Towards Data-Centric Multimodal ML: Vision-Language Dataset Distillation

Xindi Wu
Department of Computer Science
Princeton University

https://xindiwu.github.io/
xindiw@princeton.edu

PRINCETON UNIVERSITY

Princeton Visual AI Lab

Overview

1. Motivation
2. Previous Work
3. Method
4. Experiments
5. Conclusion

| Motivation

| Motivation

What is Data-centric Multimodal ML?

Why is dataset distillation important?

What is Multimodal?

A dictionary definition:

Multimodal: with multiple modalities

A research-oriented definition:

Multimodal is the science of heterogeneous and interconnected data

Liang, Zadeh, and Morency. Tutorial on Multimodal Machine Learning. CVPR 2022, NAACL 2022

Data is the cornerstone of innovation in multimodal ML

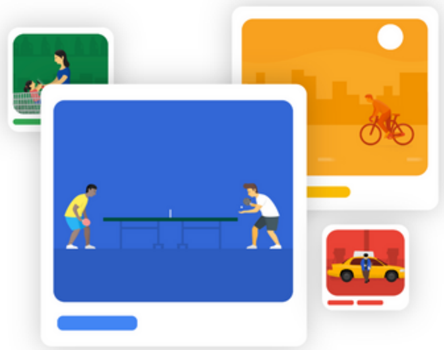
Flickr30K (2014)

Visual Genome (2016)

LAION 400M/5B (2021)

COCO (2014)

Conceptual Captions (2018)



DataComp (2023)

Data-Centric Multimodal ML

Data-Centric Multimodal ML

Data = Information + Noise

John Wright and Yi Ma. High-dimensional data analysis with low-dimensional models: Principles, computation, and applications. Cambridge University Press, 2022.

Data-Centric Multimodal ML

Data = Information + Noise

How can we identify the most critical information from datasets?

Data-Centric Multimodal ML

Data = Information + Noise

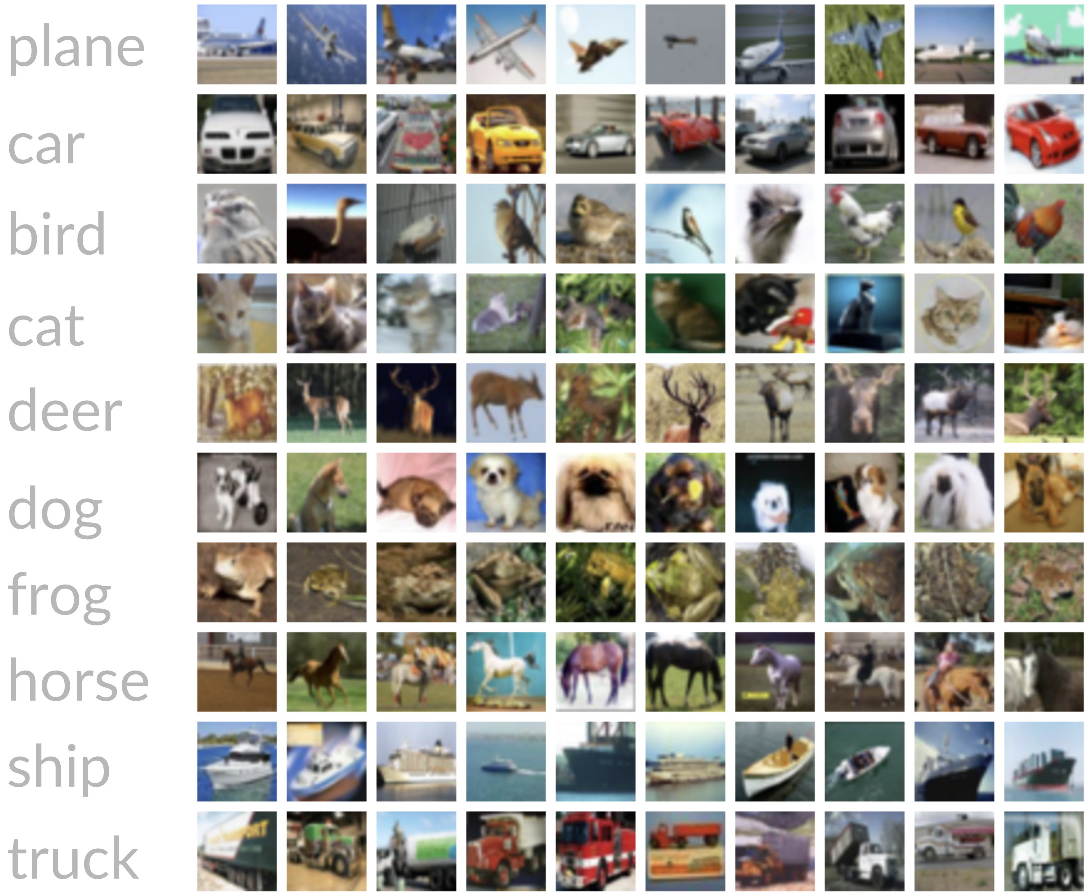
How can we identify the most critical information from datasets?

Dataset distillation **is one promising solution!**

What is Dataset Distillation?

CIFAR10 examples

N images



CIFAR10

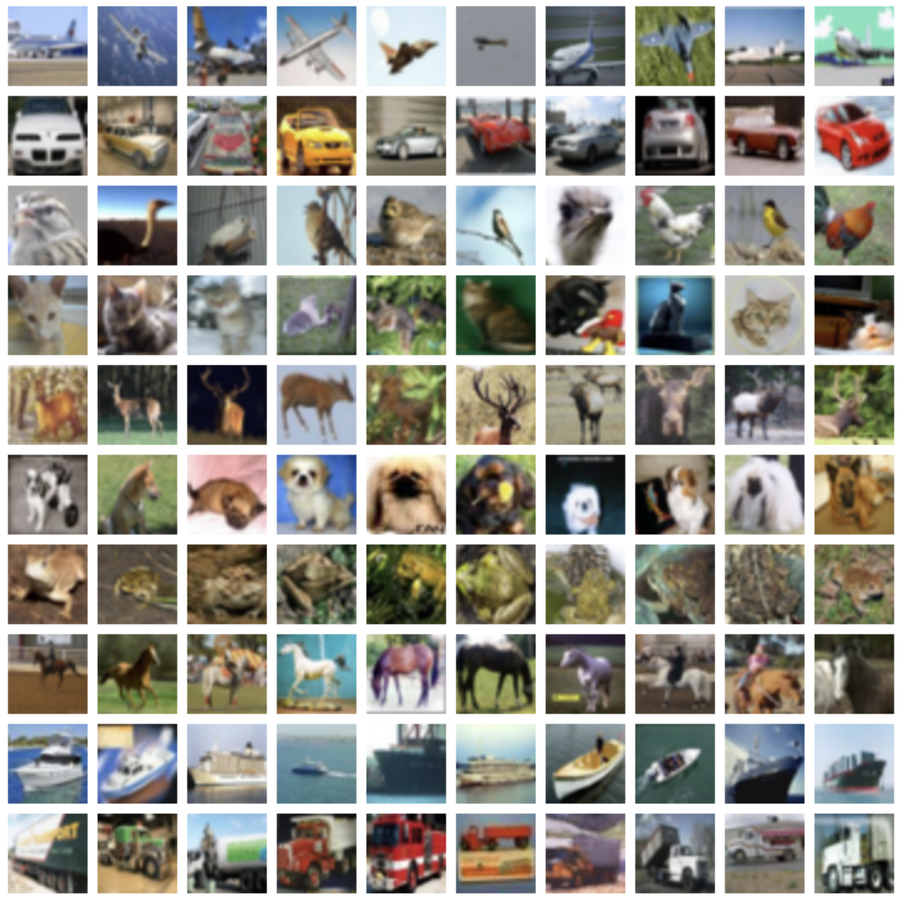
(5000 images/class)

What is Dataset Distillation?

CIFAR10 examples

N images

plane
car
bird
cat
deer
dog
frog
horse
ship
truck



CIFAR10

(5000 images/class)

N >> M

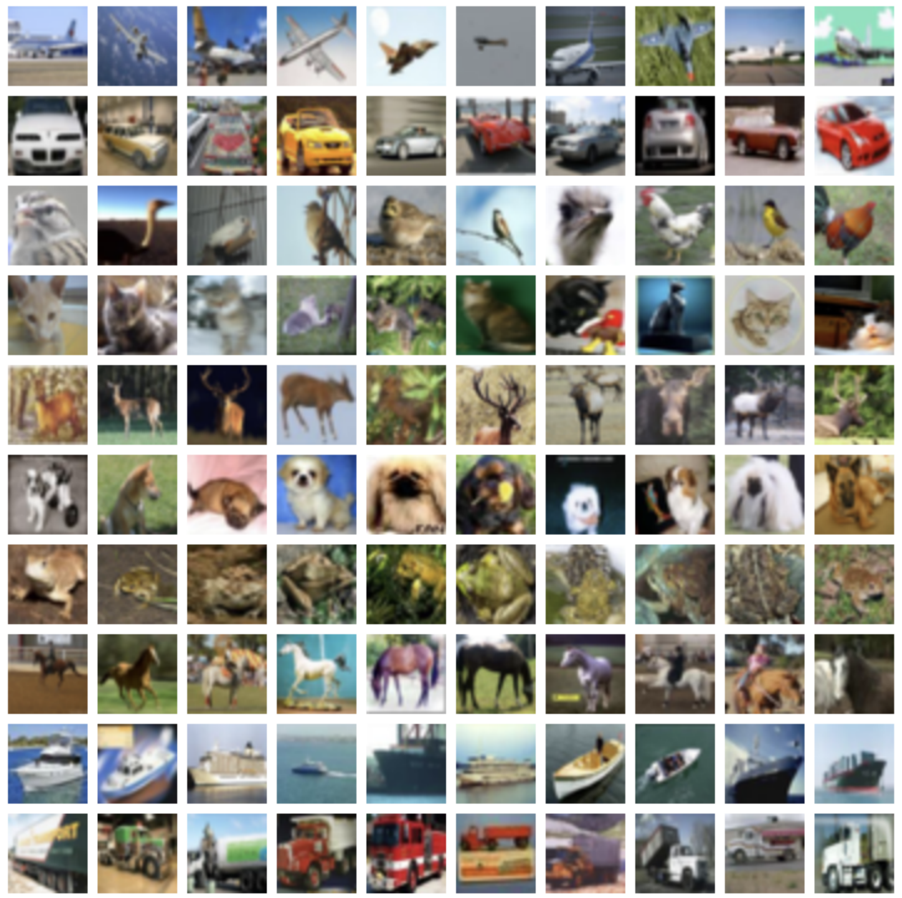
What is Dataset Distillation?

CIFAR10 examples

N images

M images

plane
car
bird
cat
deer
dog
frog
horse
ship
truck



CIFAR10

(5000 images/class)

N >> M

Distilled CIFAR10

(1 images/class)

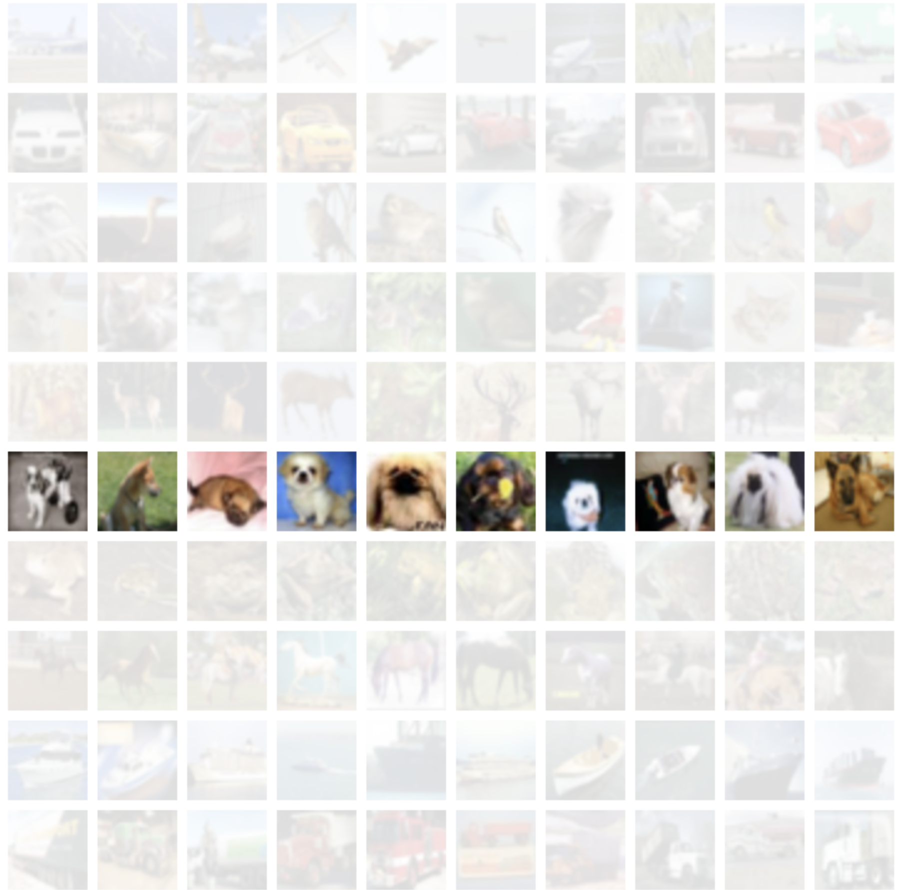
What is Dataset Distillation?

CIFAR10 examples

N images

M images

plane
car
bird
cat
deer
dog
frog
horse
ship
truck



N >> M

CIFAR10

(5000 images/class)

Distilled CIFAR10

(1 images/class)

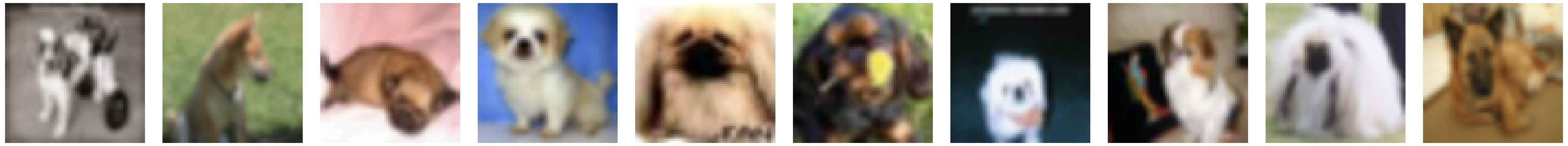
What is Dataset Distillation?

CIFAR10 examples

dog

CIFAR10

(5000 images/class)

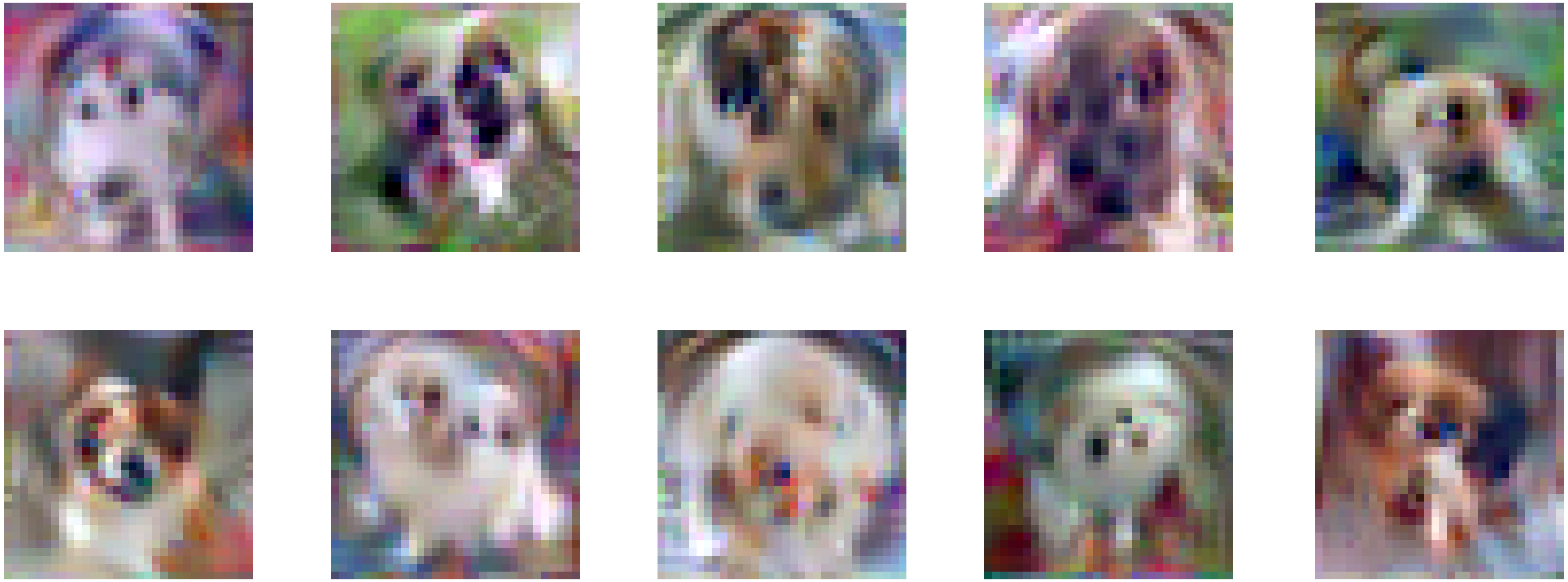


Distilled CIFAR10

(1 images/class)

What is Dataset Distillation?

CIFAR10 examples



Distilled CIFAR10 (dog)

(10 images/class)

What is Dataset Distillation?

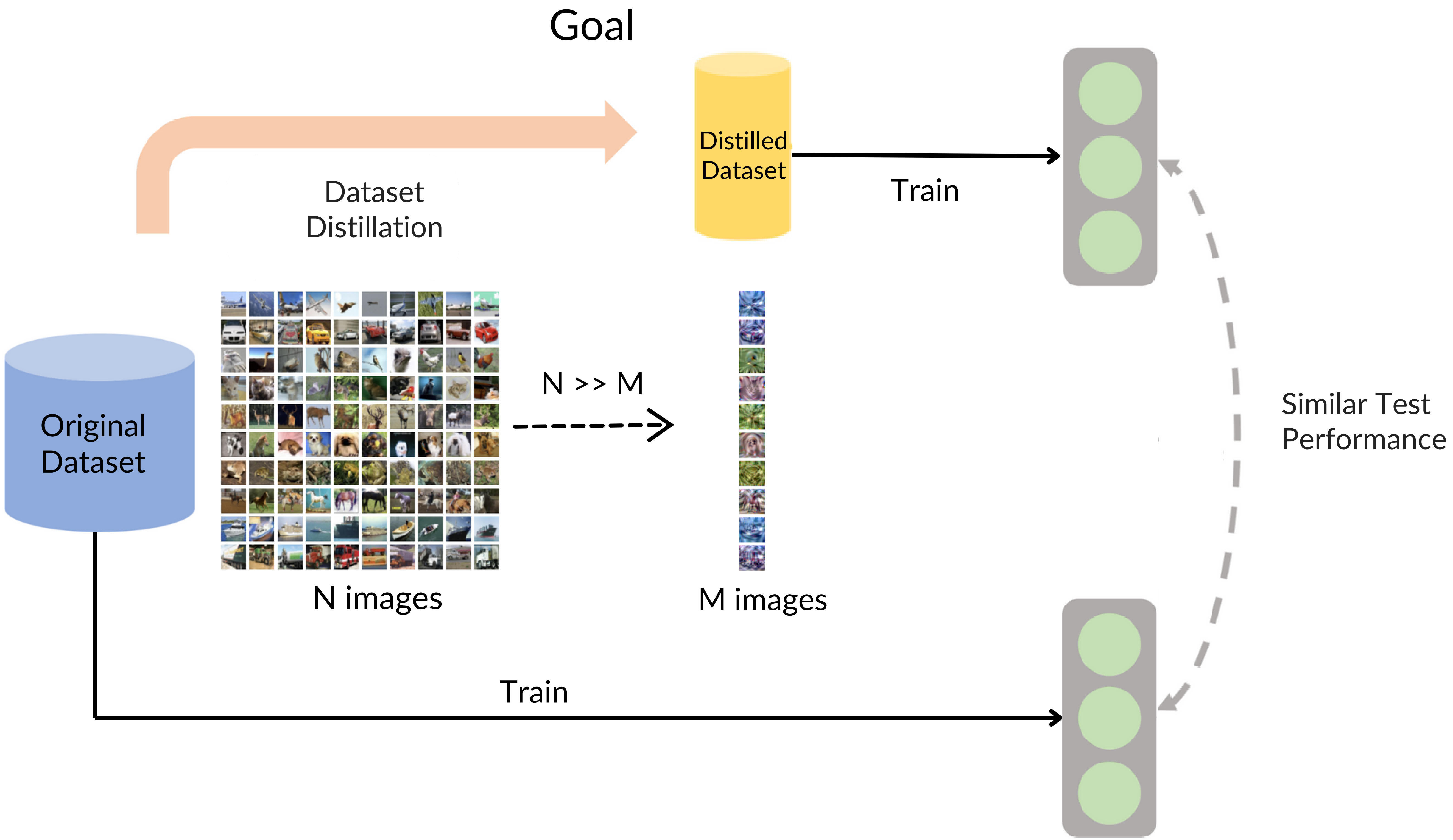


Figure credit Yu, Ruonan, Songhua Liu, and Xinchao Wang. "Dataset distillation: A comprehensive review." arXiv preprint arXiv:2301.07014 (2023).

What is Dataset Distillation?

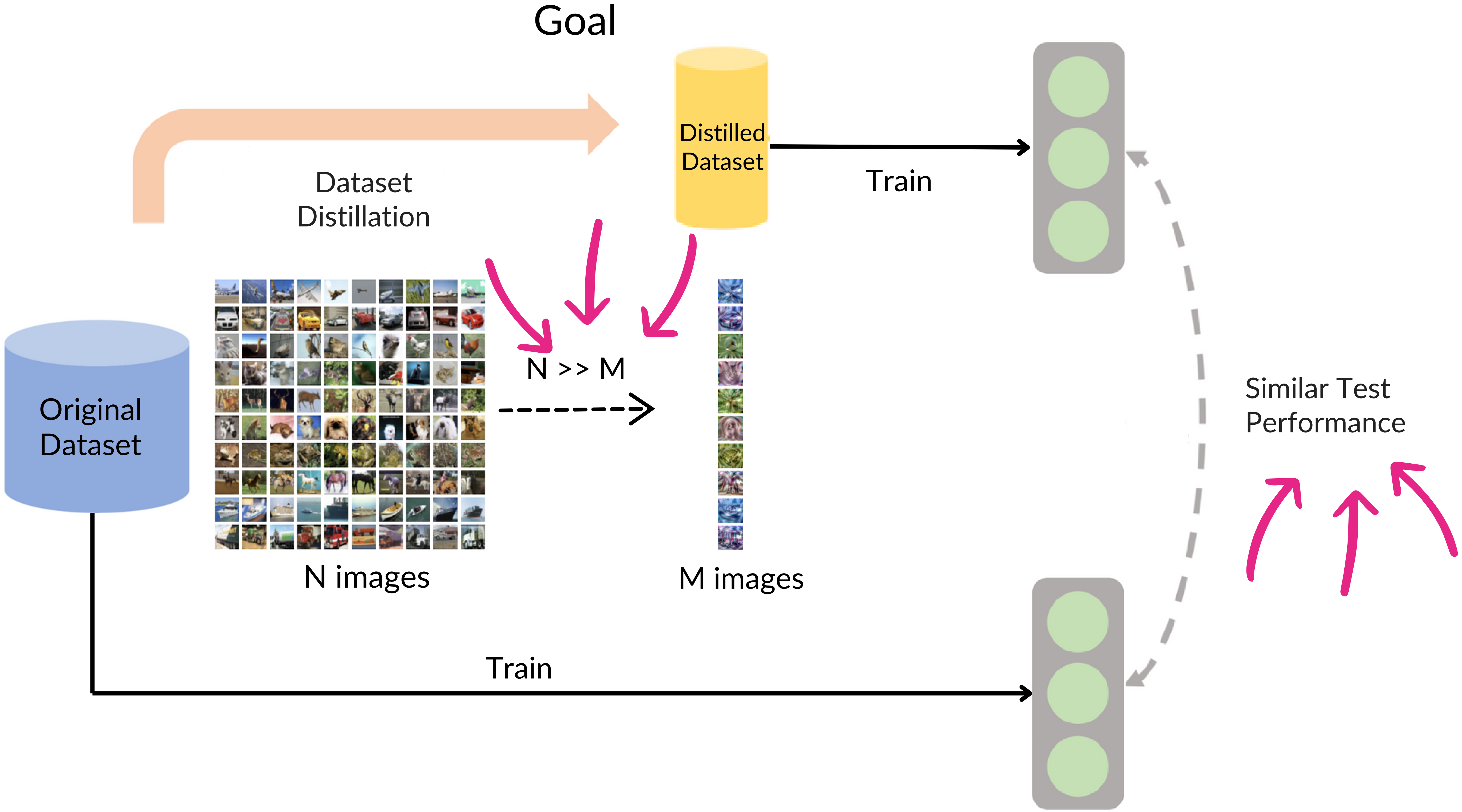


Figure credit Yu, Ruonan, Songhua Liu, and Xinchao Wang. "Dataset distillation: A comprehensive review." arXiv preprint arXiv:2301.07014 (2023).

What is Dataset Distillation?

Defination

(Loose) Definition[1]:
Approaches that aim to synthesize tiny and high-fidelity data summaries which distill the most important knowledge from a given target dataset.

[1] Sachdeva, Noveen, and Julian McAuley. "Data distillation: A survey." TMLR, 2023.

What is Dataset Distillation?

Defination

(Loose) Definition[1]:
Approaches that aim to synthesize tiny and high-fidelity data summaries which distill the most important knowledge from a given target dataset.

Such distilled summaries are optimized to serve as effective drop-in replacements of the original dataset for efficient and accurate data-usage applications.

[1] Sachdeva, Noveen, and Julian McAuley. "Data distillation: A survey." TMLR, 2023.

What is Dataset Distillation?

Application

- Neural Architecture Search
- Federated Learning
- Continual Learning
- Differential Privacy

| Previous Work

Previous Work

Image Classification

CIFAR10

(5000 images/class)

Distilled CIFAR10

(1 images/class)

Previous Work

Taxonomy

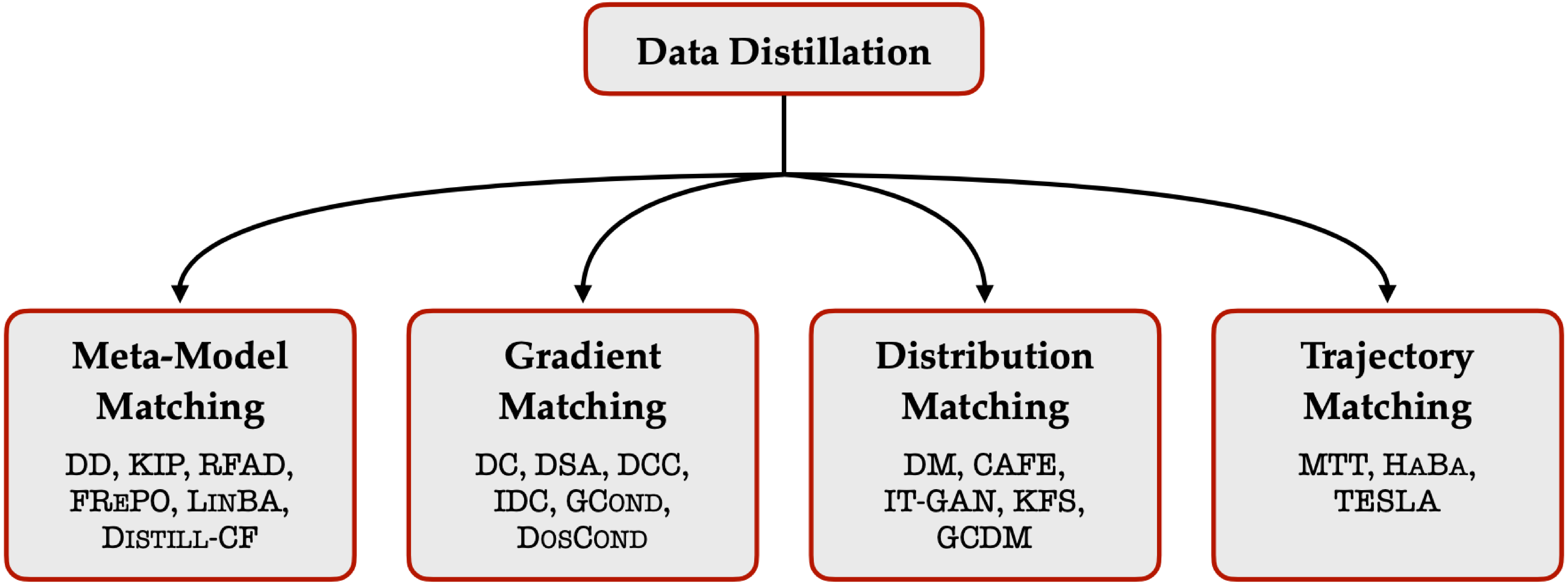


Figure credit: Sachdeva, Noveen, and Julian McAuley. "Data distillation: A survey." *TMLR*, 2023.

Previous Work

Meta-Model Matching Framework

Dataset distillation can be formulated as a bi-level meta-learning problem

Large Scale Dataset $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^{N}$

$M << N$

Small Synthetic Dataset $\hat{\mathbf{D}} = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{M}$

Previous Work

Meta-Model Matching Framework

Dataset distillation can be formulated as a bi-level meta-learning problem

Large Scale Dataset $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^{N}$

$M << N$

Small Synthetic Dataset $\hat{\mathbf{D}} = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{M}$

Model $f\left(\cdot;\theta\right)$

Generalization Loss $\ell\left(f\left(\hat{\mathbf{D}};\theta\right),\mathbf{D}\right)$

Previous Work

Meta-Model Matching Framework

Dataset distillation can be formulated as a bi-level meta-learning problem

Inner level $f\left(\hat{\mathbf{D}};\theta\right)$

Previous Work

Meta-Model Matching Framework

Dataset distillation can be formulated as a bi-level meta-learning problem

Inner level $f\left(\hat{\mathbf{D}};\theta\right)$

Outer level $\hat{\mathbf{D}}' = \underset{\hat{\mathbf{D}}}{\arg min} F\left(\hat{\mathbf{D}}\right)$

Previous Work

Meta-Model Matching Framework

Dataset distillation can be formulated as a bi-level meta-learning problem

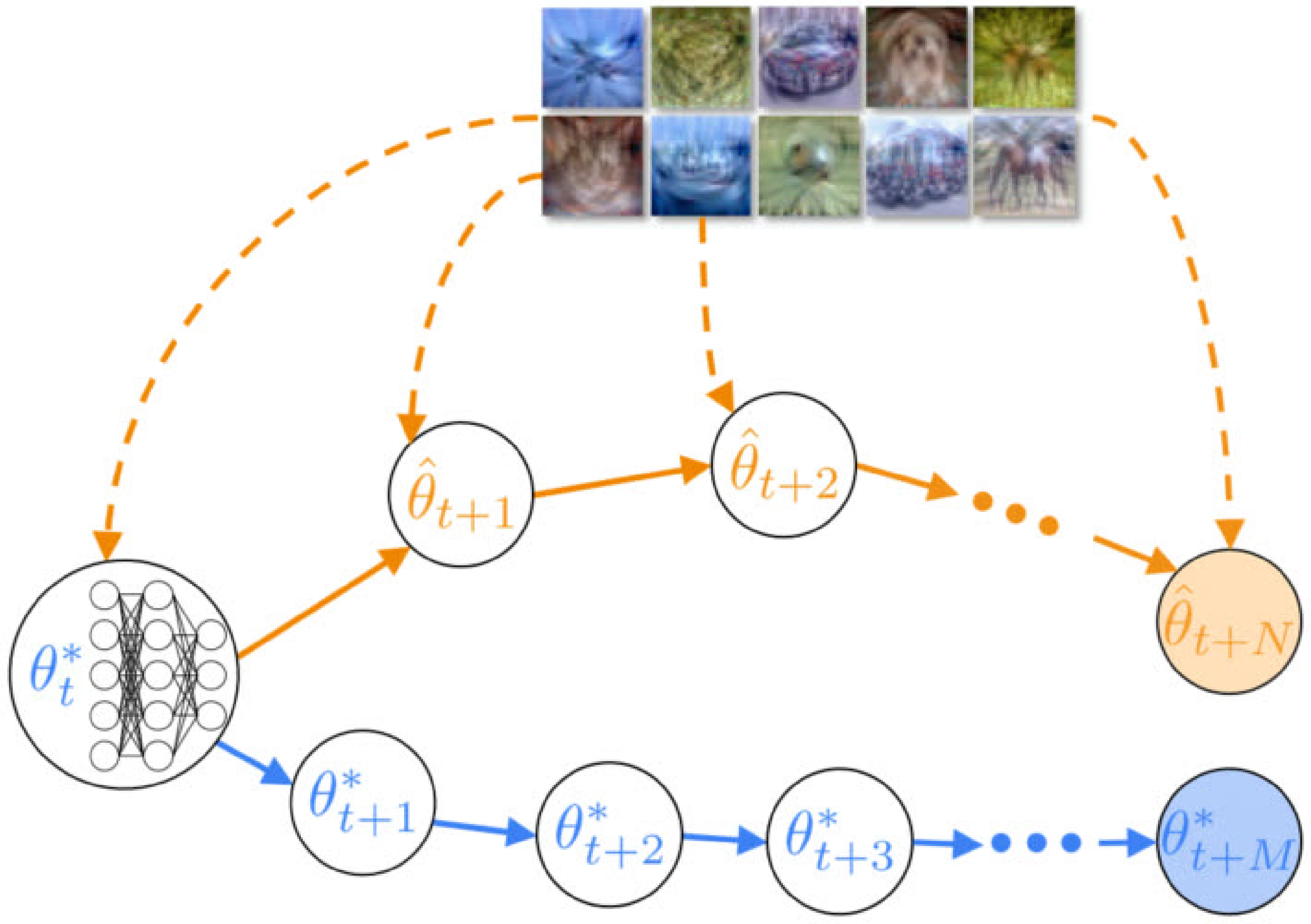
Inner level $f\left(\hat{\mathbf{D}};\theta\right)$

Outer level $\hat{\mathbf{D}}' = \arg\min_{\hat{\mathbf{D}}} F\left(\hat{\mathbf{D}}\right)$

$$F\left(\hat{\mathbf{D}}\right) = \mathbb{E}_{\theta\sim P(\theta)}\ell\left(f\left(\hat{\mathbf{D}};\theta\right),\mathbf{D}\right)$$

Previous Work

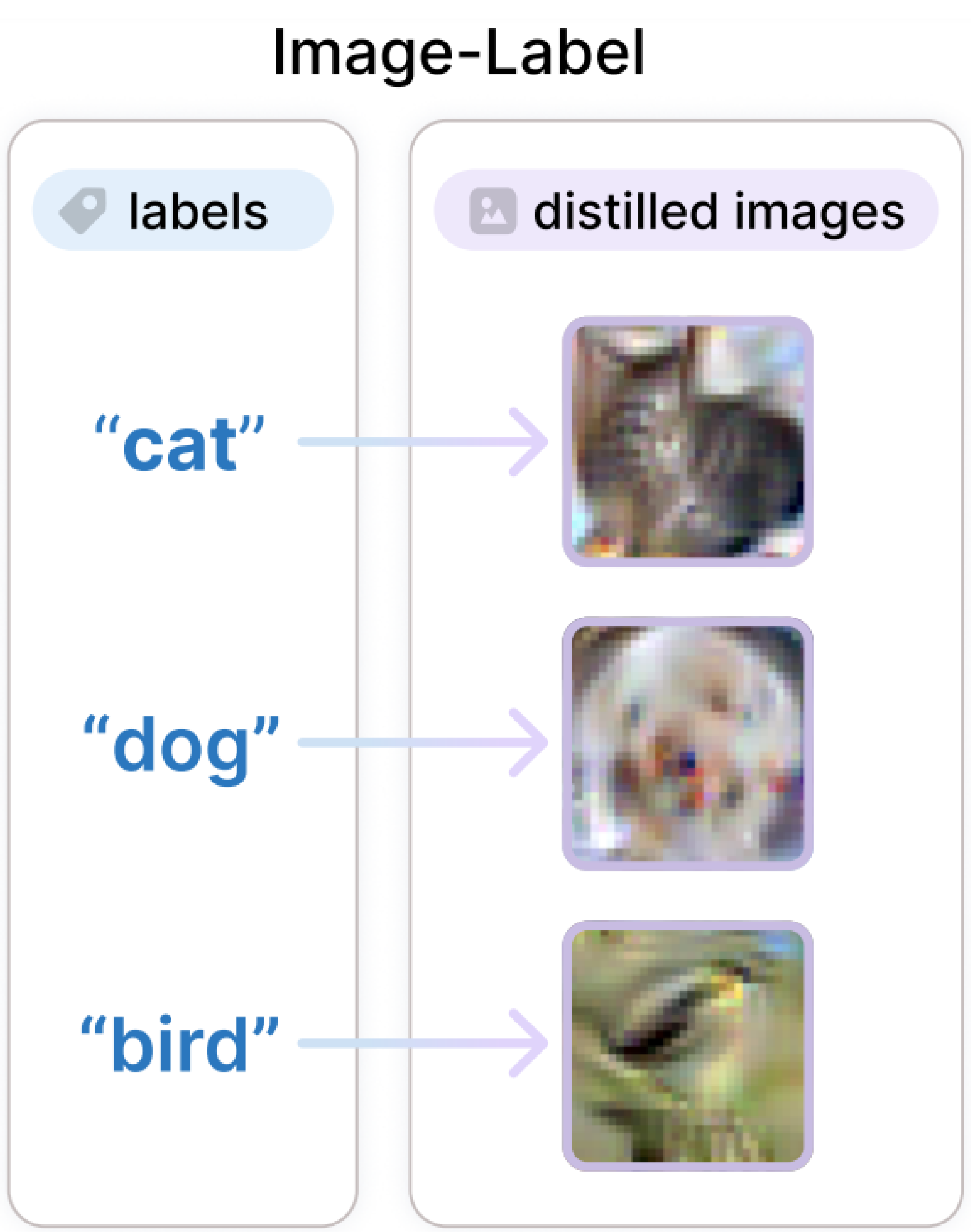
Trajectory Matching Framework



Student Trajectories are trained on Synthetic Data

Cazenavette, George, et al. "Dataset distillation by matching training trajectories." CVPR. 2022.

Comparison



Comparison

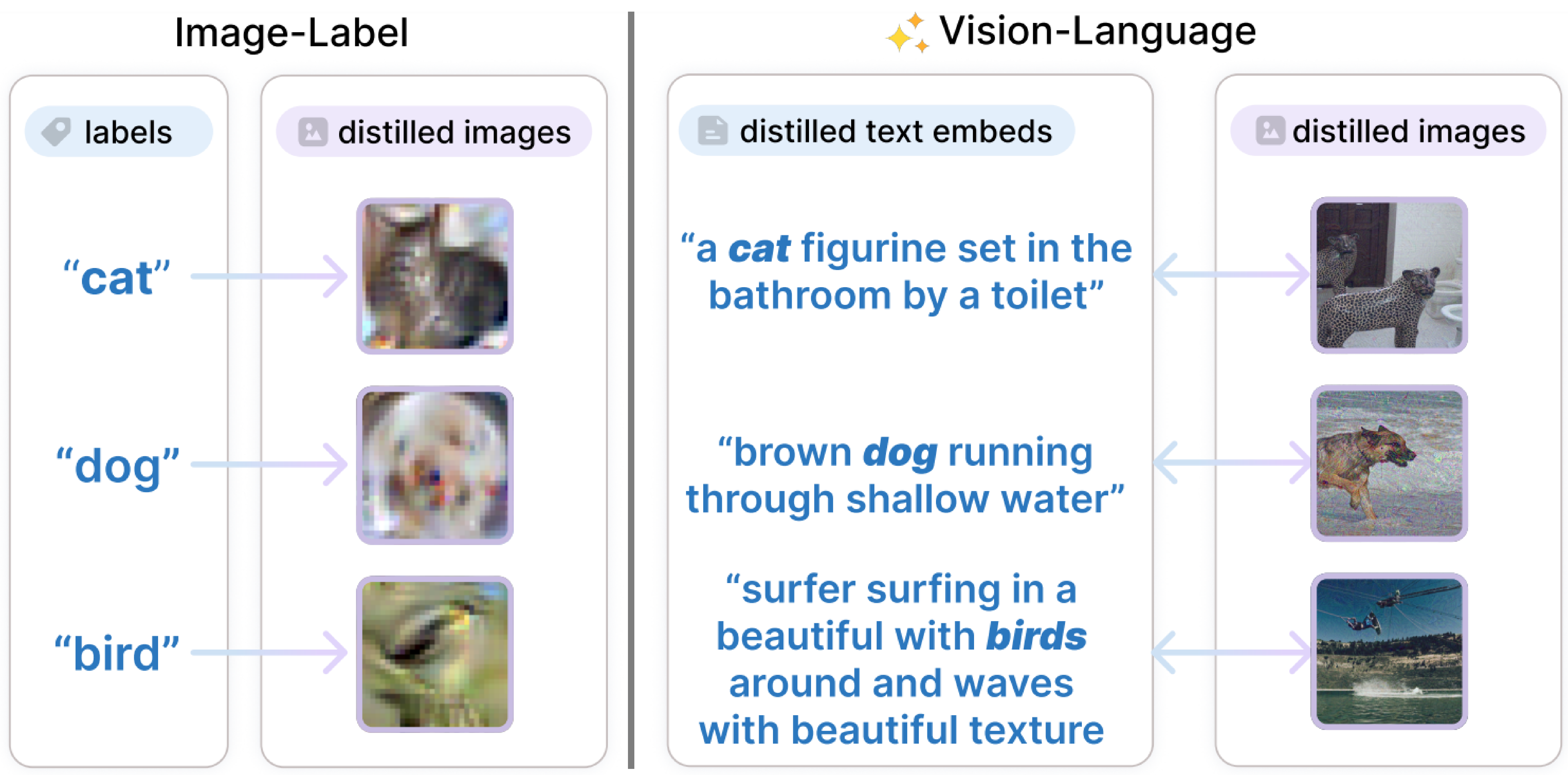


Image-Label
labels
distilled images
"cat"
"dog"
"bird"
Vision-Language
distilled text embeds
distilled images
"a cat figurine set in the bathroom by a toilet"
"brown dog running through shallow water"
"surfer surfing in a beautiful with birds around and waves with beautiful texture

Comparison

What's unique about Vision-Language Dataset?

image

text

A dog wearing a green sweater and fanny pack walks on a snow-covered field.

A small dog wearing a green sweater and a backpack walks through snow.

A dog wearing a green sweater and a backpack walking on snow.

A dog wearing a green sweater and backpack running in the snow.

A small dog wearing a sweater walking in the snow.

Comparison

What's unique about Vision-Language Dataset?

image

text

A **dog** wearing a green sweater and fanny pack walks on a snow-covered field.

A small **dog** wearing a green sweater and a backpack walks through snow.

A **dog** wearing a green sweater and a backpack walking on snow.

A **dog** wearing a green sweater and backpack running in the snow.

A small **dog** wearing a sweater walking in the snow.

Comparison

What's unique about Vision-Language Dataset?

image

text

A **dog** wearing a green sweater and fanny pack walks on a snow-covered field.

A small **dog** wearing a green sweater and a backpack walks through snow.

A **dog** wearing a green sweater and a backpack walking on snow.

A **dog** wearing a green sweater and backpack running in the snow.

A small **dog** wearing a sweater walking in the snow.

Comparison

What's unique about Vision-Language Dataset?

image

text

A **dog** wearing a green sweater and fanny pack walks on a snow-covered field.

A small dog wearing a green sweater and a backpack walks through snow.

A **dog** wearing a green sweater and a backpack walking on snow.

A **dog** wearing a green sweater and backpack running in the snow.

A small **dog** wearing a sweater walking in the snow.

Comparison

What's unique about Vision-Language Dataset?

image

text

A **dog** wearing a green sweater and fanny pack walks on a snow-covered field.

A small dog wearing a green sweater and a backpack walks through snow.

A **dog** wearing a green sweater and a backpack walking on snow.

A **dog** wearing a green sweater and backpack running in the snow.

A small **dog** wearing a sweater walking in the snow.

Comparison

What's unique about Vision-Language Dataset?

image

text

A **dog** wearing a green sweater and fanny pack walks on a snow-covered field.

A small dog wearing a green sweater and a backpack walks through snow.

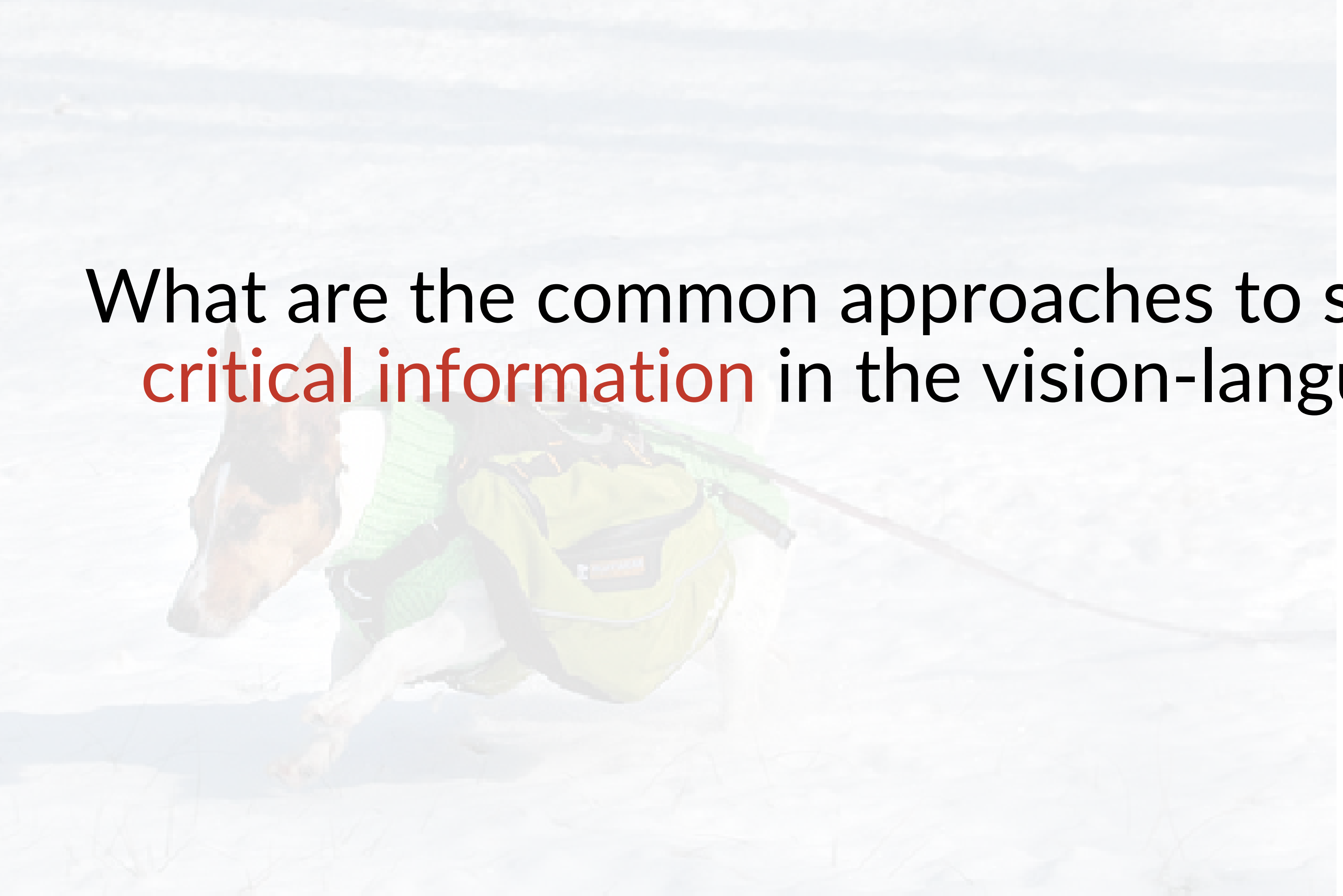
A **dog** wearing a green sweater and a backpack walking on snow.

A **dog** wearing a green sweater and backpack running in the snow.

A small **dog** wearing a sweater walking in the snow.

Comparison

image



text

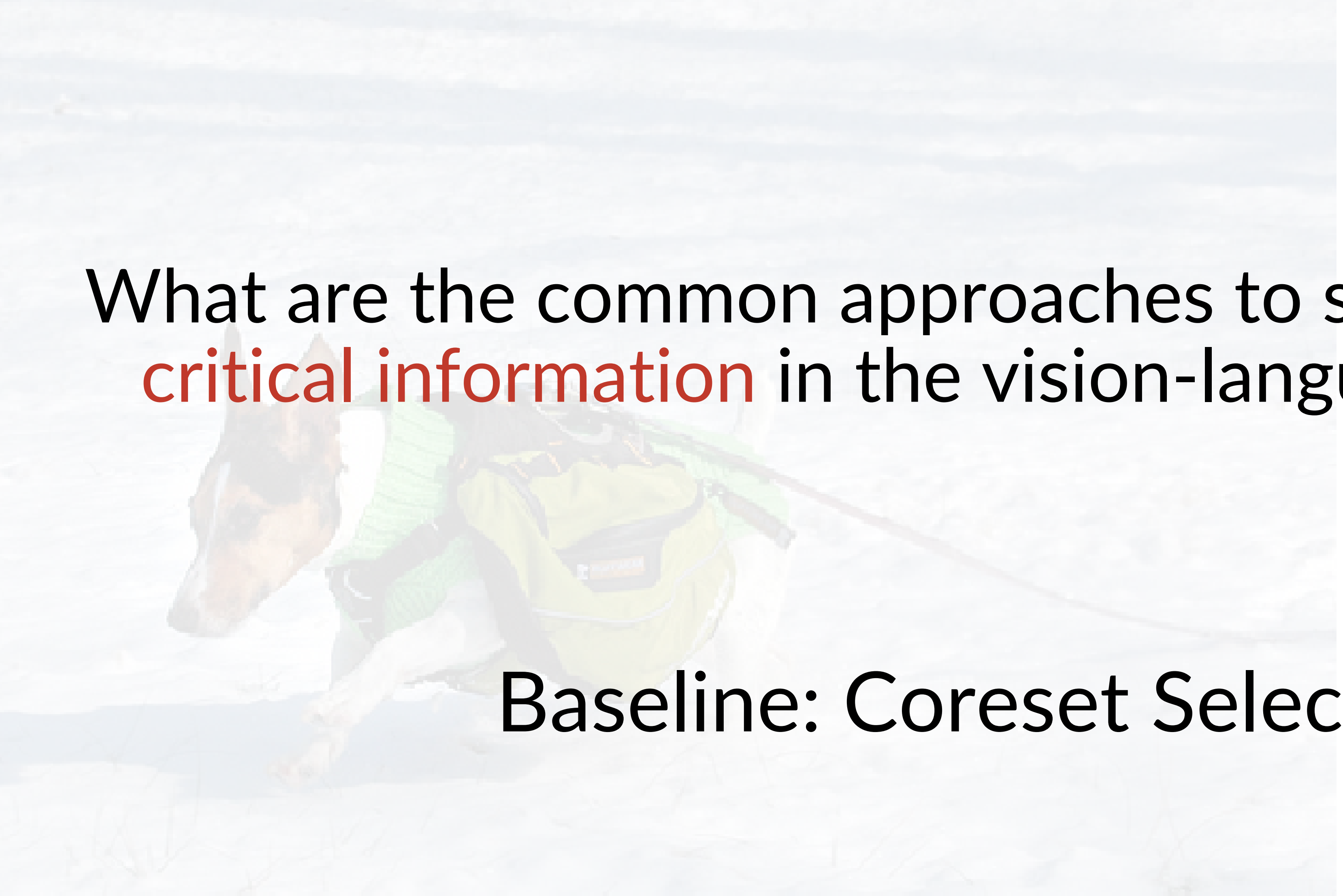
A dog wearing a green sweater and fanny pack walks on a snow-covered field.

A dog wearing a green sweater and backpack running in the snow.

A small dog wearing a sweater walking in the snow.

What are the common approaches to select the most critical information in the vision-language dataset?

Comparison

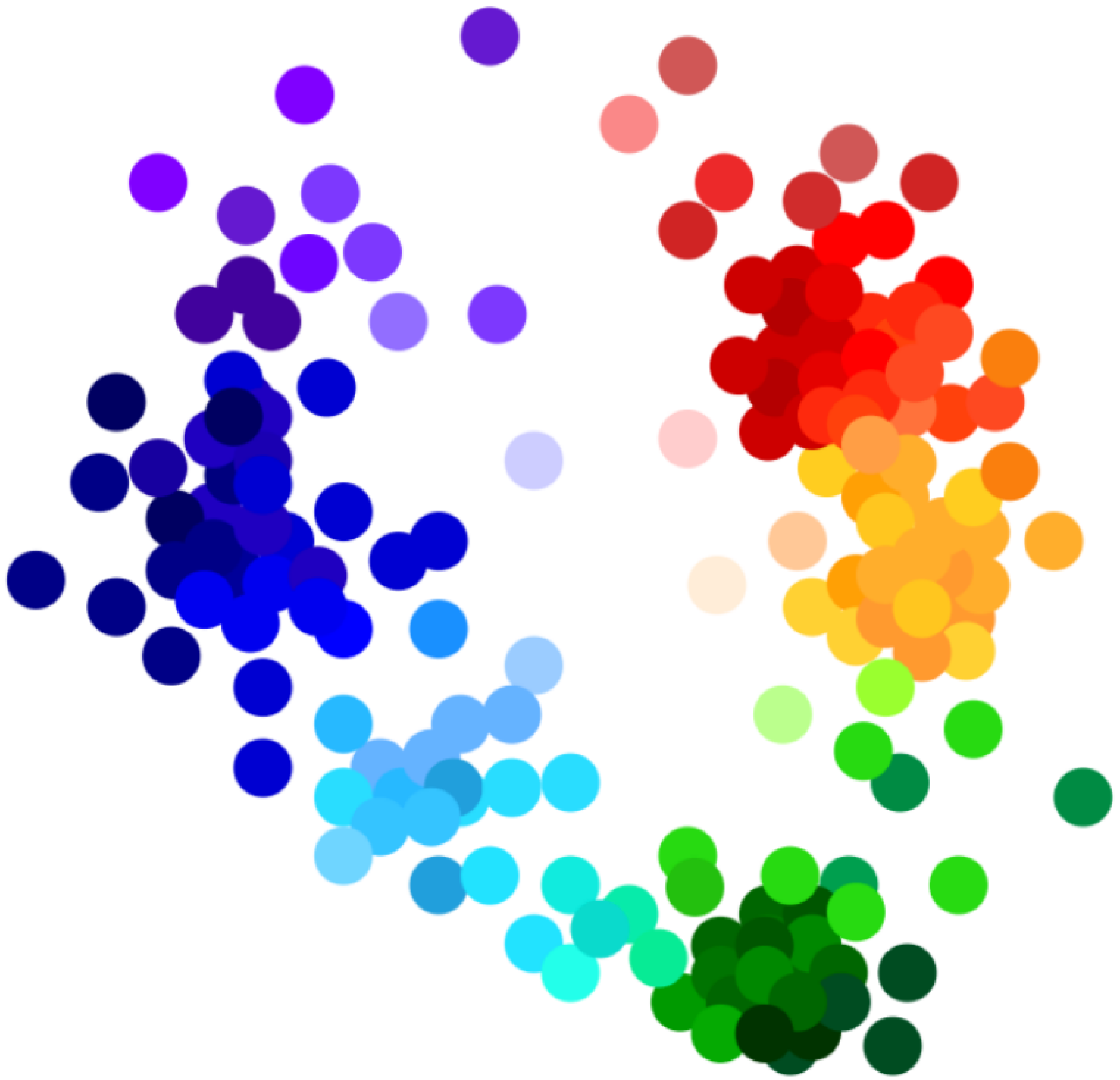


What are the common approaches to select the most critical information in the vision-language dataset?

Baseline: Coreset Selection

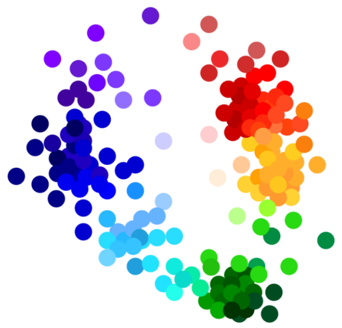
Baseline

Coreset Selection



Baseline

Coreset Selection

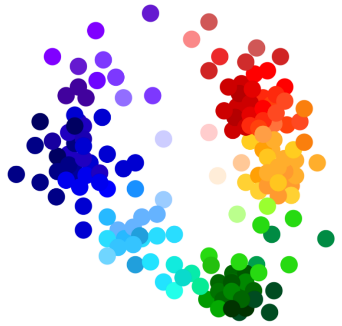


Herding

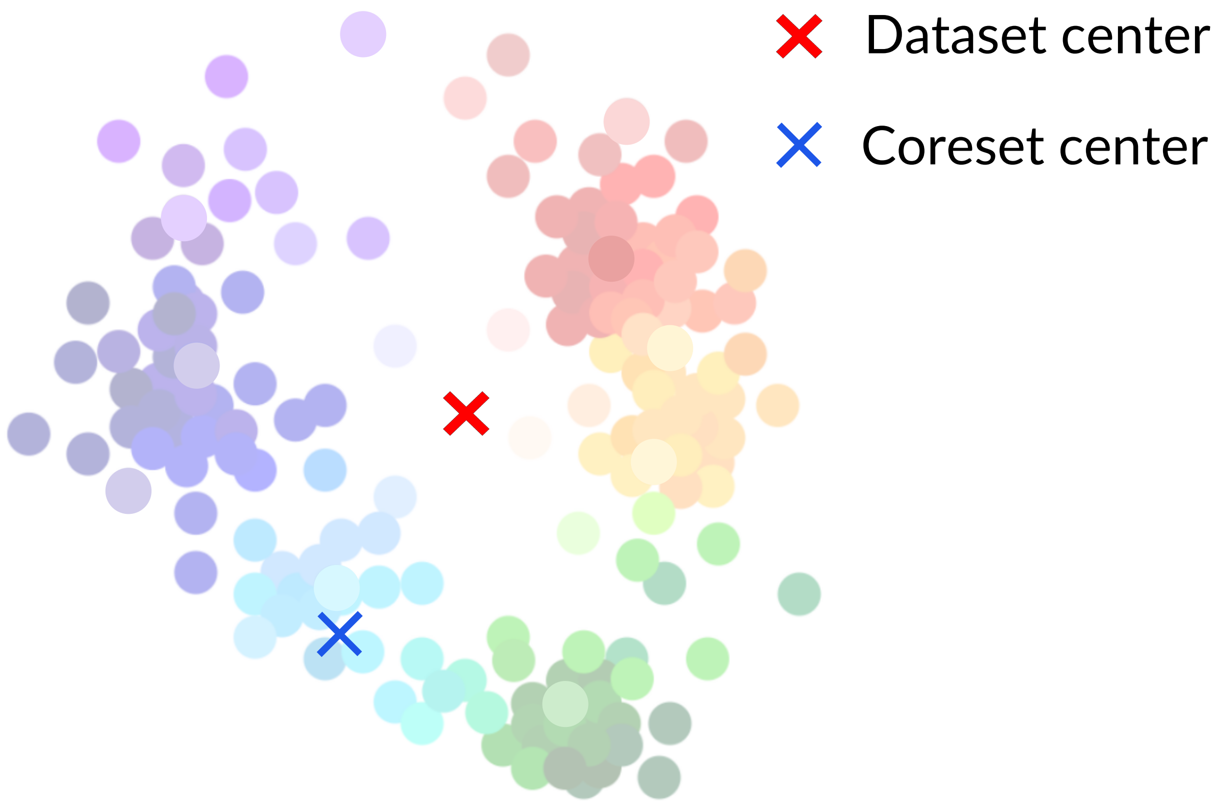
Max Welling. Herding dynamical weights to learn. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1121–1128, 2009.

Baseline

Coreset Selection



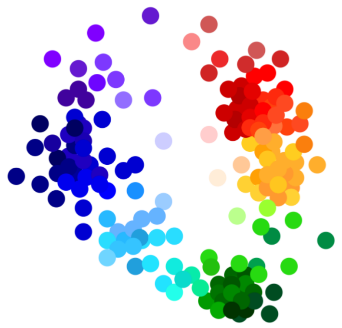
Herding



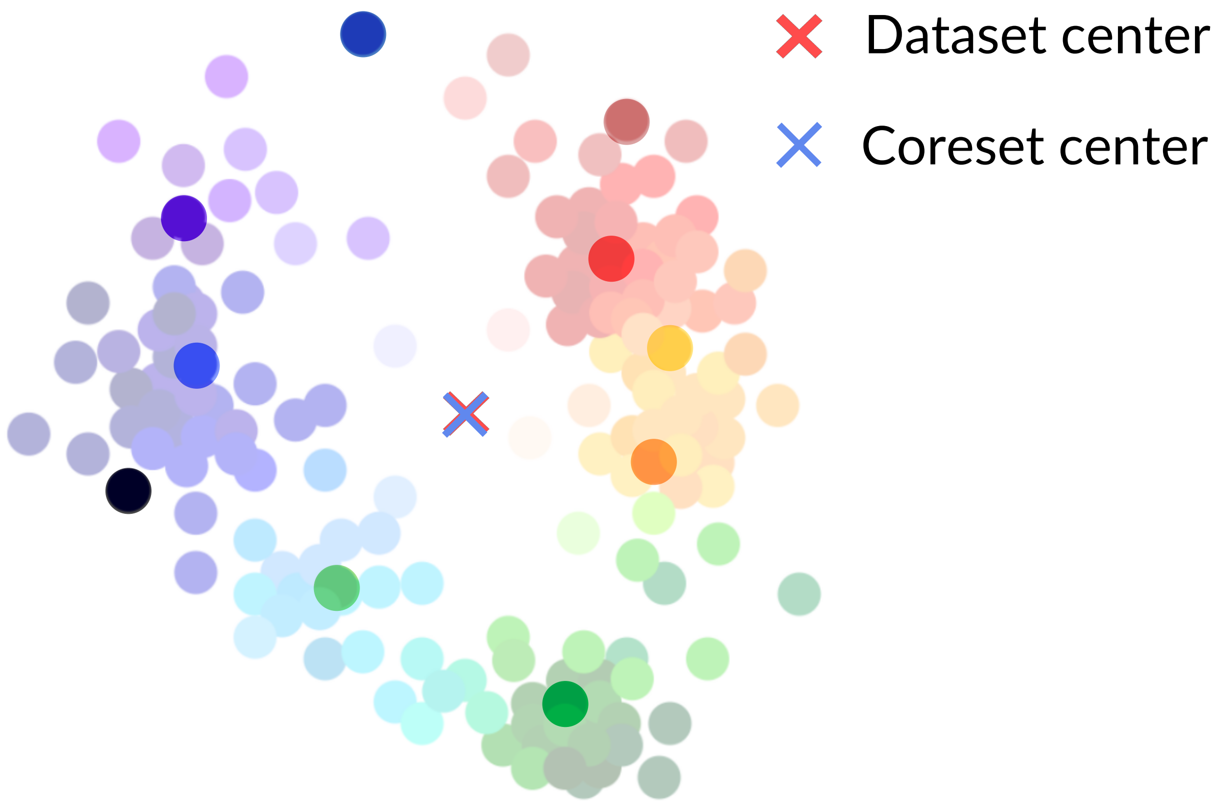
Max Welling. Herding dynamical weights to learn. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1121–1128, 2009.

Baseline

Coreset Selection



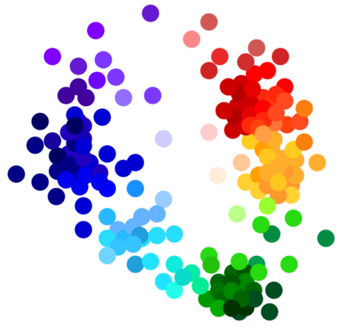
Herding



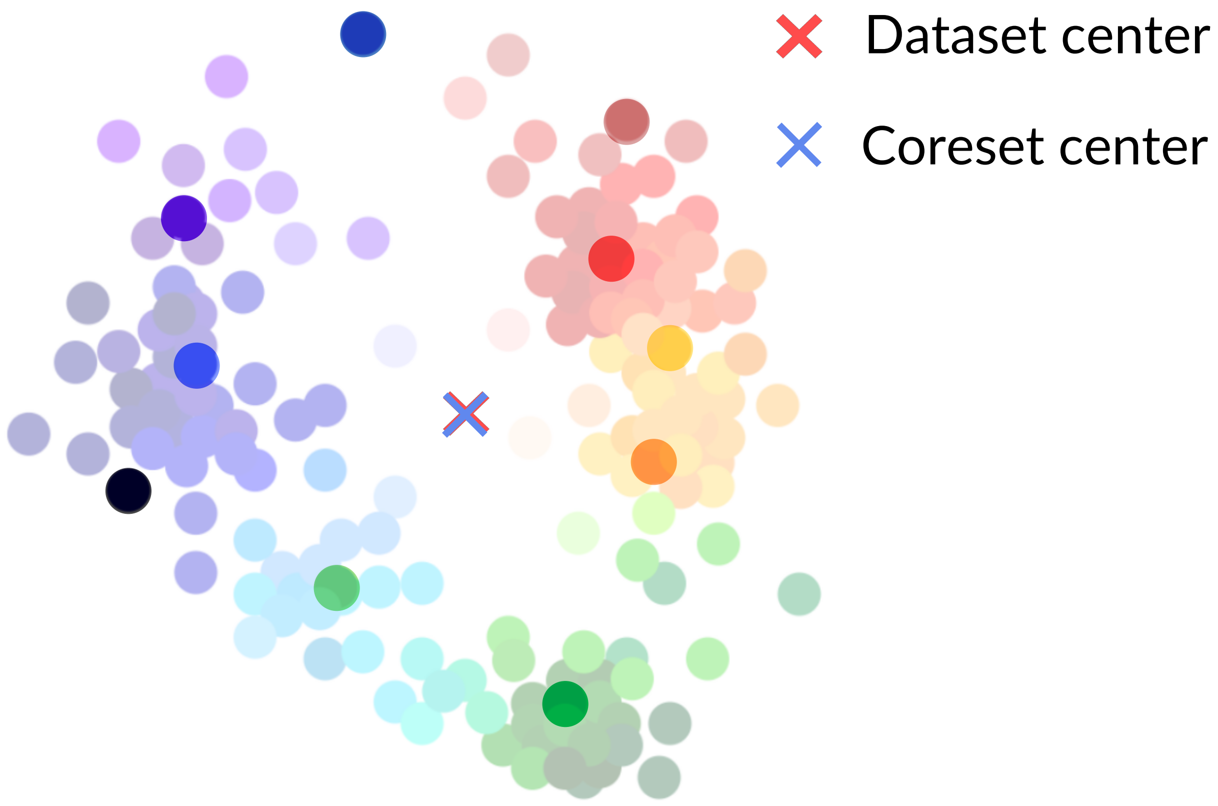
Max Welling. Herding dynamical weights to learn. In Proceedings of the 26th Annual International Conference on Machine Learning, pages 1121–1128, 2009.

Baseline

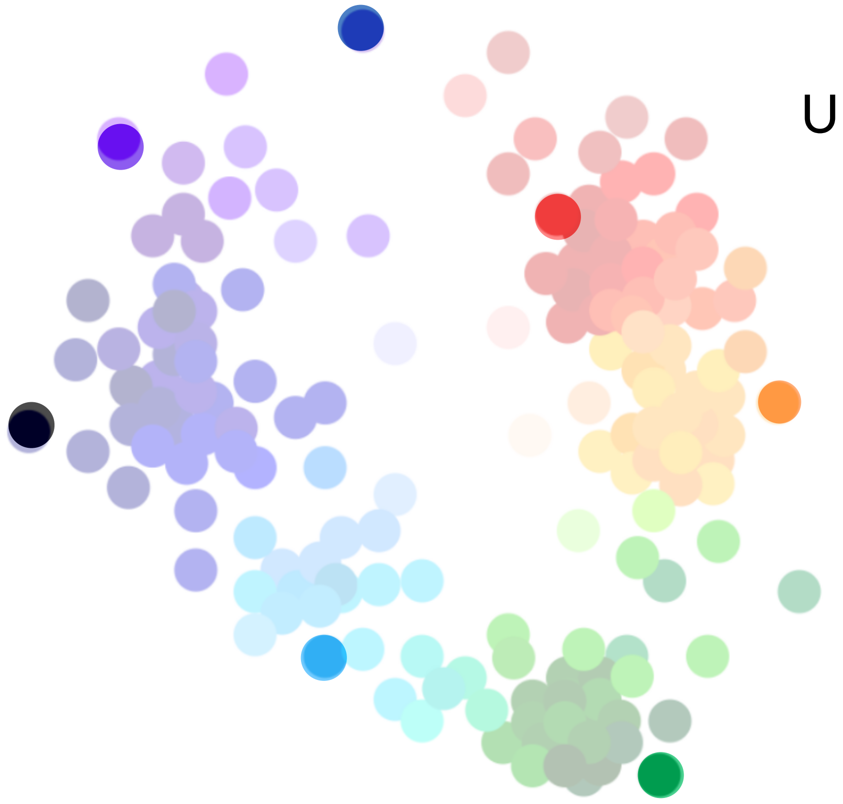
Coreset Selection



Herding



K-center

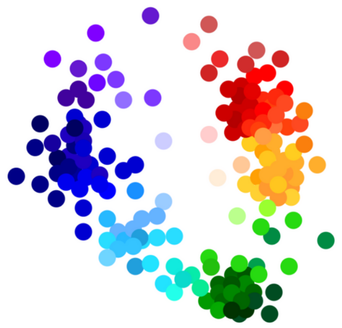


Until select K samples

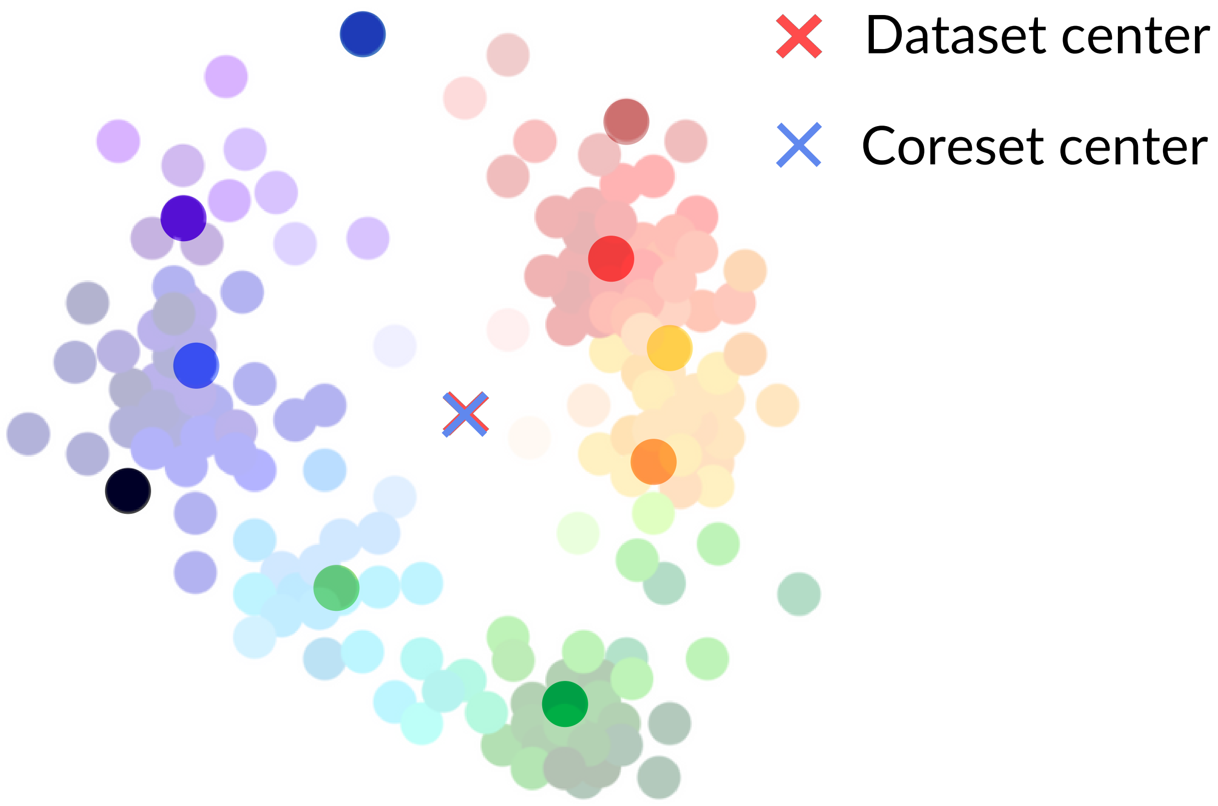
Sener, Ozan, and Silvio Savarese. "Active learning for convolutional neural networks: A core-set approach." arXiv preprint arXiv:1708.00489 (2017).
Reza Zanjirani Farahani and Masoud Hekmatfar. Facility location: concepts, models, algorithms and case studies. Springer Science & Business Media, 2009.

Baseline

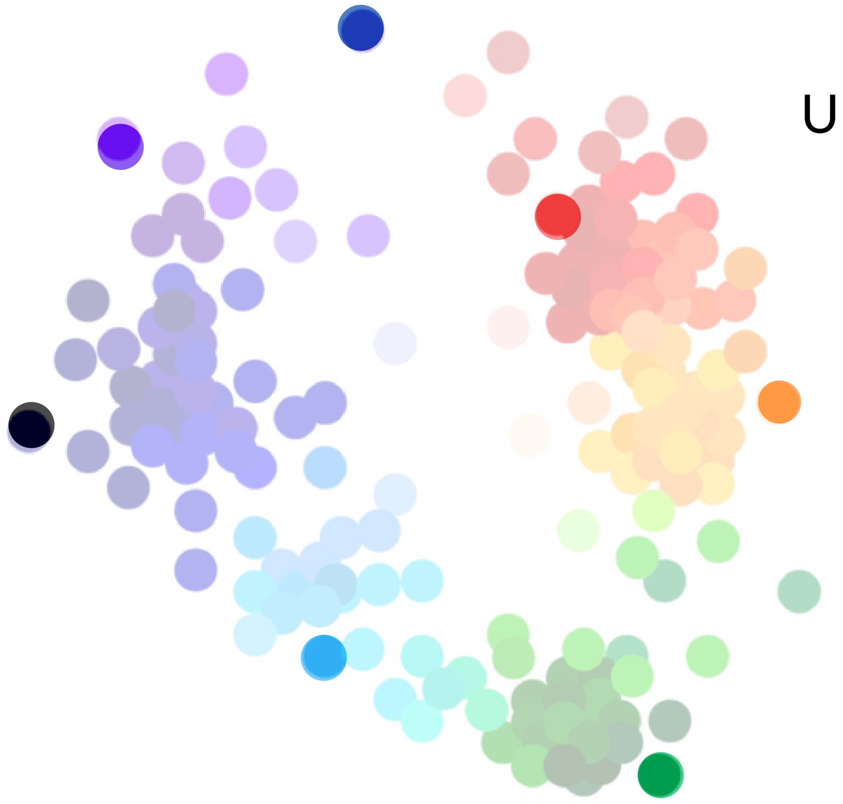
Coreset Selection



Herding



K-center



Until select K samples

Forgetting

Forgetting event:
When the model correctly predicts the example in one epoch but fails in the next.

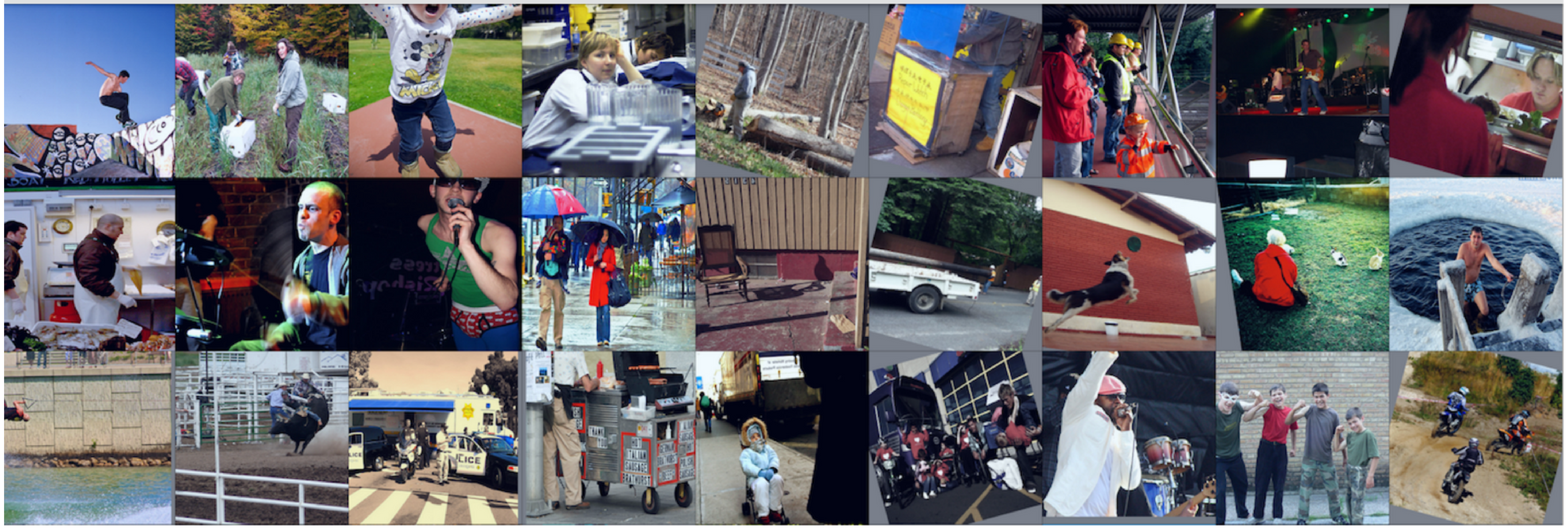
Track these forgetting events during training and identify the ones with the least forgetting events.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. arXiv preprint arXiv:1812.05159, 2018.

Our Task

Large Vision-Language Dataset

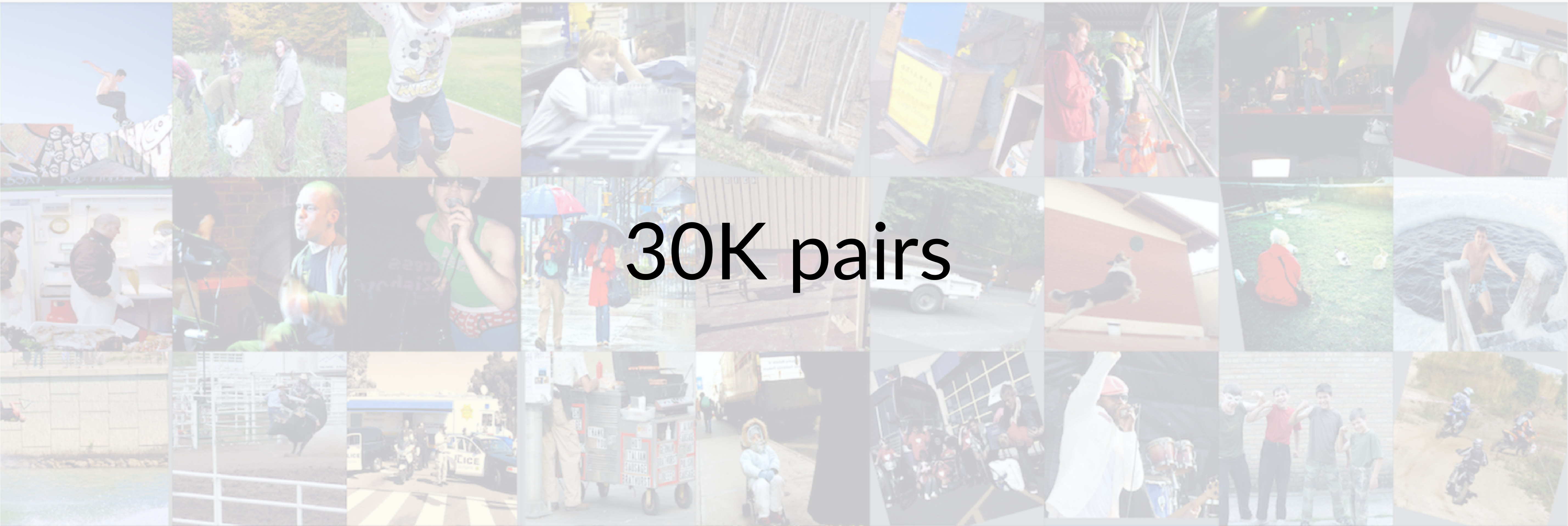
N pairs of **image + text**



Our Task

Large Vision-Language Dataset

N pairs of **image + text**



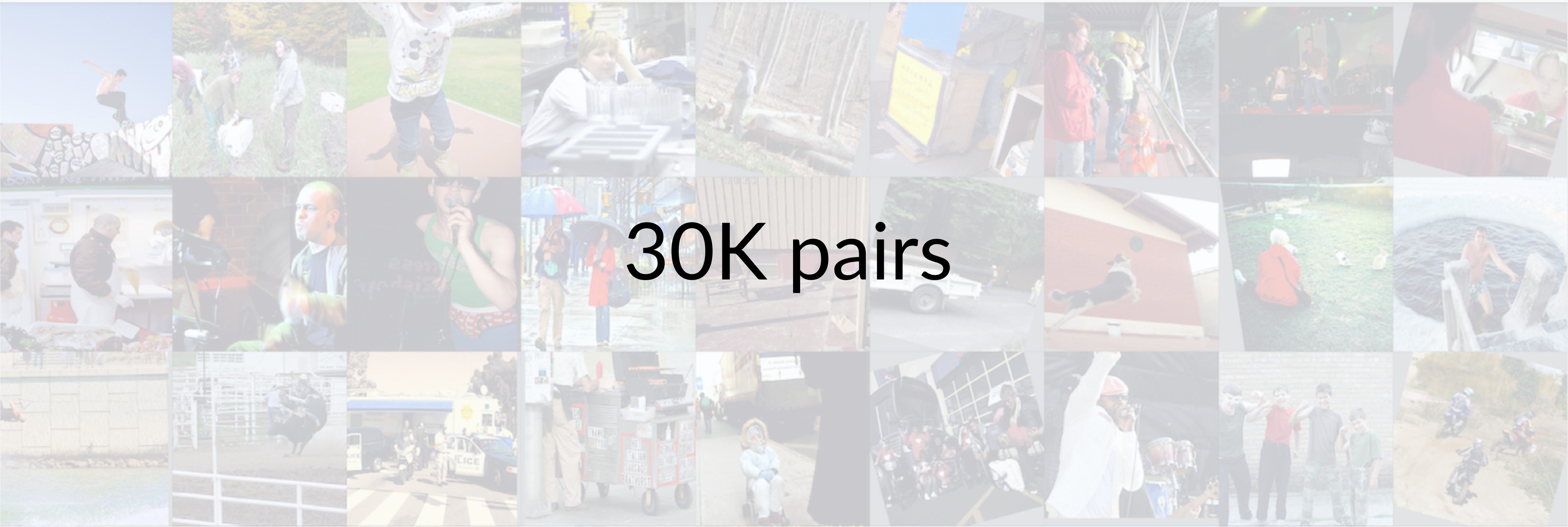
30K pairs

Flickr30K

Our Task

Large Vision-Language Dataset

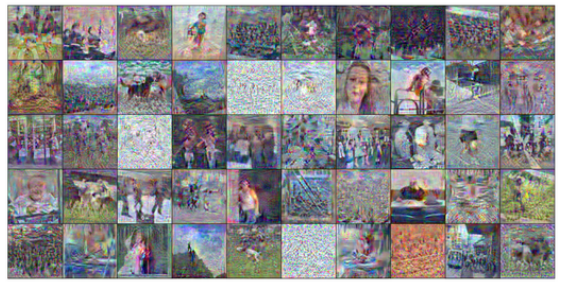
N pairs of image + text



Flickr30K

N >> M

Small Synthetic Dataset

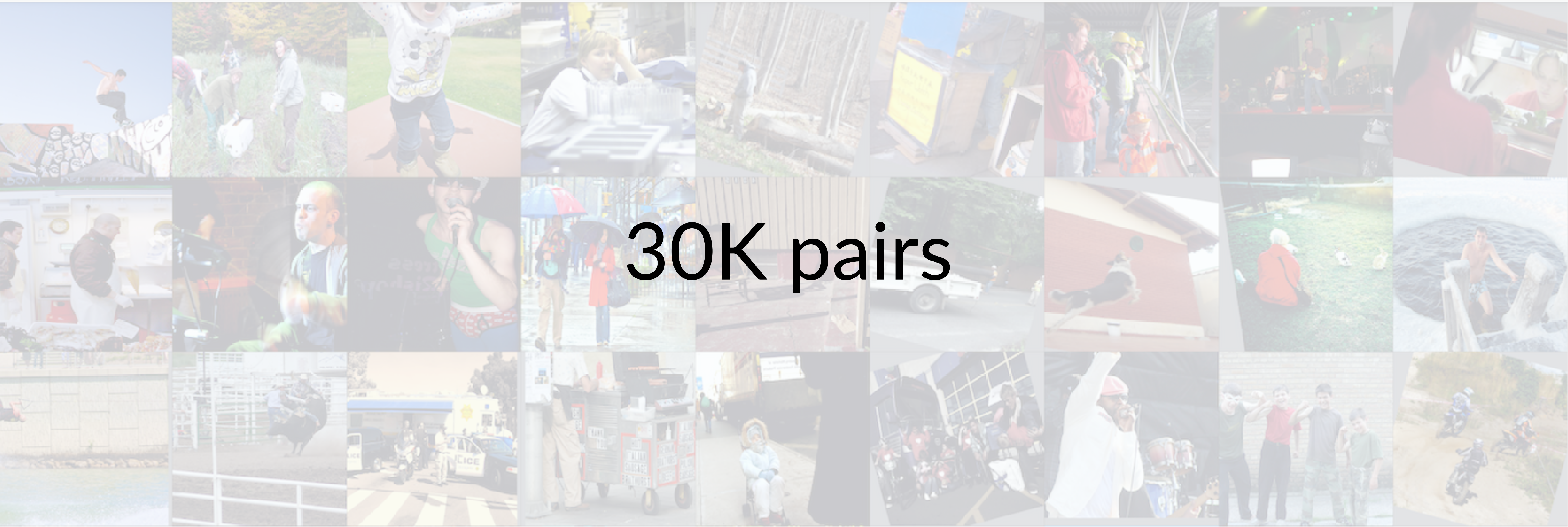


M pairs of image + text

Our Task

Large Vision-Language Dataset

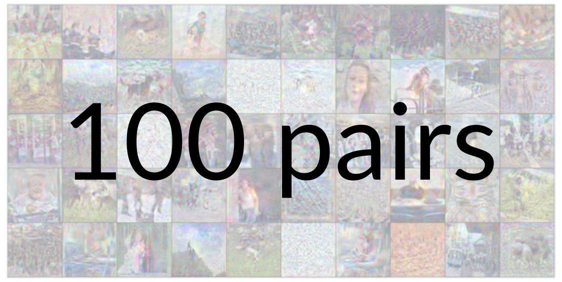
N pairs of image + text



Flickr30K

N >> M

Small Synthetic Dataset

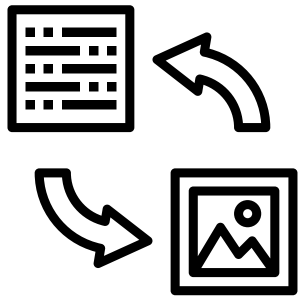


M pairs of image + text

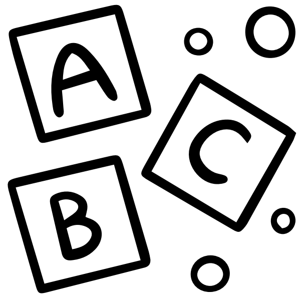
Challenges

Vision-Language Dataset Distillation

Complex Cross-Modal Relationships



Discrete text optimization issue

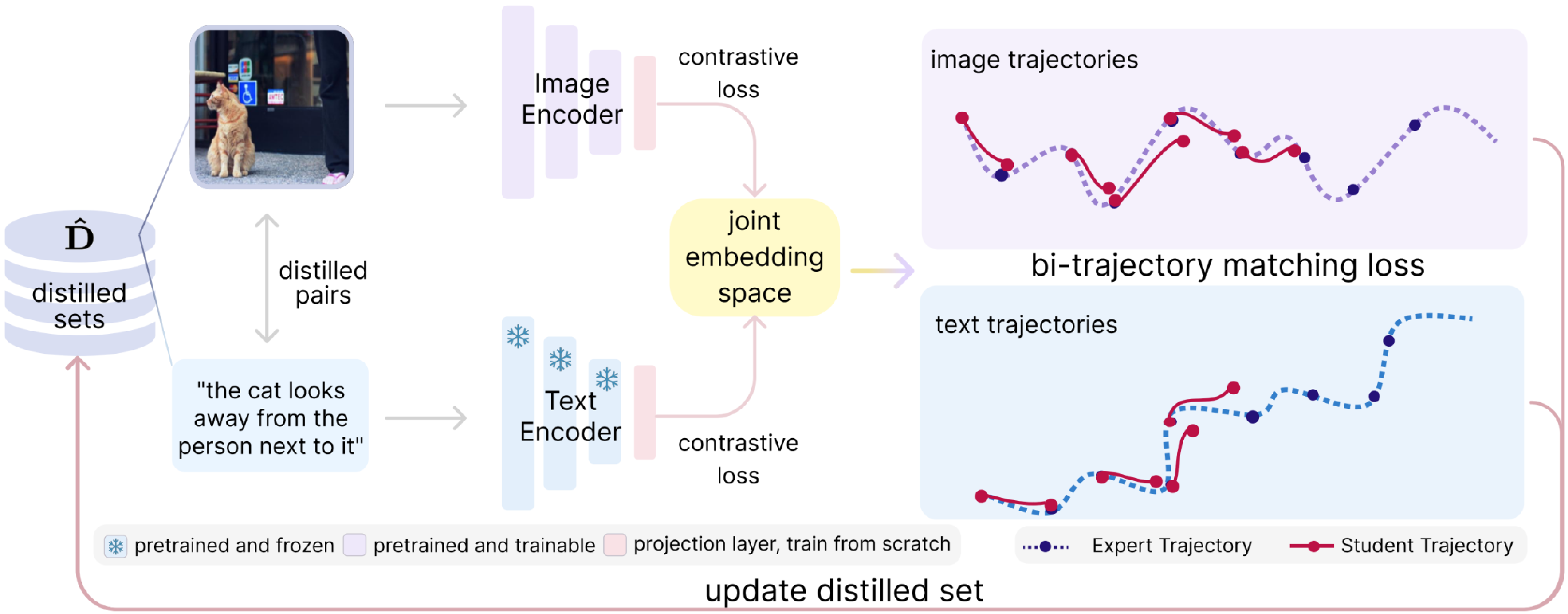


Heavy computational cost

| Method

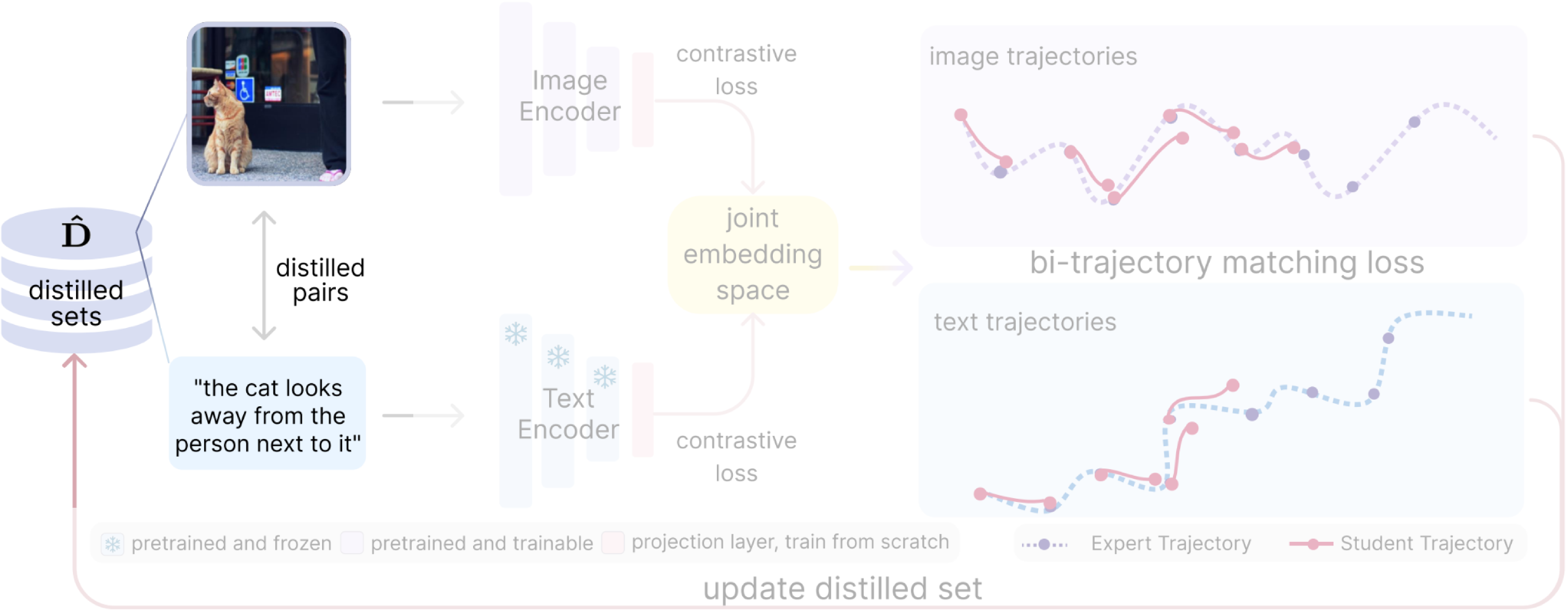
Method

Pipeline Overview



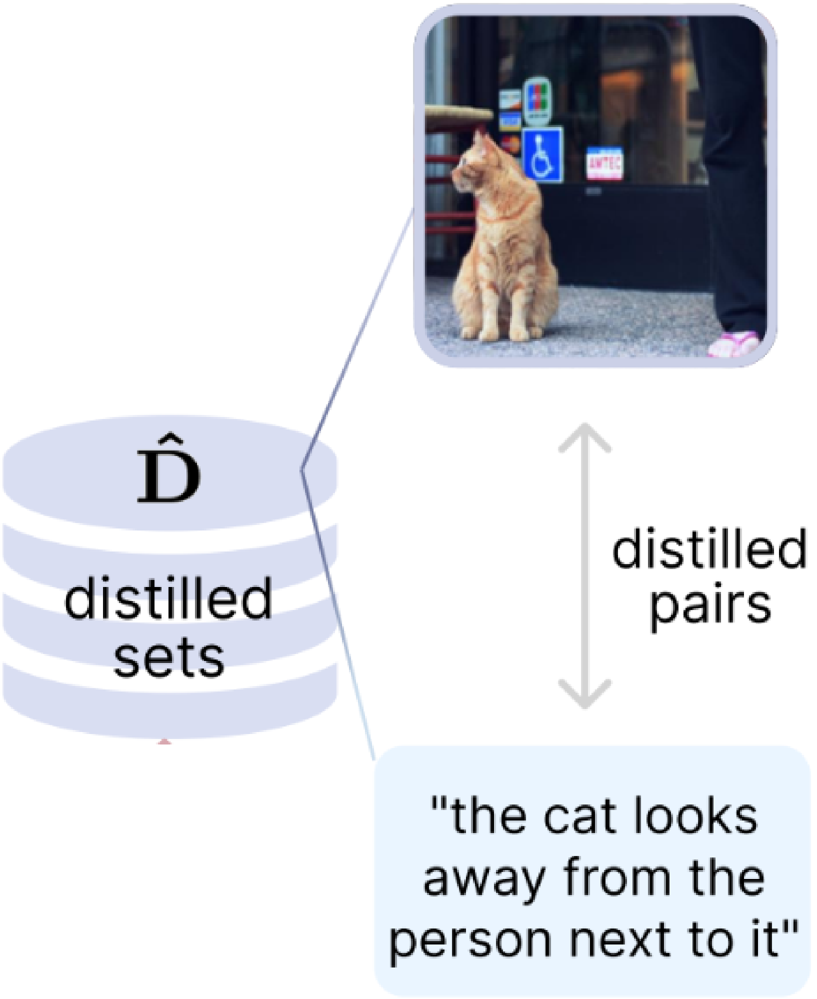
Method

Problem Formulation



Method

Problem Formulation

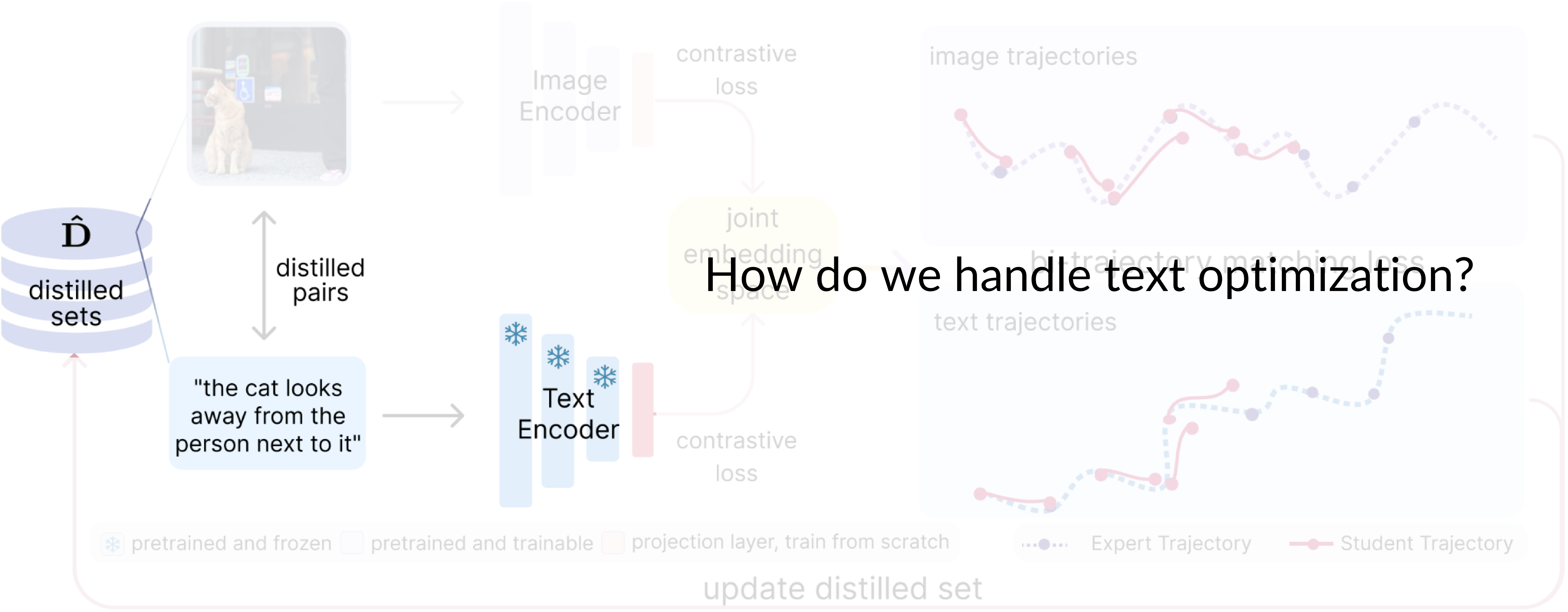


Large Scale Dataset $\mathbf{D} = \{(x_i, y_i)\}_{i=1}^{N}$

$M << N$

Small Synthetic Dataset $\hat{\mathbf{D}} = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{M}$

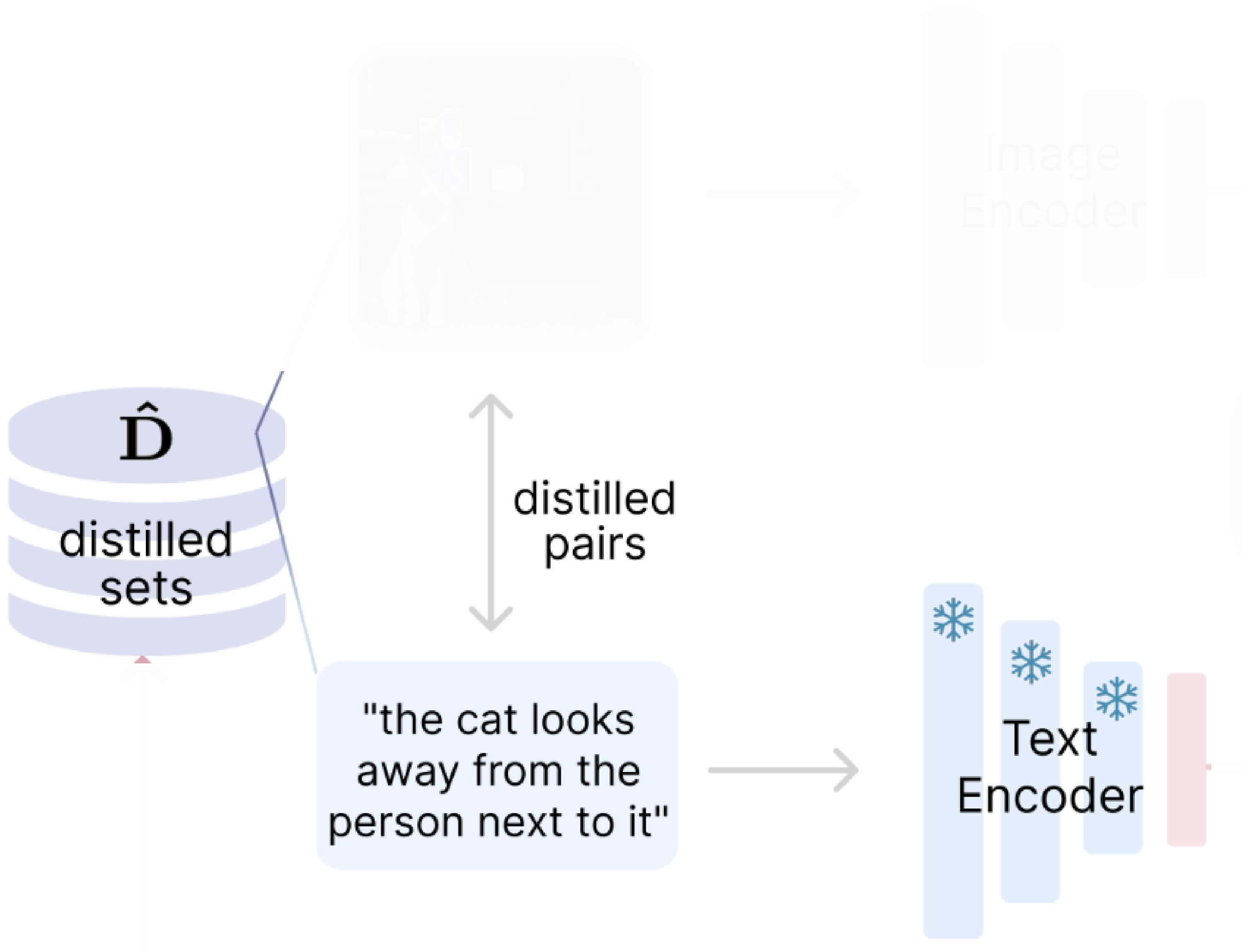
Method



How do we handle text optimization?

Method

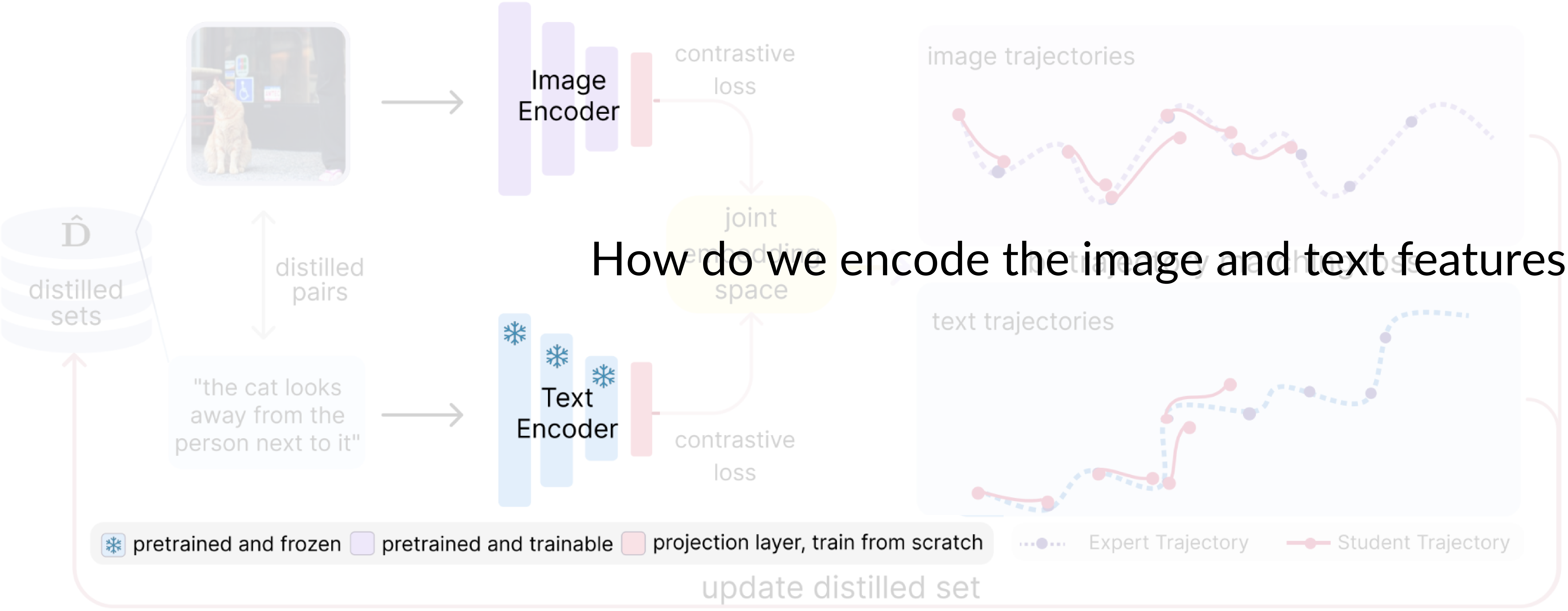
How do we handling text optimization?



We initialize the **continuous sentence embeddings** using a <u>pretrained BERT model</u> and update the distilled text in the **continuous embedding space.**

Method

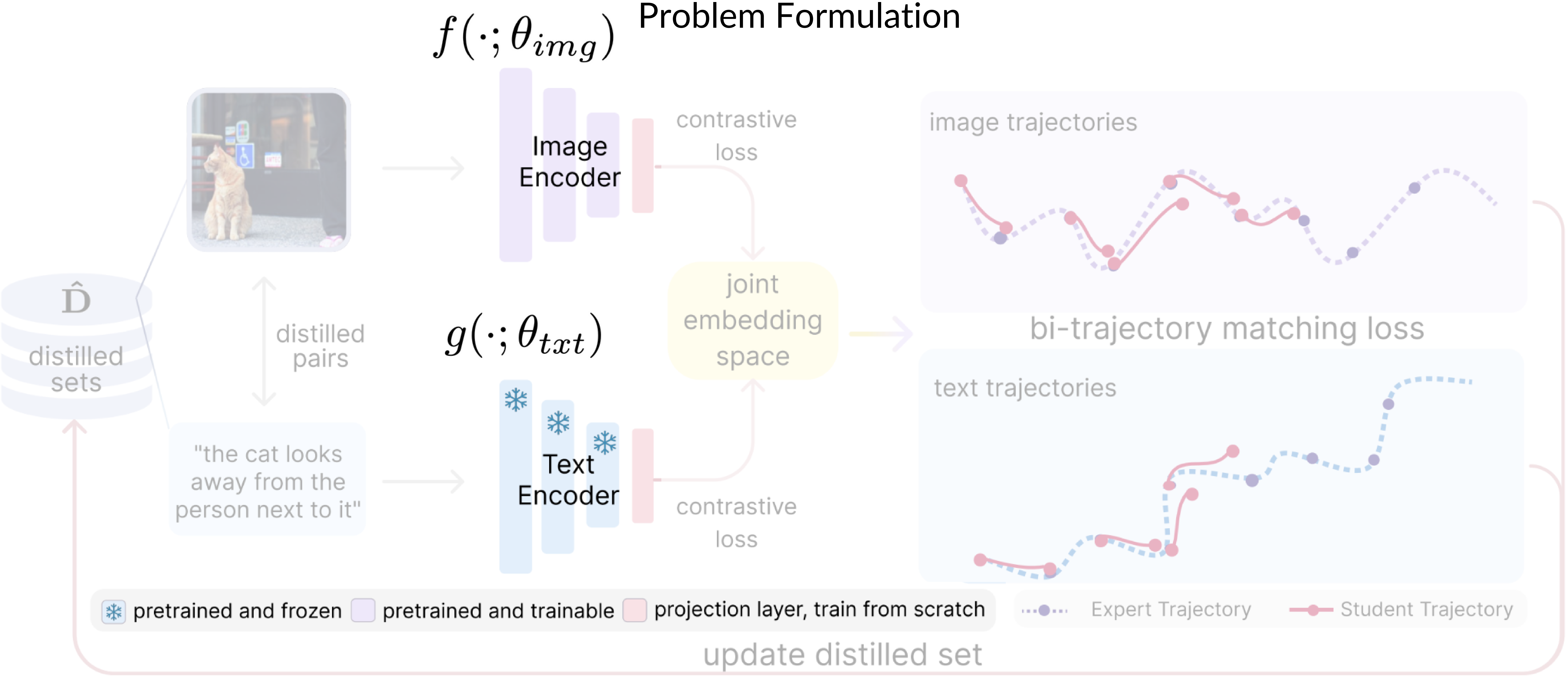
Problem Formulation



How do we encode the image and text features?

Method

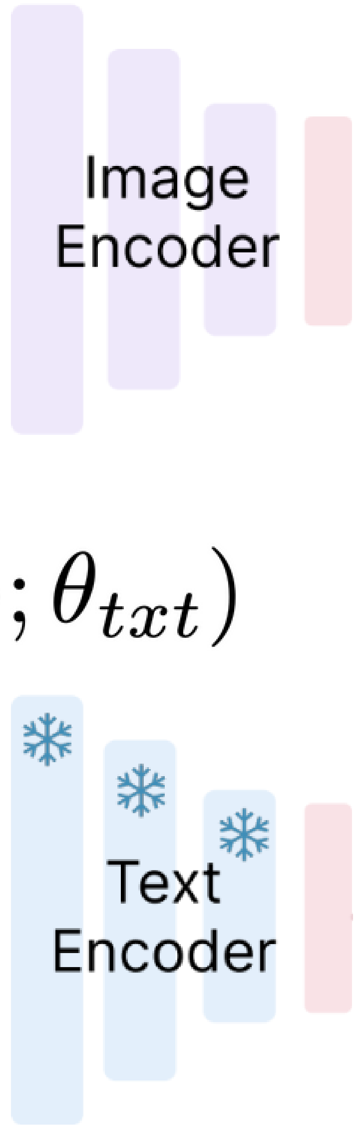
Problem Formulation



Method

Problem Formulation

$f(\cdot; \theta_{img})$



$g(\cdot; \theta_{txt})$

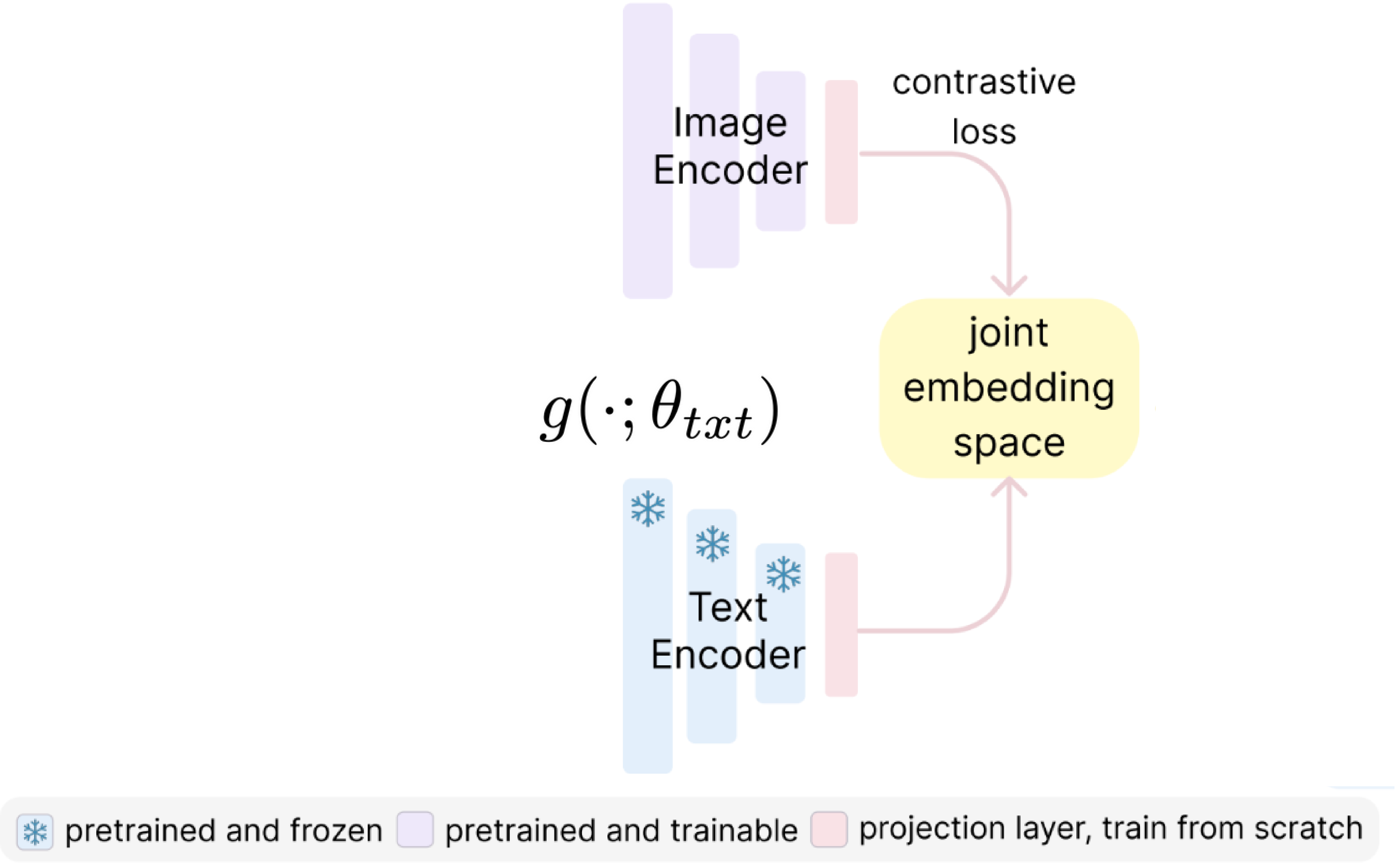
$$\theta^* \approx \arg\min_{\theta} \frac{1}{N} \sum_{i=1}^{N} \ell\left(f(x_i; \theta_{img}), g(y_i; \theta_{txt})\right)$$

pretrained and frozen | pretrained and trainable | projection layer, train from scratch

Method

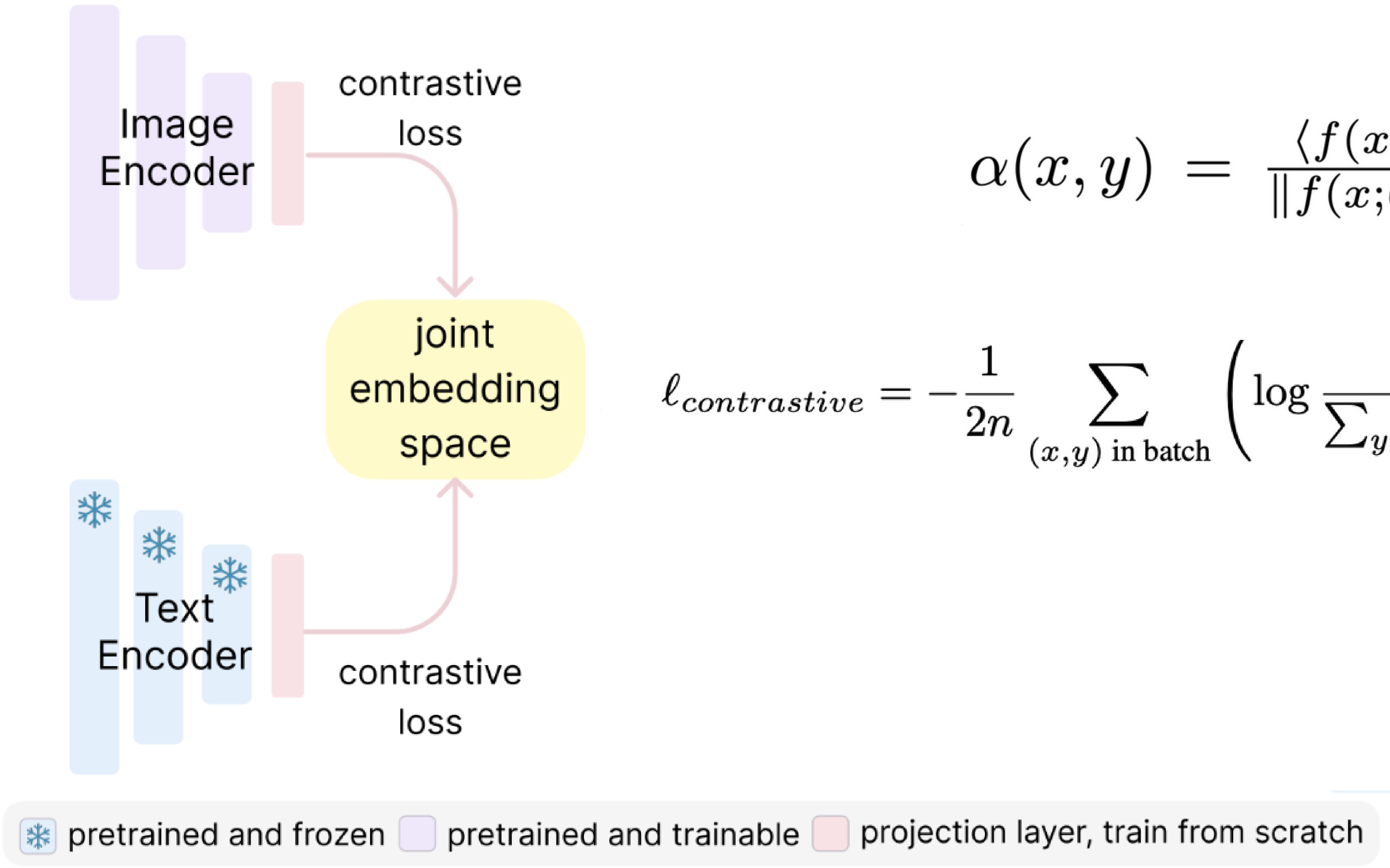
Problem Formulation

$f(\cdot;\theta_{img})$



Method

Problem Formulation



$$\alpha(x, y) = \frac{\langle f(x;\theta_{img}), g(y;\theta_{txt}) \rangle}{\|f(x;\theta_{img})\| \|g(y;\theta_{txt})\|}$$

$$\ell_{contrastive} = -\frac{1}{2n} \sum_{(x,y)\text{ in batch}} \left(\log \frac{\exp(\alpha(x,y))}{\sum_{y' \neq y} \exp(\alpha(x,y'))} + \log \frac{\exp(\alpha(x,y))}{\sum_{x' \neq x} \exp(\alpha(x',y))} \right)$$

Method

Bi-Trajectory Guided Vision-Language Co-Distillation

Concretely, our approach consists of two stages:

Stage 1 Expert training

Training multiple models for T epochs on the full dataset $\mathbf{D}$
Obtaining the expert training trajectories $\{\tau^*\}$, with each trajectory $\tau^* = \{\theta_t^*\}_{t=0}^{T}$

For our multimodal setting, the models are trained with contrastive loss.

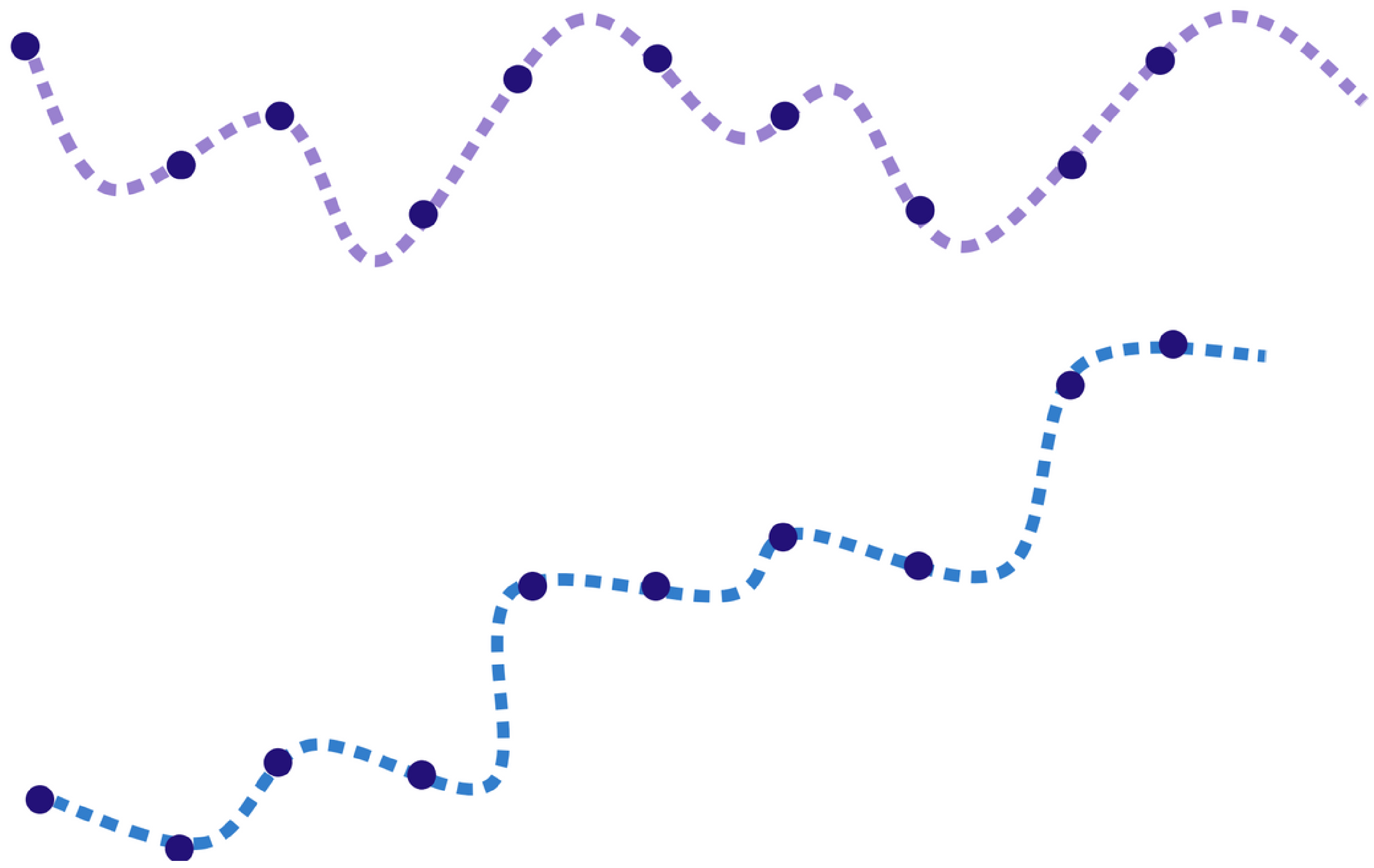
Method

Bi-Trajectory Guided Vision-Language Co-Distillation

Concretely, our approach consists of two stages:

Stage 1 Expert training

$$\tau^* = \{\theta_t^*\}_{t=0}^{T}$$



Method

Bi-Trajectory Guided Vision-Language Co-Distillation

Concretely, our approach consists of two stages:

Stage 2 Distillation

Training a set of student models on the current distilled dataset $\hat{\mathbf{D}} = \{(\hat{x}_j, \hat{y}_j)\}_{j=1}^{M}$ using the same contrastive loss.

Update $\hat{\mathbf{D}}$ based on the **bi-trajectory matching loss** of the student models' parameter trajectories and the expert trajectories τ^*.

Method

Bi-Trajectory Guided Vision-Language Co-Distillation

Bi-trajectory matching loss:

$$\ell_{trajectory} = \frac{\|\hat{\theta}_{img,s+\hat{R}} - \theta^*_{img,s+R}\|_2^2}{\|\theta^*_{img,s} - \theta^*_{img,s+R}\|_2^2} + \frac{\|\hat{\theta}_{txt,s+\hat{R}} - \theta^*_{txt,s+R}\|_2^2}{\|\theta^*_{txt,s} - \theta^*_{txt,s+R}\|_2^2}$$

Method

Bi-Trajectory Guided Vision-Language Co-Distillation

Bi-trajectory matching loss:

$$\tau^* = \{\theta_t^*\}_{t=0}^{T}$$

$$\ell_{trajectory} = \frac{\|\hat{\theta}_{img,s+\hat{R}} - \theta^*_{img,s+R}\|_2^2}{\|\theta^*_{img,s} - \theta^*_{img,s+R}\|_2^2} + \frac{\|\hat{\theta}_{txt,s+\hat{R}} - \theta^*_{txt,s+R}\|_2^2}{\|\theta^*_{txt,s} - \theta^*_{txt,s+R}\|_2^2}$$

Method

Bi-Trajectory Guided Vision-Language Co-Distillation

Bi-trajectory matching loss:

$$\tau^* = \{\theta_t^*\}_{t=0}^{T}$$

$$\ell_{trajectory} = \frac{\|\hat{\theta}_{img,s+\hat{R}} - \theta^*_{img,s+R}\|_2^2}{\|\theta^*_{img,s} - \theta^*_{img,s+R}\|_2^2} + \frac{\|\hat{\theta}_{txt,s+\hat{R}} - \theta^*_{txt,s+R}\|_2^2}{\|\theta^*_{txt,s} - \theta^*_{txt,s+R}\|_2^2}$$

Method

Bi-Trajectory Guided Vision-Language Co-Distillation

Bi-trajectory matching loss:

$$\tau^* = \{\theta_t^*\}_{t=0}^{T}$$

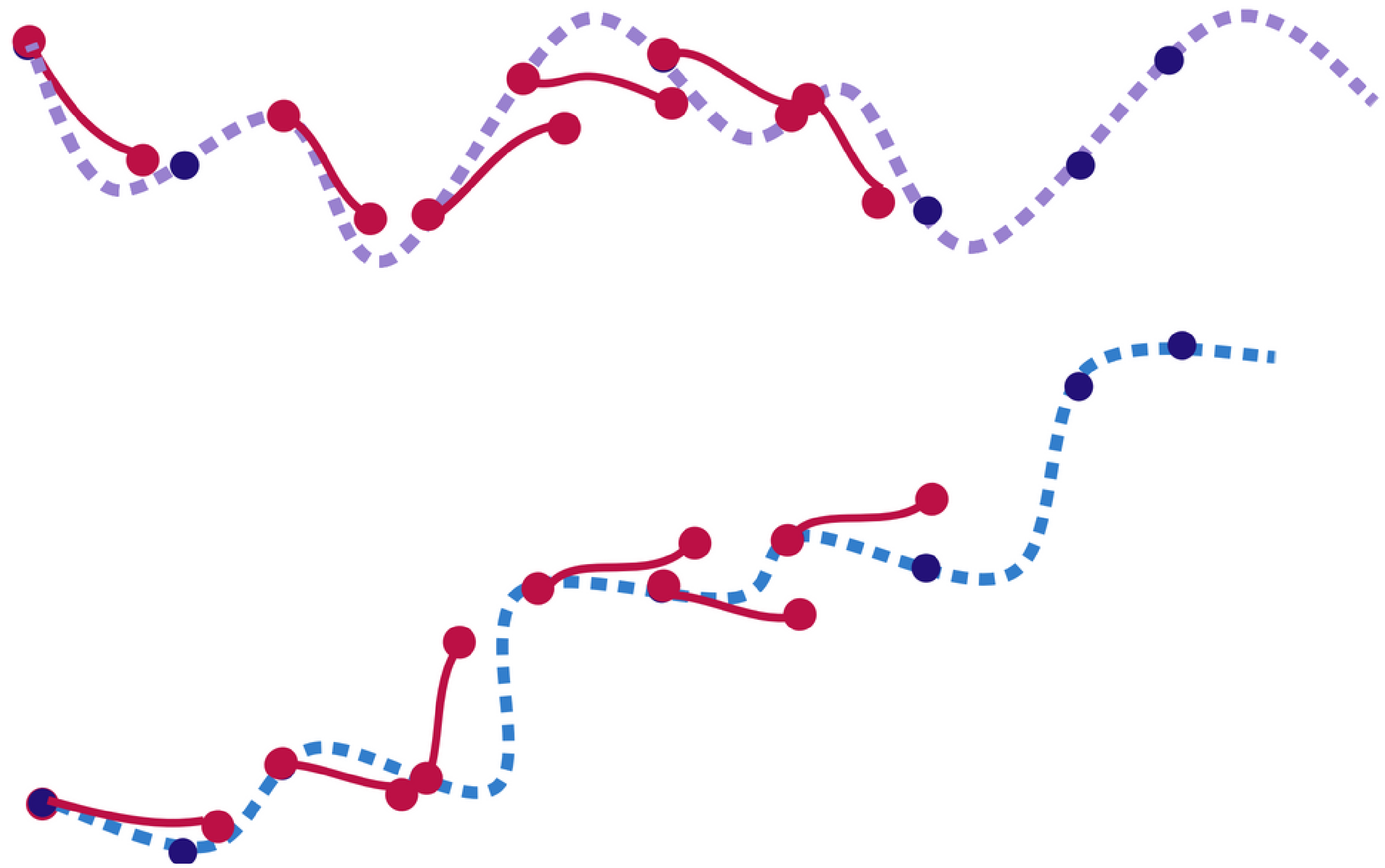
$$\ell_{trajectory} = \frac{\|\hat{\theta}_{img,s+\hat{R}} - \theta^*_{img,s+R}\|_2^2}{\|\theta^*_{img,s} - \theta^*_{img,s+R}\|_2^2} + \frac{\|\hat{\theta}_{txt,s+\hat{R}} - \theta^*_{txt,s+R}\|_2^2}{\|\theta^*_{txt,s} - \theta^*_{txt,s+R}\|_2^2}$$

Method

Bi-Trajectory Guided Vision-Language Co-Distillation

Concretely, our approach consists of two stages:

Stage 2 Distillation



Method

Low-Rank Adaptation Matching

Vision Transformers (ViTs)

1. High dimensionality of the embeddings
2. Large number of parameters

Method

Low-Rank Adaptation Matching

Consider a linear layer with
m input units, n output units
The weight matrix for this layer has dimensions m x n

$$Y = W X$$

Method

Low-Rank Adaptation Matching

Consider a linear layer with
m input units, n output units
The weight matrix for this layer has dimensions m x n

$$Y = W X$$

m = 800
n = 3200
W: 800 x 3200 = 2,560,000 weights.

Method

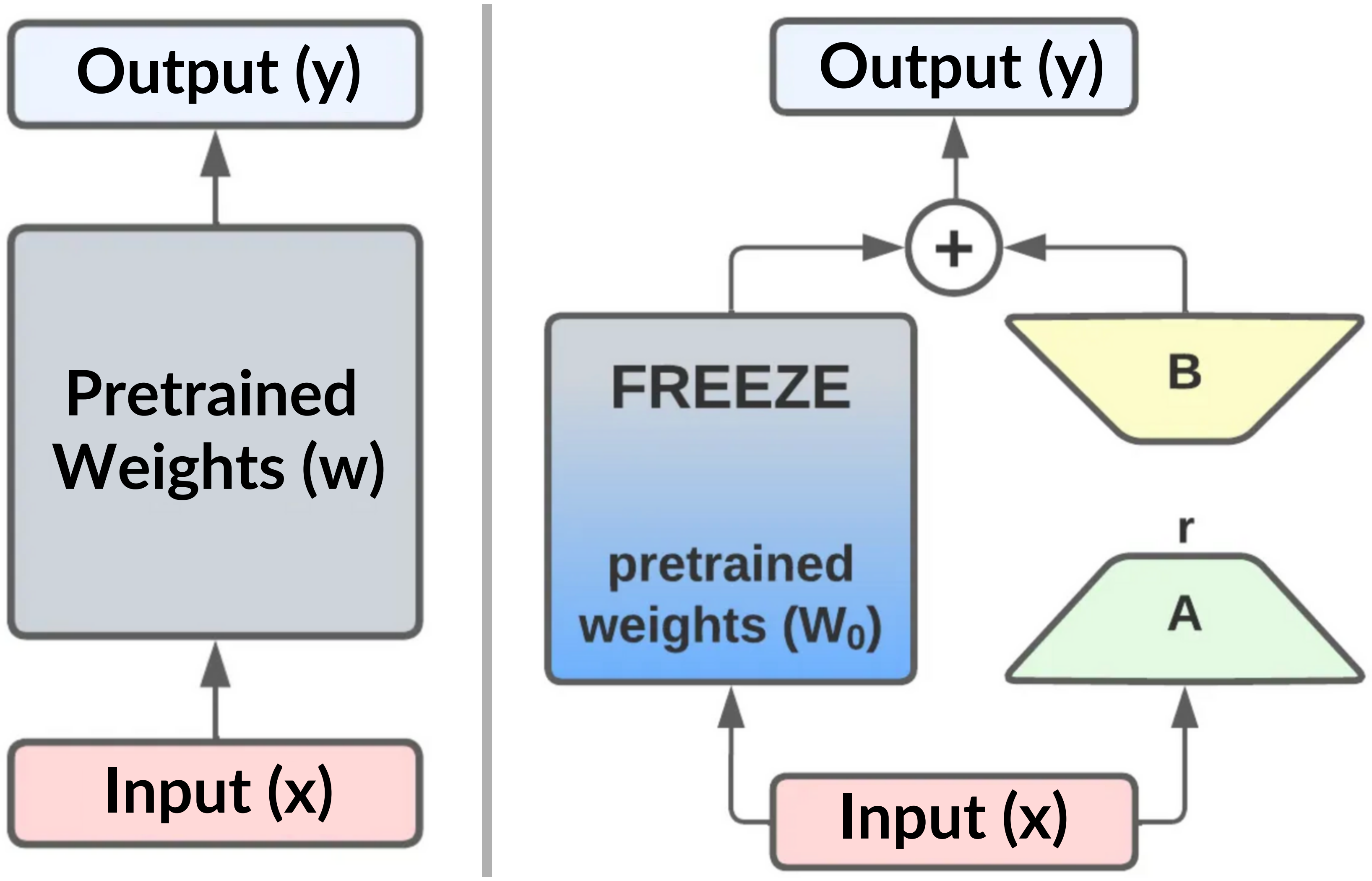
Low-Rank Adaptation Matching

Low-Rank Adaptation (LoRA)

We keep W fixed and introduce two matrices, A and B

$$Y = W X + A^{*}B X$$

Low-Rank Adaptation



Method

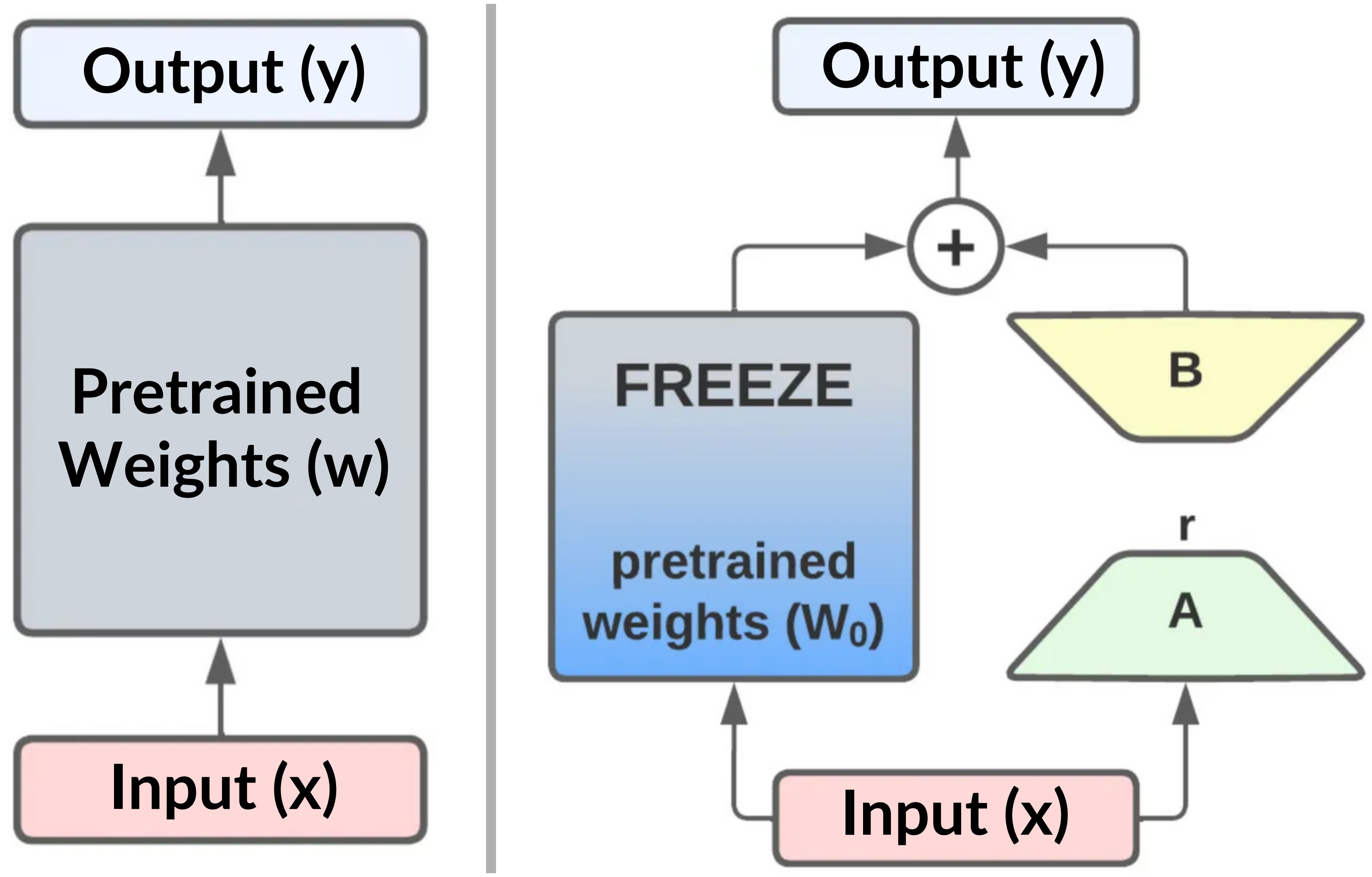
Low-Rank Adaptation Matching

Low-Rank Adaptation

Low-Rank Adaptation (LoRA)

We keep W fixed and introduce two matrices, A and B

$$Y = W X + A^*B X$$



Matrix A has a shape of 800 x r, and matrix B has a shape of r x 3200.

m = 800, n = 3200, r = 1

(800 x 1) + (1 x 3200) = 4000

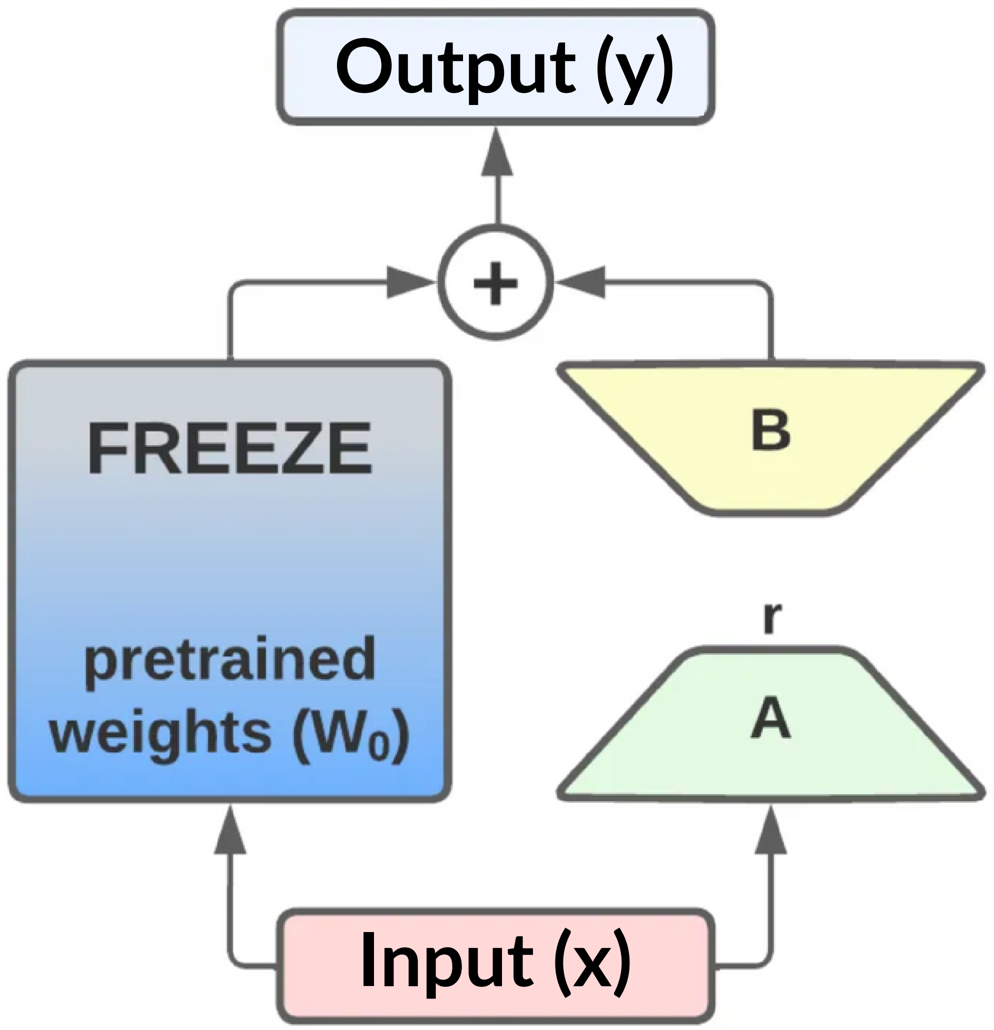
Method

Low-Rank Adaptation Matching

Low-Rank Adaptation (LoRA) Matching

LoRA matching optimizes the trajectories of low rank adapters instead of the full parameters.

Low-Rank Adaptation



Method

Low-Rank Adaptation Matching

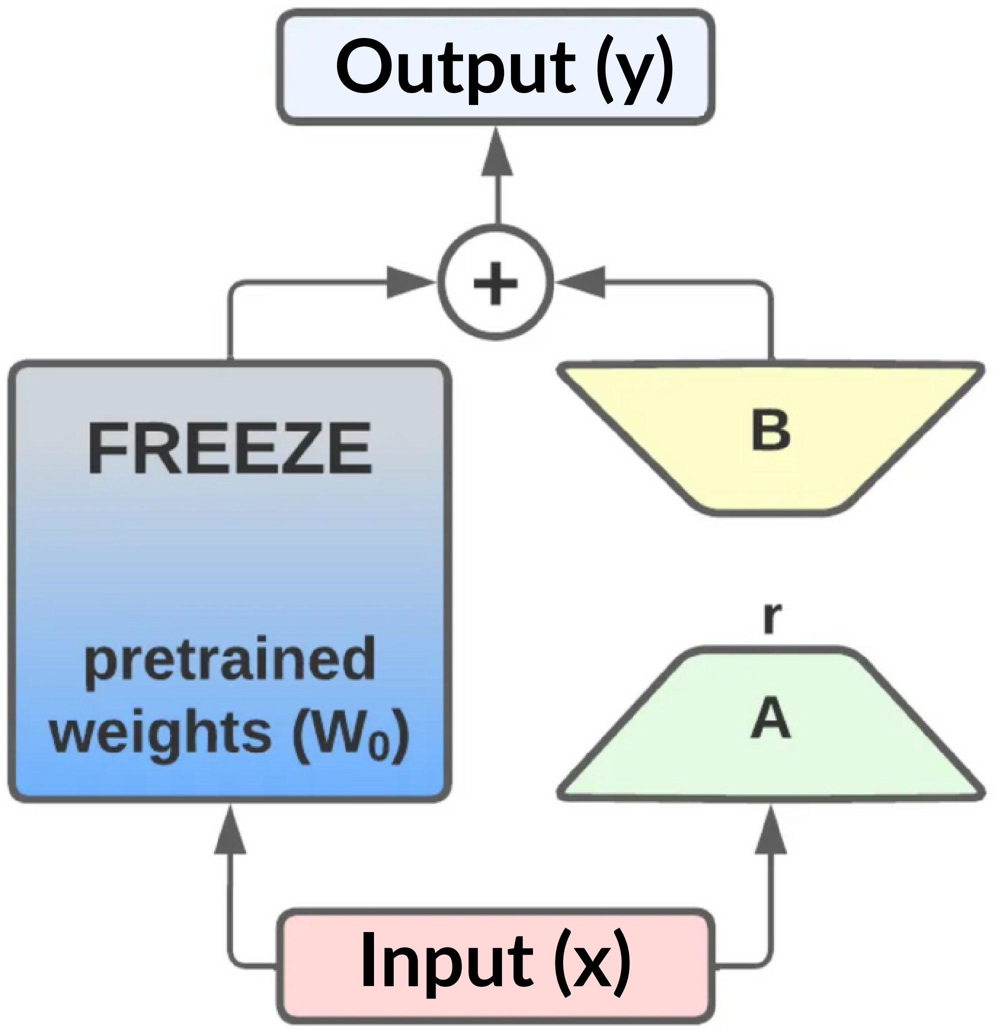
Low-Rank Adaptation

Low-Rank Adaptation (LoRA) Matching

LoRA matching optimizes the trajectories of low rank adapters instead of the full parameters.

VIT_base

86 million



Method

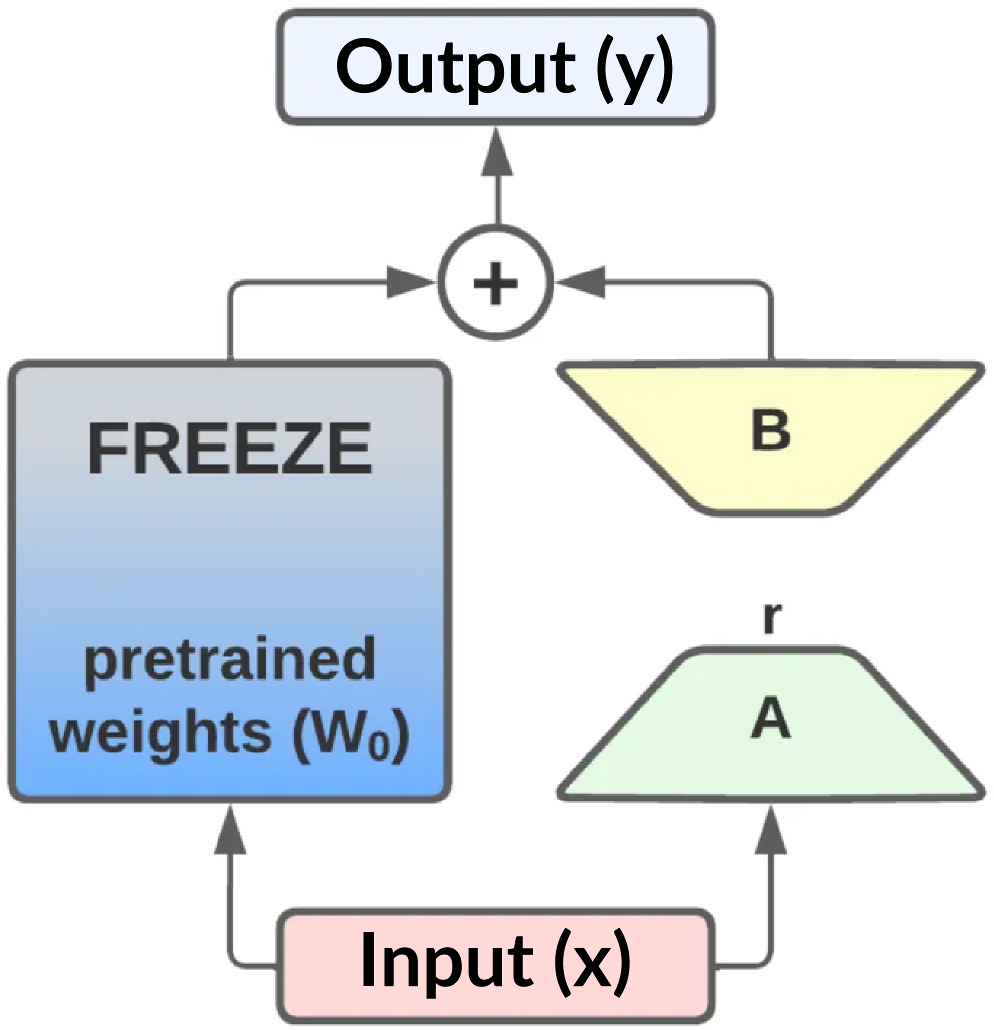
Low-Rank Adaptation Matching

Low-Rank Adaptation (LoRA) Matching

LoRA matching optimizes the trajectories of low rank adapters instead of the full parameters.

VIT_base	LoRA_VIT_base
86 million	18 million
	r=4

Low-Rank Adaptation



Method

Low-Rank Adaptation Matching

Low-Rank Adaptation (LoRA) Matching

LoRA matching optimizes the trajectories of low rank adapters instead of the full parameters.

Efficient model adaptation with minimal extra parameters.
Focus on optimizing a smaller parameter set.
Efficiently optimize trajectory loss during distillation.

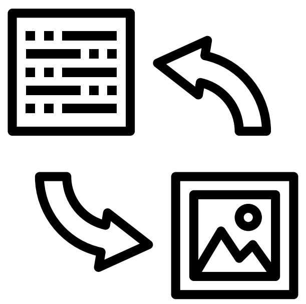
VIT_base	LoRA_VIT_base
86 million	18 million
	r=4

Low-Rank Adaptation

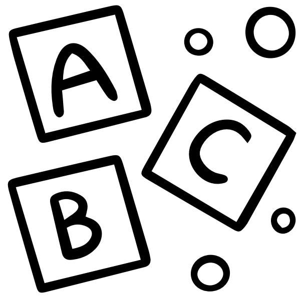
Challenges < > Solutions

Vision-Language Dataset Distillation

Complex Cross-Modal Relationships



Discrete text optimization issue

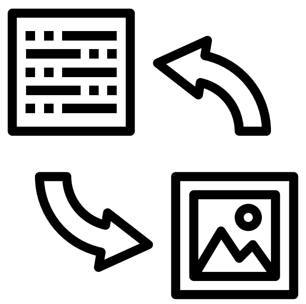


Heavy computational cost

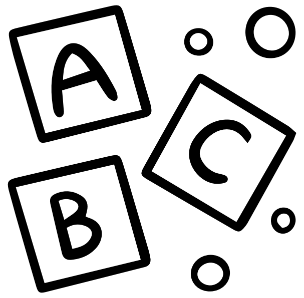
Challenges <> Solutions

Vision-Language Dataset Distillation

Complex Cross-Modal Relationships



Discrete text optimization issue



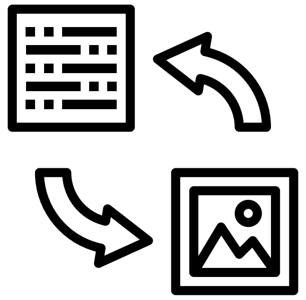
Heavy computational cost

Bi-trajectory Matching with Contrastive Loss

Challenges <> Solutions

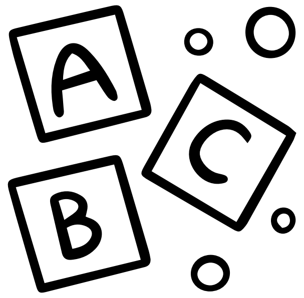
Vision-Language Dataset Distillation

Complex Cross-Modal Relationships



Bi-trajectory Matching with Contrastive Loss

Discrete text optimization issue

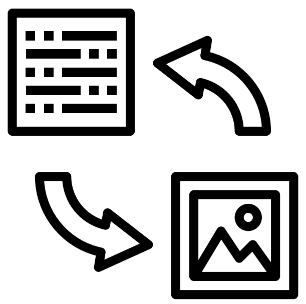
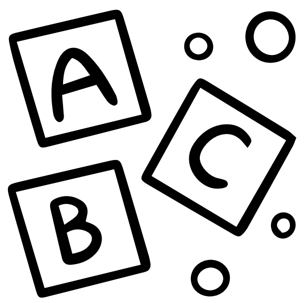


Continuous Sentence Embeddings

Heavy computational cost

Challenges <> Solutions

Vision-Language Dataset Distillation

Complex Cross-Modal Relationships	Discrete text optimization issue	Heavy computational cost
		
Bi-trajectory Matching with Contrastive Loss	Continuous Sentence Embeddings	Finetuning with Low-Rank Adaptation

| Experiments

Experiments

Evaluation Setting

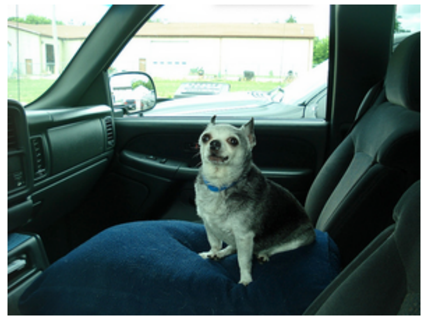
Dataset

Flickr30K

A small dog wearing a
sweater walking in the snow.
....

30K

COCO



a grey dog seated on
a chair of a vehicle
....

328K

Experiments

Evaluation Setting

Dataset

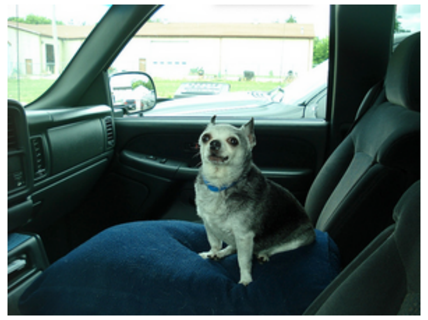
Flickr30K

A small dog wearing a sweater walking in the snow.

....

30K

COCO



a grey dog seated on a chair of a vehicle

....

328K

Task

Image-text retrieval

image-to-text retrieval (TR) &
text-to-image retrieval (IR)

R@K (with K = 1, 5, 10)
compute the fraction of times the correct result is found among the top K items.

Experiments

Evaluation Setting

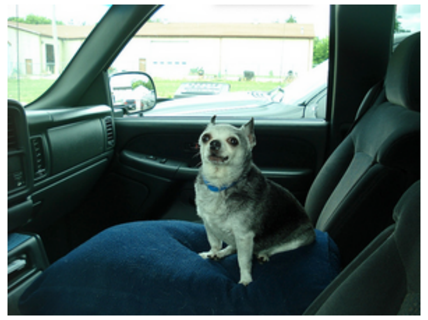
Dataset

Flickr30K

A small dog wearing a sweater walking in the snow.
....

30K

COCO



a grey dog seated on a chair of a vehicle
....

328K

Task

Image-text retrieval

image-to-text retrieval (TR) & text-to-image retrieval (IR)

R@K (with K = 1, 5, 10)
compute the fraction of times the correct result is found among the top K items.

Model

Image encoder:
- NFNet
- ViT

Language encoder:
- BERT

Experiments

Quantitative Results | Baseline Comparisons

			TR					IR				
			Coreset Selection					Coreset Selection				
Dataset	#pairs	ratio %	R	H	K	F	Dist (ours)	R	H	K	F	Dist (ours)
Flickr30K	100	0.34	1.3	1.1	0.6	1.2	**9.9** ± 0.3	1.0	0.7	0.7	0.7	**4.7** ± 0.2
	200	0.68	2.1	2.3	2.2	1.5	**10.2** ± 0.8	1.1	1.5	1.5	1.2	**4.6** ± 0.9
	500	1.67	5.2	5.1	4.9	3.6	**13.3** ± 0.6	2.4	3.0	3.5	1.8	**6.6** ± 0.3
	1000	3.45	5.2	5	5.6	3.1	**13.3** ± 1.0	3.8	4.1	4.4	3.2	**7.9** ± 0.8
COCO	100	0.08	0.8	0.8	1.4	0.7	**2.5** ± 0.3	0.3	0.5	0.4	0.3	**1.3** ± 0.1
	200	0.17	1.0	1.0	1.2	1.1	**3.3** ± 0.2	0.6	0.9	0.7	0.6	**1.7** ± 0.1
	500	0.44	1.9	1.9	2.5	2.1	**5.0** ± 0.4	1.1	1.7	1.1	0.8	**2.5** ± 0.5
	1000	0.88	1.9	2.4	2.4	1.9	**6.8** ± 0.4	1.5	1.3	1.5	0.7	**3.3** ± 0.1

Random selection of training examples (**R**) | Herding (**H**) K-center (**K**) | Forgetting (**F**)

Experiments

Quantitative Results | Baseline Comparisons

			TR					IR				
			Coreset Selection					Coreset Selection				
Dataset	#pairs	ratio %	R	H	K	F	Dist (ours)	R	H	K	F	Dist (ours)
Flickr30K	100	0.34	1.3	1.1	0.6	1.2	**9.9** ± 0.3	1.0	0.7	0.7	0.7	**4.7** ± 0.2
	200	0.68	2.1	2.3	2.2	1.5	**10.2** ± 0.8	1.1	1.5	1.5	1.2	**4.6** ± 0.9
	500	1.67	5.2	5.1	4.9	3.6	**13.3** ± 0.6	2.4	3.0	3.5	1.8	**6.6** ± 0.3
	1000	3.45	5.2	5	5.6	3.1	**13.3** ± 1.0	3.8	4.1	4.4	3.2	**7.9** ± 0.8
COCO	100	0.08	0.8	0.8	1.4	0.7	**2.5** ± 0.3	0.3	0.5	0.4	0.3	**1.3** ± 0.1
	200	0.17	1.0	1.0	1.2	1.1	**3.3** ± 0.2	0.6	0.9	0.7	0.6	**1.7** ± 0.1
	500	0.44	1.9	1.9	2.5	2.1	**5.0** ± 0.4	1.1	1.7	1.1	0.8	**2.5** ± 0.5
	1000	0.88	1.9	2.4	2.4	1.9	**6.8** ± 0.4	1.5	1.3	1.5	0.7	**3.3** ± 0.1

Random selection of training examples (**R**) | Herding (**H**) K-center (**K**) | Forgetting (**F**)

Experiments

Quantitative Results | Baseline Comparisons

Dataset	#pairs	ratio %	TR Coreset Selection R	TR Coreset Selection H	TR Coreset Selection K	TR Coreset Selection F	TR Dist (ours)	IR Coreset Selection R	IR Coreset Selection H	IR Coreset Selection K	IR Coreset Selection F	IR Dist (ours)
Flickr30K	100	0.34	1.3	1.1	0.6	1.2	**9.9** $\pm$ 0.3	1.0	0.7	0.7	0.7	**4.7** $\pm$ 0.2
	200	0.68	2.1	2.3	2.2	1.5	**10.2** $\pm$ 0.8	1.1	1.5	1.5	1.2	**4.6** $\pm$ 0.9
	500	1.67	5.2	5.1	4.9	3.6	**13.3** $\pm$ 0.6	2.4	3.0	3.5	1.8	**6.6** $\pm$ 0.3
	1000	3.45	5.2	5	5.6	3.1	**13.3** $\pm$ 1.0	3.8	4.1	4.4	3.2	**7.9** $\pm$ 0.8
COCO	100	0.08	0.8	0.8	1.4	0.7	**2.5** $\pm$ 0.3	0.3	0.5	0.4	0.3	**1.3** $\pm$ 0.1
	200	0.17	1.0	1.0	1.2	1.1	**3.3** $\pm$ 0.2	0.6	0.9	0.7	0.6	**1.7** $\pm$ 0.1
	500	0.44	1.9	1.9	2.5	2.1	**5.0** $\pm$ 0.4	1.1	1.7	1.1	0.8	**2.5** $\pm$ 0.5
	1000	0.88	1.9	2.4	2.4	1.9	**6.8** $\pm$ 0.4	1.5	1.3	1.5	0.7	**3.3** $\pm$ 0.1

Random selection of training examples (**R**) | Herding (**H**) K-center (**K**) | Forgetting (**F**)

Experiments

Qualitative Results | Distilled Examples

a man in a black wet suit is surfing a huge wave in the beautiful blue water

before the distillation

a man surfs over a huge wave in the ocean

after 2000 distillation steps

Note that the texts visualized here are <u>nearest sentence decodings</u> in the training set corresponding to the distilled text embeddings.

Experiments

Qualitative Results | Before and After Distillation

a newly married couple sharing a kiss in front of a convertible

a couple kissed in front of a beautiful three-tiered cake with blue ribbon and pink accents

a man in a black wet suite is surfing a huge wave in the beautiful blue water

a man surfs over a huge wave in the ocean

a woman in a toboggan sledding with a child that is in a toboggan also

skateboarder in jeans and t-shirt performing jump

man is sitting in a swing on a carnival ride

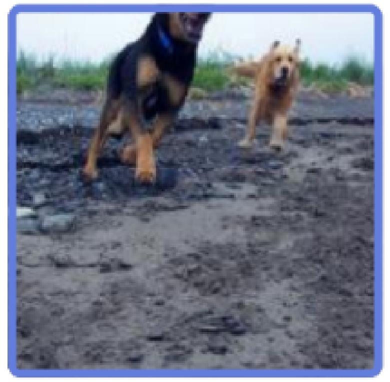
a little boy in a toy room wearing batman pajamas is building something with toys

a little boy in a white shirt is rock climbing

a boy with a blue helmet and gray pants is rock-climbing

two boys are watching another boy perform a jump on his bmx bike

kid in hoodie jumps a ramp



two dogs run through mud

the brown and white dog nips at the yellow dog

two men are competing for the ball in a game of soccer

four football players look on from the sideline while two teams are in formation at the line of scrimmage

Experiments

Qualitative Results | Before and After Distillation

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

(Here we only reports R@1, more details are in the paper.)

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

(Here we only reports R@1, more details are in the paper.)

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

Different vis./lan. encoders

Language Model	TR	IR
BERT	9.9	4.7
CLIP	**31.4**	**17.1**

Vision Model	TR	IR
NFNet	9.9	4.7
VIT_LoRA	**10.4**	**5.4**
NF_ResNet50	6.5	3.46
NF_RegNet	7.8	3.28

(Here we only reports R@1, more details are in the paper.)

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

Different vis./lan. encoders

Language Model	TR	IR
BERT	9.9	4.7
CLIP	**31.4**	**17.1**

Vision Model	TR	IR
NFNet	9.9	4.7
VIT_LoRA	**10.4**	**5.4**
NF_ResNet50	6.5	3.46
NF_RegNet	7.8	3.28

(Here we only reports R@1, more details are in the paper.)

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

Different vis./lan. encoders

Language Model	TR	IR
BERT	9.9	4.7
CLIP	**31.4**	**17.1**

Vision Model	TR	IR
NFNet	9.9	4.7
VIT_LoRA	**10.4**	**5.4**
NF_ResNet50	6.5	3.46
NF_RegNet	7.8	3.28

Single-modality vs. multi-modality

Takeaway:

- Distillation would be impossible if we solely optimize one modality

	TR	IR
T	1.3	0.5
I	3.5	1.6
Ours	9.9	4.7

T: text-only, I: image-only

(Here we only reports R@1, more details are in the paper.)

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

Different vis./lan. encoders

Language Model	TR	IR
BERT	9.9	4.7
CLIP	**31.4**	**17.1**

Vision Model	TR	IR
NFNet	9.9	4.7
VIT_LoRA	**10.4**	**5.4**
NF_ResNet50	6.5	3.46
NF_RegNet	7.8	3.28

Single-modality vs. multi-modality

<u>Takeaway</u>:

- Distillation would be impossible if we solely optimize one modality

	TR	IR
T	1.3	0.5
I	3.5	1.6
Ours	9.9	4.7

T: text-only, I: image-only

(Here we only reports R@1, more details are in the paper.)

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

Different vis./lan. encoders

Language Model	TR	IR
BERT	9.9	4.7
CLIP	**31.4**	**17.1**

Vision Model	TR	IR
NFNet	9.9	4.7
VIT_LoRA	**10.4**	**5.4**
NF_ResNet50	6.5	3.46
NF_RegNet	7.8	3.28

Single-modality vs. multi-modality

Takeaway:

- Distillation would be impossible if we solely optimize one modality

	TR	IR
T	1.3	0.5
I	3.5	1.6
Ours	9.9	4.7

T: text-only, I: image-only

Image-Text Pair Initialization

Takeaway:

- Initializing texts from scratch
- Initializing images from scratch

Real Image	Real Text	TR	IR
		0.4	0.1
	✓	1.1	0.1
✓		9	3.9
✓	✓	**9.9**	**4.7**

(Here we only reports R@1, more details are in the paper.)

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

Different vis./lan. encoders

Language Model	TR	IR
BERT	9.9	4.7
CLIP	**31.4**	**17.1**

Vision Model	TR	IR
NFNet	9.9	4.7
VIT_LoRA	**10.4**	**5.4**
NF_ResNet50	6.5	3.46
NF_RegNet	7.8	3.28

Single-modality vs. multi-modality

Takeaway:

- Distillation would be impossible if we solely optimize one modality

	TR	IR
T	1.3	0.5
I	3.5	1.6
Ours	9.9	4.7

T: text-only, I: image-only

Image-Text Pair Initialization

Takeaway:

- ✓ Initializing texts from scratch
- ✗ Initializing images from scratch

Real Image	Real Text	TR	IR
		0.4	0.1
	✓	1.1	0.1
✓		9	3.9
✓	✓	**9.9**	**4.7**

(Here we only reports R@1, more details are in the paper.)

Experiments

Ablation Studies

With and without LoRA on ViT

Dataset	#Pairs	Without LoRA TR	Without LoRA IR	With LoRA TR	With LoRA IR
Flickr30K	100	1.5	0.6	10.4 ± 0.8	5.4 ± 0.2
	1000	3.3	1.5	15.8 ± 1.4	8.1 ± 0.7

Different vis./lan. encoders

Language Model	TR	IR
BERT	9.9	4.7
CLIP	**31.4**	**17.1**

Vision Model	TR	IR
NFNet	9.9	4.7
VIT_LoRA	**10.4**	**5.4**
NF_ResNet50	6.5	3.46
NF_RegNet	7.8	3.28

Single-modality vs. multi-modality

Takeaway:

- Distillation would be impossible if we solely optimize one modality

	TR	IR
T	1.3	0.5
I	3.5	1.6
Ours	9.9	4.7

T: text-only, I: image-only

Image-Text Pair Initialization

Takeaway:

- ✅ Initializing texts from scratch
- ❌ Initializing images from scratch

image component plays a more critical role in the distilled dataset.

Real Image	Real Text	TR	IR
		0.4	0.1
	✓	1.1	0.1
✓		9	3.9
✓	✓	**9.9**	**4.7**

(Here we only reports R@1, more details are in the paper.)

| Conclusion

Summary

- In this work, we propose the first vision-language dataset distillation method.
- Our experiments show that co-distilling different modalities via bi-trajectory matching holds promise.
- We hope that the insights we gathered can serve as roadmap for future studies exploring more complex settings.

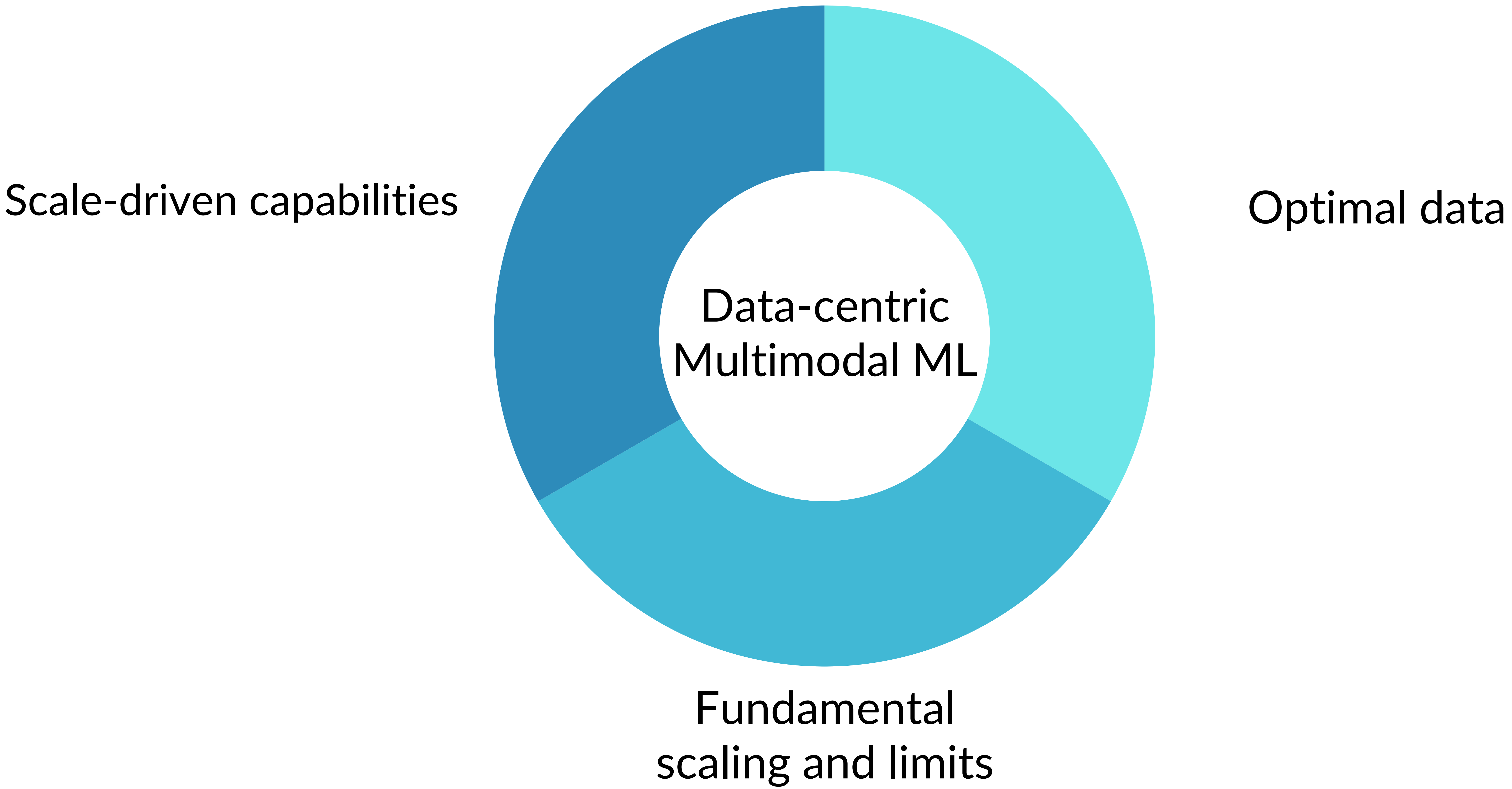

Byron Zhang


Zhiwei Deng


Olga Russakovsky

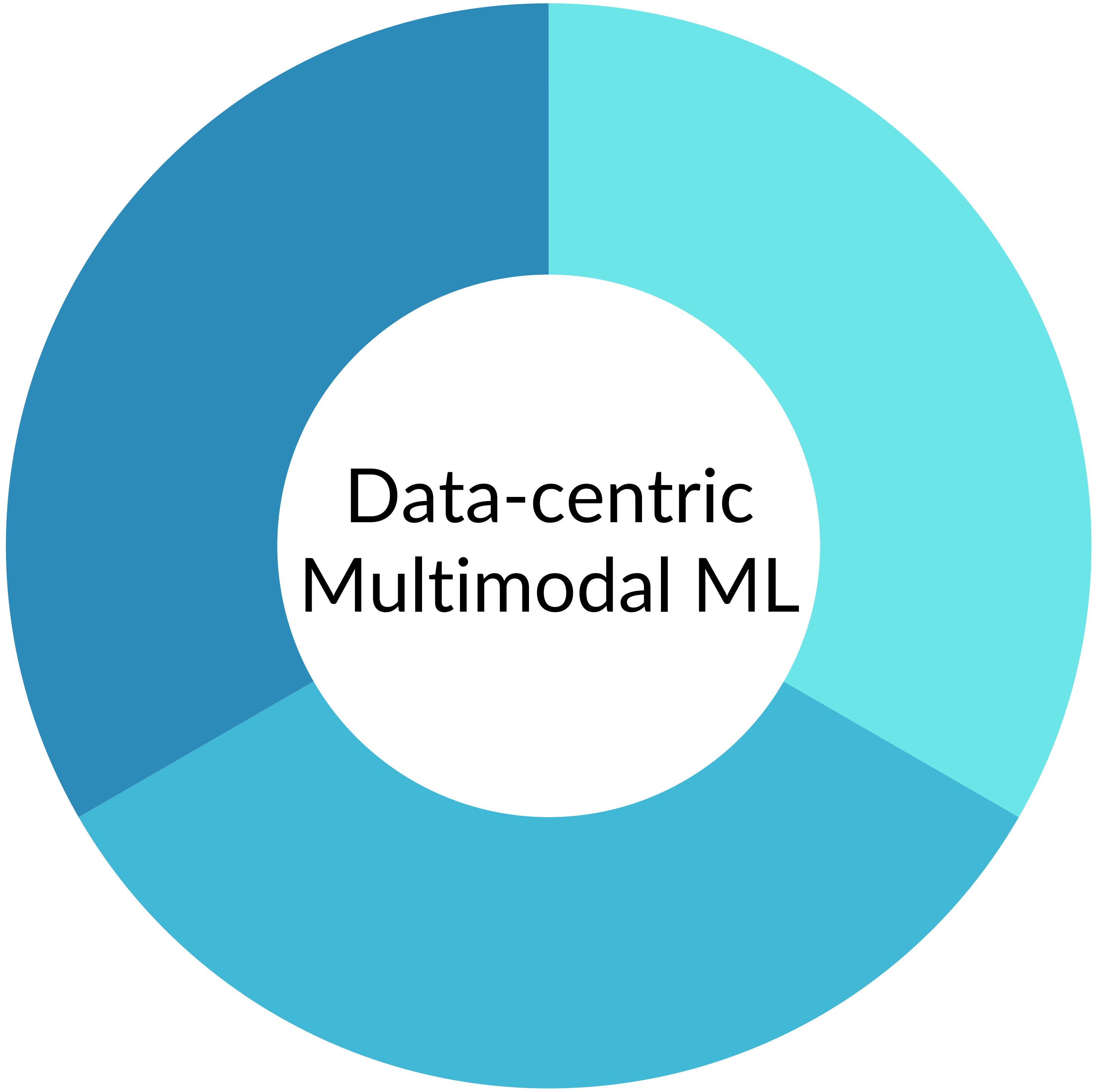
Website arXiv Code

Looking Forward



Looking Forward

Scale-driven capabilities



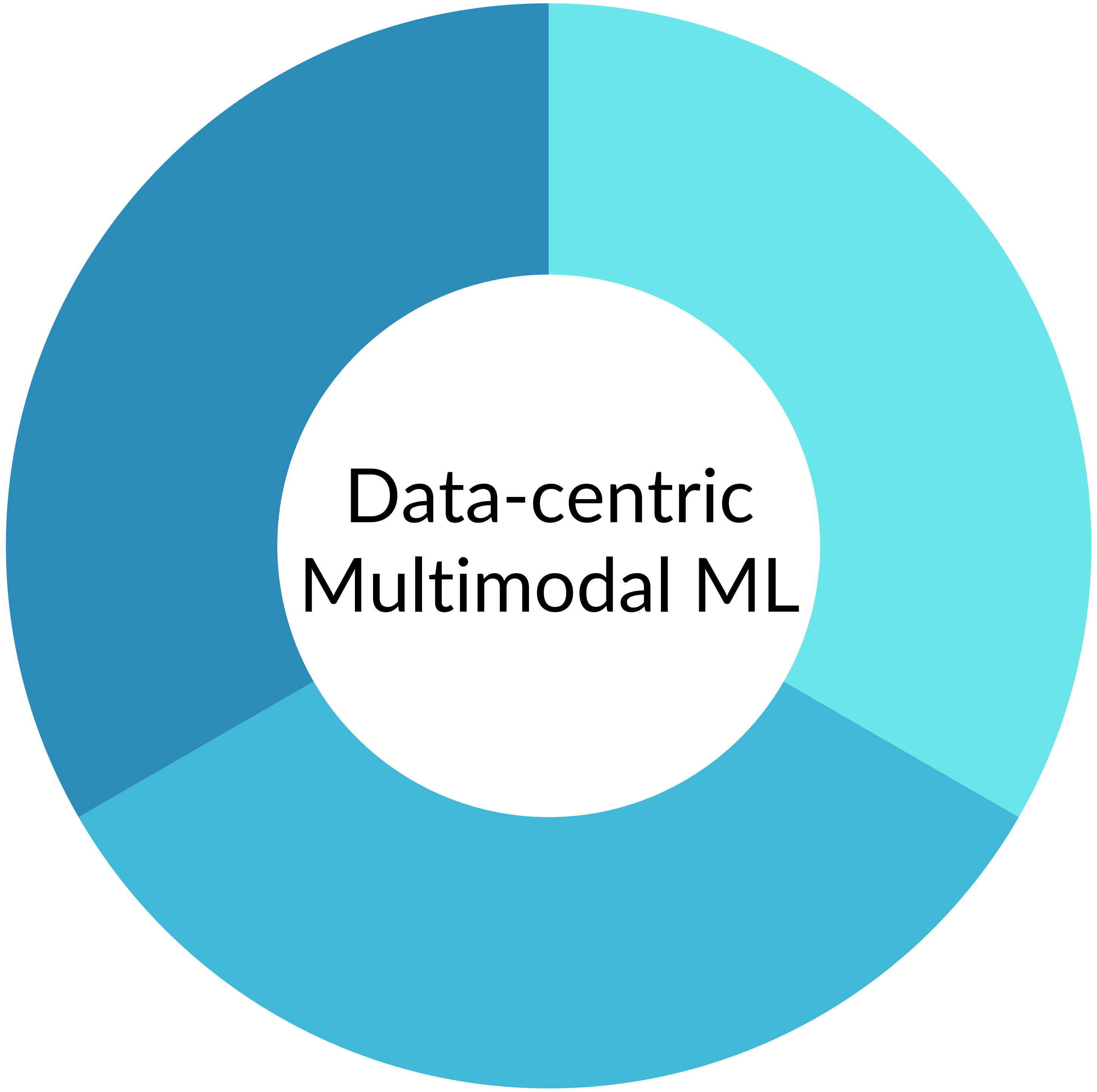
Optimal data

What subsets of the data are most important for the capabilities of foundation multimodal models?
What kind of synthetic multimodal data we can build?

Fundamental scaling and limits

Looking Forward

Scale-driven capabilities



Optimal data

What subsets of the data are most important for the capabilities of foundation multimodal models?
What kind of synthetic multimodal data we can build?

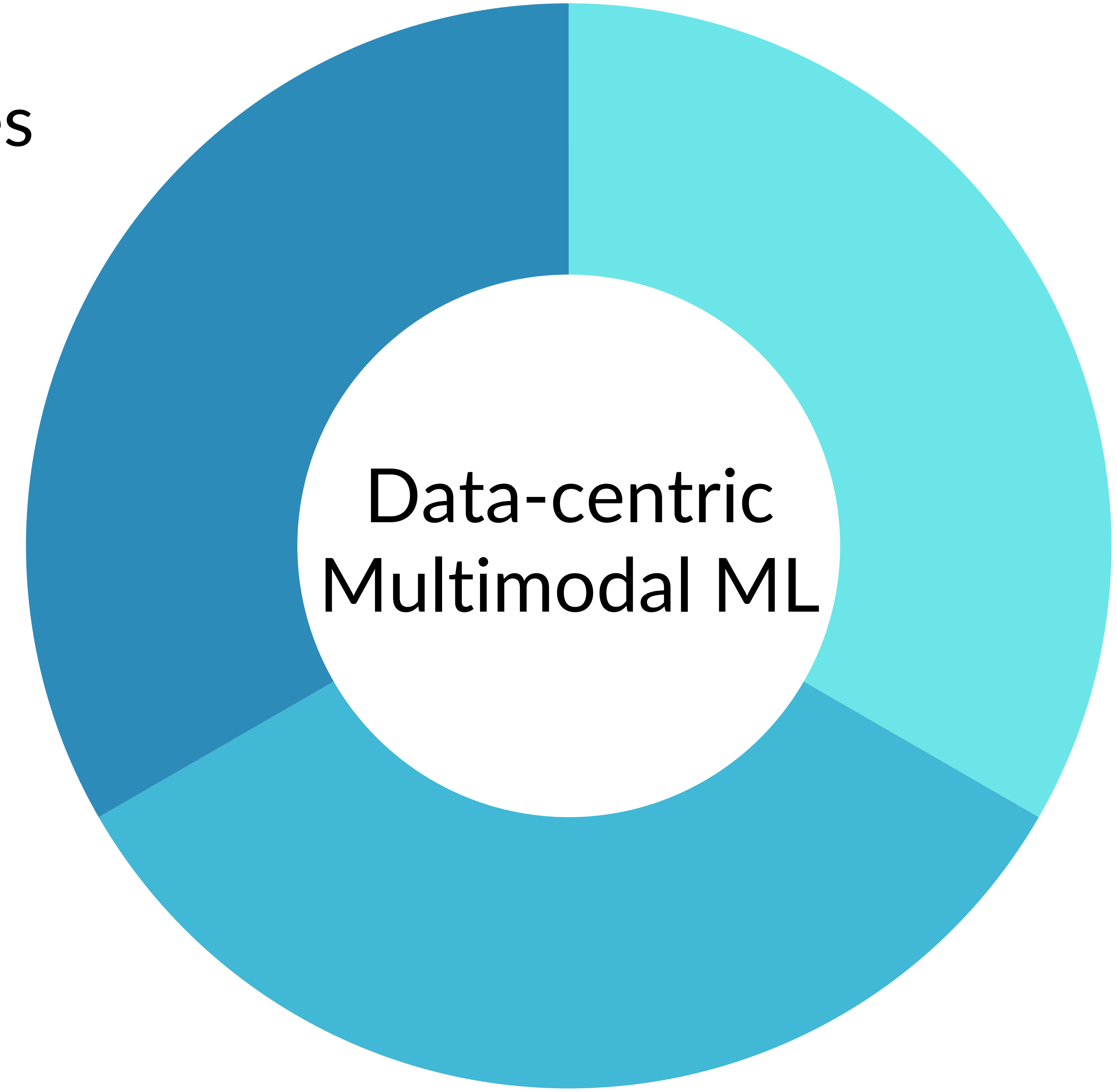
Fundamental scaling and limits

Given a fixed compute budget, is it better to increase the multimodal model size or the multimodal dataset size?

Looking Forward

Scale-driven capabilities

a more rigorous understanding of what increasing the scale does to the multimodal training procedure and how these desirable emergent capabilities come about



Optimal data

What subsets of the data are most important for the capabilities of foundation multimodal models?
What kind of synthetic multimodal data we can build?

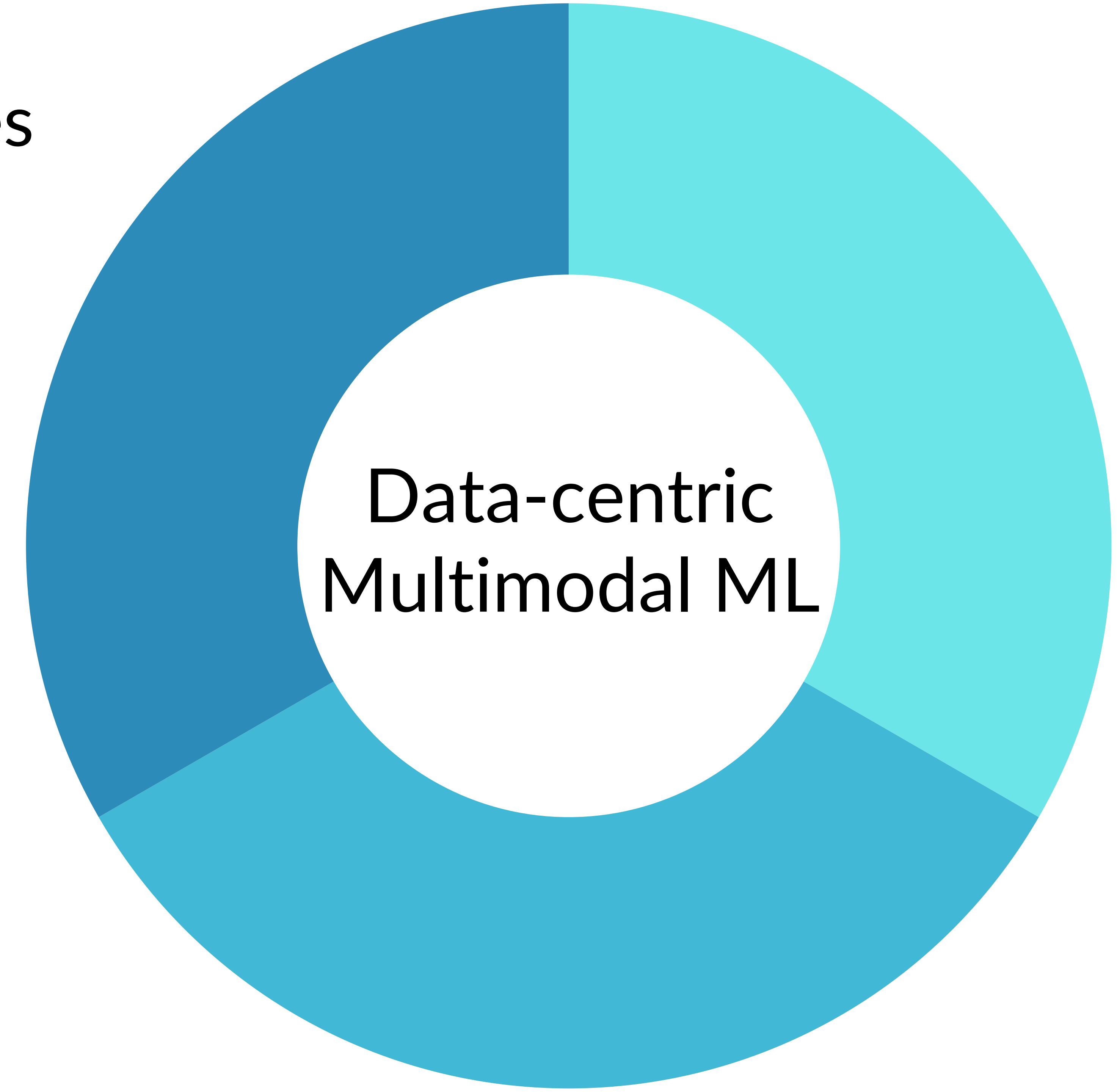
Fundamental scaling and limits

Given a fixed compute budget, is it better to increase the multimodal model size or the multimodal dataset size?

Looking Forward

Scale-driven capabilities

a more rigorous understanding of what increasing the scale does to the multimodal training procedure and how these desirable emergent capabilities come about



Optimal data

What subsets of the data are most important for the capabilities of foundation multimodal models?
What kind of synthetic multimodal data we can build?

Fundamental scaling and limits

Given a fixed compute budget, is it better to increase the multimodal model size or the multimodal dataset size?

Data-centric Multimodal ML

Summary

- In this work, we propose the first vision-language dataset distillation method.
- Our experiments show that co-distilling different modalities via bi-trajectory matching holds promise.
- We hope that the insights we gathered can serve as roadmap for future studies exploring more complex settings.

Byron Zhang

Zhiwei Deng

Olga Russakovsky

Website arXiv Code