Extending Multi-Modal Contrastive Representations

Zehan Wang, Ziang Zhang, Luping Liu, Yang Zhao, Haifeng Huang, Tao Jin, Zhou Zhao

Duke B&B

January 17, 2025

Presented by Zigui Wang

Connecting Multi-modal Contrastive Representations

Zehan Wang, Yang Zhao, Xize Chen, Haifeng Huang, Jiageng Liu, Li Tang, Linjun Li, Yongqi Wang, Aoxiong Yin, Ziang Zhang, Zhou Zhao

Duke B&B

January 17, 2025

Introduction

Challenges

- Traditional multi-modal contrastive learning methods rely on large-scale, high-quality paired data (e.g., text-image or audio-visual pairs), which are costly and impractical to obtain for many modality combinations.

Inspiration

- Modality pairs with little direct paired data often have a large number of paired dated with the same intermediate modality. (Audio-Text-Visual)
- With regard to the overlapping modality, its representations in two MCRs are just different data views sharing the same inherent semantics. So we can take them as positive pairs to connect different MCRs. As modalities within each MCR are semantically aligned, the connections built from overlapping modalities can also be applied to non-overlapping modalities.

Purpose

- Propose a **paired-data-free** and training-efficient method for MCR learning.

Background: Multi-modal Contrastive Learning

Basic ideas: Map the multi-modal data to a representation space where the similarities of positive pairs are maximized, similarities of negative pairs are minimized.

Example: Given N paired instances from two different modalities, we map the i^{th} pair to L2-normalized embeddings x_i and z_i via two encoders. Multi-modal contrastive learning aims to maximize the cosine similarity between x_i and z_i and minimize the cosine similarity between x_i and z_j where $i \neq j$. The contrastive loss can be formulated as:

$$InfoNCE(x, z) = -\frac{1}{2}\frac{1}{N}\sum_{i=1}^{N}[log\frac{exp(sim(x_i, z_i)/\tau)}{\sum_{j=1}^{N} exp(sim(x_i, z_j)/\tau)} + log\frac{exp(sim(z_i, x_i)/\tau)}{\sum_{j=1}^{N} exp(sim(z_i, x_j)/\tau)}]$$

Methods: C-MCR Diagram

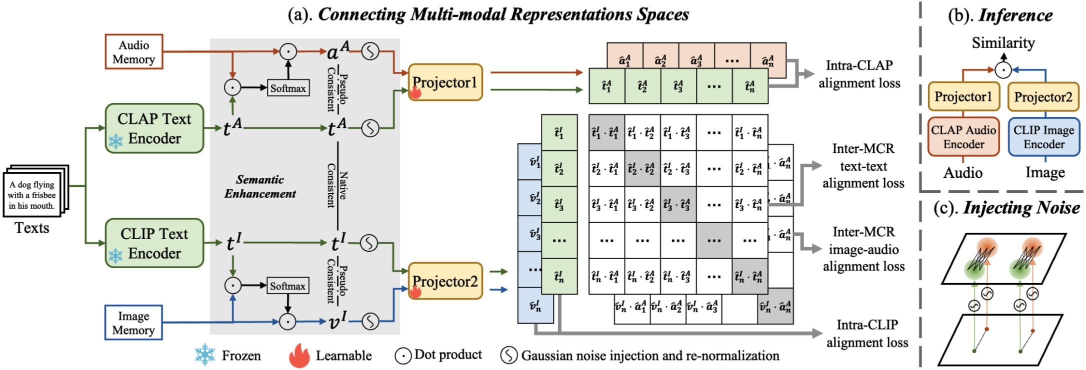
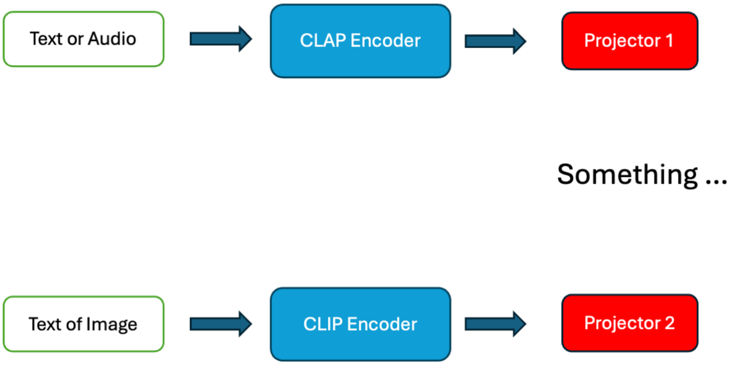


Figure: C-MCR model diagram

Methods: C-MCR Diagram



Method: C-MCR Problem Formulation

Problem Formulation

- Three modalities, T:Text, A:Audio, and V:Visual.
- Try to combine CLIP (Text vs Visual) and CLAP (Text vs Audio).

Notation

- For text inputs, the embeddings obtained by CLIP and CLAP encoder can be denoted as $t^I \in R^c$ and $t^A \in R^d$ respectively.
- All audio embeddings learned from CLAP are denoted $A = (a_1, a_2, \cdots, a_M)$.
- All image embeddings learned from CLIP are denoted $V = (v_1, v_2, \cdots, v_N)$.
- Two projectors are denoted as $f_1(\cdot)$ and $f_2(\cdot)$.

Challenges

- There is no pair information about audio and visual data.
- Embeddings in CLIP/CLAP spaces are incapable of comprehensively reflecting all the semantic information of the input, and this loss of meaning would be inherited and amplified, thereby compromising the robustness of the connection.
- MCR spaces exhibit a modality gap phenomenon, i.e., the embeddings of different modalities are located in two completely separate regions, which may let model struggle for tasks that require cross-modal understanding.

Method: Text-centric Pseudo Pair

Considering i^{th} text embeddings t_i^I and t_i^A, we can generate image embeddings v_i^I and audio embeddings a_i^A that are similar/paired to i^{th} text.

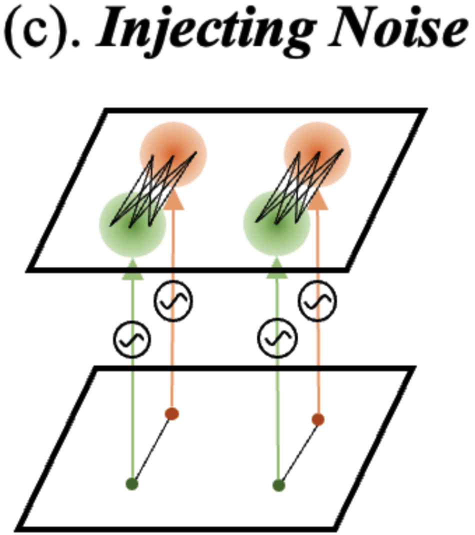
$$
\begin{aligned}
v_i^I &= \sum_{k=1}^{N} \frac{exp(sim(t_i^I, v_k)/\tau_1)}{\sum_{j=1}^{N} exp(sim(t_i^I, v_j)/\tau_1} \times v_k && (1)\\
&= softmax((t_i^I \cdot V^I)/\tau_1) \times (V^I)^T && (2)\\
a_i^A &= \sum_{k=1}^{M} \frac{exp(sim(t_i^A, a_k)/\tau_1)}{\sum_{j=1}^{M} exp(sim(t_i^A, a_j)/\tau_1} \times a_k && (3)\\
&= softmax((t_i^A \cdot T^A)/\tau_1) \times (A^A)^T && (4)
\end{aligned}
$$

Inter-modality Semantic Consistency: By dynamically absorbing information from memories based on semantic similarity to the text embeddings t_i^I and t_i^A, we can generate more diverse and accurate semantically-consistent embeddings v_i^I and a_i^A.

Method: Injecting Noise

Challenge: The semantics in the original input data are often complex, and some information is inevitably lost when encoding it into the MCR space. When connecting and aligning existing representation spaces, this loss and bias of meaning will be inherited and amplified, affecting the robustness of alignment.

$$\tilde{t}^I = \text{Normalize}(t^I + \theta_1)$$
$$\tilde{v}^I = \text{Normalize}(v^I + \theta_2)$$
$$\tilde{t}^A = \text{Normalize}(t^A + \theta_3)$$
$$\tilde{a}^A = \text{Normalize}(a^A + \theta_4)$$



Method: C-MCR Diagram

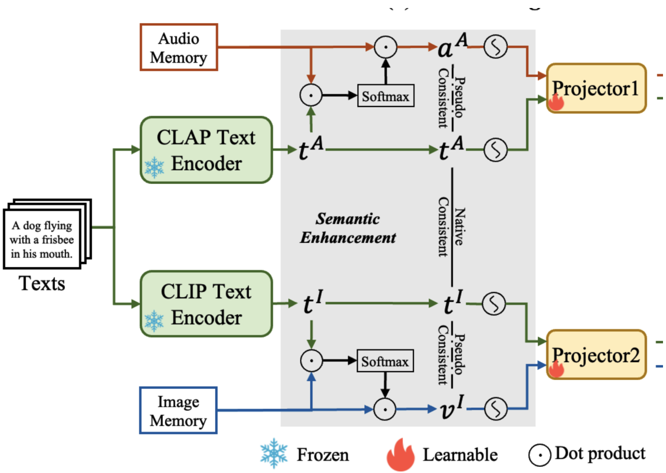


Figure: C-MCR model diagram

Method: Training Objective

Inter-modality/Across modality Alignment: Aim to establish the connection between 2 MCRs. Ensure that embeddings with similar meanings from **2 MCR spaces** are closed with each other in new space.

Intra-modality/Within modality Alignment: Aim to close the modality gap within an **single MCR** space to ensure embeddings with similar semantics (across different modalities) are placed closer together.

Method: Inter-MCR Alignment Loss

Recall we project the embeddings from CLIP and CLAP to a new shared space via two projectors $f_1(\cdot)$ and $f_2(\cdot)$.

$$\hat{t}^I = f_1(\tilde{t}^I); \hat{v}^I = f_1(\tilde{v}^I); \hat{t}^A = f_2(\tilde{t}^A); \hat{a}^A = f_2(\tilde{a}^A)$$

So the across MCR spaces data are $\hat{t}^A$ and $\hat{t}^I$, $\hat{v}^I$ and $\hat{a}^A$

$$L_{ttc} = -\frac{1}{2}\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{\exp(\mathrm{sim}(\hat{\mathbf{t}}_i^I, \hat{\mathbf{t}}_i^A)/\tau_2)}{\sum_{j=1}^{B}\exp(\mathrm{sim}(\hat{\mathbf{t}}_i^I, \hat{\mathbf{t}}_j^A)/\tau_2)} + \log\frac{\exp(\mathrm{sim}(\hat{\mathbf{t}}_i^A, \hat{\mathbf{t}}_i^I)/\tau_2)}{\sum_{j=1}^{B}\exp(\mathrm{sim}(\hat{\mathbf{t}}_i^A, \hat{\mathbf{t}}_j^I)/\tau_2)}\right]$$

$$L_{avc} = -\frac{1}{2}\frac{1}{B}\sum_{i=1}^{B}\left[\log\frac{\exp(\mathrm{sim}(\hat{\mathbf{v}}_i^I, \hat{\mathbf{a}}_i^A)/\tau_3)}{\sum_{j=1}^{B}\exp(\mathrm{sim}(\hat{\mathbf{v}}_i^I, \hat{\mathbf{a}}_j^A)/\tau_3)} + \log\frac{\exp(\mathrm{sim}(\hat{\mathbf{a}}_i^A, \hat{\mathbf{v}}_i^I)/\tau_3)}{\sum_{j=1}^{B}\exp(\mathrm{sim}(\hat{\mathbf{a}}_i^A, \hat{\mathbf{v}}_j^I)/\tau_3)}\right]$$

The inter-MCR alignment loss is defined as

$$L_{inter} = L_{ttc} + L_{avc}$$

Method: Intra-MCR Alignment Loss

From the paper *Mind the Gap: Understanding the Modality Gap in Multi-modal Contrastive Representation Learning*, the repulsive term in contrastive preserves the modality gap.

$$-\log \frac{\exp(\mathrm{sim}(\mathbf{x}_i, \mathbf{z}_i)/\tau)}{\sum_{j=1}^{N} \exp(\mathrm{sim}(\mathbf{x}_i, \mathbf{z}_j)/\tau)} = \underbrace{-\,\mathrm{sim}(\mathbf{x}_i, \mathbf{z}_i)/\tau}_{\textit{pull positive close}} + \underbrace{\log \sum_{j=1}^{N} \exp(\mathrm{sim}(\mathbf{x}_i, \mathbf{z}_j)/\tau)}_{\textit{push negative away}}$$

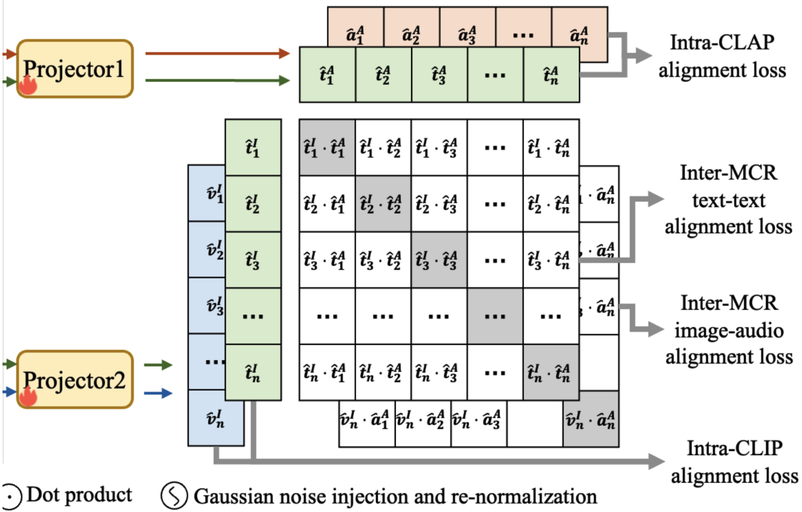
So the Intra-MCR alignment loss can be wrtten as:

$$L_{intra} = \frac{1}{2}\frac{1}{B}\sum_{i=1}^{B}(\left\|\hat{t}_i{}^I - \hat{v}_i{}^I\right\|_2 + \left\|\hat{t}_i{}^A - \hat{a}_i{}^A\right\|_2)$$

And the overall loss is

$$L = L_{inter} + \lambda L_{intra}$$

Method: C-MCR Diagram



Why We Need Extending Multimodal Contrastive Representation (Ex-MCR)?

- **Author**: C-MCR mainly focuses on learning a new space for the two non-overlapping modalities, while the original modality alignments in powerful pre-trained MCRs are forgotten. As a result of the decline of original alignment, C-MCR faces challenges in concurrently establishing connections among three or more MCRs. Therefore, C-MCR can not be used to flexibly learn a shared contrastive representation space for more than three modalities.
- **My Interpretation**: They apply C-MCR to more than 3 modalities and the results are bad.

Ex-MCR Improvement Compared to C-MCR

- **Architecture**: Instead of mapping MCRs to new space, extending one MCR space (leaf-MCR) to another fixed MCR space (called base MCR).
- **Training data**: C-MCR only uses intermediate modality-centric data pairs (Text). Ex-MCR can extract various modality-centric poseudo data pairs.
- **Learning objective**: Employ a dense contrastive loss on pseudo-pairs between all possible modalities pairs.

Method: Ex-MCR Diagram

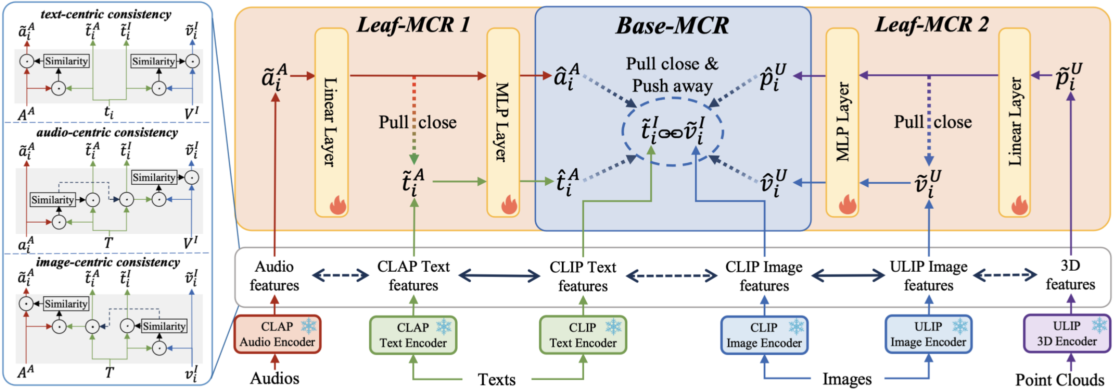


Figure: Ex-MCR model diagram

Various modality centric data

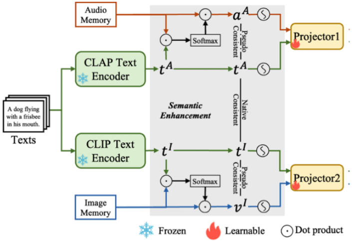
Text centric Data Considering i^{th} text embeddings $\tilde{t_i}^I$ and $\tilde{t_i}^A$, we can generate image embeddings $\tilde{v_i}^I$ and audio embeddings $\tilde{a_i}^A$ that are similar/paired to i^{th} text.

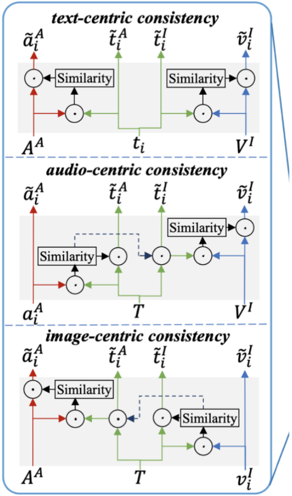
$$\tilde{t_i}^A = t_i^A ; \tilde{t_i}^I = t_i^I$$
$$\tilde{v_i}^I = softmax((\tilde{t_i}^I \cdot V^I)/\tau_1) \times (V^I)^T$$
$$\tilde{a_i}^A = softmax((\tilde{t_i}^A \cdot T^A)/\tau_1) \times (A^A)^T$$

Audio centric Data Considering i^{th} audio embeddings $\tilde{a_i}^A$, we can generate image embeddings v_i^I and text embeddings $\tilde{t_i}^A$ and $\tilde{t_i}^I$ that are similar/paired to i^{th} text.

$$\tilde{a_i}^A = a_i^A ;$$
$$\tilde{t_i}^A = softmax((a_i^A \cdot T^A)/\tau_1) \cdot (T^A)^T$$
$$\tilde{t_i}^I = softmax((a_i^A \cdot T^A)/\tau_1) \cdot (T^I)^T$$
$$\tilde{v_i}^I = softmax((\tilde{t_i}^I \cdot V^I)/\tau_1) \cdot (V^I)^T$$

Method: Ex-MCR Improvement





Intra-MCR Alignment Loss

Linear Layer: $f_l(\cdot)$

$$L_{intra} = \frac{1}{2}\frac{1}{B}\sum_{i=1}^{B}\left\| f_l(\tilde{a}_i{}^A) - \tilde{t}_i{}^A \right\|_2$$

Next: The shared MLP $f_m(\cdot)$ are employed to map both audio and text embeddings of CLAP space to the CLIP space, which can be expressed as:

$$\hat{a}_i{}^A = f_m(f_l(\tilde{a}_i{}^A))$$

$$\hat{t}_i{}^A = f_m(t_i^A)$$

Inter-MCR Alignment Loss

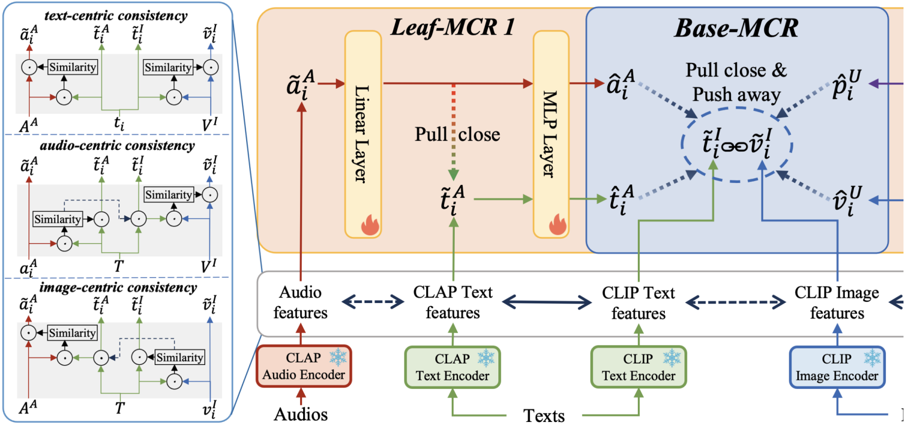
Recall we have pseudo paired data $\hat{a}_i{}^A, \hat{t}_i{}^A$ from leaf MCR CLAP, $\tilde{t}_i{}^I, \tilde{v}_i{}^I$ from base CLIP.

$$L_{avc} = InfoNCE(\hat{a}^A, \tilde{v}^I); L_{tvc} = InfoNCE(\hat{t}^A, \tilde{v}^I)$$
$$L_{atc} = InfoNCE(\hat{a}^A, \tilde{t}^I); L_{ttc} = InfoNCE(\hat{t}^A, \tilde{t}^I)$$

Therefore, overall loss is defined as:

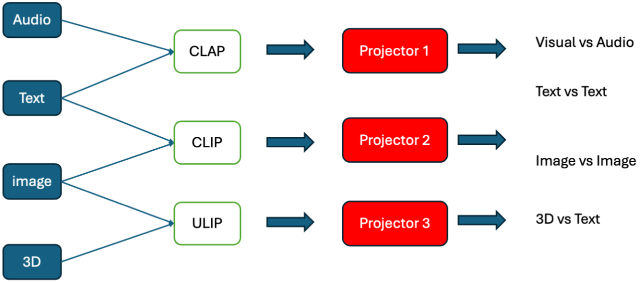
$$L = \lambda L_{intra} + \frac{1}{4}(L_{avc} + L_{atc} + L_{tvc} + L_{ttc})$$

Method: Ex-MCR Improvement



Method: Ex-MCR Improvement

As we can see, L_{atc} (Audio vs Text) and L_{tvc} (Visual vs Text) are not included.



Experimental Results 1 : Audio-Visual-Text Experiments

Author employed zero-shot audio-image,audio-text, and image-text retrieval tasks to evaluate the audio-image-text representations of Ex-MCR of extending CLAP to CLIP.

Table 1: Results of audio-visual-text experiments. The best results are **bolded**.

Method	Audio-Image FlickrNet		Audio-Image AVE		Audio-Image VGGSS		Audio-Text AudioCaps		Image-Text COCO	
	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5
CLAP	-	-	-	-	-	-	21.98	35.23	-	-
CLIP	-	-	-	-	-	-	-	-	44.57	57.62
AudioCLIP	3.81	4.91	2.33	2.65	3.10	3.94	2.23	2.68	20.14	27.42
WAV2CLIP	2.77	3.41	3.48	4.23	7.42	10.47	0.88	0.99	44.57	57.62
C-MCR	4.74	**5.97**	4.21	4.91	5.95	7.69	9.50	13.62	24.56	33.83
Ex-MCR	**4.94**	5.95	**4.46**	**4.93**	**6.39**	**8.12**	**11.19**	**16.65**	**44.57**	**57.62**

Experimental Results 2 : 3D-Visual-Text Experiments

Author employed zero-shot 3D-object(text), 3D-image, and image-text retrieval tasks to evaluate the 3D-image-text representations of Ex-MCR of extending ULIP to CLIP.

Table 2: Results of 3d-visual-text experiments.

Method	3D-Text ModelNet40			3D-Image Objaverse-LVIS			Image-Text COCO		
	Acc@1	Acc@3	Acc@5	mAP	R@1	R@5	mAP	R@1	R@5
CLIP	-	-	-	-	-	-	44.57	32.58	57.62
ULIP	60.40	79.00	84.40	3.54	1.45	4.51	34.42	22.92	46.33
ULIP v2	**73.06**	86.39	91.50	**11.41**	**6.00**	**15.63**	34.42	22.92	46.33
C-MCR	64.90	87.00	92.80	3.84	1.36	4.80	24.23	14.34	33.19
Ex-MCR	66.53	**87.88**	**93.60**	6.23	2.54	8.25	**44.57**	**32.58**	**57.62**

Experimental Results 3 : Ablation Study

Author employed zero-shot 3D-object(text), 3D-image, and image-text retrieval tasks to evaluate the 3D-image-text representations of Ex-MCR of extending ULIP to CLIP.

Table 3: Data modality-centric. A, I, and T represent audio-centric, image-centric, and text-centric data, respectively.

	AVE	AudioCaps
A	4.10	11.11
I	3.41	5.54
T	4.17	9.89
A+I	4.11	11.09
A+T	4.12	10.88
I+T	4.05	8.39
A+I+T	**4.46**	**11.19**

Table 4: Alignment objective. A-T, T-T, A-V, and T-V represent the alignment objective between audio-text, text-text, audio-image, and text-image, respectively.

	AVE	AudioCaps
A-T	4.00	10.82
T-T	4.15	11.30
A-V	3.97	7.49
T-V	4.18	7.68
All	**4.46**	**11.19**

Table 5: Structure of $f_1(\cdot)$

$f_1(\cdot)$	AVE	AudioCaps
Linear	**4.46**	**11.19**
1 MLP	4.16	10.25
2 MLP	4.04	9.93

Table 6: Structure of $f_m(\cdot)$

$f_m(\cdot)$	AVE	AudioCaps
Linear	3.70	11.15
1 MLP	4.15	10.53
2 MLP	**4.46**	11.19
3 MLP	4.31	**11.30**
4 MLP	4.35	11.07
5 MLP	4.42	10.93

Recommendations

Pros

- Easy to implement, do not required pair label.
- Computation efficient.
- Method Diagram figure is really good.

Cons

- Performance relies on pre-trained model.
- Have some doubt on its ability to learn over 3 modalities.