TANGENTBIND: UNLOCKING THE POTENTIAL OF EMERGENT ALIGNMENT IN MULTIMODAL MODEL

Anonymous authors
Paper under double-blind review

ABSTRACT

Improving the alignment of modalities has proven effective across various downstream tasks in multimodal models. Currently, modality alignment follows two main research directions: aligning all modalities simultaneously or binding the others by aligning to a core modality. The first ensures direct alignment, but it is difficult to extend to new modalities. The second is scalable but weak in emergent ability due to needing more direct inter-modality alignment. To address these problems, we propose the **TangentBind**. Specifically, we first align all modalities to a core modality, e.g., image or text. Then, we introduce a generative network that generates the embeddings of the second modality, e.g., text or image, based on the core modality embedding. Thirdly, other modalities, such as audio, are aligned to the core modality and generative embedding, improving emergent ability while retaining alignment with the core modality. During training, in addition to infoNCE, the Tangent Term is introduced to align the new modalities with the generated embeddings. This addresses accuracy issues caused by using generated vectors as representations for modalities. With VISION and TEXT as the core modality, our experiments include other modalities such as AUDIO, DEPTH, and INFRARED. Eventually, our experiments show that the emergent ability of TangentBind significantly outperforms the original benchmark on 9 datasets.

1 INTRODUCTION

Unifying multimodal representation aims to learn a shared semantic representation space across various modalities, such as audio, RGB images, text, depth images, and heatmaps(Wang et al., 2023a; Girdhar et al., 2023; Guzhov et al., 2021; Wu et al., 2021; 2022; Liu et al., 2023). A unified multimodal space is essential as a critical foundation for multimodal understanding and generation (Karpathy et al., 2014; Mithun et al., 2018; Lu, 2023). However, the previous multimodal requires all modes to coexist, which is challenging and labor-intensive(Guzhov et al., 2021; Radford et al., 2021; Moon et al., 2022).

Due to the limitations of existing datasets, recent approaches only utilize pairs of modalities or a few visual modalities. Consequently, the resulting embeddings are confined to the modality pairs used during training and lack alignment for other modalities (Zhu et al., 2024). Recent works have introduced more flexible alignment strategies to address this issue.(Gao et al., 2024; Wang et al., 2024a; Zhu et al., 2024; Dhakal et al., 2024; Wang et al., 2024b; Lyu et al., 2024). Among them, ImageBind was the first work in this direction, introducing a core modality alignment framework to reduce paired data requirements, where only the image as core modality is directly aligned with other modalities. The concept of **emergent** alignment pertains to the indirect alignment performance, and it also facilitates a degree of alignment between modalities that are not directly trained together. However, when other modalities are indirectly aligned through text, the emergent effect exists only in limited form, which typically results in suboptimal zero-shot performance.

To address the weak zero-shot performance issue of ImageBind, LanguageBind (Zhu et al., 2024) selects text as the core modality and employs generative data to enhance text-related zero-shot capabilities. However, LanguageBind struggles with retrieval tasks involving non-text modalities due to the limitations of the binding method. Specifically, text needs more of the fine-grained details presented in images. Using text as the core modality significantly reduces retrieval performance across various image-based modalities. Besides, generating large amounts of data requires much

human work and computational resources. A vital limitation of these binding methods relying on indirect alignment lies in the potential degradation of emergent or zero-shot performance, especially for modalities that lack direct alignment(Zhu et al., 2024).

In this paper, we propose **TangentBind**. This method is capable of mapping all modalities into a unified embedding space and enhancing the emergent capabilities of models using an embedding generative network. Furthermore, our method does not require extra datasets where all modalities are present simultaneously, nor does it rely on massive synthetic data across different modalities. Instead, we train a generative network to generate the embedding vectors of modalities that have already been aligned with the core modality. Based on the embedding vector of the core modality, these generated embeddings can be used to align with other modalities. Notably, when the generated embedding is used to align the generated embedding vector with other modalities, it can affect the alignment with the core modality. Therefore, we propose the **Tangent Term** for preventing this problem, and we demonstrate the effectiveness of the **Tangent Term** in section 3.2. Additionally, we demonstrate that TangentBind can be initialized using large-scale pre-trained multimodal models, such as CLIP(Radford et al., 2021), LanguageBind(Zhu et al., 2024), and ImageBind(Girdhar et al., 2023), and that **Tangent Term** can massively increase the emergence capacity used to initialize the model. We utilized Image and Text as the core modalities, respectively, and ensured the extension of vision and language to audio, depth, and infrared modalities. The model demonstrated strong emergent capabilities in tasks involving each modality that was not directly trained. In figure.2, we compare **TangentBind** with ImageBind and LanguageBind and show the advantages of TangentBind. **TangentBind**, with the image as the core modality, achieved emergent classification top-1 accuracy. Figure.1 demonstrates the powerful emergent capabilities of **TangentBind** with image core modality. Our experiment results surpass ImageBind by 8.5%, 10.3%, and 16.1%, respectively, on ESC, LLVIP, and NYU-D benchmarks. Similarly, **TangentBind**, with text as the core modality, also achieved emergent retrieval Recall@1 results of 25.1, 12.8, and 23.9 on image retrieval tasks in VGG-S, LLVIP, and NYU-D benchmarks surpassing LanguageBind by 15.1, 5.3, and 6.0, respectively. Therefore, TangentBind can compete with or outperform specialized models trained with direct supervision.

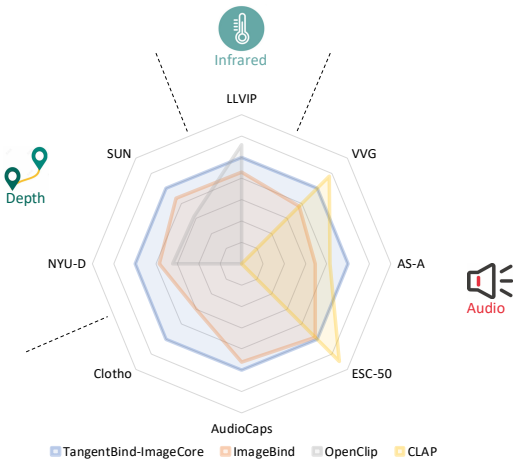

Figure 1: **Zero-shot** Language-related task performance. TangentBind with image core modality was demonstrated as a powerful emergent ability for indirect alignment modalities.

Our primary contributions are listed as follows:

- We propose **TangentBind**, a multimodal pre-train method based on latent space generation. During training, all modalities are aligned with the core modality through contrastive learning, while the generative model enhances the emergent capabilities of the indirect alignment modality.
- We propose **Tangent Term**, a loss function term that enhances the emergent ability of the model while maintaining core modality alignment.
- Extensive experiments validate the effectiveness of our method, demonstrating significant performance improvements in the emergent capabilities of bind-type models.

2 RELATED WORK

2.1 MULTIMODAL LEARNING

CLIP(Radford et al., 2021) is a pioneering multimodal learning method that aligns images and text for constructing cross-modal representations. Various methods, such as CLIP4Clip(Luo et al., 2022)

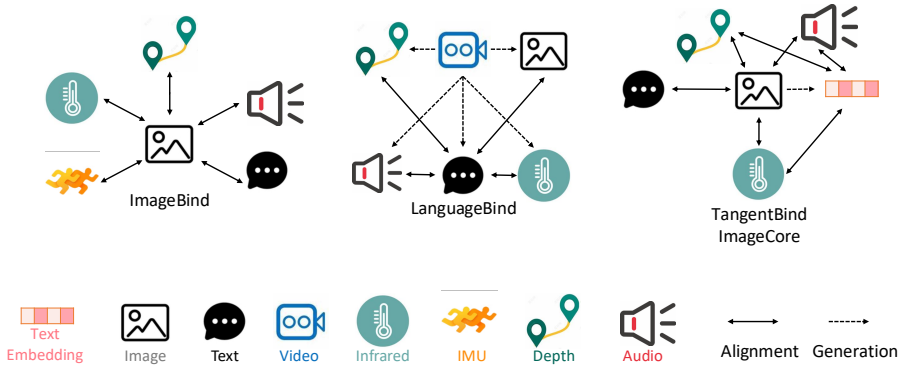

Figure 2: TangentBind vs. ImageBind and LanguageBind. The image on the left shows ImageBind's indirect alignment of the modal by aligning it directly with image. The center image shows LanguageBind augmenting the zero-shot capability of the model by generating data. The right image demonstrates TangentBind, with image as the core modality, which enhances the emergent zero-shot capabilities by simply generating embedding.

and Clip2Video(Fang et al., 2021), etc, adapt CLIP to extract semantic vision representations. Recent efforts have comprehensively explored multimodal alignment through pretraining(Yin et al., 2023; Xu et al., 2023; Wu et al., 2023). In addition to language and vision modalities, AudioCLIP(Guzhov et al., 2021) adds audio as an additional modality with the CLIP framework, enabling zero-shot audio classification. Imagebind(Girdhar et al., 2023) expands multi-modal alignment pretraining by aligning all modalities with the vision modality. ImageBind-LLM(Han et al., 2023) uses the joint embedding space in the pre-trained ImageBind to fine-tune LLaMA efficiently. NeuroBind(Yang et al., 2024) learns a general representation based on pre-trained image embedding space that unifies multiple types of brain signals. UniBind(Lyu et al., 2024) adaptively build LLM-augmented classwise embedding centers and learn to achieve a unified and balanced representation space. To enhance the performance on language-related tasks, MEDBind(Gao et al., 2024), LanguageBind(Zhu et al., 2024) use text data as the core modality to align other modalities. The methods mentioned above, however, did not investigate strategies to enhance the model's emergent capabilities for previously untrained modality pairs.

2.2 Contrastive Learning

Contrastive learning has been remarkably successful in learning representations from multimodal data pairs (Logeswaran & Lee, 2018; He et al., 2020). The primary motivation behind these work is maximizing the mutual information between two views (Tian et al., 2020; Bachman et al., 2019; Tamkin et al., 2020). The loss functions, such as NCE (Gutmann & Hyvärinen, 2010), infoNCE (van den Oord et al., 2018) and MIL-NCE (Miech et al., 2020), have been proposed for contrastive learning. However, these loss functions focus on the alignment of two modalities. To extend the model to multiple modalities, (Guzhov et al., 2021; Alayrac et al., 2020) propose a simple summation of loss functions for joint learning across various modalities. According to Wang & Isola (2020), the loss function of contrastive learning can be split into the alignment and uniformity parts. The alignment part is responsible for the alignment when using loss function summation for multimodal alignment. We analyze the direct summation of two loss functions, and then the two alignment parts will interfere in appendix A.2.

3 METHOD

We present TangentBind, a multimodal pretraining approach to align different modalities and enhance cross-modal retrieval and emergent classification. Figure 3 shows the process of aligning text

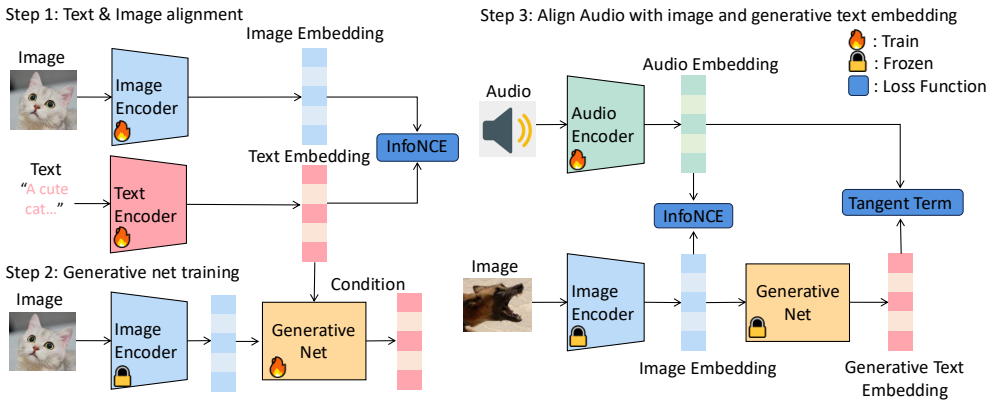

Figure 3: An example of TangentBind overview with $\mathcal{M}_a, \mathcal{M}_b$ and $\mathcal{C}$ are text, audio, and image, respectively. Firstly, we align image with text with infoNCE loss function. Secondly, our work trains generative network fed by image embedding conditions on text embedding with frozen image encoder and text encoder parameters. Finally, our method uses infoNCE and **tangentbind** to align the audio with the image and the generated text embedding.

with image and then aligning audio with image and generative text embedding, with the image as the core modality. **TangentBind** consists of three steps:

1. Align modality $\mathcal{M}_a$ with the core modality $\mathcal{C}$
2. Train a generative network to produce $\mathcal{M}_a$ embedding using core modality embedding.
3. Align modality $\mathcal{M}_b$ with both the $\mathcal{C}$ and the generative modality

When using TangentBind to bind $N+1$-th modality, this process only needs to replace $\mathcal{M}_a$ and $\mathcal{M}_b$ with any modality that has been aligned with $\mathcal{C}$ and $N+1$-th modality, and then perform step 2 and 3.

3.1 Aligning Core Modality and Generative Network

Following ImageBind (Girdhar et al., 2023) and LanguageBind (Zhu et al., 2024), TangentBind trains each encoder separately for respective modality. First, step 1 uses a contrastive learning method to train $\mathcal{M}_a$ encoder and $\mathcal{C}$ encoder for alignment. Then TangentBind uses aligned modality $\mathcal{M}_a$ to train a latent generative model of modality pair $\langle x_i^a, c_i \rangle$. We train our model to predict the unnoised x_i^a directly and use a mean-squared error loss on this prediction:

$$\mathcal{L} = \mathrm{E}_{t\sim[1,T]} \|G_\Theta(\hat{x}_i^{a,(t)}, t, c_i) - x_i^a\|^2. \tag{1}$$

In formula 1, G is generative net and Θ is the training coefficient. Finally, they can be normalized on the hypersphere, and we can get the generated embedding $\hat{x}_i^a$. When aligning other modalities with $\mathcal{C}$, we also use the **Tangent Term** to align with the generated $\mathcal{M}_a$ embedding in the tangent space.

3.2 Aligning Core Modality and Generative Modality

The alignment performance will be weakened if we directly align with the generated inaccurate embedding (Oussidi & Elhassouny, 2018). Therefore, **Tangent Term**, an improvement of infoNCE, is proposed to reduce the adverse effects of inaccurate embeddings.

The definition of infoNCE is given by the equation:

$$\mathcal{L}^{\text{infoNCE}}_{\mathcal{M}_b \to \mathcal{C}} = -\frac{1}{N} \sum_i^N \log\left(\frac{\exp(\text{sim}(x_i^b, c_i)/\tau)}{\sum_j^N \exp(\text{sim}(x_i^b, c_j)/\tau)}\right). \tag{2}$$

In equation 2 x_i^b, c_i is $i-$th embedding vector of $\mathcal{M}_b, \mathcal{C}$ respectively, τ is temperature and $\text{sim}(\cdot,\cdot)$ is similarity function. According to Wang & Isola (2020), infoNCE can be divided into the align part and the uniform part as shown in equation 3 and 4 respectively.

$$\mathcal{L}^{\text{align}} = -\text{sim}(x_i^b, c_i)/\tau, \tag{3}$$

$$\mathcal{L}^{\text{uniform}} = \log(\sum_j \exp(\text{sim}(x_i^b, c_j)/\tau)). \tag{4}$$

The align part is responsible for aligning the features, and the uniform part makes the embedding space more evenly distributed on the hypersphere.

$$\bar{c}_i \triangleq -\frac{\partial \mathcal{L}^{\text{align}}}{\partial x_i^b} = \frac{\partial \text{sim}(x_i^b, c_i)}{\partial x_i^b}/\tau. \tag{5}$$

To achieve higher similarity between x_i^b and c_i, the moving direction of x_i^b and the inner product of $-\partial\mathcal{L}^{\text{align}}/\partial x_i$ are greater than 0 after updating the parameters. We denote $\bar{c}_i$ in equation 5[1].

Tangent Term (see equation 6) is proposed for keeping the similarity between x_i^b and c_i increasing while aligning $\mathcal{M}_b$ with the generative $\mathcal{M}_a$ embedding $\hat{x}_i^a$.

$$\mathcal{L}^{\text{tan}}_{\mathcal{M}_b \to \mathcal{M}_a} = -\frac{1}{N}\sum_i^N \log(\frac{\exp(\text{sim}(T_{\bar{c}_i}(x_i^b), \hat{x}_i^a)/\tau)}{\sum_j^N \exp(\text{sim}(T_{\bar{c}_i}(x_i^b), \hat{x}_j^a)/\tau)}). \tag{6}$$

Tangent Term In equation 6, $\hat{x}_i^a$ is generated from c_i by the generative network in 3.1 and $T_{\bar{c}_i}(\cdot)$ is tangent normalize function. It is the crucial function in **Tangent Term**. Its functionality is mapping the embedding to the space tangent to $\bar{c}_i$ and scaling it to the unit hypersphere.Thus, $T_{\bar{c}_i}(\cdot)$ is defined as $T_{\bar{c}_i}(x) \triangleq \text{normalize}(\left(I - \frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2}\right)x)$, where $\left(I - \frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2}\right)$ project x into the orthogonal complement space of $\bar{c}_i$ which means $\bar{c}_i^T\left(I - \frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2}\right) = 0$ as shown in Figure.4.

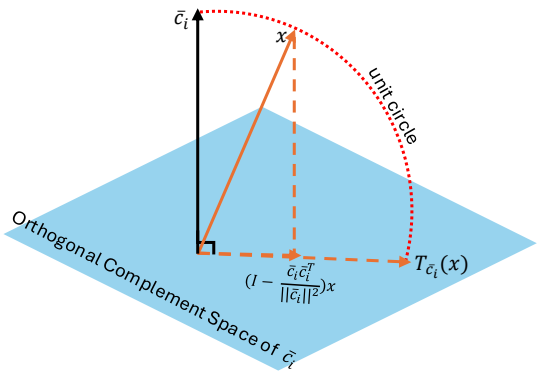

Figure 4: $T_{\bar{c}_i}$ maps x to the orthogonal space of $\bar{c}_i$ and normalizes it.

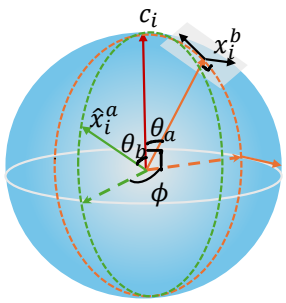

Figure 5: The process of aligning x_i^b with c_i and $\hat{x}_i^a$ simultaneously. θ_a, θ_b represent the angles of c_i with $\hat{x}_i^a$, x_i^b, respectively. ϕ is denoted as the angle between $\hat{x}_i^a$ and x_i^b in the tangent space of c_i.

Finally, the loss function can be rewritten into the following form:

$$\mathcal{L} = \mathcal{L}^{\text{infoNCE}} + \lambda\mathcal{L}^{\text{tan}}. \tag{7}$$

In equation 7, λ is the hyperparameter , $\mathcal{L}^{\text{tan}} = (\mathcal{L}^{\text{tan}}_{\mathcal{M}_b\to\mathcal{M}_a} + \mathcal{L}^{\text{tan}}_{\mathcal{M}_a\to\mathcal{M}_b})/2$ and $\mathcal{L}^{\text{infoNCE}} = (\mathcal{L}^{\text{infoNCE}}_{\mathcal{M}_b\to\mathcal{C}} + \mathcal{L}^{\text{infoNCE}}_{\mathcal{C}\to\mathcal{M}_b})/2$.

Figure 5 illustrates the loss function 7 enables x_i^b to simultaneously align with both c_i and $\hat{x}_i^a$ while using cosine similarity. The black arrow toward c_i represents the infoNCE objective, bringing x_i^b

[1]It is worth noting that the higher the similarity, the lower $\mathcal{L}^{\text{align}}$

and c_i closer together. The other black arrow indicates the **Tangent Term**, which drives x_i^b toward $\hat{x}_i^a$ within the space orthogonal to c_i. As this orthogonal space on the sphere corresponds to the tangent space at c_i (do Carmo, 2016), we refer to the method as **TangentBind**.

4 THEORETICAL ANALYSIS

This analysis mainly illustrates the effect of the **Tangent Term** on the similarity of x_i^b and c_i while ignoring the effect of $\mathcal{L}^{\text{uniform}}$. Since the optimization methods(Kingma & Ba, 2014; Ruder, 2016), are all gradient-based optimization, our analysis is based on gradients primarily. The update parameter $\Delta\Theta$ can be simply written as $\delta(\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta}+\lambda\frac{\partial\mathcal{L}^{\text{tan}}}{\partial\Theta})$ where δ is step size. To clarify the effect of $\mathcal{L}^{\text{tan}}$ on $\mathcal{L}^{\text{align}}$ during gradient descent, we introduce Theorem 1, and the proof is shown in Appendix.A.1.

Theorem 1. *If* $\lambda \leq \frac{\|\bar{c}_i\|}{\|\frac{\partial\mathcal{L}^{tan}}{\partial T_{\bar{c}_i}(x_i^b)}\|}$, *then we have*

$$(\frac{\partial\mathcal{L}^{align}}{\partial\Theta}+\lambda\frac{\partial\mathcal{L}^{tan}}{\partial\Theta})^T\frac{\partial\mathcal{L}^{align}}{\partial\Theta}\geq 0. \tag{8}$$

According to Theorem.1 and Taylor expansion(Rudin, 1976), when updating the coefficients, we have

$$\mathcal{L}^{\text{align}}(\Theta+\Delta\Theta)\approx\mathcal{L}^{\text{tan}}(\Theta)-\delta(\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta}+\lambda\frac{\partial\mathcal{L}^{\text{tan}}}{\partial\Theta})^T\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta}\leq\mathcal{L}^{\text{align}}(\Theta), \tag{9}$$

which means the similarity between x_i^b and c_i will not decrease. To be specific, when $\text{sim}(\cdot,\cdot)$ is cosine similarity, substituting $\frac{\partial\mathcal{L}^{\text{tan}}}{\partial x_i^b}$ with $\bar{c}_i$ according to equation.5 we have

$$\|\bar{c}_i\|=\|\frac{\partial\mathcal{L}^{\text{align}}}{\partial x_i^b}\|=\|\frac{\partial(c_i^T\cdot x_i^b)}{\partial x_i^b}\|/\tau=\|c_i^T\|/\tau=\frac{1}{\tau}, \tag{10}$$

and

$$\|\frac{\partial\mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}\|=\|\frac{\partial((\hat{x}_i^a)^T\cdot T_{\bar{c}_i}(x_i^b))}{\partial T_{\bar{c}_i}(x_i^b)}\|/\tau=\|\hat{x}_i^a\|/\tau=\frac{1}{\tau}. \tag{11}$$

Thus, cosine similarity means $\lambda \leq 1$ can ensure that $(\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta}+\lambda\frac{\partial\mathcal{L}^{\text{tan}}}{\partial\Theta})^T\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta}\geq 0$ due to $\|\bar{c}_i\|/\|\frac{\partial\mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}\|=1$.

In the above analysis, we ignore the effect of $\mathcal{L}^{\text{uniform}}$. However, during the training process, $\mathcal{L}^{\text{uniform}}$ may cause similarity to decrease. The detailed analysis for $\mathcal{L}^{\text{uniform}}$ has been studied in previous work (Liang et al., 2022; Wang & Isola, 2020).

5 EXPERIMENTS AND RESULTS

This section includes an evaluation of the effectiveness of TangentBind in various downstream tasks such as RGB image, depth image, infrared image, and audio. The effectiveness of the generative network has also been tested. We also conduct the ablation study to analyze the impact of **Tangent Term** and different parameter configurations on the performance of TangentBind. For the dataset and experimental implementation details, please refer to the Appendix.B and C.

5.1 IMPLEMENTATION DETAILS

Only two modalities (image and language) have datasets paired with multiple other modalities, so we only show the results using **TangentBind** with Image and Text as the core modality, respectively. The $\text{sim}(\cdot,\cdot)$ notation in **TangentBind** is cosine similarity. To demonstrate the adaptability of **Tangent Term**, we use the pre-trained multimodal models ImageBind-Huge(Girdhar et al., 2023) and LanguageBind(Zhu et al., 2024) respectively for initialization. Since the embedding generated by the diffusion model is different at each time, we generate $\mathcal{M}_a$ embeddings 100 times based on each c_i and calculate the mean of the generated $\mathcal{M}_a$ embeddings. We use AudioSet(Gemmeke et al., 2017), SUN(Song et al., 2015), and LLVIP(Jia et al., 2021) to fine-tune models. While fine-tuning

Table 1: X-Language classification. * donates **Emergent Zero-shot**. We report the top-1 classification accuracy(%) for all datasets except AudioSet (mAP). The SOTA of NYU-D, SUN, AudioSet, ESC-50, and VGG-S come from (Girdhar et al., 2022)(Girdhar et al., 2022; Koutini et al., 2021; Chen et al., 2022; Kazakos et al., 2021) respectively. DepthSwin model(Girdhar et al., 2022) is finetuned from the ImageSwin model(Liu et al., 2021). Our results are highlighted in bold.

Method	Infrared	Depth		Audio		
	LLVIP	NYU-D	SUN	AudioSet	ESC-50	VGG-S
OpenCLIP(Cherti et al., 2023)	82.2	45.4	25.4	-	-	-
DepthSwin(Girdhar et al., 2022)	-	72.5	63.1	-	-	-
JointCRF(Wang et al., 2017)	-	65.8	63.6	-	-	-
DFCR(Cao et al., 2018)	-	65.3	56.3	-	-	-
AudioCLIP(Guzhov et al., 2021)	-	-	-	28.4	68.6	47.4
CLAP(Elizalde et al., 2023)	-	-	-	23.1	**92.6**	46.2
WAV2CLIP(Wu et al., 2021)	-	-	-	0.71	41.4	10.0
LanguageBind	87.2	65.1	-	27.7	91.8	28.9
Tan-LanguageCore	**85.1**	**65.8**	**-**	**28.1**	**92.0**	**29.3**
ImageBind*	63.4	54.0	35.1	17.6	66.9	27.8
Tan-ImageCore*	**73.7**	**70.1**	**40.3**	**25.4**	**68.4**	**36.3**
Absolute SOTA	-	79.4	64.9	49.6	97.0	52.5

the encoders of other modalities , we freeze the image and text encoder parameters of ImageBind and LanguageBind. The temperature of **Tangent Term** and infoNCE is set to the same to balance the functionality between them. According to Sec.4, to ensure the alignment with core modality, we make $\lambda = 1$. We use a $6-$layer decode-only transformer architecture diffusion model in Ramesh et al. (2022) with 100 time steps for the generative network,[2]. To demonstrate the enhancement of the emergent capability with **Tangent Term**, we use only generative networks that produce either image or text for our downstream tasks, which are all text-related or image-related.

5.2 Tangent Term Augmenting ImageCore Model

Table 2: **Zero-shot** Audio-Language retrieval. * donates **Emergent Zero-shot**. Our results are highlighted in bold.

Method	Clotho		AudioCaps	
	R@1	R@10	R@1	R@10
AVFIC(Nagrani et al., 2022)	3.0	17.5	8.7	37.7
AudioClip(Guzhov et al., 2021)	3.20	20.3	3.53	31.6
WAV2Clip(Wu et al., 2021)	0.78	12.1	0.88	15.3
C-MCR(Wang et al., 2023b)	8.37	36.7	15.76	48.1
LanguageBind	12.1	44.0	12.2	53.2
Tan-LanguageCore	**11.7**	**41.8**	**11.8**	**52.1**
ImageBind*	6.0	28.4	9.3	42.3
Tan-ImageCore*	**10.0**	**33.4**	**10.1**	**49.2**

Emergent Ability As shown in Table 1, we tested the effect of **Tangent Term** on ImageCore on the emergent classification task on 6 datasets. On the emergent zero-shot classification tasks of Audio, Depth, and Infrared (VGG-S(Chen et al., 2020), NYU-D(Nathan Silberman & Fergus, 2012), and LLVIP(Jia et al., 2021)), **TangentBind** top-1 accuracy outperforms ImageBind 8.5%, 16.1%, and 10.3% respectively. In addition, as shown in Table 2, we also test **TangentBind** on the Audio-Language emergent retrieval task. The recall@10 on the Clotho(Drossos et al., 2019), AudioCaps(Kim et al., 2019) datasets are 5% and 6.9% higher than the ImageBind after the introduction of **Tangent Term**. The emergent capability shows significant improvement across all benchmarks, achieving performance levels that closely approximate those achieved by incorporating text features. These experiment results suggest that **TangentBind** effectively aligns multiple modalities within Tangent Space, thereby significantly enhancing the emergent zero-shot capabilities.

Core Modality Alignment Ability Table 3 presents the performance of TangentBind for zero-shot retrieval using RGB images. The experimental results demonstrate that the introduction of

[2]The code can be found in https://github.com/lucidrains/DALLE2-pytorch

the **Tangent Term** leads to improvements in Recall@1 for AVE (Tian et al., 2018), VGG-S (Chen et al., 2020), LLVIP (Jia et al., 2021), and NYU-D (Nathan Silberman & Fergus, 2012) by 0.1, 1.4, 1.1, and 0.5, respectively, compared to ImageBind model. This indicates that incorporating the Tangent Term does not degrade the alignment of the various modalities with the core modality.

5.3 Tangent Term Augmenting LanguageCore Model

Emergent Ability As presented in Table 3, we evaluated the impact of the **Tangent Term** on LanguageCore in the emergent RGB-related retrieval task across four datasets. For the modalities of Audio, Depth, and Infrared, **TangentBind** demonstrated superior performance in Recall@1 on AVE (Tian et al., 2018), VGG-S (Chen et al., 2020), NYU-D (Nathan Silberman & Fergus, 2012), and LLVIP (Jia et al., 2021), with improvements of 4.5, 5.1, 6.0, and 5.3, respectively, surpassing LanguageBind model. The emergent capability is significantly enhanced across all benchmarks. It demonstrates that the **Tangent Term** remains effective even when the core modality is altered.

Table 3: Comparison of RGB→X retrieval.* donates **Emergent**. Our results are highlighted in bold.

Dataset	Method	Task	R@1
AVE	ImageBind	RGB→A	36.9
	LanguageBind*		10.6
	Tan-ImageCore		**37.0**
	Tan-LanguageCore*		**15.1**
VGG-S	Imagebind	RGB→A	28.7
	LanguageBind*		10.0
	Tan-ImageCore		**30.1**
	Tan-LanguageCore*		**15.1**
LLVIP	Imagebind	RGB→I	26.3
	LanguageBind*		7.5
	Tan-ImageCore		**27.4**
	Tan-LanguageCore*		**12.8**
NYU-D	Imagebind	RGB→D	34.7
	LanguageBind*		17.9
	Tan-ImageCore		**35.2**
	Tan-LanguageCore*		**23.9**

Core Modality Alignment Ability As shown in Table 1, to verify that the alignment between various modalities and the core text modality is not compromised by the introduction of the **Tangent Term**, we evaluated its effect on LanguageCore model across six datasets in the zero-shot classification task. For the modalities of Audio, Depth, and Infrared, **TangentBind** top-1 accuracy outperforms LanguageBind on VGG-S (Chen et al., 2020), NYU-D (Nathan Silberman & Fergus, 2012), and ESC-50 (Piczak, 2015), with marginal improvements of 0.4%, 0.7%, and 0.2%, respectively. On LLVIP (Jia et al., 2021), there is only a minor top-1 accuracy decrease of 2.1%. To further assess the impact of the **Tangent Term** on the core text modality, Table 2 highlights its effect on the audio-language emergent retrieval task. Following the introduction of the **Tangent Term**, the maximum recall value on Clotho (Drossos et al., 2019) and AudioCaps (Kim et al., 2019) decreases by only 2.2. These results demonstrate that the negative impact of the **Tangent Term** on tasks where text serves as the core modality is minimal.

5.4 Ablation Study

The effect of generative networks To illustrate the efficacy of the diffusion model, we substituted it with ResNet(He et al., 2015), VAE(Kingma & Welling, 2013), and C-MCR(Wang et al., 2023b) in Step 2. In Figure 6, we display the cumulative distribution function (CDF) curves of cosine similarity between the embeddings generated by these methods and the actual data embeddings for VGG-S(Chen et al., 2020) and SUN(Song et al., 2015). When the CDF curve approaches 1.0, it indicates the generated embeddings are similar to the actual data embeddings. As illustrated in Figure 6, the embeddings produced by the diffusion model closely resemble the actual embeddings. Notably, the C-MCR method failed on the SUN dataset. Detailed descriptions of each method's implementation are provided in Appendix C. To demonstrate the necessity of Step 2 and the robust capability of **TangentTerm** to maintain alignment with the core modality, we introduced Gaussian noise with a mean of 0 and a variance of 1e-3 to the core modal embeddings and normalized them as a disturbance sample. Table 4 shows that the image core model with the diffusion model achieved top-1 classification accuracy of 36.3%, 73.7%, and 70.1% on **emergent** text-related classification tasks across the VGG-S, LLVIP, and NYU-D datasets, respectively, outperforming other methods.

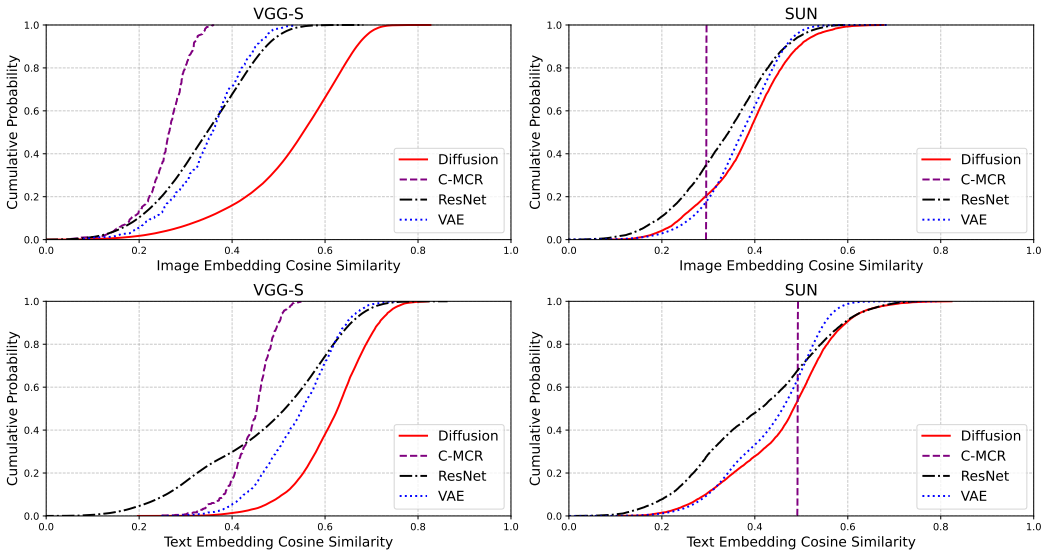

Figure 6: CDF Curve of Similarity on VGG-S and SUN. The top/bottom row displays curves representing the cosine similarity between text/image embeddings generated by various methods in Step 2 and the actual text/image embeddings, with image/text as the core modality. The solid, dashed, dash-dotted, and dotted lines correspond to the diffusion, C-MCR, ResNet, and VAE methods, respectively.

Table 4: **Diffusion Model Embedding** vs. **Other Method Embedding**. Top-1 accuracy on $X \rightarrow T$ tasks and Recall@1 on $RGB \rightarrow X$ tasks. Diff denotes the generation of embeddings using a diffusion model (Ramesh et al., 2022) in Step 2. The best results are highlighted in bold. A gray background indicates that the result is from an **Emergent Zero-shot** task.

Dataset	Task	ImageCore+					LanguageCore+				
		Diff	ResNet	VAE	Noise	C-MCR	Diff	ResNet	VAE	Noise	C-MCR
VGG-S	A→T	**36.3**	30.7	32.1	23.7	26.7	**29.3**	27.2	28.3	26.4	29.1
	RGB→A	**30.1**	29.3	28.1	27.7	28.3	**15.1**	12.3	13.5	8.1	9.5
LLVIP	I→T	**73.7**	68.1	64.3	59.1	61.7	85.1	83.1	87.5	86.2	**86.9**
	RGB→I	**27.4**	26.5	27.1	26.1	25.9	**12.8**	10.3	9.1	5.2	8.9
NYU-D	D→T	**70.1**	67.2	69.3	50.2	58.4	**65.8**	65.7	64.8	64.1	65.4
	RGB→D	**35.2**	35.0	33.9	33.1	34.4	**23.9**	20.7	22.4	14.6	18.9

Similarly, Table 4 reveals that the text core model with the diffusion model reached the highest Recall@1 scores of 15.1, 12.8, and 23.9 on **emergent** RGB-related retrieval tasks across the same datasets. Moreover, it is noteworthy that models using the diffusion model in Step 2 achieved optimal performance in all core modality-related tasks, which is white background in Table 4, except for a slight decline in the $I \rightarrow T$ task on the LLVIP dataset compared to the best result.

Replacing Tangent Term by infoNCE To visualize the ability of **Tangent Term** to maintain the core modality alignment, Table 5 presents the core modality alignment results of models with image and text as the core modalities. In Table 5, $\mathcal{L}^{\text{infoNCE}}$ is used to directly align with the generated embeddings instead of $\mathcal{L}^{\text{tan}}$. In detail, the **Tangent Term** $\mathcal{L}^{\text{tan}}_{\mathcal{M}_b \rightarrow \mathcal{M}_a} + \mathcal{L}^{\text{tan}}_{\mathcal{M}_a \rightarrow \mathcal{M}_b}$ is replaced by $\mathcal{L}^{\text{infoNCE}}_{\mathcal{M}_b \rightarrow \mathcal{M}_a} + \mathcal{L}^{\text{infoNCE}}_{\mathcal{M}_a \rightarrow \mathcal{M}_b}$ in loss function. For consistency, the value of λ is set to 1. As shown in Table 5, compared to **Tangent Term**, infoNCE Recall@1 scores on the image-based retrieval tasks on the VGG-S, AVE, NYU-D, and LLVIP datasets drop by 4.9, 4.7, 6.1, and 5, respectively. This indicates a direct use of infoNCE can severely disrupt the alignment of core modalities. A similar pattern is observed when text is the core modality, as Table 5 demonstrates that if infoNCE is used directly then the top-1 accuracy of text-based classification on the VGG-S, ESC-50, NYU-D, and LLVIP datasets will drop significantly to 22.2%, 72.3%, 55.7% and 64.1%. These results suggest that the **Tangent Term** enhances emergent capabilities and mitigates the negative impact on core modality alignment compared to direct alignment using infoNCE.

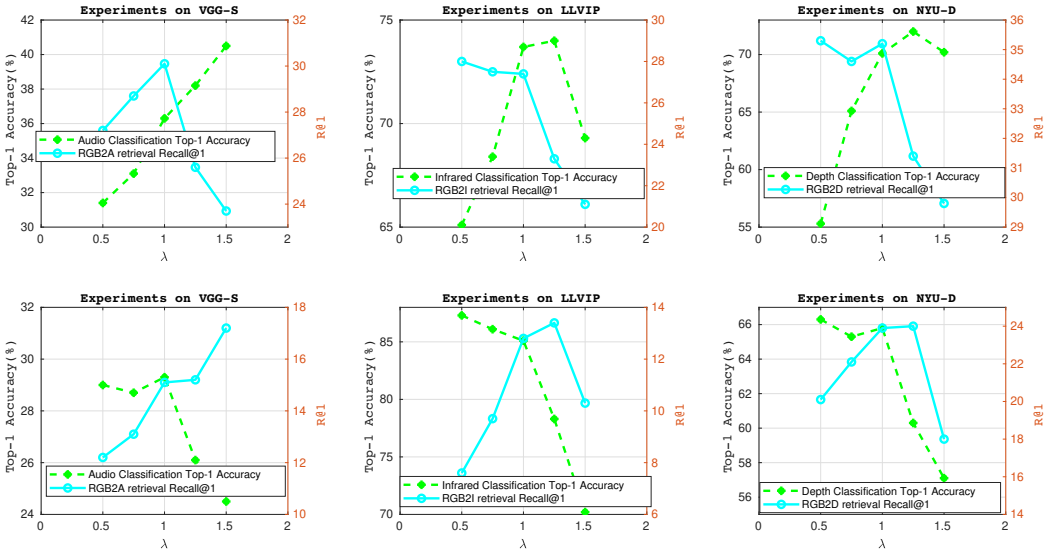

Figure 7: Experimental results on classification and retrieval tasks after varying hyperparameters λ. The first row of figures from left to right shows the results on the VGG-S, LLVIP, and NYU-D datasets with images as the core modality. The second row of figures from left to right shows the results on VGG-S, LLVIP, and NYU-D datasets with text as the core modality, respectively.

Changing Hyperparameter λ Figure.7 presents the top-1 accuracy and Recall@1 values for classification and retrieval on various datasets as a function of varying λ during training. As observed in Figure.7, the emergent capability of the model with the image core modality on LLVIP improves with increasing values of λ, reaching its peak at $\lambda = 1.25$. However, when $\lambda > 1.25$, the emergent performance gradually declines on such a model. In contrast, the Recall@1 value for image retrieval consistently decreases as λ increases. This trend aligns with Theorem 1, which indicates that the **Tangent Term** negatively impacts core modality alignment when λ becomes excessively large.

6 CONCLUSION

In this work, we introduce **TangentBind**, an emergent enhancement method for multimodal pretraining. To improve the integrity of modality, we train a generative network that indirectly aligns modality embeddings. Additionally, to prevent the generated embeddings from compromising alignment with the core modality, we propose the **Tangent Term** for aligning the generated modality embeddings. Extensive experiments, including the use of multiple core modalities and ablation studies, demonstrate that the **Tangent Term** can enhance the emergent capabilities of the multimodal alignment model while preserving alignment with the core modality.

Table 5: **Image based retrieval** and **text based classification** on VGG-S, AVE, NYU-D, LLVIP and ESC-50. We replace **Tangent Term** by infoNCE in **image core** and **text core** modality modal during training, respectively, and compare core modality alignment performance. We report the Recall@1 score for image based retrieval tasks and top-1 classification accuracy(%) for text based classification tasks.

Dataset	**Task**	**Core**	infoNCE+ **Tan Term**	infoNCE
VGG-S	RGB→A	Image	**30.1**	25.2
	A→T	Text	**29.3**	22.2
AVE	RGB→A	Image	**37.0**	32.3
NYU-D	RGB→D	Image	**35.2**	29.1
	D→T	Text	**65.8**	55.7
LLVIP	RGB→I	Image	**27.4**	22.4
	I→T	Text	**85.1**	64.1
ESC-50	A→T	Text	**92.0**	72.3

REFERENCES

Jean-Baptiste Alayrac, Adria Recasens, Rosalia Schneider, Relja Arandjelović, Jason Ramapuram, Jeffrey De Fauw, Lucas Smaira, Sander Dieleman, and Andrew Zisserman. Self-supervised multimodal versatile networks. *Advances in neural information processing systems*, 33:25–37, 2020.

Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *ArXiv*, abs/1810.02281, 2018. URL https://api.semanticscholar.org/CorpusID:52922363.

Philip Bachman, R Devon Hjelm, and William Buchwalter. Learning representations by maximizing mutual information across views. *Advances in neural information processing systems*, 32, 2019.

Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

Yuanzhouhan Cao, Zifeng Wu, and Chunhua Shen. Estimating depth from monocular images as classification using deep fully convolutional residual networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 28(11):3174–3182, 2018. doi: 10.1109/TCSVT.2017.2740321.

Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 721–725, 2020. URL https://api.semanticscholar.org/CorpusID:216522760.

Ke Chen, Xingjian Du, Bilei Zhu, Zejun Ma, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Hts-at: A hierarchical token-semantic audio transformer for sound classification and detection. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 646–650. IEEE, 2022.

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2818–2829, 2023.

Aayush Dhakal, Subash Khanal, Srikumar Sastry, Adeel Ahmad, and Nathan Jacobs. Geobind: Binding text, image, and audio through satellite images. *arXiv preprint arXiv:2404.11720*, 2024.

M.P. do Carmo. *Differential Geometry of Curves and Surfaces: Revised and Updated Second Edition*. Dover Books on Mathematics. Dover Publications, 2016. ISBN 9780486806990. URL https://books.google.com/books?id=uXF6DQAAQBAJ.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: an audio captioning dataset. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740, 2019. URL https://api.semanticscholar.org/CorpusID:204800739.

Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1675–1685. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/du19c.html.

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)*, pp. 1–5. IEEE, 2023.

Han Fang, Pengfei Xiong, Luhui Xu, and Yu Chen. Clip2video: Mastering video-text retrieval via image clip. *ArXiv*, abs/2106.11097, 2021. URL https://api.semanticscholar.org/CorpusID:235490558.

Yuan Gao, Sangwook Kim, David E Austin, and Chris McIntosh. Medbind: Unifying language and multimodal medical data embeddings. *ArXiv*, abs/2403.12894, 2024. URL https://api.semanticscholar.org/CorpusID:268532501.

Jort F. Gemmeke, Daniel P. W. 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 *Proc. IEEE ICASSP 2017*, New Orleans, LA, 2017.

Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2232–2241. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ghorbani19b.html.

Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens Van Der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16102–16112, 2022.

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, 2023. URL https://arxiv.org/abs/2305.05665.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 297–304. JMLR Workshop and Conference Proceedings, 2010.

Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio, 2021. URL https://arxiv.org/abs/2106.13043.

Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, Xudong Lu, Shuai Ren, Yafei Wen, Xiaoxin Chen, Xiangyu Yue, Hongsheng Li, and Yu Jiao Qiao. Imagebind-llm: Multi-modality instruction tuning. *ArXiv*, abs/2309.03905, 2023. URL https://api.semanticscholar.org/CorpusID:261582620.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 770–778, 2015. URL https://api.semanticscholar.org/CorpusID:206594692.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9729–9738, 2020.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

Xinyu Jia, Chuang Zhu, Minzhen Li, Wenqi Tang, and Wenli Zhou. Llvip: A visible-infrared paired dataset for low-light vision. *2021 IEEE/CVF International Conference on Computer Vision Workshops (ICCVW)*, pp. 3489–3497, 2021. URL https://api.semanticscholar.org/CorpusID:237278539.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2014.

Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Slow-fast auditory streams for audio recognition. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 855–859. IEEE, 2021.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. AudioCaps: Generating captions for audios in the wild. In Jill Burstein, Christy Doran, and Thamar Solorio (eds.), *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)*, pp. 119–132, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics. doi: 10.18653/v1/N19-1011. URL https://aclanthology.org/N19-1011.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL https://api.semanticscholar.org/CorpusID:6628106.

Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL https://api.semanticscholar.org/CorpusID:216078090.

Khaled Koutini, Jan Schlüter, Hamid Eghbal-Zadeh, and Gerhard Widmer. Efficient training of audio transformers with patchout. *arXiv preprint arXiv:2110.05069*, 2021.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.

Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, H. Cai, Fatih Murat Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *ArXiv*, abs/2305.10764, 2023. URL https://api.semanticscholar.org/CorpusID:258762826.

Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 10012–10022, 2021.

Lajanugen Logeswaran and Honglak Lee. An efficient framework for learning sentence representations. *arXiv preprint arXiv:1803.02893*, 2018.

Zhou Lu. A theory of multimodal learning. *ArXiv*, abs/2309.12458, 2023. URL https://api.semanticscholar.org/CorpusID:262217483.

Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508: 293–304, 2022.

Yuanhuiyi Lyu, Xu Zheng, Jiazhou Zhou, and Lin Wang. Unibind: Llm-augmented unified and balanced representation space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26752–26762, 2024.

Charles H. Martin and Michael W. Mahoney. Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning. *Journal of Machine Learning Research*, 22(165):1–73, 2021. URL http://jmlr.org/papers/v22/20-410.html.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9879–9889, 2020.

Niluthpol Chowdhury Mithun, Juncheng Li, Florian Metze, and Amit K. Roy-Chowdhury. Learning joint embedding with multimodal cues for cross-modal video-text retrieval. In *Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval*, ICMR '18, pp. 19–27, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450350464. doi: 10.1145/3206025.3206064. URL https://doi.org/10.1145/3206025.3206064.

Seungwhan Moon, Andrea Madotto, Zhaojiang Lin, Alireza Dirafzoon, Aparajita Saraf, Amy Bearman, and Babak Damavandi. Imu2clip: Multimodal contrastive learning for imu motion sensors from egocentric videos and text. *ArXiv*, abs/2210.14395, 2022. URL https://api.semanticscholar.org/CorpusID:253117171.

Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manén, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *European Conference on Computer Vision*, 2022. URL https://api.semanticscholar.org/CorpusID:247939759.

Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

Andreea-Maria Oncescu, A Koepke, Joao F Henriques, Zeynep Akata, and Samuel Albanie. Audio retrieval with natural language queries. *arXiv preprint arXiv:2105.02192*, 2021.

Achraf Oussidi and Azeddine Elhassouny. Deep generative models: Survey. In *2018 International conference on intelligent systems and computer vision (ISCV)*, pp. 1–8. IEEE, 2018.

Karol J. Piczak. Esc: Dataset for environmental sound classification. *Proceedings of the 23rd ACM international conference on Multimedia*, 2015. URL https://api.semanticscholar.org/CorpusID:17567398.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL https://arxiv.org/abs/2103.00020.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *ArXiv*, abs/2204.06125, 2022. URL https://api.semanticscholar.org/CorpusID:248097655.

Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016.

W. Rudin. *Principles of Mathematical Analysis*. International series in pure and applied mathematics. McGraw-Hill, 1976. ISBN 9780070856134. URL https://books.google.com.sg/books?id=kwqzPAAACAAJ.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 567–576, 2015.

Alex Tamkin, Mike Wu, and Noah D. Goodman. Viewmaker networks: Learning views for unsupervised representation learning. *ArXiv*, abs/2010.07432, 2020. URL https://api.semanticscholar.org/CorpusID:222381644.

Yapeng Tian, Jing Shi, Bochen Li, Zhiyao Duan, and Chenliang Xu. Audio-visual event localization in unconstrained videos. In *ECCV*, 2018.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pp. 776–794. Springer, 2020.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748, 2018. URL https://api.semanticscholar.org/CorpusID:49670925.

Jianhua Wang, Chuanxia Zheng, Weihai Chen, and Xingming Wu. Learning aggregated features and optimizing model for semantic labeling. *The Visual Computer*, 33:1587–1600, 2017.

Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities. *ArXiv*, abs/2305.11172, 2023a. URL https://api.semanticscholar.org/CorpusID:258762390.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, 2020. URL https://api.semanticscholar.org/CorpusID:218718310.

Zehan Wang, Yang Zhao, Xize Cheng, Haifeng Huang, Jiageng Liu, Lilian H. Y. Tang, Lin Li, Yongqiang Wang, Aoxiong Yin, Ziang Zhang, and Zhou Zhao. Connecting multi-modal contrastive representations. *ArXiv*, abs/2305.14381, 2023b. URL https://api.semanticscholar.org/CorpusID:258866011.

Zehan Wang, Ziang Zhang, Xize Cheng, Rongjie Huang, Luping Liu, Zhenhui Ye, Haifeng Huang, Yang Zhao, Tao Jin, Peng Gao, and Zhou Zhao. Freebind: Free lunch in unified multi-modal space via knowledge fusion. *ArXiv*, abs/2405.04883, 2024a. URL https://api.semanticscholar.org/CorpusID:269626610.

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*, 2024b.

Ho-Hsiang Wu, Prem Seetharaman, Kundan Kumar, and Juan Pablo Bello. Wav2clip: Learning robust audio representations from clip. *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567, 2021. URL https://api.semanticscholar.org/CorpusID:239616434.

Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and S Yu Philip. Multimodal large language models: A survey. In *2023 IEEE International Conference on Big Data (BigData)*, pp. 2247–2256. IEEE, 2023.

Yusong Wu, K. 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. *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2022. URL https://api.semanticscholar.org/CorpusID:253510826.

Peng Xu, Xiatian Zhu, and David A Clifton. Multimodal learning with transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10):12113–12132, 2023.

Fengyu Yang, Chao Feng, Daniel Wang, Tianye Wang, Ziyao Zeng, Zhiyang Xu, Hyoungseob Park, Pengliang Ji, Hanbin Zhao, Yuanning Li, and Alex Wong. Neurobind: Towards unified multimodal representations for neural signals. *ArXiv*, abs/2407.14020, 2024. URL https://api.semanticscholar.org/CorpusID:271310533.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, HongFa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Wancai Zhang, Zhifeng Li, Wei Liu, and Li Yuan. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment, 2024. URL https://arxiv.org/abs/2310.01852.

A Additional Theoretical Analysis

A.1 Proof of Theorem1

Proof. First, we apply the chain rule to get the gradient of $\mathcal{L}^{\text{tan}}$ with respect to the parameter Θ:

$$\frac{\partial \mathcal{L}^{\text{tan}}}{\partial \Theta} = \frac{\partial x_i^b}{\partial \Theta} \frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b} \frac{\partial \mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}. \tag{12}$$

Similarly, using the chain rule to calculate the gradient of $\mathcal{L}_{align}$ with respect to Θ, we get $\frac{\partial \mathcal{L}^{\text{align}}}{\partial \Theta} = \frac{\partial x_i^b}{\partial \Theta} \frac{\partial \mathcal{L}^{\text{tan}}}{\partial x_i^b}$. Substituting $\frac{\partial \mathcal{L}^{\text{tan}}}{\partial x_i^b}$ with $\bar{c}_i$ according to equation.5, we get

$$\frac{\partial \mathcal{L}^{\text{align}}}{\partial \Theta} = -\frac{\partial x_i^b}{\partial \Theta} \bar{c}_i. \tag{13}$$

Take the inner product $\frac{\partial \mathcal{L}^{\text{align}}}{\partial \Theta} + \lambda \frac{\partial \mathcal{L}^{\text{tan}}}{\partial \Theta}$ and $\frac{\partial \mathcal{L}^{\text{align}}}{\partial \Theta}$ to get formula.14.

$$(\frac{\partial \mathcal{L}^{\text{align}}}{\partial \Theta} + \lambda \frac{\partial \mathcal{L}^{\text{tan}}}{\partial \Theta})^T \frac{\partial \mathcal{L}^{\text{align}}}{\partial \Theta} = \bar{c}_i^T \frac{\partial x_i^b}{\partial \Theta}^T \frac{\partial x_i^b}{\partial \Theta} \bar{c}_i + \lambda \frac{\partial \mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}^T \frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b}^T \frac{\partial x_i^b}{\partial \Theta}^T \frac{\partial x_i^b}{\partial \Theta} \bar{c}_i, \tag{14}$$

Notably, we expand $\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b}$ as in equation.15, and find that $\bar{c}_i^T \frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b} = 0$ due to $\bar{c}_i^T \left(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2} \right) = 0$.

$$\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b} = \frac{\partial \left(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2} \right) x_i^b}{\partial x_i^b} \frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial \left(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2} \right) x_i^b} = \left(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2} \right) \frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial \left(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2} \right) x_i^b}. \tag{15}$$

Next, we introduce Lemma.2

Lemma 2. *We have $x^T A^T A y \geq -(\kappa(A^T A) - 1)^{\frac{1}{2}} \|Ay\|^2 \frac{\|x\|}{\|y\|}$, if $y^T x = 0$ and $\kappa(\cdot)$ is condistion number.*

Proof. Rather than proving this inequality directly, we turn to the follow the lower bound for the optimization problem.16.

$$\begin{aligned} \min_{x} \quad & x^T A^T A y, \\ \text{subject to} \quad & y^T x = 0, \\ & x^T x = 1, \end{aligned} \tag{16}$$

To solve optimization problem16, we write the Lagrangian function in 17,

$$L(x, \mu_1, \mu_2) = x^T A^T A y + \mu_1 x^T y + \mu_2 (x^T x - 1). \tag{17}$$

We can easily get the Lagrangian dual function 18 form 17,

$$L(\mu_1, \mu_2) = \inf_x L(x, \mu_1, \mu_2) = -\frac{\|(A^T A - \mu_1 I) y\|^2}{4\mu_2} - \mu_2, \tag{18}$$

Thus, we obtain the unconstrained Lagrangian dual problem 19

$$\max_{\mu_1, \mu_2} L(\mu_1, \mu_2). \tag{19}$$

Besides, we have

$$\max_{\mu_2} L(\mu_1, \mu_2) = \|(A^T A - \mu_1 I) y\| = \left(\mu_1^2 y^T y - 2\mu_1 y^T A^T A y + y^T A^T A A^T A y \right)^{\frac{1}{2}}. \tag{20}$$

Thus, we get

$$\max_{\mu_1, \mu_2} L(\mu_1, \mu_2) = \max_{\mu_1} \max_{\mu_2} L(\mu_1, \mu_2) = -\frac{\left(\|A^T A y\|^2 \|y\|^2 - \|Ay\|^4 \right)^{\frac{1}{2}}}{\|y\|}. \tag{21}$$

Since the maximum eigenvalue of A^TA divided by the minimum eigenvalue of A^TA is less than 2, we have

$$\|A^TAy\|^2\|y\|^2 \leq \|A\|^2\|Ay\|^2\|y\|^2 = \|Ay\|^4\frac{\|A\|^2\|y\|^2}{\|Ay\|^2} \leq \kappa(A^TA)\|Ay\|^4. \tag{22}$$

After that, according to 21 and 22, we have

$$\max_{\mu_1,\mu_2} L(\mu_1,\mu_2) \geq -(\kappa(A^TA)-1)^{\frac{1}{2}}\frac{\|Ay\|^2}{\|y\|}. \tag{23}$$

According to the duality principle (Boyd & Vandenberghe, 2004), we have $\frac{x^T}{\|x\|}A^TAy \geq -(\kappa(A^TA)-1)^{\frac{1}{2}}\frac{\|Ay\|^2}{\|y\|}$ which completes the proof. □

According to Martin & Mahoney (2021); Arora et al. (2018); Du et al. (2019); Ghorbani et al. (2019), $\kappa\left(\frac{\partial x_i^b}{\partial\Theta}^T\frac{\partial x_i^b}{\partial\Theta}\right)$ will converge and less than 2. Consider $\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b}\frac{\partial\mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}$, $\bar{c}_i$ and $\frac{\partial x_i^b}{\partial\Theta}$ as x, y, and A in Lemma.2, respectively. According to Lemma.2, we can get inequality.24

$$(\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta}+\lambda\frac{\partial\mathcal{L}^{\text{tan}}}{\partial\Theta})^T\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta} \geq \|\frac{\partial x_i^b}{\partial\Theta}\bar{c}_i\|^2(1-\lambda\frac{\|\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b}\frac{\partial\mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}\|}{\|\bar{c}_i\|}). \tag{24}$$

Moreover, expanding $\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial(I-\frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2})x_i^b}$ in equation.15 we have

$$\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial(I-\frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2})x_i^b} = I-\frac{(I-\frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2})x_i^b\left((I-\frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2})x_i^b\right)^T}{\|(I-\frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2})x_i^b\|^2}. \tag{25}$$

Furthermore, we find $(I-\frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2})$ and $\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial(I-\frac{\bar{c}_i\bar{c}_i^T}{\|\bar{c}_i\|^2})x_i^b}$ are both projection matrix (Horn & Johnson, 2012), and both spectral radius are less than 1. Thus, we have

$$\|\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b}\frac{\partial\mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}\| \leq \|\frac{\partial\mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}\|, \tag{26}$$

due to the spectral radius of $\frac{\partial T_{\bar{c}_i}(x_i^b)}{\partial x_i^b}$ is less than 1. Thus, according to inequality .24 and 26, we have

$$(\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta}+\lambda\frac{\partial\mathcal{L}^{\text{tan}}}{\partial\Theta})^T\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta} \geq \|\frac{\partial x_i^b}{\partial\Theta}\bar{c}_i\|^2(1-\lambda\frac{\|\frac{\partial\mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}\|}{\|\bar{c}_i\|}), \tag{27}$$

which means $(\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta}+\lambda\frac{\partial\mathcal{L}^{\text{tan}}}{\partial\Theta})^T\frac{\partial\mathcal{L}^{\text{align}}}{\partial\Theta} \geq 0$ when $\lambda \leq \frac{\|\bar{c}_i\|}{\|\frac{\partial\mathcal{L}^{\text{tan}}}{\partial T_{\bar{c}_i}(x_i^b)}\|}$ and complete the proof. □

A.2 INFONCE VS. TANGENTTERM

In this section, we roughly analyze why directly using infoNCE to align with generative embeddings results in a degradation of the core modal alignment capability. For ease of understanding, we consider the case where similarity function is cosine similarity and $\tau = 1$. As analyzed in Sec.3.2, $\mathcal{L}^{\text{tan}}$ aligns the two modes, while $\mathcal{L}^{\text{uniform}}$ just plays a role of regular term. Thus, the align parts of the two infoNCEs are added together and we get

$$\mathcal{L}^{\text{align}}(\hat{x}_i^a,x_i^b)+\mathcal{L}^{\text{align}}(c_i,x_i^b) = -(\hat{x}_i^a)^Tx_i^b-c_i^Tx_i^b = -(\hat{x}_i^a+c_i)^Tx_i^b, \tag{28}$$

which means that when x_i^b is neither aligned with c_i nor with $\hat{x}_i^a$, but instead with normalize$(\hat{x}_i^a+c_i)$. As a result, the ability to align with the core modality is destroyed. However, if we use **Tangent Term** instead of infoNCE, then the loss function for the alignment part becomes

$$\mathcal{L}^{\text{align}} = -c_i^Tx_i^b-(\hat{x}_i^a)^T\text{normalize}((I-\frac{c_ic_i^T}{\|c_i\|^2})x_i^b). \tag{29}$$

It is worth noting that to simplify the notation, we still use $\mathcal{L}^{\text{align}}$ in equation.29, where there is a slight difference between equation.29 and equation.3. Letting $\mathcal{L}^{\text{align}}$ take the derivative of x_i^b, we get

$$\frac{\partial \mathcal{L}^{\text{align}}}{\partial x_i^b} = -c_i - (I - \frac{x_i^b (x_i^b)^T}{\|x_i^b\|^2})(I - \frac{c_i c_i^T}{\|c_i\|^2})(I - \frac{(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b \left((I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b\right)^T}{\|(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b\|^2})\hat{x}_i^a. \quad (30)$$

Since x_i^b is confined to the hypersphere, $(I - \frac{x_i^b (x_i^b)^T}{\|x_i^b\|^2})$ appears in equation.30. Taking an inner product of c_i and $\frac{\partial \mathcal{L}^{\text{align}}}{\partial x_i^b}$, we have

$$c_i^T \frac{\partial \mathcal{L}^{\text{align}}}{\partial x_i^b} = -c_i^T c_i - c_i^T (I - \frac{x_i^b (x_i^b)^T}{\|x_i^b\|^2})(I - \frac{c_i c_i^T}{\|c_i\|^2})(I - \frac{(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b \left((I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b\right)^T}{\|(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b\|^2})\hat{x}_i^a. \quad (31)$$

It is easy to find that

$$c_i^T (I - \frac{x_i^b (x_i^b)^T}{\|x_i^b\|^2}) = c_i^T - \frac{c_i^T x_i^b}{\|x_i^b\|^2} (x_i^b)^T, \quad (32)$$

where

$$c_i^T (I - \frac{c_i c_i^T}{\|c_i\|^2}) = 0, \quad (33)$$

and

$$\frac{c_i^T x_i^b}{\|x_i^b\|^2} \left((x_i^b)^T (I - \frac{c_i c_i^T}{\|c_i\|^2}) \right) (I - \frac{(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b \left((I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b\right)^T}{\|(I - \frac{\bar{c}_i \bar{c}_i^T}{\|\bar{c}_i\|^2}) x_i^b\|^2}) = 0. \quad (34)$$

Thus, we have

$$c_i^T \frac{\partial \mathcal{L}^{\text{align}}}{\partial x_i^b} = -c_i^T c_i \leq 0, \quad (35)$$

which means that the similarity between x_i^b and c_i is on increase.

B Downstream Datasets

AudioSet (Gemmeke et al., 2017) contains 10s videos from YouTube annotated into 527 classes. It consists of the following: a balanced subset containing about 20,000 videos, a test subset containing 18,000 videos, and an unbalanced training subset containing about 2 million videos. In image and text core modality training, we use the balanced set of 16,000 for audio-video and audio-text alignment respectively. For the zero-shot evaluation in Table.1, we use the test set and compute logits for each class using textual class names. 16,000 data pairs are used for training. During the text core modality model training and zero-shot evaluation we use prompt templates for class names as described later in Appendix.C.1. The metric used is mAP.

AudioCaps (Kim et al., 2019) is a dataset of audio-visual clips from YouTube with textual descriptions. It consists of clips from the AudioSet dataset. Following ImageBind, we used the splitting method provided in Oncescu et al. (2021) to remove clips that overlap with the VGGSound dataset. We obtain 48,198 training segments, 418 validation segments, and 796 test segments. We use only the test set for zero-shot evaluation of our model. The task is text $\rightarrow$ audio retrieval and is evaluated using recall@K.

ESC-50 (Piczak, 2015) is used to perform a zero-shot evaluation of the learned representations. The task here is "Environmental Sound Categorization" (ESC). It consists of 2000 5 s audio clips organized into 50 categories. In this work, we make zero-shot predictions for evaluation. The metric used is the accuracy of the top-1 accuracy.

VGG-S (Chen et al., 2020) contains approximately 200,000 video clips of 10 seconds in length annotated with 309 sound categories, including human actions, sound-producing objects, and human-object interactions. We performed zero-shot classification and RGB$\rightarrow$ Audio retrieval using only the audio from the test set. Evaluations are performed using top-1 accuracy for zero-shot classification and Recall@K for RGB$\rightarrow$ Audio retrieval.

Table 6: Training Settings in ImageCore model

Config	**Audio**	**Depth**	**Infrared**
Encoder	ViT-Huge		
Number of Heads	12	8	12
Optimizer	AdamW		
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.95$		
Epochs	8	2	2
Learning rate	5e-4	5e-4	1e-4
Temperature	0.07	0.2	0.1
Weight decay	0.2	0.2	0.05
Batch size	512	256	256
Learning rate schedule	Cosine decay		

AVE (Tian et al., 2018) contains 4,143 YouTube videos across 28 event categories and videos in the AVE dataset that are temporally labeled with audiovisual event boundaries. Evaluations were performed using top-1 accuracy for zero-shot classification and Recall@K for RGB→ Audio retrieval with the highest accuracy.

Clotho (Drossos et al., 2019) is an audio dataset with textual descriptions from the Freesound platform. It consists of a development set and a test set containing 2893 audio clips and 1045 audio clips respectively, each associated with 5 descriptions. We consider the text → audio retrieval task and treat each of the 5 associated descriptions as a separate test query, which is then retrieved from the set of audio clips. The metric used is recall@K, i.e., a given test query is assumed to be solved correctly if the base fact audio is retrieved in the first K audio clips retrieved.

SUN (Song et al., 2015) contains about 10,000 RGB-D images. We follow ImageBind to post process the depth maps in three steps- 1) in-filled depth values, 2) convert them to disparity for scale normalization and 3) limited the minimum and maximum depth to 0.01 and 10 meters respectively. We use training split (about 5,000 data pair) for training models. Specific, for text core modality model training, we use prompt templates for the class names as described later in Appendix.C.1.

NYU-D (Nathan Silberman & Fergus, 2012) is used to evaluation by 80% samples. Through preprocessing, we limited the minimum and maximum depth of the depth images to 0.01 and 10 meters respectively. Following ImageBind, we performed a classification and reorganization process which produced a total of 10 scene categories. For zero-shot evaluation and RGB → Depth retrieval task, we use top-1 accuracy and Recall@1. We use prompt templates as described later in Appendix.C.1 in RGB → Depth retrieval task.

LLVIP (Jia et al., 2021) is an infrared spectral pedestrian object detection dataset. Following the ImageBind method, we extracted all people in the image and designated all other objects as background elements. This process resulted in a dataset containing 7622 "background" categories and 7954 "people" categories, which were subsequently used for binary classification tests. About 5,000 Infrared-RGB pairs are used to training. Besides, prompt templates as described later in Appendix.C.1 is used in zero-shot classification task. Since LLVIP is intended to be used for detect tasks and each RGB image is not text labeled, we use GPT4o to generate text annotations for each RGB image during the training process for text core modality.

Imagenet-1K Russakovsky et al. (2015) encompasses 1,000 object classes and comprises 1.28M images for training, 5000 images for validation, and 100,000 images for testing. Building upon this foundation, Imagenet-1K-VL-Enriched [3] enhances Imagenet-1K dataset by including image captions, bounding boxes, and corrected label information. Caption & image pairs from Imagenet-1K-VL-Enriched training split are used for training diffusion model.

[3]Dataset can be found in https://huggingface.co/datasets/visual-layer/imagenet-1k-vl-enriched

Table 7: Training Settings in LanguageCore model

Config	**Audio**	**Depth**	**Infrared**
Encoder	ViT-Huge		
Number of Heads	12	8	12
Optimizer	AdamW		
Optimizer Momentum	$\beta_1 = 0.9, \beta_2 = 0.95$		
Epochs	8	4	4
Learning rate	1e-4	5e-4	1e-4
Temperature	0.05	0.2	0.2
Weight decay	0.2	0.1	0.05
Batch size	512	256	256
Learning rate schedule	Cosine decay		

C Implementation Details

We now describe the implementation details used in this work. In table 6 and 7, we detail the settings used to train each of the modalities. Our experiments were done on 4 24GB 4090 GPUs and 4 48GB A40 GPUs.

C.1 Prompt Templates

For all evaluations, we use the default set of templates from CLIP(Radford et al., 2021). It is worth mentioning that we use the same templates for non visual modalities like audio and depth as well since we only use semantic/textual supervision associated with images.

C.2 Model Architecture

Diffusion Model We deployed two symmetric decode-only diffusion models, one for generating text embeddings from image embeddings and another for the reverse. Both models use the same transformer architecture with a 1024-dimensional embedding, 8 attention heads of 128 dimensions each, and 6 layers and the number of training parameters is about 30M. Training involves 100 timesteps, with a 5% embedding dropout to facilitate classifier-free guidance, optimized using an AdamW optimizer, a learning rate of 1e-4, and a batch size of 128, following the loss function1 of our methodology documentation. For the model where the image is the core modality, the image embedding serves as the input, and the text embedding is the output to generate. Conversely, when text is the core modality, text embedding is used as input to generate the image embedding. This approach allows us to efficiently handle and transform data within the latent space, reducing computational demands. This setup not only ensures the capability of model in cross-modal generation but also enhances its performance in generating high-quality, contextually accurate outputs across different modalities.

Multimodal Encoders For image core model and language core model, we use the same structure of encoder on the same modality. Following Girdhar et al. (2023) we use 12-layer, 1024-dimensional vision transformer with a patch size of 16 and a stride of 10 for the VISION, AUDIO, DEPTH, and INFRARED modalities. For video data, our strategy includes capturing two frames every two seconds to optimize processing efficiency. Besides, we used 128 mel-spectrogram partitions to convert 2 seconds of audio sampled at 16kHz into a spectrogram. Similarly, thermal and depth images are treated as single-channel inputs and encoded using the same ViT architecture, facilitating consistent handling across these modalities. The encoders for both the image and language core models are initialized with weights from the ImageBind-Huge and LanguageBind pre-trained models respectively for enhancing learning efficiency and demonstrating the transferability of our TangentBind approach. This leverages their advanced pre-trained features to accelerate convergence and improve generalization across varied multimodal applications.

Temperature **Tangent Term** 6 and infoNCE 2 use the same temperature τ during training for encoder of the same modality. In our experiments, we found that fixed temperatures worked best by

comparing learnable and fixed temperatures. The experiments show the temperature with the best effect for each modality in table.6 and .7.

C.3 Ablation Details

ResNet We employed the standard ResNet50(He et al., 2015) architecture provided by PyTorch. The ResNet50 model has a nearly number of training parameters with the diffusion model employed, suggesting that both architectures share comparable complexity and computational demands, which allows for an equitable comparison of their performance in analogous tasks. To tailor the model for our embedding-based task, the output dimension of the final fully connected layer was modified to produce embeddings with a dimensionality of 1024. These embeddings are subsequently normalized to lie on the unit hypersphere to facilitate the use of cosine similarity measures in subsequent analyses. For training the modified ResNet50, we utilized the L2 loss function, which is well-suited for embedding normalization by encouraging the model to minimize the Euclidean distance between the predicted and target embeddings. The optimizer of choice was AdamW, the learning rate was set to 1e-4. The training was conducted with a batch size of 128. Our training procedure mirrored that of the diffusion model in terms of dataset usage; specifically, we trained on the train split of the ImageNet-1K-VL-Enriched dataset. During training, the core modality embeddings were utilized as inputs, while the embeddings of an alternative modality served as labels.

C-MCR We employed the C-MCR(Wang et al., 2023b) method to generate cross-modal embeddings, leveraging the training split of the ImageNet-1K-VL-Enriched dataset to obtain paired $\langle x_i^{\text{Image}}, x_i^{\text{text}} \rangle$ embeddings that serve as image and text memories. This approach is rooted in the framework established by the C-MCR methodology, where embeddings for both modalities are computed based on

$$\hat{x}_i^{\text{image}} = \sum_{i=1}^{N} \frac{\exp(\text{sim}(x^{\text{text}}, x_i^{\text{image}}))}{\sum_{j=1}^{N} \exp(\text{sim}(x^{\text{text}}, x_j^{\text{image}}))} x_i^{\text{image}}; \quad \hat{x}_i^{\text{text}} = \sum_{i=1}^{N} \frac{\exp(\text{sim}(x^{\text{image}}, x_i^{\text{text}}))}{\sum_{j=1}^{N} \exp(\text{sim}(x^{\text{image}}, x_j^{\text{text}}))} x_i^{\text{text}}, \tag{36}$$

and subsequently normalize to lie on the unit hypersphere. As illustrated in Figure 6, the CDF curve corresponding to the C-MCR method approximates a straight line. This outcome primarily arises due to the SUN dataset, which consists of scene data that does not align well with the ImageNet-1K-VL-Enriched dataset. During our experiments, we observed that whether text or image served as the core modality, the weights $\frac{\exp(\text{sim}(x^{\text{text}}, x_i^{\text{image}}))}{\sum_{j=1}^{N} \exp(\text{sim}(x^{\text{text}}, x_j^{\text{image}}))}$ and $\frac{\exp(\text{sim}(x^{\text{image}}, x_i^{\text{text}}))}{\sum_{j=1}^{N} \exp(\text{sim}(x^{\text{image}}, x_j^{\text{text}}))}$ in (36) assigned to each embedding from the SUN dataset on the ImageNet1K were minutely different. This minimal variation led to the generation of nearly identical embeddings for both modalities, hence the nearly linear CDF curve observed. This phenomenon underscores a critical aspect of the C-MCR method: its strong dependence on the memory.

VAE We implemented a VAE(Kingma & Welling, 2013) where both the encoder and decoder components are constructed using convolutional neural networks (CNNs)(LeCun et al., 1998). The architecture of the encoder and decoder are symmetric, each comprising ten layers with the following input channels of convolutional layers: [32, 32, 64, 64, 128, 128, 256, 256, 512, 512]. This VAE model possesses a comparable number of training parameters to the diffusion model used, indicating that both architectures are similarly complex and computationally demanding, facilitating a fair comparison of their performance across similar tasks. Each convolutional layer is defined with a kernel size of 3, a stride of 2, and padding of 1. Post convolution, the decoder maps the latent space representation to embedding with dimensionality of 1024. This representation is then normalized to lie on the unit hypersphere. The loss function employed is the L2 loss, similar to that used in ResNet50 architectures, which helps in minimizing the distance between the reconstructed outputs and the actual inputs, thereby ensuring better fidelity in the generated samples. AdamW optimizer is chosen with a learning rate of 1e-4 and a batch size of 128. Training was conducted using the training split of the ImageNet-1K-VL-Enriched dataset. In this setup, core modality embeddings were used as inputs, and embeddings from another modality served as labels. This training approach not only facilitates effective learning of cross-modal representations but also ensures that the VAE is

capable of generating high-quality embeddings that are highly representative of the input data across different modalities.