Modality-agnostic Self-supervised Learning with Meta-learned Masked Auto-encoder

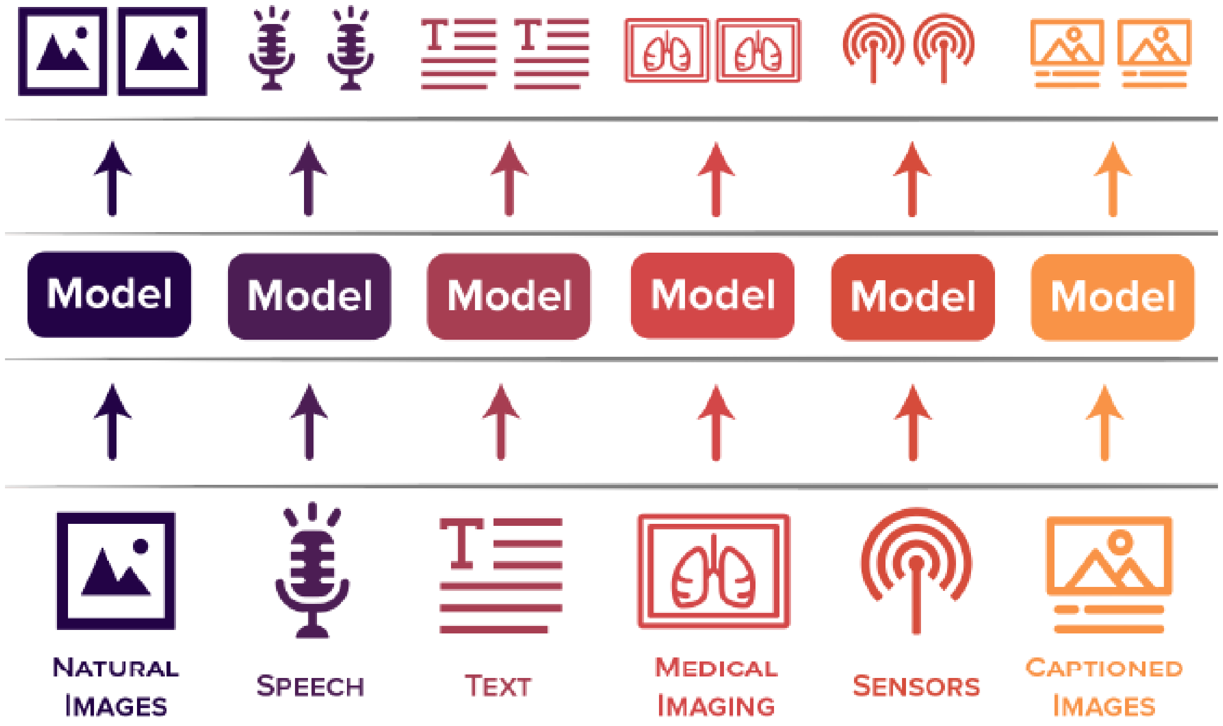
Huiwon Jang[A*]   Jihoon Tack[A*]   Daewon Choi[B]   Jongheon Jeong[A]   Jinwoo Shin[A]

[A]Korea Advanced Institute of Science and Technology (KAIST)
[B]Korea University

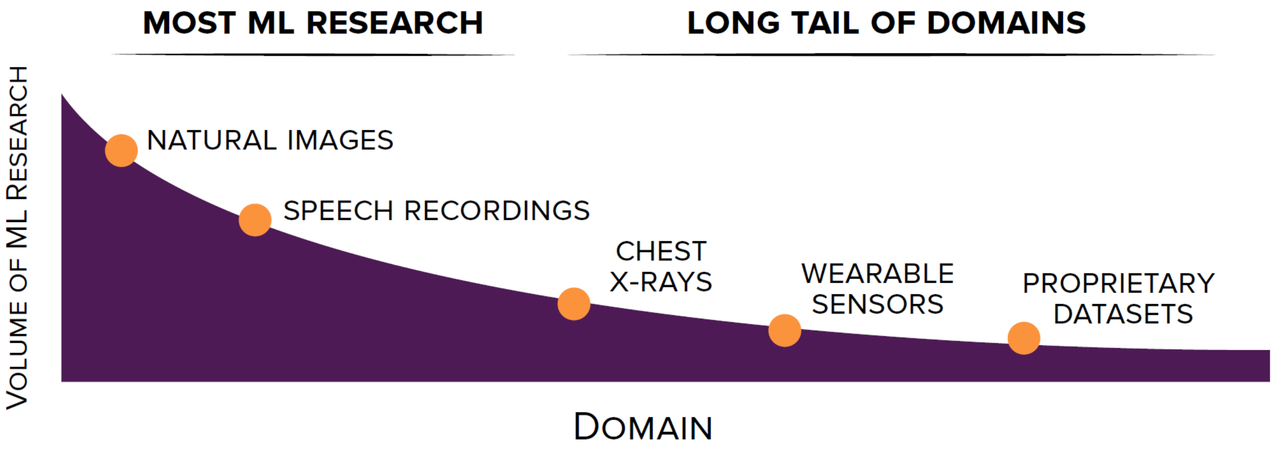
*Equal contribution

Importance of modality-agnostic self-supervised learning

- **Modality-agnostic SSL** learns representation **without modality-specific inductive bias**
  - SSL has achieved a remarkable success in various fields: **Vision** (SimCLR, MAE), **NLP** (BERT, GPT), …
  - **Benefit:** We can apply SSL approach to pretrain **new & long tail of** modality or domain



[1] Modality-agnostic SSL

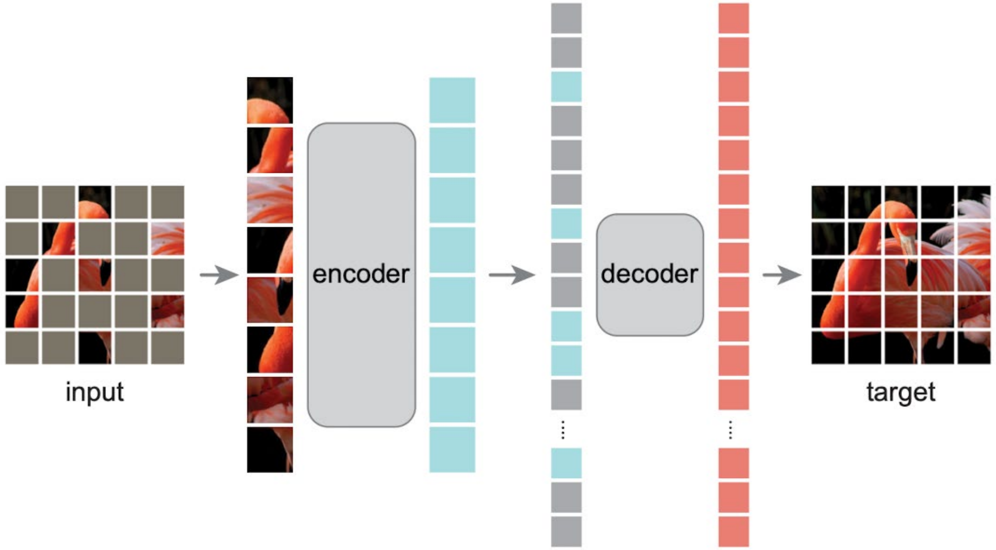


[2] Benefit of modality-agnostic SSL

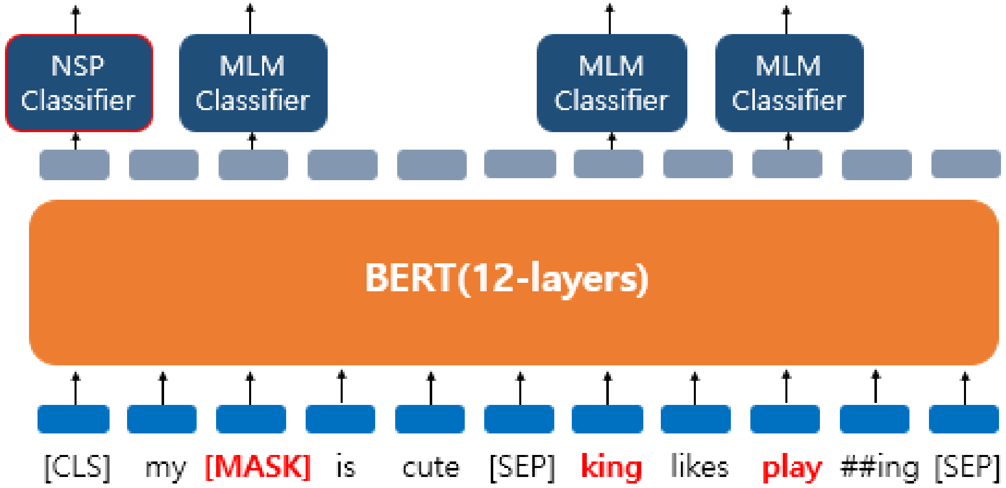
[1-2] Tamkin et al., DABS: A Domain-agnostic Benchmark for Self-supervised Learning, NeurIPS Datasets and Benchmarks, 2021

(Motivation) Masked Auto-Encoder

- **MAE** is a powerful SSL framework for various domains
  - MAE do **not need any domain-specific inductive bias**
  - Not only image domain (MAE), but also Language (BERT), Tabular (Vime), Audio (AudioMAE)



MAE [1] (Modality: Image)



BERT [2] (Modality: Language)

Research Questions

🤔 Is **MAE** indeed a **modality-agnostic** with a <u>proper decoder</u>?

🤔 How can we **improve MAE** in a modality-agnostic manner?

[1] He et al., Masked Autoencoders are Scalable Vision Learners, CVPR 2022

[2] Devlin et al., BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, NAACL 2019

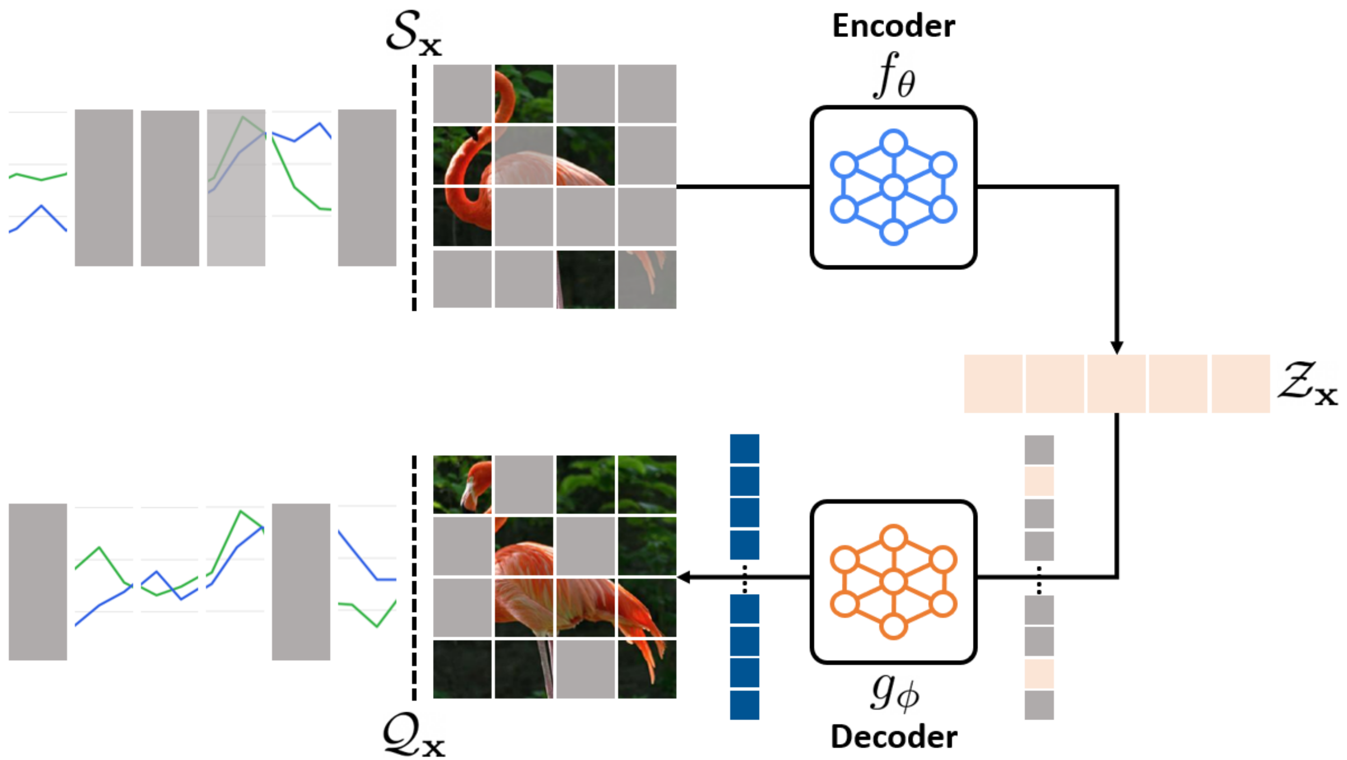
(Motivation) Masked Auto-Encoder

🤔 Is **MAE** indeed a **modality-agnostic** with a proper decoder?

💡 **Observation:** MAE with a **<u>proper decoder size</u>** outperforms previous approaches

- Improving MAE must be a promising direction to be better modality-agnostic SSL

decoder size	EuroSAT	Pfam	LibriSpeech
prev. best	**87.4**	54.7	60.2
0	86.3	44.7	33.3
2	86.7	**61.4**	68.1
4	**87.4**	61.3	64.1
6	86.7	**61.4**	**74.1**



(Motivation) Masked Auto-Encoder

🤔 How can we **improve MAE** in a modality-agnostic manner?

💡 **MAE** can be interpreted as an **amortization-based meta-learner**

- We can improve MAE by **leveraging the advances of meta-learning**

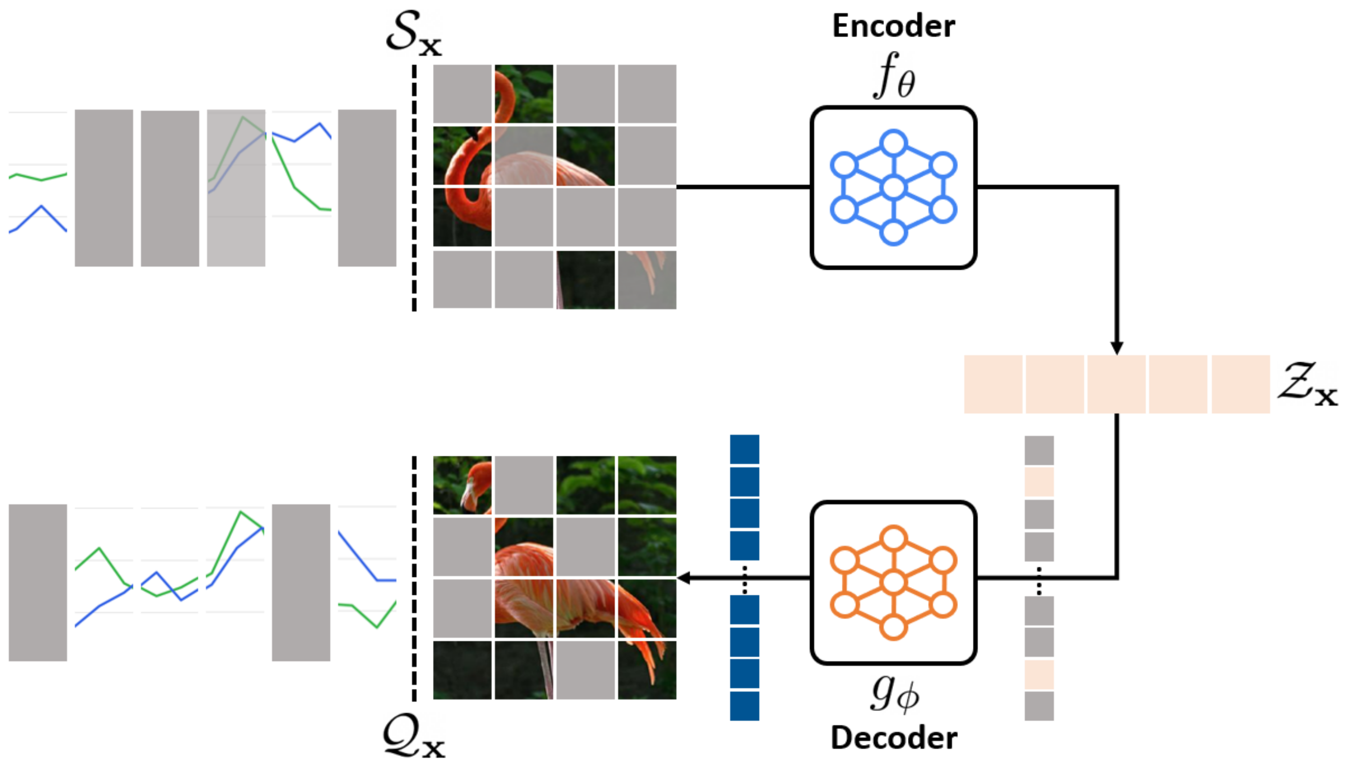
Amortization-based meta-learning (# task = 1)

- $\mathcal{S} \cup \mathcal{Q} = \{(\mathbf{x}_i, \mathrm{y}_i)\}_{i=1}^{N} \sim \mathcal{T}$: Sampling task
- $\mathcal{Z} = f_\theta(\mathcal{S})$: Memory
- $\mathrm{y}^{(q)} = g_\phi\big(\mathbf{x}^{(q)}; \mathcal{Z}\big)$

⬍

Task formulation of MAE (# task = 1)

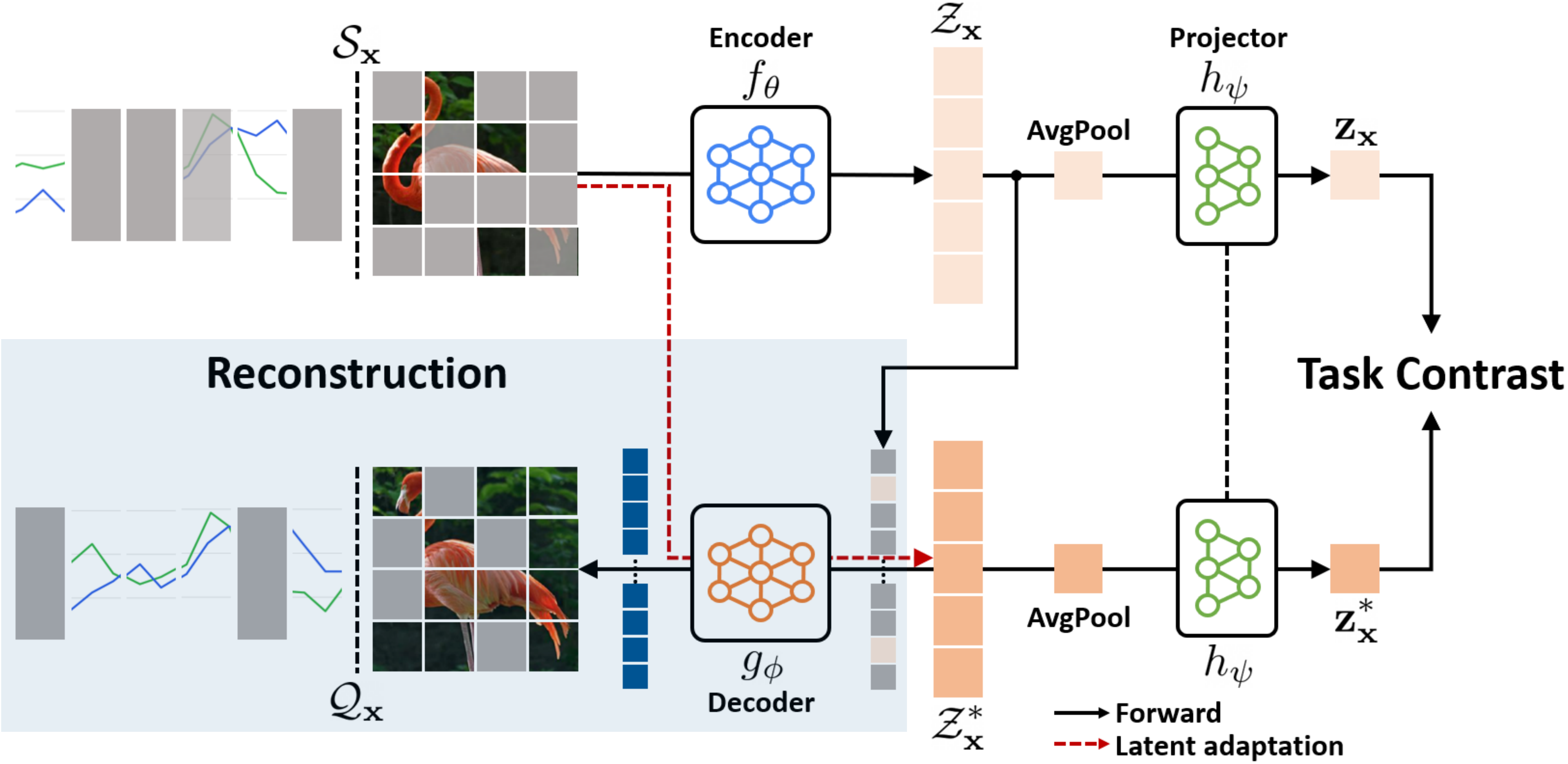
- $\mathrm{Tokenize}(\mathbf{x}) \coloneqq \big\{\big(m, \bar{\mathbf{x}}^{(m)}\big)\big\}_{m=1}^{M} = \mathcal{S}_\mathbf{x} \cup \mathcal{Q}_\mathbf{x}$
- $\mathcal{Z}_\mathbf{x} = f_\theta(\mathcal{S}_\mathbf{x})$
- $\bar{\mathbf{x}}^{(q)} = g_\phi^{(q)}(\mathcal{Z}_\mathbf{x}) \coloneqq g_\phi(q; \mathcal{Z}_\mathbf{x})$



Method: Meta-learned Masked Auto-Encoder (MetaMAE)

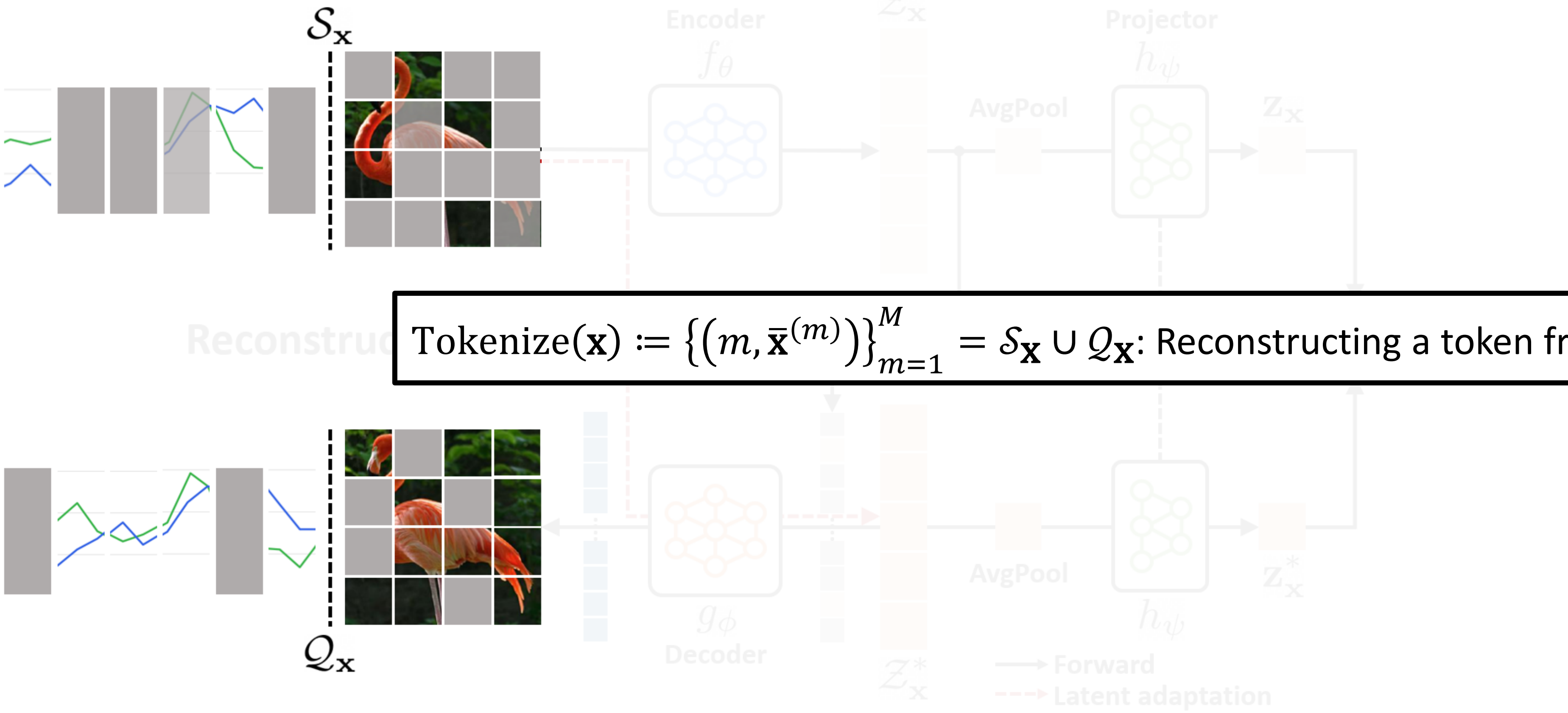
🤔 How can we **improve MAE** in a modality-agnostic manner?

- **Idea: Reconstruction** from adapted latent representations + **Task contrast**



Method: Meta-learned Masked Auto-Encoder (MetaMAE)

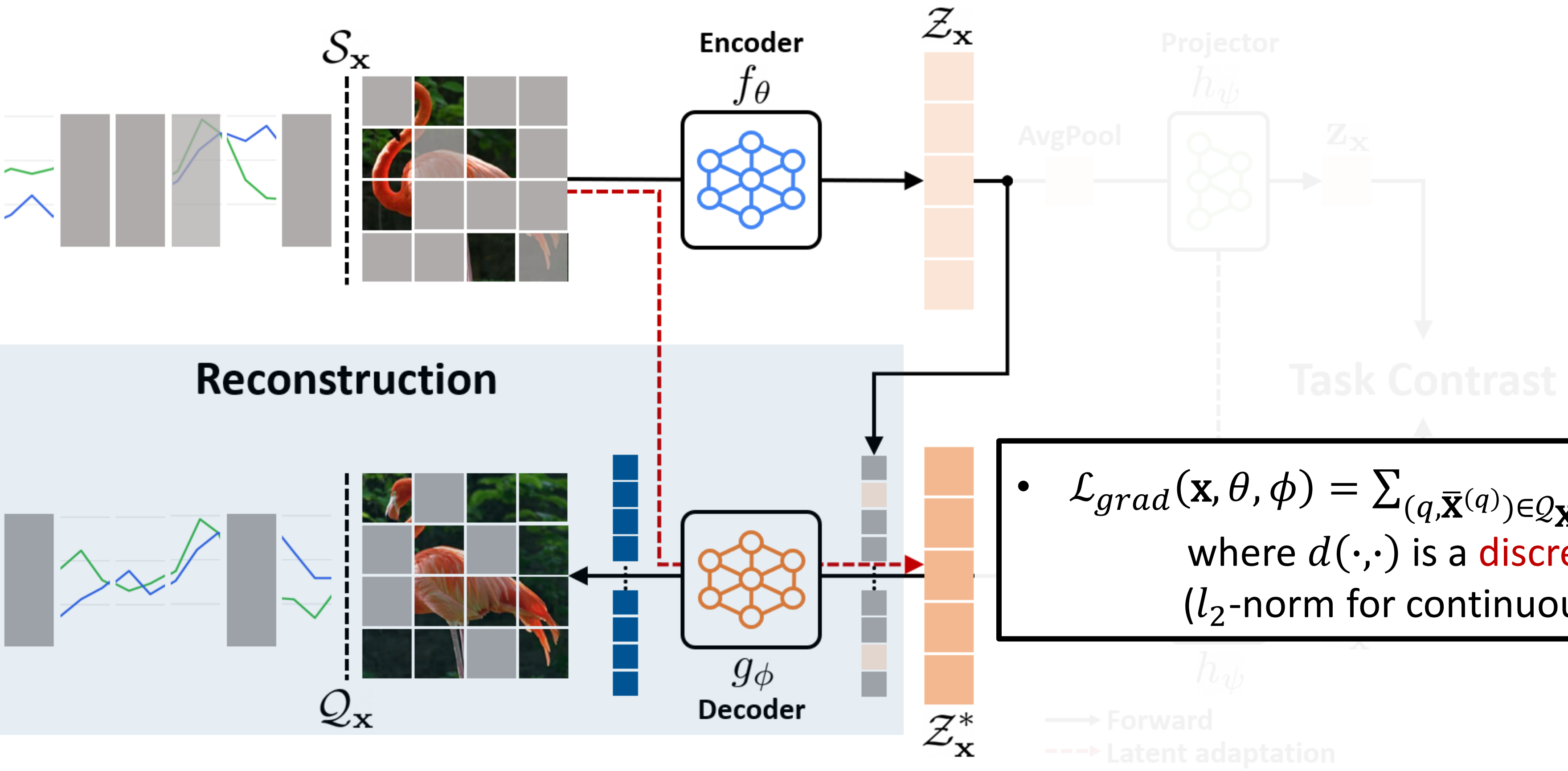
- **Idea: Reconstruction** from adapted latent representations + **Task contrast**
  - We assume that tokenized $\mathbf{x}$ is a **few-shot prediction task**



$\mathrm{Tokenize}(\mathbf{x}) \coloneqq \{(m, \bar{\mathbf{x}}^{(m)})\}_{m=1}^{M} = \mathcal{S}_{\mathbf{x}} \cup \mathcal{Q}_{\mathbf{x}}$: Reconstructing a token from an index is task

Method: Meta-learned Masked Auto-Encoder (MetaMAE)

- **Idea: Reconstruction** from adapted latent representations + **Task contrast**
  - We assume that tokenized $\mathbf{x}$ is a **few-shot prediction task**
  - Latent adaptation via **Gradient-based meta-learning** to predict queries
    - $\mathcal{Z}_{\mathbf{x}}^{*} = \mathcal{Z}_{\mathbf{x}} - \alpha \nabla_{\mathcal{Z}_{\mathbf{x}}} \mathcal{L}_{MAE}(\theta, \phi; \tilde{\mathcal{S}}_{\mathbf{x}})$



- $\mathcal{L}_{grad}(\mathbf{x}, \theta, \phi) = \sum_{(q, \bar{\mathbf{x}}^{(q)}) \in \mathcal{Q}_{\mathbf{x}}} d(\bar{\mathbf{x}}^{(q)}, g_\phi^{(q)}(\mathcal{Z}_{\mathbf{x}}^{*}))$

  where $d(\cdot,\cdot)$ is a discrepancy function

  (l_2-norm for continuous & cross-entropy for discrete)

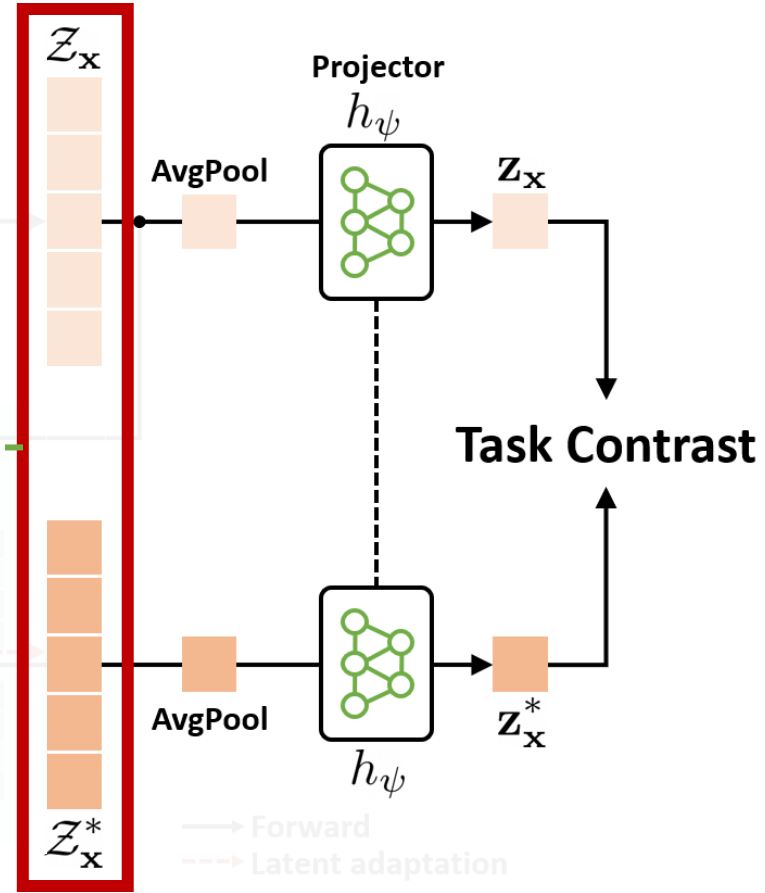
Method: Meta-learned Masked Auto-Encoder (MetaMAE)

- **Idea: Reconstruction** from adapted latent representations + **Task contrast**
  - We assume that <u>tokenized $\mathbf{x}$</u> is a **few-shot prediction task**
  - Latent adaptation via **Gradient-based meta-learning** <u>to predict queries</u>
  - **Task contrastive learning** between task-agnostic and task-specific representations

- $\mathbf{z}_\mathbf{x}$ and $\mathbf{z}^*_\mathbf{x}$: Task representation
- $\mathcal{T} = \cup_\mathbf{x}\{\mathbf{z}_\mathbf{x}, \mathbf{z}^*_\mathbf{x}\}$: Collection of all representations of tasks
- $\mathcal{L}_{task-con}(\mathbf{x},\theta,\phi) = \frac{1}{2}[l_{con}(\mathbf{z}_\mathbf{x};\mathbf{z}^*_\mathbf{x},\mathcal{T}\setminus\{\mathbf{z}^*_\mathbf{x}\}) + l_{con}(\mathbf{z}^*_\mathbf{x};\mathbf{z}_\mathbf{x},\mathcal{T}\setminus\{\mathbf{z}_\mathbf{x}\})]$

  where $l_{con}(\mathbf{z};\mathbf{z}^+,\{\mathbf{z}^-\}) = -\log\frac{\exp(sim(\mathbf{z},\mathbf{z}^+)/\tau)}{\exp(sim(\mathbf{z},\mathbf{z}^+)/\tau)+\sum_{\mathbf{z}^-}\exp(sim(\mathbf{z},\mathbf{z}^-)/\tau)}$

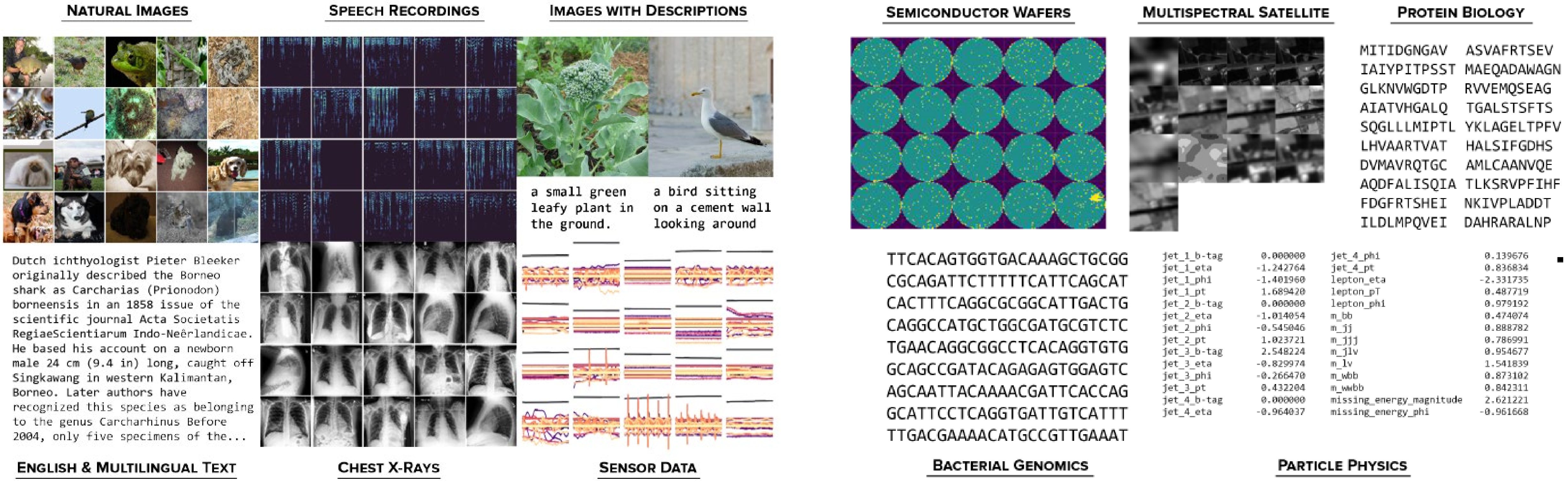
Representations from the same task



Final loss term: $\mathcal{L}_{grad}(\mathbf{x},\theta,\phi) + \lambda\mathcal{L}_{task-con}(\mathbf{x},\theta,\phi)$

Experiment: Setup

- **DABS 1.0 and 2.0 benchmarks**
  - Various modalities: time-series, tabular, multi-spectral image, token, speech, and RGB image
  - Various downstream tasks, including cross-domain tasks
  - Multi-modal tasks to verify the possibility for tackling unified SSL



Experiment: In-domain linear evaluation

- **MetaMAE** achieves <u>state-of-the-art</u> performance on in-domain linear evaluation

Modality	**Time-series**	**Tabular**	**MS Image**	**Token**		**Speech**	**RGB Image**
Dataset	PAMAP2	HIGGS	EuroSAT	Genom	Pfam	Libri	WaferMap
Random initialization							
Baseline	69.8†	54.8†	62.3†	37.2†	30.1	17.1*	77.7†
Self-supervised learning Framework							
e-Mix	80.1	65.7	87.4	40.5	31.3	60.2	92.6
ShED	85.2	68.0†	61.5†	33.6	54.7	34.8*	92.4†
Capri	-	-	67.4†	23.5†	27.4	25.4	92.5†
MAE	85.3†	70.0†	86.3†	53.6	44.7	46.0	93.9†
MetaMAE	**89.3**	**71.5**	**88.5**	**69.4**	**62.3**	**79.8**	**95.5**

- **MetaMAE** achieves <u>state-of-the-art</u> performance on vision-language

Table 3: Linear classification accuracy (%) pretrained on a vision-language dataset, MSCOCO.

			SSL Framework				
Pretrain data	**Transfer data**	Baseline	e-Mix	ShED	Capri	MAE	**MetaMAE**
MSCOCO	VQA	53.4	57.6	53.1	52.9	54.2	**69.7**
	Mismatched-caption	49.8	50.1	50.6	49.6	49.3	**70.5**

Experiment: Cross-domain linear evaluation

- **MetaMAE** achieves <u>state-of-the-art</u> performance on cross-domain linear evaluation
  - MetaMAE can be transferred to various cross-domain transfer learning scenarios across the modalities

			SSL Framework				
Pretrain data	**Transfer data**	Baseline	e-Mix	ShED	Capri	MAE	MetaMAE
Genomics	Genomics-OOD	8.6	9.7	7.3	5.5	22.2	**37.2**
Pfam	SCOP	8.0	5.7	10.7	2.0	7.9	**11.8**
	Secondary	52.4	53.7	**67.6**	49.5	62.5	65.9
	Stability	0.31	0.39	**0.53**	0.26	0.40	**0.53**
	Fluorescence	0.04	0.20	0.27	0.06	0.06	**0.31**
LibriSpeech	Audio MNIST	33.1*	80.4*	67.3*	53.6	45.1	**89.5**
	Fluent Loc	62.1*	60.9*	60.2*	59.8	61.7	**66.7**
	Fluent Act	26.2*	29.9*	30.5*	28.3	26.8	**38.4**
	Fluent Obj	30.1*	39.9*	39.4*	33.1	32.0	**49.3**
	Google Speech	4.9*	19.2*	20.7*	13.7	9.5	**46.8**
	VoxCeleb1	0.6*	2.4*	2.8*	1.6	1.6	**7.4**
ImageNet32	CIFAR-10	24.2*	39.4*	39.6*	48.7	46.0	**59.2**
	CUB	1.6*	3.9*	3.0*	3.7	3.1	**6.3**
	VGG Flowers	9.0*	26.0*	13.0*	18.6	22.2	**36.3**
	DTD	7.4*	8.8*	18.4*	14.7	14.2	**20.9**
	Traffic Sign	14.3*	65.1*	27.5*	28.0	32.0	**67.1**
	Aircraft	2.7*	10.2*	5.6*	6.4	5.9	**16.4**

Conclusion

We propose **MetaMAE:** a novel and effective modality-agnostic SSL framework

- We interpret mask reconstruction task of MAE as a meta-learning
  to suggest an integration with advanced modality-agnostic meta-learning methods

- We show that MetaMAE significantly improves the performance across a diverse range of modalities
- We verify the possibility of MetaMAE for tackling unified multi-modal SSL

Thank you for your attention!