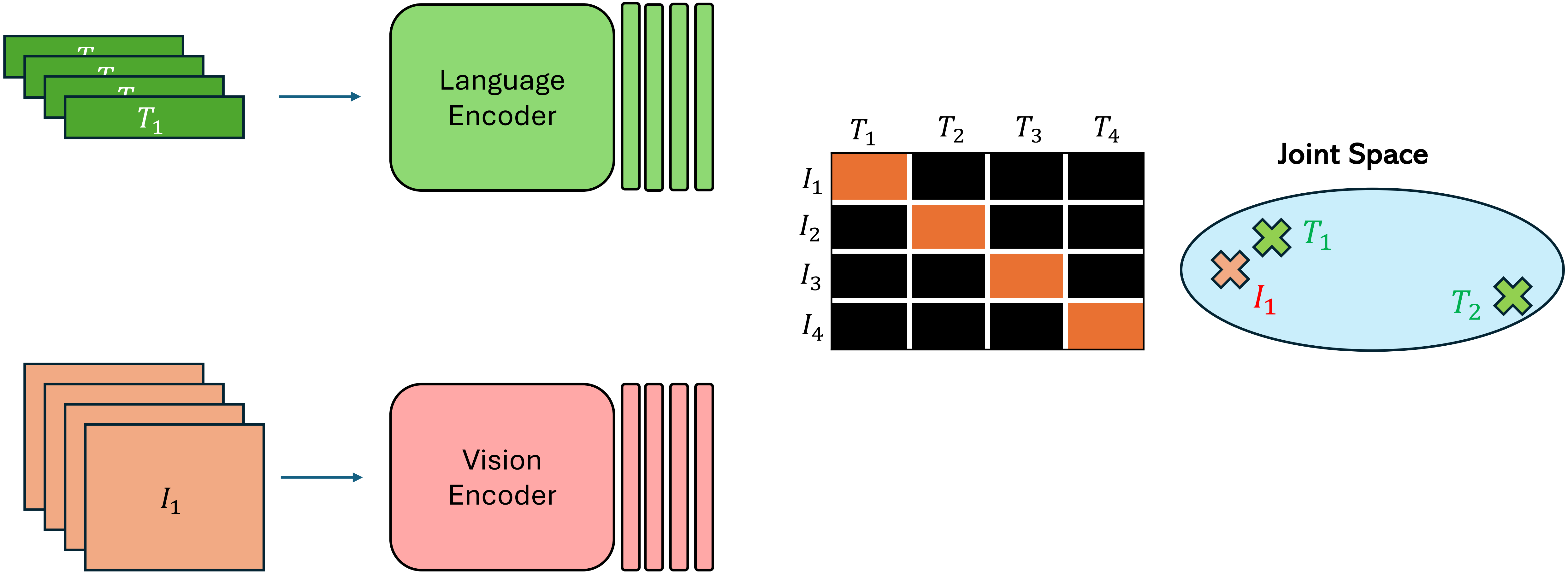
VRIJE UNIVERSITEIT BRUSSEL

NEURAL INFORMATION PROCESSING SYSTEMS

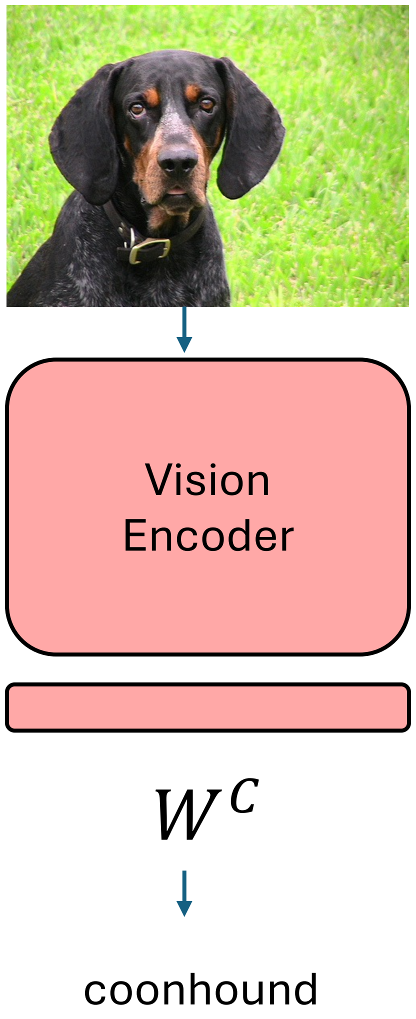
Interpreting and Analysing CLIP's Zero-Shot Image Classification via Mutual Knowledge

Fawaz Sammani, Nikos Deligiannis

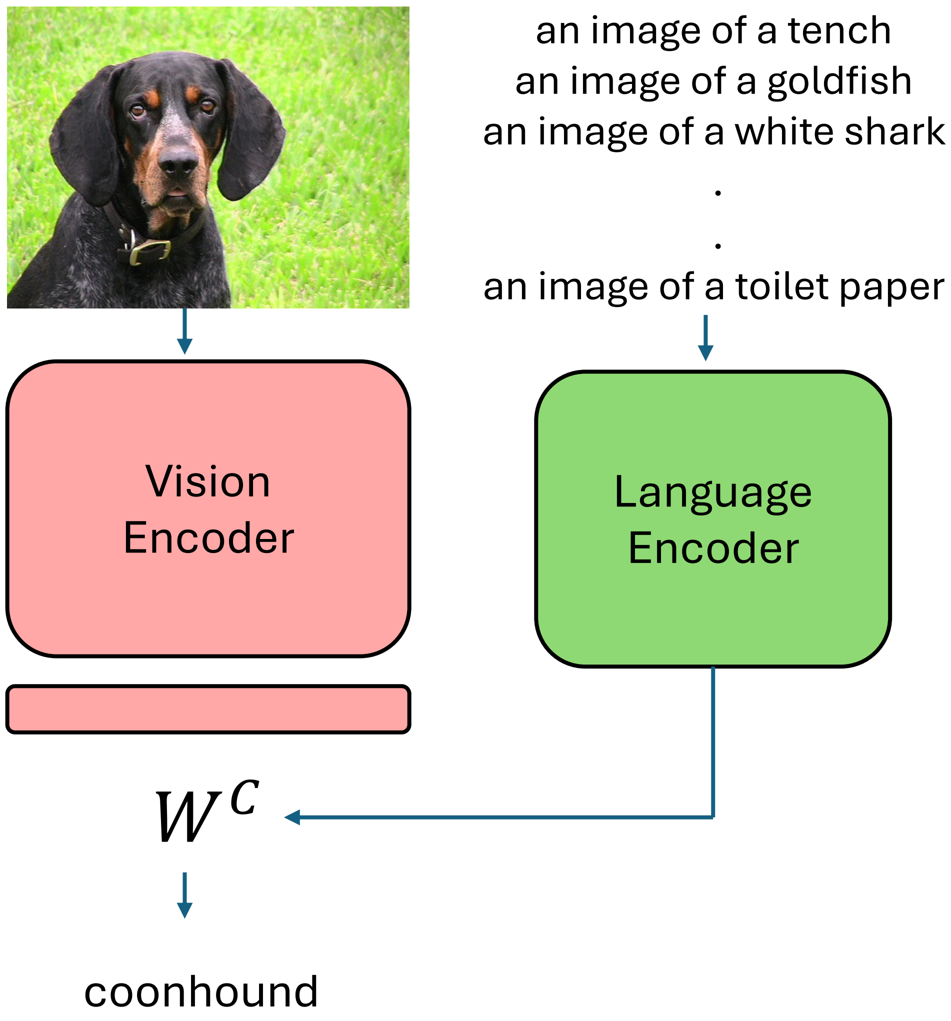
CLIP
Contrastive Language Image Pretraining

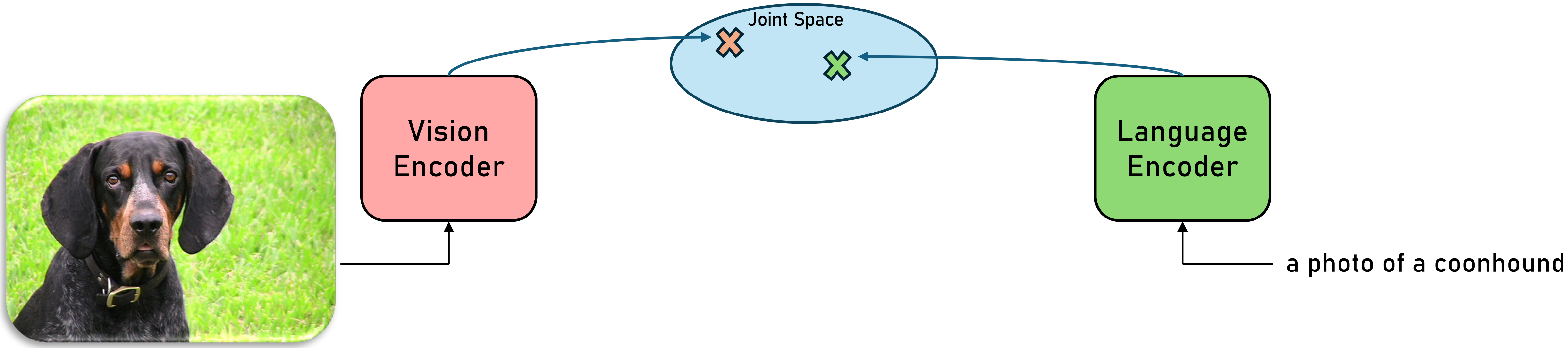


Normal Classification

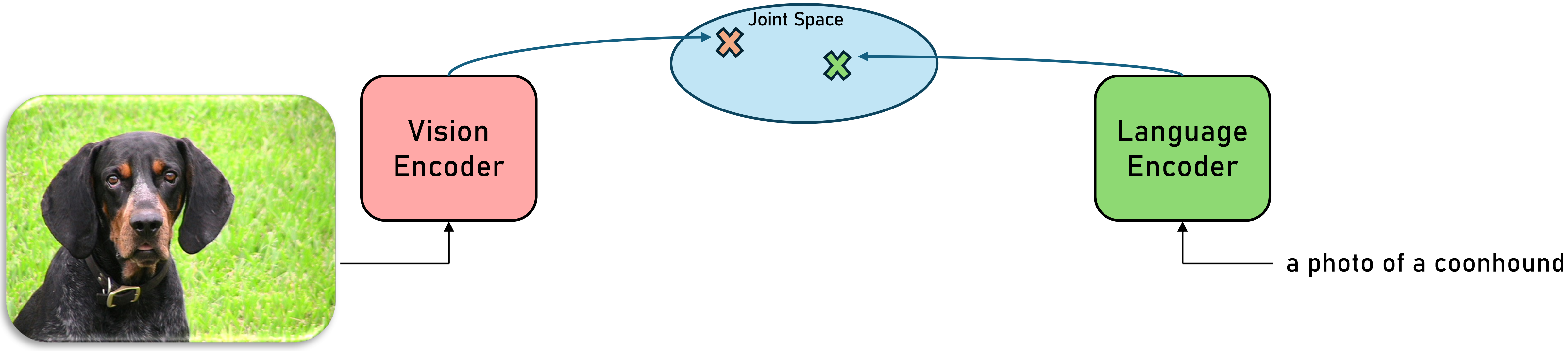


CLIP Zero-Shot Classification



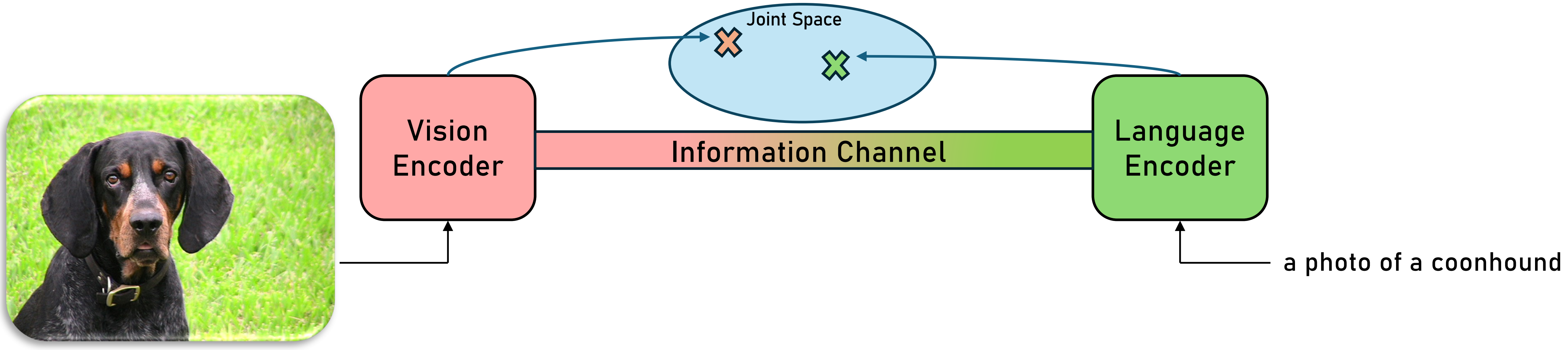


What do the vision and language encoders of CLIP learn in common, causing image-text points to be closer or further apart in the joint space?

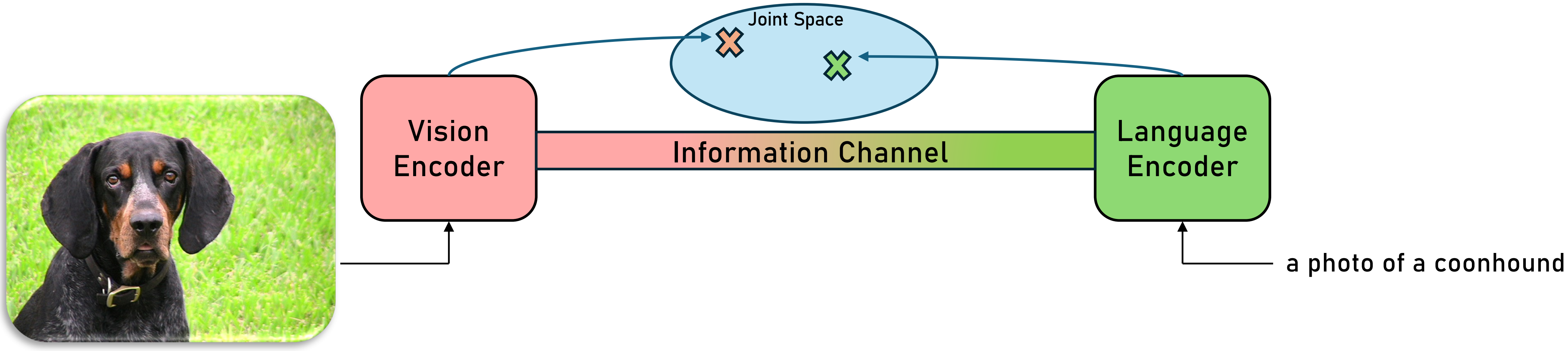


What do the vision and language encoders of CLIP **learn in common**, causing image-text points to be closer or further apart in the joint space?

Mutual Information



vision-language information channel



How to enable this?

The vision and language interpretations must be in the same space

How to do this efficiently?

Discrete Units (simple MI calculation in discrete space, correctly models ‘’bits” of the channel)

How to make it understandable?

Human-Friendly interpretations

Textual Concepts

Descriptors; short descriptions in natural language
Covers many objects in the world

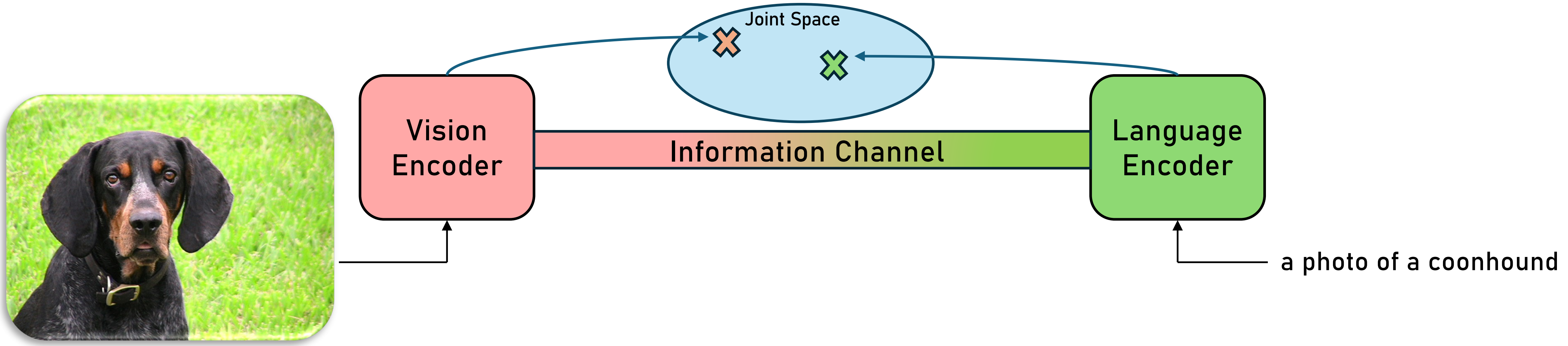
a long snout

feathered ears

sharp teeth

pointy snout

.
.
.



How to enable this?

Space of Textual Concepts

The vision and language interpretations must be in the same space

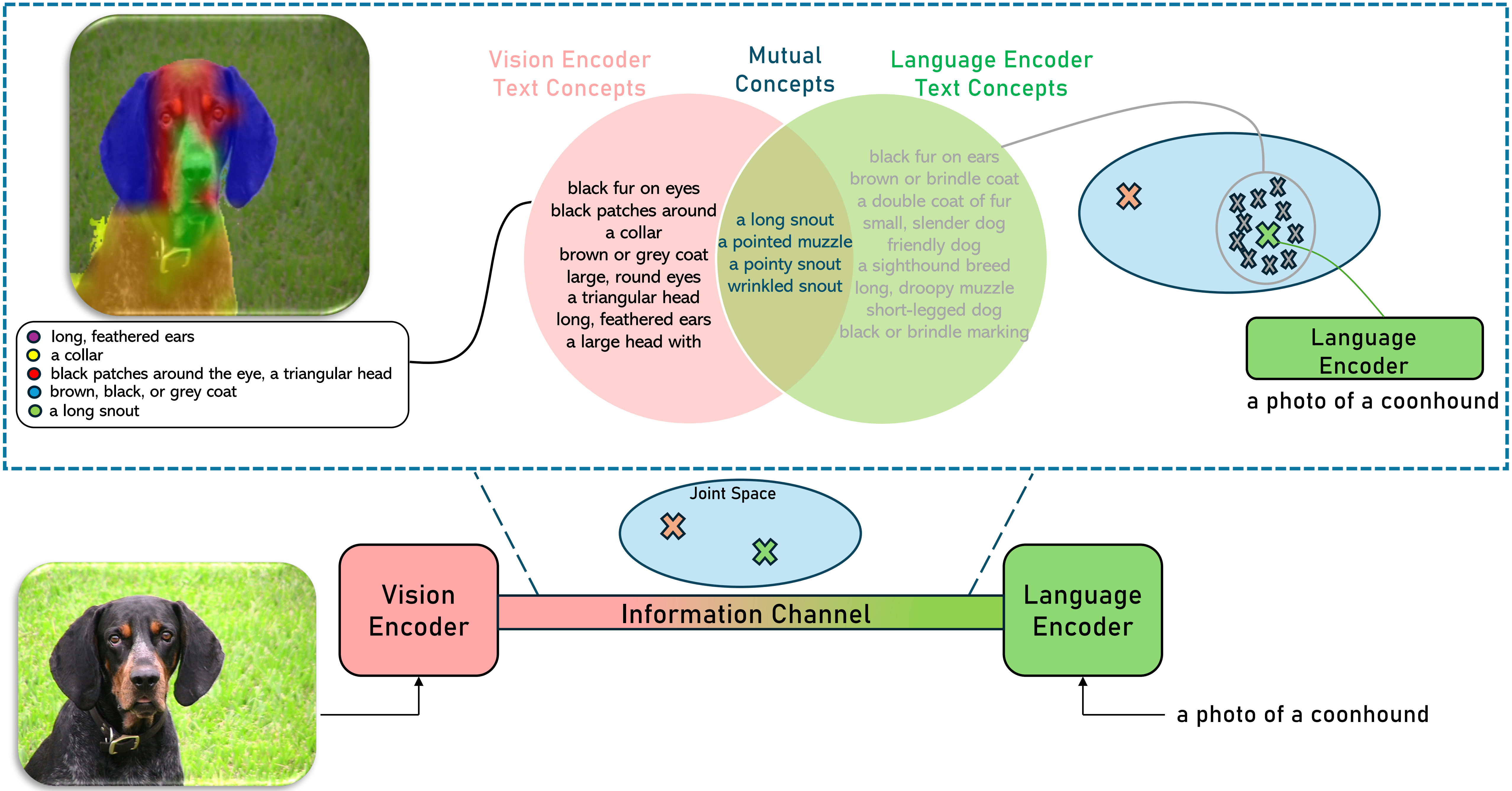
How to do this efficiently?

a long snout → 0
feathered ears → 1

Discrete Units (simple MI calculation in discrete space, correctly models ''bits" of the channel)

How to make it easily understandable?

Human-Friendly interpretations



Vision Encoder Text Concepts
Mutual Concepts
Language Encoder Text Concepts
black fur on eyes
black patches around
a collar
brown or grey coat
large, round eyes
a triangular head
long, feathered ears
a large head with
a long snout
a pointed muzzle
a pointy snout
wrinkled snout
black fur on ears
brown or brindle coat
a double coat of fur
small, slender dog
friendly dog
a sighthound breed
long, droopy muzzle
short-legged dog
black or brindle marking
long, feathered ears
a collar
black patches around the eye, a triangular head
brown, black, or grey coat
a long snout
Language Encoder
a photo of a coonhound
Joint Space
Vision Encoder
Information Channel
Language Encoder
a photo of a coonhound

- large eyes
- pointed ears; long, feathered ears
- a long, snout-like nose; long whiskers
- a thick, double coat of fur that is black and silver
- a black head with a white stripe behind the eye

- often has spots or stripes
- a small head with a red and yellow bill
- birds or other animals nesting on the cliff

- blue plumage; long, narrow tail
- a black back and wings; a long, thin strip of feathers
- long, red bill; a red beak

Evaluating Multimodal Concepts

Baseline 1: Multimodal Concept Bottleneck Models

Baseline 2: Neuron Annotation

Table 1: Evaluation scores of our multimodal explanations compared to the baselines established. All use the same features, model and textual concept bank for fair comparison.

Explanation	Requires Training	**Delet.**↓	**Insert.**↑	**AccDrop**↓	**AccInc**↑
MM-CBM	Yes	3.147	**3.385**	2.634	1.013
MM-ProtoSim	Yes	3.149	3.358	2.665	0.943
Feature Maps	Yes	2.921	3.114	2.283	1.233
Ours (PCA)	No	2.460	3.168	1.582	**1.849**
Ours (K-means)	No	**2.422**	3.122	**1.555**	1.781

Evaluation with CLIP Classification via Descriptions

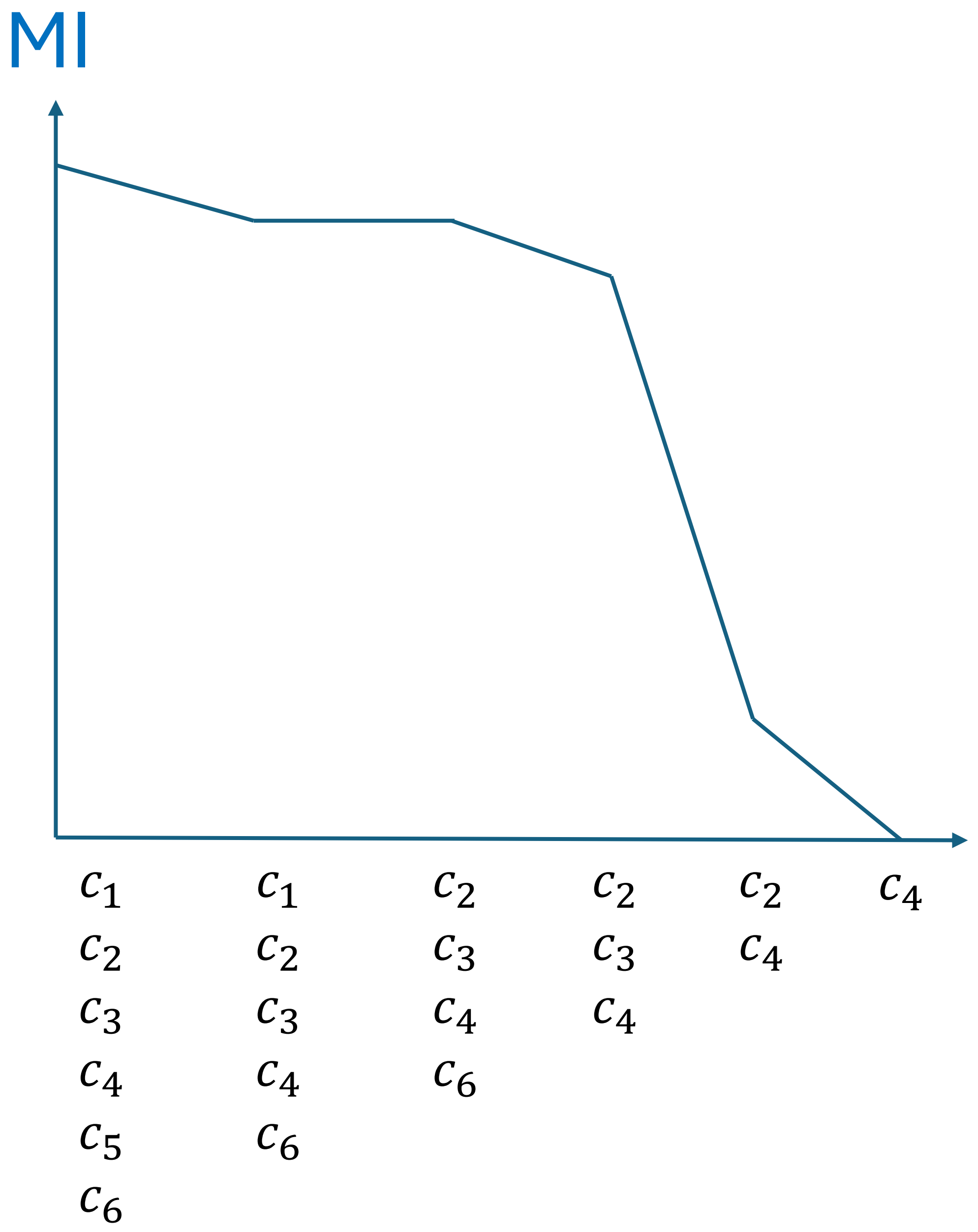
Table 2: Effectiveness and Relevancy of our multimodal concepts in boosting zero-shot accuracy of both ResNet and ViT CLIP models on the ImageNet validation set compared to baselines [36, 43].

ResNets	Base	Ours	Δ	ViTs	Base	Ours	Δ
RN50	59.54	**61.85**	+2.31	ViT-B/16	67.93	**70.28**	+2.35
RN50x4	64.36	**67.93**	+3.57	ViT-B/32	63.28	**65.58**	+2.30
RN50x16	68.47	**72.22**	+3.75	ViT-L/14	74.69	**76.74**	+2.05
RN101	60.68	**64.14**	+3.46	ViT-L/14@336px	75.49	**77.64**	+2.15

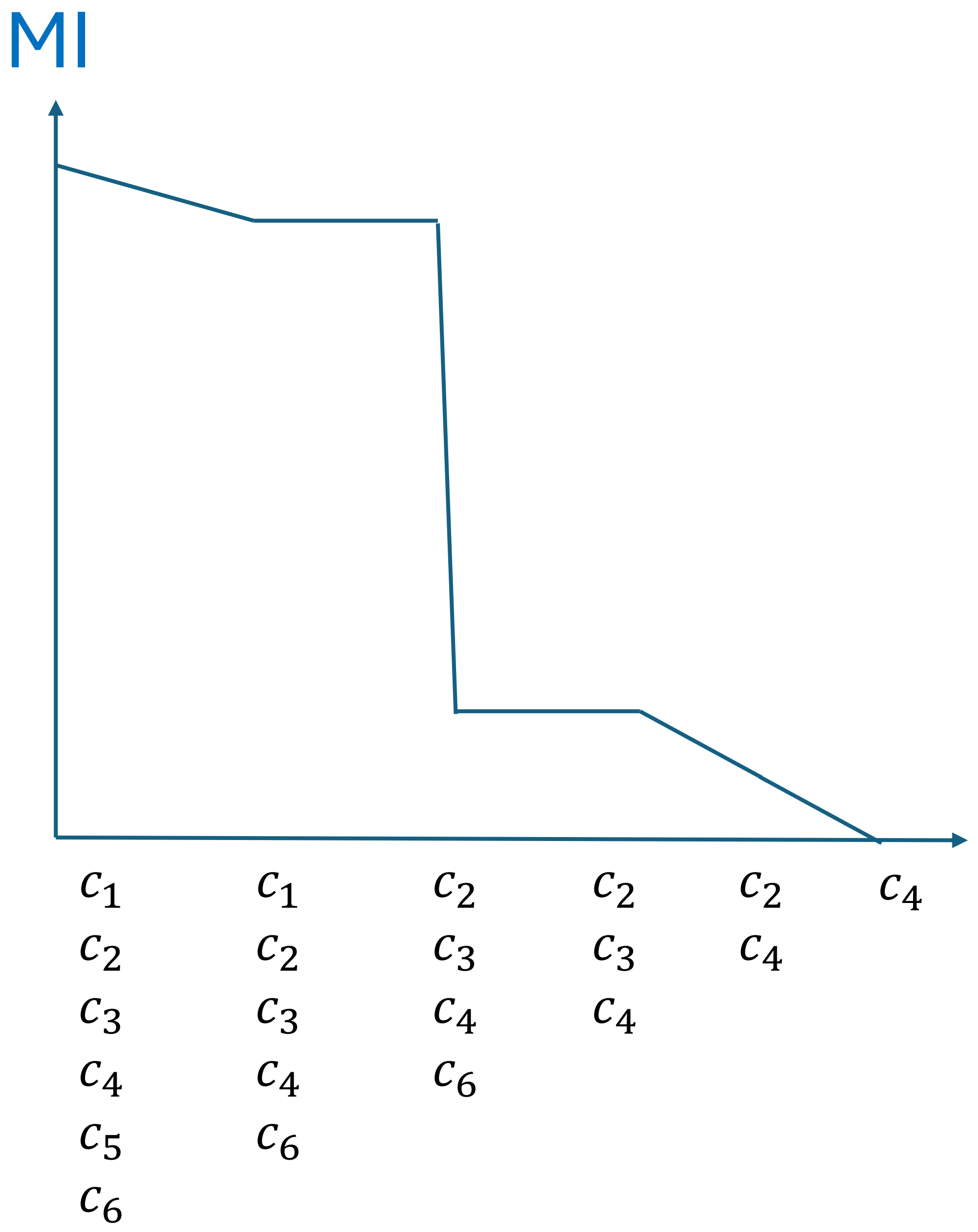
Defining Mutual Knowledge

Concepts are entangled. It only makes sense to consider them as a whole, or in relation to each other.

We define that two sources have a strong shared knowledge when a source retains knowledge about the other, despite removing important information units from it.



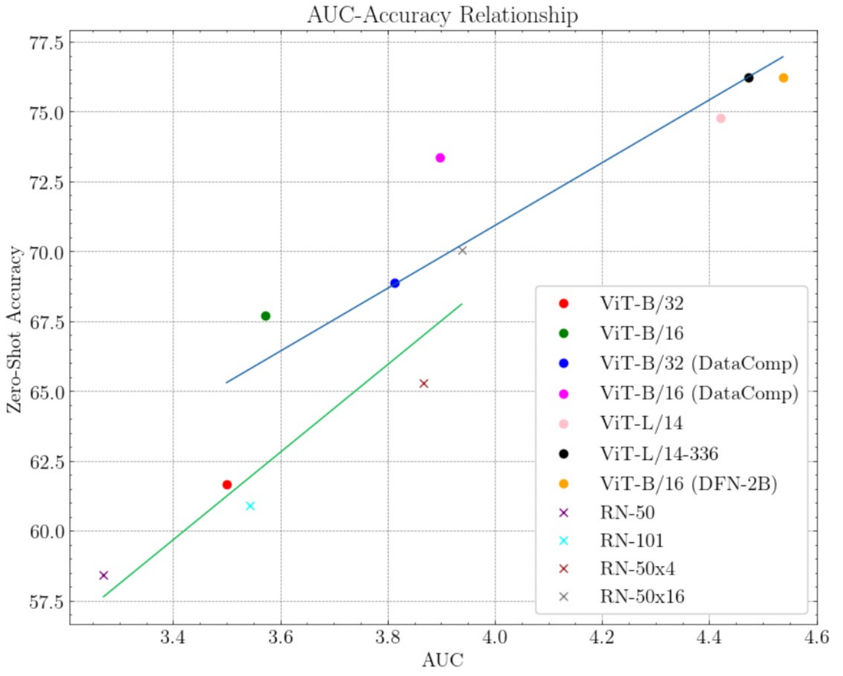
A higher AUC indicates gradual or late drops of MI in the curve, and thus stronger shared knowledge



A lower AUC indicates sharp or early drops of MI, and thus weaker shared knowledge

Table 3: MI and AUC scores for different model families using PCA and K-means evaluated on the full ImageNet validation split, along with the pretraining data and Top-1 accuracy.

Model Family	Model	Data Size	Top-1 (%)	MI		AUC	
				PCA	K-means	PCA	K-means
ViTs	ViT-B/32	400M	61.66	7.40	7.26	3.61	3.39
	ViT-B/16	400M	67.70	7.50	7.44	3.62	3.53
	ViT-B/32-dcp	1B	68.88	7.79	7.65	3.93	3.70
	ViT-B/16-dcp	1B	73.37	7.68	7.58	3.99	3.81
	ViT-L/14	400M	74.77	7.94	7.89	4.47	4.37
	ViT-L/14↑	400M	76.23	7.96	7.93	4.51	4.44
	ViT-B/16-dfn	2B	**76.24**	**8.19**	**8.11**	**4.62**	**4.46**
ResNets	RN-50	400M	58.42	7.14	7.20	3.23	3.32
	RN-101	400M	60.90	7.43	7.53	3.49	3.60
	RN-50×4	400M	65.28	**7.53**	7.58	3.84	3.90
	RN-50×16	400M	**70.04**	7.51	**7.63**	**3.85**	**4.03**
ConvNeXTs	CNeXt-B1	400M	65.36	6.47	6.66	2.54	2.80
	CNeXt-B2	13B	**71.22**	**7.16**	**7.56**	**3.19**	**3.74**



Mutual Knowledge is also an evaluation of the Mutual Concepts, by assuming correlation with accuracy

Code available:

`https://github.com/fawazsammani/clip-interpret-mutual-knowledge`