# Probabilistic Language-Image Pre-training









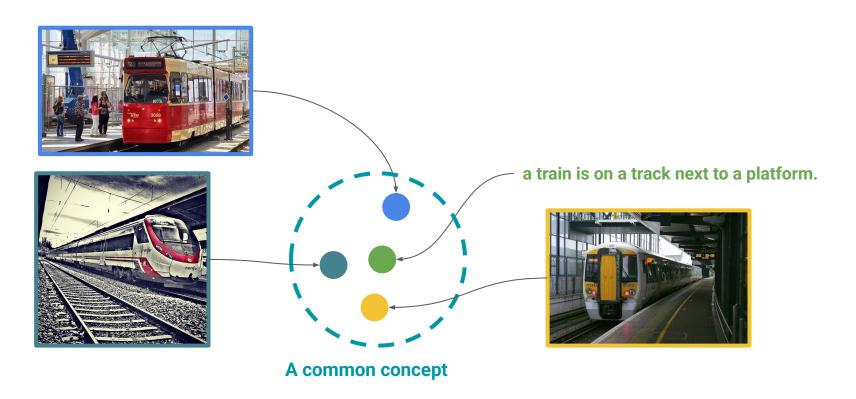
**Song Park** 



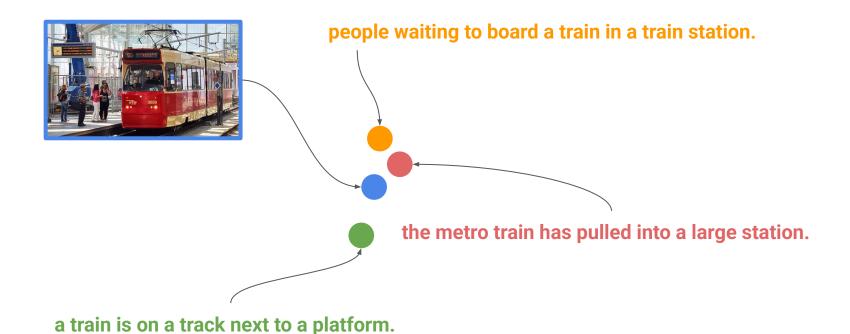
Sangdoo Yun



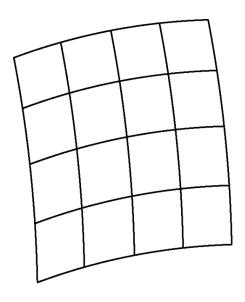
#### Image-Text matching is inherently many-to-many



#### Image-Text matching is inherently many-to-many



## How deterministic space oversimplifies multiplicity?



"Person"

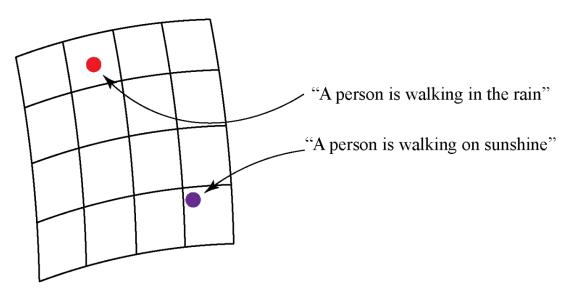
"A person is walking in the rain"

"A person is walking on sunshine"

"A person is walking"

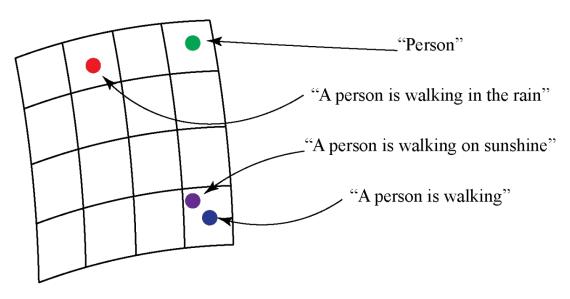
(a) Deterministic embedding

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



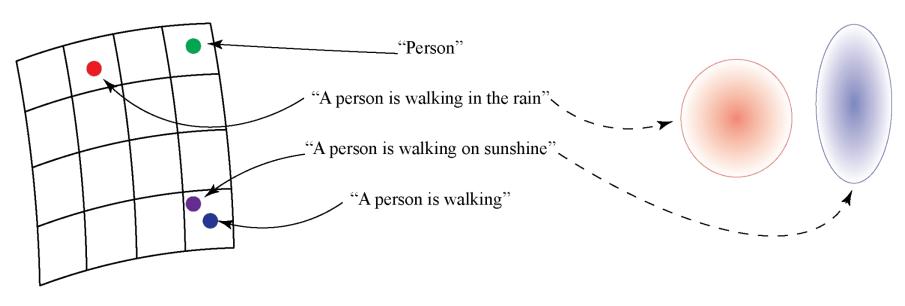
(a) Deterministic embedding

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



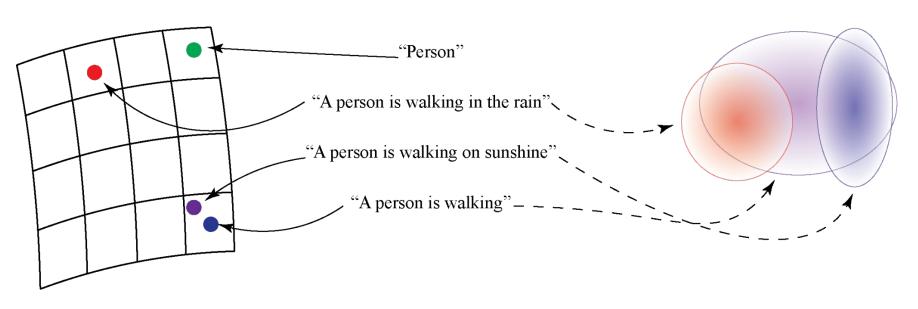
(a) Deterministic embedding

#### Prob. space has additional info. axis, "uncertainty"



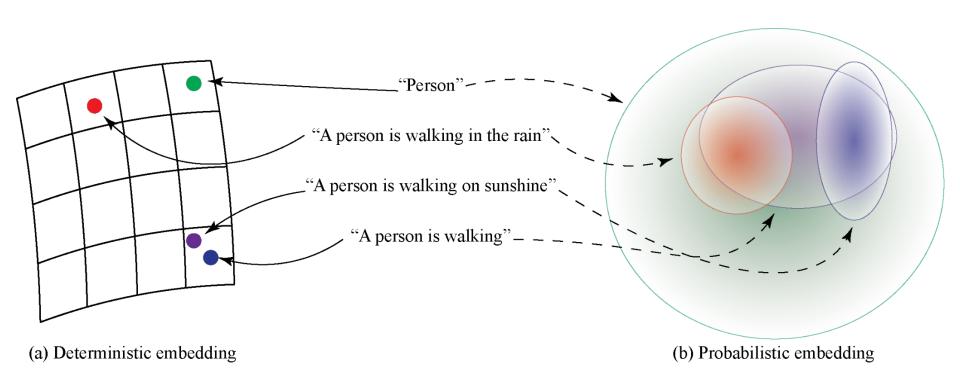
(a) Deterministic embedding

#### An uncertain input will have a large uncertainty value

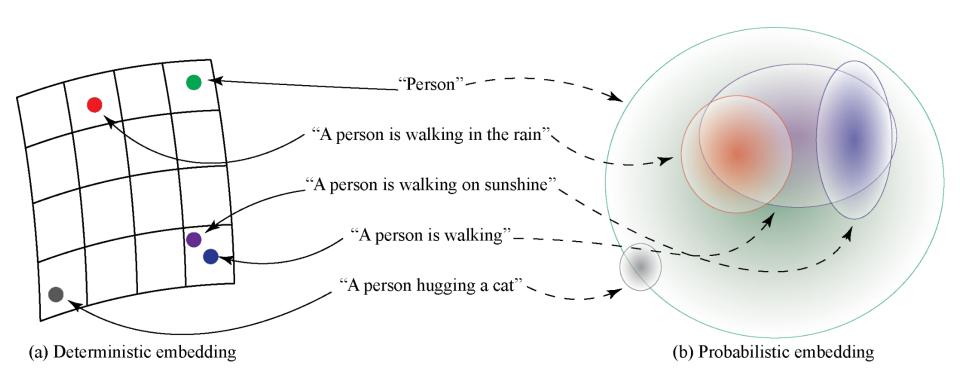


(a) Deterministic embedding

#### We can easily capture input uncertainty by "variance"



#### Real-world scenarios require handling uncertainty

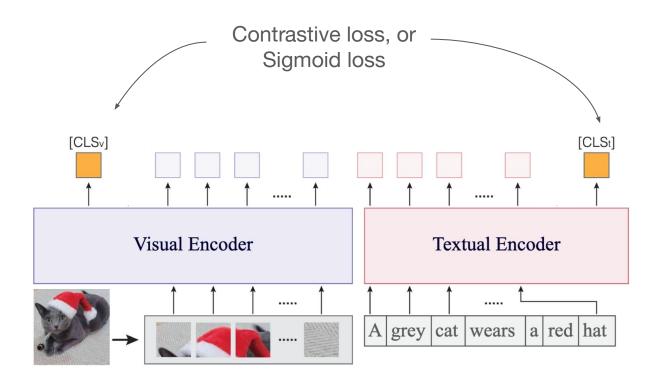


#### Contribution

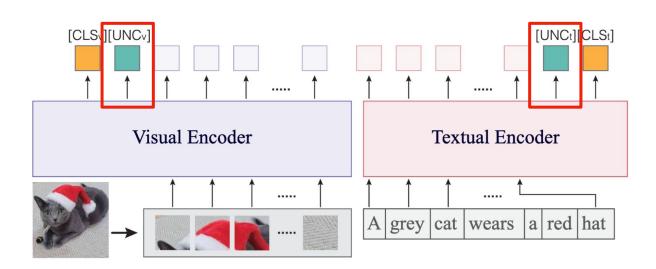
Method	Pre-trained model	Uncertainty architecture	Optimization objective	Training dataset	Evaluation dataset
PCME (Chun et al., 2021)	ResNet-152 (ImageNet) GloVe (pre-trained)	PIE-Net (Song & Soleymani, 2019)	MCSoftContrastive (Oh et al., 2019)	MS-COCO / CUB Caption	MS-COCO / CUB Caption
ProbVLM (Upadhyay 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.

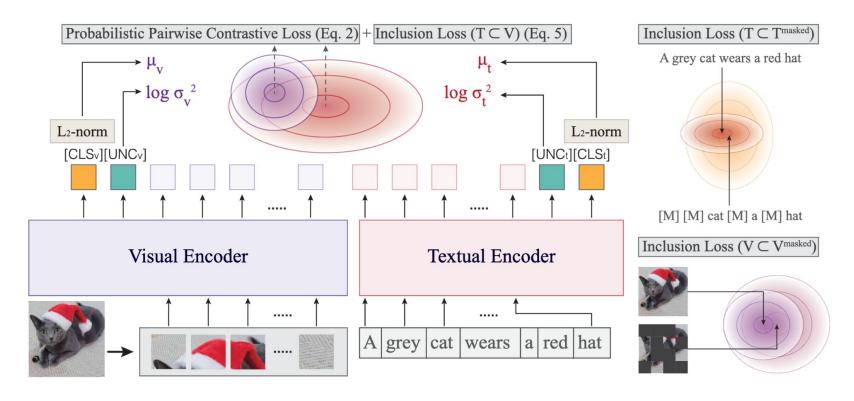
### **Standard Language-Image Pre-training**



#### Efficient uncertainty architecture by [UNC] token

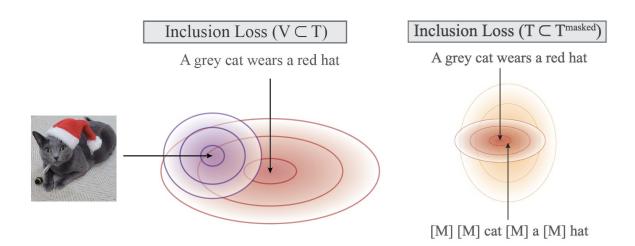


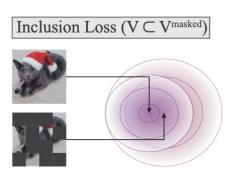
#### **ProLIP Architecture & Objective functions**



#### Inclusion loss for enforcing inclusive relationship

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





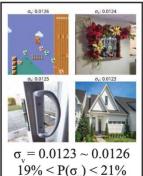
#### **Zero-shot classification**

		# 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

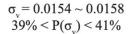
 All models are available at <a href="https://huggingface.co/collections/SanghyukChun/prolip-6712595dfc87fd85">https://huggingface.co/collections/SanghyukChun/prolip-6712595dfc87fd85</a>
 <a href="https://prolip-6712595dfc87fd85">97350291</a>

#### Certain & uncertain samples

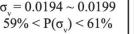




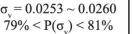


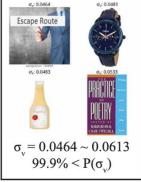












#### ← More certain samples

 $P(\sigma_{...}) < 0.1\%$ 

- 3BR house 10mins to Gaisano Mall & the city proper - Butuan City - 独立屋
- Afia Schwar wildly jubilates after Mahama won NDC presidential primaries
- Felipe da. Male, 27, \$350, No pets.
- No children, and Non-smoker Valeri Bozhinov: I hope God is a Bulgarian
- Philippa Hearnden, Head of Business, moving on to new opportunity.
- "I was lucky to meet you." said former NBC News producer Truus Bos, who first J.B. Rutagarama as a teenager at a Rwandan refugee camp.

$$\sigma_{t} = 0.0264 \sim 0.0303$$
 $P(\sigma_{t}) < 0.1\%$ 

- extech rh520a chart recorder, humidity & temperature
- Malibu Beach Inn Updated 2020
- Cat on books seen in Ischia Ponte. Campagna/Italy
- Living Room Table Design Wooden Dining Room Comely Image Of Furniture For Home Interior And
- Maruti Suzuki Alto K10 (AC) (Fully loaded) 2010 Petrol Well Maintained
- Kyoto of Japan Round [10"] Pine Cone Vegetable Dish
- 173 Crescent St, Rockland, MA 02370

$$\sigma_{t} = 0.0369 \sim 0.0371$$
  
 $19\% < P(\sigma_{t}) < 21\%$ 

- Measuring the Effectiveness of Your Church Website (Google Analytics) (1).png
- Mary Gypsioti
- robe longue de plage beauty roxy marshmallow land of tehotihuac roxy la redoute. Black Bedroom Furniture Sets. Home Design Ideas
- Lancashire Tourism Awards 2015
- Signature Line Celery 1 x 3 Glass Subway Tile in Green/Brown/Gray by Susan Jablon
- Catherine J Terry
- The Baroque route, Bolivia

$$\sigma_{t} = 0.0391 \sim 0.0393$$
  
 $39\% < P(\sigma_{t}) < 41\%$ 

- Places To Go In South Island
- huawei mate xs
- LEGO Pirates Torso without Arms
- Chocolate Frosted Cupcakes on a Plate Photo Classic Round Sticker
- Boxer Watercolor Greeting Card by Naxart Studio
- The Sabir Bey Show apk screenshot
- Primary photo for Pat Hogan, Deceased
- Boston skyline in purple radiant orchid vector
- Kitting For<br/>br /> All Industries
- Down by 2 in the past 45 days

$$\sigma_{t} = 0.0415 \sim 0.0418$$
 $59\% < P(\sigma_{t}) < 61\%$ 

- And the obligatory coffee
- Jay Jay Ink
- Men's summer sport coats - Chefs for Seals Facebook Page
- EZ(11枚目)
- Mary Thompson
- smart lighting for home
- Auto Speaker- screenshot thumbnail
- Hotel photo: My Hostel
- The Life and Selected Writings Classic ankle boots - grey
- ©TecRail
- Severs Table 1.0

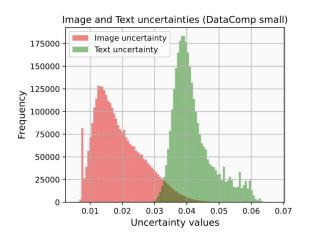
$$\sigma_{t} = 0.0460 \sim 0.0468$$
 $79\% < P(\sigma_{t}) < 81\%$ 

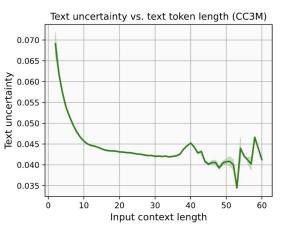
#### More uncertain samples→

- Screenshot
- channel image
- [graphic]
- : facebook :
- Security Image
- #original - [graphic]
- ##########
- downtown...
- Student Day
- Series image
- Security Image
- Upcoming Events

 $\sigma = 0.0625 \sim 0.0673$  $99.9\% < P(\sigma)$ 

#### **Property of learned uncertainty**

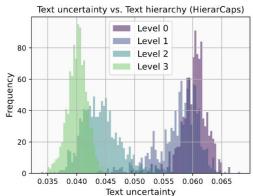




- Texts include image
- Shorter texts are more uncertain than longer texts

#### Image traversal with uncertainty





#### 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: <a href="https://github.com/naver-ai/prolip">https://github.com/naver-ai/prolip</a>
  - 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)