FreeBind: Free Lunch in Unified Multimodal Space via Knowledge Fusion

Zehan Wang [* 1 2] **Ziang Zhang** [* 1] **Xize Cheng** [1] **Rongjie Huang** [1] **Luping Liu** [1] **Zhenhui Ye** [1] **Haifeng Huang** [1] **Yang Zhao** [3] **Tao Jin** [1] **Peng Gao** [2] **Zhou Zhao** [1 2]

Abstract

Unified multi-model representation spaces are the foundation of multimodal understanding and generation. However, the billions of model parameters and catastrophic forgetting problems make it challenging to further enhance pre-trained unified spaces. In this work, we propose **FreeBind**, an idea that treats multimodal representation spaces as basic units, and freely augments pre-trained unified space by integrating knowledge from extra expert spaces via "space bonds". Specifically, we introduce two kinds of basic space bonds: 1) **Space Displacement Bond** and 2) **Space Combination Bond**. Based on these basic bonds, we design **Complex Sequential & Parallel Bonds** to effectively integrate multiple spaces simultaneously. Benefiting from the modularization concept, we further propose a coarse-to-fine customized inference strategy to flexibly adjust the enhanced unified space for different purposes. Experimentally, we bind ImageBind with extra image-text and audio-text expert spaces, resulting in three main variants: ImageBind++, InternVL$_{IB}$ and InternVL$_{IB}$++. These resulting spaces outperform ImageBind on 5 audio-image-text downstream tasks across 9 datasets. Moreover, via customized inference, it even surpasses the advanced audio-text and image-text expert spaces. Our code and checkpoints are released at https://github.com/zehanwang01/FreeBind

1. Introduction

Unified multimodal representation aims to learn a semantically shared representation space for many modalities (such as audio, image, language and 3D point cloud) (Girdhar et al., 2023; Wang et al., 2023a; Guzhov et al., 2022; Wu

---
[*]Equal contribution [1]Zhejiang University [2]Shanghai AI Lab [3]ByteDance. Correspondence to: Zhou Zhao <zhaozhou@zju.edu.cn>.
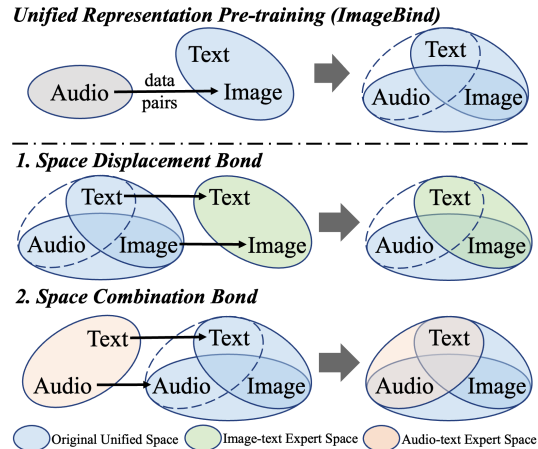
Figure 1. High-level overview of FreeBind. We propose two basic kinds of space bonds: space displacement bond and space combination bond, to efficiently augment unified space by integrating knowledge of extra expert spaces.

et al., 2022; 2023b; Xue et al., 2023a;b; Liu et al., 2023b). As an important foundation for multimodal understanding (Liu et al., 2023a; Zhu et al., 2023b; Wang et al., 2023b; Huang et al., 2023a; Han et al., 2023) and generation (Huang et al., 2023b; Tang et al., 2023; Wu et al., 2023a; Rombach et al., 2022; Podell et al., 2023), a unified multimodal space is crucial for artificial general intelligence.

Existing advanced unified multimodal representation space (Zhu et al., 2023a; Girdhar et al., 2023; Zhou et al., 2023) are built on billion-level data and parameters. Learning such a unified space demands exceedingly costly computational resources, and further enhancing the pre-trained space often requires huge training resources or faces the catastrophic forgetting problem. These challenges limit the further development of unified multimodal representation.

In this paper, we propose **FreeBind**, an efficient knowledge fusion scheme to enhance pre-trained unified space. Specifically, we propose to bind unified space (i.e., space for many modalities) with expert space (i.e., space focus on single modalities pair) via two basic "space bonds":

1) **Space Displacement Bond.** We align the unified space to the expert space to inherit all the knowledge of the expert space. However, remapping the entire unified space compromises the knowledge of unified space. Additionally, when

1

integrating multiple expert spaces, cascaded displacements are susceptible to cumulative errors. Overall, displacement bond is a radical knowledge fusion solution that sacrifices some information from the unified space in exchange for full expert knowledge.

2) **Space Combination Bond.** Complementary to displacement bond, we also propose a moderate knowledge fusion scheme called combination bond, which aligns expert space to unified space. Since unified space is frozen, its knowledge can be preserved and we can combine multiple expert spaces in parallel. However, as the expert space is reprojected, the combination bond can only partially integrate the knowledge of expert space.

Based on these two complementary basic bonds, we further propose **Complex Sequential & Parallel bonds** to effectively integrate multiple expert spaces simultaneously. Specifically, due to the pivotal role of image-text representations in unified spaces, we first integrate the unified space with advanced image-text expert space via displacement bond and tune the product to repair its lost knowledge. Then, we combine extra expert spaces via combination bond in parallel to further enhance the unified space. For the final resulting space, we design a coarse-to-fine customized inference strategy to flexibly suit different applications by selecting modules and adjusting combining factors.

To demonstrate the effectiveness of FreeBind, we study practical application on the audio-image-text unified space of ImageBind (Han et al., 2023). By integrating one image-text and two audio-text expert spaces, we construct state-of-the-art audio-image-text space that significantly surpasses ImageBind. Furthermore, leveraging the flexibility of customized inference, we achieve even better performance in image-text or audio-text tasks than the source expert spaces.

Our contributions can be summarized as follows:

- We present **FreeBind**, an approach that conceptualizes multimodal spaces as basic unit and fuses the knowledge of multimodal representation spaces through space bonds.

- We propose two complementary basic bonds between two spaces: displacement and combination bond. Building on these foundations, we further introduce complex sequential & parallel bonds for integrating multiple spaces simultaneously.

- We design a simple yet effective projector learning pipeline and propose a mixture-of-projectors strategy to strengthen the robustness of space alignments.

- We employ **FreeBind** on ImageBind to verify its effectiveness. By integrating advanced image-text and audio-text expert spaces, we establish a state-of-the-art audio-image-text space with limited resources.

2. Related work

2.1. Multimodal Representation Space

Multimodal representation space aims to embed different modality inputs into a joint space. Recent multimodal space research mainly focuses on two aspects: building stronger alignment between two modalities (i.e., expert spaces) or enabling more modalities input (i.e., unified spaces).

Current expert space achieves impressive performance on various modality pairs. By collecting a large collection of image-text pairs, CLIP (Radford et al., 2021) and ALIGN (Jia et al., 2021) show impressive performance and generalization ability. The recent InternVL (Chen et al., 2023) scale up the visual encoder to 6 billion parameters and achieves the most advanced performance on most vision-language downstream tasks. The success of vision-language representation inspires more research to explore contrastive representation on other modality pairs. CLAP (Wu et al., 2023b) learns high-quality audio-text representation space via massive audio-text pairs, while VideoCLIP (Xu et al., 2021) obtains shared video and text representations from video-text data. In addition to general multimodal representations, some recent researches (Zhang et al., 2022; Elizalde et al., 2023a;b) attempt to develop domain-specific pre-trained multimodal spaces, such as music or speech versions of CLAP (Wu et al., 2023b), and image-text space specifically learned on medical images (Zhang et al., 2022).

On the other hand, many recent works have tried to develop a unified representation space for more than three modalities to support more diverse applications. These unified space learning approaches collect massive multimodal data pairs and train encoders to align new modalities with a pre-trained image-text space. AudioCLIP (Guzhov et al., 2022) and WAV2CLIP (Wu et al., 2022) align audio inputs to CLIP by constructing audio-text-image data. Recent ImageBind (Girdhar et al., 2023) collects and organizes image-paired data of four modalities, and learns encoders of these modalities that aligned to CLIP space. Similarly, LanguageBind (Zhu et al., 2023a) align encoders of different modalities to CLIP via constructing language paired data.

Our method aims to integrate the knowledge of expert spaces into a pre-trained unified space, thereby enhancing the unified space with limited resources and enabling it to benefit from breakthroughs of expert spaces. Moreover, via customizing the inference process, the augmented unified space can even surpass expert spaces in terms of their expertise.

2.2. Knowledge Fusion in Multimodal Representation

Recent C-MCR (Wang et al., 2023d) and Ex-MCR (Wang et al., 2023c) first study how to learn new knowledge by integrating multiple expert spaces. Specifically, C-MCR builds expert space by connecting two expert spaces with

one shared modality. Subsequently, Ex-MCR proposes extending one space to another instead of connecting both to build a new one. This extending paradigm facilitates better modality scalability and can build a unified space by extending multiple expert spaces into a based expert space via their shared modalities.

Although these methods also focus on knowledge fusion in multimodal space, our method is fundamentally different from them. C-MCR and Ex-MCR are specifically designed for expert spaces with one and only one shared modality. Such strict usage requirements limit their application. In contrast, our method aims to augment pre-trained unified spaces with expert spaces, which involve multiple shared modalities and more general application scenarios.

3. Method

We introduce **FreeBind**, a training-efficient method designed to enhance pre-trained unified space through knowledge fusion. This section explores its application in augmenting audio-image-text unified space with image-text and audio-text expert spaces. Initially, we formulate the problem, followed by outlining two basic bonds and their composition. Finally, we delve into the customized coarse-to-fine inference strategy.

3.1. Problem formulation

The audio-image-text unified space are denoted as $\mathcal{A}^u \mathcal{V}^u \mathcal{T}^u$. Correspondingly, the image-text and audio-text expert spaces can be represented as $\mathcal{V}^{vt} \mathcal{T}^{vt}$ and $\mathcal{A}^{at} \mathcal{T}^{at}$ respectively. The superscripts ^{u}, ^{vt}, and ^{at} signify the unified space, image-text and audio-text expert space respectively. With these symbols, the displacement and combination bonds in Figure 1 can be expressed as:

$$\mathcal{A}^u \mathcal{V}^u \mathcal{T}^u + \mathrm{d}(\mathcal{V}^{vt} \mathcal{T}^{vt}) \rightarrow \hat{\mathcal{A}}^u (\hat{\mathcal{V}}^u_{1-\lambda_v} \mathcal{V}^{vt}_{\lambda_v})(\hat{\mathcal{T}}^u_{1-\lambda_t} \mathcal{T}^{vt}_{\lambda_t}) \quad (1)$$

$$\mathcal{A}^u \mathcal{V}^u \mathcal{T}^u + \mathrm{c}(\mathcal{A}^{at} \mathcal{T}^{at}) \rightarrow (\mathcal{A}^u_{1-\sigma_a} \hat{\mathcal{A}}^{at}_{\sigma_a}) \mathcal{V}^u (\mathcal{T}^u_{1-\sigma_t} \hat{\mathcal{T}}^{at}_{\sigma_t}) \quad (2)$$

where superscript $\hat{}$ means the representations are remapped, $\mathrm{d}(\cdot)$ and $\mathrm{c}(\cdot)$ indicates displacement and combination bond, respectively. (λ_v, λ_t) and (σ_a, σ_t) are the combining factors of expert spaces. The output spaces in Equation 1 and 2 are illustrated in the 1.3 and 2.3 part of Figure 2, and the $(\hat{\mathcal{V}}^u_{1-\lambda_v} \mathcal{V}^{vt}_{\lambda_v})$ can be formulated as:

$$(\hat{\mathcal{V}}^u_{1-\lambda_v} \mathcal{V}^{vt}_{\lambda_v}) = (1-\lambda_v)\hat{\mathcal{V}}^u + \lambda_v \mathcal{V}^{vt} \quad (3)$$

To reflect the pre-trained knowledge of unified space, some unpaired images V, texts T, and audios A are encoded into the unified space. The corresponding features are denoted as $\mathbf{V}^u \in \mathbb{R}^{n_v \times d_u}$, $\mathbf{T}^u \in \mathbb{R}^{n_t \times d_u}$, $\mathbf{A}^u \in \mathbb{R}^{n_a \times d_u}$, where d_u is the dimension of the unified space. At the same time, the same data is also encoded into expert

spaces, serving as bonds between expert spaces and unified space. For image-text expert space, the embeddings are denoted as $\mathbf{V}^{vt} \in \mathbb{R}^{n_v \times d_{vt}}$, $\mathbf{T}^{vt} \in \mathbb{R}^{n_t \times d_{vt}}$, while the embeddings in audio-text expert space can be represented as $\mathbf{A}^{at} \in \mathbb{R}^{n_a \times d_{at}}$, $\mathbf{T}^{at} \in \mathbb{R}^{n_t \times d_{at}}$.

3.2. Basic Space Bonds

3.2.1. PSEUDO DATASETS COLLECTION

To fuse different multimodal spaces, the initial step involves capturing correlations between different spaces and modalities. To this end, we collect robust and diverse pseudo datasets to bond two different spaces.

Taking the collection of pseudo datasets collection between image-text expert space and unified space as an example, the embeddings of the expert and unified spaces are $\mathbf{T}^{vt}$, $\mathbf{V}^{vt}$, $\mathbf{T}^u$, $\mathbf{V}^u$ and $\mathbf{A}^u$. The correlation between different modalities can be obtained through the inherent multimodal semantic alignment of embeddings $\mathbf{T}^{vt}$-$\mathbf{V}^{vt}$ and $\mathbf{T}^u$-$\mathbf{V}^u$-$\mathbf{A}^u$ within each space. On the other hand, the correlation between different spaces can be established via the native semantic consistency of $\mathbf{T}^{vt}$-$\mathbf{T}^u$, $\mathbf{V}^{vt}$-$\mathbf{V}^u$ due to the same data source. Combining these two kinds of correlation, we can obtain pseudo multimodal pairs from unpaired or partially-paired data. Furthermore, we retrieve pseudo pairs starting from different modalities respectively, which brings more diverse and comprehensive datasets.

When integrating the unified and image-text spaces, text and image are the shared modalities. The pseudo pairs aggregation process starting from shared modalities (i.e., text and image) can be respectively expressed as:

$$\tilde{\mathbf{T}}^{vt} = \mathbf{T}^{vt}; \ \tilde{\mathbf{V}}^{vt} = \mathrm{softmax}(\tilde{\mathbf{T}}^{vt}\mathbf{V}^{vt\top})\mathbf{V}^{vt};$$
$$\tilde{\mathbf{T}}^u = \mathbf{T}^u; \ \tilde{\mathbf{V}}^u = \mathrm{softmax}(\tilde{\mathbf{T}}^u\mathbf{V}^{u\top})\mathbf{V}^u; \quad (4)$$
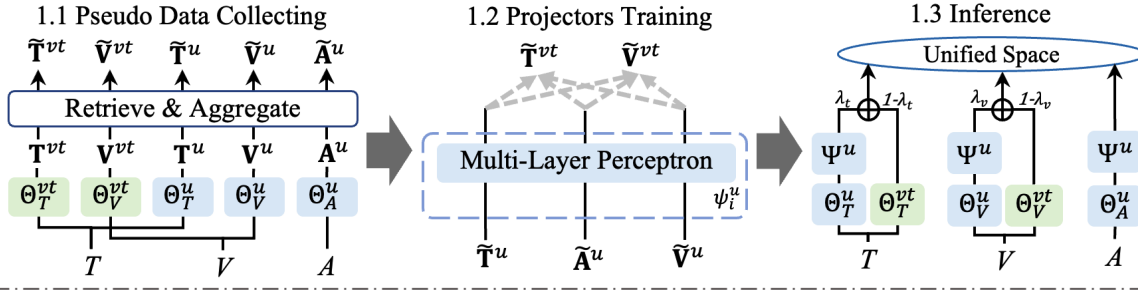$$\tilde{\mathbf{A}}^u = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{A}^{u\top})\mathbf{A}^u$$

$$\tilde{\mathbf{V}}^{vt} = \mathbf{V}^{vt}; \ \tilde{\mathbf{T}}^{vt} = \mathrm{softmax}(\tilde{\mathbf{V}}^{vt}\mathbf{T}^{vt\top})\mathbf{T}^{vt};$$
$$\tilde{\mathbf{V}}^u = \mathbf{V}^u; \ \tilde{\mathbf{T}}^u = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{T}^{u\top})\mathbf{T}^u; \quad (5)$$
$$\tilde{\mathbf{A}}^u = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{A}^{u\top})\mathbf{A}^u$$

The dataset collection from non-shared modality (i.e., audio) can be formulated as:

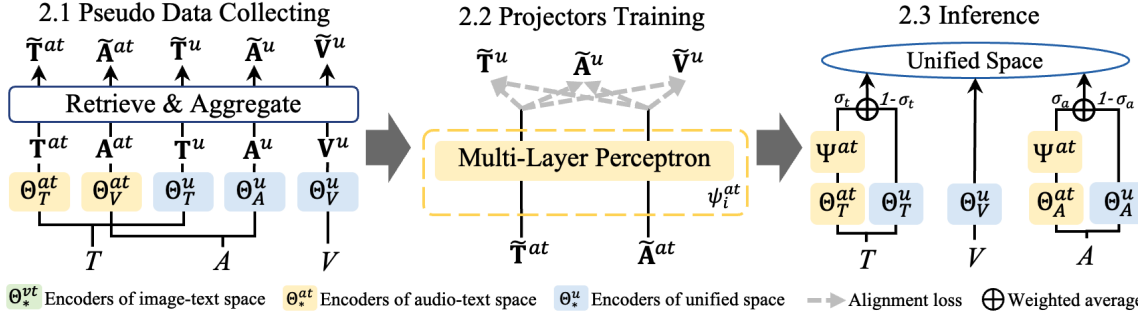$$\tilde{\mathbf{A}}^u = \mathbf{A}^u;$$
$$\tilde{\mathbf{V}}^u = \mathrm{softmax}(\tilde{\mathbf{A}}^u\mathbf{V}^{u\top})\mathbf{V}^u; \ \tilde{\mathbf{V}}^{vt} = \mathrm{softmax}(\tilde{\mathbf{A}}^u\mathbf{V}^{u\top})\mathbf{V}^{vt};$$
$$\tilde{\mathbf{T}}^u = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{T}^{u\top})\mathbf{T}^u; \ \tilde{\mathbf{T}}^{vt} = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{T}^{u\top})\mathbf{T}^{vt}$$
$$(6)$$

where the superscript $\tilde{}$ indicates embeddings are processed to be pseudo embedding pairs. The sets of pseudo pairs $(\tilde{\mathbf{T}}^b, \tilde{\mathbf{V}}^b, \tilde{\mathbf{T}}^u, \tilde{\mathbf{V}}^u, \tilde{\mathbf{A}}^u)$ collected from text, image and audio are denoted as D_T, D_V and D_A, respectively.

Figure 2. The pipeline of basic space displacement bond and space combination bond.

When integrating an audio-text expert space with a unified space, the shared modalities are audio and text. The overall pseudo dataset collection process is similar to the above, and the detailed equations can be found in the Appendix.

3.2.2. SPACE ALIGNMENTS

Single Projector Training The previous space alignment methods, C-MCR and Ex-MCR, utilize intricate inter-space and intra-space alignment loss to train their well-designed projector. Their tasks aims to align two expert spaces with one and only one shared modality, and the intra-space alignment loss is used to better transfer the robust connections learned from the shared modality to non-shared modalities.

In contrast, our objective is to enhance a pre-trained unified space by integrating expert spaces. Given unified space typically covers most modality inputs, and the modalities of expert spaces are the subset of unified space. The space alignment learned from the multiple shared modalities is much stronger than that learned from only one shared modality. Therefore, there is no motivation for using intra-space alignment loss here, and previous complex learning pipeline may introduce a negative impact on generalization.

As a result, we propose a more plain space alignment pipeline, which experimentally shows better performance. One projector ψ_i consists of simple multi-layer perceptrons (MLP). For the learning objective, we only compute the InfoNCE loss, denoted as info$(\cdot, \cdot)$, between features of different spaces. The training loss for displacement bond in

Figure 2 can be expressed as:

$$
\begin{aligned}
L = {}& \text{info}(\tilde{\mathbf{T}}^{vt}, \psi_i^u(\tilde{\mathbf{T}}^u)) + \text{info}(\tilde{\mathbf{T}}^{vt}, \psi_i^u(\tilde{\mathbf{V}}^u)) \\
& + \text{info}(\tilde{\mathbf{T}}^{vt}, \psi_i^u(\tilde{\mathbf{A}}^u)) + \text{info}(\tilde{\mathbf{V}}^{vt}, \psi_i^u(\tilde{\mathbf{T}}^u)) \quad (7) \\
& + \text{info}(\tilde{\mathbf{V}}^{vt}, \psi_i^u(\tilde{\mathbf{V}}^u)) + \text{info}(\tilde{\mathbf{V}}^{vt}, \psi_i^u(\tilde{\mathbf{A}}^u))
\end{aligned}
$$

and the loss for the combination bond in Figure 2 is:

$$
\begin{aligned}
L = {}& \text{info}(\psi_i^{at}(\tilde{\mathbf{T}}^{at}), \tilde{\mathbf{T}}^u) + \text{info}(\psi_i^{at}(\tilde{\mathbf{T}}^{at}), \tilde{\mathbf{V}}^u) \\
& + \text{info}(\psi_i^{at}(\tilde{\mathbf{T}}^{at}), \tilde{\mathbf{A}}^u) + \text{info}(\psi_i^{at}(\tilde{\mathbf{V}}^{at}), \tilde{\mathbf{T}}^u) \quad (8) \\
& + \text{info}(\psi_i^{at}(\tilde{\mathbf{V}}^{at}), \tilde{\mathbf{V}}^u) + \text{info}(\psi_i^{at}(\tilde{\mathbf{V}}^{at}), \tilde{\mathbf{A}}^u)
\end{aligned}
$$

Mixture-of-Projectors Strategy Inspired by the ensemble learning and mixture-of-expert methods, we propose the mixture-of-projectors strategy, which learns multiple projectors with different training data and ensembles them to achieve more robust alignment and more discriminative representations. Specifically, we first sample t subsets from the whole dataset D, denoted as $\{D_1, D_2, \ldots, D_t\}$. Then we train projector ψ_i on D_i respectively, and finally get a group of projectors $\Psi = \{\psi_1, \psi_2, \ldots, \psi_t\}$. The output of Ψ is the mean pool of all t projectors.

3.2.3. INFERENCE

In the product space, one modality may have multiple representations from different sources. As illustrated in parts 1.3 and 2.3 in Figure 2, we simply weighted average the representations of the same modality but from different sources.

3.3. Complex Sequential & Parallel Bonds

Based on these two basic bonds, we can easily construct various complex bonds, but which way is more effective for integrating multiple spaces still needs to be explored.

Typical unified space learning method aligns encoders of other modalities to pre-trained image-text space via massive paired data. Therefore, image-text representation is the foundation of unified spaces and directly determines its potential. Considering the properties of basic bonds and the importance of image-text space, we propose sequential & parallel bonds, which consist of two stages:

1) Sequential Displacement. Given the pivotal role of image-text representation and the value of image-text knowledge (requiring training encoders of billion-level parameters on billion-level data), we integrate advanced image-text space via displacement bond and tuning on data of other modalities to repair the missing knowledge of unified space.

2) Parallel Combination. After obtaining stronger image-text representations, we integrate expert spaces of other modalities in parallel via combination bonds. Since these expert spaces are independently connected to the same frozen unified space, we can further enhance the unified space and perform flexible customized inference.

Take the integration of advanced image-text space and n audio-text spaces as an example. Based on the displacement product, $\hat{\mathcal{A}}^u(\hat{\mathcal{V}}^u_{1-\lambda_v}\mathcal{V}^{vt}_{\lambda_v})(\hat{\mathcal{T}}^u_{1-\lambda_t}\mathcal{T}^{vt}_{\lambda_t})$, the combination bond of the i-th audio-text space can be formulated as:

$$\hat{\mathcal{A}}^u(\hat{\mathcal{V}}^u_{1-\lambda_v}\mathcal{V}^{vt}_{\lambda_v})(\hat{\mathcal{T}}^u_{1-\lambda_t}\mathcal{T}^{vt}_{\lambda_t}) + c(\mathcal{A}^{at_i}\mathcal{T}^{at_i}) \rightarrow$$
$$(\hat{\mathcal{A}}^u_{1-\sigma_a}\hat{\mathcal{A}}^{at_i}_{\sigma_a})(\hat{\mathcal{V}}^u_{1-\lambda_v}\mathcal{V}^{vt}_{\lambda_v})[(\hat{\mathcal{T}}^u_{1-\lambda_t}\mathcal{T}^{vt}_{\lambda_t})_{1-\sigma_t}\hat{\mathcal{T}}^{at_i}_{\sigma_t}] \quad (9)$$

Since n audio-text spaces are aligned to the same unified space, $\hat{\mathcal{A}}^{at}_i$ and $\hat{\mathcal{T}}^{at}_i$ can be flexibly combined during inference to obtain customized representations. The space combined all the n audio-text space can be formulated as:
$(\hat{\mathcal{A}}^u_{1-\sigma_a}\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{A}}^{at_i}_{\sigma_a})(\hat{\mathcal{V}}^u_{1-\lambda_v}\mathcal{V}^{vt}_{\lambda_v})[(\hat{\mathcal{T}}^u_{1-\lambda_t}\mathcal{T}^{vt}_{\lambda_t})_{1-\sigma_t}\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{T}}^{at_i}_{\sigma_t}]$,
and its combining process can be expressed as:

$$(\hat{\mathcal{A}}^u_{1-\sigma_a}\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{A}}^{at_i}_{\sigma_a}) = (1-\sigma_a)\hat{\mathcal{A}}^u + \frac{\sigma_a}{n}\sum_{i=1}^{n}\hat{\mathcal{A}}^{at_i};$$

$$[(\hat{\mathcal{T}}^u_{1-\lambda_t}\mathcal{T}^{vt}_{\lambda_t})_{1-\sigma_t}\frac{1}{n}\sum_{i=1}^{n}\hat{\mathcal{T}}^{at_i}_{\sigma_t}] = (1-\sigma_t)(\hat{\mathcal{T}}^u_{1-\lambda_t}\mathcal{T}^{vt}_{\lambda_t}) + \frac{\sigma_t}{n}\sum_{i=1}^{n}\hat{\mathcal{T}}^{at_i} \quad (10)$$

3.4. Coarse-to-Fine Customized Inference

In addition to the computationally efficient training process, the product of FreeBind can customize its inference to various applications. To fully realize its potential, we propose a coarse-to-fine customized inference strategy:

1) Coarse-grained Combined Modules Selection. Combination bonds align multiple expert spaces into a unified space. Therefore, during inference, we can flexibly select any aligned expert spaces to obtain gains of specific aspects.

2) Fine-grained Combining Factors Adjustment. In addition to selecting different modules, we can also customize the enhanced unified space in a fine-grained manner by changing the combination weights of different expert spaces.

Using the inference process in Equation 10 as an example, we can freely select any combination of the n aligned audio-text spaces to construct unified spaces tailored to specific aspects. Additionally, a small (σ_a, σ_t) implies partial absorption of audio-text knowledge, and moderate knowledge fusion can enhance both audio-text and audio-image performance while maintain advanced image-text ability. Conversely, a larger value for (σ_a, σ_t) leads to superior audio-text performance at the expense of other alignments. Notably, the impact of combination factors on performance is regular and robust. As depicted in Figure 3, most settings either yield a versatile space that surpasses the original unified space or generate an expertise space that exceeds the source expert spaces in their fields.

4. Experiment and Discussions

4.1. Implementation Details

Data and Pre-trained Models For both bonds, we employ 2.3M unpaired texts, 1.3M images, and 1.8M audios, following (Wang et al., 2023d). We optionally use the audio-image pairs in AudioSet (Gemmeke et al., 2017) (the audio pre-training dataset of ImageBind) to fine-tune the audio encoder. We enhance the unified audio-image-text space of ImageBind by integrating one image-text expert space, InternVL-C (Chen et al., 2023) and two audio-text expert spaces, two versions of CLAPs (Wu et al., 2023b).

Training and Inference For both kinds of basic bond, the temperature of softmax in data collection is 1/100, and the temperature of InfoNCE loss is 1/50. We leverage the all possible combination of the elements D_T, D_V and D_A as the sampled subsets in mixture-of-projector (i.e., D_T, D_V, D_A, D_{TV}, D_{TA}, D_{VA}, D_{TVA}), resulting in 7 projectors of each Ψ. All our experiments are conducted on a single 4090 GPU. We use Adam (Kingma & Ba, 2014) optimizer with a learning rate of 1e-3 and batch size of 4096 for both bond. The displacement bond is trained for 5 epochs, while the combination bond is trained for 20 epochs.

Evaluation Protocols We comprehensively evaluate Free-Bind on nine datasets over five zero-shot downstream tasks. The evaluation tasks, datasets, metrics, and the number of test samples are summarized in Table 1.

Table 1. Summary of downstream tasks and datasets.

Task	Dataset	Metric	#Samples
Audio-Text Retrieval	AudioCaps (Kim et al., 2019)	Recall	964
	Clotho (Drossos et al., 2020)	Recall	1045
Audio-Image Retrieval	VGG-SS (Chen et al., 2021)	Recall	5158
	FlickrNet (Senocak et al., 2018)	Recall	5000
Image-Text Retrieval	COCO (Lin et al., 2014)	Recall	5000
	Flickr30k (Young et al., 2014)	Recall	1000
Audio Classification	ESC-50 (Piczak, 2015)	Acc	400
	AudioSet (Gemmeke et al., 2017)	mAP	19048
Image Classification	ImageNet (Deng et al., 2009)	Acc	50000

Table 2. Notations of different bond processes and their corresponding resulting space.

Bonds Combination	Product
ImageBind +c(CLAP$_m$)+c(CLAP$_g$)	ImageBind++
ImageBind+d(InternVL)	InternVL$_{IB}$
InternVL$_{IB}$+c(CLAP$_m$) + c(CLAP$_g$)	InternVL$_{IB}$++
InternVL$_{IB}^{\dagger}$+c(CLAP$_m$)+c(CLAP$_g$)	InternVL$_{IB}^{\dagger}$++

4.2. Augmenting ImageBind

To show the effectiveness of the proposed methods, we augment the audio-image-text space of ImageBind with InternVL (image-text space), CLAP$_g$ (audio-text space for general purpose), and CLAP$_m$ (audio-text space for music purpose). For simplicity of expression, we summarize the notations of output space for different bonds combination in Table 2. The InternVL$_{IB}$ tuned on the audio-image dataset is denoted as InternVL$_{IB}^{\dagger}$. There are two standard settings of combining factors: Versatile (*Ver.*) and Audio-Text Expertise (*AT E.*)[1]. The zero-shot classification results are presented in Table 3, and the multimodal retrieval results can be found in Table 4.

Displacement Bond By integrating InternVL with ImageBind via displacement bond, the resulting unified space InternVL$_{IB}$ shows significantly better image-text performance than ImageBind. Additionally, its image-text retrieval accuracy even surpasses the source image-text space, InternVL. More importantly, despite the audio representation in InternVL$_{IB}$ is a remapped and degraded version of ImageBind's audio representations, InternVL$_{IB}$ achieves comparable audio-text and audio-image performance.

Combination Bond We try to integrate CLAPs for three kinds of unified space: ImageBind, InternVL$_{IB}$ and InternVL$_{IB}^{\dagger}$. The products are denoted as ImageBind++, InternVL$_{IB}$++ and InternVL$_{IB}^{\dagger}$++, respectively.

[1] The CLAPs combining factors (σ_a, σ_t) are (0.5, 0.1) for Versatile (*Ver.*) and (0.8, 0.5) for Audio-Text Expertise (*AT E.*). The InternVL combining factors (λ_v, λ_t) in InternVL$_{IB}$ are (0.9, 0.9).

Table 3. Results of zero-shot classification.

Models		Audio		Image
		ESC-50	AudioSet	ImageNet
Pre-trained Expert Space	InternVL	-	-	81.70
	CLAP$_g$	90.95	23.36	-
	CLAP$_m$	92.60	23.08	-
Pre-trained Unified Space	WAV2CLIP	17.40	0.71	60.58
	AudioCLIP	11.45	5.65	24.14
	C-MCR	66.05	11.15	23.16
	Ex-MCR	68.35	6.67	60.58
	ImageBind	67.25	13.96	76.31
Enhanced Unified Space	InternVL$_{IB}$	66.05	11.82	<u>81.54</u>
	InternVL$_{IB}^{\dagger}$	66.75	11.82	**81.54**
	ImageBind++(*AT E.*)	92.80	25.35	74.83
	InternVL$_{IB}$++(*AT E.*)	**93.60**	<u>25.35</u>	73.33
	InternVL$_{IB}^{\dagger}$++(*AT E.*)	<u>93.00</u>	**26.45**	72.07
	ImageBind++(*Ver.*)	88.55	19.69	76.32
	InternVL$_{IB}$++(*Ver.*)	88.30	18.93	81.49
	InternVL$_{IB}^{\dagger}$++(*Ver.*)	87.70	19.23	81.42

The (*Ver.*) variants achieve much better audio classification and audio-text retrieval performance than their corresponding source unified space while maintaining the image-text capabilities of source space. More importantly, although the audio representation in CLAP is learned by aligning with text, absorbing it can even improve audio-image alignment in the unified space. This discovery highlights cross-modal knowledge transfer capabilities, further broadening the potential of knowledge fusion in multimodal representations. In summary, combining expert spaces with appropriate factor can significantly enhance corresponding aspects without incurring extra costs, akin to a free lunch.

Additionally, increasing CLAP's combining weights yields an audio-text expertise unified space, denoted as (*AT E.*). This variant achieves even better audio-text retrieval and audio classification accuracy than CLAPs, while maintaining competitive performance for other multimodal alignments.

Complex Sequential & Parallel Bonds As a result of our complex sequential & parallel bonds, InternVL$_{IB}^{\dagger}$++ exhibits a significant advantage in image-text fields compared to ImageBind++, while achieving similar state-of-the-art performance in audio-related tasks. Besides, the overall advantage of InternVL$_{IB}^{\dagger}$++ over InternVL$_{IB}$++ demonstrates that simply tuning the small audio encoder with limited resources can effectively repair the lost knowledge.

Notably, considering the pivotal role of image-text representation in unified space, tuning the image or text encoder not only demands massive computing resources but also potentially compromises the foundation of the unified space. Therefore, repairing lost knowledge through fine-tuning is only suitable for modalities other than image or text, which further emphasizes the essential of preserving the advanced image-text expert knowledge.

Table 4. Results of zero-shot cross-modal retrievals. The best result is **bolded**, and the second best result is <u>underlined</u>.

| | Models | Audio-Text | | | | Audio-Image | | | | Image-Text | | | |
| | | AudioCaps | | Clotho | | VGG-SS | | FlickrNet | | COCO | | Flickr30K | |
		R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5	R@1	R@5
Pre-trained Expert Space	InternVL	-	-	-	-	-	-	-	-	61.05	82.18	89.29	98.19
	CLAP$_g$	40.25	76.21	18.46	41.70	-	-	-	-	-	-	-	-
	CLAP$_m$	39.00	74.71	17.90	40.57	-	-	-	-	-	-	-	-
Pre-trained Unified Space	WAV2CLIP	0.88	4.22	0.78	2.60	2.51	9.28	0.82	3.41	40.24	64.78	71.89	90.55
	AudioCLIP	3.53	11.30	3.20	9.97	1.25	3.91	1.37	4.91	17.51	37.50	38.89	64.92
	C-MCR	15.76	41.37	8.37	24.86	1.94	7.69	1.39	5.97	16.67	37.04	34.16	63.64
	Ex-MCR	19.07	47.05	7.01	22.04	2.13	8.13	1.57	5.94	40.24	64.78	71.89	90.55
	ImageBind	9.24	27.47	6.64	17.28	14.82	35.67	7.68	20.79	57.28	79.54	86.04	96.97
Enhanced Unified Space	InternVL$_{IB}$	11.17	32.03	6.60	18.27	13.77	34.56	7.37	20.80	<u>61.49</u>	<u>82.17</u>	<u>89.61</u>	<u>98.14</u>
	InternVL$_{IB}^{\dagger}$	11.20	31.84	6.42	17.45	14.47	36.44	<u>8.02</u>	20.90	**61.49**	**82.17**	**89.61**	**98.14**
	ImageBind++(*AT E.*)	<u>42.50</u>	77.31	<u>19.49</u>	43.57	12.38	33.47	5.63	17.35	47.61	71.82	78.52	94.37
	InternVL$_{IB}$++(*AT E.*)	41.41	<u>77.68</u>	**20.19**	**46.03**	10.99	30.62	5.24	17.65	54.87	78.00	85.45	96.81
	InternVL$_{IB}^{\dagger}$++(*AT E.*)	**43.10**	**78.31**	19.38	<u>45.17</u>	12.23	32.68	5.71	18.23	54.75	77.87	85.08	96.99
	ImageBind++(*Ver.*)	29.16	62.98	13.67	33.19	<u>15.48</u>	**39.26**	8.01	<u>21.87</u>	57.01	79.23	85.91	97.03
	InternVL$_{IB}$++(*Ver.*)	29.11	62.30	12.66	32.75	14.40	36.78	7.74	21.85	61.07	82.00	89.30	98.09
	InternVL$_{IB}^{\dagger}$++(*Ver.*)	29.42	62.10	13.48	33.69	**15.48**	39.14	**8.32**	**22.40**	61.13	82.05	89.33	98.11

Table 5. **Study of Basic Displacement and Combination Bond.** The reported retrieval metric is R@1. "ACaps" stands for Audio-Caps. The combining factors are set to (0.9, 0.9).

| Setting | Audio-Text | | Audio-Image | | Image-Text | |
	ACaps	Clotho	VGGSS	FlickrNet	COCO	Flickr30K
ImageBind	9.24	6.64	14.82	7.68	57.28	86.04
+Cr(InternVL)	9.25	**6.64**	**16.14**	**8.08**	58.90	87.93
+Dr(InternVL)	**11.17**	6.60	13.77	7.37	**61.49**	**89.61**

Table 6. **Study of Different Complex Bonds.** "Only seq" means sequentially integrating InternVL, CLAP$_m$, and CLAP$_g$ via displacement. "Only Para." means aligning in parallel with combination. "Seq.&Para." refers to our complex sequential & parallel bonds. The combining factors follow (*Ver.*) variant.

| Route | Audio-Text | | Audio-Image | | Image-Text | |
	ACaps	Clotho	VGGSS	FlickrNet	COCO	Flickr30K
Only Seq.	**36.03**	**16.62**	9.24	4.89	33.77	66.55
Only Para.[2]	29.14	13.48	**16.06**	8.20	58.83	87.35
Seq.&Para.	29.42	13.48	15.48	**8.32**	**61.13**	**89.33**

4.3. Discussion

Displacement or Combination We integrate InternVL and ImageBind through two basic bonds to further reveal their properties. As shown in Table 5, the displacement bond inherits InternVL's advanced image-text capabilities. Despite audio embeddings are re-projected, the resulting space still achieves comparable performance in audio-text and audio-image retrieval. Meanwhile, the combination bond yield slight but consistent improvements over Image-Bind. These observations reinforce our analysis of the basic bonds: the displacement bond is a radical knowledge fusion

[2]Given InternVL are remapped, its combining factor is set to (0.1, 0.1), which corresponds to (0.9, 0.9) in InternVL$_{IB}$.

Table 7. **Study of Corase-grain Module Selection.** The combining factors follow the (*Ver.*) variant.

| Setting | Classification | | Retrieval | |
	ESC	AudioSet	ACaps	Clotho
ImageBind	67.25	13.96	9.24	6.64
+Cr(CLAP$_g$)	86.65	19.09	28.63	13.24
+Cr(CLAP$_m$)	88.40	18.89	27.33	12.70
+Cr(CLAP$_g$)+Cr(CLAP$_m$)	**88.55**	**19.69**	**29.16**	**13.67**

scheme, whereas the combination bond is more moderate.

Different Complex Bonds We compare different complex bonds in Table 6. Our complex sequential & parallel bonds achieve more balanced and stable improvements than pure sequential or parallel routes that rely on only one bond. These results confirm our analysis in Section 3.3 and emphasize the importance of combining the complementary basic bonds when designing complex bonds.

Corase-grain Combined Module Selection Table 7 report the results of employing different aligned audio-text experts to enhance ImageBind. The results reveal that combining different modules exhibits varying abilities. Integrating CLAP$_m$ yields more gains on ESC, while CLAP$_g$ improves performance more on other general audio datasets. Employing both together brings better overall performance.

Fine-grained Combining Factors Adjustment To explore and provide insights about combining factors adjustment, we comprehensively display the effect of (σ_a, σ_t) on InternVL$_{IB}^{\dagger}$ in Figure 3. There are three main observations: 1) All (σ_a, σ_t) can significantly enhance audio-text alignment, and when they are set larger than 0.5, the enhanced
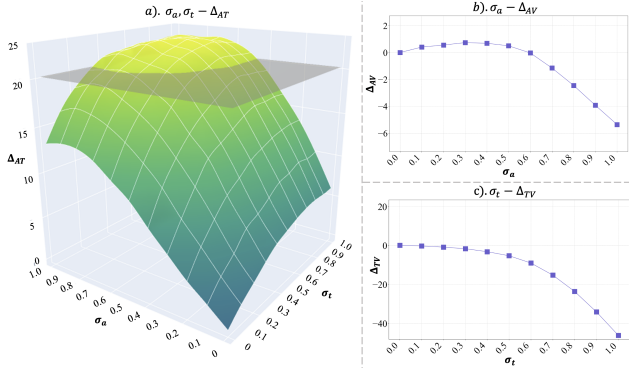
Figure 3. **Analysis of CLAPs' combining factors (σ_a, σ_t) on InternVL$_{IB}^{\dagger}$++.** $\Delta_{AT}, \Delta_{AV}, \Delta_{TV}$ represents the average R@1 variance between InternVL$_{IB}^{\dagger}$++ and InternVL$_{IB}^{\dagger}$ on audio-text, audio-image and image-text retrieval tasks, respectively. Positive Δ_* signifies improvements in the corresponding task, while negative values indicate reductions. The gray plane in the 3D figure $a)$ denotes the audio-text performance of CLAP$_g$.

unified space even outperforms CLAP$_g$ in audio-text field. 2) When σ_a takes a moderate value (around 0.5), the audio-image performance can be improved. 3) Since CLAP's text representation is aligned to audio, large σ_t may hurt the image-text alignment, but when it is set to small value (around 0.2), the negative effect is negligible.

Generally speaking, the combining factors adjustment are logical and insensitive. Most settings either bring an overall stronger unified space or provide superior expertise in a certain aspect. We conduct more analyses and visualizations on all the resulting space in the Appendix C, which further prove the regularity and insensitivity of combining factors.

Mixture-of-Projectors Results in Table 8 illustrate that combining all projectors yields substantial performance benefits, which prove that our mixture-of-projectors strategy enhances alignment and fosters more discriminative representations. Noteworthy, each projector typically consists of about 2M parameters, therefore multiple projectors will only incur minimal extra inference costs.

Projector Design We investigate various projector structures and learning objective designs, and the results are reported in Table 9. Compared with the projector learning pipeline proposed in the previous advanced space integration methods C-MCR and Ex-MCR, our simpler pipeline achieves better overall results in both basic bonds. The multiple shared modalities between unified space and expert spaces can sufficiently align the spaces. In this scenario, complex learning pipelines and intra-space loss may hinder alignment generalization. Our straightforward design is better suited for unified space scenarios.

Table 8. **Study of Mixture-of-Projector.** The results are obtained on bond: ImageBind+Dr(InternVL). ψ_* represents the projector trained on subset D_*. Ψ indicates the whole projectors group.

Used Projector	Audio-Text		Audio-Image		Image-Text	
	ACaps	Clotho	VGGSS	FlickrNet	COCO	Flickr30K
ψ_A	9.59	5.79	12.61	6.39	61.44	89.10
ψ_V	9.79	5.53	10.83	5.97	61.30	89.32
ψ_T	10.07	5.57	10.08	6.17	61.47	89.24
ψ_{AT}	9.96	5.86	12.60	6.68	61.44	89.26
ψ_{VA}	9.77	5.91	12.93	6.74	61.44	89.33
ψ_{VT}	9.91	5.46	11.12	6.43	61.49	89.24
ψ_{VAT}	10.23	5.98	12.70	6.78	61.42	89.30
Ψ	**11.17**	**6.60**	**13.77**	**7.37**	**61.49**	**89.61**

Table 9. **Study of Projector Design.** The projector structure and learning objective of C-MCR, Ex-MCR and Ours are used for two basic bonds, respectively.

Training Pipeline	Audio-Text		Audio-Image		Image-Text	
	ACaps	Clotho	VGGSS	FlickrNet	COCO	Flickr30K
Combination Bond — IB + Cr(InternVL)						
C-MCR	**9.26**	6.64	15.99	**8.12**	58.62	87.77
Ex-MCR	9.25	6.64	15.60	7.65	58.47	87.11
Ours	9.25	**6.64**	**16.14**	8.08	**58.90**	**87.93**
Displacement Bond — IB + Dr(InternVL)						
C-MCR	10.43	5.55	11.60	6.32	61.28	89.13
Ex-MCR	8.97	5.27	8.78	5.09	61.26	89.48
Ours	**11.17**	**6.60**	**13.77**	**7.37**	**61.49**	**89.61**

Computing Resource Collecting a group of pseudo datasets takes about 10 hours on a single 4090, while using 12GB GPU memory. The training times for single displacement and combination bond are approximately 6 hours and 1.5 hours, respectively, on a single 4090, and it only requires 3GB of GPU memory. Tuning the displacement product consumes 15 hours on single 4090.

5. Conclusion

This paper proposes **FreeBind** to enhance pre-trained unified multimodal representations by binding the knowledge of extra expert spaces. Based on the concept of viewing multimodal spaces as basic unit, we design two basic "space bonds": displacement and combination bond. With these foundations, we introduce complex sequential & parallel bonds to effectively combine multiple spaces simultaneously. After training, a coarse-to-fine customized inference strategy is employed to flexibly enhance unified space for different applications. Experimentally, we integrate ImageBind's audio-image-text space with multiple advanced spaces. The resulting space: ImageBind++, InternVL$_{IB}$ and InternVL$_{IB}$++ comprehensively surpass ImageBind. Moreover, via customized inference, it even outperforms state-of-the-art image-text and audio-text expert models in their respective domains.

Impact Statement

FreeBind enables flexible augment pre-trained unified space with very limited computing resources. Under appropriate usage, this technique can help quickly develop stronger unified multimodal representation with little training costs, and provide a powerful and accessible foundation for different customized multimodal application scenarios. However, low-cost unified representation learning methods could be misused to support unethical multi-modal applications. To prevent this, we plan to add unethical data detection to the pseudo dataset collection stage, thereby preventing representations from acquiring capabilities related to unethical applications.

Acknowledgments

References

Changpinyo, S., Sharma, P., Ding, N., and Soricut, R. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3558–3568, 2021.

Chen, H., Xie, W., Afouras, T., Nagrani, A., Vedaldi, A., and Zisserman, A. Localizing visual sounds the hard way. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16867–16876, 2021.

Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Muyan, Z., Zhang, Q., Zhu, X., Lu, L., et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Drossos, K., Lipping, S., and Virtanen, T. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Elizalde, B., Deshmukh, S., Al Ismail, M., and Wang, H. 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, 2023a.

Elizalde, B., Deshmukh, S., and Wang, H. Natural language supervision for general-purpose audio representations, 2023b. URL https://arxiv.org/abs/2309.05767.

Gemmeke, J. F., Ellis, D. P., Freedman, D., Jansen, A., Lawrence, W., Moore, R. C., Plakal, M., and Ritter, M. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 776–780. IEEE, 2017.

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., and Misra, I. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15180–15190, 2023.

Guzhov, A., Raue, F., Hees, J., and Dengel, A. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 976–980. IEEE, 2022.

Han, J., Zhang, R., Shao, W., Gao, P., Xu, P., Xiao, H., Zhang, K., Liu, C., Wen, S., Guo, Z., et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv preprint arXiv:2309.03905*, 2023.

Huang, H., Wang, Z., Huang, R., Liu, L., Cheng, X., Zhao, Y., Jin, T., and Zhao, Z. Chat-3d v2: Bridging 3d scene and large language models with object identifiers. *arXiv preprint arXiv:2312.08168*, 2023a.

Huang, R., Huang, J., Yang, D., Ren, Y., Liu, L., Li, M., Ye, Z., Liu, J., Yin, X., and Zhao, Z. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. *arXiv preprint arXiv:2301.12661*, 2023b.

Jia, C., Yang, Y., Xia, Y., Chen, Y.-T., Parekh, Z., Pham, H., Le, Q., Sung, Y.-H., Li, Z., and Duerig, T. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pp. 4904–4916. PMLR, 2021.

Kim, C. D., Kim, B., Lee, H., and Kim, G. 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)*, pp. 119–132, 2019.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023a.

Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., and Su, H. Openshape: Scaling up 3d shape representation towards open-world understanding. *arXiv preprint arXiv:2305.10764*, 2023b.

Piczak, K. J. Esc: Dataset for environmental sound classification. In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 1015–1018, 2015.

Podell, D., English, Z., Lacey, K., Blattmann, A., Dockhorn, T., Müller, J., Penna, J., and Rombach, R. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*, 2023.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., and Ommer, B. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Senocak, A., Oh, T.-H., Kim, J., Yang, M.-H., and Kweon, I. S. Learning to localize sound source in visual scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4358–4366, 2018.

Sharma, P., Ding, N., Goodman, S., and Soricut, R. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2556–2565, 2018.

Soldan, M., Pardo, A., Alcázar, J. L., Caba, F., Zhao, C., Giancola, S., and Ghanem, B. Mad: A scalable dataset for language grounding in videos from movie audio descriptions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5026–5035, 2022.

Tang, Z., Yang, Z., Zhu, C., Zeng, M., and Bansal, M. Any-to-any generation via composable diffusion. *arXiv preprint arXiv:2305.11846*, 2023.

Wang, P., Wang, S., Lin, J., Bai, S., Zhou, X., Zhou, J., Wang, X., and Zhou, C. One-peace: Exploring one general representation model toward unlimited modalities. *arXiv preprint arXiv:2305.11172*, 2023a.

Wang, Z., Huang, H., Zhao, Y., Zhang, Z., and Zhao, Z. Chat-3d: Data-efficiently tuning large language model for universal dialogue of 3d scenes. *arXiv preprint arXiv:2308.08769*, 2023b.

Wang, Z., Zhang, Z., Liu, L., Zhao, Y., Huang, H., Jin, T., and Zhao, Z. Extending multi-modal contrastive representations. *arXiv preprint arXiv:2310.08884*, 2023c.

Wang, Z., Zhao, Y., Cheng, X., Huang, H., Liu, J., Tang, L., Li, L., Wang, Y., Yin, A., Zhang, Z., et al. Connecting multi-modal contrastive representations. *arXiv preprint arXiv:2305.14381*, 2023d.

Wu, H.-H., Seetharaman, P., Kumar, K., and Bello, J. P. Wav2clip: Learning robust audio representations from clip. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 4563–4567. IEEE, 2022.

Wu, S., Fei, H., Qu, L., Ji, W., and Chua, T.-S. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*, 2023a.

Wu, Y., Chen, K., Zhang, T., Hui, Y., Berg-Kirkpatrick, T., and Dubnov, S. 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)*, pp. 1–5. IEEE, 2023b.

Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., and Feichtenhofer, C. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021.

Xu, J., Mei, T., Yao, T., and Rui, Y. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 5288–5296, 2016.

Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J. C., and Savarese, S. 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*, pp. 1179–1189, 2023a.

Xue, L., Yu, N., Zhang, S., Li, J., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J. C., and Savarese, S. Ulip-2: Towards scalable multimodal pre-training for 3d understanding. *arXiv preprint arXiv:2305.08275*, 2023b.

Young, P., Lai, A., Hodosh, M., and Hockenmaier, J. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78, 2014.

Zhang, Y., Jiang, H., Miura, Y., Manning, C. D., and Langlotz, C. P. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pp. 2–25. PMLR, 2022.

Zhou, J., Wang, J., Ma, B., Liu, Y.-S., Huang, T., and Wang, X. Uni3d: Exploring unified 3d representation at scale. *arXiv preprint arXiv:2310.06773*, 2023.

Zhu, B., Lin, B., Ning, M., Yan, Y., Cui, J., Wang, H., Pang, Y., Jiang, W., Zhang, J., Li, Z., et al. Language-bind: Extending video-language pretraining to n-modality by language-based semantic alignment. *arXiv preprint arXiv:2310.01852*, 2023a.

Zhu, D., Chen, J., Shen, X., Li, X., and Elhoseiny, M. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023b.

A. Pseudo Dataset between Unified and Audio-Text Expert Spaces

Considering the source embeddings: $\mathbf{T}^{at}$, $\mathbf{A}^{at}$, $\mathbf{T}^u$, $\mathbf{V}^u$ and $\mathbf{A}^u$, the pseudo dataset starting from texts (i.e., D_T) can be expressed as:

$$\tilde{\mathbf{T}}^{at} = \mathbf{T}^{at}; \quad \tilde{\mathbf{A}}^{at} = \mathrm{softmax}(\tilde{\mathbf{T}}^{at}\mathbf{A}^{at\top})\mathbf{A}^{at};$$

$$\tilde{\mathbf{T}}^u = \mathbf{T}^u; \quad \tilde{\mathbf{A}}^u = \mathrm{softmax}(\tilde{\mathbf{T}}^{at}\mathbf{A}^{at\top})\mathbf{A}^u; \quad \tilde{\mathbf{V}}^u = \mathrm{softmax}(\tilde{\mathbf{T}}^u\mathbf{V}^{u\top})\mathbf{V}^u \quad (11)$$

The pseudo dataset from audios (i.e., D_A) can be expressed as:

$$\tilde{\mathbf{A}}^{at} = \mathbf{A}^{at}; \quad \tilde{\mathbf{T}}^{at} = \mathrm{softmax}(\tilde{\mathbf{A}}^{at}\mathbf{T}^{at\top})\mathbf{T}^{at};$$

$$\tilde{\mathbf{A}}^u = \mathbf{A}^u; \quad \tilde{\mathbf{T}}^u = \mathrm{softmax}(\tilde{\mathbf{A}}^{at}\mathbf{T}^{at\top})\mathbf{T}^u; \quad \tilde{\mathbf{V}}^u = \mathrm{softmax}(\tilde{\mathbf{A}}^u\mathbf{V}^{u\top})\mathbf{V}^u \quad (12)$$

The pseudo dataset from non-shared image modality (i.e., D_V) can be expressed as:

$$\tilde{\mathbf{V}}^u = \mathbf{V}^u; \quad \tilde{\mathbf{T}}^u = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{T}^{u\top})\mathbf{T}^u; \quad \tilde{\mathbf{A}}^u = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{A}^{u\top})\mathbf{A}^u;$$

$$\tilde{\mathbf{T}}^{at} = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{T}^{u\top})\mathbf{T}^{at}; \quad \tilde{\mathbf{A}}^{at} = \mathrm{softmax}(\tilde{\mathbf{V}}^u\mathbf{A}^{u\top})\mathbf{A}^{at} \quad (13)$$

B. Training Datasets

Unimodal data   Following (Wang et al., 2023d), we employ the texts of COCO (Lin et al., 2014), CC3M (Changpinyo et al., 2021; Sharma et al., 2018), MSRVTT (Xu et al., 2016), MAD (Soldan et al., 2022), AudioCaps (Kim et al., 2019) and Clotho (Drossos et al., 2020) as the unimodal source text. There are 2.33M text samples in total (only 1M texts are selected from CC3M). All the unpaired image data are from ImageNet (Deng et al., 2009) training set, which consists of 1.3M images without any annotations. The audios are sourced from AudioSet (Gemmeke et al., 2017) training set, total in 2M audio clips.

Paired data   Optionally, we utilize the 2 million audio-image pairs from the unbalanced training set of AudioSet to tune the audio encoder for the displacement bond product. Notably, AudioSet is also the training set of ImageBind. Therefore, utilizing AudioSet for tuning does not introduce any new knowledge. The purpose of further tuning is to repair the representation damage caused by the displacement bond process.

C. Further Analysis of Combining Factors

To more comprehensively demonstrate the impact of the CLAP's combining factors on the product, we also analyzed CLAPs' combining factors (σ_a, σ_t) on InternVL$_{IB}$++ and ImageBind++, which are presented in Figure 4 and 5. The curves and surfaces in these figures are similar to that of Figure 3. This observation further demonstrates the regularity and insensitivity of combining factors, as discussed in Section 4.3.

Moreover, we further display the impact of the InternVL's combining factor (λ_t, λ_v) on the performance of InternVL$_{IB}$ in Figure 6. Generally speaking, since ImageBind's representations are remapped, the greater (λ_t, λ_v), the higher the overall performance, which is also consistent with the definition of displacement.

D. Limitations and Future Work

This paper introduces FreeBind, a promising and cost-effective unified space augmentation and knowledge fusion solution, and provides an in-depth and comprehensive analysis and discussion of the key design. However, the current FreeBind is only utilized to enhance the most basic unified audio-image-text space, whereas the most advanced unified space methods, such as ImageBind and LanguageBind, have achieved unified representations of six or seven modalities. Further research to incorporate FreeBind for more modalities would be an interesting direction.

In light of our experiments on displacement bond, which have demonstrated its capability to substitute a stronger image-text space for the unified space and effectively repair the lost knowledge through tuning, and combination bonds with small combining factors can yield an enhanced unified space with stable gains and no negative consequences. FreeBind shows promise for broader applications.
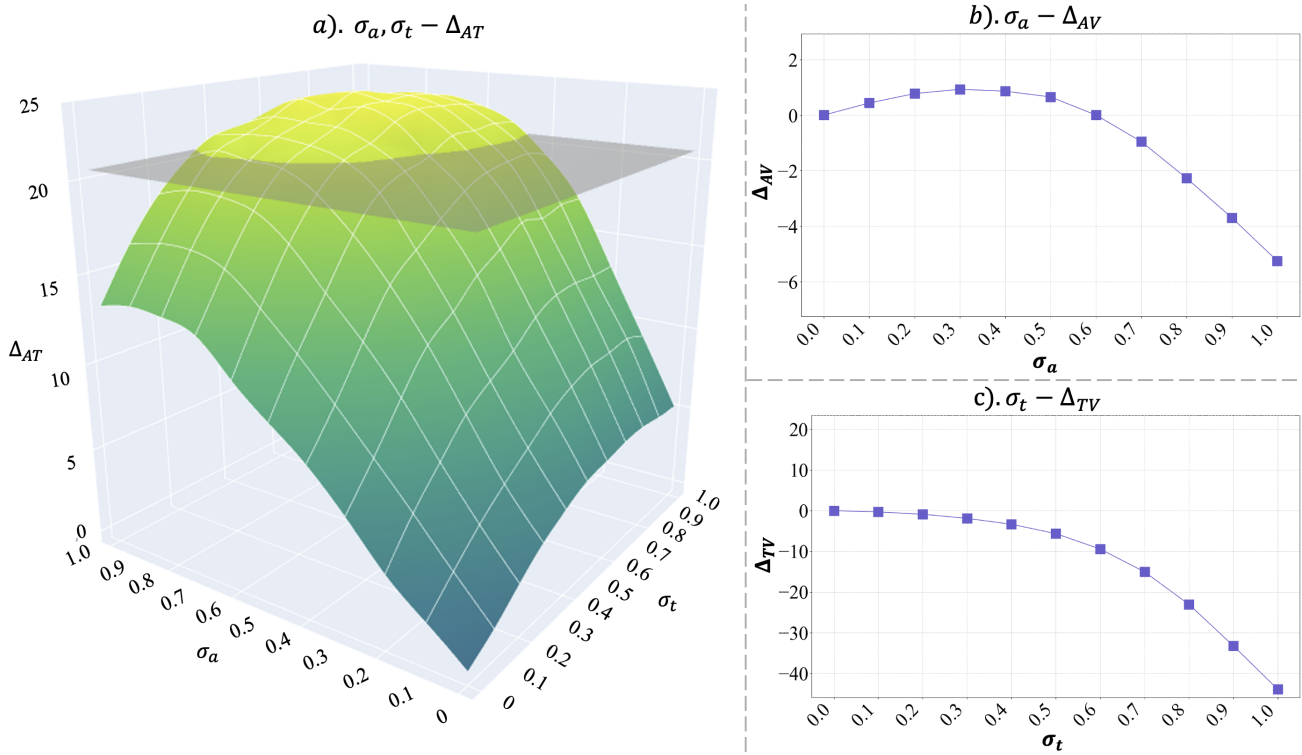
Figure 4. Analysis of CLAPs' combining factors (σ_a, σ_t) on InternVL$_{IB}$++. $\Delta_{AT}, \Delta_{AV}, \Delta_{TV}$ represents the average R@1 variance between InternVL$_{IB}$++ and InternVL$_{IB}$ on audio-text, audio-image and image-text retrieval tasks, respectively. The gray plane in the 3D figure $a)$ denotes the audio-text performance of CLAP$_g$.
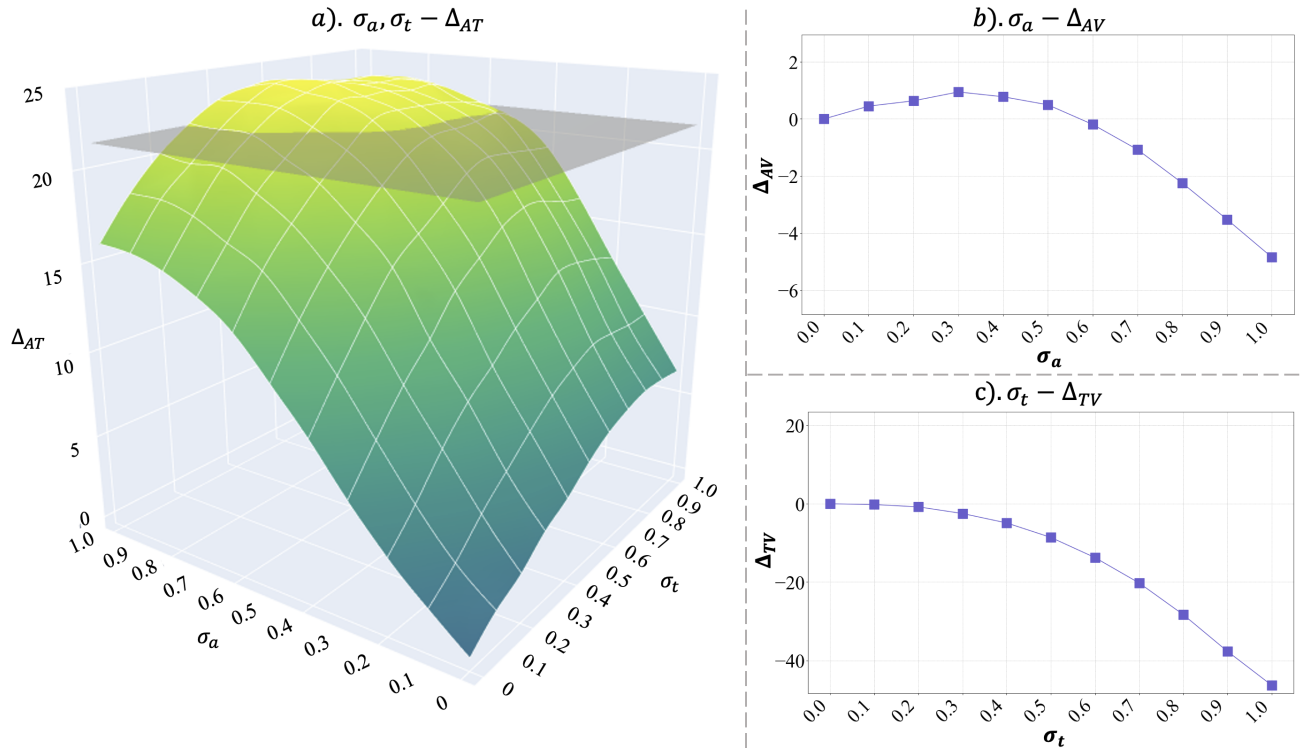

Figure 5. Analysis of CLAPs' combining factors (σ_a, σ_t) on ImageBind++. $\Delta_{AT}, \Delta_{AV}, \Delta_{TV}$ represents the average R@1 variance between ImageBind++ and ImageBind on audio-text, audio-image and image-text retrieval tasks, respectively. The gray plane in the 3D figure $a)$ denotes the audio-text performance of CLAP$_g$.
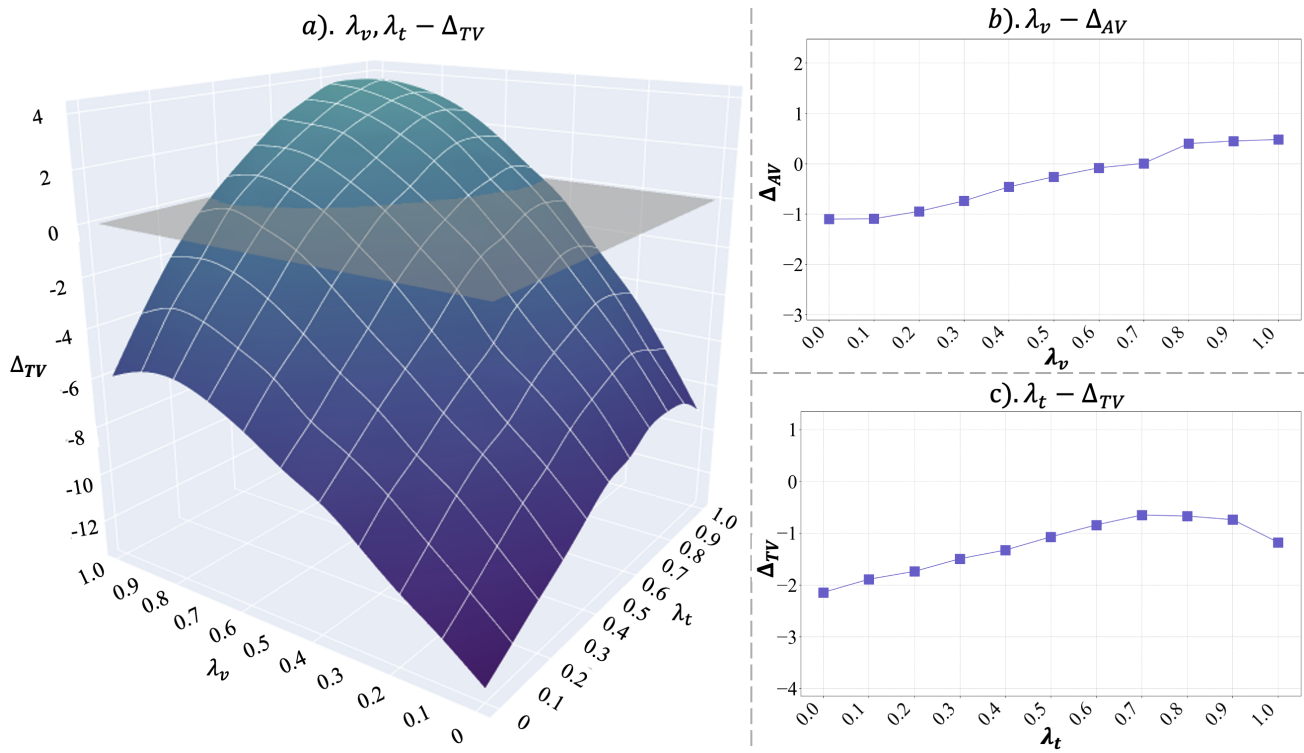
Figure 6. Analysis of InternVL's combining factors (λ_v, λ_t) on InternVL$_{IB}$. $\Delta_{AT}, \Delta_{AV}, \Delta_{TV}$ represents the average R@1 variance between InternVL$_{IB}$ and ImageBind on audio-text, audio-image and image-text retrieval tasks, respectively. Positive Δ_* signifies improvements in the corresponding task, while negative values indicate reductions. The gray plane in the 3D figure $a)$ denotes the image-text performance of ImageBind.