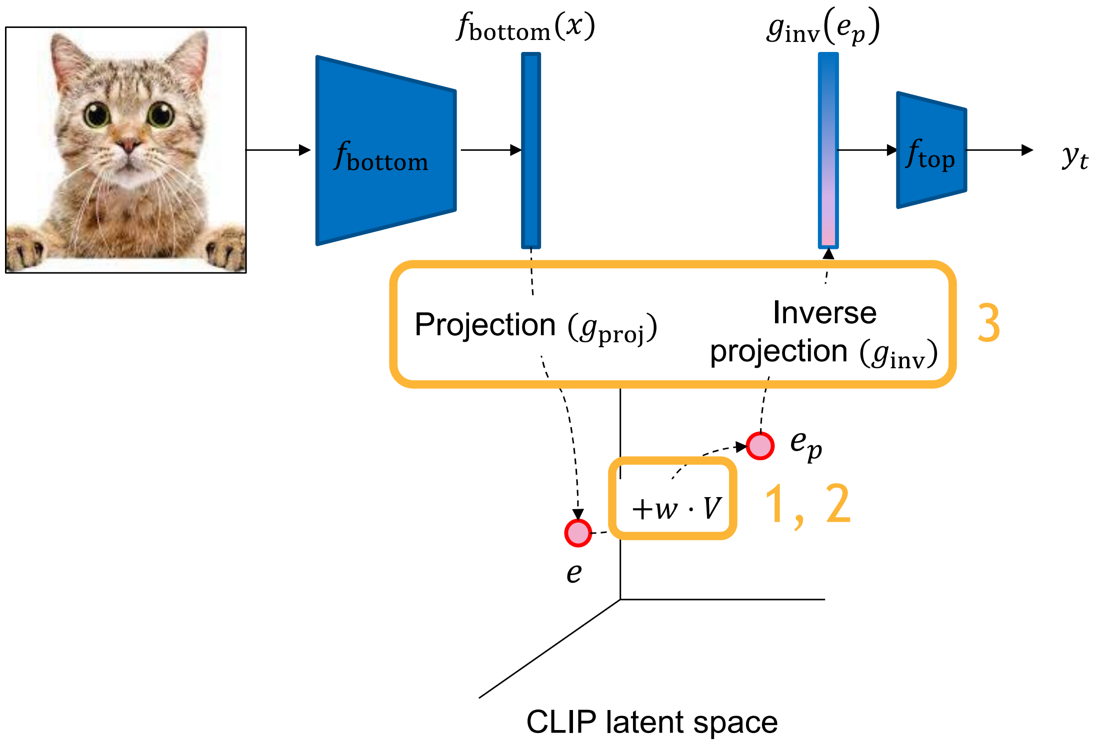
Category	Prompt template
Color	"A photo of {} object"
Texture	"A photo of {} object"
Scene	"A photo of object on {}"
Material	"A photo of object made of {}"
Part	"A photo of object containing {}"
Object	"A photo of object along with {}"



source:

Research Questions

1. How can we obtain concept direction bank V?
2. How can we obtain weight w?
3. How can we implement projection and inverse projection?



source:

2. Optimize w until the prediction changes to y_t

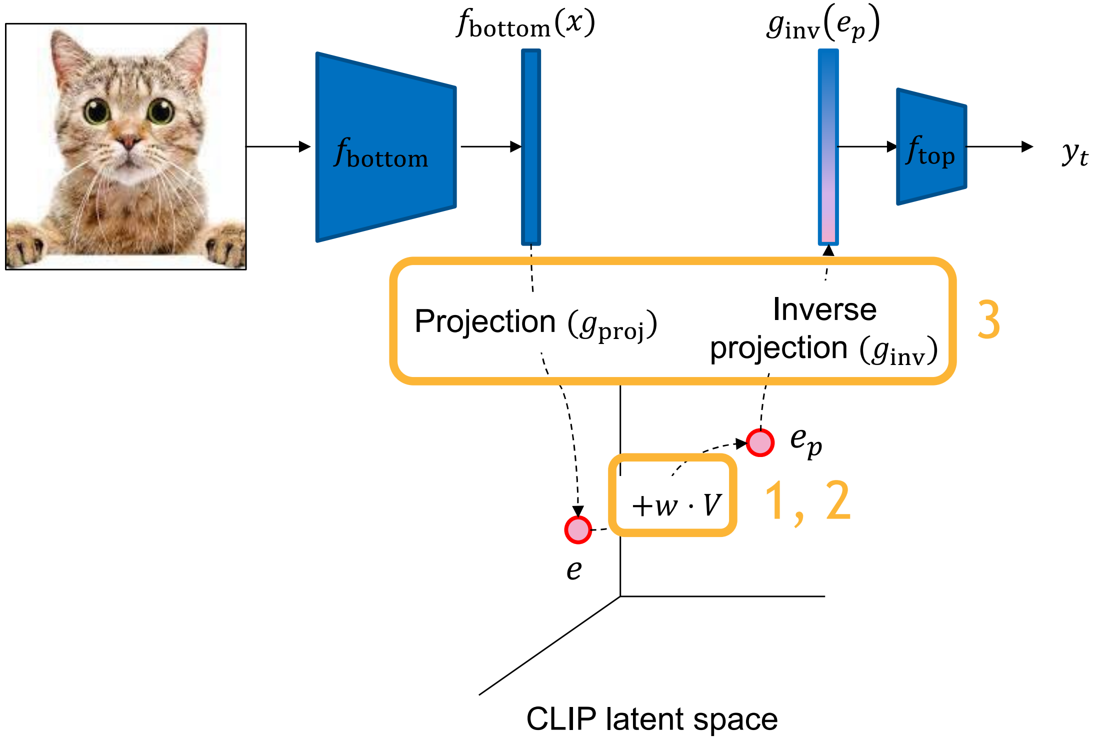
2. Optimize w with the objective function below

$$\min_{w} \mathcal{L} = \mathcal{L}_{\text{CE}} + \alpha \cdot \mathcal{L}_{\text{reg}} + \beta \cdot \mathcal{L}_{\text{id}}$$

- $\mathcal{L}_{\text{CE}} = \text{CrossEntropy}\left(f_{\text{top}}\left(g_{\text{inv}}(e_p)\right), y_t\right)$
  - Change the prediction to the target class by minimizing the cross-entropy loss
- $\mathcal{L}_{\text{reg}}$
  - Elastic net regularization: regularize concept importance to be 1) sparse and 2) unique minimum
- $\mathcal{L}_{\text{id}} = \left|\left|e_p - e\right|\right|^2$
  - Counterfactual approach requires the minimal modification
  - Identity loss to constrain the minimal perturbation

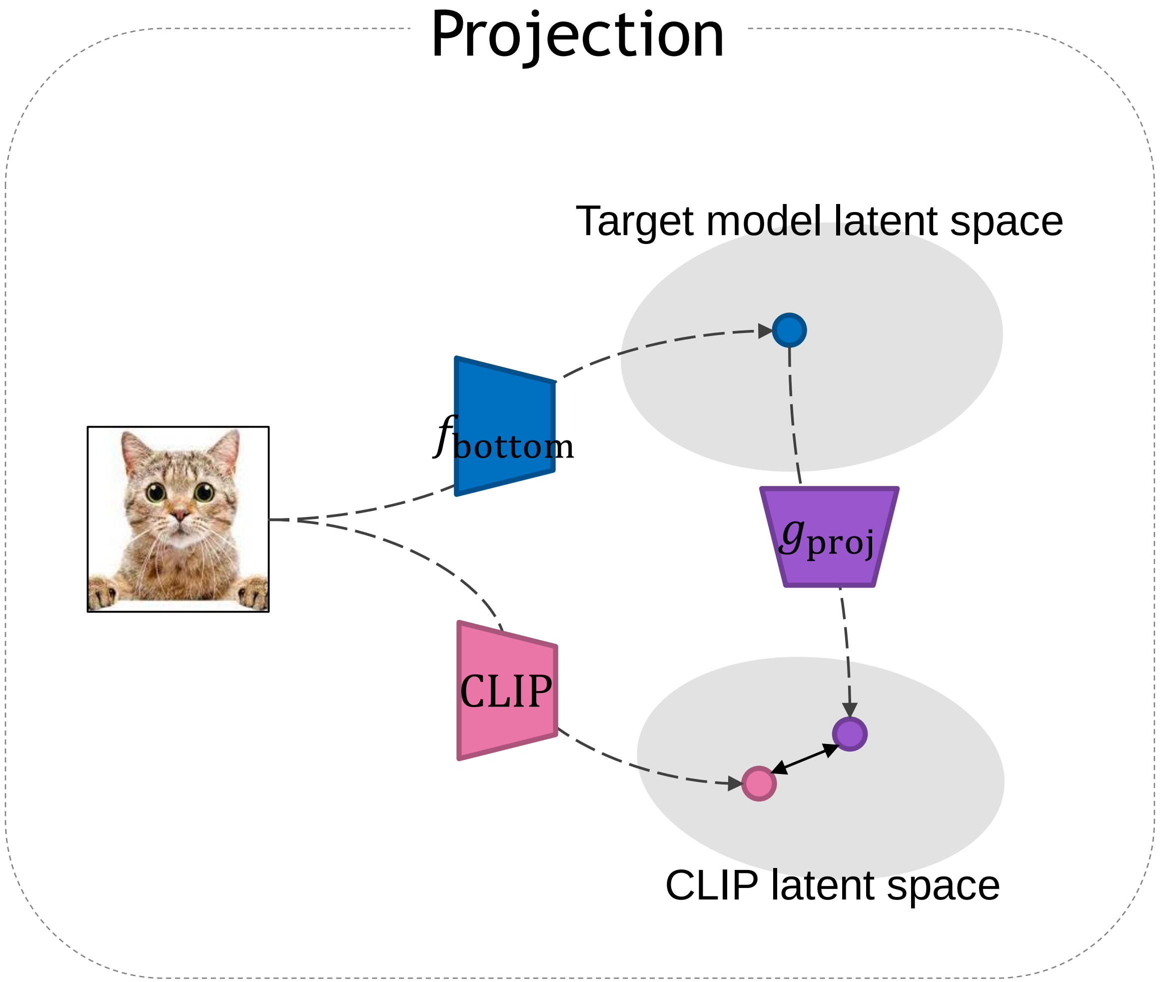
Research Questions

1. How can we obtain concept direction bank V?
2. How can we obtain weight w?
3. How can we implement projection and inverse projection?

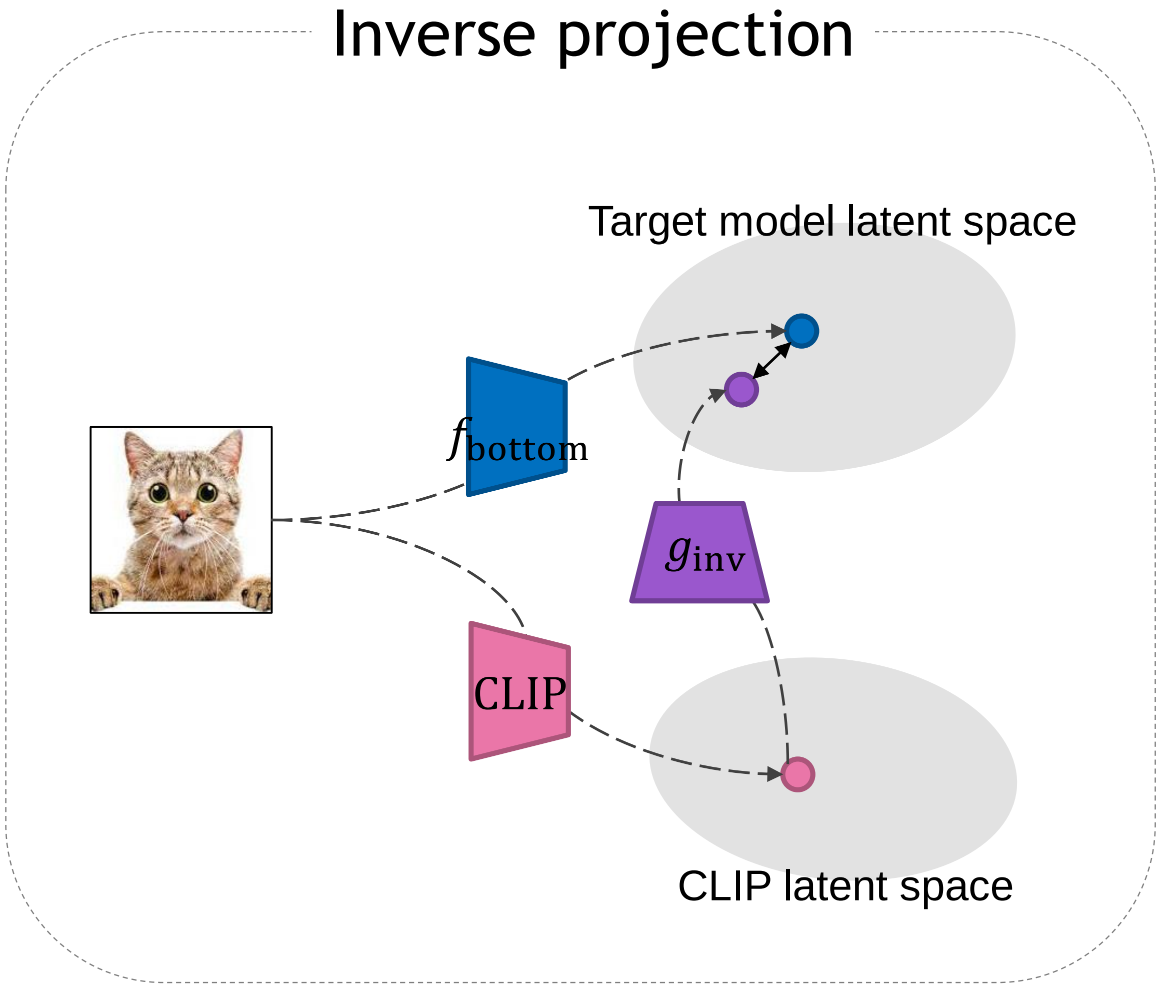


source:

3. Prepare projector and inverse projector



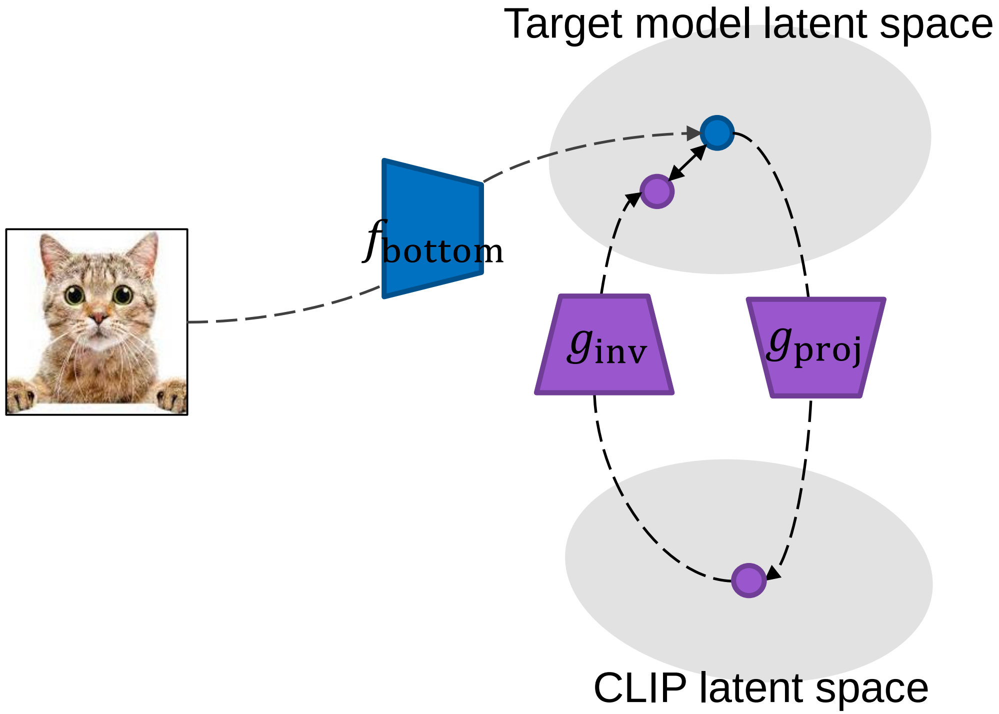
$$\mathcal{L}_{\text{proj}} = \left|\left| g_{\text{proj}}\left(f_{\text{bottom}}(x)\right) - \text{CLIP}(x) \right|\right|^2$$



$$\mathcal{L}_{\text{inv}} = \left|\left| g_{\text{inv}}\left(\text{CLIP}(x)\right) - f_{\text{bottom}}(x) \right|\right|^2$$

3. Prepare projector and inverse projector

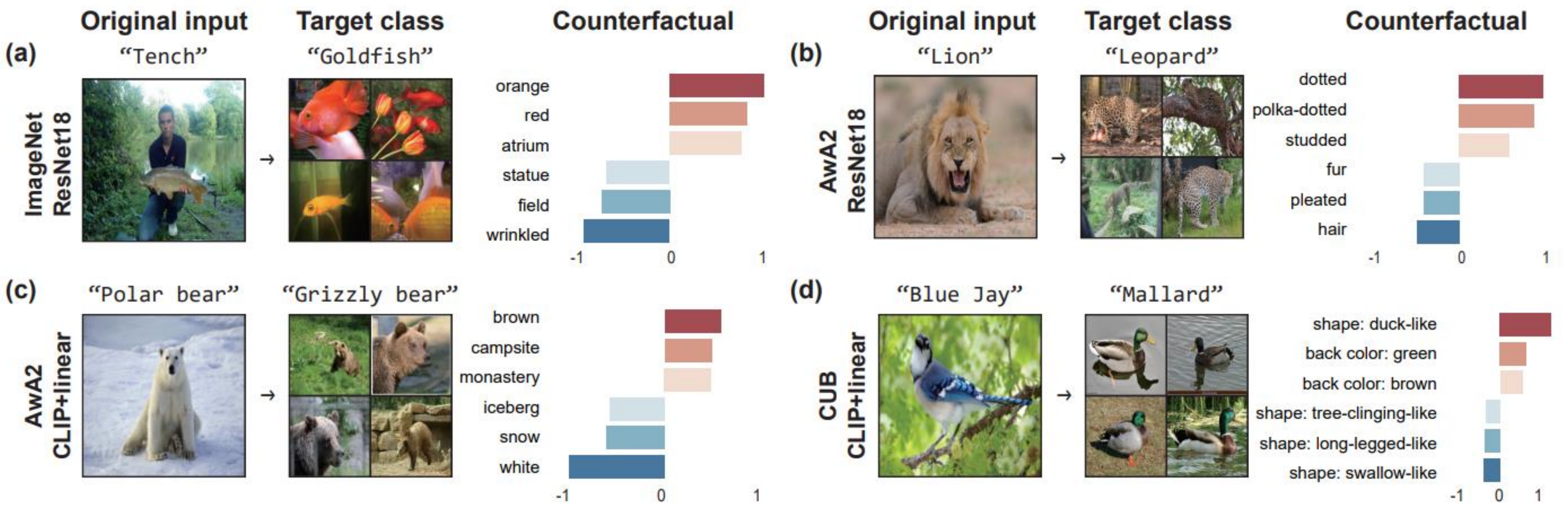
- Additionally finetune with cycle consistency loss
  - $\mathcal{L}_{\text{finetune}} = \mathcal{L}_{\text{proj}} + \mathcal{L}_{\text{inv}} + \mathcal{L}_{\text{cycle}}$
  - $\mathcal{L}_{\text{cycle}} = \left\| f_{\text{bottom}}(x) - g_{\text{inv}}\left(g_{\text{proj}}\left(f_{\text{bottom}}(x)\right)\right) \right\|^2$



Experimental setup

- Target model
  - CLIP+linear models
    - Trained on ImageNet/Animal with Attributes2 (AWA2)/Caltech-UCSD Birds-200-2011 (CUB)
    - Shares the embedding space with CLIP → Projection is not needed
  - ResNet18 models
    - Trained on ImageNet/AWA2/CUB
    - Does not share the embedding space with CLIP → Projection is needed
- Concept library C
  - `Reduce BRODEN` for ImageNet-trained models
  - `AWA2, CUB` for AWA2-trained and CUB-trained models, respectively

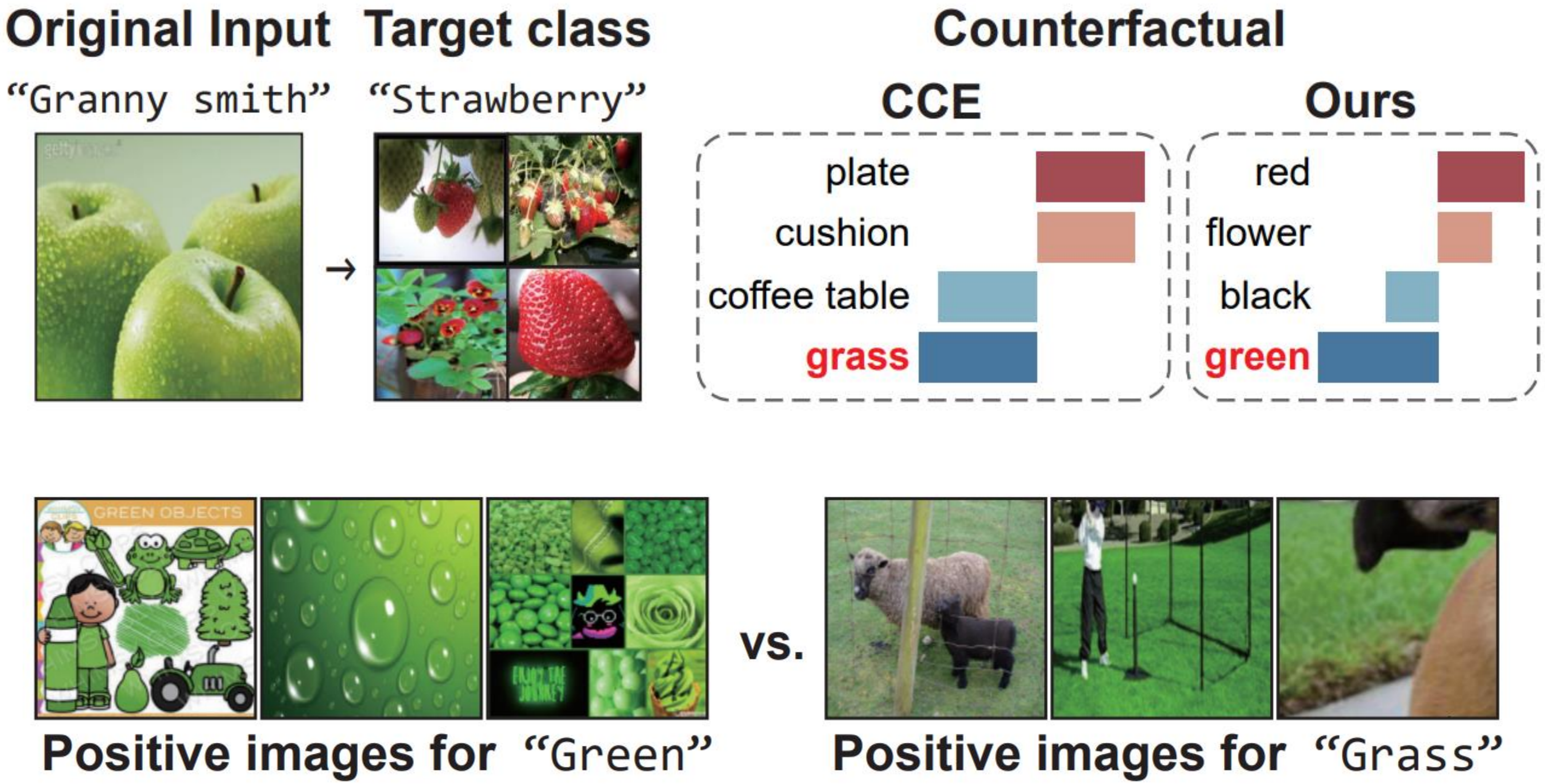
Qualitative evaluation



source:

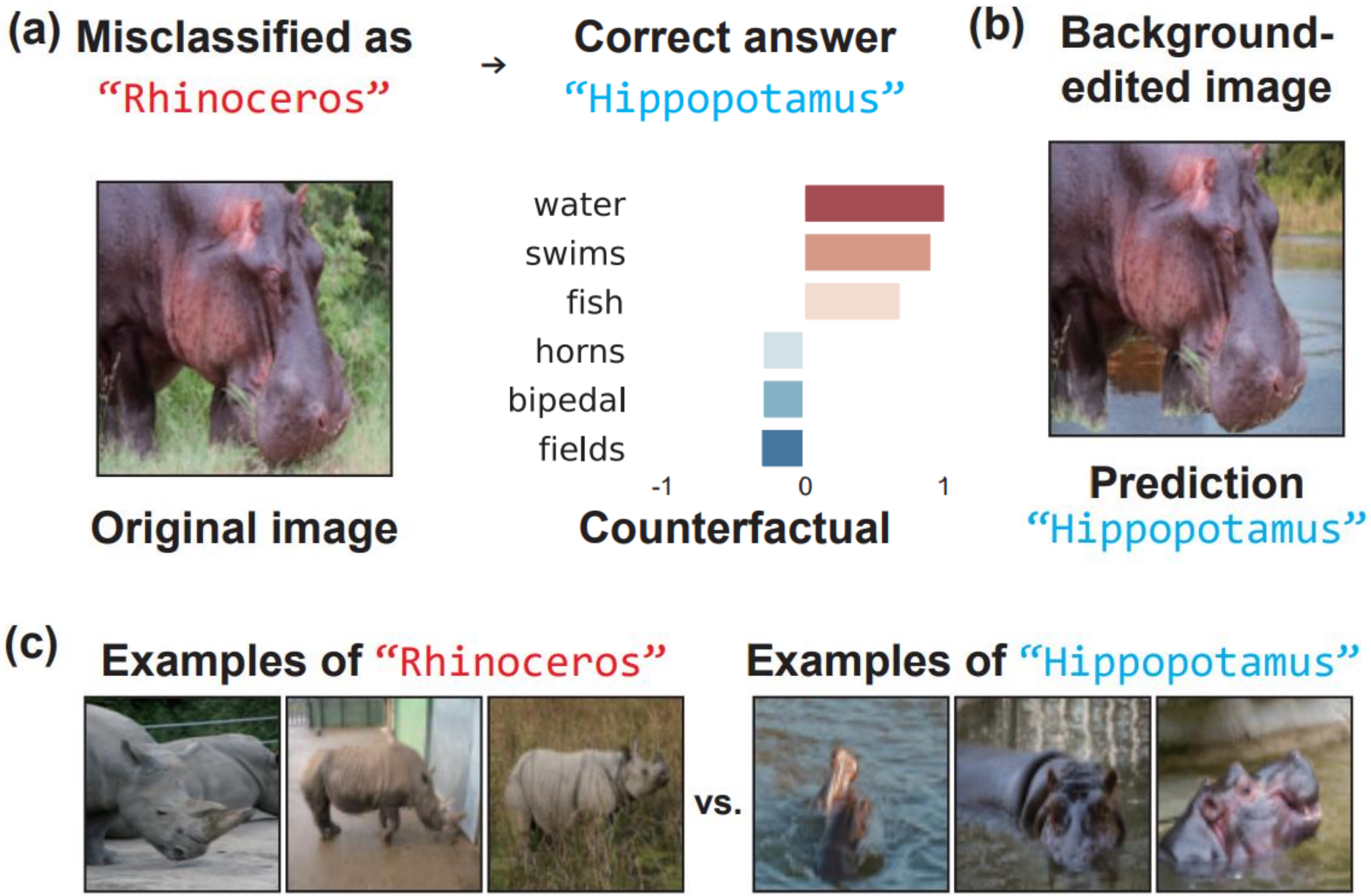
Qualitative Comparison between Ours and CCE

- Spurious correlation in dataset collection of CCE led to inaccurate interpretation



source:

Debugging misclassification cases



source:

Quantitative Evaluation

- Lack of ground truth explanation
  - Especially for conceptual explanation
  - Some previous works often skip quantitative evaluation

- Repurpose existing datasets with class-wise attributes
  1. AWA2 dataset
     - 85 binary attributes are provided for 50 animal classes
  2. CUB dataset
     - 312 continuous attributes are provided for 200 bird classes

source:

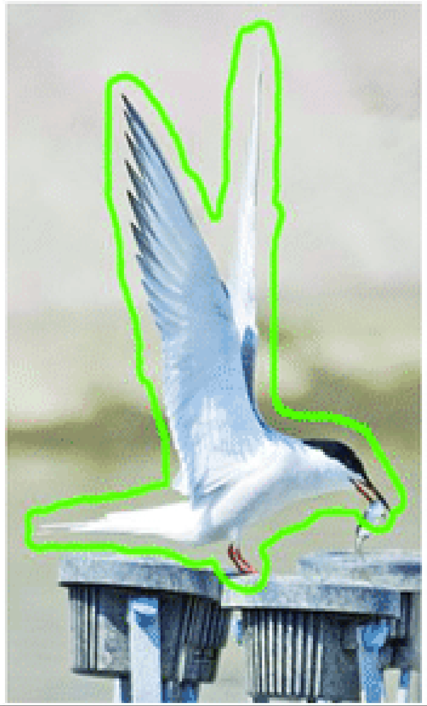
Examples of Ground Truth Attributes

AWA2 attribute labels

Class-wise binary attributes

	stripe	arctic	furry	white	black	red	brown	water
Zebra	1	0	0	1	1	0	0	0
Polar bear	0	1	1	1	0	0	0	1

CUB attribute labels



forehead_color	black	black	black
breast_pattern	solid	solid	solid
breast_color	white	white	white
head_pattern	plain	capped	plain
back_color	white	white	black
wing_color	grey/white	grey	white
leg_color	orange	orange	orange
size	medium	large	medium
bill_shape	needle	dagger	dagger
wing_shape	pointed	tapered	long
...	...	...	...
primary_color	white	white	white

source:

Quantitative Evaluation Protocol

- Since our method adopts *counterfactual* approach,
  - A_{y_o}: Attributes of the original class, A_{y_t}: Attributes of the target class
  - Only the difference between two attributes is our interest
  - Consider only if the two elements of the attributes are different
  - $A_{\mathrm{GT}} = [A_{y_t}[i] - A_{y_o}[i] \text{ for i in range}\left(\mathrm{len}\left(A_{y_o}\right)\right) \text{ if } A_{y_o}[i] \neq A_{y_t}[i]]$

Class-wise binary attributes

	stripe	arctic	furry	white	black	red	brown	water
y_o: Zebra	1	0	0	1	1	0	0	0
y_t: Polar bear	0	1	1	1	0	0	0	1

Ground truth explanation

A'_{zebra}	1	0	0	1	0
$A'_{\mathrm{polar\ bear}}$	0	1	1	0	1
A_{GT}	-1	1	1	-1	1

  - Report $\mathrm{AUROC}(A_{\mathrm{GT}}, I)$

source:

Quantitative Comparison w/ Baseline-CCE

- 26% improved AUROC compared to Baseline-CCE

Target model	Dataset	Library	CCE	Ours
CLIP+Linear	AwA2	C_{AwA2}	0.6436	**0.8132**
	CUB	C_{CUB}	0.7066	**0.7891**
ResNet18	AwA2	C_{AwA2}	0.6113	**0.7314**
	CUB	C_{CUB}	0.6979	**0.7750**
ResNet50	AwA2	C_{AwA2}	0.5811	**0.7316**
	CUB	C_{CUB}	0.6811	**0.7336**

source:

Thank You!