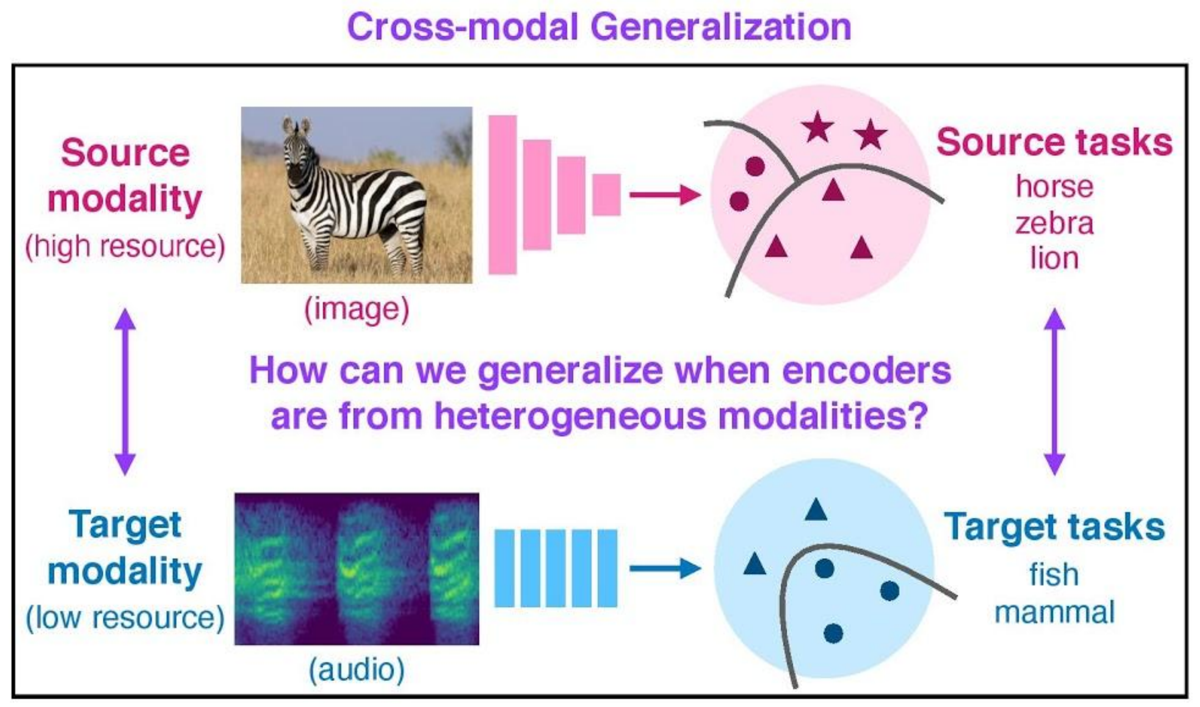
Cross-Modal Generalization:
Learning in Low Resource Modalities via Meta-Alignment

Paul Pu Liang, Peter Wu, Liu Ziyin, Louis-Philippe Morency, Ruslan Salakhutdinov

pliang@cs.cmu.edu @pliang279

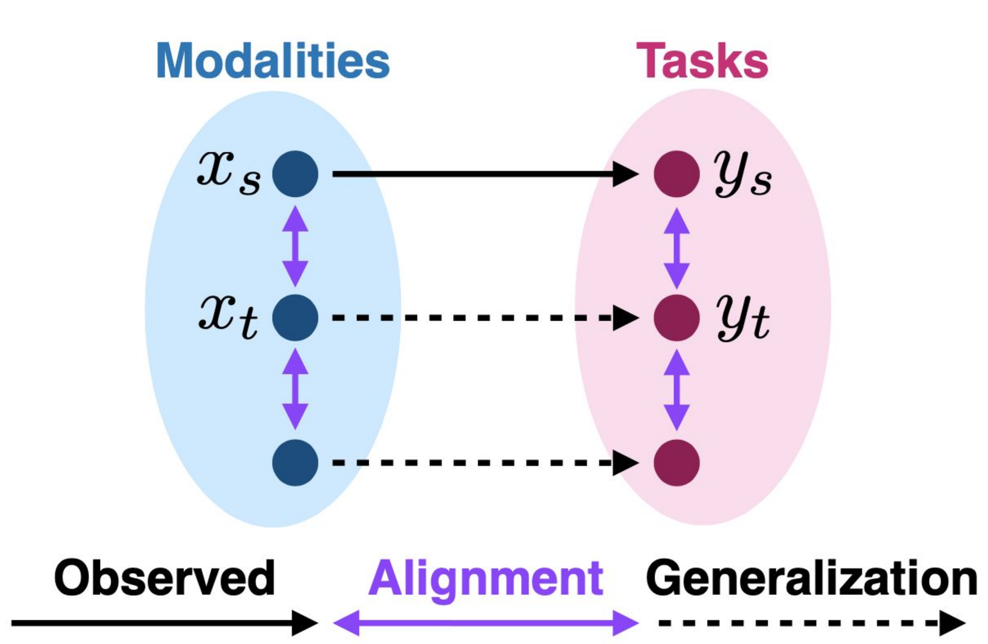
Cross-modal Generalization



How can we learn in low-resource target modalities?

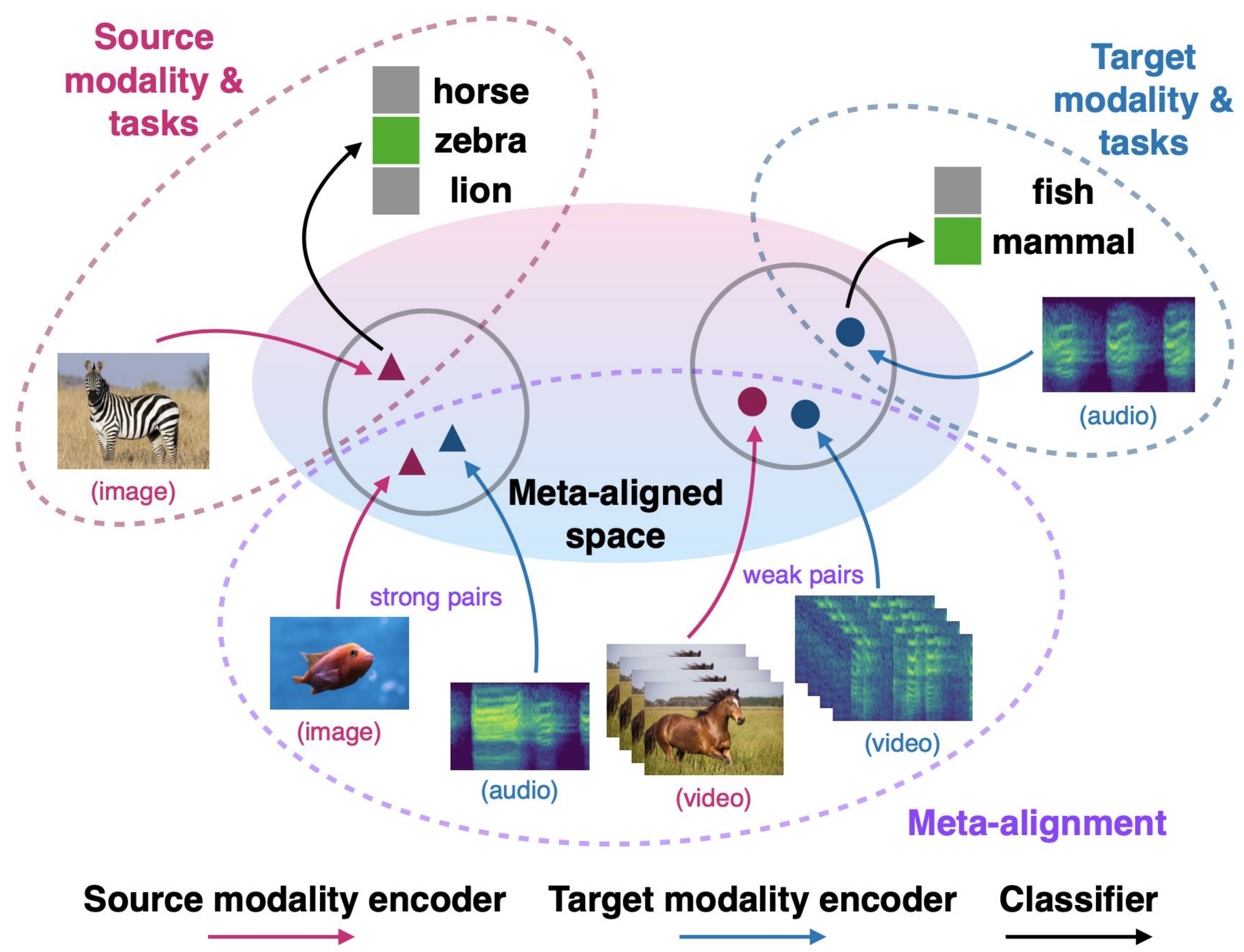
Approaches	(Meta-)Train Modality	Data	Labels
Transfer learning [3]	Target	Many	None
Unsupervised pre-training [12]	Target	Many	None
Unsupervised meta-learning [26]	Target	Many	None
Domain adaptation [59]	Target	Many	Many
Few-shot learning [17]	Target	Many	Many
Within modality + cross-modal learning [14, 56, 58, 64, 66]	Source Target	Many Many	None Many
Cross-modal generalization (ours)	Source **Target**	Many **Few**	Many **None**

Requires only a few samples and no labels in the target modality beyond those used for few-shot fine tuning.

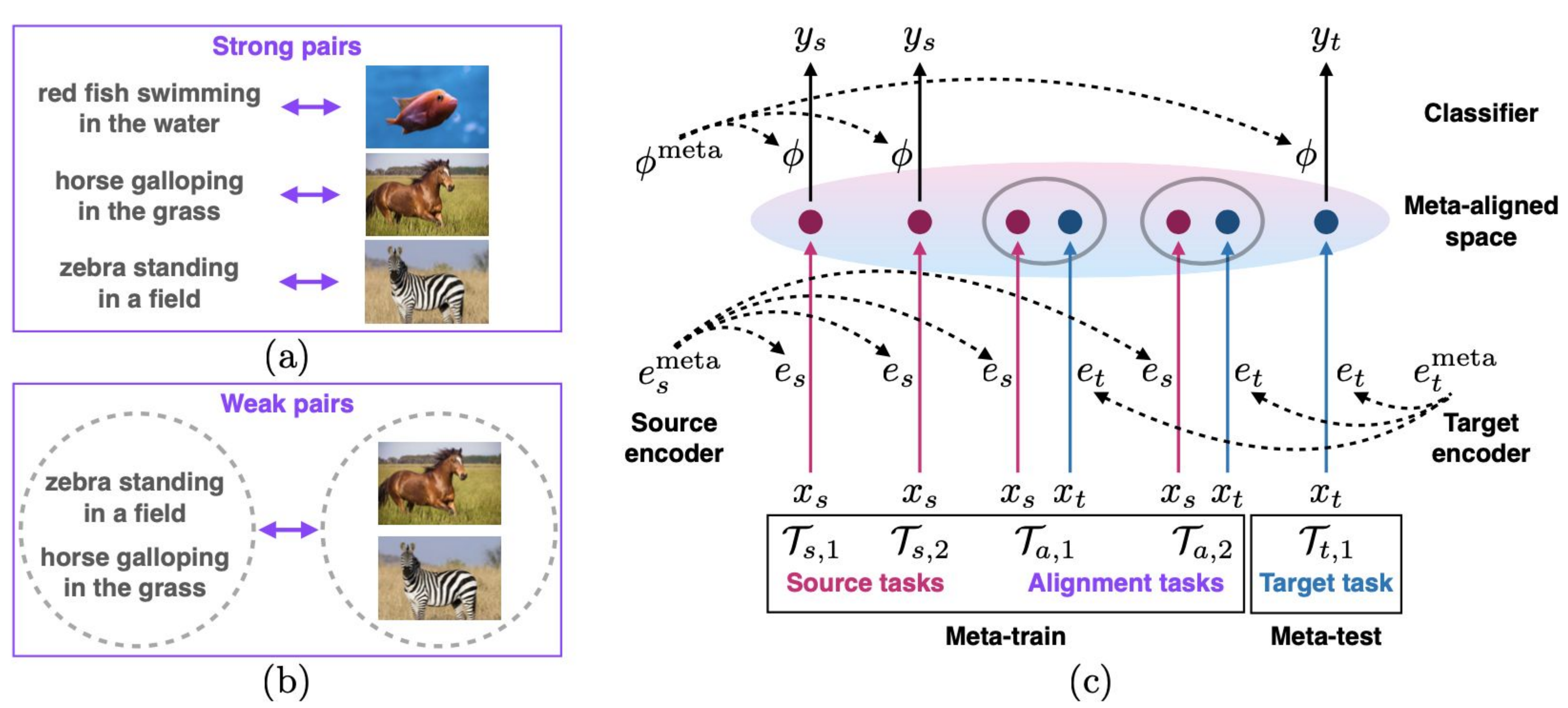


Cross-modal supervision is required under partial observability across modalities and tasks.

CroMA: Cross-modal Meta Alignment



Using Strong and Weak Cross-modal Pairs



Algorithm 1 CroMA: Cross-modal Meta-Alignment

Initialize meta-alignment encoders e_s^{meta} and e_t^{meta}, meta-classifier ϕ^{meta}.
for iteration $= 1, 2, \ldots$ **do**
Sample alignment task $\mathcal{T}_a$ with train $\mathcal{D}_{\text{train}}^{\mathcal{T}_a}$ and test data $\mathcal{D}_{\text{test}}^{\mathcal{T}_a}$ of pairs $\{x_s, x_t\}$.
Initialize $e_s := e_s^{\text{meta}}, e_t := e_t^{\text{meta}}$ and compute alignment loss (2) on train data $\mathcal{D}_{\text{train}}^{\mathcal{T}_a}$.
Compute $\widetilde{e}_s$ and $\widetilde{e}_t$ after gradient updates using alignment loss wrt e_s and e_t.
Update meta-alignment encoders $e_s^{\text{meta}} \leftarrow e_s^{\text{meta}} + \epsilon(\widetilde{e}_s - e_s^{\text{meta}})$, $e_t^{\text{meta}} \leftarrow e_t^{\text{meta}} + \epsilon(\widetilde{e}_t - e_t^{\text{meta}})$.
Sample source modality task $\mathcal{T}_s$ with train $\mathcal{D}_{\text{train}}^{\mathcal{T}_s}$ and test data $\mathcal{D}_{\text{test}}^{\mathcal{T}_s}$ of pairs $\{x_s, y_s\}$.
Initialize $\phi := \phi^{\text{meta}}$ and compute classification loss on train data $\mathcal{D}_{\text{train}}^{\mathcal{T}_s}$.
Compute $\widetilde{\phi}$ after gradient updates using classification loss wrt ϕ.
Update meta-classifier $\phi^{\text{meta}} \leftarrow \phi^{\text{meta}} + \epsilon(\widetilde{\phi} - \phi^{\text{meta}})$.

Results on Cross-modal Generalization

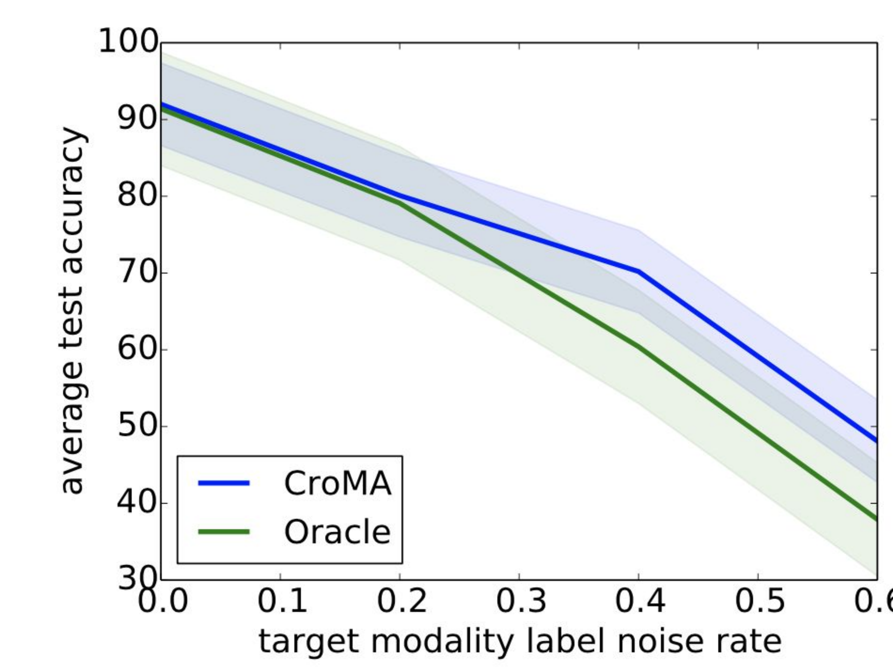
Task	Type	Approach	1-Shot	5-Shot	10-Shot	#Target (labels)
Text (Yummly) ↓ Image (Yummly)	Unimodal	Pre-training [3, 12]	33.1 ± 2.8	36.4 ± 3.5	49.0 ± 3.8	0(0)
		Unsup. meta-learning [26] (reconstruct)	37.4 ± 0.6	41.7 ± 3.7	49.0 ± 1.0	5131(0)
	Cross-modal	Align + Classify [10, 24, 50, 59, 62]	37.1 ± 3.0	40.0 ± 2.7	47.8 ± 6.6	5131(0)
		Align + Meta Classify [53]	39.4 ± 2.5	40.0 ± 2.3	48.8 ± 7.8	5131(0)
		CroMA (ours)	$\mathbf{39.7 \pm 1.3}$	$\mathbf{47.1 \pm 3.3}$	$\mathbf{51.1 \pm 2.1}$	5131(0)
	Oracle	Within modality generalization [17, 45]	38.9 ± 2.1	42.1 ± 1.4	47.9 ± 5.6	5131(5131)

Few-shot Cross-modal Retrieval

K	Experiment	R@1 ↑	R@5 ↑	R@10 ↑	Rank ↓	Cos. ↓
	No align	1.0%	2.0%	5.5%	101	0.428
5	Align	2.0%	5.5%	8.5%	103	0.272
	CroMA	**4.0%**	**19.5%**	**39.0%**	**13**	**0.003**
	No align	0.5%	3.0%	4.5%	101	0.399
10	Align	1.5%	11.0%	18.5%	52	0.222
	CroMA	**3.5%**	**17.5%**	**35.0%**	**15**	**0.004**

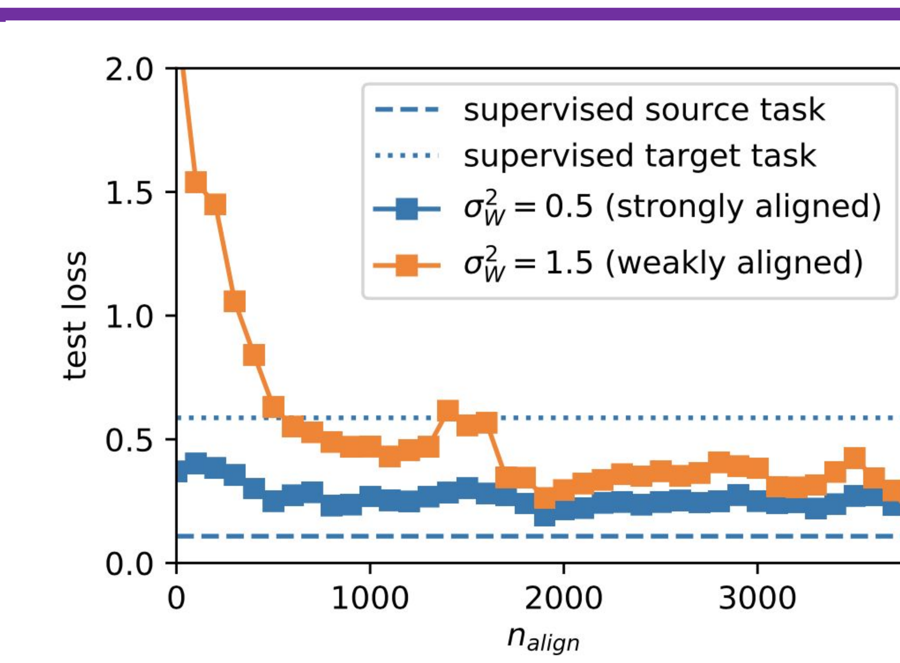
Meta-alignment is able to perform cross-modal retrieval at fine granularities.

Noisy Target Labels



More robust to noisy labels in the low-resource target modality.

A Simple Analysis: Alignment vs Supervision



1. More alignment pairs help, but at most by the performance of the high-resource source task.
2. Quality of alignment matters - less noise in alignment data is better.
3. Even weak alignment is preferable to supervised learning with enough weakly paired data.