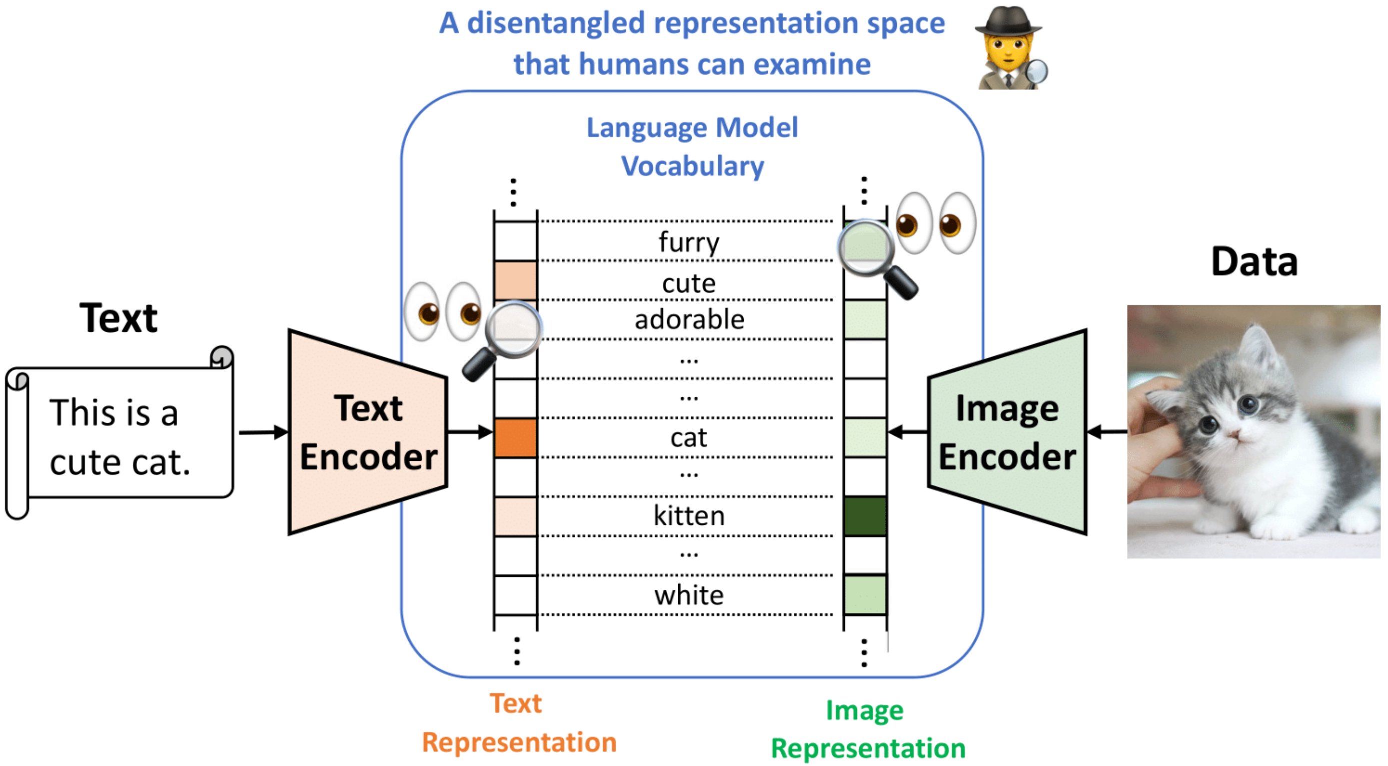
Retrieval-based Disentangled Representation Learning with Natural Language Supervision

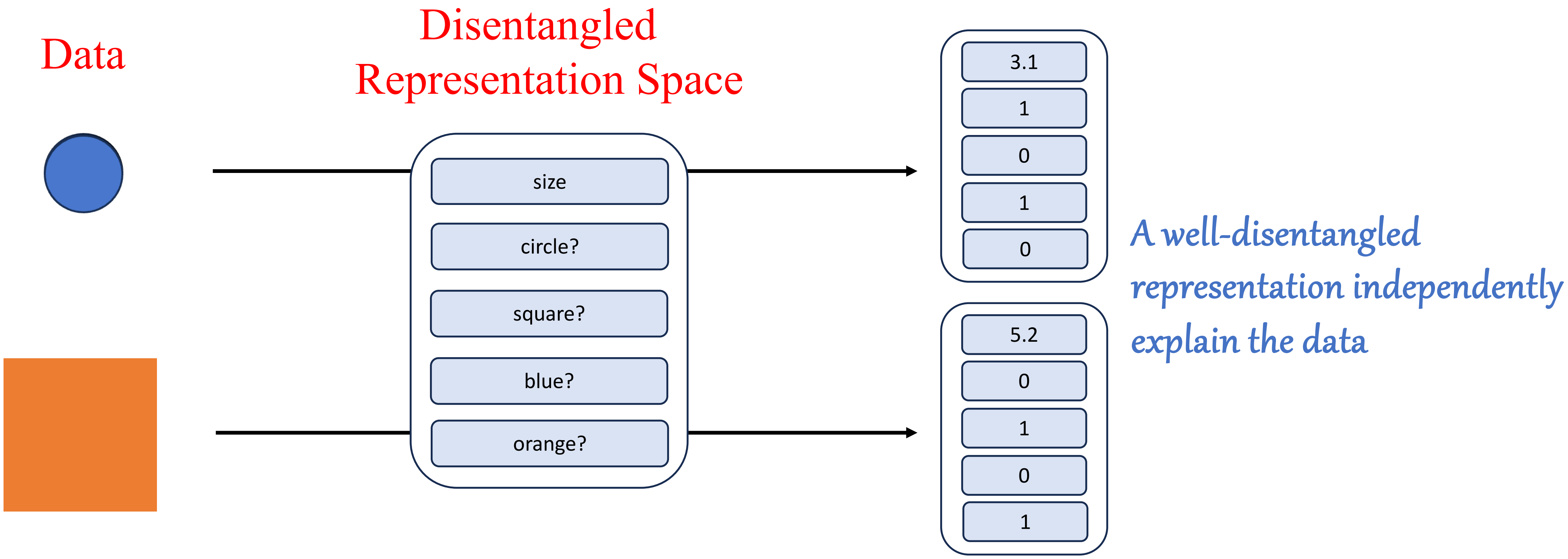
Jiawei Zhou[1], Xiaoguang Li[2], Lifeng Shang[2], Xin Jiang[2], Qun Liu[2], Lei Chen[1]

[1]The Hong Kong University of Science and Technology
[2]Huawei Noah's Ark Lab



Background: Disentangled Representation

Disentangled representation learning aims to identify the underlying factors of variations within data and correlate them to distinct units of the learned representation.



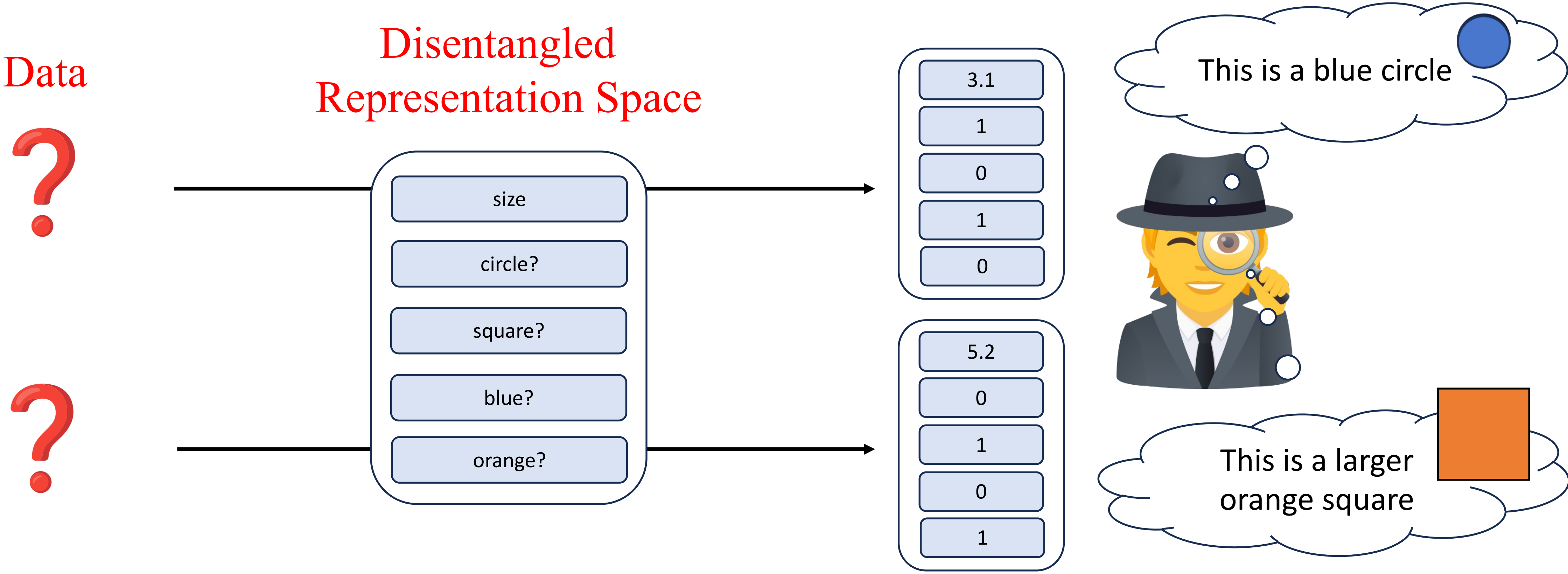
Yoshua Bengio, et al., 2013. Representation learning: A review and new perspectives.

Challenge

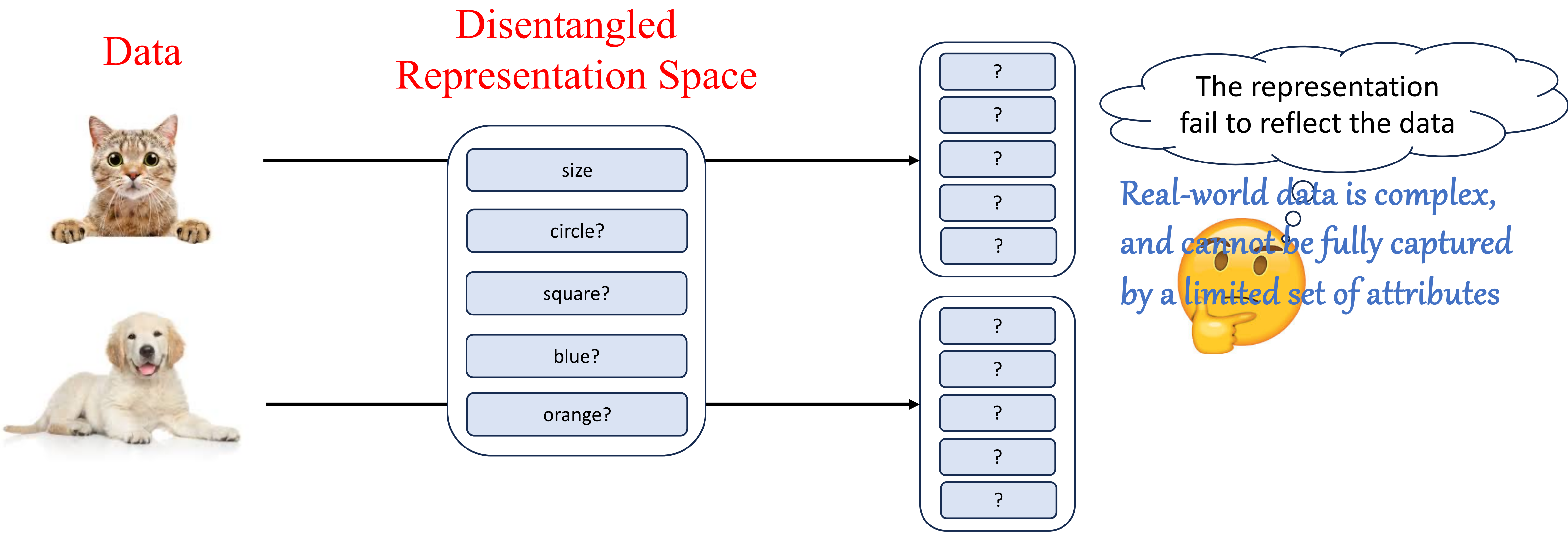
1. How to define an effective **disentangled representation space**?
(i.e., how to use finite attributes to differentiate between diverse real-world objects?)

2. How to induce **dimension-wise supervision** on disentangled representation space?

Background: Disentangled Representation

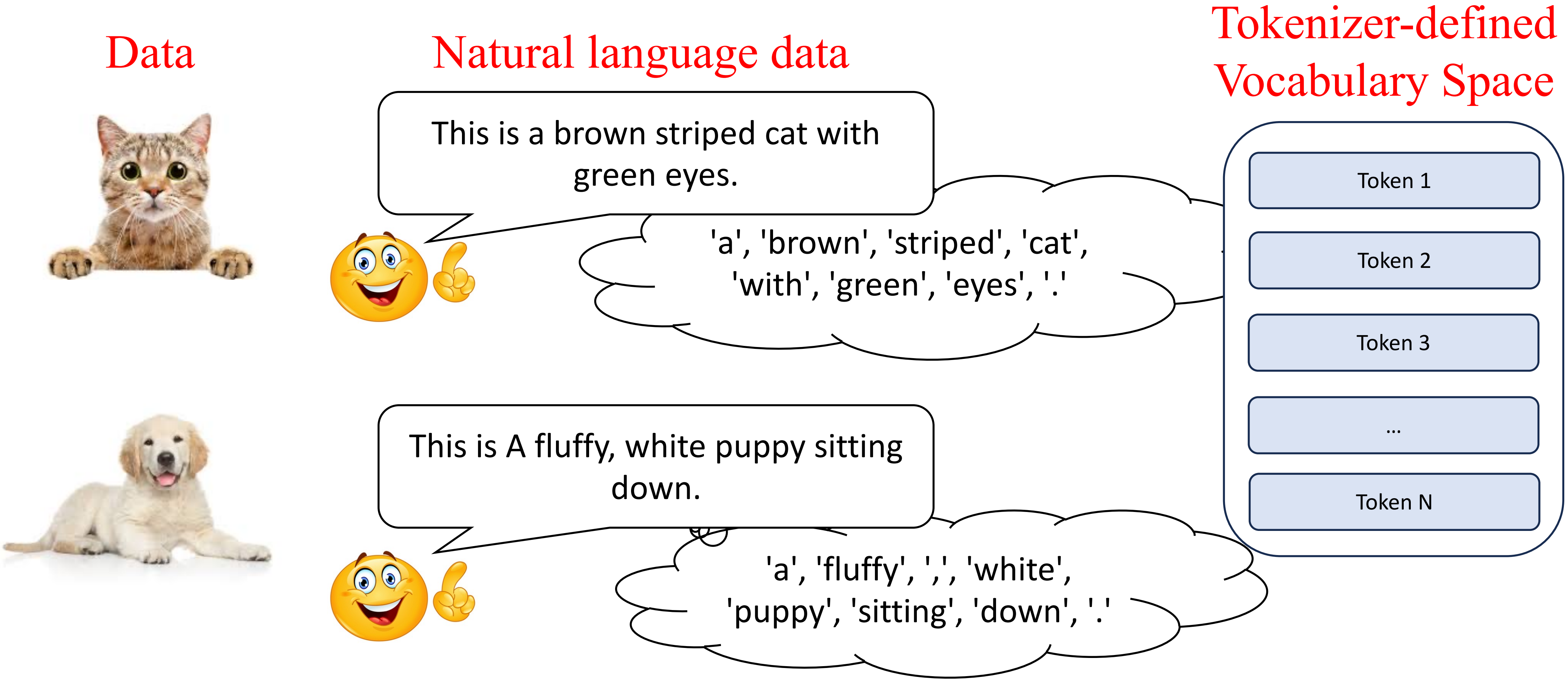


Challenge: Disentangled Representation Space



Solution: Disentangled Representation Space

Real-world objects can be differentiated through natural language descriptions.
Natural language data can be tokenized into a finite set of tokens.



Solution: Disentangled Representation Space

1. How to define an effective **disentangled representation space**?
(i.e., how to use finite attributes to differentiate between diverse real-world objects?)

natural language expression = proxy of the input data

tokenizer vocabulary space = disentangled representation space

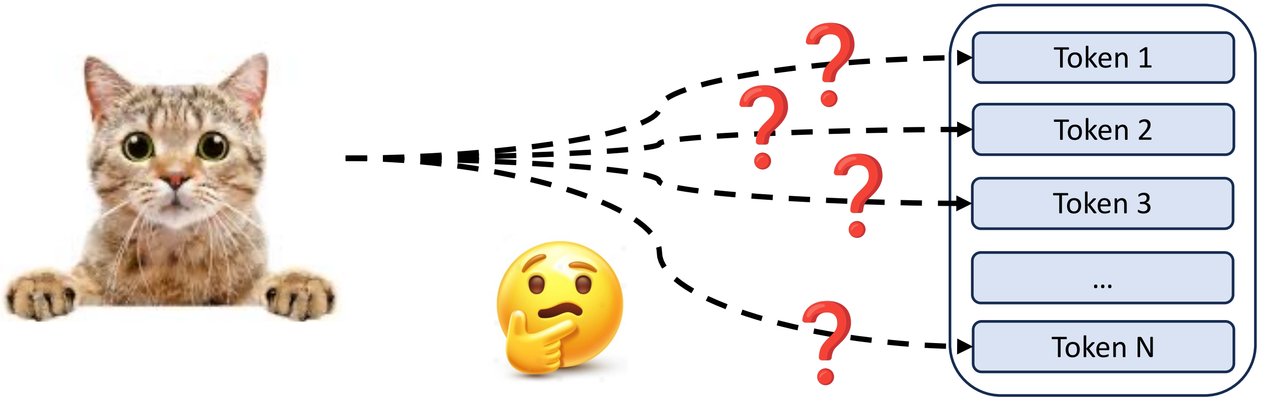
Challenge: Dimension-wise Supervision

1. How to define an effective **disentangled representation space**?
(i.e., how to use finite attributes to differentiate between diverse real-world objects?)

natural language expression = proxy of the input data

tokenizer vocabulary space = disentangled representation space

2. How to induce **dimension-wise supervision** on disentangled representation space?



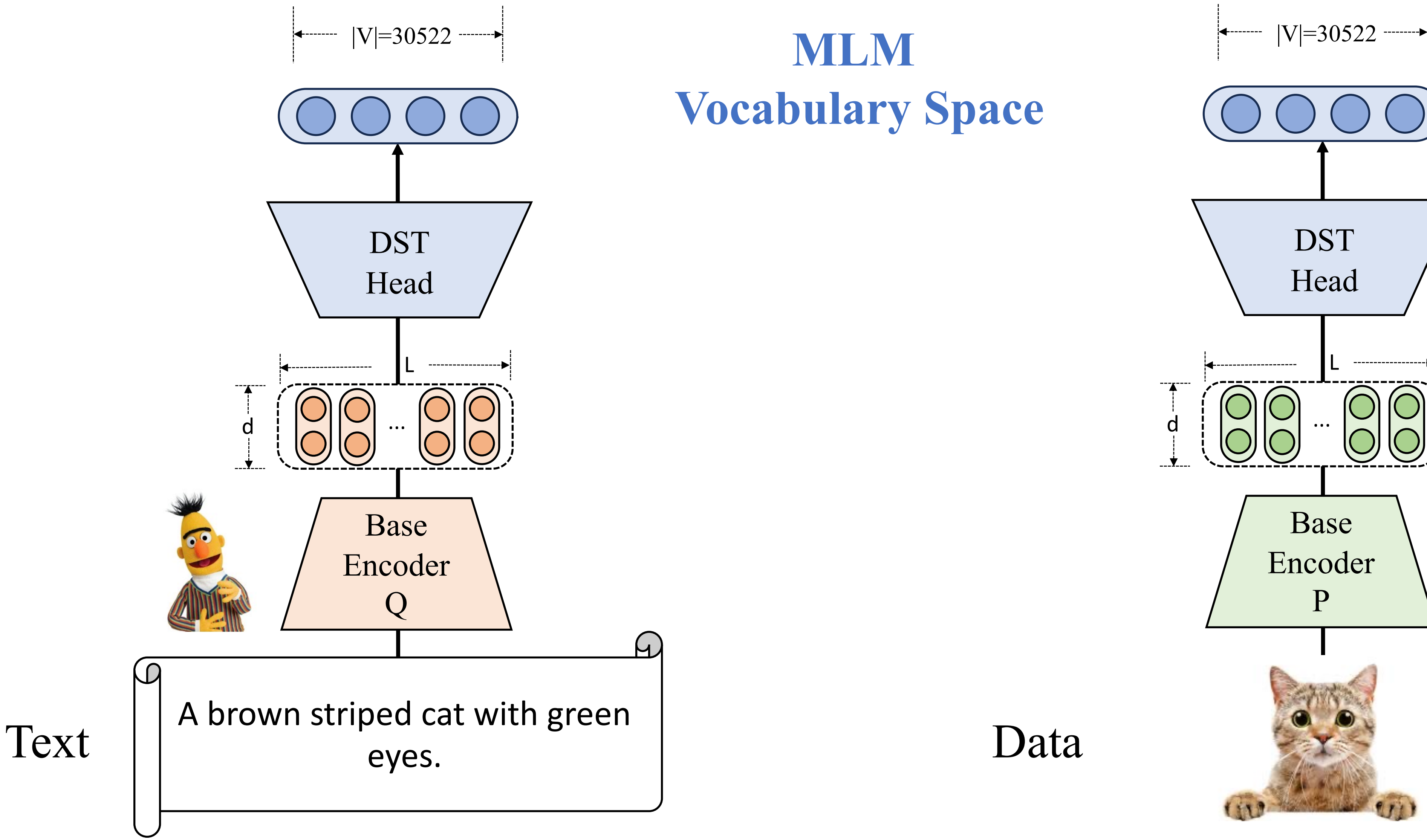
Solution: Dimension-wise Supervision

Key components :

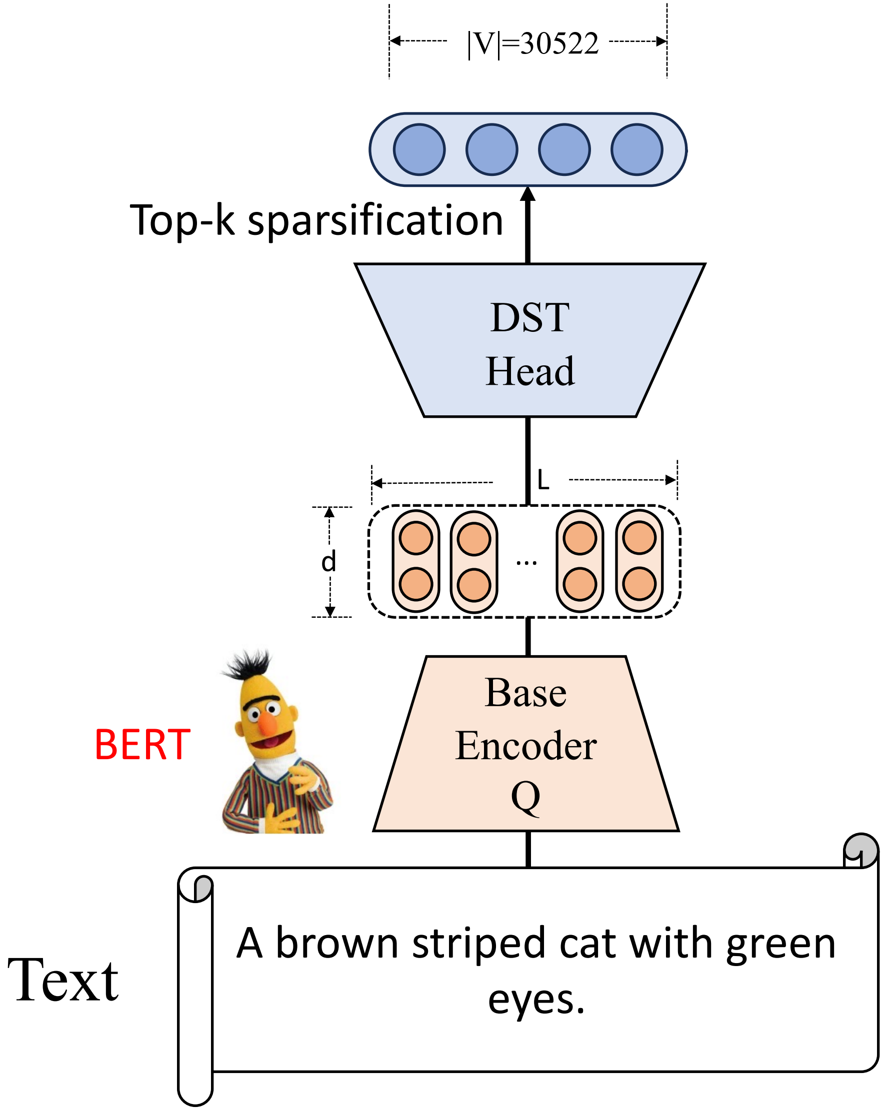
1. Pre-trained Masked Language Model
2. Sparse Bi-encoder Framework
3. Contrastive Learning

Let's see how VDR works!

Bi-encoder Framework



Bi-encoder Framework (Text Encoder)



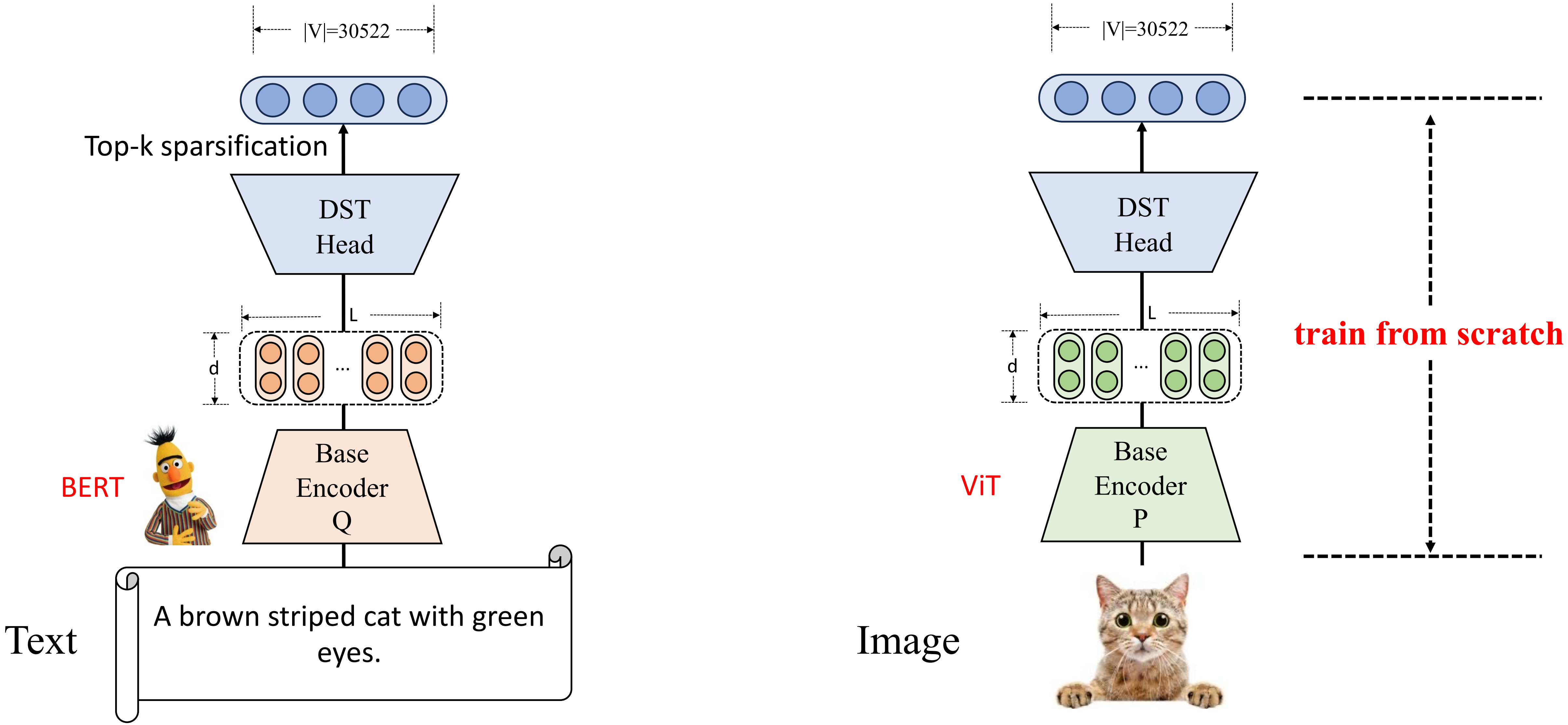
1. Use output token probability of pre-trained MLM, and replace the softmax to elu1p activation in MLM head

$$elu1p(x) = \begin{cases} x+1 & \text{if } x >= 0 \\ e^x & \text{otherwise} \end{cases}$$

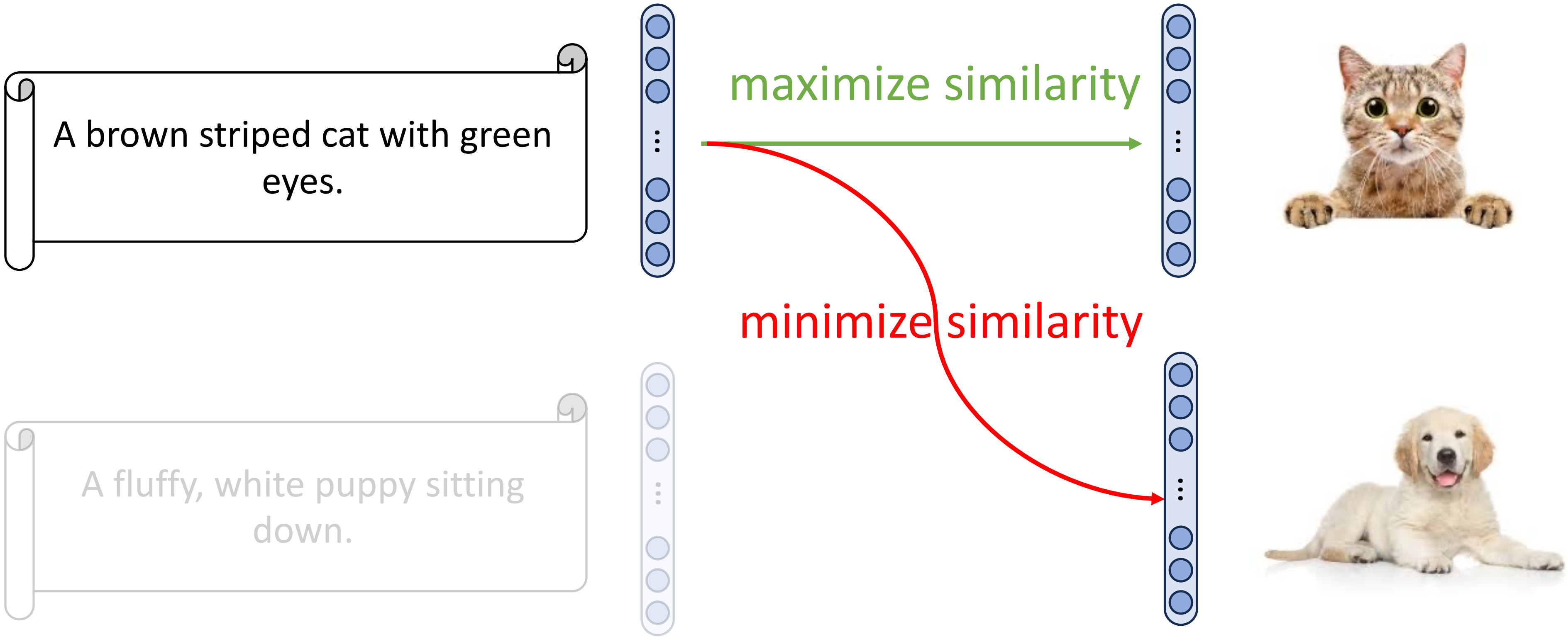
2. Apply max pooling to aggregate token representations

3. Top-k sparsification

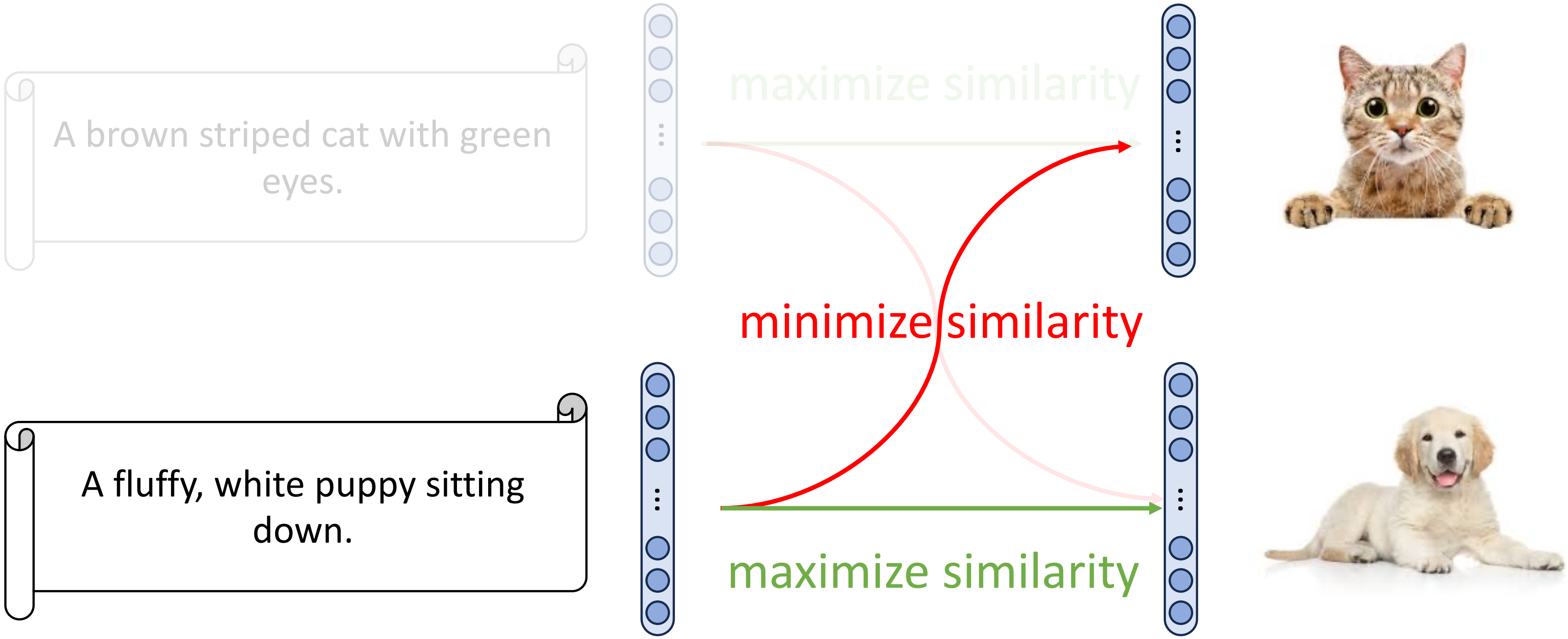
Bi-encoder Framework (Image Encoder)



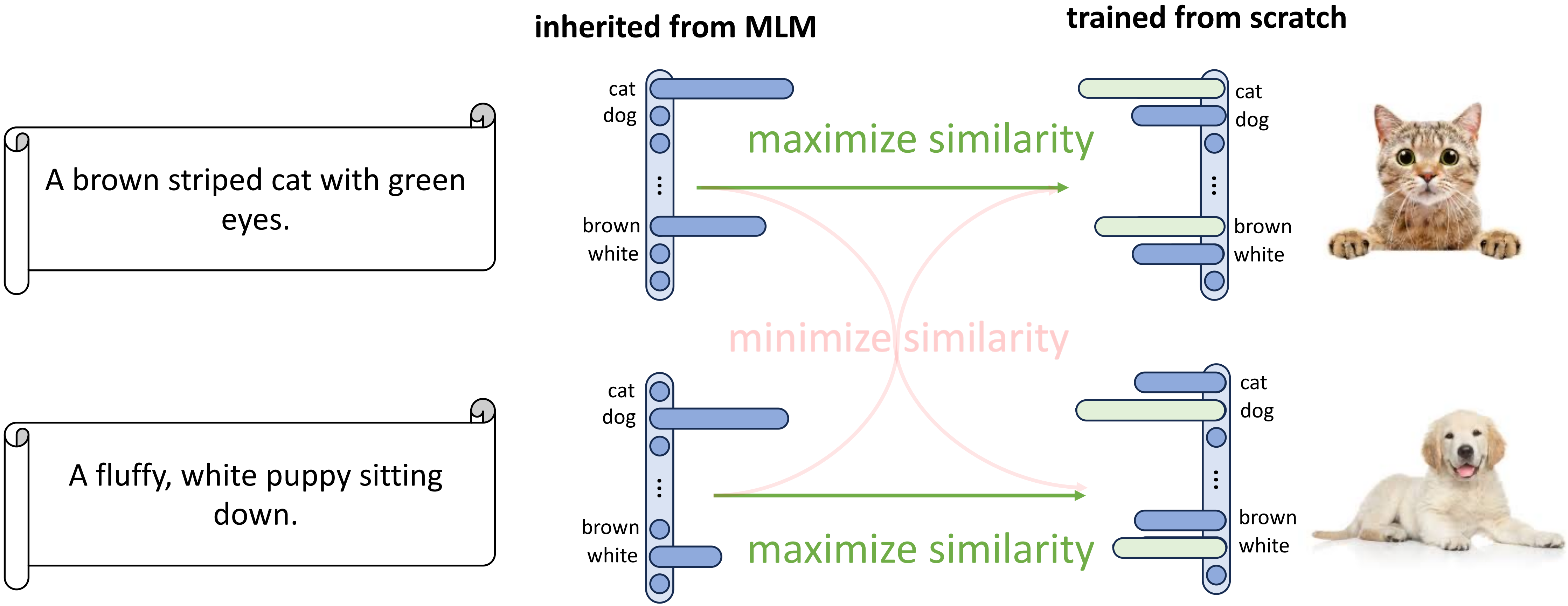
Contrastive Learning



Contrastive Learning

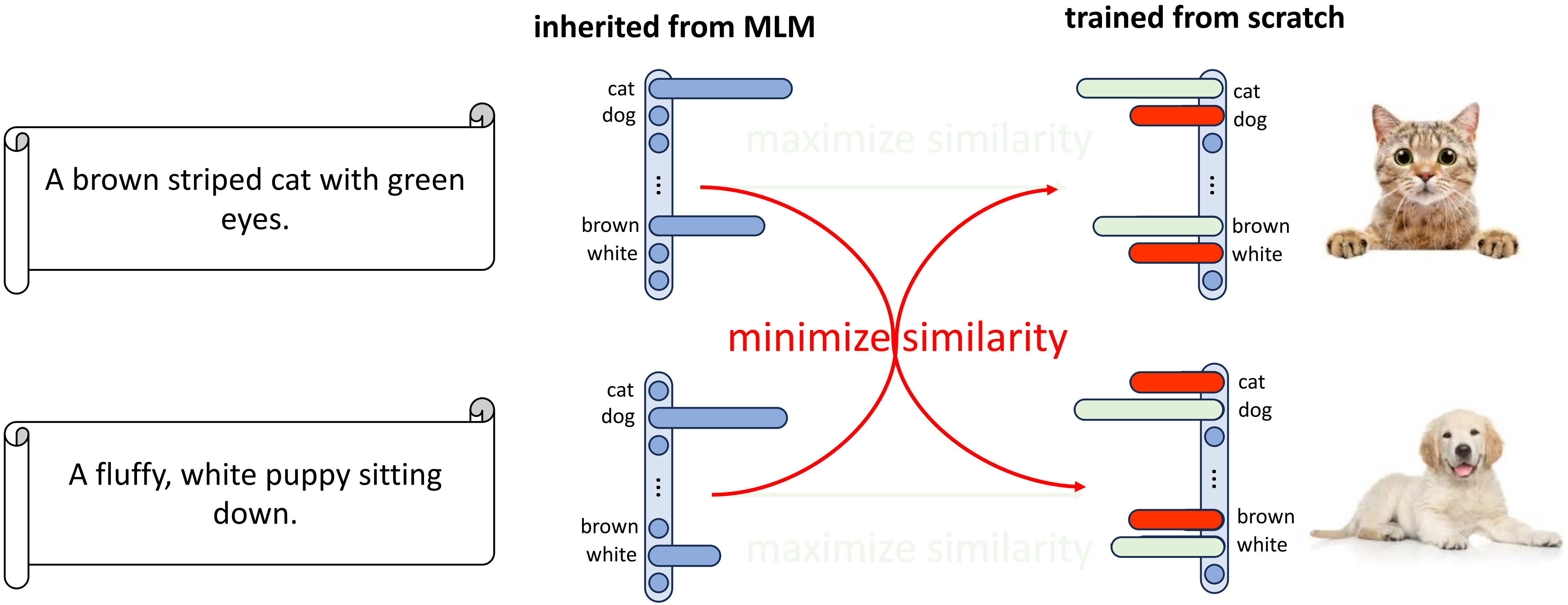


Contrastive Learning



Similarity measured as inner product of representations

Contrastive Learning



Similarity measured as inner product of representations

Experimental Setup

Our experiments cover both text-to-text retrieval scenarios and cross-modal retrieval scenarios.

Model

Text-to-text Retrieval (2 text encoders)

- 20 epochs
- batch size 256

Cross-modal Retrieval (1 text encoder + 1 image encoder)

- 20 epochs
- batch size 4096

Dataset

Text-to-text Retrieval

- Train on MS MARCO
- Eval on BEIR benchmark

Cross-modal Retrieval

- Train on YFCC15m
- Eval on ImageNet, MSCOCO, Flickr30k

Experimental Results (text-to-text)

Model	BM25	SPLADE	†DPR	†$\text{VDR}^{\alpha}_{\text{t2t}}$	†VDR_{t2t}	ANCE	UnifieR	Contriever	SimLM	MASTER	RetroMAE	LexMAE	E5_{base}
Retrieval Pre-training			✘				✔	✔	✔	✔	✔	✔	✔
Special Negatives			✘			✔	✔		✔	✔		✔	✔
Distillation			✘						✔	✔		✔	✔
Wikipedia Access			✘					✔	✔	✔	✔		✔
ArguAna	31.5	43.9	40.8	**48.8**	48.6	41.5	39.0	44.6	42.1	39.5	43.3	50.0	51.4
Climate-FEVER	**21.3**	19.9	16.2	18.1	17.6	19.8	17.5	23.7	16.3	21.5	23.2	21.9	15.4
DBPedia	31.3	36.6	30.4	37.6	**39.0**	28.1	40.6	41.3	34.5	39.9	39.0	42.4	41.0
FEVER	**75.3**	73.0	63.8	74.8	74.0	66.9	69.6	75.8	65.7	69.2	77.4	80.0	58.2
FiQA	23.6	28.7	23.7	**29.3**	28.8	29.5	31.1	32.9	29.2	32.8	31.6	35.2	36.4
HotpotQA	60.3	63.6	45.2	**68.4**	65.5	45.6	66.1	63.8	58.1	58.9	63.5	71.6	62.2
NFCorpus	32.5	31.3	26.1	32.7	**33.0**	23.7	32.9	32.8	32.3	33.0	30.8	34.7	36.6
NQ	32.9	46.9	43.2	45.8	**47.2**	44.6	51.4	49.8	47.7	51.6	51.8	56.2	60.0
SCIDOCS	**15.8**	14.5	10.9	15.4	15.3	12.2	15.0	16.5	14.5	14.1	15.0	15.9	19.0
SciFact	66.5	62.8	47.4	**67.6**	67.3	50.7	68.6	67.7	58.8	63.7	65.3	71.7	73.1
TREC-COVID	65.6	67.3	60.1	**69.0**	67.8	65.4	71.5	59.6	63.7	62.0	77.2	76.3	79.6
Touché-2020	**36.7**	20.1	22.1	27.7	29.8	28.4	30.2	23.0	29.2	32.0	23.7	29.0	28.3
Avg.	41.1	42.4	35.8	**44.6**	44.5	38.0	44.5	44.3	44.4	43.1	45.1	48.7	46.8
Avg. (w/o NQ)	-	-	-	44.5	44.3	-	-	43.8	40.4	42.4	44.5	-	45.6

- VDR_{t2t} outperform DPR by 8.7% with similar model size and training costs.
- VDR_{t2t} achieve comparable performance to other advanced retrievers.

Experimental Results (cross-modal)

Model	ImageNet		MSCOCO							Flickr30k						
			image-to-text			text-to-image				image-to-text			text-to-image			
	Top1	Top5	R@1	R@5	R@10	R@1	R@5	R@10	R-mean	R@1	R@5	R@10	R@1	R@5	R@10	R-mean
CLIP	32.8†	57.4†	20.8	43.9	55.7	13.0	31.7	42.7	32.6	34.9	63.9	75.9	23.4	47.2	58.9	50.7
†CLIP-BERT	32.4	56.1	23.9	47.8	60.3	13.6	33.8	45.1	37.4	44.1	71.2	80.7	27.8	54.7	65.9	57.4
†VDR_{cm}	**38.7**	**63.6**	**30.9**	**54.5**	**65.4**	**17.4**	**38.1**	**49.7**	**42.7**	**51.0**	**79.3**	**86.7**	**32.4**	**60.1**	**70.7**	**63.4**
†$\text{VDR}_{\text{cm}}^{\text{np}}$	-	-	-	-	-	11.8	28.6	38.6	-	-	-	-	21.1	42.3	52.8	-
SLIP	33.6†	58.6†	27.7	52.6	63.9	18.2	39.2	51.0	42.1	47.8	76.5	85.9	32.3	58.7	68.8	61.7
†FILIP	39.1	64.4	21.6	46.7	59.0	13.7	31.7	41.6	35.7	46.3	74.4	83.2	30.7	58.2	68.6	60.2
ProtoCLIP	32.0	-	30.2	55.1	66.5	16.9	37.9	49.4	42.7	-	-	-	-	-	-	-
†DeCLIP	43.2	69.4	25.3	51.2	63.4	16.6	35.2	45.4	39.5	51.3	80.7	88.5	35.5	63.0	73.0	**65.3**

- VDR_{cm} outperform CLIP {6.2%, 5.3%, 6.0%} on {ImageNet, MSCOCO, Flickr30k}, respectively.

Image Disentanglement

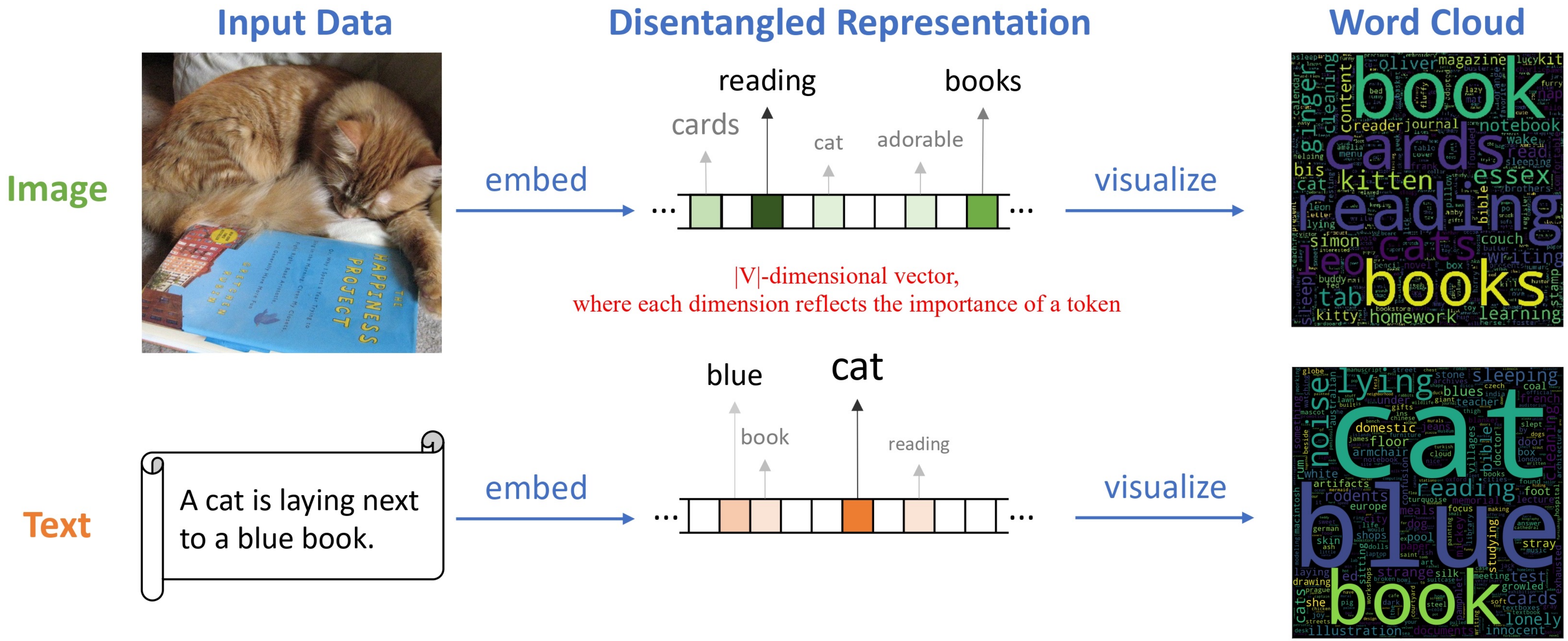
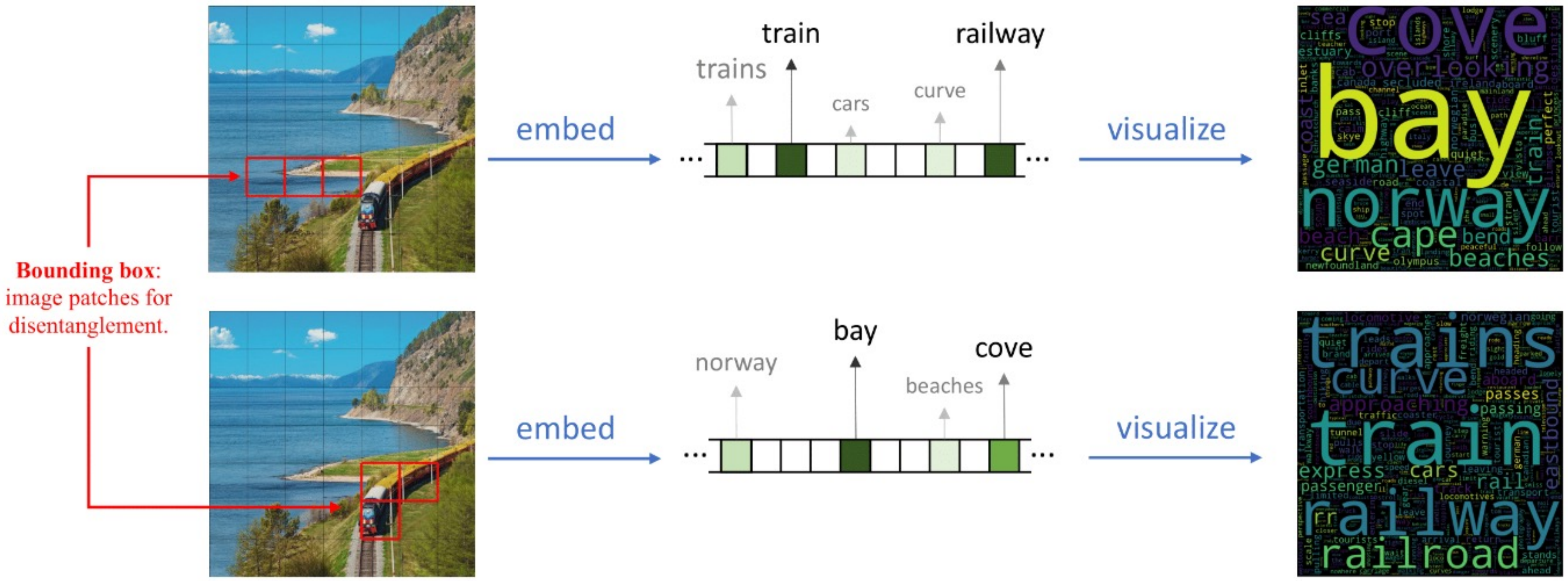
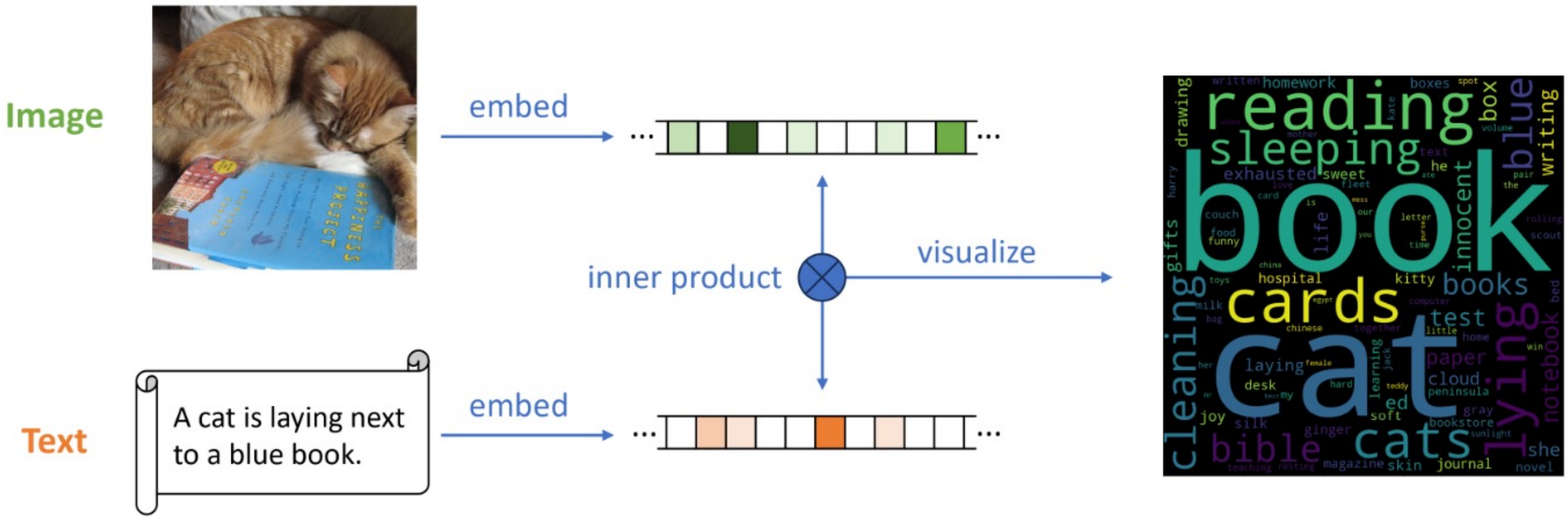


Image Disentanglement

Image Disentanglement (Patch-level)



Retrieval Reasoning



Q & A

Code: https://github.com/jzhoubu/VDR

Email: jzhoubu@connect.ust.hk