

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

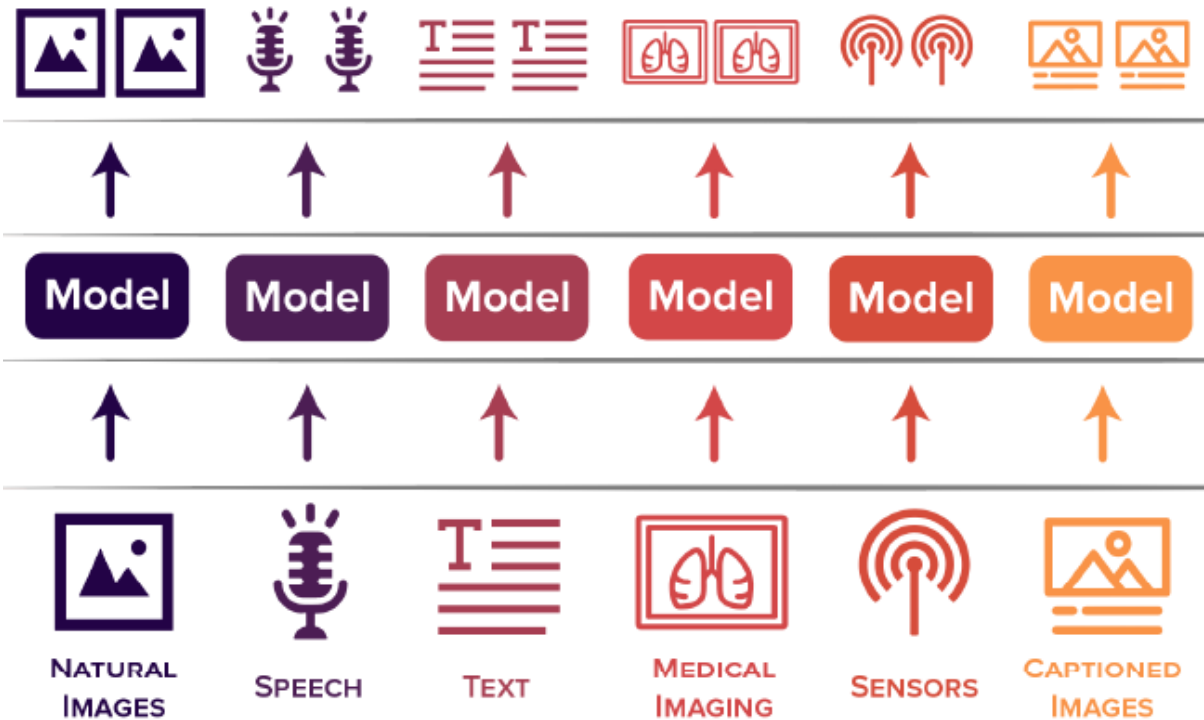
Huiwon Jang<sup>A\*</sup>   Jihoon Tack<sup>A\*</sup>   Daewon Choi<sup>B</sup>   Jongheon Jeong<sup>A</sup>   Jinwoo Shin<sup>A</sup>

<sup>A</sup>Korea Advanced Institute of Science and Technology (KAIST)

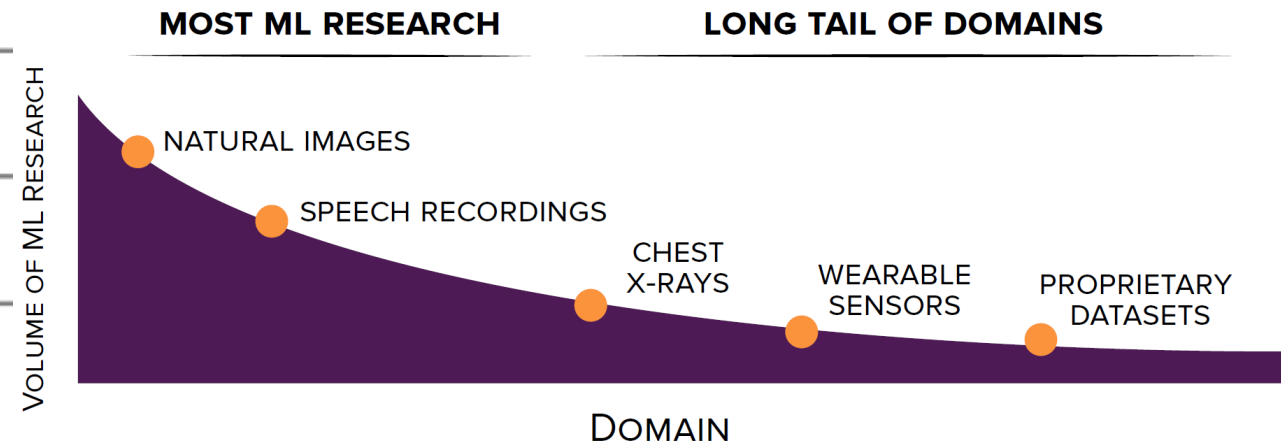
<sup>B</sup>Korea University

# 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



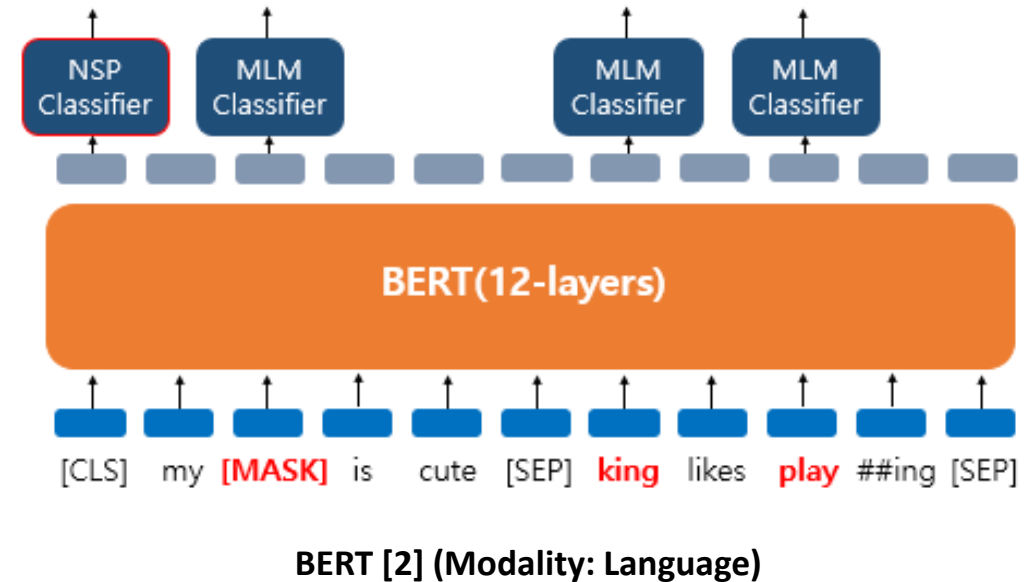
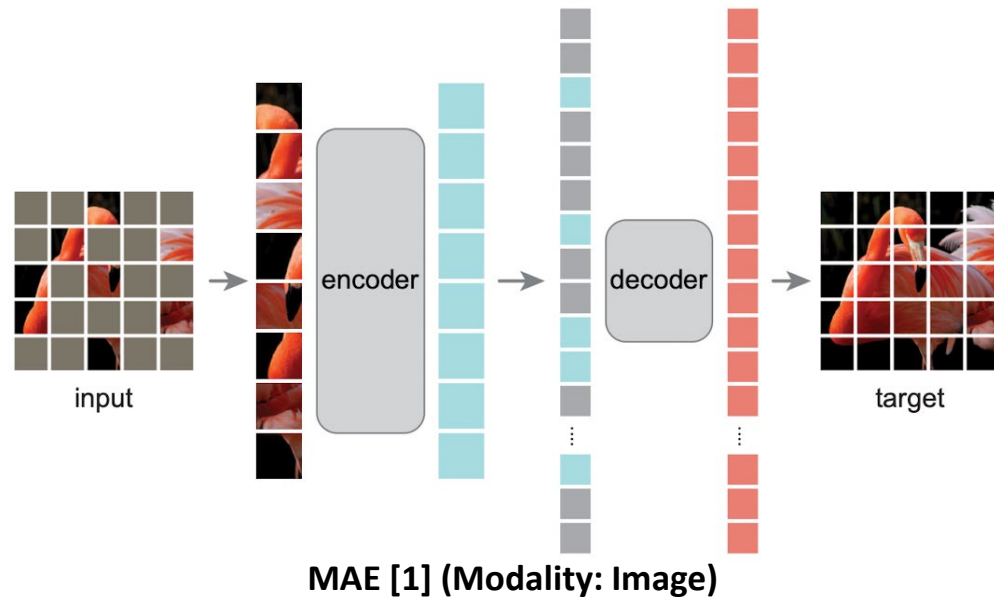
[1] Modality-agnostic SSL



[2] Benefit of modality-agnostic SSL

# (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)



## Research Questions

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

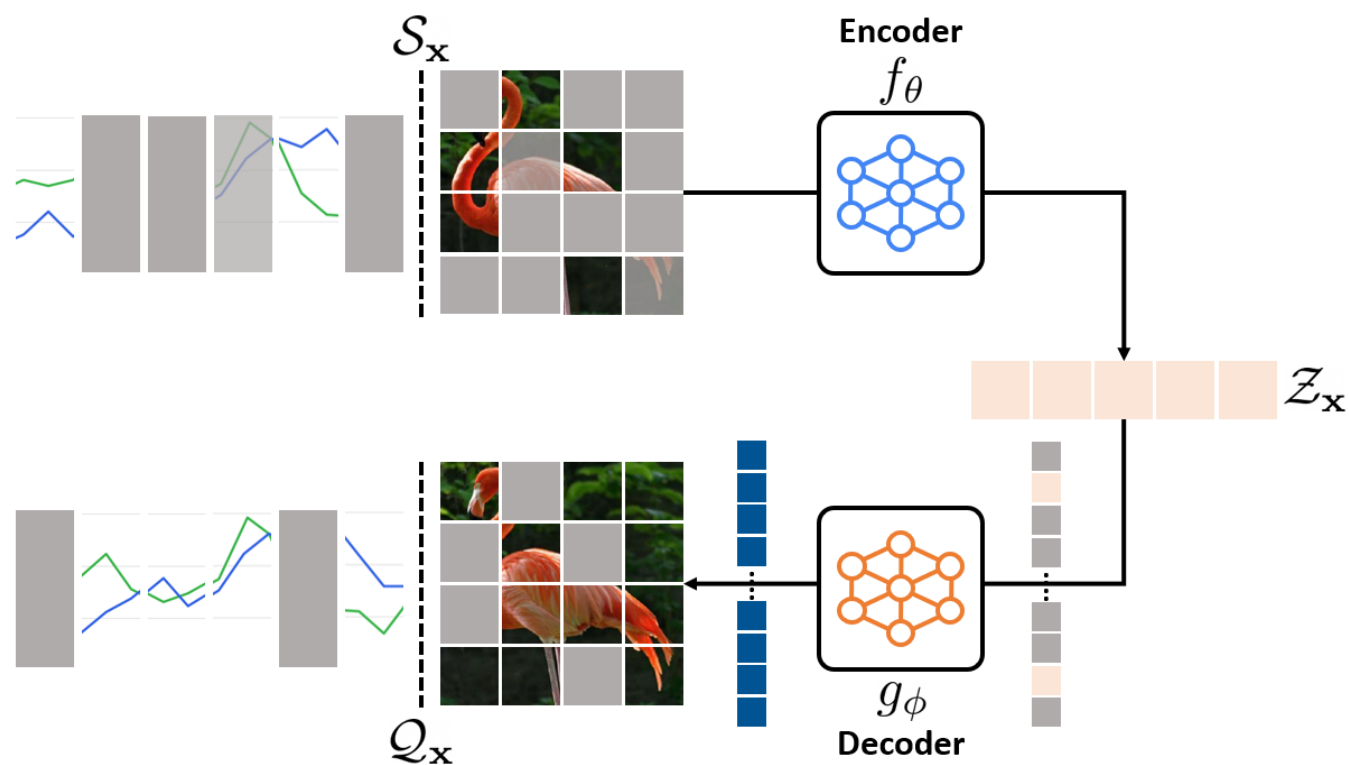
# (Motivation) Masked Auto-Encoder

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

💡 **Observation:** MAE with a proper decoder size outperforms previous approaches

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

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



# (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**

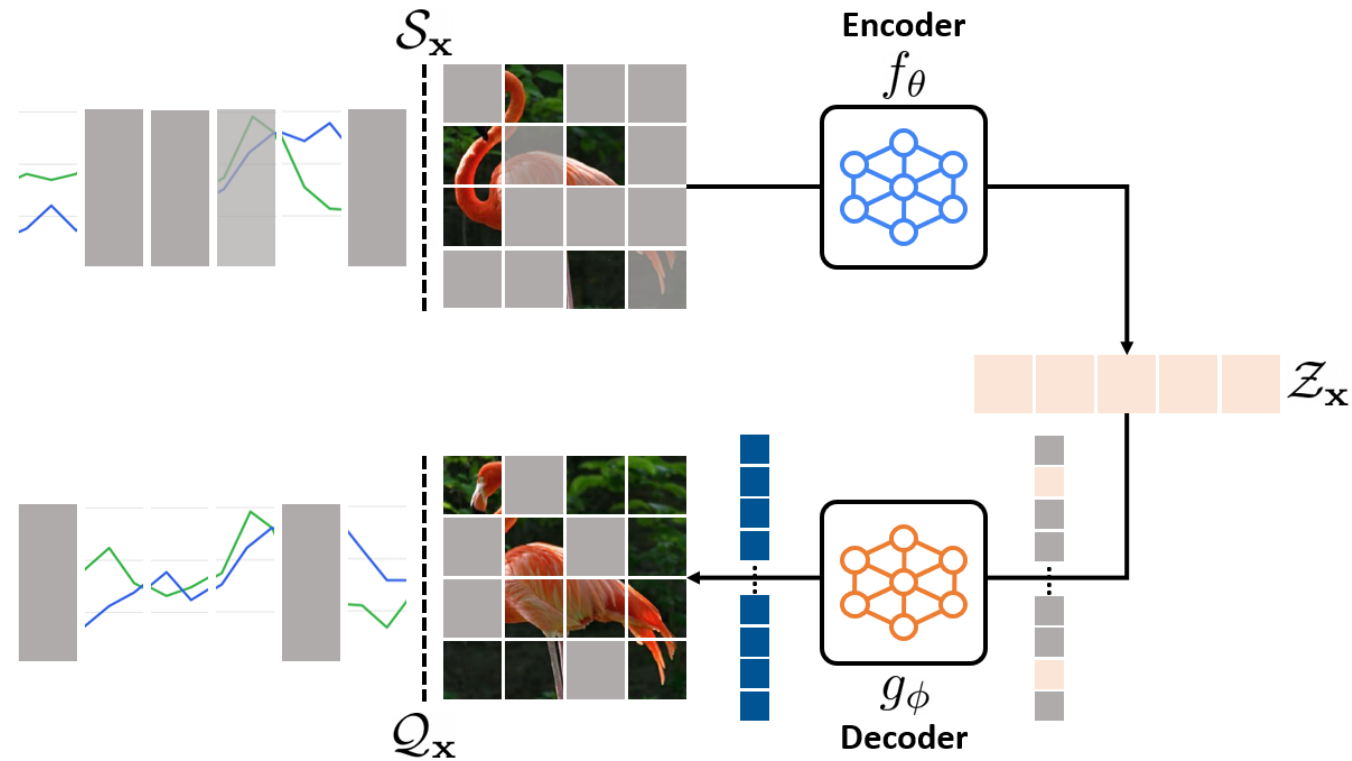
## Amortization-based meta-learning (# task = 1)

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



## Task formulation of MAE (# task = 1)

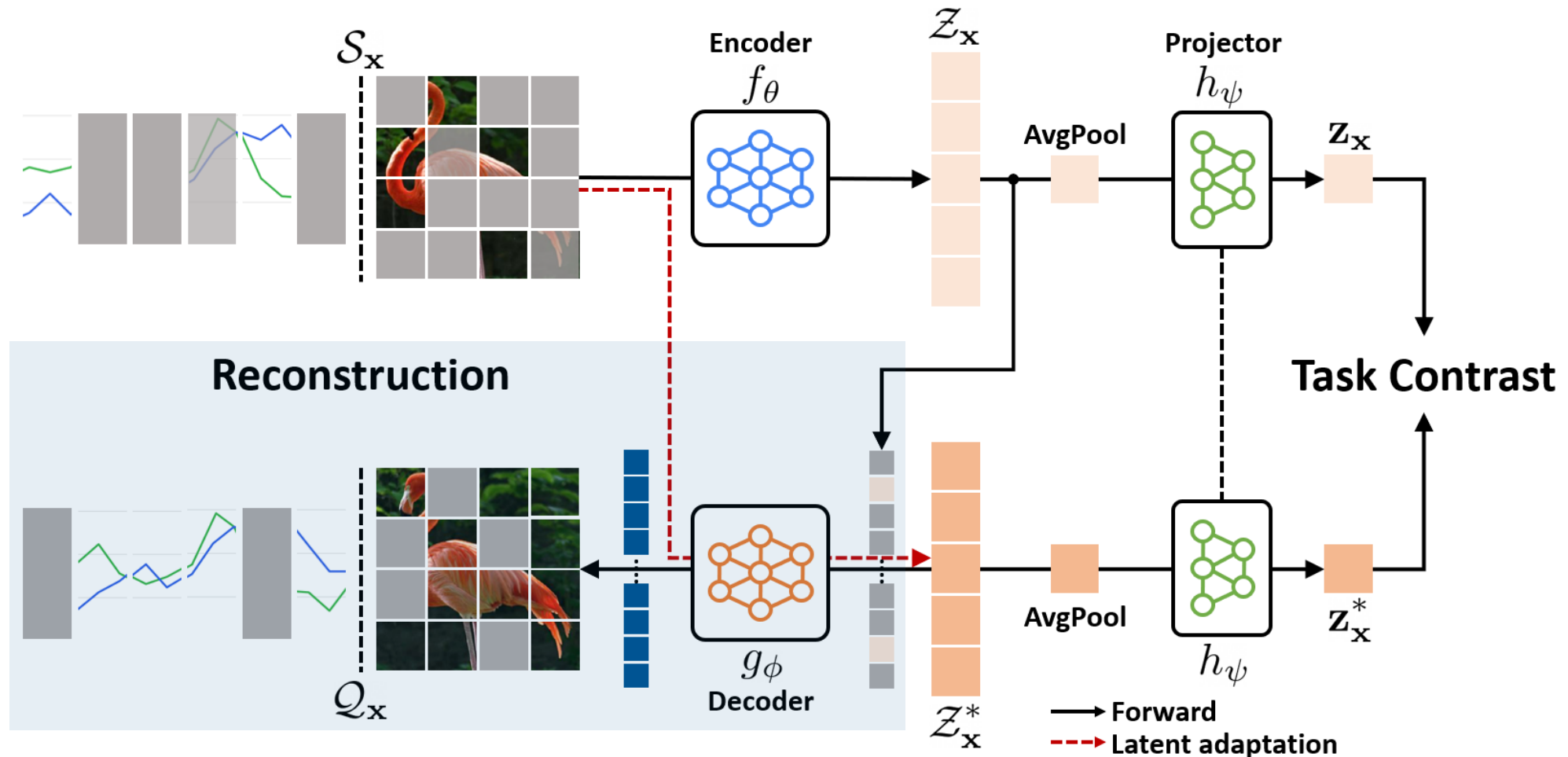
- $\text{Tokenize}(\mathbf{x}) := \{(m, \bar{\mathbf{x}}^{(m)})\}_{m=1}^M = \mathcal{S}_\mathbf{x} \cup \mathcal{Q}_\mathbf{x}$
- $\mathcal{Z}_\mathbf{x} = f_\theta(\mathcal{S}_\mathbf{x})$
- $\bar{\mathbf{x}}^{(q)} = g_\phi^{(q)}(\mathcal{Z}_\mathbf{x}) := g_\phi(q; \mathcal{Z}_\mathbf{x})$



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

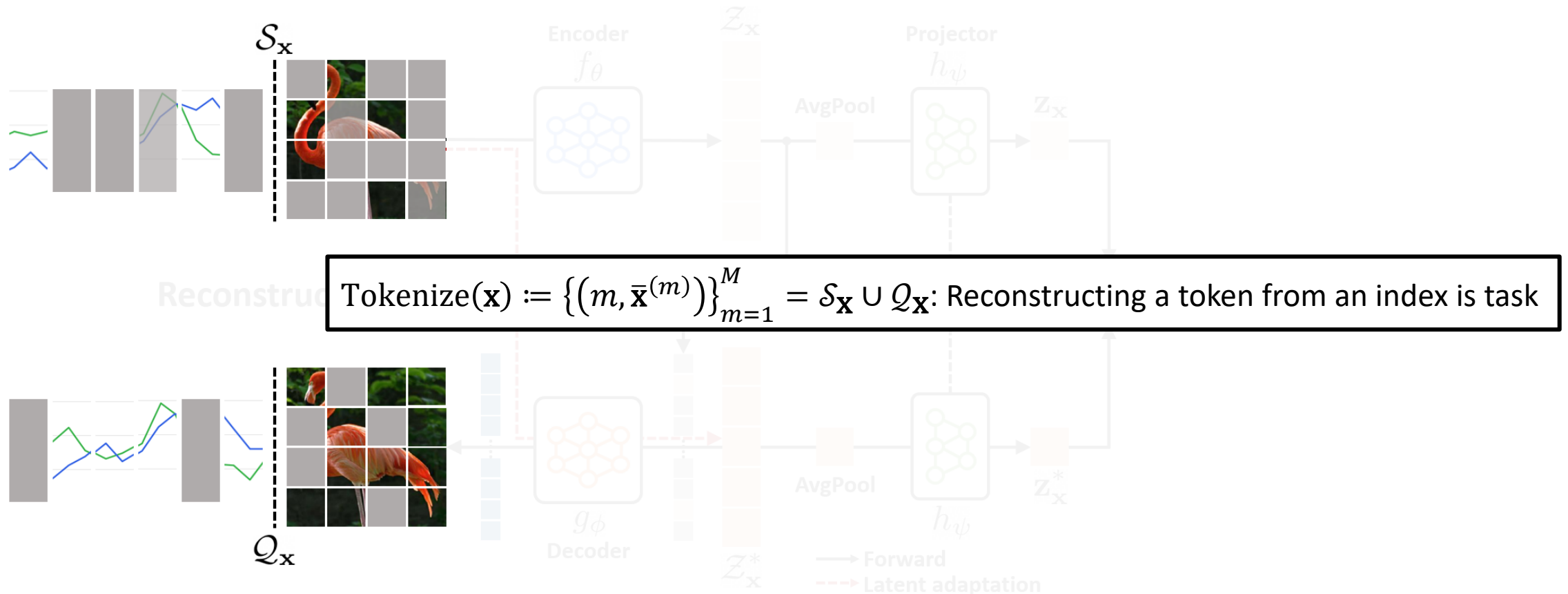
🤔 How can we **improve MAE** in a modality-agnostic manner?

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



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

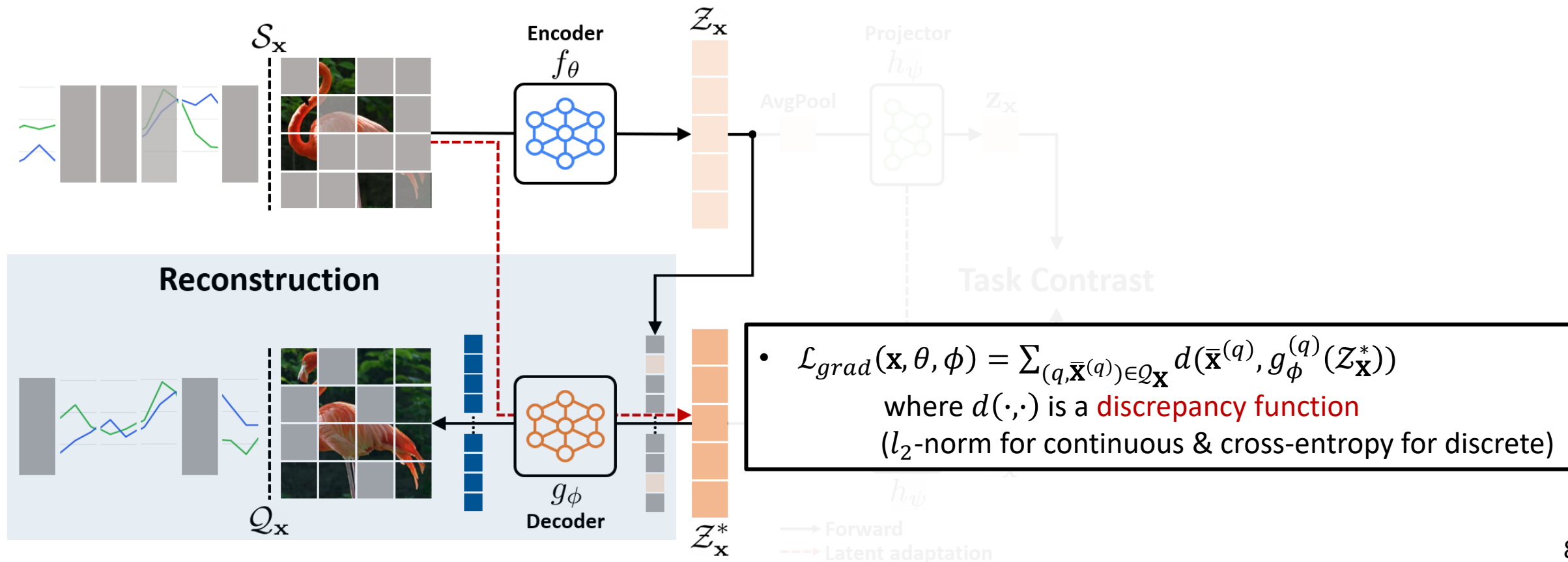
- **Idea: Reconstruction** from adapted latent representations + **Task contrast**
  - We assume that tokenized  $\mathbf{x}$  is a **few-shot prediction task**



# 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
  - $Z_{\mathbf{x}}^* = Z_{\mathbf{x}} - \alpha \nabla_{Z_{\mathbf{x}}} \mathcal{L}_{MAE}(\theta, \phi; \tilde{\mathcal{S}}_{\mathbf{x}})$

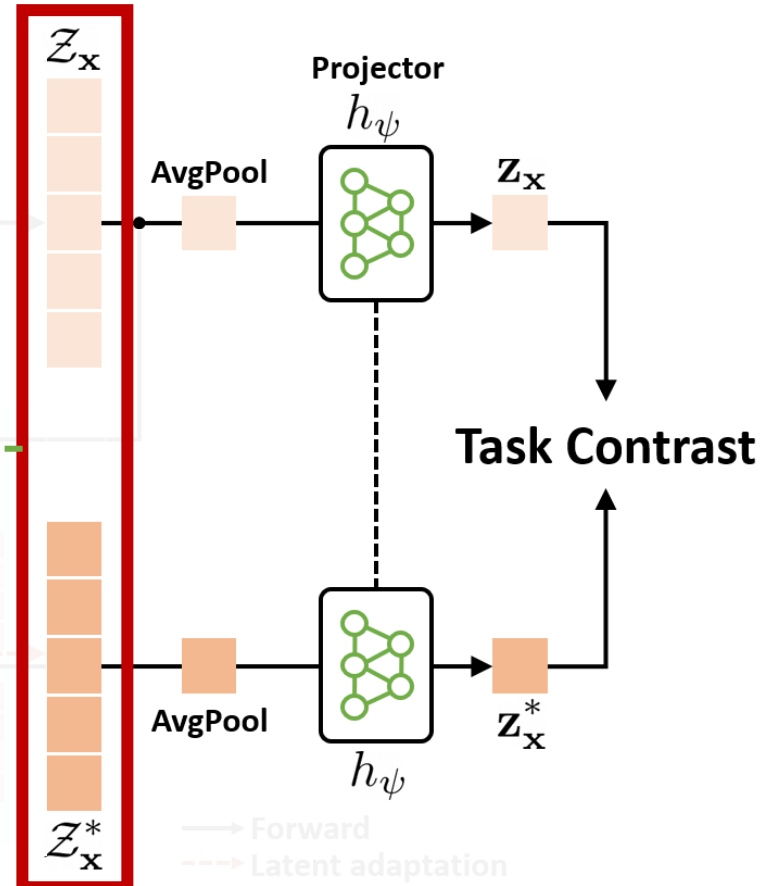




# Method: Meta-learned Masked Auto-Encoder (MetaMAE)

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

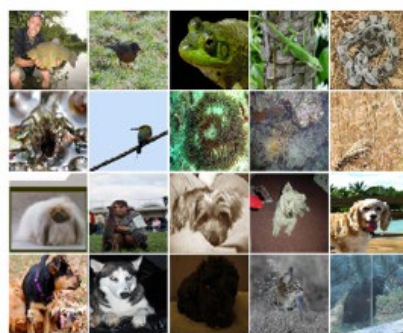
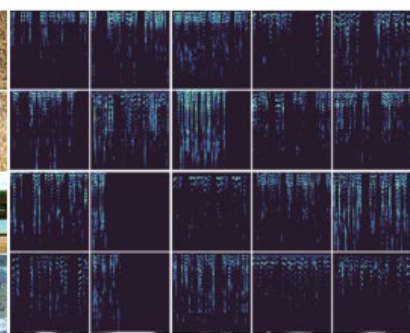

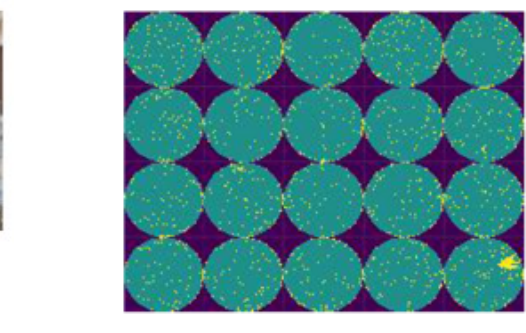
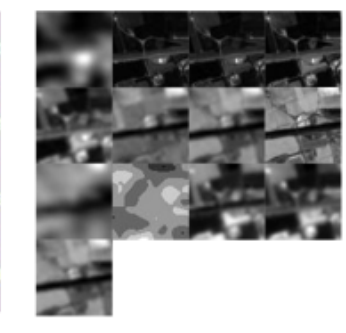
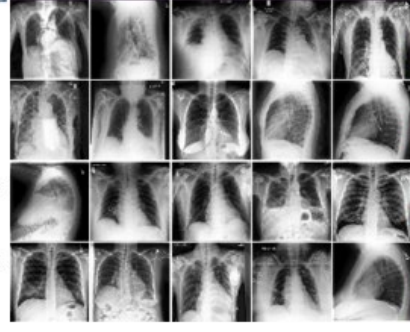
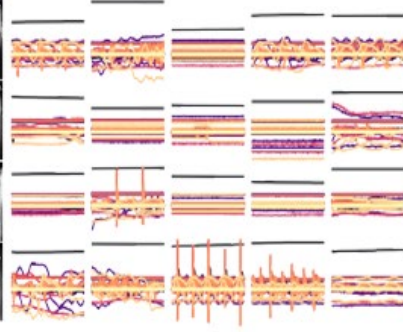


Final loss term:  $\mathcal{L}_{grad}(\mathbf{x}, \theta, \phi) + \lambda \mathcal{L}_{task-con}(\mathbf{x}, \theta, \phi)$

# 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

<u>NATURAL IMAGES</u>	<u>SPEECH RECORDINGS</u>	<u>IMAGES WITH DESCRIPTIONS</u>	<u>SEMICONDUCTOR WAFERS</u>	<u>MULTISPECTRAL SATELLITE</u>	<u>PROTEIN BIOLOGY</u>																																																								
		 <p>a small green leafy plant in the ground.</p> <p>a bird sitting on a cement wall looking around</p>			<table border="0"> <tr><td>MITIDGNGAV</td><td>ASVAFRTSEV</td></tr> <tr><td>IAIYPITPST</td><td>MAEQADAWAGN</td></tr> <tr><td>GLKNVWGDT</td><td>RVVEMQSEAG</td></tr> <tr><td>AIATVHGALQ</td><td>TGALSTSFTS</td></tr> <tr><td>SQGLLLMIPTL</td><td>YKLAGELTPFV</td></tr> <tr><td>LHVAARTVAT</td><td>HALSIFGDHS</td></tr> <tr><td>DVMAVRQTGC</td><td>AMLCANVQE</td></tr> <tr><td>AQDFALISQIA</td><td>TLKSRVFFIHF</td></tr> <tr><td>FDGFRTSHEI</td><td>NKIVPLADDT</td></tr> <tr><td>ILDLMQVEI</td><td>DAHRARALNP</td></tr> </table>	MITIDGNGAV	ASVAFRTSEV	IAIYPITPST	MAEQADAWAGN	GLKNVWGDT	RVVEMQSEAG	AIATVHGALQ	TGALSTSFTS	SQGLLLMIPTL	YKLAGELTPFV	LHVAARTVAT	HALSIFGDHS	DVMAVRQTGC	AMLCANVQE	AQDFALISQIA	TLKSRVFFIHF	FDGFRTSHEI	NKIVPLADDT	ILDLMQVEI	DAHRARALNP																																				
MITIDGNGAV	ASVAFRTSEV																																																												
IAIYPITPST	MAEQADAWAGN																																																												
GLKNVWGDT	RVVEMQSEAG																																																												
AIATVHGALQ	TGALSTSFTS																																																												
SQGLLLMIPTL	YKLAGELTPFV																																																												
LHVAARTVAT	HALSIFGDHS																																																												
DVMAVRQTGC	AMLCANVQE																																																												
AQDFALISQIA	TLKSRVFFIHF																																																												
FDGFRTSHEI	NKIVPLADDT																																																												
ILDLMQVEI	DAHRARALNP																																																												
<p>Dutch ichthyologist Pieter Bleeker originally described the Borneo shark as <i>Carcharias (Prionodon) borneensis</i> in an 1858 issue of the scientific journal <i>Acta Societatis RegiaeScientiarum Indo-Neerlandicae</i>. He based his account on a newborn male 24 cm (9.4 in) long, caught off Singkawang in western Kalimantan, Borneo. Later authors have recognized this species as belonging to the genus <i>Carcharhinus</i> Before 2004, only five specimens of the...</p>			<pre> TTCACAGTGGTGACAAAGCTGCGG CGCAGATTCTTTTTCATTGAGCAT CACTTTCAGGCGCGGCATTGACTG CAGGCCATGCTGGCGATGCGTCTC TGAACAGGCGGCCTCACAGGTGTG GCAGCCGATACAGAGAGTGGAGTC AGCAATTACAAAACGATTACCAG GCATTCTCAGGTGATTGTCATTT TTGACGAAAAACATGCCGTTGAAAT                     </pre>	<table border="0"> <tr><td>jet_1_b-tag</td><td>0.000000</td><td>jet_4_phi</td><td>0.139676</td></tr> <tr><td>jet_1_ota</td><td>-1.242764</td><td>jet_4_pt</td><td>0.836834</td></tr> <tr><td>jet_1_phi</td><td>-1.401960</td><td>lepton_eta</td><td>-2.331735</td></tr> <tr><td>jet_1_pt</td><td>1.689420</td><td>lepton_pT</td><td>0.487719</td></tr> <tr><td>jet_2_b-tag</td><td>0.000000</td><td>lepton_phi</td><td>0.979192</td></tr> <tr><td>jet_2_eta</td><td>-1.014054</td><td>m_bb</td><td>0.474074</td></tr> <tr><td>jet_2_phi</td><td>-0.545046</td><td>m_jj</td><td>0.888782</td></tr> <tr><td>jet_2_pt</td><td>1.023721</td><td>m_jjj</td><td>0.786991</td></tr> <tr><td>jet_3_b-tag</td><td>2.548224</td><td>m_jlv</td><td>0.954677</td></tr> <tr><td>jet_3_eta</td><td>-0.829974</td><td>m_lv</td><td>1.541839</td></tr> <tr><td>jet_3_phi</td><td>-0.266470</td><td>m_wbb</td><td>0.873102</td></tr> <tr><td>jet_3_pt</td><td>0.432204</td><td>m_wbb</td><td>0.842311</td></tr> <tr><td>jet_4_b-tag</td><td>0.000000</td><td>missing_energy_magnitude</td><td>2.621221</td></tr> <tr><td>jet_4_eta</td><td>-0.964037</td><td>missing_energy_phi</td><td>-0.961668</td></tr> </table>	jet_1_b-tag	0.000000	jet_4_phi	0.139676	jet_1_ota	-1.242764	jet_4_pt	0.836834	jet_1_phi	-1.401960	lepton_eta	-2.331735	jet_1_pt	1.689420	lepton_pT	0.487719	jet_2_b-tag	0.000000	lepton_phi	0.979192	jet_2_eta	-1.014054	m_bb	0.474074	jet_2_phi	-0.545046	m_jj	0.888782	jet_2_pt	1.023721	m_jjj	0.786991	jet_3_b-tag	2.548224	m_jlv	0.954677	jet_3_eta	-0.829974	m_lv	1.541839	jet_3_phi	-0.266470	m_wbb	0.873102	jet_3_pt	0.432204	m_wbb	0.842311	jet_4_b-tag	0.000000	missing_energy_magnitude	2.621221	jet_4_eta	-0.964037	missing_energy_phi	-0.961668	
jet_1_b-tag	0.000000	jet_4_phi	0.139676																																																										
jet_1_ota	-1.242764	jet_4_pt	0.836834																																																										
jet_1_phi	-1.401960	lepton_eta	-2.331735																																																										
jet_1_pt	1.689420	lepton_pT	0.487719																																																										
jet_2_b-tag	0.000000	lepton_phi	0.979192																																																										
jet_2_eta	-1.014054	m_bb	0.474074																																																										
jet_2_phi	-0.545046	m_jj	0.888782																																																										
jet_2_pt	1.023721	m_jjj	0.786991																																																										
jet_3_b-tag	2.548224	m_jlv	0.954677																																																										
jet_3_eta	-0.829974	m_lv	1.541839																																																										
jet_3_phi	-0.266470	m_wbb	0.873102																																																										
jet_3_pt	0.432204	m_wbb	0.842311																																																										
jet_4_b-tag	0.000000	missing_energy_magnitude	2.621221																																																										
jet_4_eta	-0.964037	missing_energy_phi	-0.961668																																																										
<u>ENGLISH &amp; MULTILINGUAL TEXT</u>	<u>CHEST X-RAYS</u>	<u>SENSOR DATA</u>	<u>BACTERIAL GENOMICS</u>	<u>PARTICLE PHYSICS</u>																																																									

# Experiment: In-domain linear evaluation

- **MetaMAE** achieves state-of-the-art performance on in-domain linear evaluation

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

- **MetaMAE** achieves state-of-the-art performance on vision-language

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

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

# Experiment: Cross-domain linear evaluation

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

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

# 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!