Position: Multiplicity is an Inevitable and Inherent Challenge in Multimodal Learning

Sanghyuk Chun
sanghyukc@princeton.edu

Olga Russakovsky

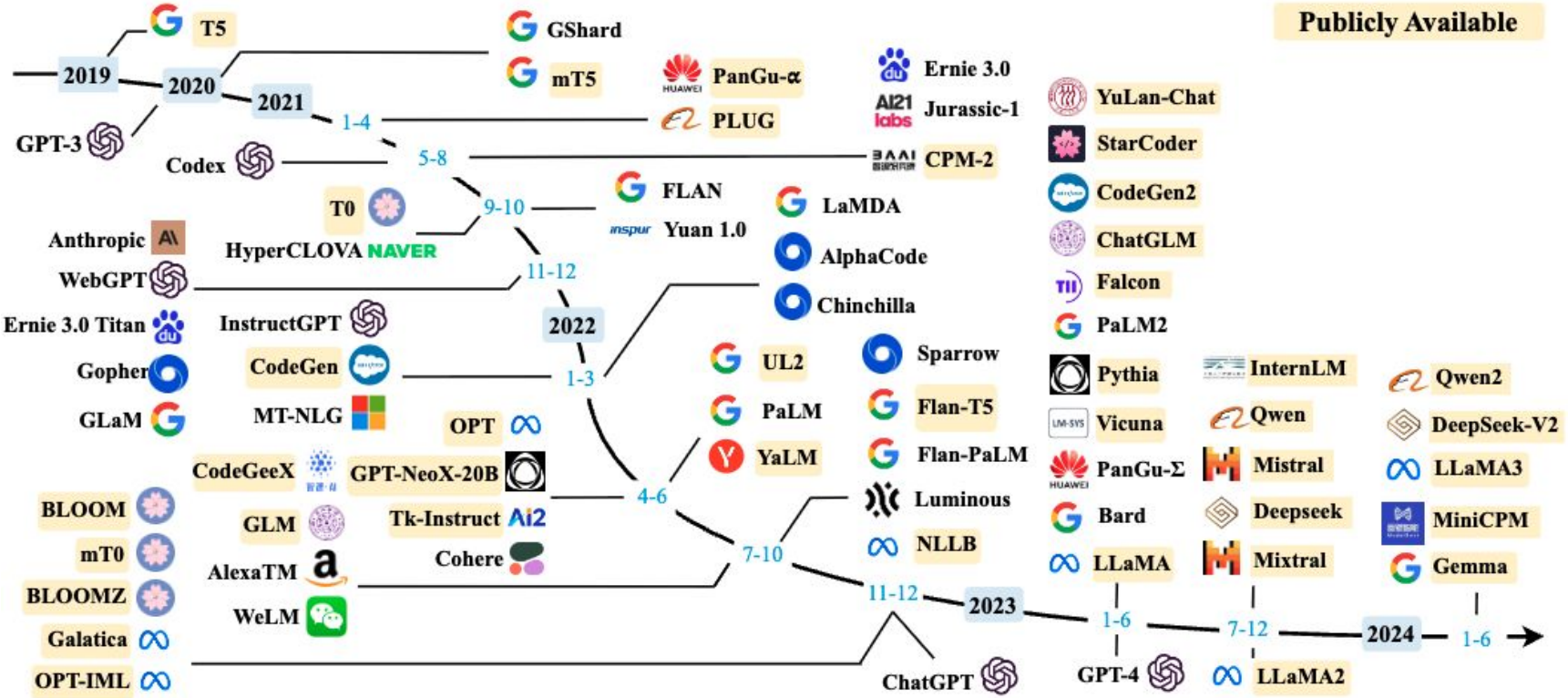
Contents

- Introduction: **The multiplicity problem** in multimodal learning
- How can we handle **the multiplicity in contrastive learning**?
- How can we handle **the multiplicity in evaluation**?
- How can we handle **the multiplicity in dataset construction**?
- Conclusion

Contents

- Introduction: **The multiplicity problem** in multimodal learning
- How can we handle **the multiplicity in contrastive learning**?
- How can we handle **the multiplicity in evaluation**?
- How can we handle **the multiplicity in dataset construction**?
- Conclusion

Large foundation models

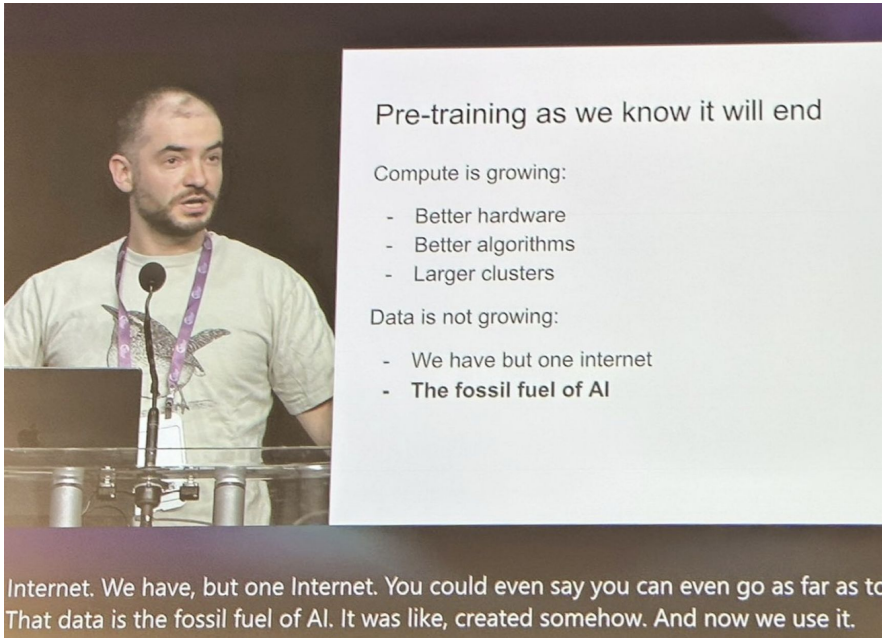


A Survey of Large Language Models, https://arxiv.org/abs/2303.18223

Large foundation models

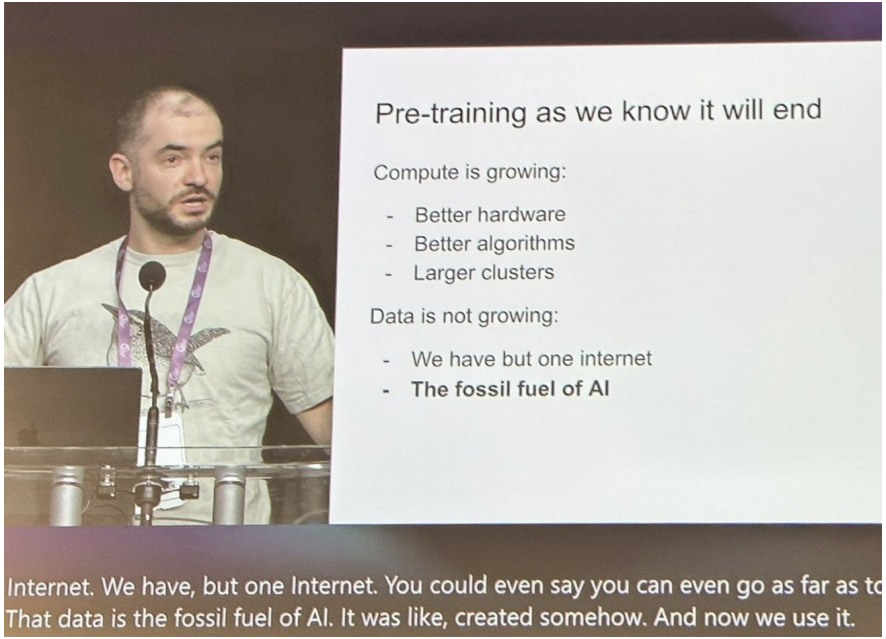
User	What is funny about this image? Describe it panel by panel. 
GPT-4	The image shows a "Lightning Cable" adapter package with three panels. Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port. Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it. Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end. The humour in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

Where are we going?



* During NeurIPS 2024 Test-of-time award talk

New capability and challenges



Dhruv Batra @DhruvBatraDB · 10m

Brilliant talk by @ilyasut, but he's wrong on one point.

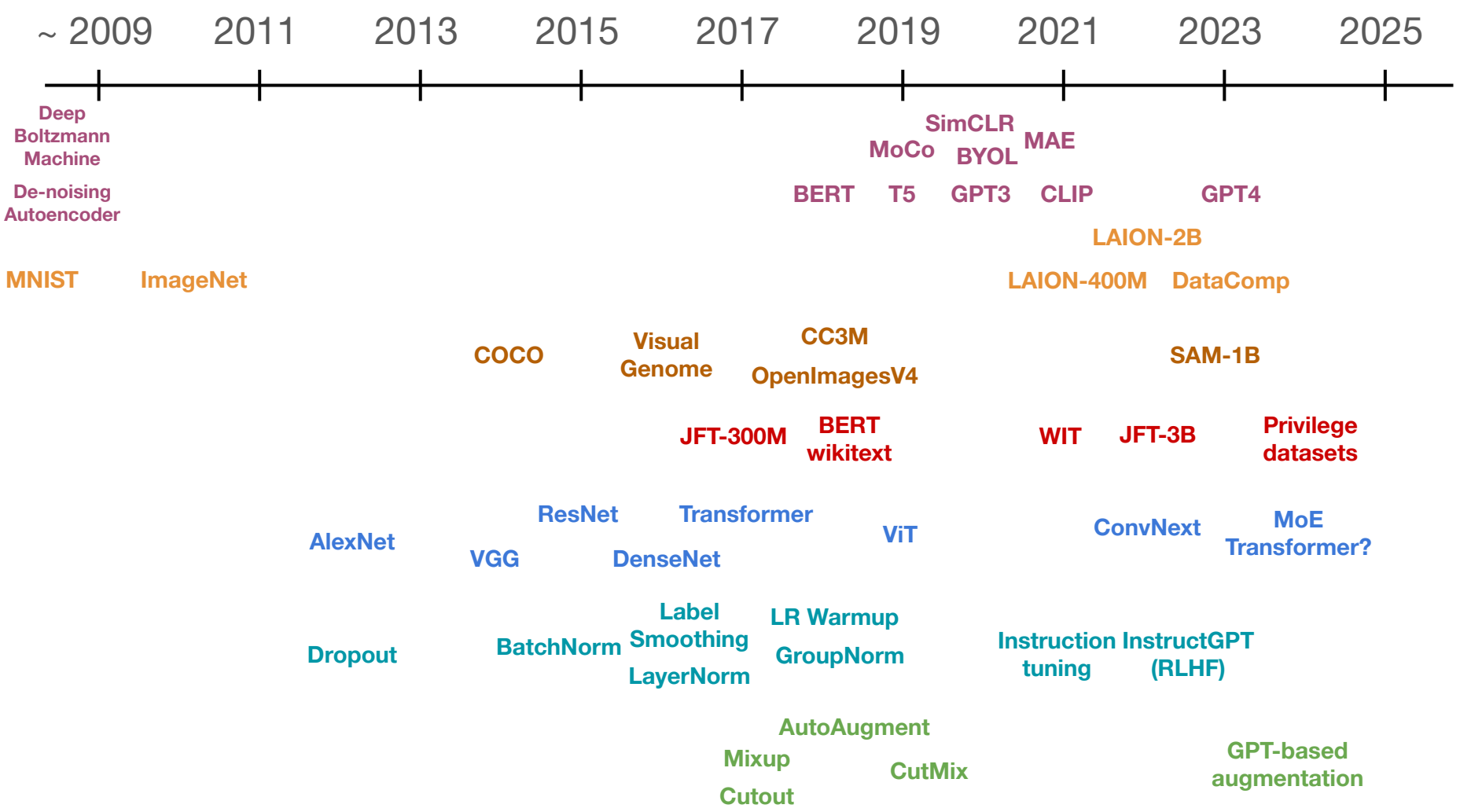
We are NOT running out of data. We are running out human-written text.

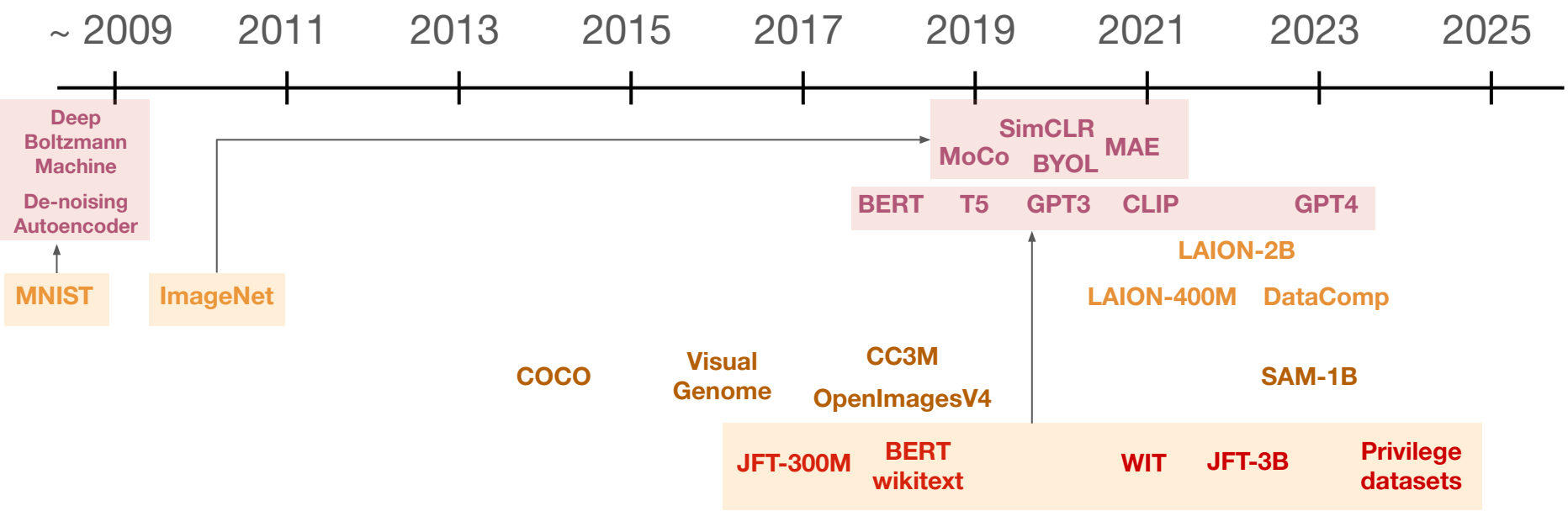
We have more videos than we know what to do with. We just haven't solved pre-training in vision.

Just go out and sense the world. Data is easy.

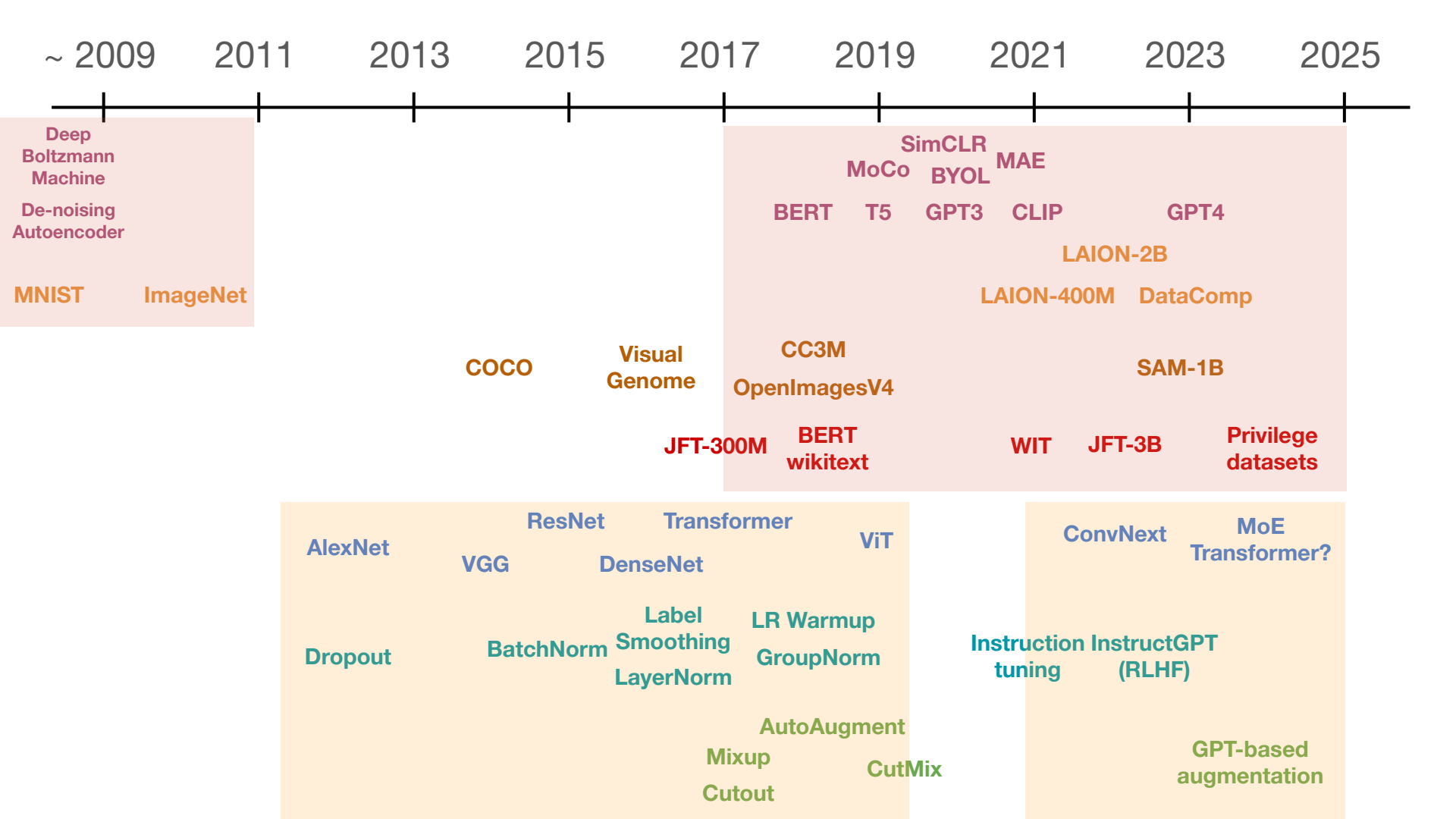
https://x.com/DhruvBatraDB/status/1868009853324865762

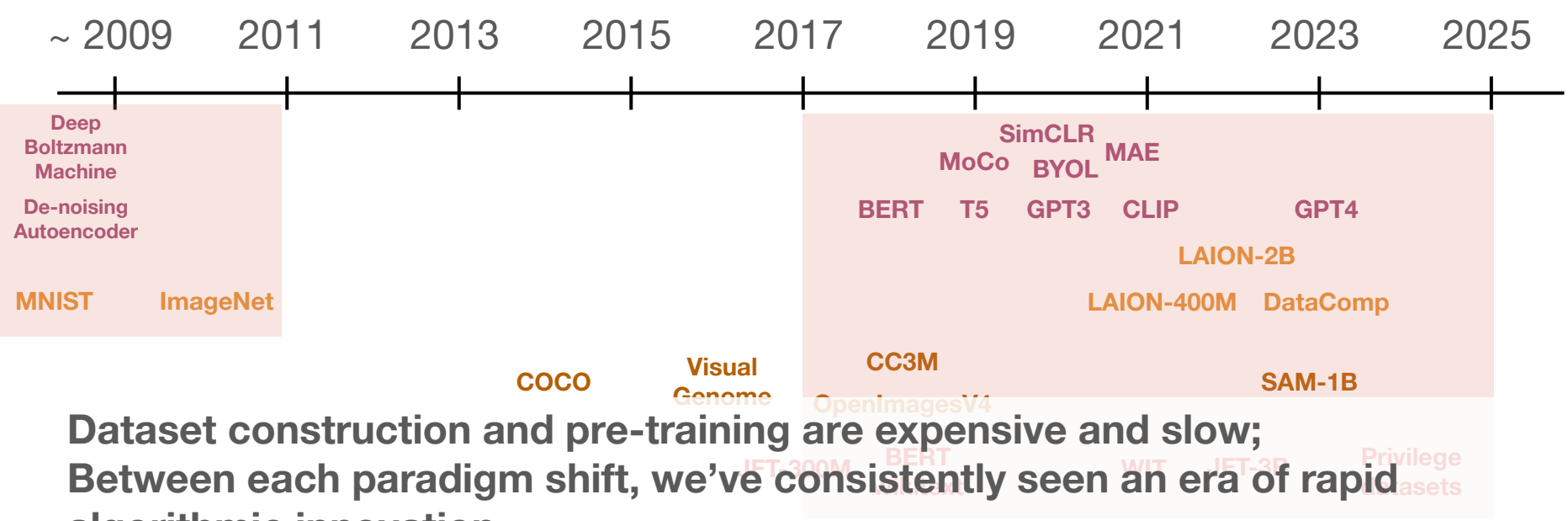
* During NeurIPS 2024 Test-of-time award talk


~ 2009
2011
2013
2015
2017
2019
2021
2023
2025
Deep Boltzmann Machine
De-noising Autoencoder
SimCLR
MoCo
BYOL
MAE
BERT
T5
GPT3
CLIP
GPT4
LAION-2B
MNIST
ImageNet
LAION-400M
DataComp
COCO
Visual Genome
CC3M
OpenImagesV4
SAM-1B
JFT-300M
BERT wikitext
WIT
JFT-3B
Privilege datasets
ResNet
Transformer
AlexNet
VGG
DenseNet
ViT
ConvNext
MoE Transformer?
Label Smoothing
LR Warmup
Dropout
BatchNorm
LayerNorm
GroupNorm
Instruction tuning
InstructGPT (RLHF)
AutoAugment
Mixup
CutMix
Cutout
GPT-based augmentation



(Empirically)
Large-scale datasets typically come first, followed by pre-training

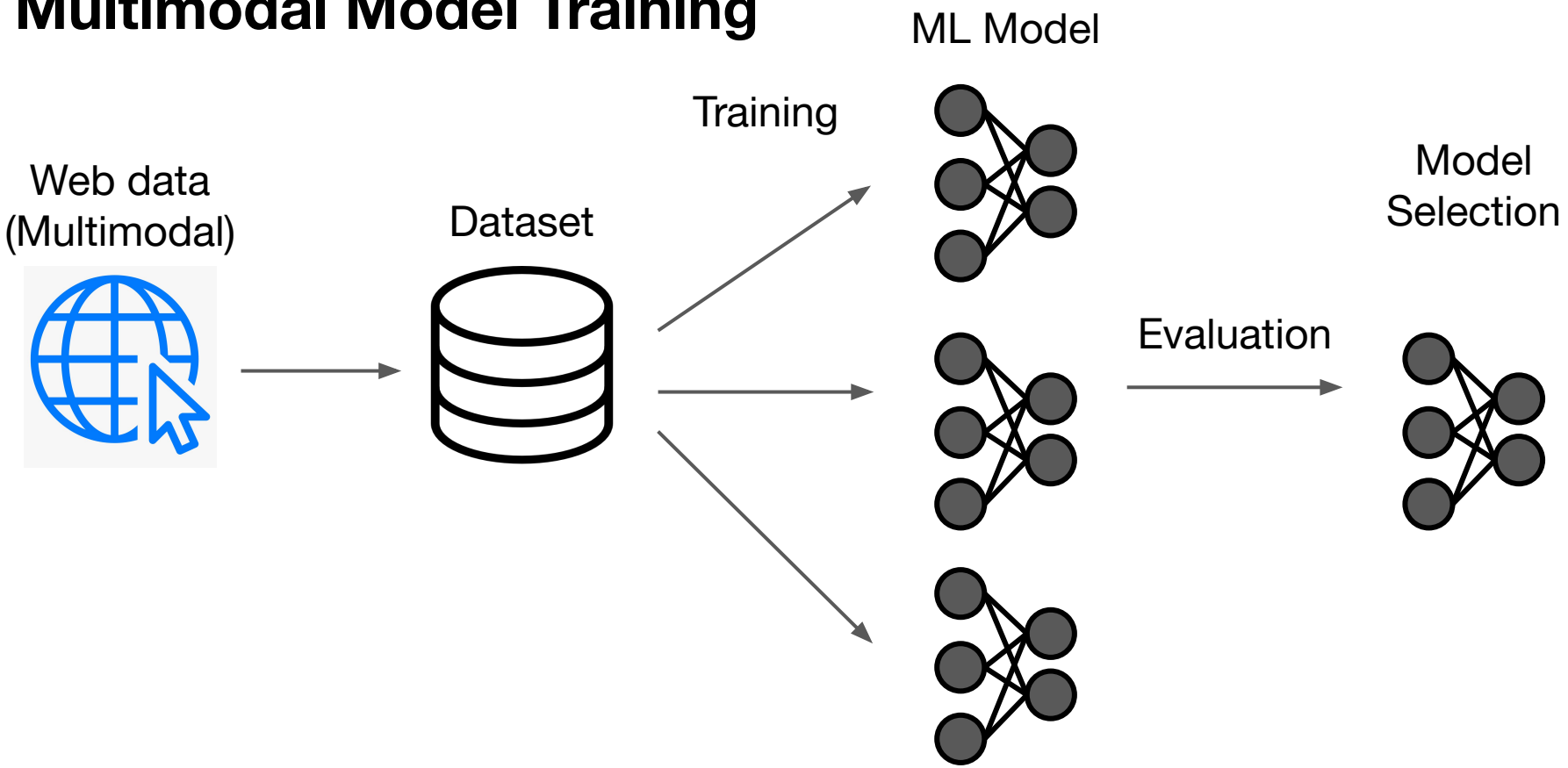

~ 2009
2011
2013
2015
2017
2019
2021
2023
2025
Deep Boltzmann Machine
De-noising Autoencoder
MNIST
ImageNet
SimCLR
MoCo
BYOL
MAE
BERT
T5
GPT3
CLIP
GPT4
LAION-2B
LAION-400M
DataComp
COCO
Visual Genome
CC3M
OpenImagesV4
SAM-1B
JFT-300M
BERT wikitext
WIT
JFT-3B
Privilege datasets
AlexNet
VGG
ResNet
DenseNet
Transformer
ViT
ConvNext
MoE Transformer?
Dropout
BatchNorm
Label Smoothing
LayerNorm
LR Warmup
GroupNorm
Instruction tuning
InstructGPT (RLHF)
AutoAugment
Mixup
Cutout
CutMix
GPT-based augmentation



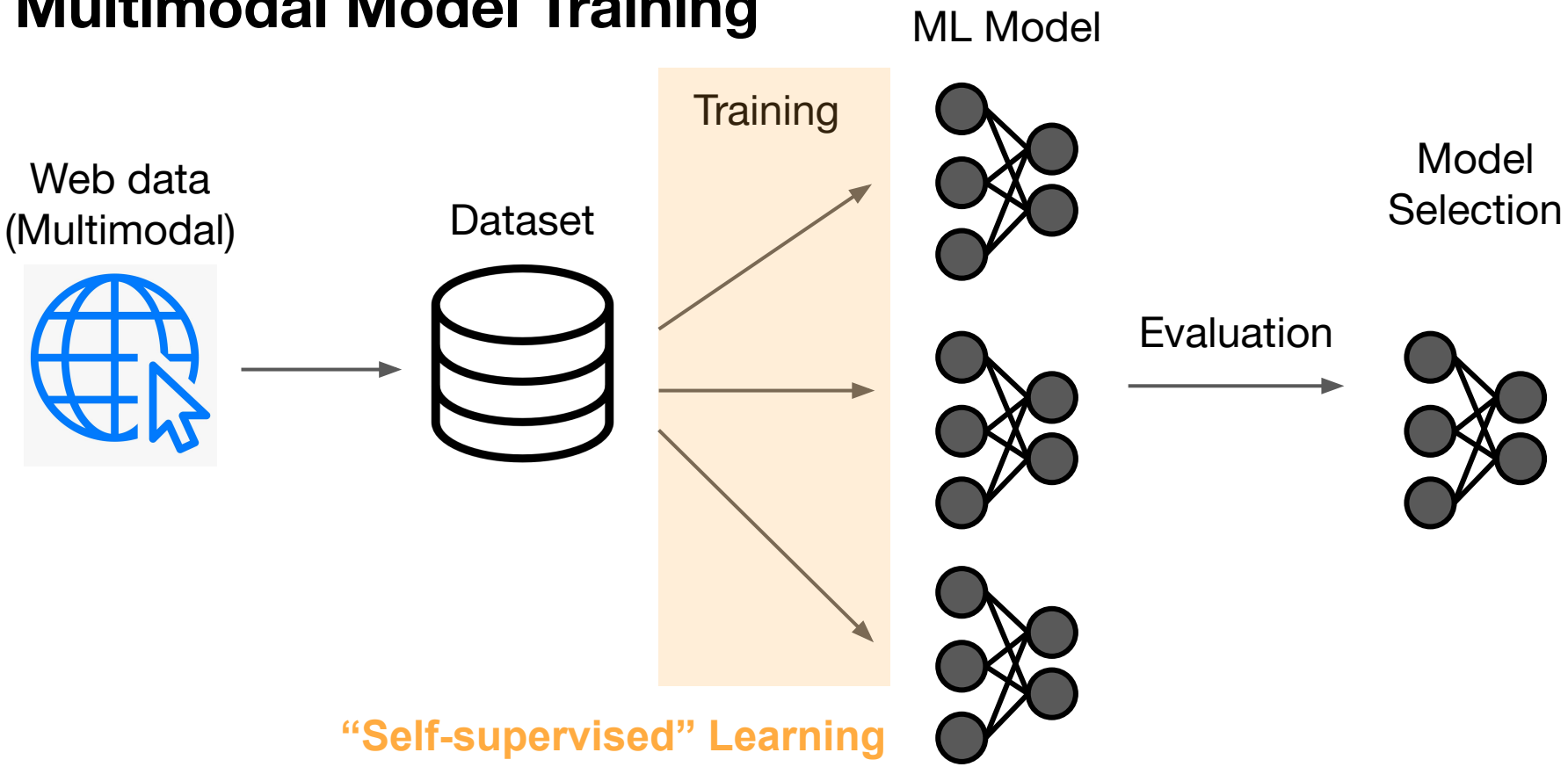
Dataset construction and pre-training are expensive and slow; Between each paradigm shift, we've consistently seen an era of rapid algorithmic innovation.

Question: What are the current challenges in multimodal learning?

Multimodal Model Training

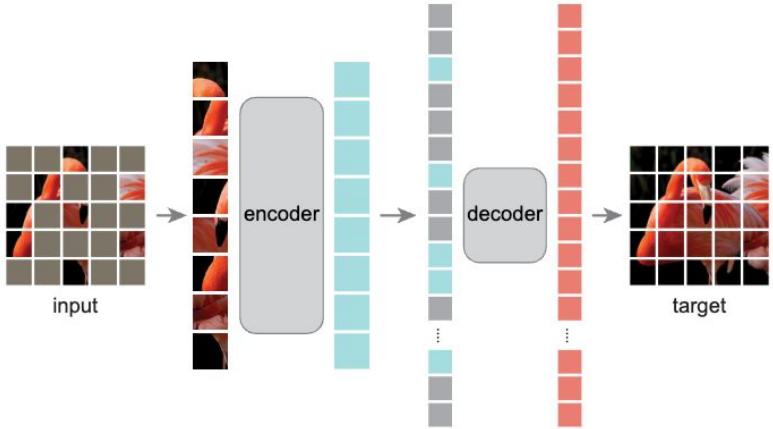


Multimodal Model Training



Self-Supervised Learning

Visual data (Masked Autoencoder)

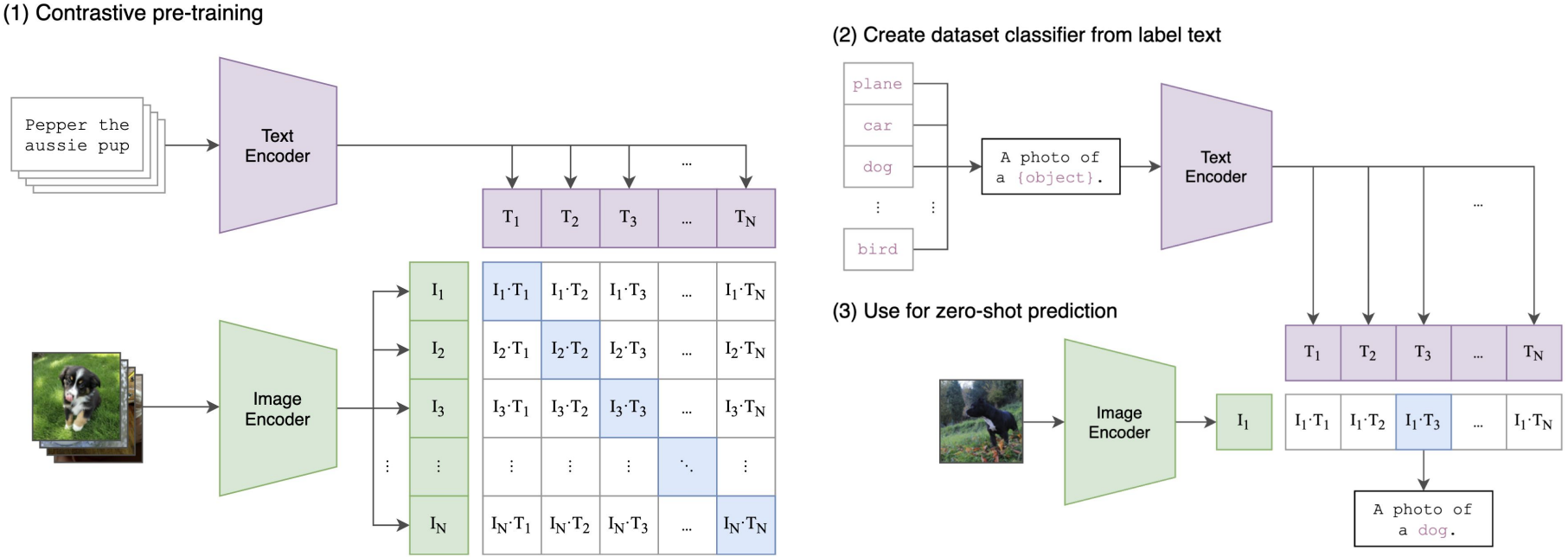


Text data (GPT)

He et al., “Masked Autoencoders Are Scalable Vision Learners”, CVPR 2022.

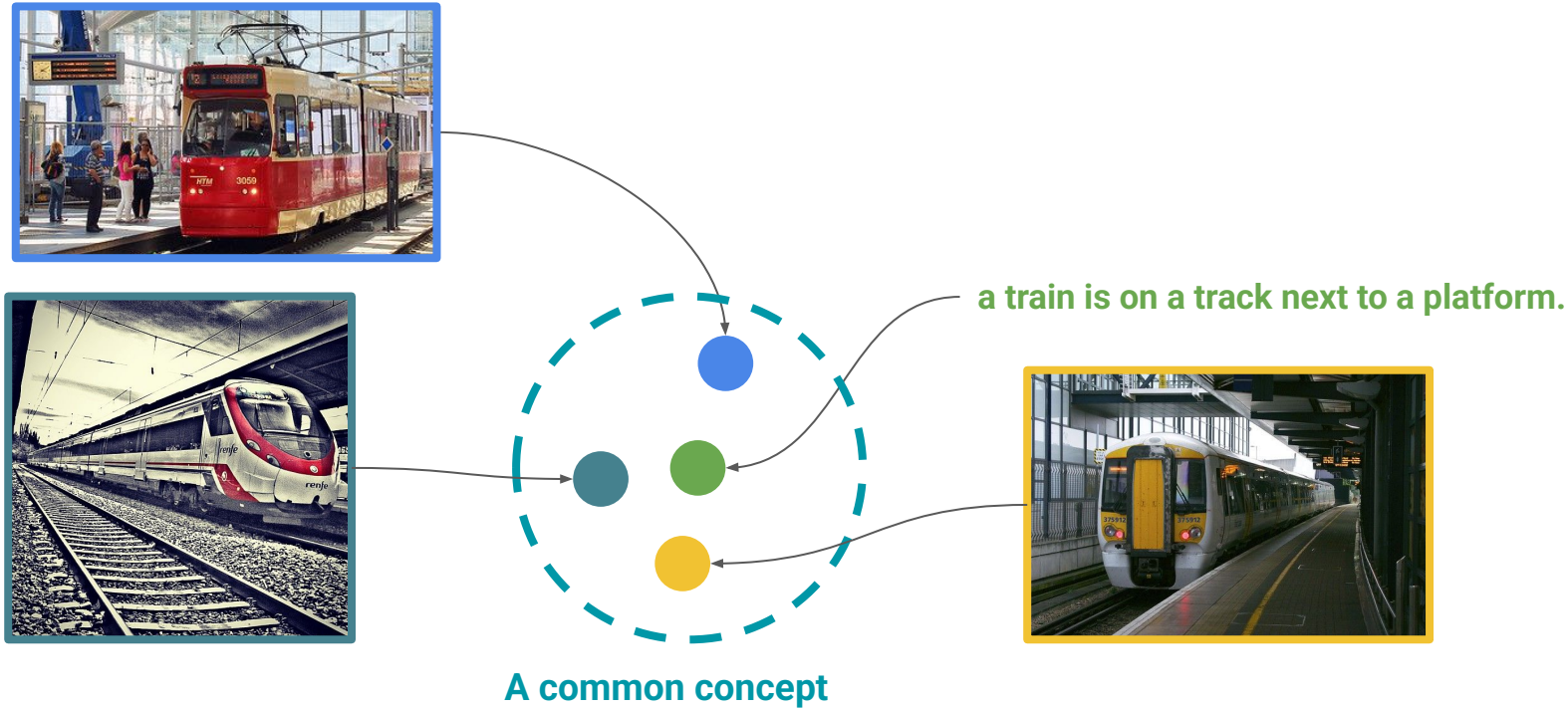
Contrastive Language-Image Pre-training (CLIP)

- CLIP and zero-shot prediction

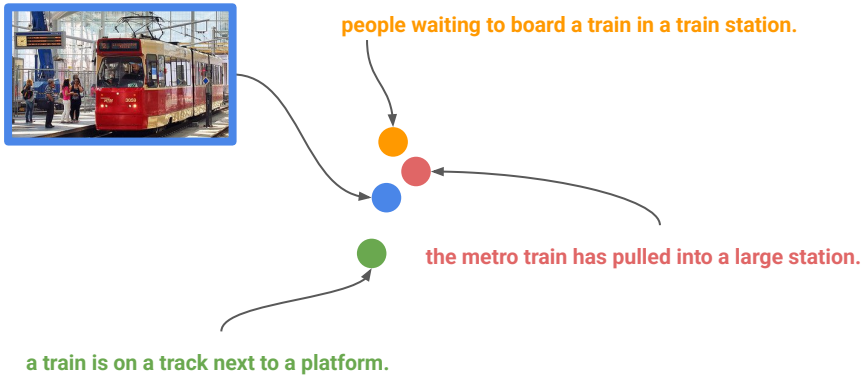


Radford et al., “Learning Transferable Visual Models From Natural Language Supervision”, ICML 2021.

Multiplicity (many-to-many) problem



Chun et al., “Probabilistic Embeddings for Cross-Modal Retrieval”, CVPR 2021.



“Many-to-many mapping” challenges in language-image training:

- An image potentially can be matched with a number of different captions.

“Many-to-many mapping” challenges in language-image training:

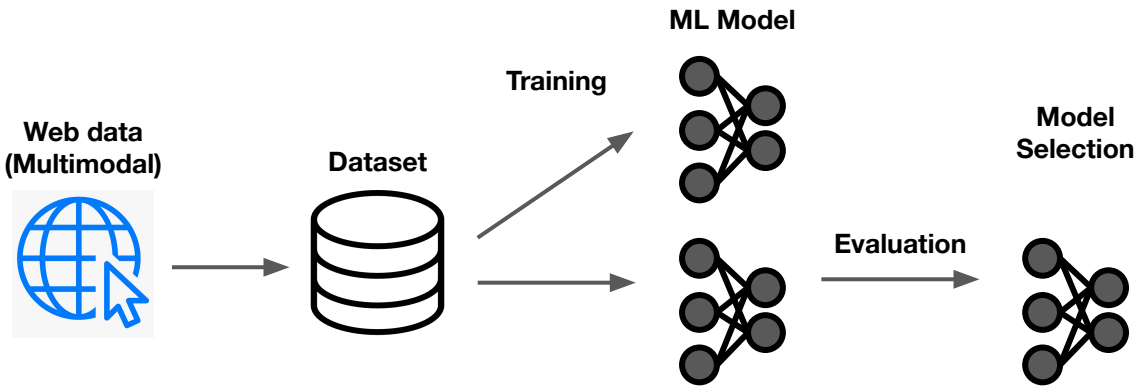
- An image potentially can be matched with a number of different captions.
- While an image is taken by thoroughly captured in a photograph, language descriptions are the product of conscious choices of the key relevant concepts to report from a given scene.

“Many-to-many mapping” challenges in language-image training:

- An image potentially can be matched with a number of different captions.
- While an image is taken by thoroughly captured in a photograph, language descriptions are the product of conscious choices of the key relevant concepts to report from a given scene.

Multiplicity: A Core Challenge in Multimodal Learning

- Multiplicity makes multimodal learning challenging in all the core components (Dataset construction, Contrastive learning, Evaluation)

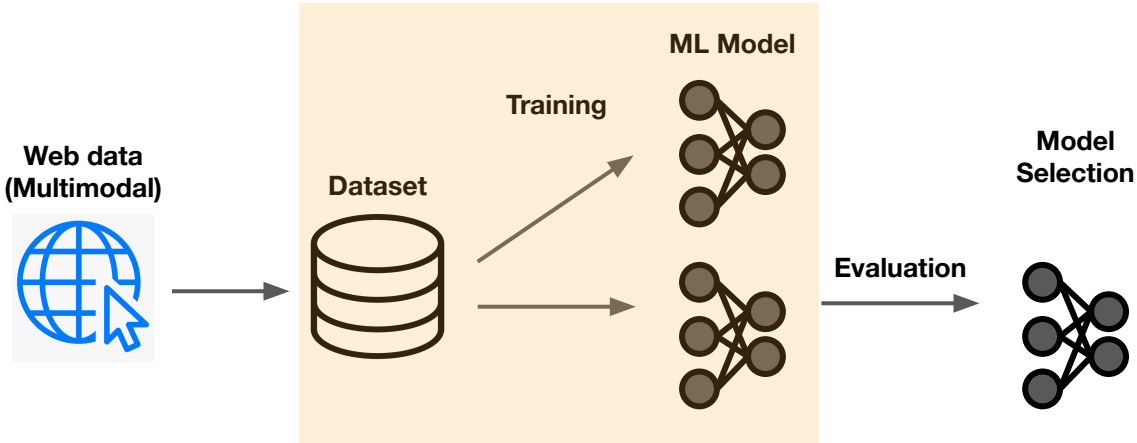


Contents

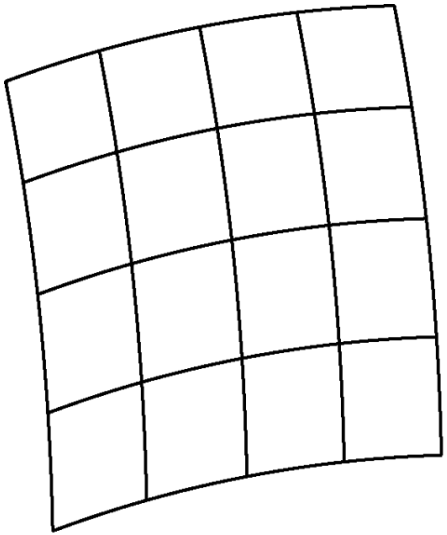
- Introduction: **The multiplicity problem** in multimodal learning
- How can we handle **the multiplicity in contrastive learning**?
- How can we handle **the multiplicity in evaluation**?
- How can we handle **the multiplicity in dataset construction**?
- Conclusion

Multiplicity: A Core Challenge in Multimodal Learning

- Multiplicity makes multimodal learning challenging in all the core components (Dataset construction, **Contrastive learning**, Evaluation)
- **Contrastive learning** might become difficult when there exist multiple possible other “ground-truths”.
- How can we handle this?



How deterministic space oversimplifies multiplicity?



(a) Deterministic embedding

"Person"

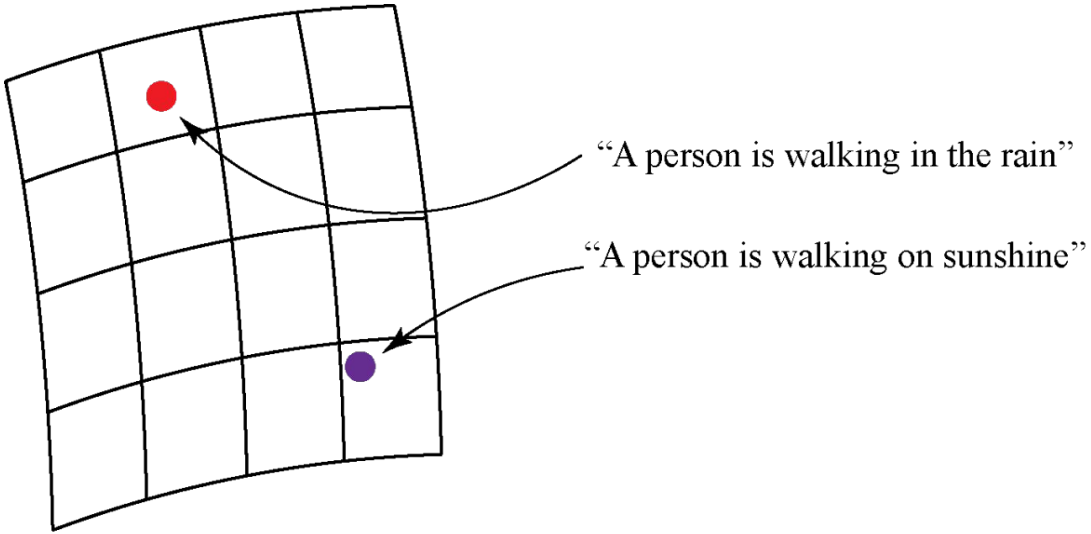
"A person is walking in the rain"

"A person is walking on sunshine"

"A person is walking"

(b) Probabilistic embedding

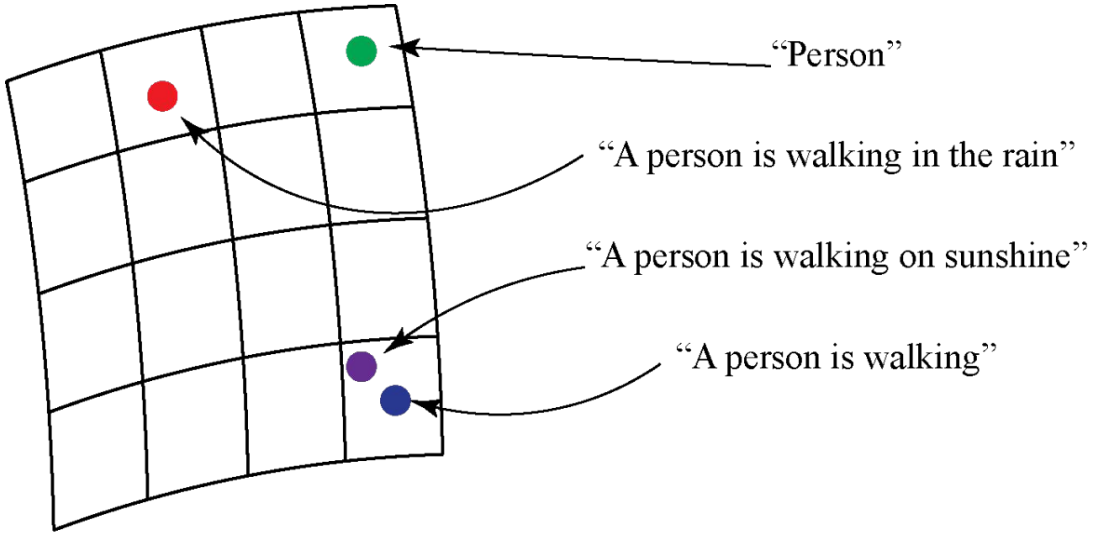
Det. space maps an input to a specific vector coord.



(a) Deterministic embedding

(b) Probabilistic embedding

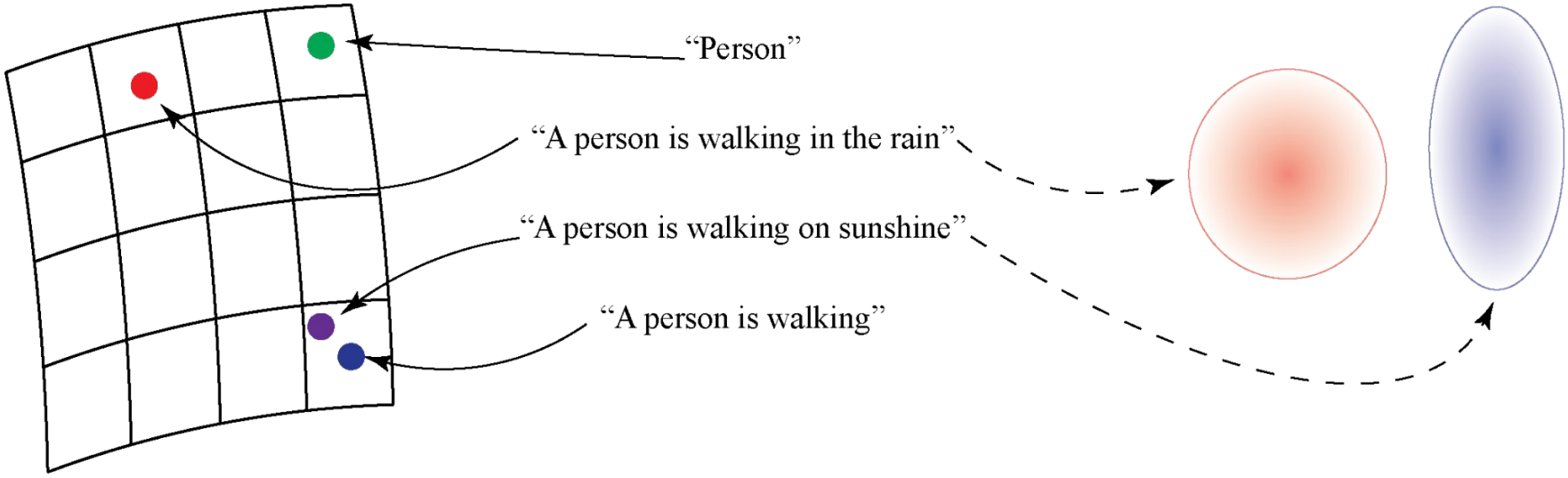
Even if an input can be matched to multiple instances, it will be mapped to a specific coord.



(a) Deterministic embedding

(b) Probabilistic embedding

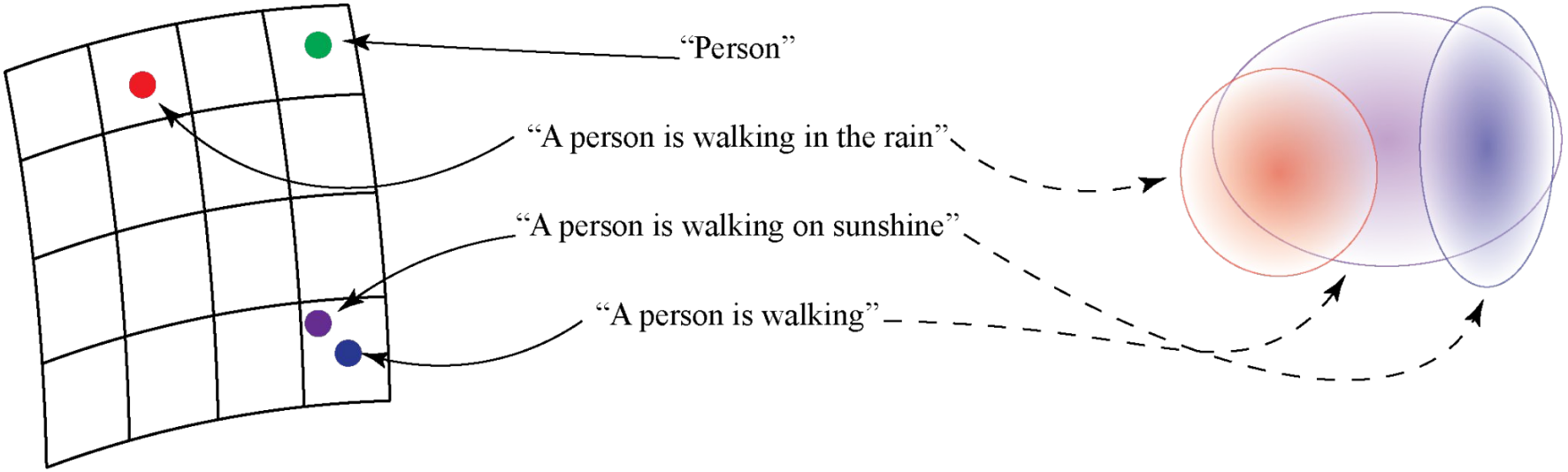
Prob. space has additional info. axis, “uncertainty”



(a) Deterministic embedding

(b) Probabilistic embedding

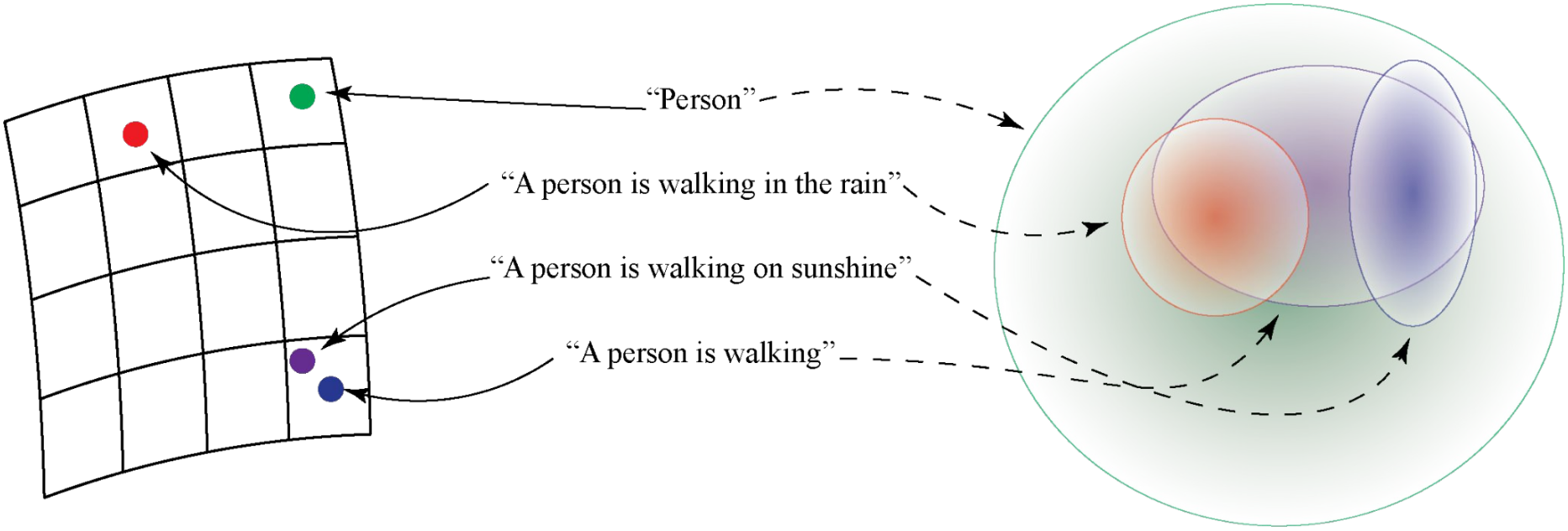
An uncertain input will have a large uncertainty value



(a) Deterministic embedding

(b) Probabilistic embedding

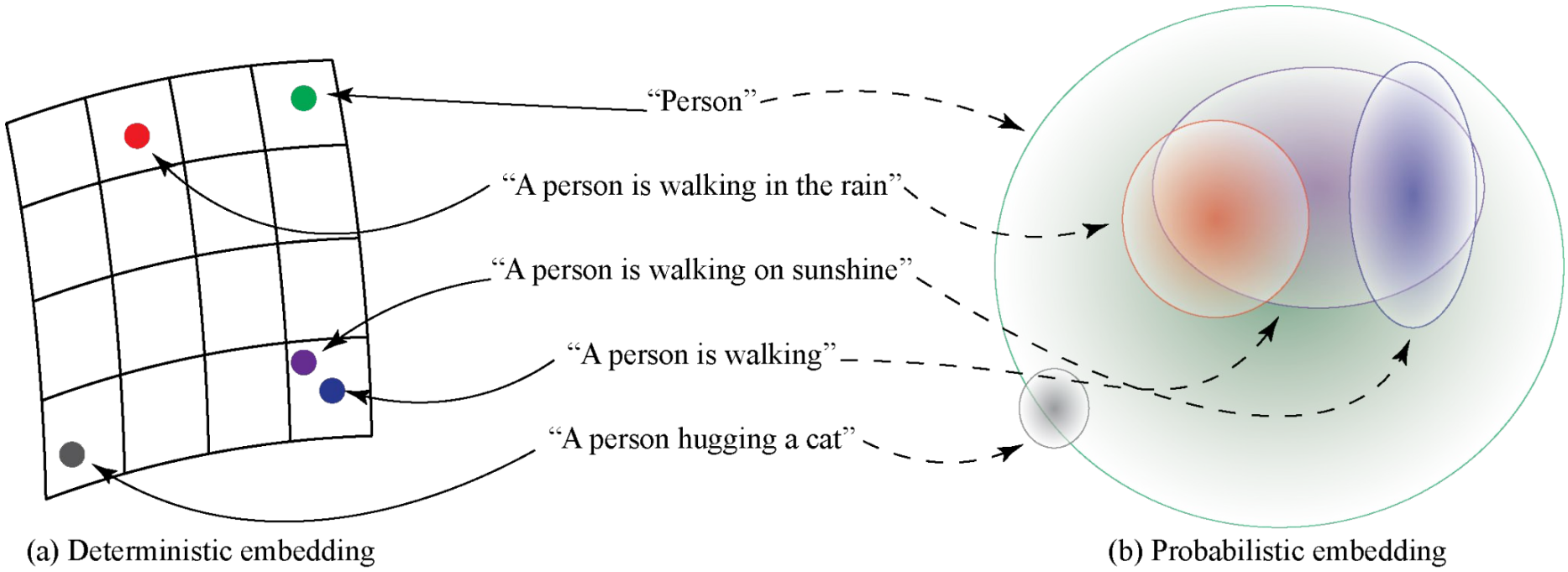
We can easily capture input uncertainty by “variance”



(a) Deterministic embedding

(b) Probabilistic embedding

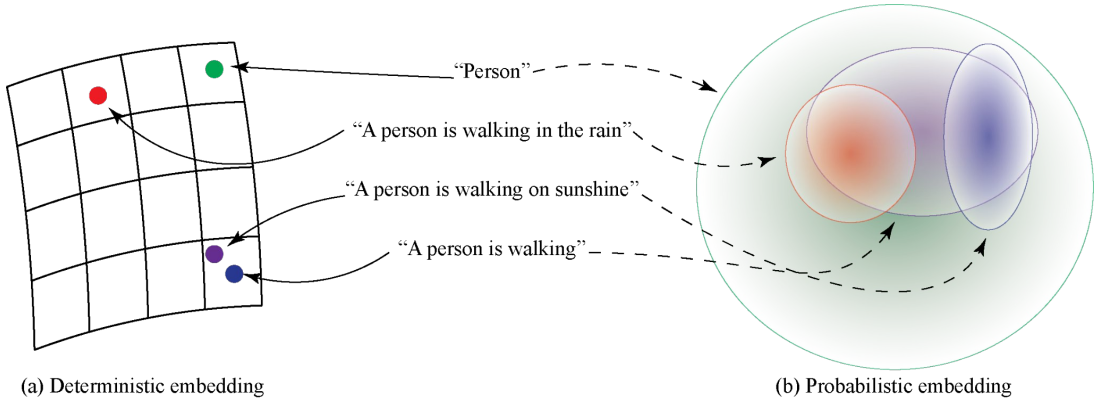
Real-world scenarios require handling uncertainty



Chun et al., "Probabilistic Language-Image Pre-Training", ICLR 2025.

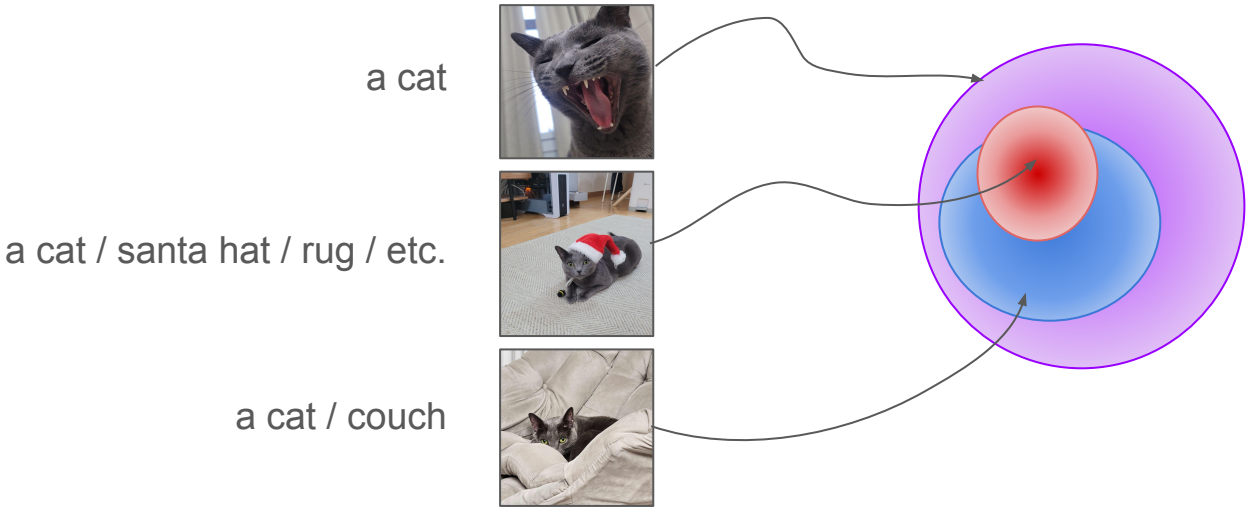
What is “Uncertainty” in CLIP emb. space?

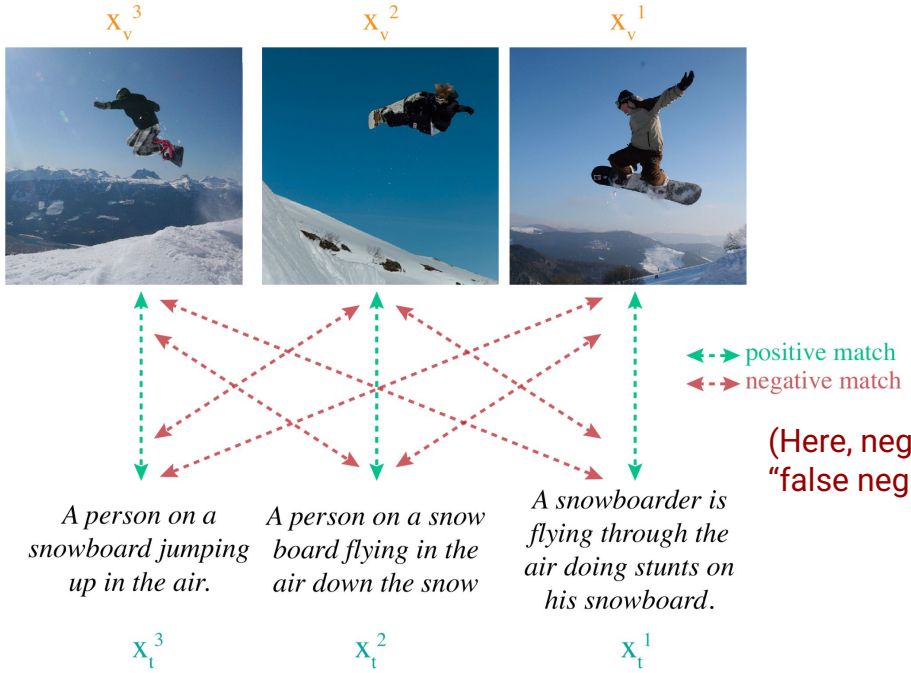
- Text embeddings
  - A general caption → can matched with many images → high uncertainty (e.g., “A photo”)
  - A detailed caption → can matched with few images → low uncertainty (e.g., “A photo of a cat sitting in a chair”)



What is “Uncertainty” in CLIP emb. space?

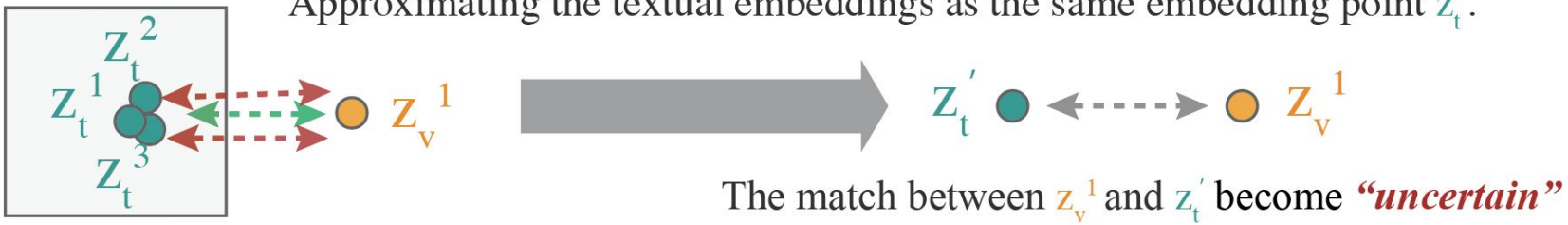
- Image embeddings
  - A simple/general image → matched to many texts → high uncertainty
  - A complex/detailed image → matched with few texts → low uncertainty





(Here, negative matches denote "false negatives")

Approximating the textual embeddings as the same embedding point z_t'.



The match between z_v^1 and z_t' become ***"uncertain"***

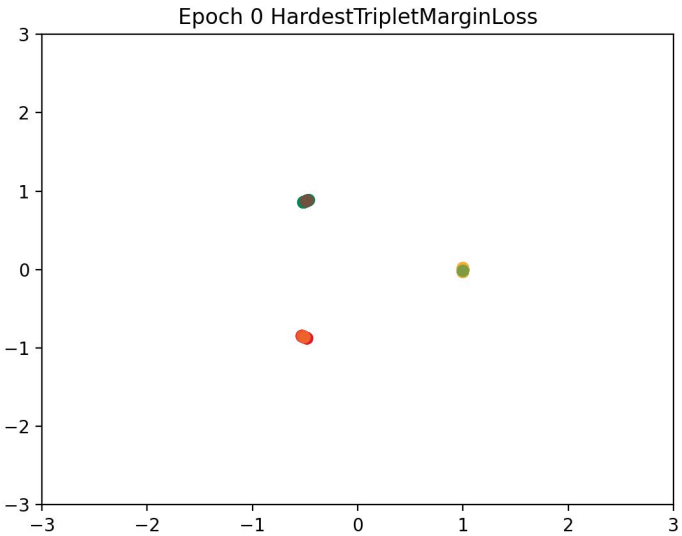
There are a lot of false negatives in popular VL datasets

Dataset	# positive images	# positive captions
Original MS-COCO Caption	1,332	6,305 (=1,261×5)
CxC [32]	1,895 (×1.42)	8,906 (×1.41)
Human-verified positives	10,814 (×8.12)	16,990 (×2.69)
ECCV Caption	11,279 (×8.47)	22,550 (×3.58)

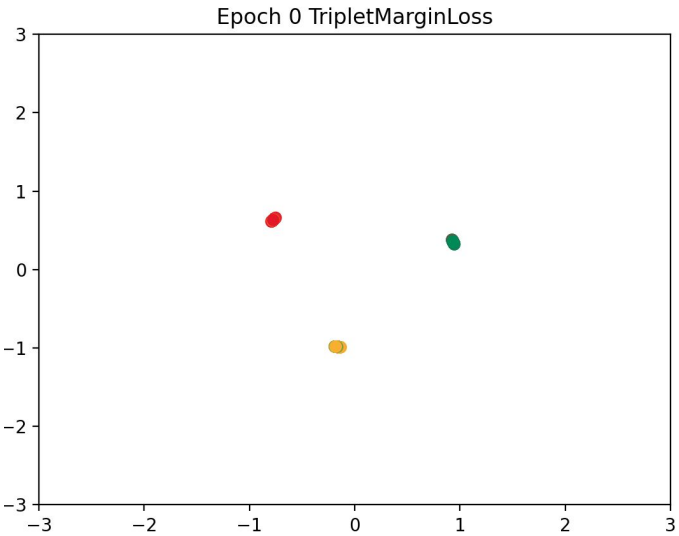
Chun et al., “ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO”, ECCV 2022.

What will be happened if we train embeddings with deterministic embeddings and negative mining?

With Hardest Negative Mining (HNM)



Without HNM



Red, **Green**, **Yellow**: Certain classes

Orange: Either **Red** or **Yellow**

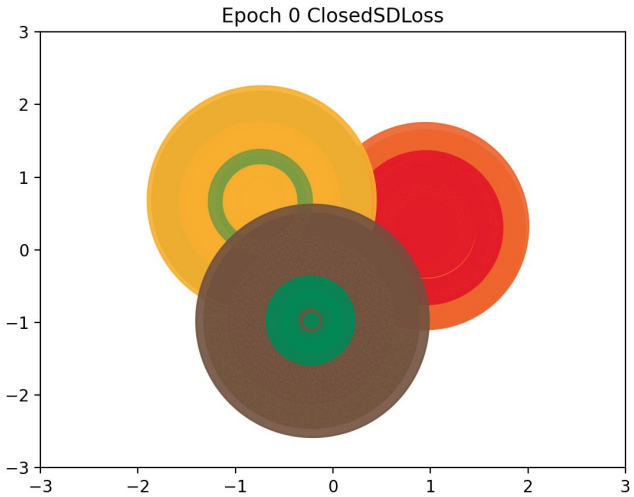
Brown: Either **Red** or **Green**

Yellowish Green: Either **Green** or **Yellow**

2-D toy dataset (animation also can be found in https://naver-ai.github.io/pcmepp/)

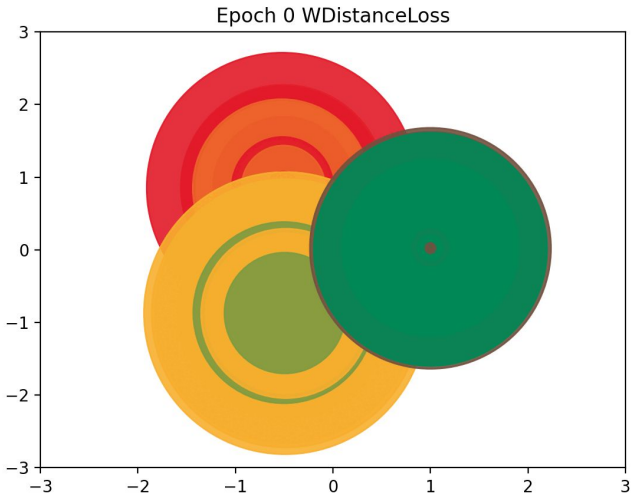
Toy experiments with probabilistic embeddings

CSD (proposed)



Small variances for certain classes,
Large variances for uncertain classes

Wasserstein distance



Both certain / uncertain classes have
the similar variances

Red, **Green**, **Yellow**:
Certain classes

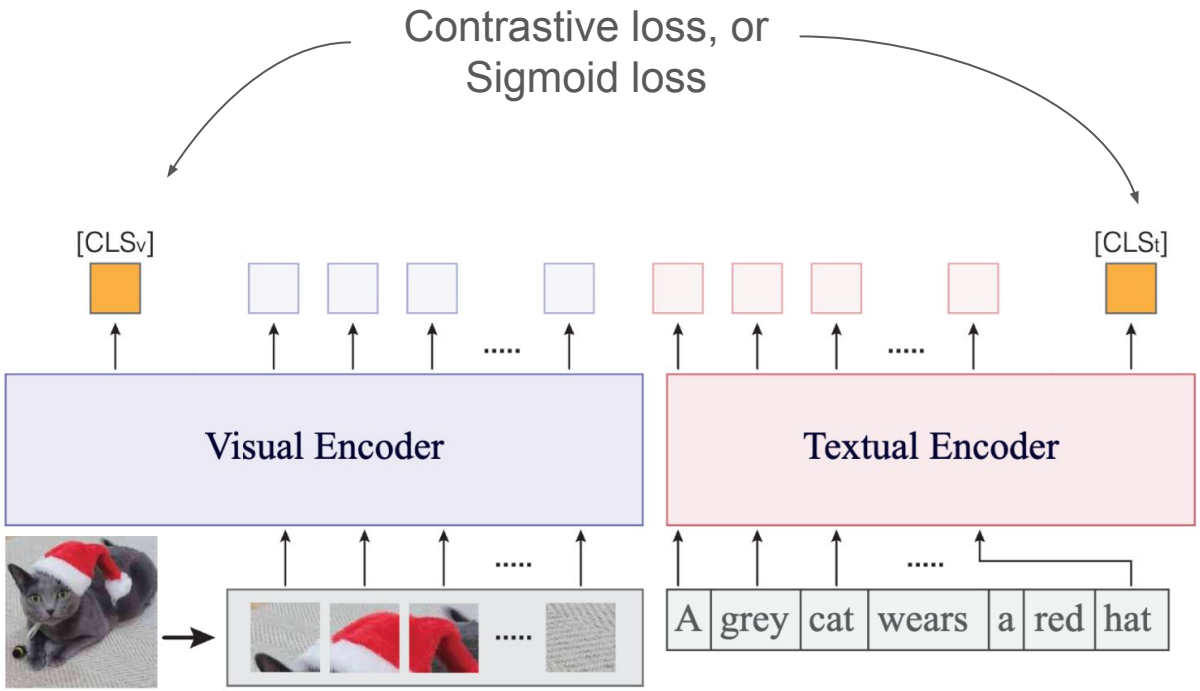
Orange: Either
Red or **Yellow**

Brown: Either
Red or **Green**

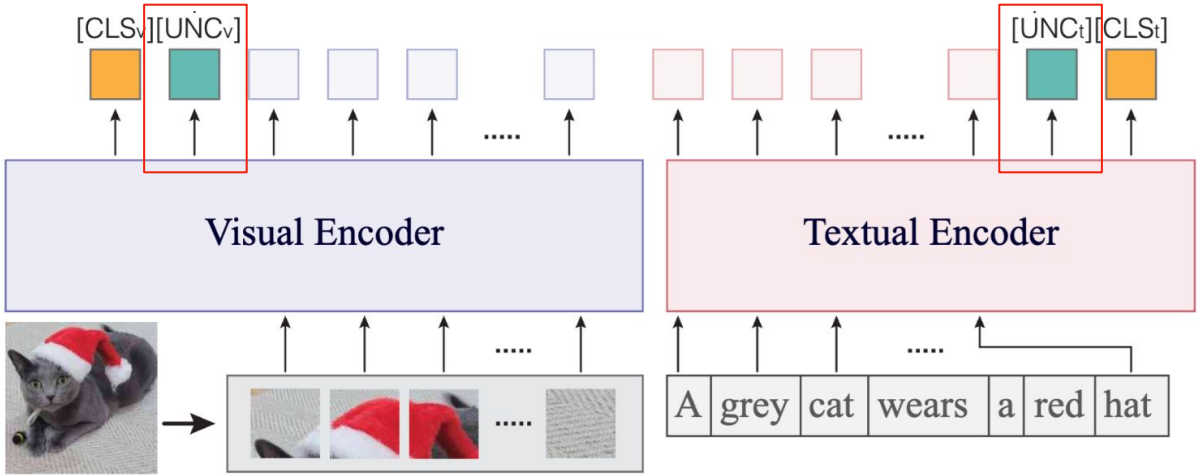
Yellowish Green:
Either **Green** or **Yellow**

2-D toy dataset (animation also can be found in https://naver-ai.github.io/pcmepp/)

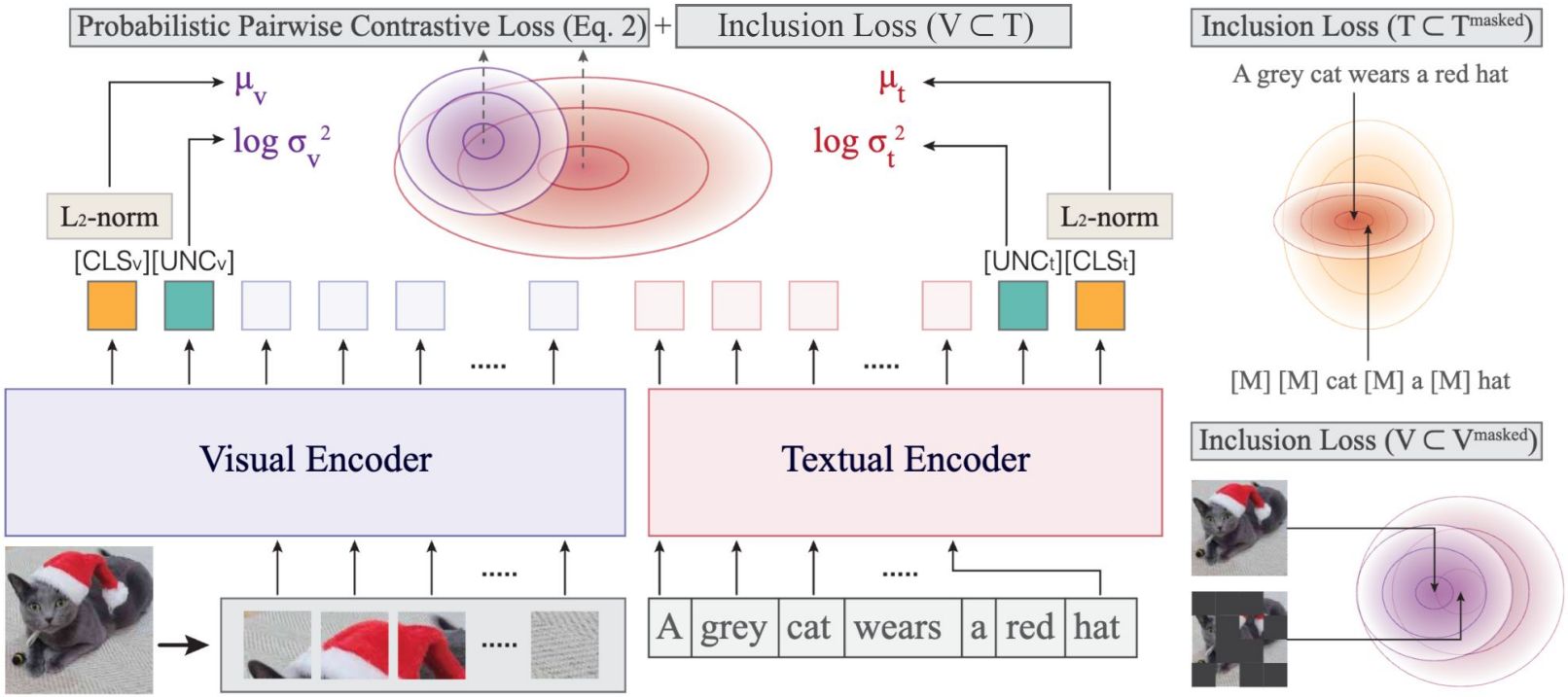
Standard Language-Image Pre-training



Efficient uncertainty architecture by [UNC] token



Efficient uncertainty architecture by [UNC] token

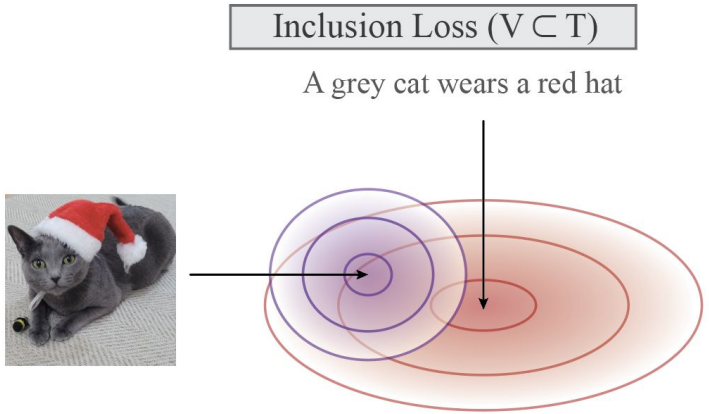


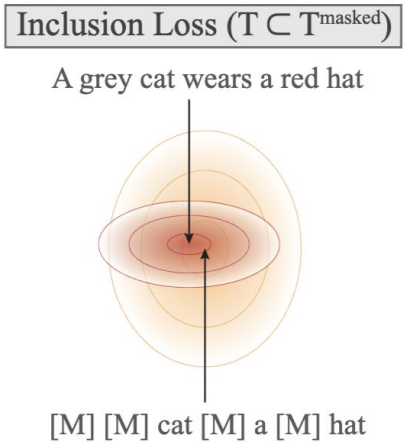
Probabilistic Pairwise Contrastive Loss (PPCL)

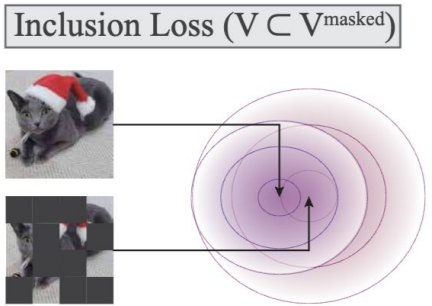
- How can we define “distance” between two random variables?
- Our desired properties:
  - If the match of Z_v and Z_t is certain, then Z_v and Z_t has small variance
  - If the match of Z_v and Z_t is uncertain, then Z_v and Z_t has large variance
- Closed-form sampled distance (CSD)
  - $d(\mathbf{Z}_v, \mathbf{Z}_t) = \mathbb{E}_{\mathbf{Z}_v, \mathbf{Z}_t} \|\mathbf{Z}_v - \mathbf{Z}_t\|_2^2 = \|\mu_v - \mu_t\|_2^2 + \|\sigma_v^2 + \sigma_t^2\|_1$
- PPCL is a pairwise contrastive loss (or sigmoid loss) using CSD as distance between instances

Inclusion loss for enforcing inclusive relationship

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







Certain & uncertain samples

$\sigma_v = 0.0054 \sim 0.0072$
$P(\sigma_v) < 0.1\%$

$\sigma_v = 0.0123 \sim 0.0126$
$19\% < P(\sigma_v) < 21\%$

$\sigma_v = 0.0154 \sim 0.0158$
$39\% < P(\sigma_v) < 41\%$

$\sigma_v = 0.0194 \sim 0.0199$
$59\% < P(\sigma_v) < 61\%$

$\sigma_v = 0.0253 \sim 0.0260$
$79\% < P(\sigma_v) < 81\%$

$\sigma_v = 0.0464 \sim 0.0613$
$99.9\% < P(\sigma_v)$

← More certain samples

More uncertain samples→

- 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
 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

$\sigma_t = 0.0460 \sim 0.0468$
$79\% < P(\sigma_t) < 81\%$

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

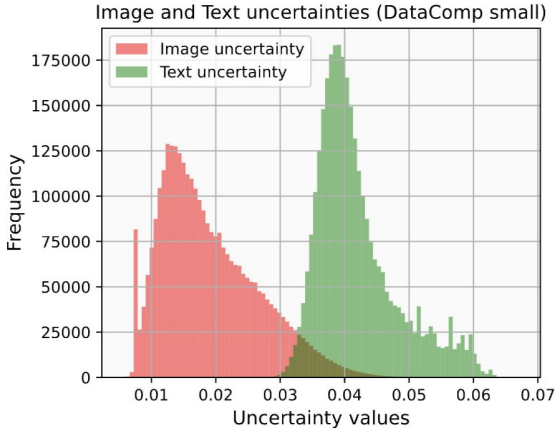
$\sigma_t = 0.0625 \sim 0.0673$
$99.9\% < P(\sigma_t)$

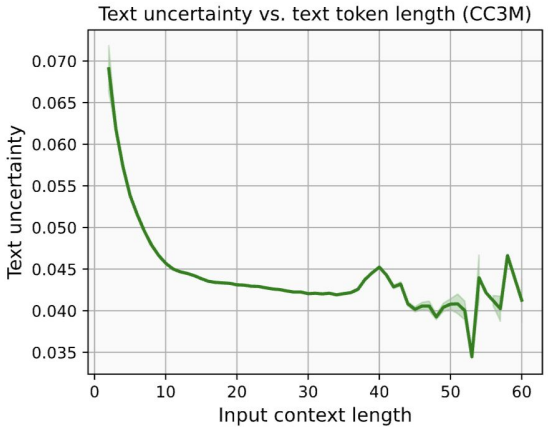
Zero-shot classification

		# 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

- Pre-trained on the DataComp-1B dataset; ProLIP is the first large-scale pre-trained probabilistic vision-language model
- All models are available at https://huggingface.co/collections/SanghyukChun/prolip-6712595dfc87fd8597350291

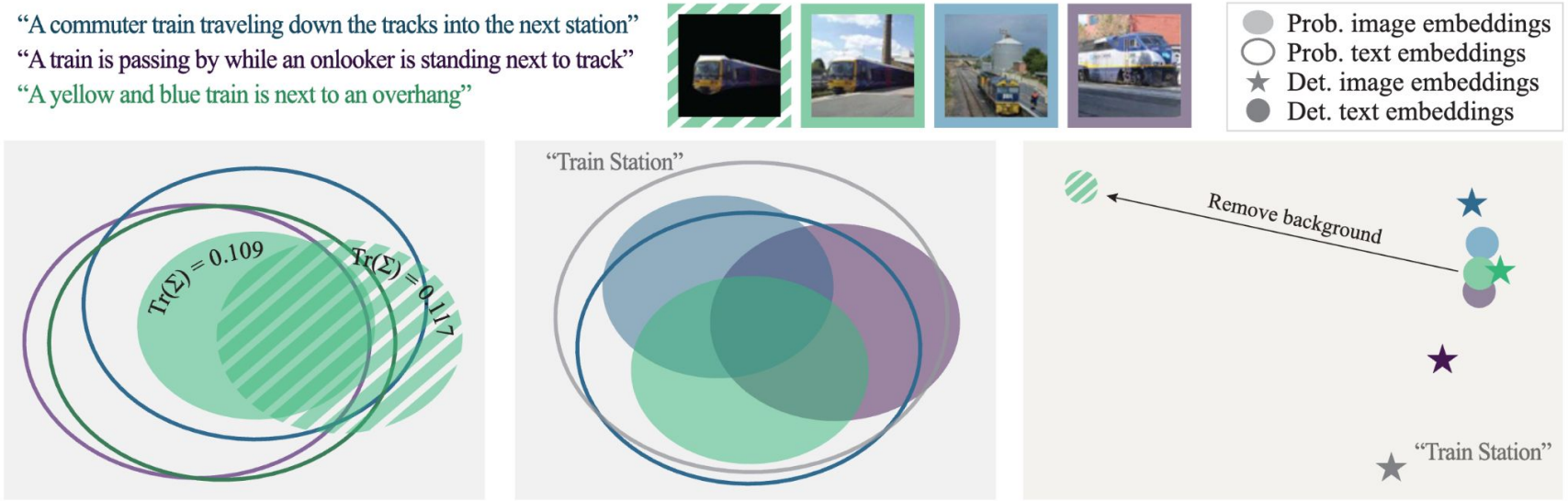
Property of learned uncertainty





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

Visualization of learned uncertainty



(a) ProLIP embedding space

(b) Deterministic embedding space

References

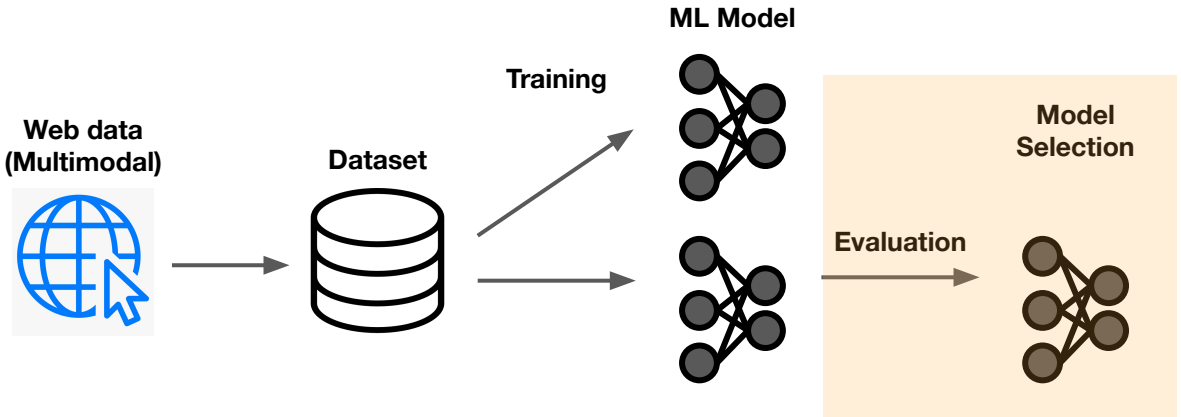
- **[CVPR 2021] Probabilistic Embeddings for Cross-Modal Retrieval.**
  Sanghyuk Chun, Seong Joon Oh, Rafael Sampaio de Rezende, Yannis Kalantidis, Diane Larlus
- **[ECCV 2022] ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO.**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

Contents

- Introduction: **The multiplicity problem** in multimodal learning
- How can we handle **the multiplicity in contrastive learning**?
- How can we handle **the multiplicity in evaluation**?
- How can we handle **the multiplicity in dataset construction**?
- Conclusion

Multiplicity: A Core Challenge in Multimodal Learning

- Multiplicity makes multimodal learning challenging in all the core components (Dataset construction, Contrastive learning, **Evaluation**)
- **Evaluation** might become difficult because we have potential plausible matches in the dataset, which is marked as “negative” (i.e., False Negatives)
- How can we handle this?



MS-COCO Caption dataset for Image-Text Matching (ITM)

- MS-COCO Caption collects captions by human annotators
- During training or evaluating models on COCO Caption, the collected image-text pairs are treated as the only positive.

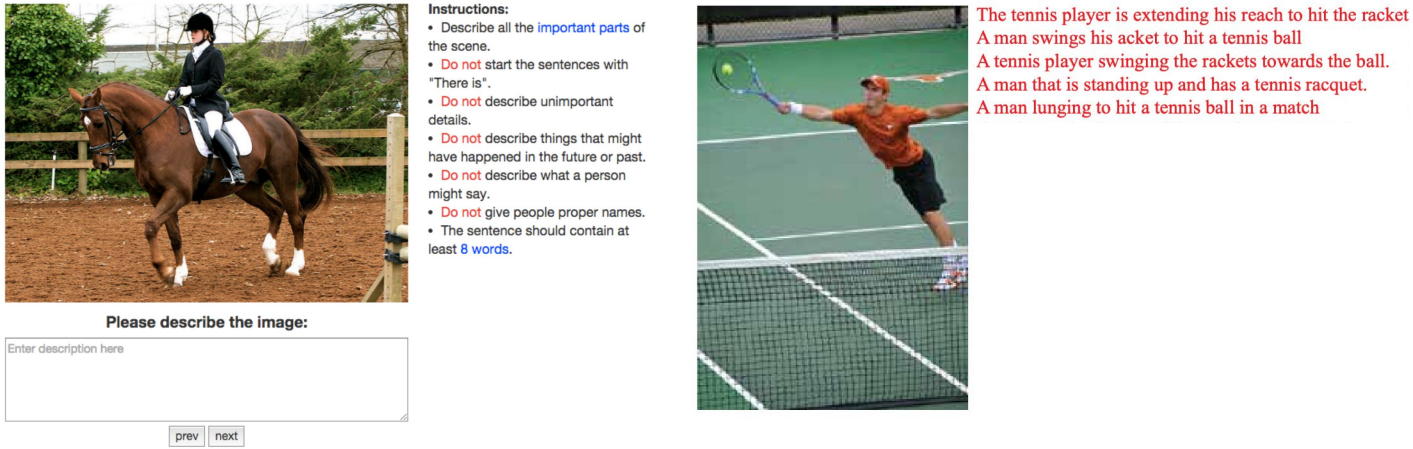


Fig. 2: Example user interface for the caption gathering task.

More than one captions can describe to describe one image

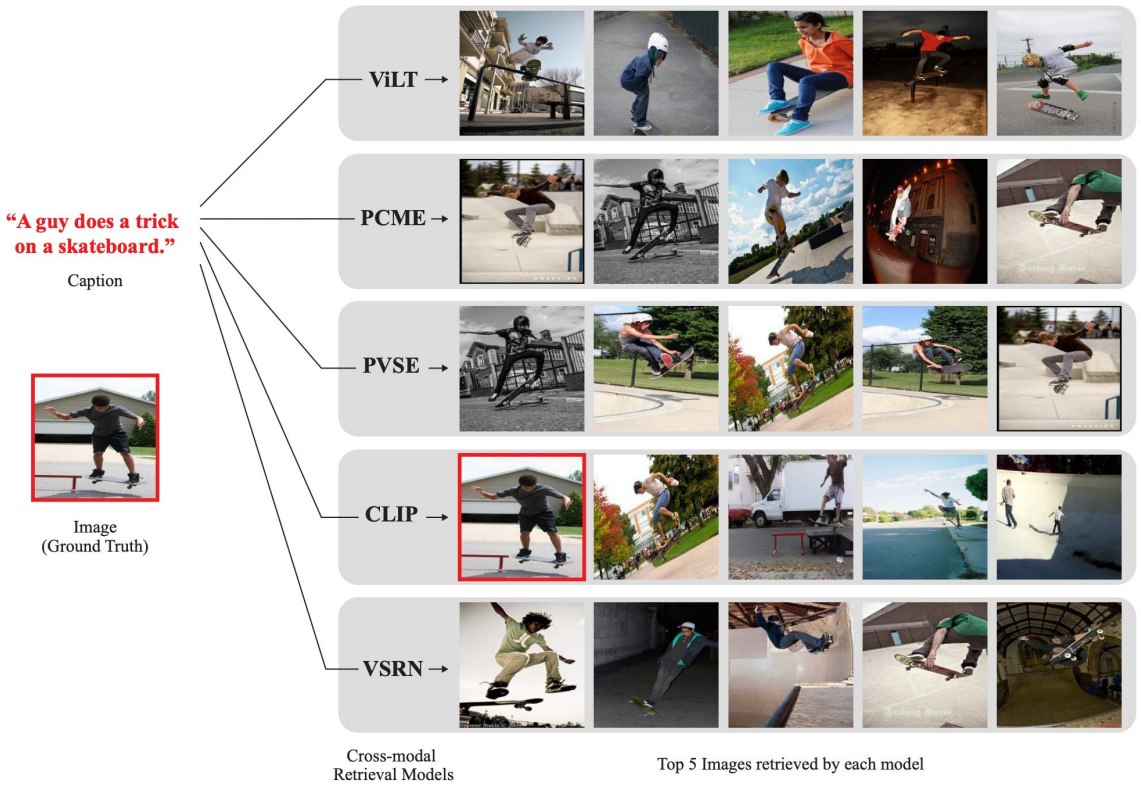
The tennis player is extending his reach to hit the racket
A man swings his acket to hit a tennis ball
A tennis player swinging the rackets towards the ball.
A man that is standing up and has a tennis racquet.
A man lunging to hit a tennis ball in a match

More than one captions can describe to describe one image

The tennis player is extending his reach to hit the racket.
A man swings his acket to hit a tennis ball
A tennis player swinging the rackets towards the ball.
A man that is standing up and has a tennis racquet.
A man lunging to hit a tennis ball in a match
A male tennis player walking on the tennis court.
A man on a court swinging a racket at a ball.
The man is playing tennis with a racket.
A man standing on a tennis court holding a racquet.
A man on a tennis court trying to hit the ball
A man taking a swing at a tennis ball
A man taking a swing at a tennis ball
A man throwing a tennis ball in the air for him to hit it with his racket.
A man hitting a tennis ball with a racquet.
A man with a tennis racket is running on a court
A man swinging his racket to hit the ball.
The tennis player is hitting the ball with his racket
A tennis player caught jumping up to hit the ball
A man is holding a tennis racquet prepared to hit the incoming ball.
A man holding a tennis racquet as a ball clears the net.
A man with a racket prepares to hit a tennis ball
A man in shorts and a long sleeve shirt playing tennis.

A man stands on a tennis court hitting a ball with a racket.
A man plays a game of tennis during the day.
A man with a tennis racket swings at a tennis ball
A man with a tennis racket on a court.
A man playing tennis on the tennis court
A person hitting a tennis ball with a tennis racket
A man playing tennis and holding back his racket to hit the ball.
A male tennis player swinging his tennis racket.
A man swinging at a tennis ball with a tennis racket.
A person hitting a tennis ball with a tennis racket.
A man on a tennis court that has a racquet.
A man in a head band hits a tennis ball
A man standing on top of a tennis court holding a racquet.
A male in a blue shirt playing tennis on a tennis court
A man holding a tennis racket on a tennis court.
A tennis player swinging a racket at a ball
A man holding a racquet on top of a tennis court.
A boy hitting a tennis ball on the tennis court.
A man on a court swinging a tennis racket.
A man in white shirt and shorts playing tennis.
A guy in a maroon shirt is holding a tennis racket out to hit a tennis ball.
A person hitting a tennis ball with a tennis racket on a tennis court.
The man is playing tennis on the court.

False Negatives (FNs) can lead to wrong evaluation results



The only "correct" model by the original "positive"

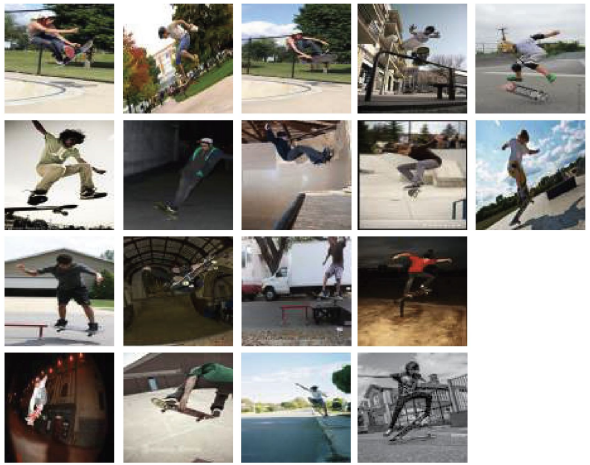
Naive solution: Validating all possible image-text pairs whether each of them is positive or not

AS-IS

Caption: “A guy does a trick on a skateboard.”
Positive images:

TO-BE

Caption: “A guy does a trick on a skateboard.”
Positive images:



Validating all possible pairs is extremely expensive (5,000 * 25,000 = 125M pairs)
=> We need a model-based filtering

ECCV Caption: Extended COCO Caption Validation dataset

"A guy does a trick on a skateboard."

Caption

Image
(Ground Truth)

ViLT →

PCME →

PVSE →

CLIP →

VSRN →

ECCV Caption: Extended COCO Caption Validation dataset

"A guy does a trick on a skateboard."

Caption

Image
(Ground Truth)

ViLT →

PCME →

PVSE →

CLIP →

VSRN →

ECCV Caption: Extended COCO Caption Validation dataset

"A guy does a trick on a skateboard."

Caption

Image
(Ground Truth)

ViLT →

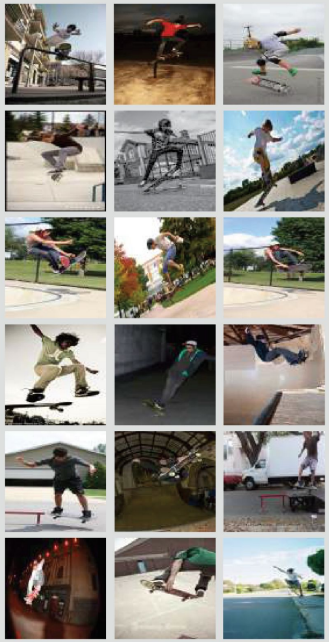
PCME →

PVSE →

CLIP →

VSRN →

New positives validated by crowdworkers



ECCV Caption: Extended COCO Caption Validation dataset

Dataset	# positive images	# positive captions
Original MS-COCO Caption	1,332	6,305 (=1,261×5)
CxC	1,895 (×1.42)	8,906 (×1.41)
Human-verified positives	10,814 (×8.12)	16,990 (×2.69)
ECCV Caption	11,279 **(×8.47)**	22,550 **(×3.58)**

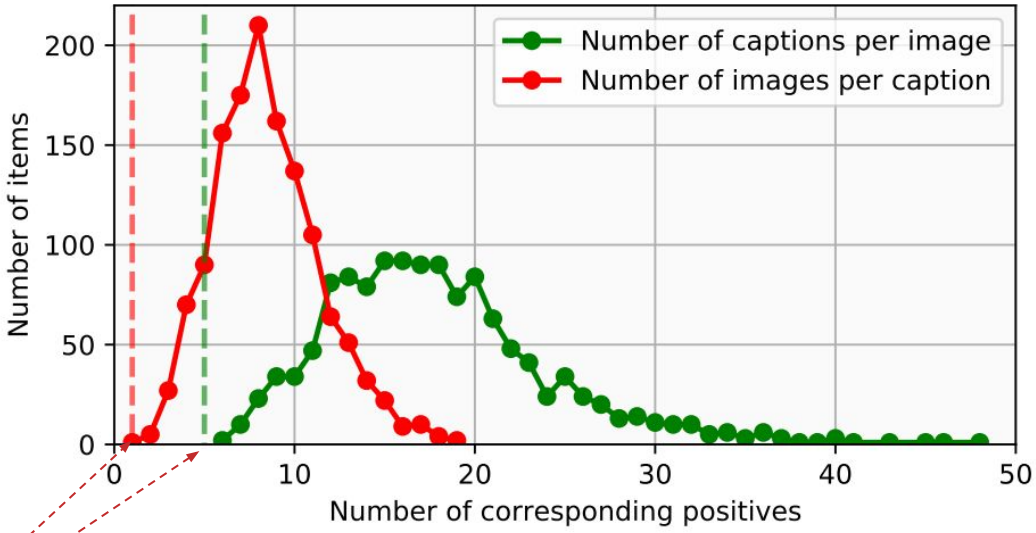
We build ECCV Caption dataset by merging our human annotations and CxC.

ECCV Caption: Extended COCO Caption Validation dataset

Dataset	I2T Precision	I2T Recall	T2I Precision	T2I Recall
Original MS-COCO Caption	96.9	21.1	96.4	13.8
CxC	95.5	23.0	93.2	16.1
Plausible Match ($\zeta = 0$)	65.3	56.5	56.6	61.8

COCO / CxC shows high precision (> 93) but low recall (< 25, i.e., many false negatives).
PM shows high recall (> 55) but low precision (< 65, i.e., many false positives).

ECCV Caption: Extended COCO Caption Validation dataset



(a) Number of positive pairs.

the original COCO positives

Metric also matters; R@k is not enough

GT images for “A train on a train track near many trees”.

(A) Only top-1 is wrong.

R@1	R@5	mAP@R	Preference score
0.0	100.0	68.6	70.85

(A) Only top-1 is correct.

R@1	R@5	mAP@R	Preference score
100.0	100.0	11.1	10.66

(C) Top-1 to -5 are wrong.

R@1	R@5	mAP@R	Preference score
0.0	0.0	14.1	13.15

(D) Only top-5 is correct.

R@1	R@5	mAP@R	Preference score
0.0	100.0	2.2	4.89

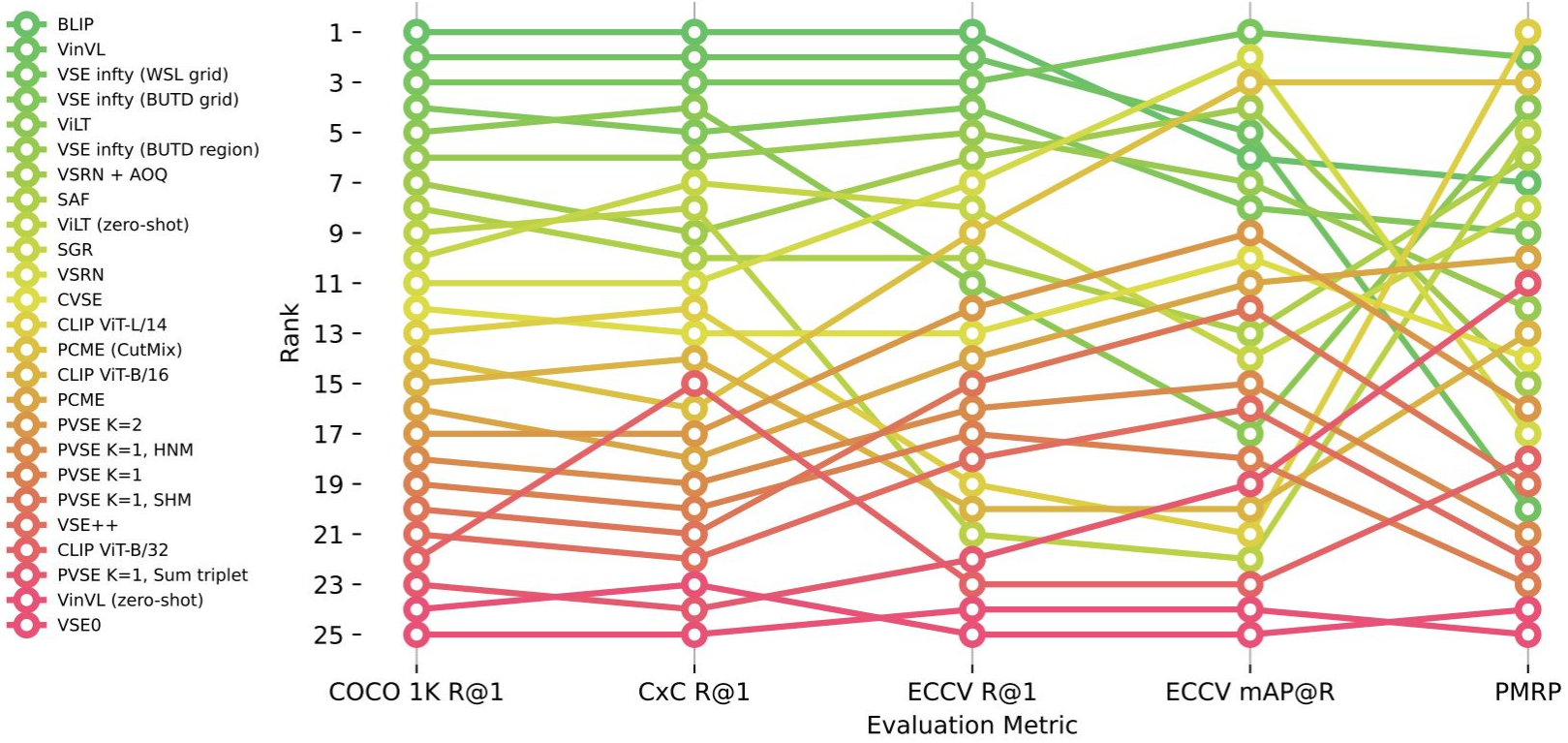
Re-evaluating Vision-Language (VL) models with …

- ECCV Caption mAP@R, R-Precision and Recall@1
- CxC Recall@1
- MS-COCO Recall@1, Recall@5
  - The full 5K test images
  - 1K cross-validation images (splitting 5K images to 5 splits and average their results)
- Easy to use evaluation code https://github.com/naver-ai/eccv-caption
  - `pip3 install eccv_caption`

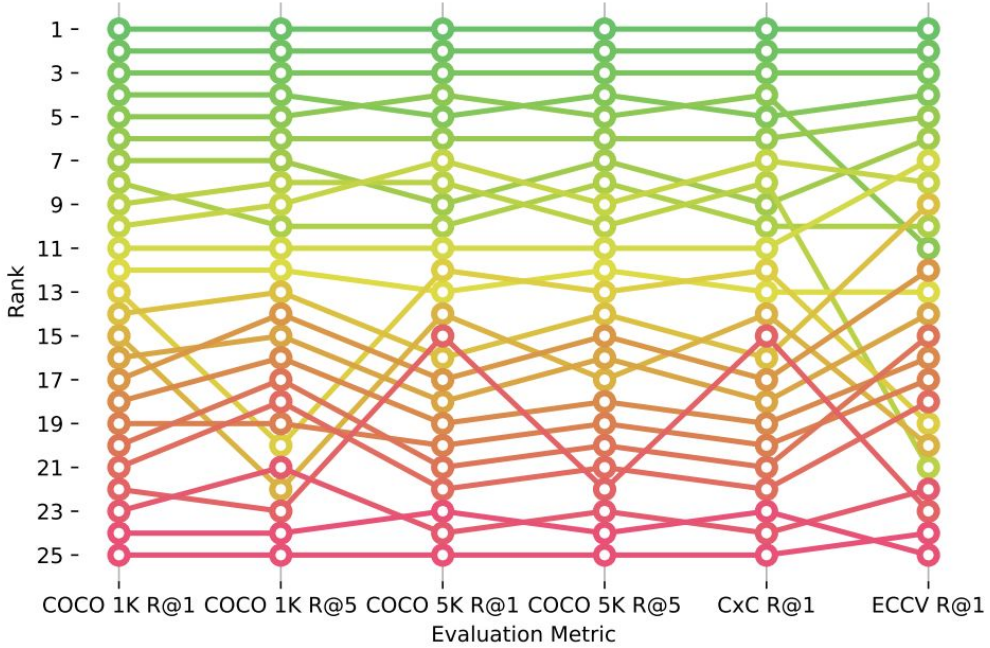
```python
from eccv_caption import Metrics

metric = Metrics()

# Get i2t, t2i retrived items from your own model
# i2t = {query_image_id: [sorted_caption_id_by_similarity]}
# t2i = {query_caption_id: [sorted_image_id_by_similarity]}
# See the example code for how to compute i2t / t2i from your model

scores = metric.compute_all_metrics(
        i2t, t2i,
        target_metrics=('eccv_r1', 'eccv_map_at_r', 'eccv_rprecision',
                        'coco_1k_recalls', 'coco_5k_recalls', 'cxc_recalls'),
        Ks=(1, 5, 10),
        verbose=False
)
print(scores)
```

```
{
    'coco_1k_r1': {'i2t': 0.7425999999999999, 't2i': 0.5544},
    'coco_1k_r5': {'i2t': 0.9282, 't2i': 0.82324},
    'coco_1k_r10': {'i2t': 0.9658, 't2i': 0.90152},
    'coco_5k_r1': {'i2t': 0.5636, 't2i': 0.3658},
    'coco_5k_r5': {'i2t': 0.7928, 't2i': 0.61048},
    'coco_5k_r10': {'i2t': 0.867, 't2i': 0.71148},
    'cxc_r1': {'i2t': 0.5804, 't2i': 0.38314912702226495},
    'cxc_r5': {'i2t': 0.8142, 't2i': 0.6385151369533878},
    'cxc_r10': {'i2t': 0.8862, 't2i': 0.7425116130065673},
    'eccv_map_at_r': {'i2t': 0.23988490345859761, 't2i': 0.3195708452398554},
    'eccv_rprecision': {'i2t': 0.3381634023972558, 't2i': 0.4177097641300428},
    'eccv_r1': {'i2t': 0.7129262490087233, 't2i': 0.7297297297297297}
}
```

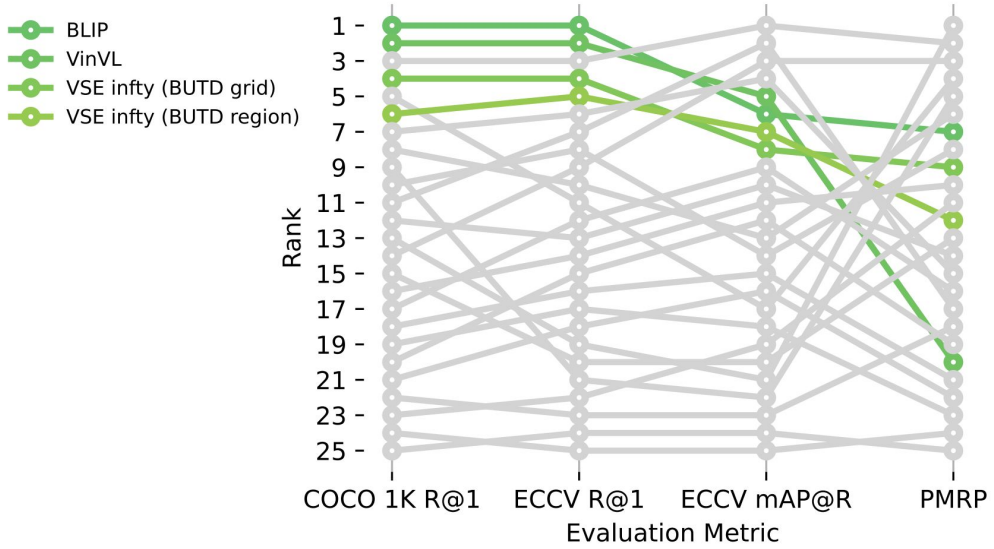
Message 1: ECCV mAP@R changes COCO R1 ranks a lot



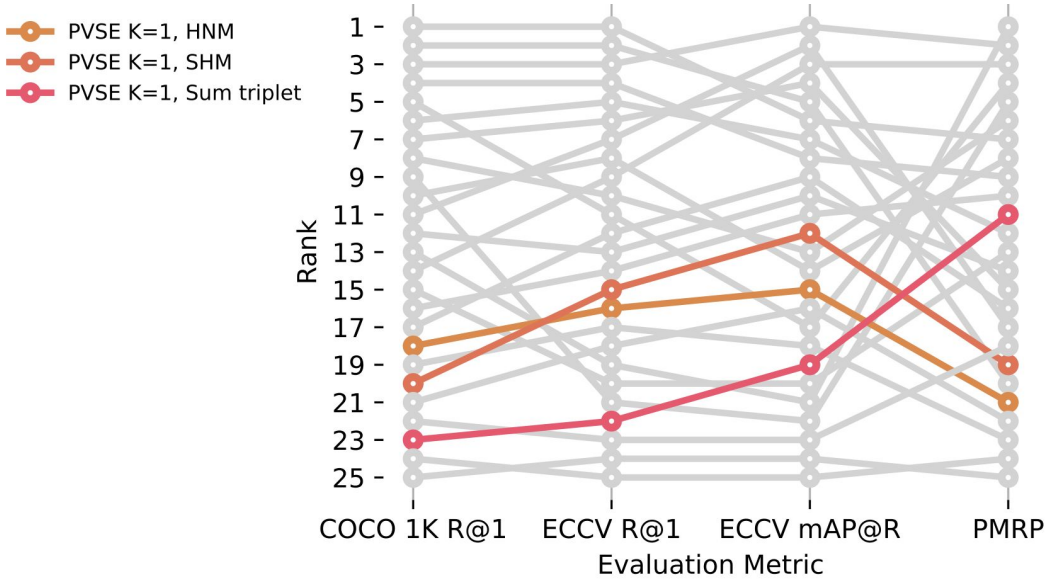
Message 2: R@Ks are not informative



Message 3: Best R@1 models could be worse in mAP@R



Message 4: Our training strategy could be overfitted to R1



References

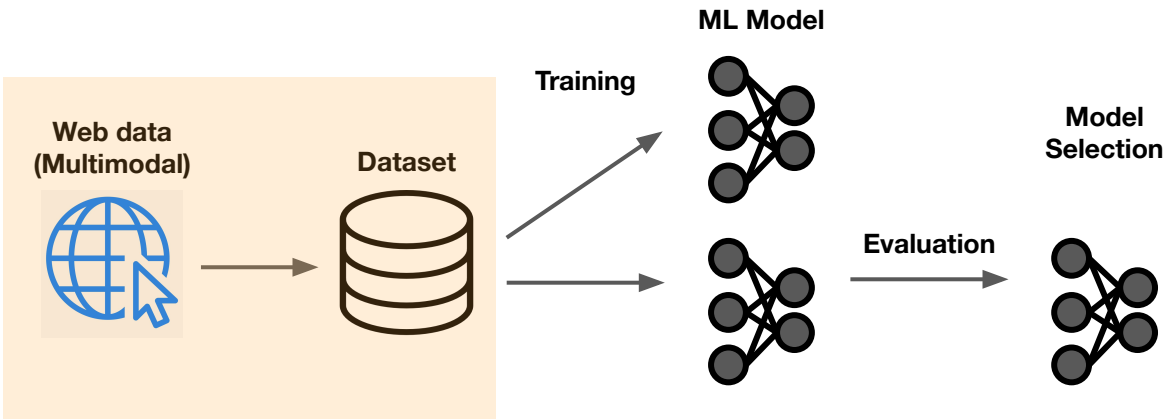
- **[ECCV 2022] ECCV Caption: Correcting False Negatives by Collecting Machine-and-Human-verified Image-Caption Associations for MS-COCO.**
  Sanghyuk Chun, Wonjae Kim, Song Park, Minsuk Chang, Seong Joon Oh
- **[ECCV 2024 SyntheticData4CV Workshop] RoCOCO: Robust Benchmark of MS-COCO to Stress-test Robustness of Image-Text Matching Models.**
  Seulki Park, Daeho Um, Hajung Yoon, **Sanghyuk Chun**, Sangdoo Yun

Contents

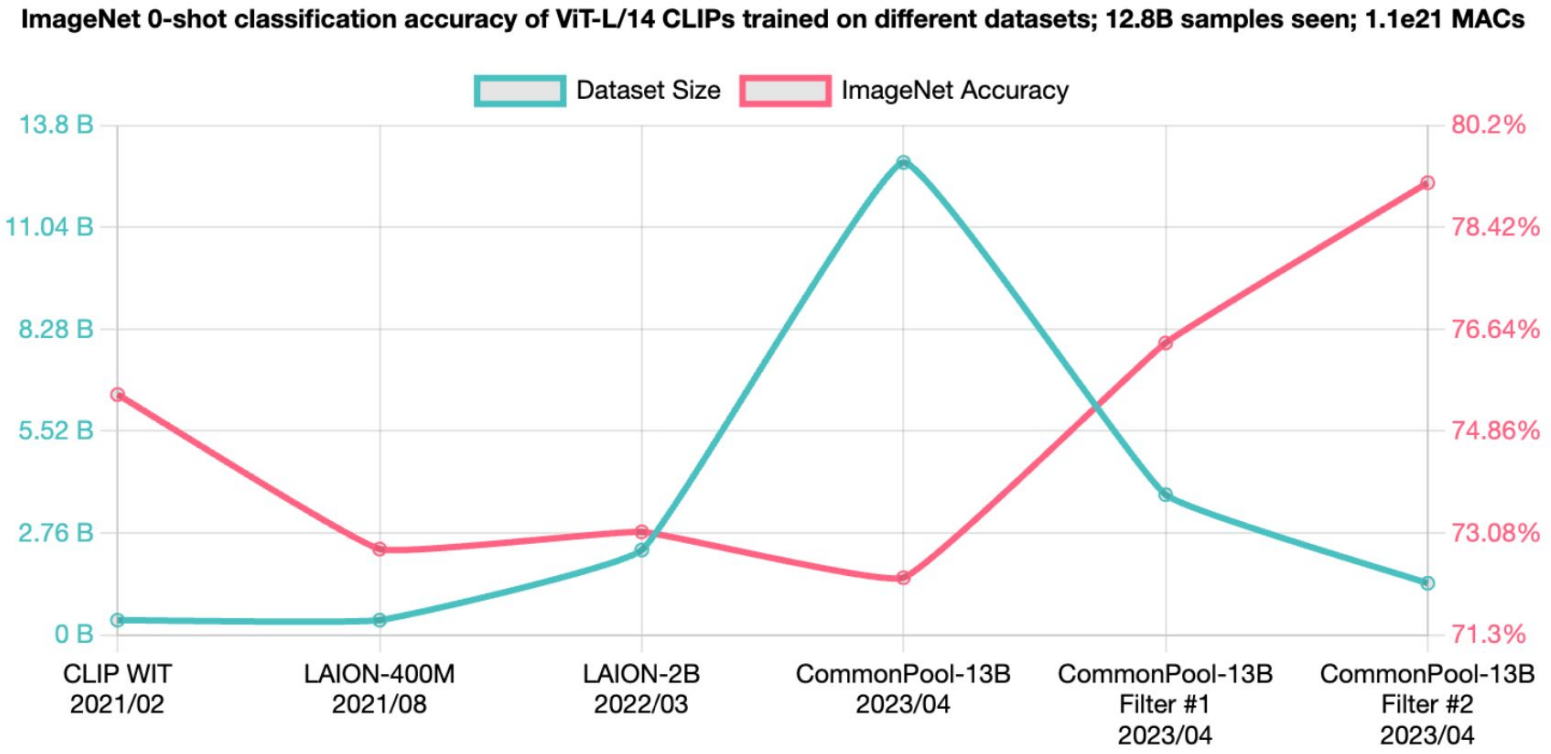
- Introduction: **The multiplicity problem** in multimodal learning
- How can we handle **the multiplicity in contrastive learning**?
- How can we handle **the multiplicity in evaluation**?
- How can we handle **the multiplicity in dataset construction**?
- Conclusion

Multiplicity: A Core Challenge in Multimodal Learning

- Multiplicity makes multimodal learning challenging in all the core components (**Dataset construction**, Contrastive learning, Evaluation)
- Challenging to **construct a high-quality large VL dataset** due to the “hidden GTs” in the dataset; especially for “general” and “ambiguous” images/texts
- How can we handle this?

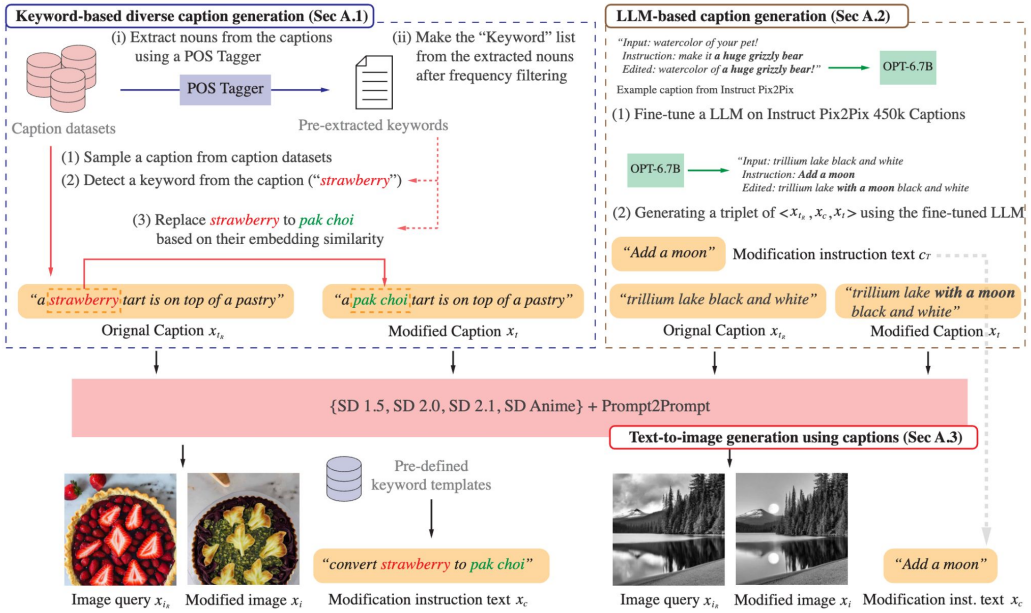


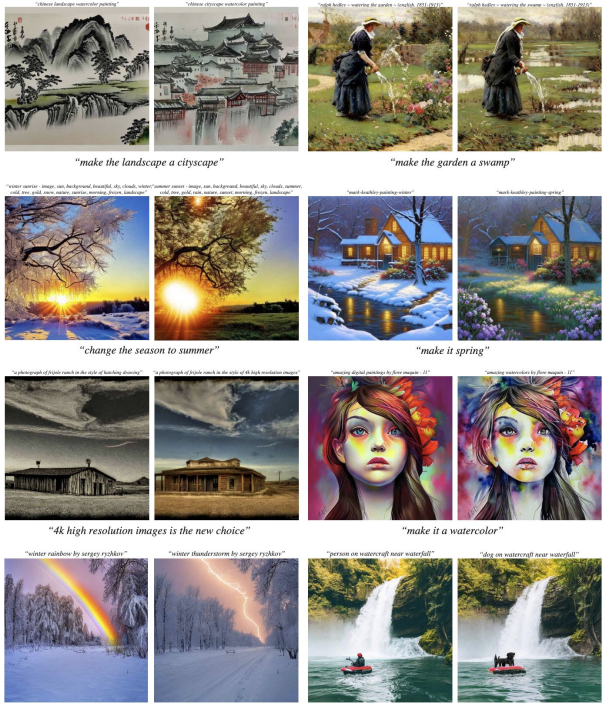
Multimodal Model Training – Scale and quality matter



Can syntactic data be a shortcut?

- We now have very strong text-to-image models as well as instruction-based image editing models
- We can directly use the models to construct a large-scale dataset





Gu and **Chun** et al., "CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion", TMLR 2024.

Pitfall of syntactic datasets

- As datasets grow larger, the performance improvements tend to plateau
- The data quality can be poor; especially when we re-use data multiple times

	IP2P(1M)	1M	5M	10M	18.8M
FashionIQ Avg(R@10, R@50)					
ARTEMIS	26.03	27.44	36.17	41.35	40.62
Combiner	29.83	29.64	35.23	41.81	41.84
CompoDiff	27.24	31.91	38.11	42.41	42.33
CIRR Avg(R@1, R_s@1)					
ARTEMIS	14.91	15.12	15.84	17.56	17.35
Combiner	16.50	16.88	17.21	18.77	18.47
CompoDiff	27.42	28.32	31.50	37.25	37.83

Gu and **Chun** et al., “CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion”, TMLR 2024.

Difficulty of multimodal data construction compared to unimodal data construction

- Unimodal data: we only consider the “quality” of the input, e.g., image resolution, text length, popularity, …
- Multimodal data: we should consider both modality and define “correct matches” between modalities
  - Furthermore, the "paired" or "positive" relationship between two modalities is often ambiguous (which depends on “task”)
- A possible idea:
  - Use only easy-to-scale and easy-to-collect unimodal data to train multimodal models (e.g., LinCIR [Gu and **Chun** et al. 2024])
  - We can easily scale up the number of training samples by this

Gu and **Chun** et al., “Language-only Efficient Training of Zero-shot Composed Image Retrieval”, CVPR 2024.

What would be the correct alignment?

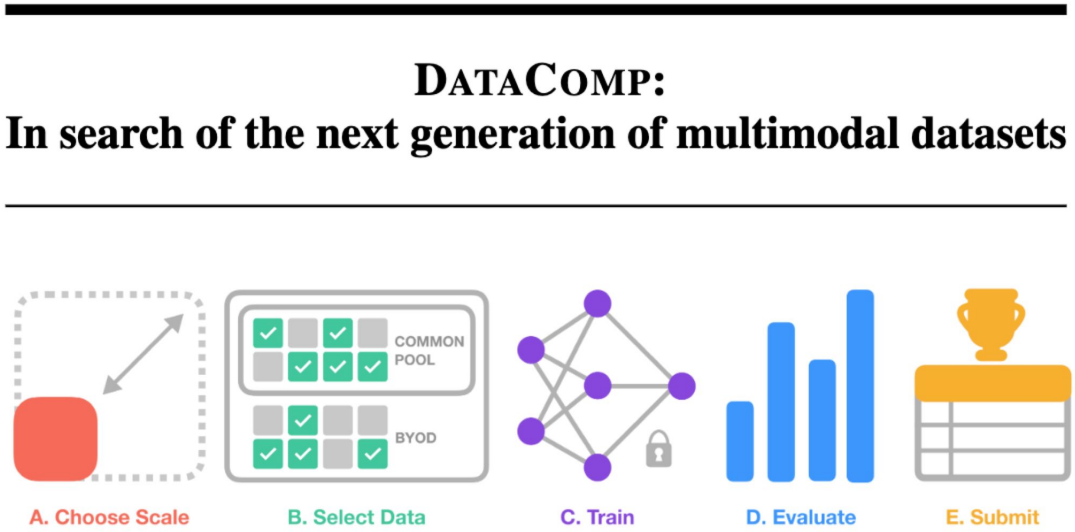
- **It depends on the task!**
- Vision-Language alignment
  - Only describing category of the main object?
  - Exhaustively describing all the possible objects or backgrounds?
  - Describing what happens next based on the visual information?
- Audio-visual alignment
  - On-screen sound: sounds produced by the main object in the scene (e.g., door closing sound)
  - The speech of the talking person, considering their lip movements
  - Off-screen sounds: sounds produced by occluded objects
  - Sounds considering ambient or context (e.g., background music, diegetic sound)
  - Mixture of them
  - => We may need a text condition to specify the alignment (Jeong et al. 2025)

Jeong et al., “Read, Watch and Scream! Sound Generation from Text and Video”, AAAI 2025.

Current approaches for large-scale data construction

- Step 1. Collect billion-scale multimodal pairs from web
- Step 2. Apply filtering algorithms to the collected pairs
  - Convention 1. Apply heuristic filtering algorithms such as, “English-only filter”, “text-length filter” or “image quality filter” for modality-wise filter
  - Convention 2. Apply “alignment-based filtering” using a strong pre-trained CLIP model

DataComp – Multimodal dataset filtering benchmark



CommonPool dataset, unfiltered **12.8B** image-text pairs.

Evaluation: Constrains the model size and the number of samples seen during training.

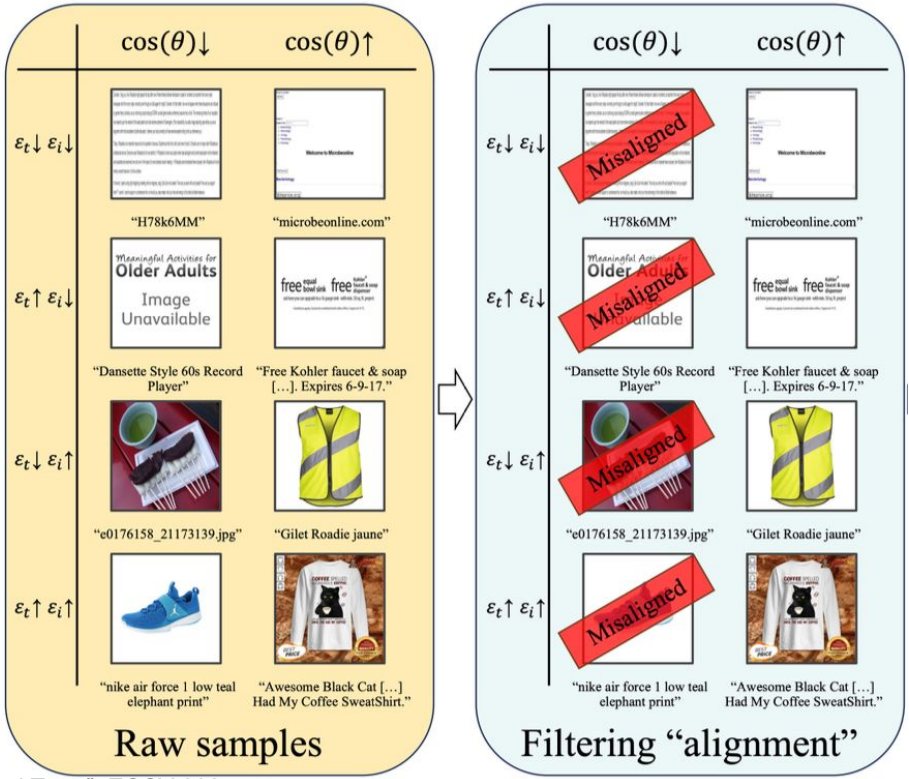
DataComp: In search of the next generation of multimodal datasets

DataComp – Naive filtering

Filtering by CLIP score

Use a CLIP model trained with clean image-text pairs

However, CLIP scores would be noisy (Many-to-many nature)



Kim et al., "HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts", ECCV 2024

Cleanness of data – Specificity

Previous filtering methods empirically shown so far,

- Less ambiguous images and texts
- Long texts are better than short texts
- Content about well-known topics (named entities)

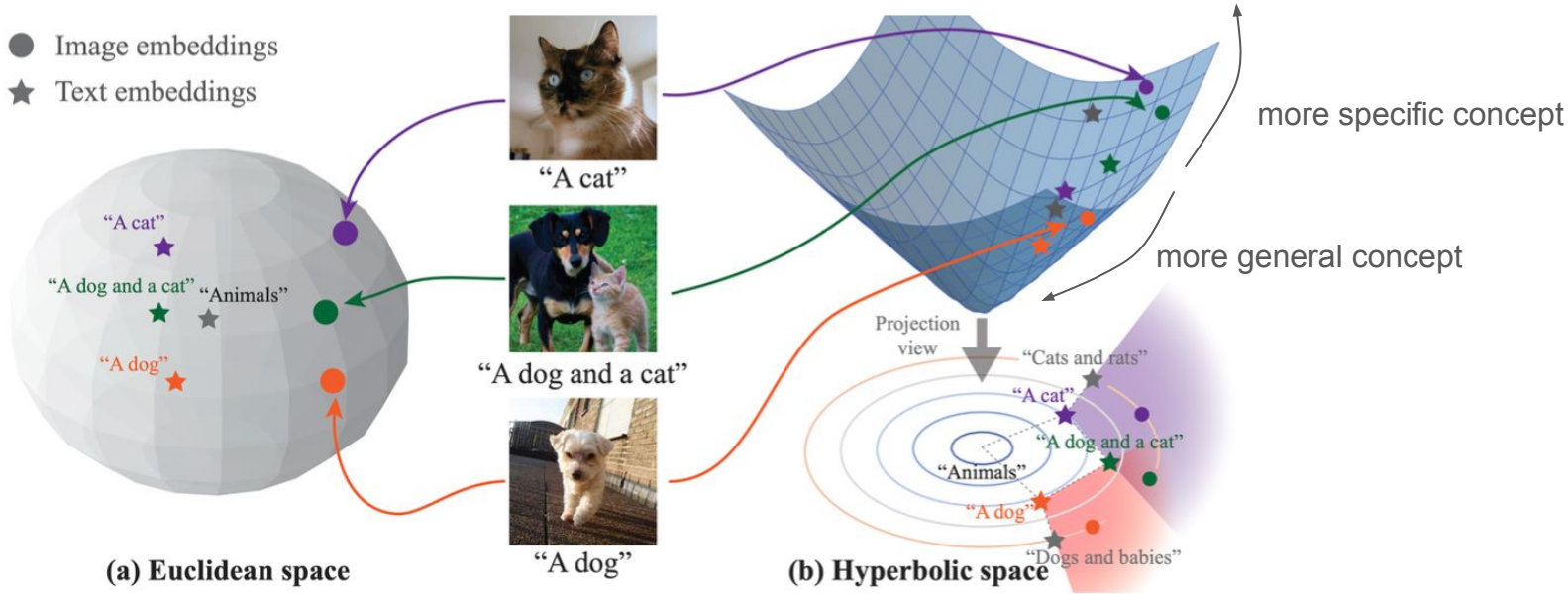
are better.

We assume they are high **specificity** data points.

(i.e., it should have clear concept and with limited correspondences)

How to get the specificity?

Modeling hierarchical relationship

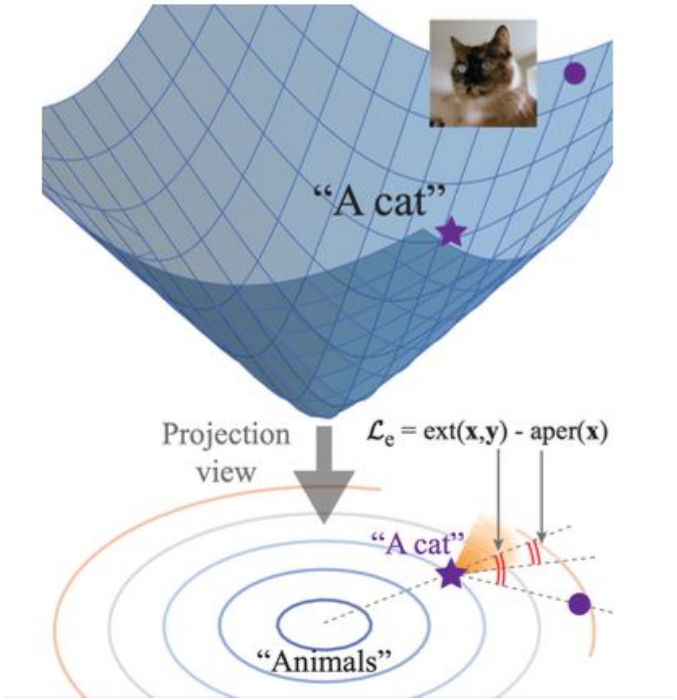


Kim et al., "HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts", ECCV 2024

Use Pre-trained Hyperbolic CLIP[1]

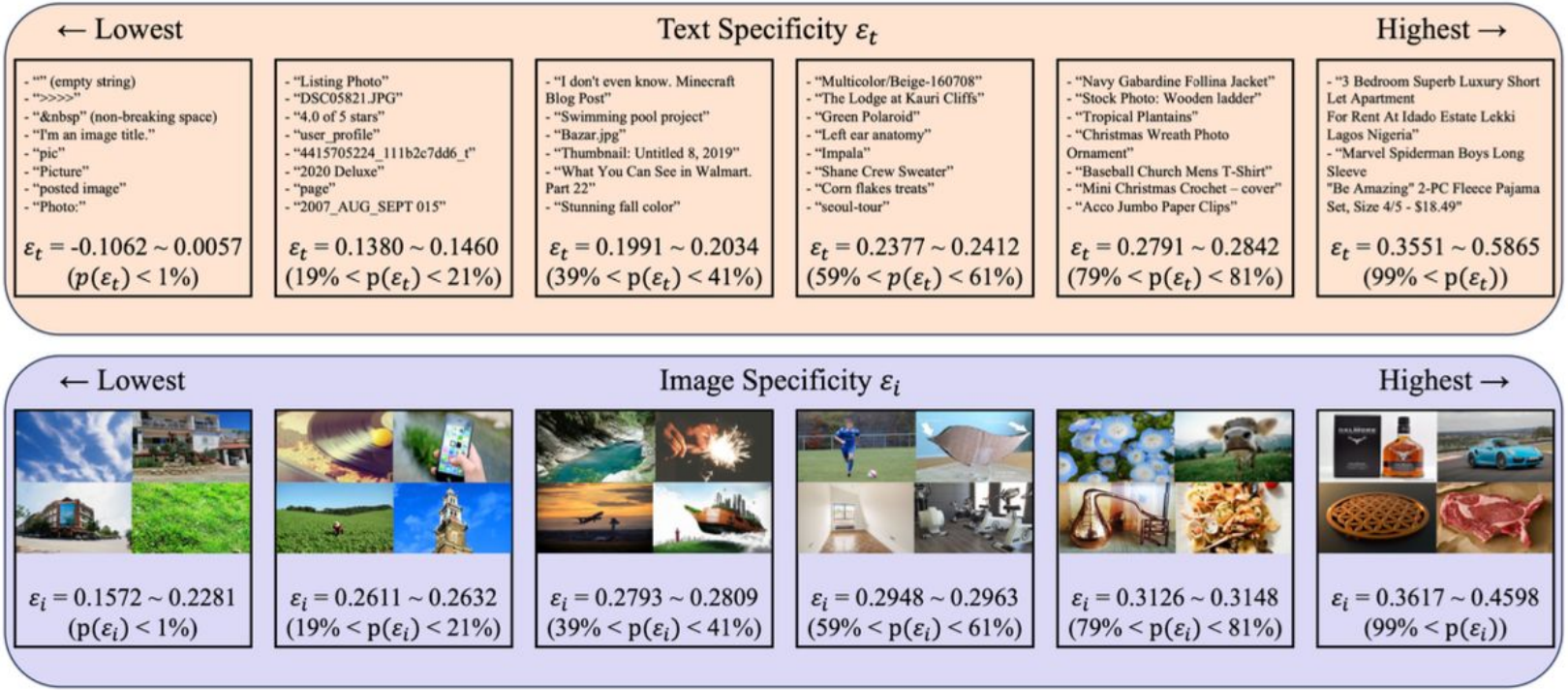
For each data point in hyperbolic embedding space, we compute the **entailment** loss, measuring how much it entails or is entailed by others on average.

→ We define this as "**specificity**" of the data point



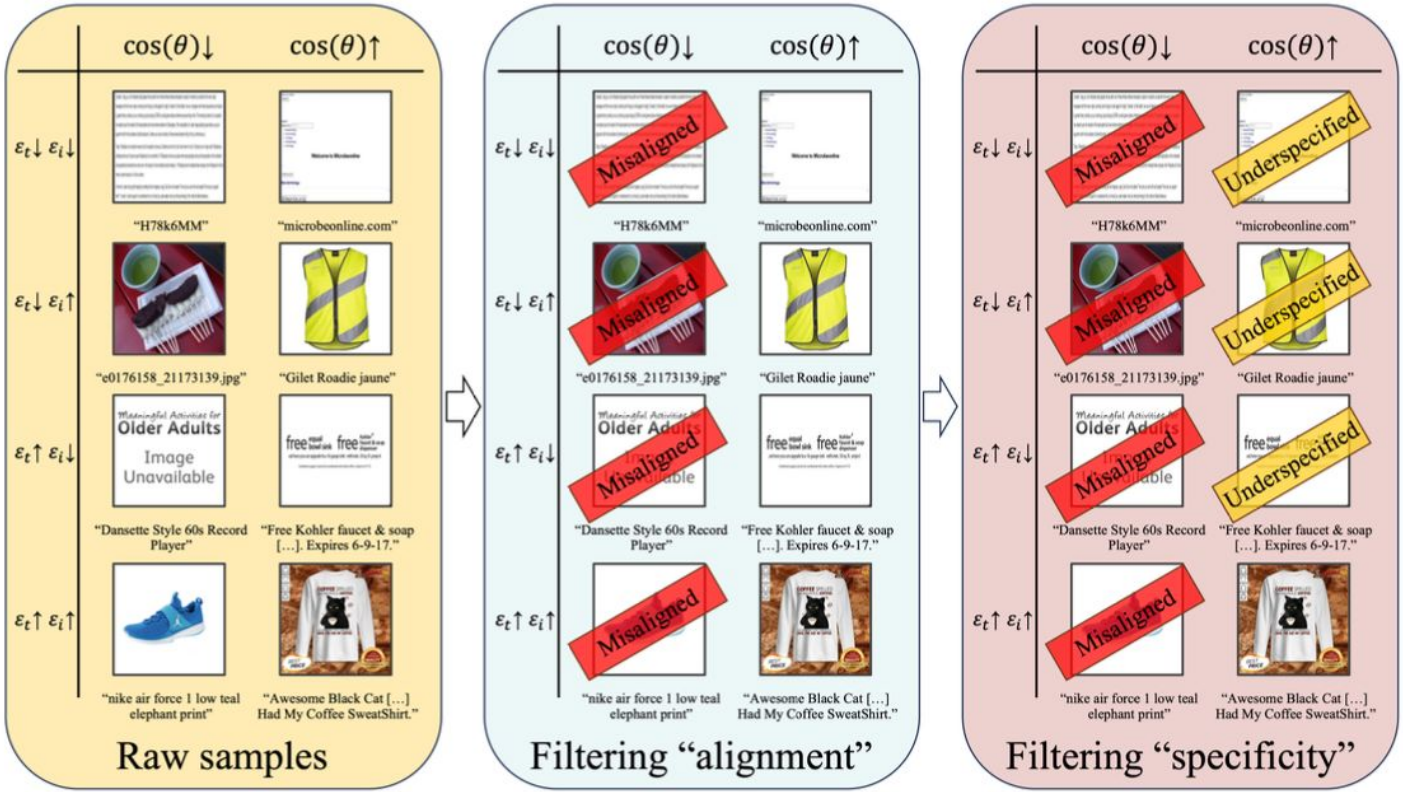
[1] Desai et al., "Hyperbolic Image-Text Representations", ICML 2023.

Specificity



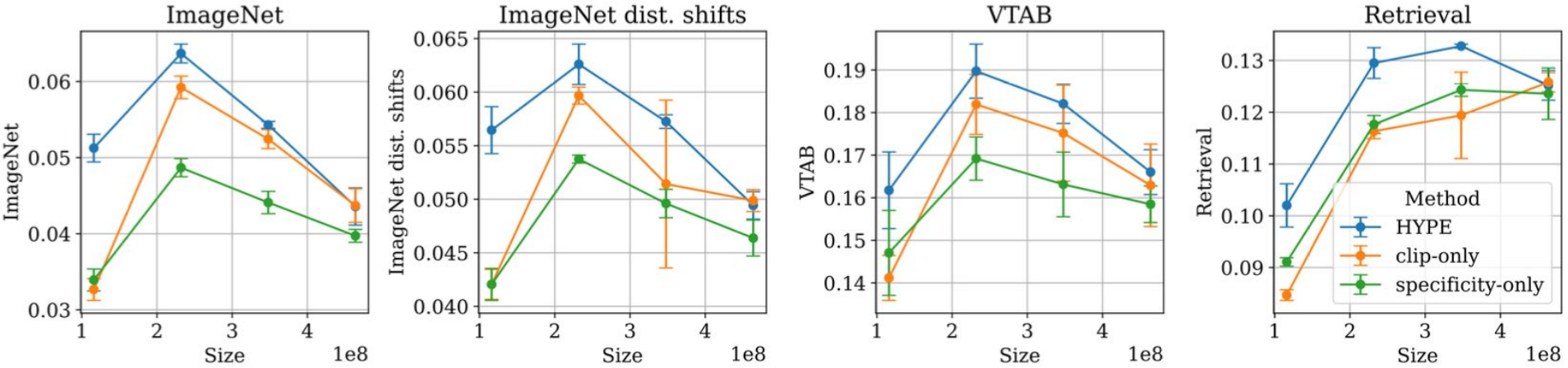
Kim et al., "HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts", ECCV 2024

Our filtering – HYPE (CLIP + Entailment)



Kim et al., "HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts", ECCV 2024

Our filtering – HYPE (CLIP + Entailment)



Kim et al., “HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts”, ECCV 2024

Our filtering – HYPE (CLIP + Entailment)

Leaderboard

Rank	Created	Submission	ImageNet acc.	Average perf.
1	06-21-2024	negCLIPLoss & NormSim + DFN + HYPE	0.382	0.388
2	06-21-2024	negCLIPLoss & NormSim + DFN	0.375	0.386
3	11-08-2023	Hype sampler + DFN	0.382	0.379
4	11-07-2023	Hype sampler	0.346	0.373
5	10-02-2023	Data Filtering Networks	0.371	0.373
6	09-08-2023	The Devil Is in the Details	0.320	0.371
7	09-08-2023	TMARS + SSFT	0.338	0.362
8	08-17-2023	T-MARS: Improving Visual Representations by Circumventing Text Feature Learning	0.330	0.361
9	09-08-2023	The Devil Is in the Details - ImageNet best	0.336	0.355

References

- **[ECCV 2024] HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts.**
  Wonjae Kim, **Sanghyuk Chun**, Taekyung Kim, Dongyoon Han, Sangdoo Yun
- **[TMLR 2024] CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion.**
  Geonmo Gu*, **Sanghyuk Chun***, Wonjae Kim, HeeJae Jun, Yoohoon Kang, Sangdoo Yun
- **[CVPR 2024] Language-only Efficient Training of Zero-shot Composed Image Retrieval.**
  Geonmo Gu*, **Sanghyuk Chun***, Wonjae Kim, Yoohoon Kang, Sangdoo Yun
- **[AAAI 2025] Read, Watch and Scream! Sound Generation from Text and Video.**
  Yujin Jeong, Yunji Kim, **Sanghyuk Chun**, Jiyoung Lee

Contents

Multiplicity is a Core Challenge in Multimodal Learning

- Multiplicity complicates multimodal learning at every level;
  from data collection to training objectives to evaluation
- Probabilistic representation learning could be helpful for fixing the inherent limitation of deterministic representations under multiplicity
- We need to consider many-to-many evaluation suites rather than simply focusing on one-to-one relationship and “Recall@K”
- Until now, the rule-of-thumb of data construction is: collecting large-scale multimodal pairs from the web and then filtering the collected pairs.
  - We have to consider ambiguity or specificity of each data point

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 Workshop 2025] LongProLIP: A Probabilistic Vision-Language Model with Long Context Text.**
  Sanghyuk Chun, Sangdoo Yun

References

- **[ECCV 2024 SyntheticData4CV Workshop] RoCOCO: Robust Benchmark of MS-COCO to Stress-test Robustness of Image-Text Matching Models.**
  Seulki Park, Daeho Um, Hajung Yoon, **Sanghyuk Chun**, Sangdoo Yun
- **[ECCV 2024] HYPE: Hyperbolic Entailment Filtering for Underspecified Images and Texts.**
  Wonjae Kim, **Sanghyuk Chun**, Taekyung Kim, Dongyoon Han, Sangdoo Yun
- **[TMLR 2024] CompoDiff: Versatile Composed Image Retrieval With Latent Diffusion.**
  Geonmo Gu*, **Sanghyuk Chun***, Wonjae Kim, HeeJae Jun, Yoohoon Kang, Sangdoo Yun
- **[CVPR 2024] Language-only Efficient Training of Zero-shot Composed Image Retrieval.**
  Geonmo Gu*, **Sanghyuk Chun***, Wonjae Kim, Yoohoon Kang, Sangdoo Yun
- **[AAAI 2025] Read, Watch and Scream! Sound Generation from Text and Video.**
  Yujin Jeong, Yunji Kim, **Sanghyuk Chun**, Jiyoung Lee

Resources

- ProLIP
  - Inference & training code: https://github.com/naver-ai/prolip
  - 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
- ECCV Caption
  - https://github.com/naver-ai/eccv-caption
  - `pip3 install eccv_caption`
- HYPE
  - Filtered uuids: https://github.com/naver-ai/hype