

Probabilistic Language-Image Pre-training



Sanghyuk Chun



Wonjae Kim



Song Park



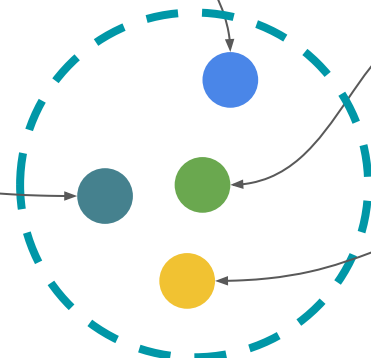
Sangdoo Yun



NAVER AI LAB

* slide can be found in <https://sanghyukchun.github.io/home/>

Image-Text matching is inherently many-to-many



a train is on a track next to a platform.



A common concept

Image-Text matching is inherently many-to-many



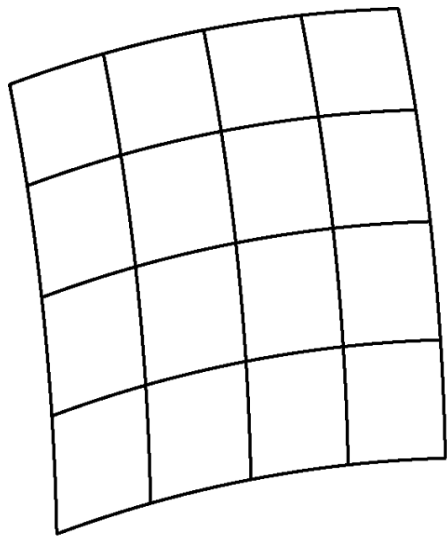
people waiting to board a train in a train station.



the metro train has pulled into a large station.

a train is on a track next to a platform.

How deterministic space oversimplifies multiplicity?



(a) Deterministic embedding

“Person”

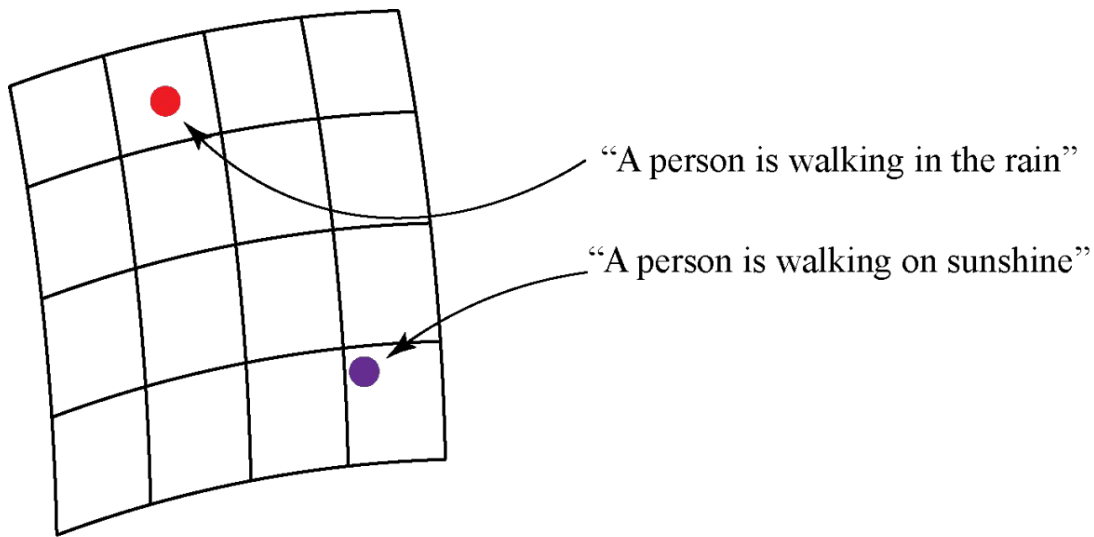
“A person is walking in the rain”

“A person is walking on sunshine”

“A person is walking”

(b) Probabilistic embedding

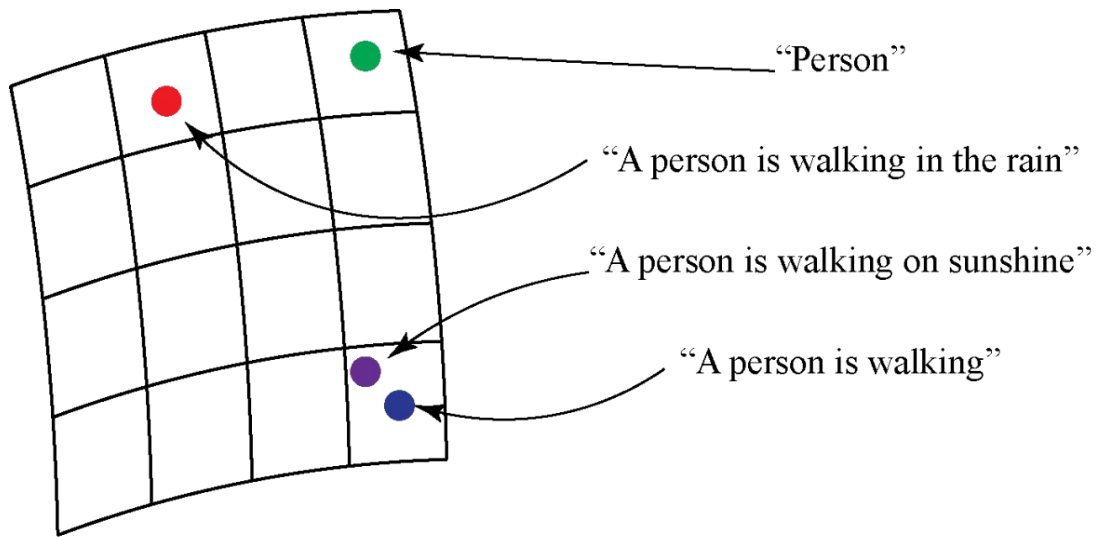
Det. space maps an input to a specific vector coord.



(a) Deterministic embedding

(b) Probabilistic embedding

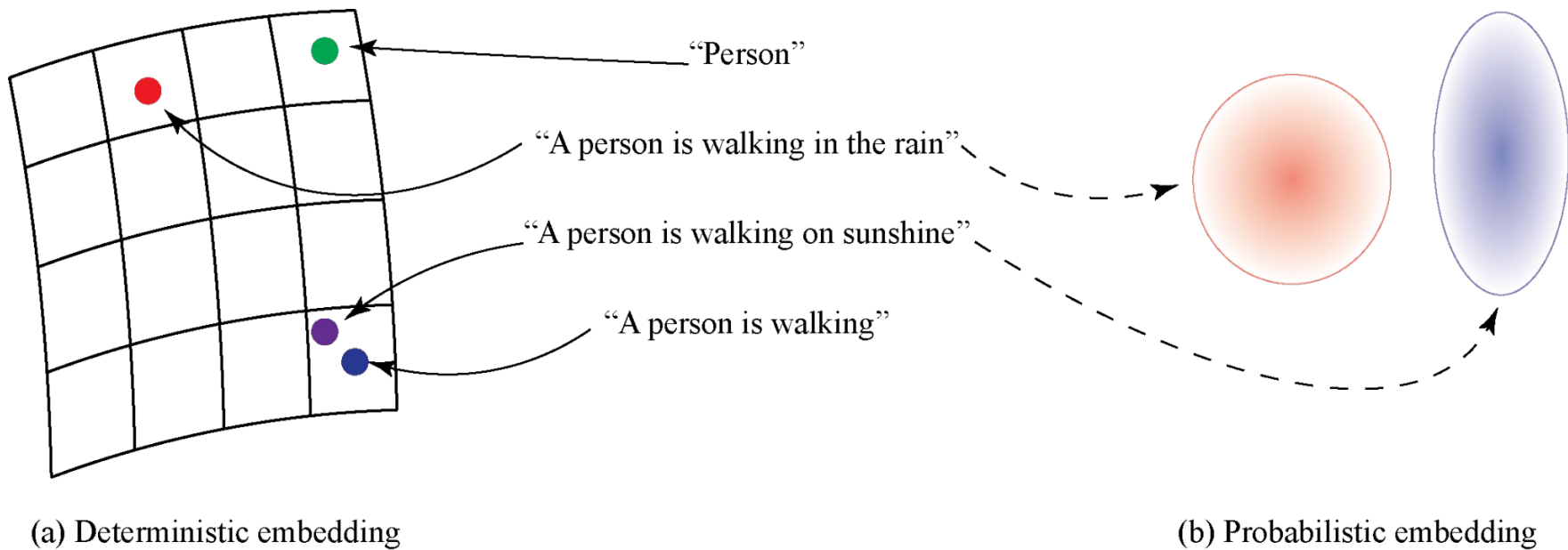
Even if an input can be matched to multiple instances, it will be mapped to a specific coord.



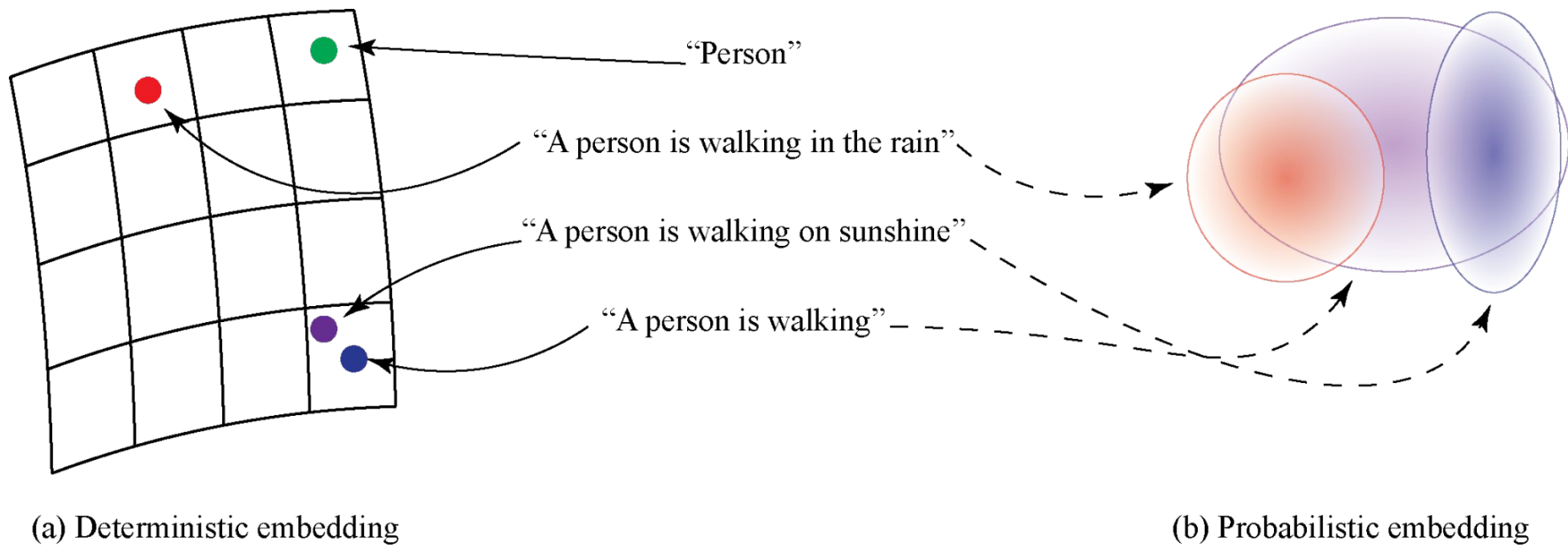
(a) Deterministic embedding

(b) Probabilistic embedding

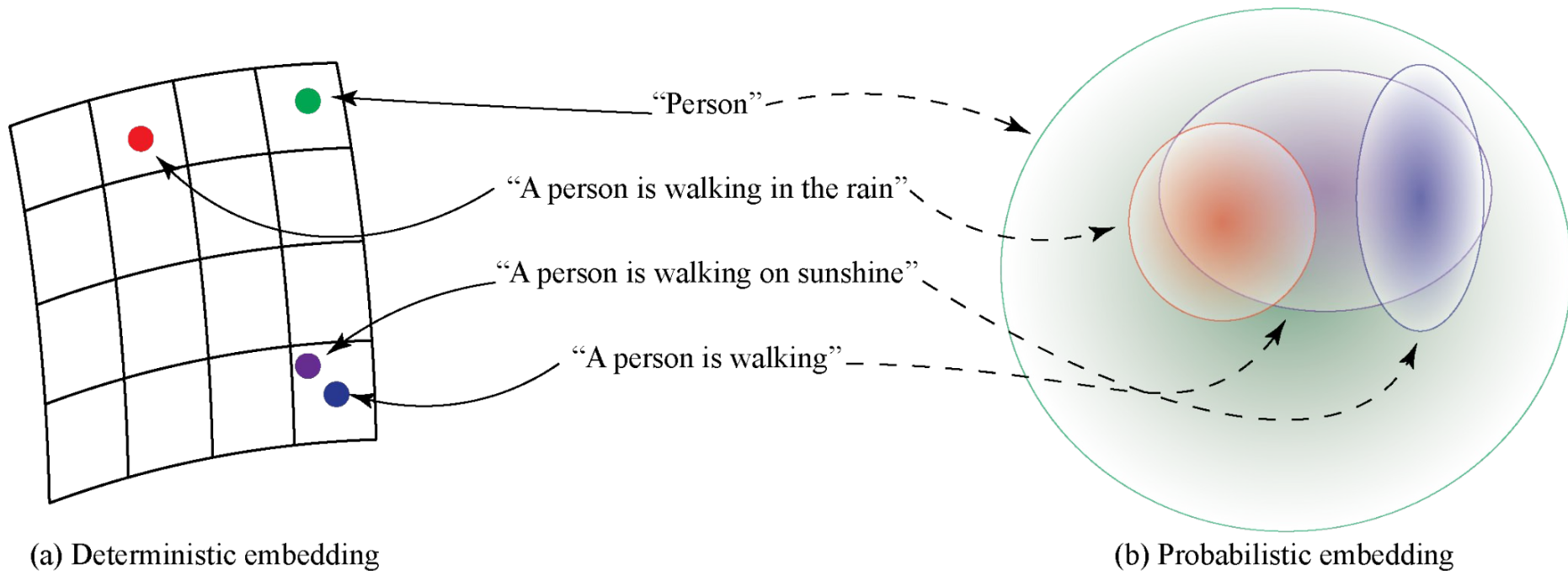
Prob. space has additional info. axis, “uncertainty”



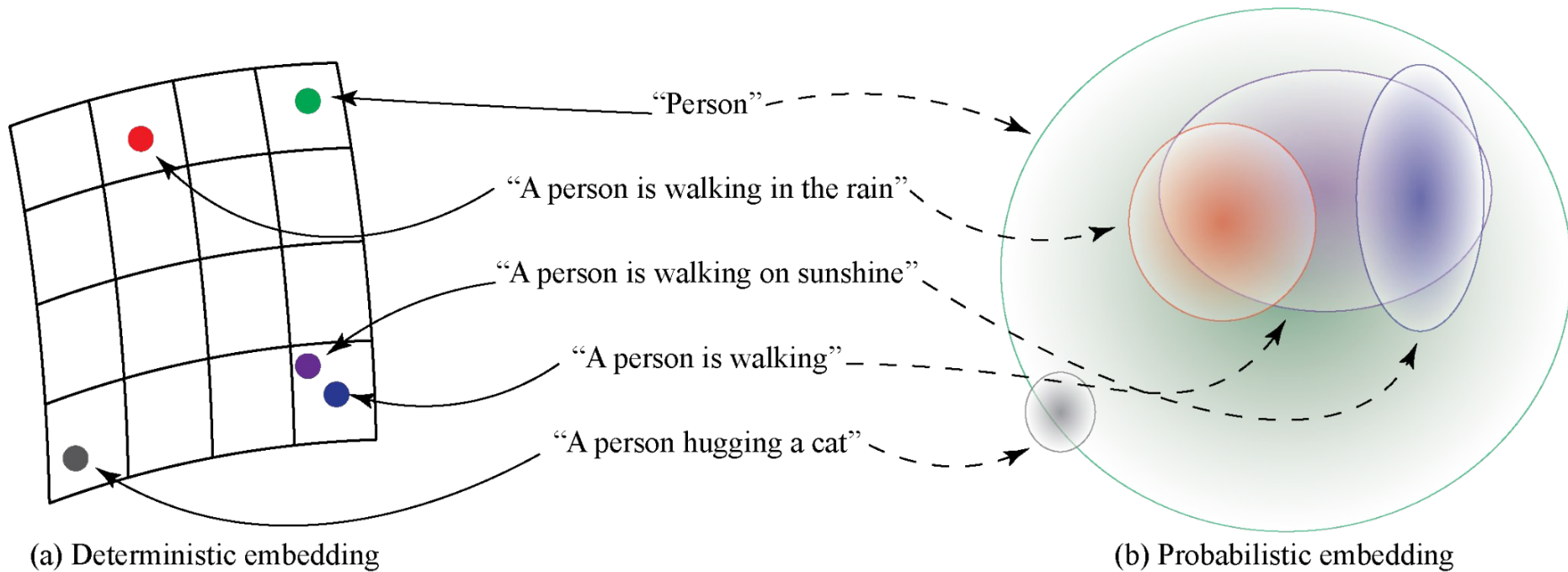
An uncertain input will have a large uncertainty value



We can easily capture input uncertainty by “variance”



Real-world scenarios require handling uncertainty

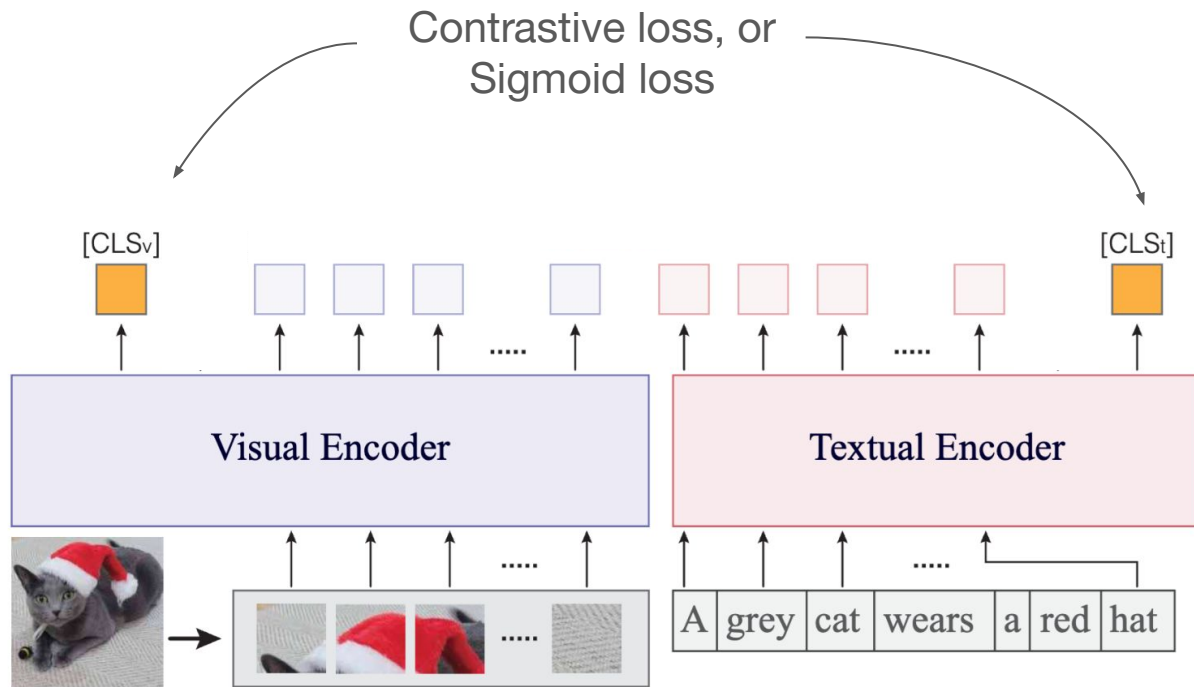


Contribution

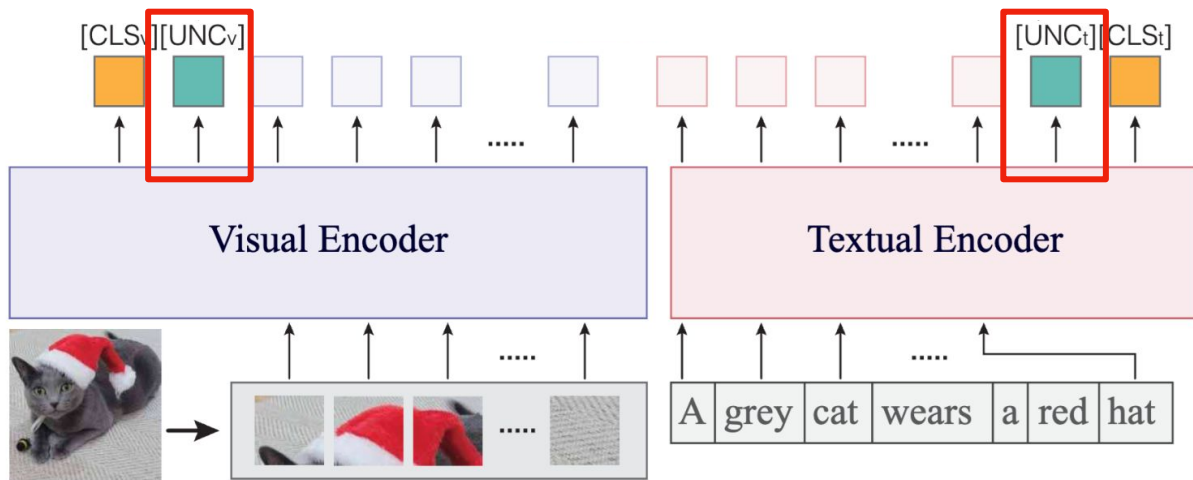
Method	Pre-trained model	Uncertainty architecture	Optimization objective	Training dataset	Evaluation dataset
PCME (Chun et al., 2021)	ResNet-152 (ImageNet) GloVe (pre-trained)	PIE-Net (Song & Soleymani, 2019)	MCSofContrastive (Oh et al., 2019)	MS-COCO / CUB Caption	MS-COCO / CUB Caption
ProbVLM (Upadhyay et al., 2023)	CLIP (OpenAI) (Frozen)	MLP (Generalized Gaussian)	intra-modal alignment & cross-modal alignment	MS-COCO / CUB Caption	MS-COCO / CUB Caption
PCME++ (Chun, 2024)	CLIP (OpenAI)	1-layer Transformer	BCE with CSD	MS-COCO	ECCV Caption (Chun et al., 2022)
PCME++ (Chun, 2024)	-	1-layer Transformer	BCE with CSD & CLIP loss	CC3M + CC12M + RedCaps	ImageNet ZS
ProLIP (Chun et al., 2024)	-	[UNC] token	PPCL & Inclusion loss	DataComp 1B	38 ZS tasks

- ProLIP is the first PrVLM pre-trained on a billion-scale image-text dataset using only probabilistic objectives, and shows strong ZS classification even compared to CLIP/SigLIP
- We tackle three issues of the previous PrVLMs: **(1) efficient uncertainty architecture**, **(2) better optimization objective**, and **(3) enforces a desired behavior by additional “inclusion” loss**.

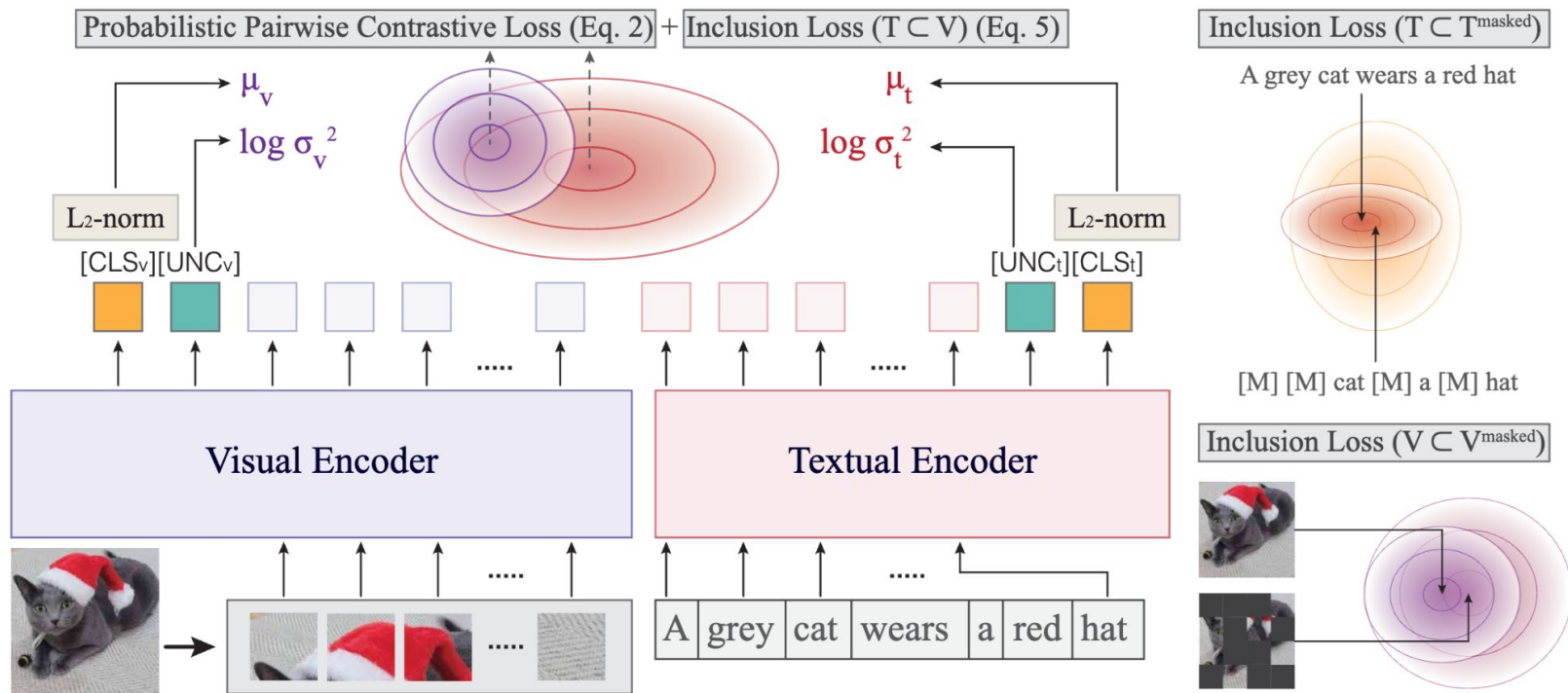
Standard Language-Image Pre-training



Efficient uncertainty architecture by [UNC] token

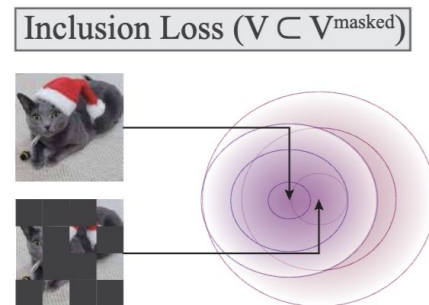
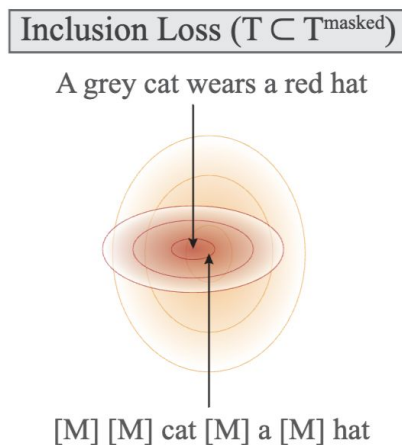
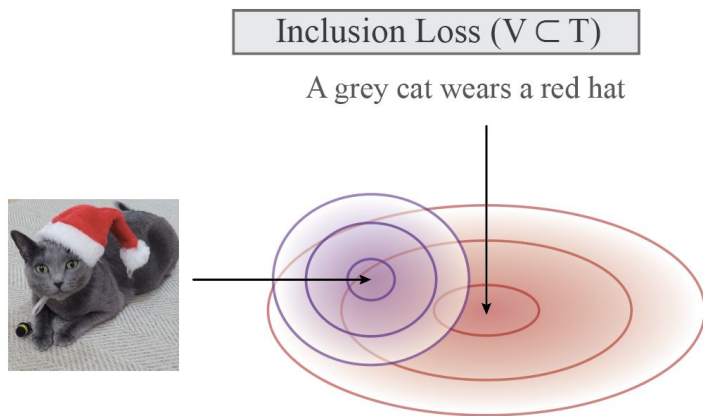


ProLIP Architecture & Objective functions



Inclusion loss for enforcing inclusive relationship

- Two inclusive relationships
 - Texts include images
 - Masked input includes the original input

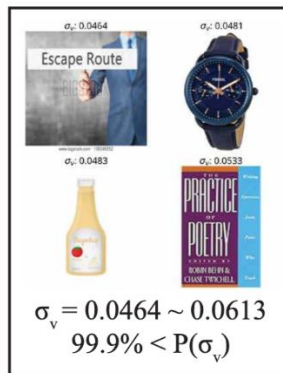
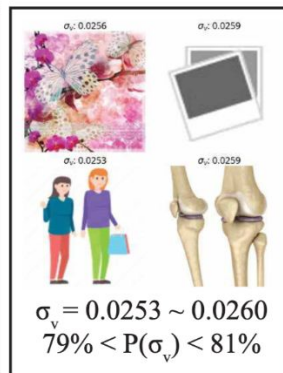
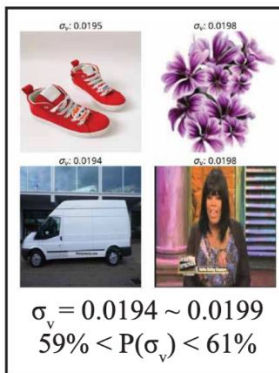
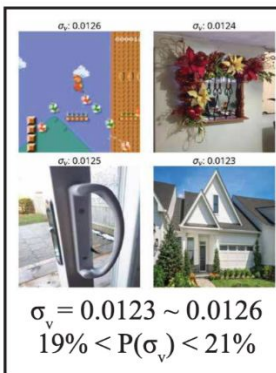
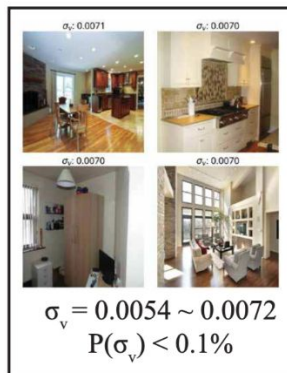


Zero-shot classification

		# Samples Seen	ImageNet	IN dist. shifts	VTAB	Retrieval	Average
ViT-B/16	CLIP	1.28B	67.2	55.1	56.9	53.4	57.1
	SigLIP	1.28B	67.4	55.4	55.7	53.4	56.7
	ProLIP	1.28B	67.8	55.3	58.5	53.0	57.9
	ProLIP	12.8B	74.6	63.0	63.7	59.6	63.3
ViT-L/16	ProLIP	1.28B*	79.4	68.6	64.0	61.3	65.9
ViT-SO400M/14	ProLIP	1.28B*	79.3	69.0	65.1	62.5	66.6

- All models are available at <https://huggingface.co/collections/SanghyukChun/prolip-6712595dfc87fd8597350291>

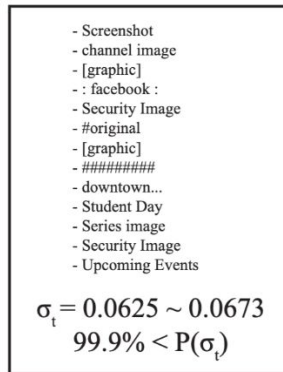
Certain & uncertain samples



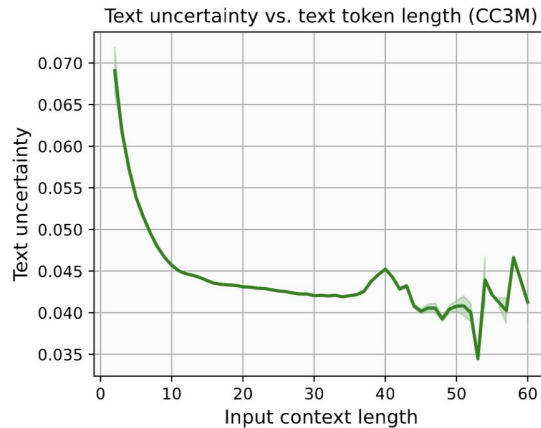
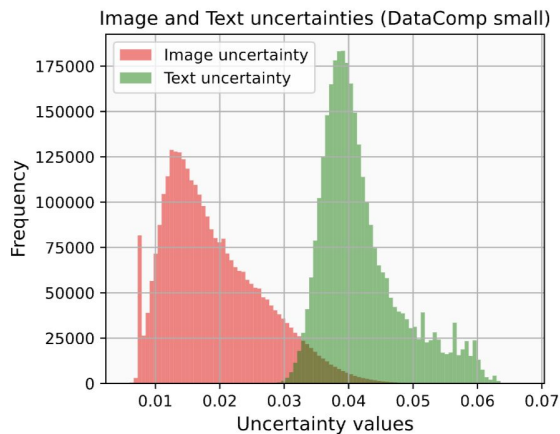
← More certain samples



More uncertain samples→



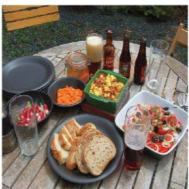




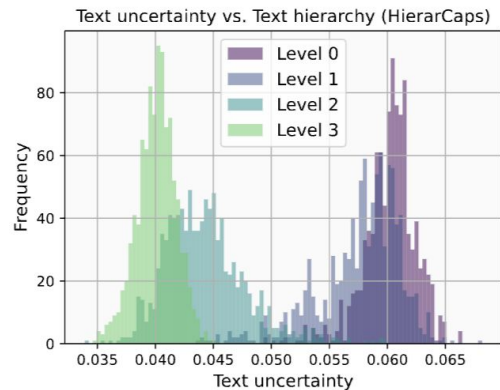
Property of learned uncertainty



- Texts include image
- Shorter texts are more uncertain than longer texts

Image traversal with uncertainty

Query images						
Top-1 Retrieved captions	↓ Most inclusive captions	Kite surfer in the air on top of a red board	A blue bird sitting on the top of a branch with autumn leaves	An outdoor table containing assorted bowls of food and beer	A large red chair has a horse statue on it	Leaves and purple flowers come out of a brown vase on a desk
		Kite surfer in the air on top of a red board	A blue bird sitting on the top of a branch with autumn leaves	An outdoor table containing assorted bowls of food and beer	A large red chair has a horse statue on it	Leaves and purple flowers come out of a brown vase on a desk
		Kite surfer in the air	Blue bird sitting on top of branch	Outdoor picnic	Red chair has a horse statue	Vase with flowers in it
		kite surfing	bird	picnic	red chair	vase
HierarCaps GTs	Level 3	Kite surfer in the air on top of a red board	A blue bird sitting on the top of a branch with autumn leaves	An outdoor table containing assorted bowls of food and beer	A large red chair has a horse statue on it	There is a small glass vase that has purple flowers in it
	Level 2	Kite surfer on top of the board	Blue bird sitting on top of branch	Outdoor table with food and beer	Red chair has a horse statue	Vase with flowers in violet
	Level 1	kite surfing	blue bird	outdoor table	red chair	vase
	Level 0	water sports	bird	outdoor	chair	object



References

- **[CVPR 2021] Probabilistic Embeddings for Cross-Modal Retrieval.**
Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, Diane Larlus
- **[ECCV 2022] ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO.**
Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, Seong Joon Oh
- **[ICLR 2024] Improved Probabilistic Image-Text Representations.**
Sanghyuk Chun
- **[ICLR 2025] Probabilistic Language-Image Pre-training.**
Sanghyuk Chun, Wonjae Kim, Song Park, Sangdoo Yun
- **[ICLR WS 2025] LongProLIP: A Probabilistic Vision-Language Model with Long Context Text.**
Sanghyuk Chun, Sangdoo Yun

Resources

- Github: <https://github.com/naver-ai/prolip>
 - Inference code & training code are available
- Model collections:
<https://huggingface.co/collections/SanghyukChun/prolip-6712595dfc87fd8597350291>
 - ViT-B/16 from-scratch model
 - ViT-L/16, ViT-SO400M/14, ViT-H/14 fine-tuned models
 - LongProLIP models (ProLIP with longer text context)