Extending Multi-modal Contrastive Representations

Ziang Zhang[†1, 2] **Zehan Wang**[†1, 2] **Luping Liu**[1] **Rongjie Huang**[1] **Xize Cheng**[1]
Zhenhui Ye[1] **Wang Lin**[1] **Huadai Liu**[1] **Haifeng Huang**[1]
Yang Zhao[1] **Tao Jin**[1] **Siqi Zheng**[3] **Zhou Zhao**[1, 2*]
[1]Zhejiang University [2]Shanghai AI Laboratory [3]Alibaba Group
{ziangzhang, wangzehan01}@zju.edu.cn

Abstract

Multi-modal contrastive representation (MCR) of more than three modalities is critical in multi-modal learning. Although recent methods showcase impressive achievements, the high dependence on large-scale, high-quality paired data and the expensive training costs limit their further development. Inspired by recent C-MCR, this paper proposes **Ex**tending **M**ultimodal **C**ontrastive **R**epresentation (Ex-MCR), a training-efficient and paired-data-free method to build unified contrastive representation for many modalities. Since C-MCR is designed to learn a new latent space for the two non-overlapping modalities and projects them onto this space, a significant amount of information from their original spaces is lost in the projection process. To address this issue, Ex-MCR proposes to extend one modality's space into the other's, rather than mapping both modalities onto a completely new space. This method effectively preserves semantic alignment in the original space. Experimentally, we extend pre-trained audio-text and 3D-image representations to the existing image-text space. Without using paired data, Ex-MCR achieves comparable performance to advanced methods on a series of audio-image-text and 3D-image-text tasks and achieves superior performance when used in parallel with data-driven methods. Moreover, semantic alignment also emerges between the extended modalities (e.g., audio and 3D). Our project page is available at https://github.com/MCR-PEFT/Ex-MCR.

1 Introduction

Multi-modal Contrastive Representation (MCR) learning aims to align inputs from diverse modalities within a shared representation space. Recently, the high-quality contrastive representations of more than three modalities attract increasing attention [1, 2, 3, 4, 5, 6, 7], and play a fundamental role in many application scenarios of multi-modal understanding [8, 9, 10, 11, 12] and generation [13, 14, 15, 16, 17, 18]. Previous methods focused on collecting a large amount of paired data for cross-modal semantic alignment. However, as the number of modalities increases, the costs associated with data preparation and model training to learn a contrastive representation space escalate significantly.

Recently, [19] introduces a novel training-efficient method, called C-MCR, for learning contrastive representations between modalities that lack paired data by mining knowledge from existing semantic-aligned spaces. It connects two pre-trained spaces onto a new shared space via overlapping modalities. Since the modalities of pre-trained spaces are intrinsically aligned, the connection learned from overlapping modalities can also be transferred to non-overlapping modalities. Experimentally, without using image-audio and 3D-text data pairs, C-MCR demonstrates advanced performance in image-audio and 3D-text downstream tasks.

*Corresponding author. †Equal Contribution

Despite the remarkable flexibility and performance of C-MCR, its broader applications are hindered by a critical limitation: C-MCR mainly focuses on learning a new space for the two non-overlapping modalities, while the modality alignments in powerful original pre-trained spaces are forgotten. As a result, C-MCR faces challenges in conducting continuous connection operations and fully utilizing all the knowledge in unified representation spaces. Therefore, it is difficult for C-MCR to build a unified embedding space, especially with more than three modalities.

This paper introduces **Ex**tending **M**ulti-modal **C**ontrastive **R**epresentations (Ex-MCR), a novel training-efficient and paired-data-free unified representation learning method with excellent modality extensibility. Ex-MCR better preserves the alignment within the original pre-trained space and enhances the overall learning pipeline to align different spaces more robustly. By inheriting and reorganizing existing knowledge of the representation space, Ex-MCR achieves low training costs and data requirements. Furthermore, when used in conjunction with large-scale pre-training methods, Ex-MCR can complementarily enhance the unified representation space. Specifically, the two important designs of Ex-MCR are discussed in detail below:

1. We extend one space (called leaf space) into another fixed space (called base space) rather than connecting two pre-trained spaces to a new space. Such a simple yet effective approach maximizes the preservation of modality alignment within base space, demonstrating great potential for augmenting existing unified space and integrating more pre-trained spaces.

2. We enhance the whole learning pipeline to promote stronger alignment across different spaces. Specifically: From a training data perspective, since another modality cannot fully represent semantic information in one modality, we treat different modalities as queries to retrieve pseudo-data pairs (so-called different mode-centric data) and combine them to form a comprehensive view of multimodal semantic alignment. From the architecture perspective, we propose a decoupled projector, which reduces interference among different optimization objectives. From the learning objective perspective, we employ a dense contrastive loss on pseudo-pairs between all possible modalities pairs, further enhancing the stability of learned alignments.

Utilizing Ex-MCR, we can flexibly align multiple leaf spaces onto the same base space without any paired data and with extremely low training costs. To evaluate the effectiveness of our Ex-MCR, we try to extend pre-trained 3D-image and audio-text spaces onto image-text space via the overlapping image and text modality, which derive unified audio-image-text-3D representations. Without using any paired data, Ex-MCR attains state-of-the-art performance results across various zero-shot tasks, including audio-visual, 3D-image, audio-text, visual-text retrieval, and 3D object classification. More importantly, semantic alignment is also observed between extended modalities (e.g., audio-3D), which highlights the potential of Ex-MCR in modality extensibility.

Our contributions can be summarized as three-fold:

(1) We propose **Ex**tending **M**ulti-modal **C**ontrastive **R**epresentations (Ex-MCR), a novel training-efficient and paired-data-free representation learning method for more than three modalities. Moreover, Ex-MCR is orthogonal and complementary to previous data-driven methods, combining both can bring an enhanced space.

(2) We comprehensively augment the entire space alignment learning pipeline from the perspectives of training data, architecture, and learning objectives. These novel designs offer valuable insights about effectively integrating knowledge within existing spaces.

(3) Leveraging pre-trained models like CLIP, CLAP, and ULIP, we extend audio and 3D to image-text space and obtain high-quality unified audio-image-text-3D representations. These representations exhibit advanced performance on a series of tasks.

2 Related Works

2.1 Multi-Modal Contrastive Representations

Multi-modal Contrastive Representations (MCR) learning aims to acquire semantically aligned cross-modal representations by pretraining the model on large-scale paired data. These aligned representations play a pivotal role in downstream comprehension and generation tasks. Inspired by the success of CLIP [20], many works try to learn contrative representations for two modalities [20, 21, 22, 23, 24]. CLIP [20] and ALIGN [25] learn shared image-text representations from million-level image-text pairs. CLAP [26, 27] learns the audio-text representation, and CAV-MAE [28] focus on acquiring shared audio-visual feature space. C-MCR [19] focuses on learning new representation space by connecting the pre-trained spaces through overlapping modality.

Apart from aligning two modalities, shared representations for more than three modalities attract increasing attention. AudioCLIP [2] and WAV2CLIP [29] train an audio encoder aligned with CLIP using audio-text-image triplets data. ULIP [3, 4] and openshape [5] construct 3D-image-text triplets data through rendering 3D mesh into 2D images and captioning images for textual description, thereby learning a corresponding 3D encoder for image-text MCR space. Furthermore, Imagebind [12] exclusively utilizes data pairs between various modalities and images to expand CLIP with multiple modal alignment encoders.

However, these methods heavily rely on large-scale, high-quality paired data collected from the internet or generated automatically and exceptionally high computational resources. Due to the lack of high-quality paired data for more modal combinations, such as audio-visual and text-3D, the extensibility of representation learning is notably constrained. Furthermore, the exceedingly high computational costs also diminish the flexibility of MCR learning.

2.2 Audio-Visual-Text and 3D-Visual-Text Learning

Audio-visual-text and 3D-visual-text learning have significant applications in multi-modal recognition [30, 31, 32, 33], localization [34, 35, 36, 37, 38, 39, 40, 41**?**], question-answer [11, 10, 42, 43, 44], and generation [45, 46, 47, 48, 49, 50]. They also play important roles in robot-related tasks such as human-machine interaction and synthetical information obtaining in complex environments [51, 52].

Previous unified spaces, such as AudioCLIP [2] and ULIP [3, 4], mainly focus on automatically collecting or generating more paired data, but they are limited by the relatively low quality of the training datasets. Imagebind [12] employed individual vision-aligned data instead of triplets but pre-training the encoders from scratch results in high computational costs. FreeBind [53] and OmniBind [54] achieve strong modality alignment by integrating representation spaces that simultaneously contain multiple instances of the same modality. These two methods mainly focus on enhancing modality alignment within existing spaces, whereas Ex-MCR is a training paradigm designed to construct new modality alignments. Our approach uses paired-free data and minimal computational resources, yet it still achieves superior performance in audio-image-text and 3D-image-text retrieval. More importantly, Ex-MCR is orthogonal to existing data-driven solutions, allowing it to be flexibly used in parallel with the large-scale pre-training unified space for even stronger performance.

3 Extending Multi-modal Contrastive Representations

3.1 Extending Rather Than Connecting

Given two pre-trained MCR spaces on modalities $(\mathcal{A}, \mathcal{B})$ and $(\mathcal{B}, \mathcal{C})$, C-MCR [19] employs two projectors to map them into a new shared space, where the alignment of different spaces can be learned from overlapping modality $\mathcal{B}$. Since each pre-trained space intrinsically contains the alignment of $(\mathcal{A}, \mathcal{B})$ and $(\mathcal{B}, \mathcal{C})$, the alignment learned from overlapping modality theoretically can be transferred to the non-overlapping modalities.

Specifically, for aligning different spaces, the embeddings of $\mathcal{B}$ are aligned in the new space, and pseudo $(\mathcal{A}, \mathcal{C})$ pairs retrieved by the same data of $\mathcal{B}$ are also aligned for a more comprehensive inter-space alignment. Moreover, the embeddings of different modalities within the same space are realigned to close the modality gap [55], which significantly enhances the transferability of learned inter-space alignment. C-MCR shows remarkable flexibility and versatility since connecting two existing spaces only requires two learnable projectors and unpaired unimodal data.

However, as C-MCR is designed to learn a new latent space for the two non-overlapping modalities ($\mathcal{A}$, $\mathcal{C}$) and projects them onto this space, a significant amount of information from their original spaces is lost in the projection process.. As a result, it faces challenges in concurrently establishing connections among three or more spaces. Therefore, C-MCR is not suitable for learning a unified representation space for more than three modalities.

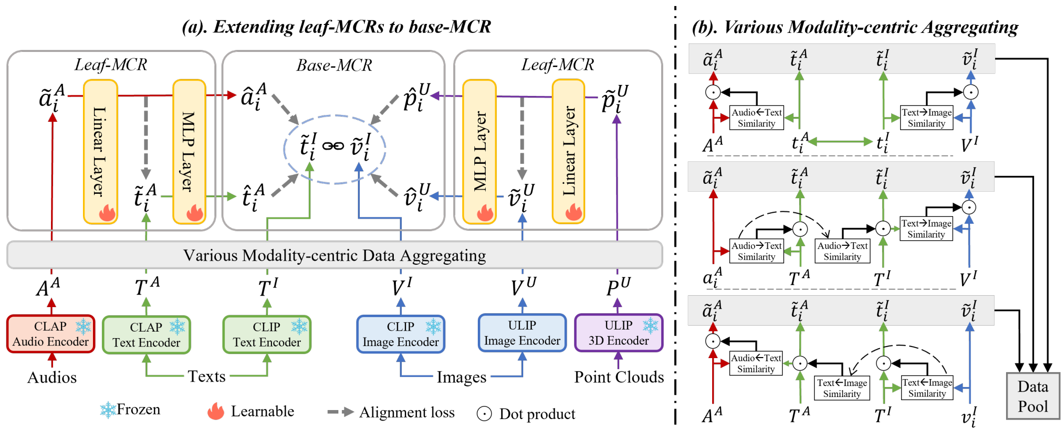

Figure 1: **The pipeline of Ex-MCR.** (a) We extend leaf spaces to base space via the overlapping modalities. The base space is frozen and the leaf spaces are aligned to the base space via projectors. (b) When extending the audio-text space to the text-image space, we iteratively use texts, audio, and images as queries to retrieve and aggregate the corresponding semantically consistent embeddings. The pseudo embedding pairs generated from different modality data are shuffled together to build the final various modality-centric data pool.

To learn unified multi-modal representations in a training-efficient and paired-data-free manner, we propose to extend one space into another space rather than connect two spaces to a new space. Considering the two spaces on modalities $(\mathcal{A}, \mathcal{B})$ and $(\mathcal{B}, \mathcal{C})$, Ex-MCR chooses one as the base space $(\mathcal{A}, \mathcal{B})$, and the other as the leaf space $(\mathcal{B}, \mathcal{C})$. In the "Extending" scheme, the base space is frozen, and we only train one projector to map leaf space to base space via the overlapping modalities $\mathcal{B}$. Specifically, we employ the native pairs of $\mathcal{B}$ and pseudo pairs generated by $\mathcal{B}$ to align leaf space to base space. Simultaneously, we close the modality gap between $(\mathcal{B}, \mathcal{C})$ modalities of leaf space, thereby facilitating more transferable alignments.

In contrast to C-MCR, Ex-MCR can conveniently expand more spaces and learn unified representation for three or more modalities. Benefiting from efficient training and no need for paired data, we can flexibly align multiple leaf spaces to the same base space. In addition to explicitly establishing alignment among modalities of leaf space and base space, semantic alignment also emerges between extended modalities. Ex-MCR employs base space as a bridge for achieving semantic alignment among modalities in multiple leaf spaces.

3.2 Enhancing Alignment Learning Pipeline

Before delving into the details of our learning pipeline, we first clarify the necessary symbols and notations. We align the ULIP (3D-image) and CLAP (audio-text) onto CLIP (image-text). As shown in Fig.1 (a), the unimodal data of audios A, texts T, images V, and 3D point clouds P are input to their corresponding encoders, and the set of the extracted feature is denoted as $\mathbf{A}^A$, $\mathbf{T}^A$, $\mathbf{T}^I$, $\mathbf{V}^I$, $\mathbf{V}^U$ and $\mathbf{P}^U$, where superscripts ^{A}, ^{I}, ^{U} indicate representation space of CLAP, CLIP, ULIP, respectively. The $\mathbf{A}^A = \{\mathbf{a}_1^A, \mathbf{a}_2^A, \dots \mathbf{a}_{n_a}^A\}$ where n_a is the number of all audio data and $\mathbf{a}_i^A$ represents the CLAP feature of i-th audio. Similarly, there are $\mathbf{t}_i^A$, $\mathbf{t}_i^I$, $\mathbf{v}_i^I$, $\mathbf{v}_i^U$, $\mathbf{o}_i^U$ in $\mathbf{T}^A$, $\mathbf{T}^I$, $\mathbf{V}^I$, $\mathbf{V}^U$, $\mathbf{P}^U$ respectively.

In Ex-MCR, freezing base space allows us to maintain the original alignment of base space but also implies that the modality gap within base space is preserved. Consequently, it becomes necessary to map the leaf space to more suitable positions within the base space. To this end, we enhance the entire alignment learning pipeline from perspectives of data, architecture, and learning objectives.

3.2.1 Various Modality-centric Data

C-MCR only uses data of overlapping modality to retrieve semantically similar embeddings of other modalities and treats these generated embeddings as pseudo pairs (we call single modality-centric data). However, it is difficult to fully represent one modality with another, and retrieved embeddings by one modality often ignore some semantics of other modalities. For example, images about "mushrooms" tend to be absent when retrieving embeddings by audio, and audio of "wind noise" may be ignored in embeddings aggregated by images. Therefore, aggregating embeddings from only a single modality struggles to capture the entire representation space of different modalities.

To tackle the above problem, we propose various modality-centric data strategy. By ensembling semantic consistent embeddings aggregated by multiple modalities, the final embeddings can reflect the representation space of different modalities in different MCRs more comprehensively. As depicted in Fig.1 (b), all modalities in two spaces are iteratively employed as queries to aggregate corresponding semantic consistent embeddings. Take aligning audio-text space to text-image space as an example, the consistent embeddings based on overlapping modality (e.g., text) are aggregated as follows:

$$\begin{aligned} \tilde{\mathbf{t}}_i^A = \mathbf{t}_i^A; \quad \tilde{\mathbf{a}}_i^A = \mathrm{softmax}((\tilde{\mathbf{t}}_i^A \cdot \mathbf{T}^A)/\tau_1) \cdot (\mathbf{A}^A)^T \\ \tilde{\mathbf{t}}_i^I = \mathbf{t}_i^I; \quad \tilde{\mathbf{v}}_i^I = \mathrm{softmax}((\tilde{\mathbf{t}}_i^I \cdot \mathbf{V}^I)/\tau_1) \cdot (\mathbf{V}^I)^T \end{aligned} \tag{1}$$

where the τ_1 is the temperature parameter of softmax, and the softmax is over all the samples in used datasets. The tilde symbols mean the features are processed to be semantically consistent. The $\tilde{\mathbf{t}}_i^A$ and $\tilde{\mathbf{t}}_i^I$ are derived from the same text data, and their semantics are natively consistent. Benefiting from the modality semantic alignment within each pre-trained space, the generated $\tilde{\mathbf{a}}_i^A$ and $\tilde{\mathbf{v}}_i^I$ are also semantically relevant to the $\tilde{\mathbf{t}}_i^A$ and $\tilde{\mathbf{t}}_i^I$.

To capture the representation space of non-overlapping modality more comprehensively, we further aggregate semantic consistent embeddings via data of non-overlapping modality (e.g., audio and image). The process of generating embeddings based on audio can be expressed as:

$$\begin{aligned} \tilde{\mathbf{a}}_i^A = \mathbf{a}_i^A; \qquad\qquad\qquad\qquad \tilde{\mathbf{v}}_i^I = \mathrm{softmax}((\tilde{\mathbf{t}}_i^I \cdot \mathbf{V}^I)/\tau_1) \cdot (\mathbf{V}^I)^T \\ \tilde{\mathbf{t}}_i^A = \mathrm{softmax}((\tilde{\mathbf{a}}_i^A \cdot \mathbf{T}^A)/\tau_1) \cdot (\mathbf{T}^A)^T; \tilde{\mathbf{t}}_i^I = \mathrm{softmax}((\tilde{\mathbf{a}}_i^A \cdot \mathbf{T}^A)/\tau_1) \cdot (\mathbf{T}^I)^T \end{aligned} \tag{2}$$

Since the embeddings of $\mathbf{T}^A$ and $\mathbf{T}^I$ of overlapping modality are one-to-one matched, the similarity weights between $\tilde{\mathbf{a}}_i^A$ and $\mathbf{T}^A$ can be naturally transferred to $\mathbf{T}^I$.

Based on the aforementioned formulas, when extending audio-text to text-image, we iteratively employ texts, audios, and images as queries to aggregate corresponding semantic consistent embeddings. During training, semantic consistent embeddings from different sources are shuffled together and the final data pool of various modality-centric data can be represented as $\{\tilde{\mathbf{a}}_i^A, \tilde{\mathbf{v}}_i^I, \tilde{\mathbf{t}}_i^A, \tilde{\mathbf{t}}_i^I\}_{i=0}^n$.

3.2.2 Decoupled Projector

The main network structure of Ex-MCR is a projector, and it serves two purposes: 1) Learning the intra-space alignment to close the modality gaps within leaf space and prompt more stable alignment between spaces. 2) Learning the inter-space alignment for extending leaf space to base space. Considering these two different purposes, we propose a decoupled projector to alleviate the potential conflict between distinct optimization objectives and explore a more reasonable mapping layer design for these two purposes. As shown in Fig.1, the projector is decoupled into a linear layer $f_l(\cdot)$ for intra-space alignment and a multi-layer perceptron layer $f_m(\cdot)$ for inter-space alignment. For extending CLAP to CLIP, we first use f_l to align $\tilde{\mathbf{a}}_i^A$ to $\tilde{\mathbf{t}}_i^A$, the loss function is defined as:

$$L_{intra} = \frac{1}{2}\frac{1}{B}\sum_{i=1}^{B} \|f_l(\tilde{\mathbf{a}}_i^A) - \tilde{\mathbf{t}}_i^A)\|_2 \tag{3}$$

With the intra-space alignment loss, $f_l(\cdot)$ learns the mapping between audio subspace and text subspace within the CLAP, thereby effectively closing the modality gap. Since the subspaces of different modalities within pre-trained spaces are very similar, linear mapping is enough to bridge the modality gap. Moreover, our experiments even found that activation layers hurt bridging the modality gap.

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

$$\hat{\mathbf{a}}_i^A = f_m(f_l(\tilde{\mathbf{a}}_i^A)); \ \ \hat{\mathbf{t}}_i^A = f_m(\tilde{\mathbf{t}}_i^A) \tag{4}$$

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

Method	Audio-Image FlickrNet		Audio-Image AVE		Audio-Image VGGSS		Audio-Text AudioCaps		Image-Text COCO	
	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
CLAP	-	-	-	-	-	-	40.25	76.21	-	-
CLIP	-	-	-	-	-	-	-	-	40.24	64.78
AudioCLIP	1.37	4.91	0.61	2.65	1.25	3.94	3.53	11.30	17.51	37.50
WAV2CLIP	0.82	3.41	0.95	4.23	2.51	10.47[2]	0.88	4.22	40.24	64.78
ImageBind	7.68	20.78	**18.00**	**40.11**	14.82	35.67	9.24	27.47	**57.28**	**79.54**
C-MCR$_{CLIP-CLAP}$	1.39	5.97	1.25	4.49	1.94	7.69	15.76	41.37	16.67	37.04
Ex-MCR-base	1.57	5.95	1.40	4.94	2.13	8.12	19.07	47.05	40.24	64.78
Ex-MCR-huge	1.80	6.16	1.89	7.36	3.26	11.77	**26.95**	**59.60**	**57.28**	**79.54**
Ex-MCR-huge + ImageBind	**7.92**	**21.26**	17.11	38.95	**15.49**	**37.55**	18.34	47.44	**57.28**	**79.54**

Table 2: Results of 3D-image-text experiments.

Method	3D-Text ModelNet40			3D-Image Objaverse-LVIS		Image-Text COCO	
	Acc@1	Acc@3	Acc@5	R@1	R@5	R@1	R@5
CLIP	-	-	-	-	-	40.24	64.78
ULIP	60.40	79.00	84.40	1.45	4.51	28.69	53.14
ULIP v2	**73.06**	86.39	91.50	**6.00**	**15.63**	28.69	53.14
C-MCR$_{CLIP-ULIP}$	64.90	87.00	92.80	1.36	4.80	24.53	48.25
Ex-MCR-base	66.53	**87.88**	**93.60**	2.54	8.25	**40.24**	**64.78**

3.2.3 Dense Alignment Objective

Since the modality gap within the base space is still preserved, a more robust learning objective is needed to map leaf space to the appropriate position in the base space. To this end, we propose to learn the alignment densely among the quadruple semantic consistent embedding pairs described in Sec.3.2.1. When extending CLAP to CLIP, the dense inter-space alignment losses are defined as:

$$\begin{aligned} L_{avc} &= \text{InfoNCE}(\hat{\mathbf{a}}^A, \tilde{\mathbf{v}}^I); \quad L_{tvc} = \text{InfoNCE}(\hat{\mathbf{t}}^A, \tilde{\mathbf{v}}^I) \\ L_{atc} &= \text{InfoNCE}(\hat{\mathbf{a}}^A, \tilde{\mathbf{t}}^I); \quad L_{ttc} = \text{InfoNCE}(\hat{\mathbf{t}}^A, \tilde{\mathbf{t}}^I) \end{aligned} \tag{5}$$

where the $\text{InfoNCE}(\cdot,\cdot)$ is the standard contrastive loss function, which is defined as:

$$\text{InfoNCE}(\mathbf{x},\mathbf{z}) = -\frac{1}{2B}\sum_{i=1}^{B}\left[\log\frac{\exp((\mathbf{x}_i\cdot\mathbf{z}_i)/\tau_2)}{\sum_{j=1}^{B}\exp((\mathbf{x}_i\cdot\mathbf{z}_j)/\tau_2)} + \log\frac{\exp((\mathbf{z}_i\cdot\mathbf{x}_i)/\tau_2)}{\sum_{j=1}^{B}\exp((\mathbf{z}_i\cdot\mathbf{x}_j)/\tau_2)}\right] \tag{6}$$

where the τ_2 is the temperature parameter. The overall loss is defined as a weighted combination of the intra-space and inter-space losses:

$$L = \lambda L_{intra} + \frac{1}{4}(L_{avc} + L_{atc} + L_{tvc} + L_{ttc}) \tag{7}$$

where λ is the hyper-parameter to balance the two terms.

Various modality-centric data 3.2.1, decoupled projector 3.2.2, and dense alignment loss 3.2.3 are also symmetrically employed to extend the 3D-image space to image-text space via images. As a result, we obtain a unified 3D-image-text-audio representation. Considering audio, text, image, and 3D point cloud inputs, we use CLAP's audio encoder, CLIP's text and image encoder, and ULIP's 3D encoder to extract corresponding features $\mathbf{a}_i^A$, $\mathbf{t}_i^I$, $\mathbf{v}_i^I$, $\mathbf{p}_i^U$. The $\mathbf{t}_i^I$, $\mathbf{v}_i^I$, $f_m^A(f_l^A(\mathbf{a}_i^A))$, $f_m^U(f_l^U(\mathbf{p}_i^U))$ are the final audio-text-image-3D unified representation learned by Ex-MCR, where the $f_m^A(\cdot), f_l^A(\cdot); f_m^U(\cdot), f_l^U(\cdot)$ are the learned projectors of CLAP and ULIP respectively.

[2]WAV2CLIP is trained on VGG-Sound. Its retrieval results on VGGSS are supervised, while other results are zero-shot.

4 Experiment

4.1 Experimental Setting

Datasets For a fair comparison, we use the same unimodal datasets to C-MCR [19] for training, totaling 2.31M texts, 1.3M images, 1.8M audio, and 0.8M 3D point clouds. More details about training datasets are provided in the Appendix.

Implementation Details For Ex-MCR-base, We employ pre-trained frozen CLIP ViT-B/32 [20], CLAP [27], and ULIPv2 (PointBERT version) [4] models. We also extend CLAP's audio encoder to OpenCLIP ViT-H [56] to build the Audio-Image-Text space Ex-MCR-huge in parallel with ImageBind [12]. The temperature τ_1 in Eq.12 for embedding aggregation is set to 0.01 following [19], while the τ_2 in Eq.6 is set to 0.05. The hyper-parameter λ in Eq.7 is set to 0.1. Following [19], we also add Gaussian noise with a variance of 0.004 to the semantic consistent embeddings described in Sec.3.2.1. The linear projector $f_l(\cdot)$ is a simple linear layer, and the MLP projector $f_m(\cdot)$ is a 2-layer MLP. We train our model with a batch size of 4096 for 36 epochs. We employ the AdamW optimizer with an initial learning rate of 1e-3 and a cosine learning rate decay strategy.

4.2 Audio-Image-Text Results

Downstream Tasks We employ zero-shot audio-image, audio-text, and image-text retrieval tasks to evaluate the audio-image-text representations of Ex-MCR. For audio-image retrieval, we conduct evaluations on Flickr-SoundNet [57], VGGSS [39], and AVE [58] datasets. Due to their small dataset sizes, we utilize all their available data, comprising 5,000, 5,000, and 4,097 samples. For audio-text retrieval, we utilize the validation set from the AudioCaps [59] dataset, which includes 964 audio samples, and for each audio, there are 5 corresponding captions for retrieval. Regarding image-text retrieval, we employ the validation set of COCO [60] dataset, consisting of 5,000 images and 25,014 text captions. We calculate the cosine similarity between modalities in representation space and use Top-1 and Top-5 metrics for performance comparison.

Performance Comparison In the upper part of Fig.1, we compare Ex-MCR-base to WAV2CLIP, AudioCLIP, and C-MCR. Notably, even without using audio-image paired data, Ex-MCR-base achieves significantly better performance over WAV2CLIP and AudioCLIP, which illustrates that Ex-MCR is a more effective representation learning method when high-quality data pairs are limited. Furthermore, compared to C-MCR, Ex-MCR not only achieves better audio-image alignment but also inherits more audio-text alignment from CLAP, fully retaining CLIP's image-text modal alignment, suggesting that Ex-MCR is generally superior to C-MCR in both establishing new Spaces and maintaining original ones. We then compare the performance of Ex-MCR-huge and data-driven alignment-building methods in the bottom half of Fig.1. Inheriting the audio-text alignment of CLAP, Ex-MCR-huge achieved better results on audio-text retrieval tasks, while ImageBind, trained directly with Audio-Image pairing data, has better audio-image performance. We were pleasantly surprised to find that using Ex-MCR and data-driven methods in parallel, with very little additional cost, can complement each other to achieve a state-of-the-art unified audio-image-text representation.

4.3 3D-Image-Text Results

Downstream Tasks To evaluate the performance of 3D-image-text space learned by extending ULIP to CLIP, we conduct a zero-shot 3D object classification task to assess the alignment between 3D and text. We also perform zero-shot 3D-image and image-text retrieval tasks to evaluate the alignment between 3D and image, as well as image and text. The zero-shot 3D object classification task is carried on the ModelNet40 [61] validation set, and we use the same prompt strategy as [4]. Regarding the zero-shot 3D-image retrieval task, we use the Objaverse-LVIS dataset [62], which includes 46,054 3D objects. Additionally, we continued to use the COCO dataset's validation set for zero-shot image-text retrieval.

It is worth noting that ULIP aligns a 3D encoder to a vision-language model called SLIP [63] (not CLIP) through 3D-image-text data. Ex-MCR only uses the aligned 3D-image representation of ULIP to extend it to a different vision-language model (i.e., CLIP) via the paired-data-free way. So we are not reproducing or refining the alignment of ULIP, but building a new alignment from scratch between the 3D representation of ULIP and CLIP.

Performance Comparison From Tab.3.2.3, we can find the following key points:

1) Even without using any 3D-text data, Ex-MCR still outperforms the advanced models (ULIP and ULIP v2) trained on 3D-text pairs in most performance metrics for 3D object classification.

2) For 3D-image retrieval, since the 3D-image space of ULIPv2 is treated as leaf space, it is reasonable that Ex-MCR-base 3D-image performance is slightly lower than ULIPv2. At the same time, the better 3D-image retrieval accuracy than ULIP and C-MCR shows that Ex-MCR effectively learns strong 3D-image alignment.

3) Ex-MCR retains the best image-text retrieval accuracy compared to these previous state-of-the-art models. The leading performance on all these tasks further demonstrates the superiority of Ex-MCR in unified contrastive representation learning.

Table 3: Various modality-centric data: We report the mAP metrics on all audio-visual and audio-text retrievals. The A, I, and T represent pseudo data derived from audio, image, and text, respectively. The "+" between A, I, and T means combining these data for training.

	FlickrNet	AVE	VGGSS	AudioCaps
A	3.94	4.10	5.47	11.11
I	3.83	3.41	4.82	5.54
T	4.85	4.17	5.72	9.89
$A+I$	4.22	4.11	6.04	11.09
$A+T$	4.63	4.12	5.88	10.88
$I+T$	4.70	4.05	5.84	8.39
$A+I+T$	**4.94**	**4.46**	**6.39**	**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. "All" means using all above alignment losses simultaneously.

	FlickrNet	AVE	VGGSS	AudioCaps
$A-T$	4.01	4.00	5.70	10.82
$T-T$	4.56	4.15	5.68	**11.30**
$A-V$	4.30	3.97	5.91	7.49
$T-V$	4.77	4.18	5.43	7.68
All	**4.94**	**4.46**	**6.39**	11.19

Table 5: Structure of $f_l(\cdot)$. "Linear" means single linear layer, and "n MLP" indicates n-layer MLP.

$f_l(\cdot)$	FlickrNet	AVE	VGGSS	AudioCaps
Linear	**4.94**	**4.46**	6.39	**11.19**
1 MLP	4.54	4.16	**6.50**	10.25
2 MLP	4.36	4.04	6.00	9.93

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

$f_m(\cdot)$	FlickrNet	AVE	VGGSS	AudioCaps
Linear	3.62	3.70	5.40	11.15
1 MLP	4.62	4.15	5.81	10.53
2 MLP	4.94	**4.46**	6.39	11.19
3 MLP	4.85	4.31	6.57	**11.30**
4 MLP	4.95	4.35	6.55	11.07
5 MLP	4.79	4.42	6.59	10.93

4.4 Emergent 3D-Audio Alignment

In this section, we study whether the semantic alignment also emerges between the extended modalities (e.g., audio and 3D). We mutually retrieve audio in AudioCaps and 3D objects in Objaverse. In Fig. 4.4 and 3, we provide visualizations of some top-5 retrieval results, and audios are described by their corresponding caption annotations. These cases effectively demonstrate the emergent semantic alignment between audio-3D in Ex-MCR space. For example, the sound of a flushing toilet and water flow can retrieve 3D objects of toilets or sinks, while a sailboat 3D object can retrieve clips containing sounds of water vessels and wind.

These exciting results demonstrate that extending ULIP and CLAP onto CLIP following our Ex-MCR methods derives a 3D-image-text-audio unified contrastive representation space. In addition to the state-of-the-art performance on all possible tasks, Ex-MCR is an extremely training-efficient and paired-data-free representation learning method, which amplifies its application value in unified multi-modal representation learning. To further support the conclusion, we provide more audio-image retrieval results and the original audio files in the supplementary material.

4.5 Ablation Studies

In this section, we analyze the main components of Ex-MCR. All experiments are conducted on extending CLAP to CLIP, and we reported the average mAP of audio-visual and audio-text retrieval, respectively. In addition, we also provide ablation results on full evaluation metrics in the Appendix.

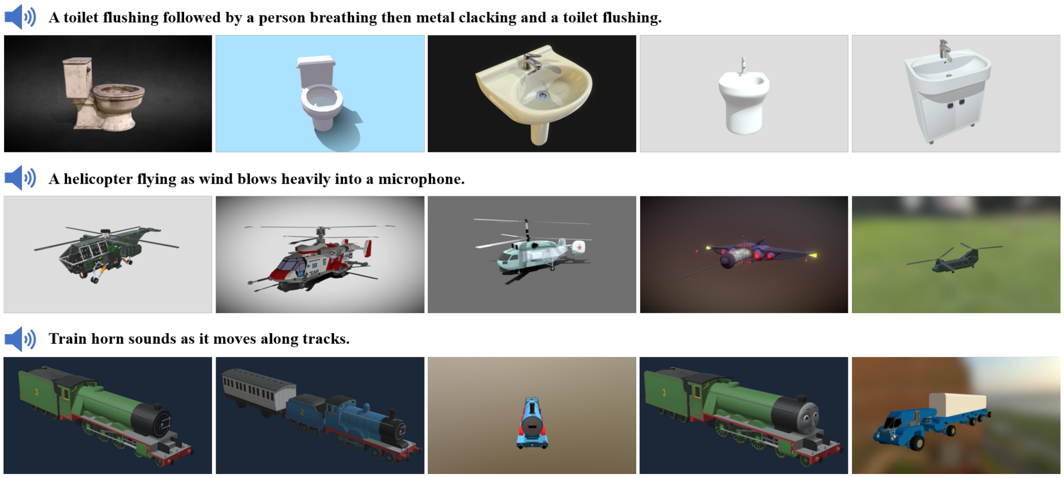

Figure 2: Visualization of Audio to 3D retrieval.

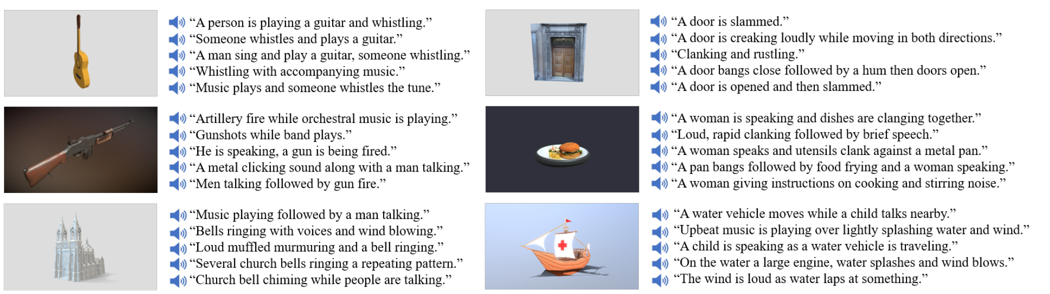

Figure 3: Visualization of 3D to Audio retrieval.

Various modality-centric data As described in Sec.3.2.1, we employ various modality-centric data to train our projectors. For investigating the effect of different modality-centric data, we ablate each modality-centric data, and the results are reported in Tab.3. The A, I, and T represent pseudo data derived from audio, image, and text respectively. Each kind of data is beneficial for audio-visual and audio-image alignment, and using all kinds of data simultaneously brings the best performance. In addition, we find that pseudo-pairs from audios are critical to the performance of audio-text retrieval, demonstrating the importance of various modality-centric data, and proving that previous single modality-centric data really can not fully reflect the audio representation space.

Dense alignment objective To analyze the impact of different alignment objectives, we train the model with each alignment objective. From the results reported in Tab.4, we find that directly aligning the pseudo audio-image or audio-text embedding pairs leads to sub-optimal audio alignment, whereas aligning spaces by overlapping text modality brings better alignment than learning alignment directly from pseudo pairs. This observation further suggests that overlapping modalities play a key pivotal role in aligning different spaces.

Structure of $f_l(\cdot)$ Tab.5 demonstrates the impact of different structures of $f_l(\cdot)$. The results prove our hypothesis: the representation structures between different modalities within one MCR space are similar, and a simple linear layer is enough to bridge the modality gap. Moreover, the activation layer of the MLP introduces non-linearity, which may disrupt the spatial structure of representations.

Structure of $f_m(\cdot)$ The ablation studies of $f_m(\cdot)$ are summarized in Tab.6. When aligning different MCR spaces, the nonlinear MLP structure with stronger expressivity is better than the simple linear layer. Besides, good results are achieved no matter how many layers of MLP, which demonstrates the robustness of our method. According to more detailed experiments in Tab.11, empirically, MLP with 2 or 3 layers achieves a good balance between expressivity and learning difficulty.

Training hyperparameters τ_2 **and** λ: The results of ablation experiments show that the performance is insensitive to the τ_2 in Eq6 and λ in Eq7. So the picked τ_2 is 0.05 which is commonly used and the picked λ is only to equal the absolute value of different loss terms. For detailed experimental results, please refer to the table in Appendix C.

5 Conclusion

This paper proposes **Ex**tending **M**ulti-modal **C**ontrastive **R**epresentations (Ex-MCR), a novel training-efficient and paired-data-free unified constrastive representation learning method for more than three modalities. Ex-MCR effectively integrates the knowledge in pre-trained spaces through overlapping modalities between these spaces. By extending ULIP and CLAP onto CLIP via the overlapping image and text modality, respectively, we derive unified and high-quality audio-image-text-3D representations. Additionally, Ex-MCR provides a new view to build unified representations. Even without using paired data, Ex-MCR still achieves competitive performance, and when combined with data-driven approaches, it complementarily enhances unified representation spaces, leading to state-of-the-art results across various tasks.

Acknowledgements

This work was supported in part by the National Key R&D Program of China under Grant No.2022ZD0162000.

References

[1] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023.

[2] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980. IEEE, 2022.

[3] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1179–1189, 2023.

[4] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023.

[5] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *arXiv preprint arXiv:2305.10764*, 2023.

[6] Deepti Hegde, Jeya Maria Jose Valanarasu, and Vishal M Patel. Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. *arXiv preprint arXiv:2303.11313*, 2023.

[7] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, and Pheng-Ann Heng. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following, 2023.

[8] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*, 2023.

[9] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*, 2023.

[10] Yang Zhao, Zhijie Lin, Daquan Zhou, Zilong Huang, Jiashi Feng, and Bingyi Kang. Bubogpt: Enabling visual grounding in multi-modal llms. *arXiv preprint arXiv:2307.08581*, 2023.

[11] Zehan Wang, Haifeng Huang, Yang Zhao, Ziang Zhang, and Zhou Zhao. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023.

[12] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

[13] Zineng Tang, Ziyi Yang, Chenguang Zhu, Michael Zeng, and Mohit Bansal. Any-to-any generation via composable diffusion, 2023.

[14] Haohe Liu, Qiao Tian, Yi Yuan, Xubo Liu, Xinhao Mei, Qiuqiang Kong, Yuping Wang, Wenwu Wang, Yuxuan Wang, and Mark D Plumbley. Audioldm 2: Learning holistic audio generation with self-supervised pretraining. *arXiv preprint arXiv:2308.05734*, 2023.

[15] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1(2):3, 2022.

[16] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022.

[17] Oran Gafni, Adam Polyak, Oron Ashual, Shelly Sheynin, Devi Parikh, and Yaniv Taigman. Make-a-scene: Scene-based text-to-image generation with human priors. In *European Conference on Computer Vision*, pages 89–106. Springer, 2022.

[18] Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023.

[19] Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Li Tang, Linjun Li, Yongqi Wang, Aoxiong Yin, Ziang Zhang, et al. Connecting multi-modal contrastive representations. *arXiv preprint arXiv:2305.14381*, 2023.

[20] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[21] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.

[22] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[23] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, Jianfeng Gao, et al. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision*, 14(3–4):163–352, 2022.

[24] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[26] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[27] Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[28] Yuan Gong, Andrew Rouditchenko, Alexander H Liu, David Harwath, Leonid Karlinsky, Hilde Kuehne, and James R Glass. Contrastive audio-visual masked autoencoder. In *The Eleventh International Conference on Learning Representations*, 2022.

[29] Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4563–4567. IEEE, 2022.

[30] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 776–780. IEEE, 2017.

[31] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 721–725. IEEE, 2020.

[32] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv preprint arXiv:1512.03012*, 2015.

[33] Angela Dai, Angel X Chang, Manolis Savva, Maciej Halber, Thomas Funkhouser, and Matthias Nießner. Scannet: Richly-annotated 3d reconstructions of indoor scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5828–5839, 2017.

[34] Dave Zhenyu Chen, Angel X Chang, and Matthias Nießner. Scanrefer: 3d object localization in rgb-d scans using natural language. In *European conference on computer vision*, pages 202–221. Springer, 2020.

[35] Panos Achlioptas, Ahmed Abdelreheem, Fei Xia, Mohamed Elhoseiny, and Leonidas Guibas. Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pages 422–440. Springer, 2020.

[36] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2928–2937, 2021.

[37] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

[38] Shentong Mo and Pedro Morgado. A closer look at weakly-supervised audio-visual source localization. *Advances in Neural Information Processing Systems*, 35:37524–37536, 2022.

[39] Honglie Chen, Weidi Xie, Triantafyllos Afouras, Arsha Nagrani, Andrea Vedaldi, and Andrew Zisserman. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16867–16876, 2021.

[40] Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. 3drp-net: 3d relative position-aware network for 3d visual grounding. *arXiv preprint arXiv:2307.13363*, 2023.

[41] Zehan Wang, Haifeng Huang, Yang Zhao, Linjun Li, Xize Cheng, Yichen Zhu, Aoxiong Yin, and Zhou Zhao. Distilling coarse-to-fine semantic matching knowledge for weakly supervised 3d visual grounding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2662–2671, 2023.

[42] Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022.

[43] Yan-Bo Lin, Yi-Lin Sung, Jie Lei, Mohit Bansal, and Gedas Bertasius. Vision transformers are parameter-efficient audio-visual learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2299–2309, 2023.

[44] Haifeng Huang, Zehan Wang, Rongjie Huang, Luping Liu, Xize Cheng, Yang Zhao, Tao Jin, and Zhou Zhao. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023.

[45] Ludan Ruan, Yiyang Ma, Huan Yang, Huiguo He, Bei Liu, Jianlong Fu, Nicholas Jing Yuan, Qin Jin, and Baining Guo. Mm-diffusion: Learning multi-modal diffusion models for joint audio and video generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10219–10228, 2023.

[46] Ben Poole, Ajay Jain, Jonathan T Barron, and Ben Mildenhall. Dreamfusion: Text-to-3d using 2d diffusion. *arXiv preprint arXiv:2209.14988*, 2022.

[47] Chen-Hsuan Lin, Jun Gao, Luming Tang, Towaki Takikawa, Xiaohui Zeng, Xun Huang, Karsten Kreis, Sanja Fidler, Ming-Yu Liu, and Tsung-Yi Lin. Magic3d: High-resolution text-to-3d content creation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 300–309, 2023.

[48] Yongqi Wang, Wenxiang Guo, Rongjie Huang, Jiawei Huang, Zehan Wang, Fuming You, Ruiqi Li, and Zhou Zhao. Frieren: Efficient video-to-audio generation with rectified flow matching. *arXiv preprint arXiv:2406.00320*, 2024.

[49] Wang Lin, Tao Jin, Wenwen Pan, Linjun Li, Xize Cheng, Ye Wang, and Zhou Zhao. Tavt: Towards transferable audio-visual text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14983–14999, 2023.

[50] Wang Lin, Tao Jin, Ye Wang, Wenwen Pan, Linjun Li, Xize Cheng, and Zhou Zhao. Exploring group video captioning with efficient relational approximation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15281–15290, 2023.

[51] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[52] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. *arXiv preprint arXiv:2302.14045*, 2023.

[53] Zehan Wang, Ziang Zhang, Xize Cheng, Rongjie Huang, Luping Liu, Zhenhui Ye, Haifeng Huang, Yang Zhao, Tao Jin, Peng Gao, et al. Freebind: Free lunch in unified multimodal space via knowledge fusion. In *Forty-first International Conference on Machine Learning*.

[54] Zehan Wang, Ziang Zhang, Hang Zhang, Luping Liu, Rongjie Huang, Xize Cheng, Hengshuang Zhao, and Zhou Zhao. Omnibind: Large-scale omni multimodal representation via binding spaces. *arXiv preprint arXiv:2407.11895*, 2024.

[55] Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning, 2022.

[56] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[57] Arda Senocak, Tae-Hyun Oh, Junsik Kim, Ming-Hsuan Yang, and In So Kweon. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4358–4366, 2018.

[58] Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *Proceedings of the European conference on computer vision (ECCV)*, pages 247–263, 2018.

[59] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132, 2019.

[60] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[61] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1912–1920, 2015.

[62] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13142–13153, 2023.

[63] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. Slip: Self-supervision meets language-image pre-training. In *European Conference on Computer Vision*, pages 529–544. Springer, 2022.

A Training Dataset

The details of our training dataset, which are mentioned in Sec.4.1, are shown below.

Text Dataset To ensure that the texts contain sufficient information for other modalities, the data of text is sourced from diverse perspectives in image-text datasets (COCO, CC3M), video-text datasets (MSRVTT, MAD), and audio-text datasets (AudioCaps, Clotho). Following [19], we select 1M texts from CC3M. There are 2.33M text samples in total. We extract their CLAP and CLIP features $\mathbf{T}^A$ and $\mathbf{T}^I$ using the CLAP and CLIP encoders, respectively.

Image Dataset For another modality in base space, Image, we utilize ImageNet1K as the data source. ImageNet1K is a large-scale image recognition dataset consisting of 1.3 million images. We extract their features to the sets $\mathbf{V}^I$, and $\mathbf{V}^U$ in CLIP and ULIP, using the CLIP Encoder and ULIP Encoder.

Audio Dataset AudioSet is a large-scale audio dataset with 2.1M audio clips from YouTube, equivalent to 5.8 thousand hours of audio and encompassing over 500 sound classes. We use the CLAP audio encoder to extract the feature set $\mathbf{A}^A$ from the audios of the training set.

3D Point Cloud Dataset For the 3D modality, we use Objaverse, the recently released and large-scale 3D objects dataset. It has approximately 800K real-world 3D objects. All 3D data are transformed into point clouds and extracted into the feature set $\mathbf{P}^U$ using the ULIP 3D encoder.

It is worth noting that we do not employ any annotations provided with the datasets mentioned above as part of our training data, which means we only use the unimodal modality of data in each dataset we selected.

B Architecture of Projectors

Table 7: Model configurations of projectors.

Module	Block	C_{in}	C_{out}
$f_1(\cdot)$	Linear	512	512
$f_m(\cdot)$	Linear	512	1024
	BatchNorm1D	1024	1024
	Relu	-	-
	Linear	1024	512
	BatchNorm1D	512	512
	Relu	-	-
	Linear	512	1024
	BatchNorm1D	1024	1024
	Relu	-	-
	Linear	1024	512
	BatchNorm1D	512	512
	Relu	-	-

The model configurations of our projectors are shown in Tab.7.

C Detailed Results of Ablation Study

As a supplement to Tab.3, Tab.4, Tab.5, and Tab.6, we provide detailed ablation experiment results on more comprehensive evaluation metrics of various datasets, as shown below.

Table 8: Detailed results of experiments on data modality-centric.

Data Perspective	FlickrNet		AVE		VGGSS		AudioCaps	
	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5
A	3.94	4.77	4.10	4.66	5.47	6.95	11.11	16.39
I	3.83	4.63	3.41	3.70	4.82	5.96	5.54	7.18
T	4.85	**5.96**	4.17	4.61	5.72	7.23	9.89	14.47
A+I	4.22	4.96	4.11	4.71	6.01	7.78	11.09	**16.91**
A+T	4.63	5.56	4.12	4.64	5.88	7.57	10.88	16.23
I+T	4.70	5.82	4.05	4.34	5.84	7.36	8.39	12.09
A+I+T	**4.94**	5.95	**4.46**	**4.93**	**6.39**	**8.12**	**11.19**	16.65

Table 9: Detailed results of experiments on alignment objective.

Objective	FlickrNet		AVE		VGGSS		AudioCaps	
	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5
A-T	4.01	4.78	4.00	4.56	5.70	7.28	10.82	15.87
T-T	4.56	5.33	4.15	4.54	5.68	6.86	**11.30**	**16.93**
A-V	4.30	5.34	3.97	4.51	5.91	7.30	7.49	10.35
T-V	4.77	**6.03**	4.18	4.92	5.43	6.93	7.68	10.36
Dense	**4.94**	5.95	**4.46**	**4.93**	**6.39**	**8.12**	11.19	16.65

Table 10: Detailed results of experiments on the structure of $f_1(\cdot)$.

$f_1(\cdot)$	FlickrNet		AVE		VGGSS		AudioCaps	
	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5
Linear	**4.94**	**5.95**	**4.46**	**4.93**	6.39	8.12	**11.19**	**16.65**
1 MLP	4.54	5.59	4.16	4.75	**6.50**	**8.54**	10.25	14.92
2 MLP	4.36	5.15	4.04	4.66	6.00	7.63	9.93	14.48

Table 11: Detailed results of experiments on the structure of $f_m(\cdot)$.

$f_m(\cdot)$	FlickrNet		AVE		VGGSS		AudioCaps	
	mAP	R@5	mAP	R@5	mAP	R@5	mAP	R@5
Linear	3.62	4.50	3.70	4.03	5.40	6.82	11.15	16.37
1 MLP	4.62	5.79	4.15	4.76	5.81	7.28	10.53	15.87
2 MLP	4.94	5.95	**4.46**	4.93	6.39	8.12	11.19	16.65
3 MLP	4.85	5.93	4.31	4.88	6.57	8.70	**11.30**	**17.10**
4 MLP	**4.95**	**6.20**	4.35	4.84	6.55	8.57	11.07	16.23
5 MLP	4.79	6.02	4.42	**5.15**	**6.59**	**8.63**	10.93	16.21

Table 12: Detailed results of experiments on the hyperparameter of τ_2.

τ_2	FlickrNet	AVE	VGGSS	AudioCaps
	R@5	R@5	R@5	R@5
0.01	4.35	5.41	9.02	61.34
0.02	4.82	5.31	9.91	62.73
0.03	5.46	6.06	10.79	62.02
0.04	5.83	6.51	11.14	61.12
0.05	6.16	7.36	11.77	59.60
0.06	6.10	6.47	11.22	57.66
0.07	6.11	6.67	10.89	56.27
0.08	6.21	6.31	10.60	55.10
0.09	6.05	6.51	10.44	55.09
0.10	5.93	6.40	10.41	53.68

Table 13: Detailed results of experiments on the hyperparameter of λ.

λ	FlickrNet	AVE	VGGSS	AudioCaps
	R@5	R@5	R@5	R@5
0.00	6.02	5.81	10.46	58.59
0.01	6.19	6.57	11.42	60.88
0.03	6.24	6.47	11.36	59.93
0.05	6.19	6.35	11.10	59.41
0.10	6.16	7.36	11.77	59.60
0.15	5.74	6.43	11.23	59.34
0.20	5.92	6.31	11.17	58.08
0.25	5.84	6.14	11.23	58.15
0.30	5.79	6.19	11.15	57.59
0.35	5.67	6.36	10.93	56.97

D Compute Resource

Collecting a group of pseudo datasets takes about 10 hours on a single 4090 while using 12GB GPU memory. The training times for projectors between two spaces are approximately 1.5 hours, on a single 4090, and it only requires 3GB of GPU memory.

E Potential Ethical Impact

This paper introduces Ex-MCR, a paired-data-free and training-efficient method for constructing a unified multimodal representation space. While this method offers flexibility in constructing a new unified representation space, its training process, which does not necessitate paired data, may inadvertently create unintended associations within the constructed representation space. The alignment of the representation space is primarily influenced by the pre-training space utilized and the unimodal data, both of which need to be restricted to prevent potential misuse for unethical applications.

F Limitation and Future Work

Currently, larger and more advanced models are continually emerging across various modalities, leading to increasingly stronger alignment in pre-trained contrastive representations. Although the experimental results indicate that Ex-MCR already demonstrates significant advantages at comparable model scales, considering its flexibility as a general paradigm, utilizing these advanced models to explore the upper limits of this learning approach would be an exciting research direction.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: See in introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: See in Appendix F.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

Justification: No need.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Code and model checkpoints are provided in supplemental material.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We are using open-source data and the code can be found in the supplemental material.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We show the experimental detail in Sec 4.1.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in the appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: The training and test sets used in the experiments in this paper are deterministic, and the results of the retrieval experiments are also discrete and deterministic, so no error analysis is performed.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)

- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: A description of the computational resources can be found in the appendix D.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: A description of the broader impacts can be found in appendix E.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: The paper poses no such risks.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All the creators or original owners of assets used in the paper are properly credited and the license and terms of use are explicitly mentioned and properly respected.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide detailed documentation in the supplementary material.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.