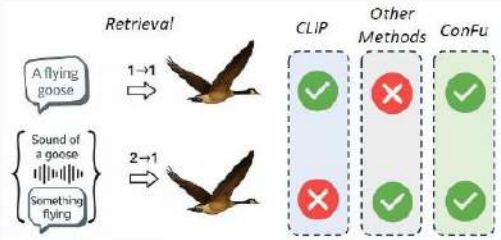
THE MORE, THE MERRIER: CONTRASTIVE FUSION FOR HIGHER-ORDER MULTIMODAL ALIGNMENT

Stefanos Koutoupis, Michaela Areti Zervou, Konstantinos Kontras, Maarten De Vos, Panagiotis Tsakalides, Grigorios Tsagkatakis

HAR.S.H.
HARdware-aware Similarity searcH

Motivation & Key Idea



Prevailing methods align only **pairs of modalities**, missing synergistic higher-order interactions.

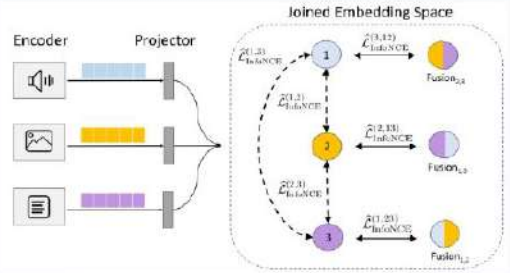
> Can a contrastive learning framework capture not only pairwise alignments but also higher-order, synergistic dependencies among modalities?

- **ConFu:** unifies pairwise (1→1) and higher-order (2→1) contrastive learning in one objective.
- **Theory:** $\mathcal{L}$ maximizes a lower bound on Total Correlation, capturing pairwise + synergistic interactions.

Key Contributions

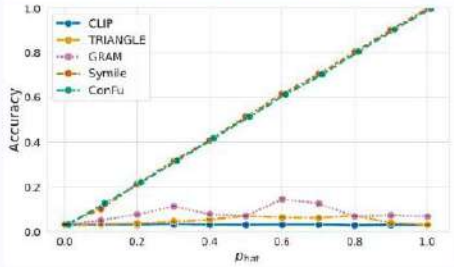
- **1→1 & 2→1 retrieval:** only method supporting both modes without requiring all modalities at inference.
- **Bird-MML dataset:** 149K image–audio–text triplets across 150 bird species for multimodal complementarity research.
- **Robustness:** shows greater stability when input involves corrupted or non-informative inputs such as missing modalities or noise.

Method



$$\mathcal{L}_{\text{fused}} = \hat{\mathcal{L}}^{(3,12)}_{\text{InfoNCE}} + \hat{\mathcal{L}}^{(2,13)}_{\text{InfoNCE}} + \hat{\mathcal{L}}^{(1,23)}_{\text{InfoNCE}}$$
$$\mathcal{L}_{\text{pair}} = \hat{\mathcal{L}}^{(1,2)}_{\text{InfoNCE}} + \hat{\mathcal{L}}^{(1,3)}_{\text{InfoNCE}} + \hat{\mathcal{L}}^{(2,3)}_{\text{InfoNCE}}$$
$$\mathcal{L} = (1-\lambda)\,\mathcal{L}_{\text{pair}} + \lambda\,\mathcal{L}_{\text{fused}}$$

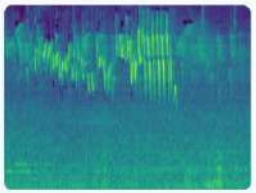
Synthetic XOR Experiment



Predict z_2 from (z_1, z_3) where pairwise MI=0.
Only synergy can solve it.

- **ConFu** ✓ strong positive trend with $\hat{\rho}$
- **Symile** ✓ also captures XOR synergy
- **GRAM/TRIANGLE** ✗ <15% even with 1024-dim embeddings
- **CLIP** ✗ ~3%, random chance level

Bird-MML dataset



Description

This red male house finch, with a distinctive black and white back stripe, enjoys a meal from a wooden bird feeder.

Results

Retrieval@10 for Target

	Target	M1			M2			M3			Mean	
	Query	M2	M3	M23	M1	M3	M13	M1	M2	M12	1→1	2→1
MOSI	Trimodal CLIP	**22.9** ± 1.9	**20.5** ± 1.6	–	**23.5** ± 3.6	**23.3** ± 1.8	–	**20.9** ± 2.0	**24.3** ± 2.1	–	**22.6**	–
	Symile [27]	-	-	16.3 ± 1.9	-	-	18.1 ± 2.4	-	-	17.1 ± 1.7	–	17.2
	GRAM [5]	6.0 ± 2.1	10.3 ± 2.9	16.3 ± 2.5	7.7 ± 2.3	0.3 ± 0.6	7.9 ± 3.4	12.2 ± 4.1	0.2 ± 0.2	12.0 ± 4.2	6.1	12.1
	TRIANGLE [6]	-	-	8.3 ± 1.3	-	-	4.9 ± 0.7	-	-	8.8 ± 1.4	–	7.3
	ConFu	21.0 ± 1.7	16.4 ± 2.3	**16.7** ± 1.3	19.2 ± 1.4	21.0 ± 2.1	**21.6** ± 1.0	16.1 ± 1.7	23.5 ± 2.7	**20.5** ± 2.0	19.5	**19.6**
UR-FUNNY	Trimodal CLIP	**3.7** ± 0.6	**4.0** ± 0.8	–	**3.8** ± 0.5	**16.2** ± 1.6	–	**4.0** ± 0.1	**16.8** ± 1.2	–	**8.1**	–
	Symile [27]	-	-	3.7 ± 0.5	-	-	15.4 ± 0.9	-	-	16.5 ± 0.6	–	11.9
	GRAM [5]	3.2 ± 0.5	3.0 ± 0.9	3.9 ± 0.8	3.2 ± 3.9	0.1 ± 0.2	3.1 ± 0.4	3.9 ± 0.4	0.1 ± 0.0	3.3 ± 0.5	2.3	3.4
	TRIANGLE [6]	-	-	**4.2** ± 0.4	-	-	3.3 ± 0.7	-	-	3.6 ± 0.9	–	3.7
	ConFu	3.2 ± 0.3	3.5 ± 0.5	3.6 ± 0.5	3.5 ± 0.2	15.1 ± 0.8	**16.9** ± 0.9	3.5 ± 0.4	15.6 ± 0.8	**20.3** ± 1.1	7.4	**13.6**
MUSTARD	Trimodal CLIP	70.7 ± 4.1	29.1 ± 3.6	–	70.5 ± 4.3	26.5 ± 3.2	–	30.4 ± 3.4	24.6 ± 2.8	–	42.0	–
	Symile [27]	-	-	61.5 ± 5.3	-	-	57.0 ± 3.0	-	-	21.3 ± 4.4	–	46.6
	GRAM [5]	61.7 ± 8.0	24.8 ± 5.3	**81.0** ± 4.1	67.2 ± 5.4	24.8 ± 3.3	67.8 ± 5.4	27.9 ± 8.7	5.1 ± 2.1	31.2 ± 9.5	35.2	60.0
	TRIANGLE [6]	-	-	68.7 ± 4.3	-	-	59.9 ± 3.7	-	-	21.9 ± 5.6	–	50.2
	ConFu	**73.8** ± 3.2	**33.6** ± 3.1	79.6 ± 3.6	**74.4** ± 2.7	**28.1** ± 3.2	**74.6** ± 4.1	**33.9** ± 4.2	**28.0** ± 2.9	**33.5** ± 3.2	**45.3**	**62.6**

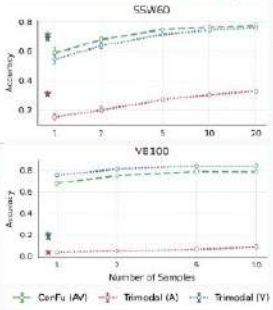
Zero-shot Classification on AV-MNIST

Method	A	V	A+V
CLIP	**41.1** ± 0.3	63.0 ± 0.4	-
Tri-CLIP	40.0 ± 0.4	62.5 ± 0.4	-
Symile [27]	-	-	70.9 ± 0.1
GRAM [5]	9.8 ± 0.4	63.9 ± 0.8	64.4 ± 0.5
TRIANGLE [6]	-	-	64.9 ± 0.1
ConFu	39.7 ± 0.1	**64.6** ± 0.3	**71.2** ± 0.2

Zero-shot Brid Classification

	SSW60 Acc. (%)			VB100 Acc. (%)		
Method	A	V	A+V	A	V	A+V
CLIP	29.9	70.1	–	4.2	20.6	–
Tri-CLIP	31.1	69.0	–	3.9	**20.7**	–
Symile [27]	-	-	60.2	-	-	13.4
TRIANGLE [6]	-	-	64.1	-	-	12.1
GRAM [5]	0.7	66.6	56.9	1.3	13.7	8.0
ConFu	30.3	69.4	**71.4**	3.4	19.3	18.1

Few-shot Linear Probing



Accuracy under Gaussian Noise

	10dB SNR		15dB SNR		20dB SNR	
Method	A deg.	V deg.	A deg.	V deg.	A deg.	V deg.
Tri-CLIP (V)	60.0	5.5	60.0	12.2	60.0	35.0
Tri-CLIP (A)	29.0	30.1	29.1	31.1	30.3	31.1
Symile [27]	58.6	21.2	59.5	27.9	60.0	46.9
GRAM [5]	58.9	4.0	58.3	8.04	58.1	25.9
TRIANGLE [6]	63.9	5.4	64.0	7.24	64.0	26.5
ConFu	71.2	30.2	71.4	33.1	71.5	45.4

Multimodal Competition

Research funded by project HAR.S.H. (project no. ΥΠ3ΤΑ-0560901), which is carried out within the framework of the National Recovery and Resilience Plan "Greece 2.0" with funding from the European Union – NextGenerationEU.