

# VIT-LENS: Towards Omni-modal Representations

Weixian Lei<sup>1,2</sup> Yixiao Ge<sup>2,3†</sup> Kun Yi<sup>2</sup> Jianfeng Zhang<sup>1</sup> Difei Gao<sup>1</sup>  
Dylan Sun<sup>2</sup> Yuying Ge<sup>3</sup> Ying Shan<sup>2,3</sup> Mike Zheng Shou<sup>1†</sup>

<sup>†</sup>Corresponding authors

<sup>1</sup>Show Lab, National University of Singapore <sup>2</sup>ARC Lab, Tencent PCG <sup>3</sup>Tencent AI Lab

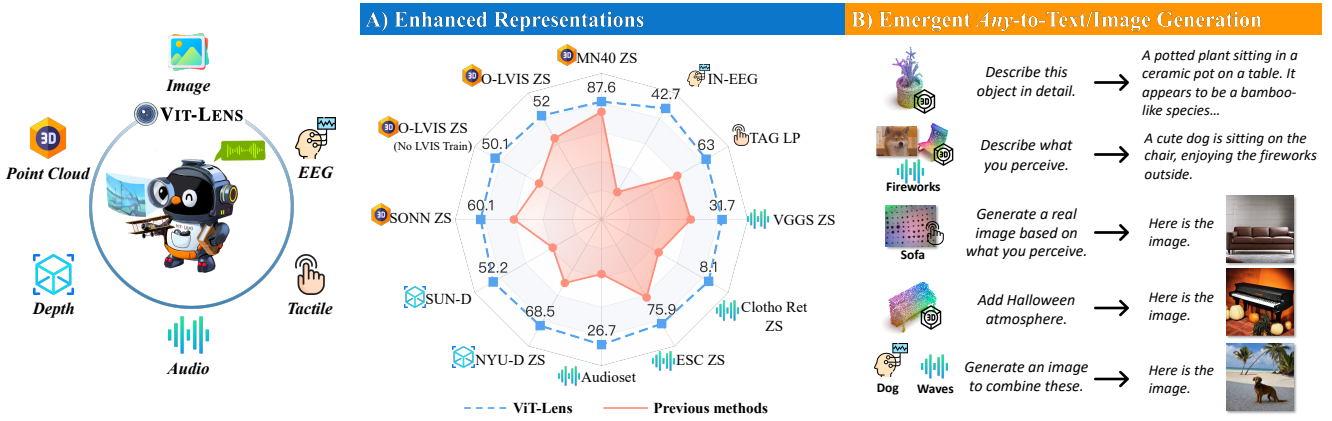


Figure 1. **VIT-LENS for omni-modal representation learning.** **A)** VIT-LENS consistently enhances the performance of understanding tasks, such as classification, zero-shot classification (ZS) and linear probing (LP), across 3D point cloud([48]), depth([28]), audio([28]), tactile([82]), and EEG([4]) modalities. The citations represent the compared previous methods. Further details in Sec. 4. **B)** By plugging VIT-LENS into multimodal foundation models, it enables emergent applications “out-of-the-box”, including Any-modality Captioning/QA, Any-modality-to-Image Generation and text-guided Any-modality-to-Image editing, to name a few.

## Abstract

Aiming to advance AI agents, large foundation models significantly improve reasoning and instruction execution, yet the current focus on vision and language neglects the potential of perceiving diverse modalities in open-world environments. However, the success of data-driven vision and language models is costly or even infeasible to be reproduced for rare modalities. In this paper, we present **VIT-LENS** that facilitates efficient omni-modal representation learning by perceiving novel modalities with a pretrained-ViT and aligning them to a pre-defined space. Specifically, the modality-specific lens is tuned to project any-modal signals to an intermediate embedding space, which are then processed by a strong ViT with pre-trained visual knowledge. The encoded representations are optimized toward aligning with the modal-independent space, pre-defined by off-the-shelf foundation models. VIT-LENS provides a unified solution for representation learning of increasing modalities with two appealing advantages: (i) Unlocking the great potential of pretrained-ViTs to novel modalities effectively with efficient

parameters and data regime; (ii) Enabling emergent downstream capabilities through modality alignment and shared ViT parameters. We tailor VIT-LENS to learn representations for 3D point cloud, depth, audio, tactile and EEG, and set new state-of-the-art results across various understanding tasks, such as zero-shot classification. By seamlessly integrating VIT-LENS into Multimodal Foundation Models, we enable Any-modality to Text and Image Generation in a zero-shot manner. Code and models are available at <https://github.com/TencentARC/ViT-Lens>.

## 1. Introduction

Humans interact with the world through various sensory systems like vision, audition, touch, smell, and taste. To advance versatile AI agents, deep learning models need to replicate these human-like multi-sensory abilities and tackle varied user-specified tasks. For instance, visually interpreting road signs to ensure our safe driving, listening to sirens to respond to emergency vehicles, and tactually assessing clothing fabric quality to offer shopping guidance. Among

these applications, omni-modal representation learning has become a focal point, which enables comprehensive perception in open-world environments.

On the way to pursuing omni-modal AI agents, the research community has utilized large-scale web data to make substantial strides in language [6, 16, 49, 55, 63, 64, 72] and vision [5, 12, 17, 20, 21, 35, 37, 59, 65, 84]. Consequently, Multimodal Foundation Models (MFMs) [1, 11, 14, 18, 24, 25, 47, 90] that integrate vision representations with Large Language Models (LLMs) have made great progress in both vision-language comprehension and generation.

However, extending the success of unleashing LLMs to comprehend and interact with a broader array of modalities remains challenging. Despite recent initiatives [28, 75] in pursuing omni-modal intelligence, their capabilities on certain modalities are often constrained by limited data used in the training phase. In contrast to image, video, and text data, which are abundant on the internet, acquiring large-scale datasets for less common modalities can be non-trivial. This scarcity of data leads to sub-optimal models with poor generalization, particularly when encountering novel categories, thereby limiting their broader real-world applications.

In this work, we present a novel perspective. Given the exceptional generalization and transfer learning capabilities of the pretrained-ViTs [7, 20, 21, 56, 65], there is promise in adapting their inherent knowledge to comprehend novel modalities. This eliminates the necessity of collecting large-scale data to train models from scratch for each modality, which demands substantial time and resources. Recognizing the rich knowledge encoded in a pretrained-ViT, we conjecture that a pretrained-ViT is able to function as a multi-modal processor – it possesses the capacity to sense and comprehend a spectrum of modalities as it interprets images.

From this standpoint, we introduce **VIT-LENS**, which encodes the out-of-image modalities through a set of pretrained-ViT parameters, with the goal of maximizing the utilization of pretrained model weights and the knowledge they encapsulate. Specifically, VIT-LENS employs a modality-specific Lens along with a lightweight modality embedding module to transform input data into an intermediate space. Subsequently, a frozen pretrained-ViT is applied for further encoding. This approach enables the encoding of diverse modalities, aligning their features with the established features of anchor data, which can range from images, text to others, from off-the-shelf foundation models.

Our proposed method offers several advantages in advancing omni-modal representation learning: **(1) Parameters and data efficient approach.** Our method adopts a shared set of pretrained-ViT parameters across various modalities, enabling an efficient utilization of model parameters. Moreover, it efficiently enhances representations for less common modalities by leveraging the advanced ViT model, reducing the demand for extensive data collection. **(2) Emergent**

**capabilities.** By training VIT-LENS with the ViT used in an off-the-shelf MFM, we can seamlessly obtain an Any-Modality MFM via VIT-LENS integration. The integrated model extends the original MFM’s capabilities to various modalities, without any specific instruction tuning. For instance, without direct training for tactile data, the model broadens its image generation capability to include tactile-to-image generation. As a result, it can generate an image of a sofa upon receiving tactile signals indicating “leather”.

We conducted comprehensive experiments across multiple modalities, extending beyond images and videos to encompass 3D point cloud, depth, audio, tactile, and EEG. This experiments were evaluated across 11 benchmarks. As is shown in Fig. 1A, VIT-LENS demonstrates state-of-the-art performance in 3D zero-shot classification. Particularly, when LVIS classes are excluded during training, VIT-LENS achieves an impressive zero-shot classification accuracy of 50.1% on Objaverse-LVIS [15], surpassing the prior SOTA by 11.0%. It consistently outperforms ImageBind [28] on depth and audio benchmarks, and surpasses previous works on tactile [82] and EEG [4] related tasks.

Beyond understanding tasks, we plug VIT-LENS into two recent MFMs, InstructBLIP [14] and SEED [24, 25]. As illustrated in Fig. 1B, this empowers the MFMs to comprehend any modality in a zero-shot manner, making Any-Captioning, Any-QA, Any-to-Image Generation and text guided Any-to-Image editing right out of the box, all without the need for specific instruction tuning.

## 2. Related Work

**Vision Language Pretraining: Advancements and Impacts.** Recent advancements in vision-language pretraining, including models such as CLIP [65], ALIGN [37], CoCa [84], Flamingo [1], and LiT [86], have leveraged image-text pairs to achieve remarkable zero-shot performance on a wide range of vision and language tasks. Meanwhile, pretrained CLIP models have served as influential teachers and their joint embedding space has demonstrated efficacy in diverse zero-shot tasks such as segmentation [40], detection [31, 89], 3D shape understanding [48, 80, 81, 87, 91], 3D open-vocabulary segmentation [58], mesh animation [83], audio understanding [33] and more. VIT-LENS extends these models’ capacities to diverse modalities by integrating pretrained-ViT, enhancing its omni-modal understanding ability and enabling superior performance across various tasks and modalities.

**Multimodal Learning.** Previous studies explored joint training across multiple modalities in both supervised [23, 27, 44] and self-supervised settings [2, 29, 45, 52, 71]. Several approaches aim at aligning various modalities to CLIP for multimodal zero-shot learning. AudioCLIP [33] adds audio to CLIP for zero-shot audio classification, while ImageBind [28] aligns six modalities to CLIP using paired image

data. Besides, ONE-PEACE [75] introduces a unified encoder that is pretrained from scratch to align vision, language, and audio. Zhang *et al.* [88] pretrain a transformer with LAION-2B, following CLIP’s methodology, for downstream supervised tasks across modalities. In contrast, VIT-LENS stands out by leveraging a pretrained-ViT to understand and unite diverse modalities without manual annotations. Its seamlessly integration with Multimodal Foundation Models (MFM) allows easy plug-and-play in emergent applications. **Multimodal Foundation Models.** Recent advancements in Large Language Models (LLMs) [55, 72] have demonstrated remarkable language understanding and reasoning abilities. Afterwards, substantial efforts [1, 43, 46, 47, 90] have been directed towards enabling LLMs to perceive and interact with the visual world with the help of visual representation models. Similar paradigms enable LLMs to understand more modalities by aligning the well-trained encoders of various modalities to the textual space of LLMs [32, 34, 68]. Beyond understanding tasks, recent works [24, 25, 69] empower LLMs with the ability to generate images, and NextGPT [76] extends the generative capabilities to encompass audio and video. Most of these Multimodal Foundation Models (MFMs) require specific instruction-following data within particular domains for training. In this study, we demonstrate that VIT-LENS can seamlessly integrate with an MFM without additional training, extending its capabilities to various modalities.

### 3. Method

**Overview.** VIT-LENS advances omni-modal representation learning by leveraging pretrained-ViT parameters to encode features for various modalities, utilizing the pre-existing knowledge in ViT. Specifically, a modality-specific encoder, composed of a modality embedding module, the modality-specific Lens, and the pretrained-ViT, is optimized to embed robust representations through the training objective of cross-modal alignment. For each modality, we consider its associated anchor modalities, like image and text, as reference points for learning. For alignment, we leverage predefined foundation models, such as CLIP [65], to extract features from the anchor modalities. Our approach leverages the extensive knowledge embedded in both the foundation models and pretrained-ViT, providing a strong basis for representation learning for each modality. This compensates for the shortage of large-scale training data available for certain modalities. We illustrate our approach in Fig. 2.

#### 3.1. Architecture

**Foundation Models for alignment.** In VIT-LENS, the new modalities are aligned to a unified feature space established by a robust foundation model. Various options exist for this model, ranging from language models [16, 49, 63, 64, 72], vision models [7, 21, 35, 56] to vision-language models [12,

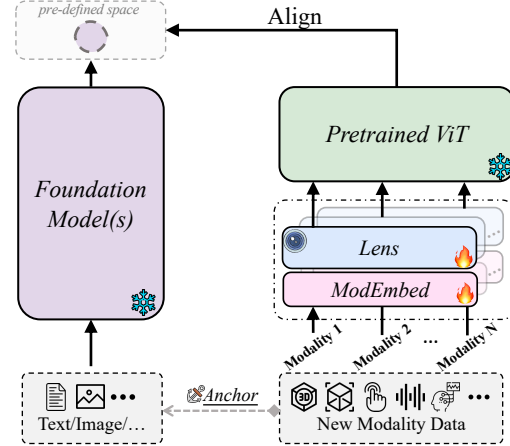


Figure 2. **Training Pipeline of VIT-LENS.** VIT-LENS extends the capabilities of a pretrained-ViT to diverse modalities. For each novel modality, it firstly employs a Modality Embedding (ModEmbed) and a Lens to learn mapping modality-specific data into an intermediate embedding space. It subsequently employs a set of pretrained-ViT layers to encode the feature. Finally, the output feature is aligned with the feature extracted from the anchor data (image, text, *etc.*) of the new modality using an off-the-shelf foundation model.

41, 42, 65]. During training, we fix the foundation model’s parameters and utilize it to encode features for the anchor data, which serves as supervision for feature alignment.

**Modality Encoder.** As is shown in Fig. 2, the modality encoder in VIT-LENS consists of a Modality Embedding Module (ModEmbed), a Lens and a set of pretrained-ViT layers. Due to the distinct characteristics of various modalities, raw signals may not match the pretrained-ViT input space. This mismatch can result in suboptimal performance, despite utilizing a powerful model. Therefore, we employ some heuristic designs: (1) *Obtain modality token embeddings*: for each modality, we adopt a specific tokenization scheme to transform raw input signals into token embeddings. (2) *Map modality token embeddings to the ViT input space*: the Lens learns to map the modality embeddings into a group of latent embeddings, thereby constructing the input for the pretrained-ViT. Subsequently, the latent embeddings are forwarded to frozen pretrained-ViT layers to obtain the final representation.

During training, the pretrained-ViT component remains frozen, and only the parameters of ModEmbed and Lens are updated. More details can be found in Supp.

**Lens: Connecting Modalities to ViT.** We introduce two variants of Lens to link modality token embeddings to ViT. We show their architectures in Fig. 3.

- **Self-attention blocks (S-Attn).** This variant involves a stack of self-attention layers [74] that transforms the input token embeddings into intermediate embeddings with equal indices. We can potentially enhance this variant’s

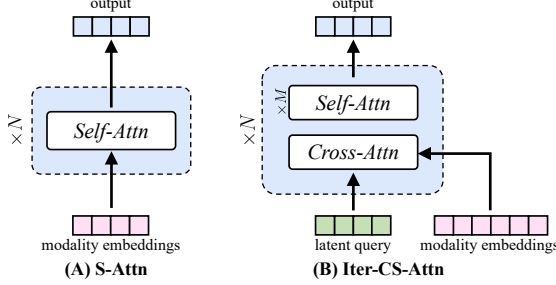


Figure 3. **Lens Architecture** used in VIT-LENS.

capability by initializing it with pretrained weights from existing ViT layers. It suits modalities structured with image-like inputs, such as depth maps.

- **Iterative cross-self-attention blocks (Iter-CS-Attn).** This variant’s basis block involves a cross-attention module coupled with a self-attention tower, inspired by [36]. It maps a latent array and input embeddings to a latent embedding of matching length within the input latent array. This manner condenses inputs of varied sized into a latent bottleneck, making it apt for lengthy input modalities like 3D point clouds. Similar architectures are employed in Vision-Language Models (VLMs) [1, 42] to extract visual information for Large Language Models (LLMs). Our innovation lies in utilizing this structure to map signals from diverse modalities into the pretrained-ViT’s input space, enabling the ViT to understand modalities beyond images.

### 3.2. Training Objective

In this work, we study modalities of 3D, depth, audio, tactile and EEG, among which all the data samples are associated with text descriptions, image appearances, or both. We use the pretrained CLIP [12, 65] as the foundation model. By default, we employ pretrained layers of ViT in the foundation CLIP as part of the modality encoder. Following the approach in previous works [33, 48, 80, 81], we adopt multi-modal contrastive learning for representation alignment.

We denote  $X = \{x_1, \dots, x_N\}$  as the collection of modality data to be learned,  $\mathcal{A} = \{A_1, \dots, A_M\}$  as the set of anchor modalities,  $a_n^m$  as the anchor data of  $x_n$  from modality  $A_m$ ,  $\mathbf{G}_A$  as the foundation model for anchor modality  $A$ , and  $\mathbf{F}$  as the modality encoder to be learned. The contrastive loss for alignment is formulated as:

$$\mathcal{L} = -\frac{1}{2B|\mathcal{A}|} \sum_{i=1}^B \sum_{k=1}^{|\mathcal{A}|} \left( \log \frac{\exp(h_i^X \cdot h_i^{A_k} / \tau)}{\sum_j \exp(h_i^X \cdot h_j^{A_k} / \tau)} + \log \frac{\exp(h_i^{A_k} \cdot h_i^X / \tau)}{\sum_j \exp(h_i^{A_k} \cdot h_j^X / \tau)} \right),$$

where  $B$  is the batch size;  $\tau$  is a learnable temperature;  $h_i^X = \text{Norm}(\mathbf{F}(x_i))$ ,  $h_i^{A_k} = \text{Norm}(\mathbf{G}_{A_k}(a_i^k))$  are normalized features of data  $x_i$  and its anchor data  $a_i^k$  from  $A_k$ .

### 3.3. Free Lunch for Multimodal Foundation Models

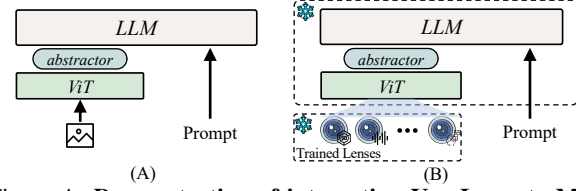


Figure 4. **Demonstration of integrating ViT-LENS to MFM.** (A) Original overall pipeline of MFM for vision; (B) Illustration of plugging well-trained Lenses of different modalities to MFM, **without** additional instruction-following training.

**Plug ViT-LENS into MFM.** Recent MFMs for vision [14, 24, 25, 46, 47, 90] enables LLMs to understand visual content. As shown in Fig. 4A, this process begins with the use of a frozen ViT to extract visual features. Subsequently, a well-trained abstractor module processes these features, constructing inputs that can be understood by the LLM.

By incorporating the ViT from MFM as part of the modality encoder and as the foundation model in VIT-LENS training, the yielded modality Lenses can be seamlessly integrated into the MFM for plug-and-play application, as depicted in Fig. 4B. In later experiments, we showcase the emergent abilities facilitated by this tuning-free adaptation.

## 4. Experiments

### 4.1. Experimental Setup

For this part, we describe the main experimental setup and provide full details in Supp.

Dataset	Task	#cls	Metric	#test
ModelNet40(MN40) [78]	3D shape cls	40	Acc	2,468
Objaverse-LVIS(O-LVIS) [15]	3D shape cls	1,156	Acc	46,832
ScanObjectNN(SONN) [73]	3D shape cls	15	Acc	581
SUN Depth-only(SUN-D) [66]	Scene cls	19	Acc	4,660
NYU-v2 Depth-only(NYU-D) [54]	Scene cls	10	Acc	654
Audioset Audio-only(AS-A) [26]	Audio cls	527	mAP	17,132 <sup>†</sup>
ESC 5-folds(ESC) [60]	Audio cls	50	Acc	2,000
Clotho(Clotho) [19]	Retrieval	-	Recall	1,046
AudioCaps(ACaps) [39]	Retrieval	-	Recall	813 <sup>†</sup>
VGGSound(VGGS) [9]	Audio cls	309	Acc	15,434 <sup>†</sup>
Touch-and-Go(TAG-M) [82]	Material cls	20	Acc	29,879
Touch-and-Go(TAG-H/S) [82]	Hard/Soft cls	2	Acc	29,879
Touch-and-Go(TAG-R/S) [82]	Rough/Smooth cls	2	Acc	8,085
ImageNet-EEG(IN-EEG) [67]	Visual Concept cls	40	Acc	1,997

Table 1. **Details of Downstream Datasets** across various modalities including 3D, depth, audio, tactile, and EEG. The evaluation of VIT-LENS is performed following feature alignment. The information presented includes the task type (classification/retrieval), the number of classes, the evaluation metric (Accuracy/mean Average Precision/Recall), and the quantity of test samples in each dataset.

**Pretraining Datasets.** Beyond image/video and text, we train VIT-LENS on a variety of modalities, including 3D

<sup>†</sup># test samples may differ from those used in previous work due to the unavailability of certain data.



	Top1	Top5
<b>Trained on ULIP-ShapeNet [80]</b>		
ULIP-PointNet++(ssg) [80]	55.7	75.7
ULIP-PointNet++(msg) [80]	58.4	78.2
ULIP-PointMLP [80]	61.5	80.7
ULIP-PointBERT [80]	60.4	84.0
VIT-LENS-B	65.4	92.7
VIT-LENS-L	<b>70.6</b>	<b>94.4</b>
<b>Trained on ULIP2-Objaverse [81]</b>		
ULIP2-PointNeXt [81]	49.0	79.7
ULIP2-PointBERT [81]	70.2	87.0
VIT-LENS-B	74.8	93.8
VIT-LENS-L	<b>80.6</b>	<b>95.8</b>

(a) Zero-shot 3D of classification on ModelNet40. Models are pretrained on triplets from ULIP-ShapeNet and ULIP2-Objaverse respectively.

	Objaverse-LVIS			ModelNet40			ScanObjectNN		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
<b>2D inference, no 3D training</b>									
PointCLIP [87]	1.9	4.1	5.8	19.3	28.6	34.8	10.5	20.8	30.6
PointCLIP v2 [91]	4.7	9.5	12.9	63.6	77.9	85.0	42.1	63.3	74.5
<b>Trained on OpenShape-Triplets (No LVIS) [48]</b>									
ULIP-PointBERT [80]	21.4	38.1	46.0	71.4	84.4	89.2	46.0	66.1	76.4
OpenShape-SparseConv [48]	37.0	58.4	66.9	82.6	95.0	97.5	54.9	76.8	87.0
OpenShape-PointBERT [48]	39.1	60.8	68.9	85.3	96.2	97.4	47.2	72.4	84.7
VIT-LENS-G	<b>50.1</b>	<b>71.3</b>	<b>78.1</b>	<b>86.8</b>	<b>96.8</b>	<b>97.8</b>	<b>59.8</b>	<b>79.3</b>	<b>87.7</b>
<b>Trained on OpenShape-Triplets [48]</b>									
ULIP-PointBERT [80]	26.8	44.8	52.6	75.1	88.1	93.2	51.6	72.5	82.3
OpenShape-SparseConv [48]	43.4	64.8	72.4	83.4	95.6	97.8	56.7	78.9	88.6
OpenShape-PointBERT [48]	46.8	69.1	77.0	84.4	96.5	98.0	52.2	79.7	88.7
VIT-LENS-G	<b>52.0</b>	<b>73.3</b>	<b>79.9</b>	<b>87.6</b>	<b>96.6</b>	<b>98.4</b>	<b>60.1</b>	<b>81.0</b>	<b>90.3</b>

(b) Zero-shot 3D classification on Objaverse-LVIS, ModelNet40 and ScanObjectNN. Models are pretrained on OpenShape Triplets. "NO LVIS" denotes excluding the Objaverse-LVIS subset.

Table 2. Zero-shot 3D classification on downstream datasets, measured in accuracy(%).

point cloud, depth, audio, tactile, and EEG data. These datasets are anchored to text descriptions, images, or both for feature alignment.

For 3D point cloud experiments, we utilize a combination of ShapeNet [8], 3D-FUTURE [22], ABO [13], and Objaverse [15]. We incorporate rendered images and text captions from previous works, resulting in three pretraining datasets: ULIP-ShapeNet [80], ULIP-2-Objaverse [81], and OpenShape-Triplets [48]. Depth data is sourced from the SUN RGB-D dataset [66], utilizing paired image and scene labels for alignment. Audio data is obtained from the Audioset dataset [26], accompanied by associated video and text label metadata. Tactile data is sourced from the Touch-and-Go dataset [82], featuring paired frame and material label text. Finally, EEG data from [67] is aligned with paired ShapeNet image and text labels.

**Evaluation on Downstream Understanding Tasks.** We evaluate VIT-LENS across diverse modalities and protocols via a comprehensive set of downstream tasks. The primary datasets used for evaluation are summarized in Tab. 1.

**Main Implementation Details.** We use the pretrained vision and text encoders from OpenCLIP [12]. We apply different model sizes: VIT-LENS-B based on ViT-B/16, VIT-LENS-L based on ViT-L/14, and VIT-LENS-G based on ViT-bigG/14.

For 3D point cloud data, we follow the baseline methods [48, 80] to uniformly sample 8,192 or 10,000 points and grouping them into sub-clouds through Farthest Point Sampling (FPS) followed by KNN grouping of neighboring points. For depth input, we follow [28] to use in-filled depth values and convert them to disparity for scale normalization. For audio data, we sample 5-second clips and extract a single frame randomly from the video clip if video serves as anchor data. The audio waveform is converted into a sequence of 128-dimensional log Mel filterbank features using a 25ms Hamming window every 10ms, following [30]. For tactile

input, we use RGB data collected from GelSight [38]. For EEG signals, we employ 128-channel temporal sequences and use the frequency range of 5-95Hz following [4].

## 4.2. Results on Understanding Tasks

**Zero-shot 3D Classification.** We follow [48, 80, 81] to use (point cloud, image, text) triplets to train VIT-LENS. We conduct zero-shot classification on downstream benchmarks. The overall results can be found in Tab. 2. In particular, when pretrained on ULIP-ShapeNet or ULIP2-Objaverse, VIT-LENS outperforms ULIP with different 3D encoders [50, 61, 62, 85], as is shown in Tab. 2a.

We present the results of training on OpenShape-Triplets in Tab. 2b. To align with [48], we adopt VIT-LENS-G and train on both "NO LVIS" (excluding all shapes from the Objaverse-LVIS subset) and the entire set for comparison. VIT-LENS outperforms models adopted in OpenShape [48]. Notably, VIT-LENS significantly improves the accuracy on the long-tail categories of Objaverse-LVIS, from 46.8% to 52.0%. Additionally, when trained on the NO LVIS subset, VIT-LENS achieves a top-1 accuracy of 50.1%. This performance beats ULIP by roughly 30% and surpasses OpenShape-PointBERT trained on the entire set by 3.3%, demonstrating the data-efficient merit of VIT-LENS. Regarding ModelNet40, VIT-LENS achieves an 87.4% accuracy, surpassing previous SOTA. Moreover, on ScannetObjectNN, which contains challenging real scans with noise and occlusion, our method exhibits decent sim-to-real transfer ability. It achieves a 60.1% zero-shot accuracy without specific sim-to-real training, surpassing the previous SOTA.

**Audio Classification and Retrieval.** In our comparison presented in Tab. 3, VIT-LENS-L consistently outperforms prior approaches in both audio classification and text-to-audio retrieval tasks. When aligned to images (I), VIT-LENS-L outperforms ImageBind based on Huge CLIP [12], and


	anchor	AudioSet mAP	VGGSound <sup>◊</sup> Top1	ESC <sup>◊</sup> Top1	Clotho <sup>◊</sup>		AudioCaps <sup>◊</sup>	
					R@1	R@10	R@1	R@10
AVFIC [53]	-	-	-	-	3.0	17.5	8.7	37.7
ImageBind-H [28]	I	17.6	27.8	66.9	6.0	28.4	9.3	42.3
VIT-LENS-L	I	23.1	28.2	69.2	6.8	29.6	12.2	48.7
AudioCLIP [33]	I+T	25.9	-	69.4	-	-	-	-
VIT-LENS-L	I+T	<b>26.7</b>	<b>31.7</b>	<b>75.9</b>	<b>8.1</b>	<b>31.2</b>	<b>14.4</b>	<b>54.9</b>
Prev. ZS SOTA	-	-	29.1/46.2* [77]	91.8 [75]	6.0	28.4 [28]	9.3	42.3 [28]

Table 3. Audio classification and retrieval on Audioset, VGGSound, ESC, Clotho and AudioCaps. <sup>◊</sup>denotes zero-shot evaluation. Gray-out denotes using larger audio-text datasets in pretraining. \*denotes using augmented captions for training.


	anchor	NYU-D	SUN-D
Text Paired [28]	T*	41.9	25.4
ImageBind-H [28]	I	54.0	35.1
VIT-LENS-L	I	64.2	37.4
VIT-LENS-L	I+T	<b>68.5</b>	<b>52.2</b>
Supervised SOTA [27]	-	76.7	64.9

Table 5. Depth-only scene classification on NYU-D and SUN-D. \* [28] rendered depth as grayscale images for direct testing. The supervised SOTA [27] used RGBD as input and extra training data.


	anchor	Material	H/S	R/S
ImageBind-B*	I	24.2	65.7	69.8
VIT-LENS-B	I	29.9	72.4	77.9
VIT-LENS-L	I	31.2	74.3	<b>78.2</b>
VIT-LENS-L	I+T	<b>65.8</b>	<b>74.7</b>	63.8
<i>Linear Probing</i>				
CMC [70, 82]	I	54.7	77.3	79.4
VIT-LENS-B	I	<b>63.0</b>	<b>92.0</b>	<b>85.1</b>

Table 6. Tactile classification on Touch-and-go. \*denotes our implementation. H/S: Hard/Soft; R/S: Rough/Smooth.


	modality	R@1	R@5	R@10
MIL-NCE [51]	V	8.6	16.9	25.8
SupportSet [57]	V	10.4	22.2	30.0
AVFIC [53]	A+V	19.4	39.5	50.3
ImageBind-H [28]	A+V	36.8	61.8	70.0
VIT-LENS-L	A+V	<b>37.6</b>	<b>63.2</b>	<b>72.6</b>
Zero-shot SOTA [10]	V	49.3	68.3	73.9

Table 4. Video Retrieval on MSRVT. V: use video; A+V: use audio and video. Gray-out means using video data in pretraining.


	anchor	Val	Test
ImageBind-B*	I	17.3	18.4
DreamDiffusion-L# [4]	I	20.4	19.2
VIT-LENS-B	I	24.6	25.3
VIT-LENS-L	I	29.3	29.2
VIT-LENS-L	I+T	<b>41.8</b>	<b>42.7</b>

Table 7. Visual concept classification on ImageNet-EEG. \*denotes our implementation. #We use the released EEG encoder and paired text encoder for inference. We report results on Val and Test set.

AVFIC [53], which leverages automatically mined audio-text pairs for alignment. When aligned to images and texts (I+T), VIT-LENS-L demonstrates stronger performance and significantly outperforms AudioCLIP [33]. Although AudioCLIP uses a audio encoder pretrained with Audioset supervised classification, it falls behind VIT-LENS-L. Additionally, on zero-shot VGGSound classification, VIT-LENS-L surpasses the SOTA [77] when class names are used as text supervision for alignment.

**Audio and Video Retrieval.** We use the MSR-VTT [79] benchmark to evaluate the text to audio and video retrieval performance, as presented in Tab. 4. We follow [28] to combine audio (A) and video (V) modalities. VIT-LENS outperforms several prior methods, even surpassing those that incorporate video data for training [51, 53, 57].

**Depth-only Scene Classification.** In Tab. 5, we present our results for depth-only classifications. VIT-LENS outperforms ImageBind across SUN-D and NYU-D. By using image and text as anchor data, VIT-LENS further improves the performance and narrows the gap with the supervised SOTA model [27] with extra training data.

**Tactile Classification Tasks.** Results for tactile tasks are displayed in Tab. 6. Across various tactile classification tasks like material, hard/soft, and rough/smooth classification, VIT-LENS-B demonstrates superior performance compared to our implementation of ImageBind-B. Even trained with appearance or text labels for material, VIT-LENS can perform well on the hard/soft and rough/smooth classification tasks. This underscores the extensive knowledge transfer by CLIP during training. Furthermore, scaling up to a larger model and incorporating text during training can further boost the performance. In comparing the image-aligned VIT-

LENS-B with CMC [82] using linear probing, we observe significantly superior performance by VIT-LENS.

**EEG Visual Concept Classification.** Results in Tab. 7 show that VIT-LENS consistently outperforms our implemented ImageBind-B. Additionally, when compared to the EEG encoder from [4], which used more EEG data for MAE-style pretraining [35] and then aligned with the CLIP-L14 image encoder, VIT-LENS demonstrates superior performance.

### 4.3. Few-shot Linear Probing

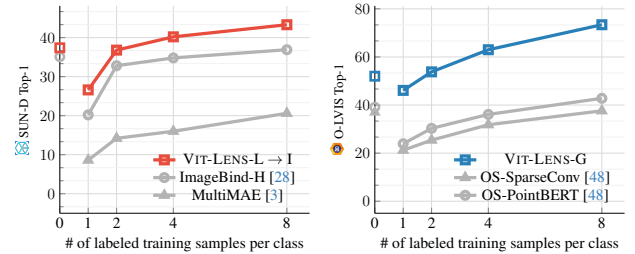


Figure 5. Few-shot linear probing on depth and 3D point cloud. We mark the zero-shot classification performance on the y-axis. We train linear classifiers on fixed features for the  $\geq 1$ -shot settings.

We evaluate the label-efficiency of VIT-LENS through few-shot linear probing using VIT-LENS-L (I) depth encoder and VIT-LENS-G 3D encoder, as shown in Fig. 5. Linear classifiers are trained on fixed representation features. For few-shot depth linear probing, we compare with ImageBind [28] and MultiMAE [3] trained on images, depth, and semantic segmentation data. VIT-LENS consistently outperforms both methods in zero-shot and few-shot settings. For few-shot 3D classification, we compare with SparseConv and

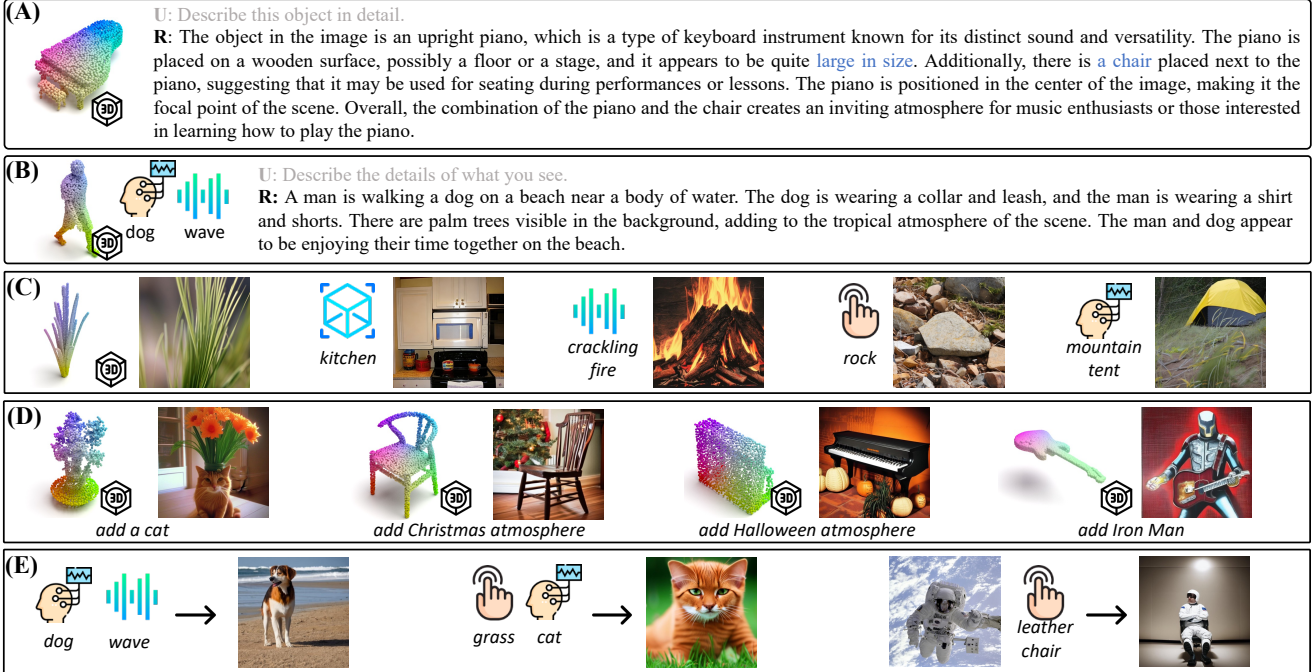


Figure 6. **Qualitative examples for plugging VIT-LENS into MFMs.** (A-B) **Integrate with InstructBLIP:** Accurately capturing concepts from single (A) or multiple modalities (B), providing detailed descriptions based on InstructBLIP’s instruction-following capability. (C-E) **Integrate with SEED:** Extending SEED’s capability to emergent compositional Any-to-image generation. (C) Single modality to image generation. (D) Text-guided any-to-image generation/editing. (E) Multi-modalities-to-image generation.

PointBERT trained in [48]. VIT-LENS significantly outperforms all methods by a large margin in all few-shot settings, showcasing its robust feature generalization capabilities.

#### 4.4. Results on VIT-LENS MFMs

In this section, we plug VIT-LENS across various modalities into off-the-shelf MFMs, and show in our experimental results that the MFMs’ capabilities can be transferred to novel modalities and their combinations, without instruction-following training. We present qualitative results below, with additional quantitative results available in Supp.

**MFM Selection in Practice.** In this work, we select InstructBLIP [14] and SEED [24, 25] to probe the emergent capabilities of the MFMs with our VIT-LENS plugged in. Both InstructBLIP and SEED utilize EVA01-g14 CLIP-ViT [21] as the visual encoder. Following the practice in Sec. 3.2, we use the same pretrained-ViT for VIT-LENS training in MFMs experiments. More details can be found in Supp.

**InstructBLIP with VIT-LENS.** InstructBLIP [14] introduced a framework for instruction tuning in a vision-language model, demonstrating its capabilities in tasks like complex visual reasoning and image descriptions. We show in our experiment that these capabilities can be effectively extended to novel modalities through the integration of VIT-LENS. Qualitative examples in Fig. 6 (A-B) showcase the model’s ability to follow instructions across various modalities,

enabling Any-modality QA, captioning, *etc.* Additionally, the model demonstrates precise and detailed descriptions, such as identifying a small “chair” next to a giant piano in (A), emphasizing the superior alignment achieved by VIT-LENS.

**SEED with VIT-LENS.** SEED-LLaMA [25] is an MFM distinguished by its capacity for multimodal comprehension and image generation. This is achieved through multimodal pretraining and instruction tuning along with its SEED tokenizer [24]. We present qualitative results of integrating VIT-LENS with SEED in Fig. 6 (E-G). The outcomes illustrate how the combined model extends SEED’s capabilities to diverse modalities. Examples in (E-G) show the ability of compositional any-to-image generation [25]. It can translate input from any modality into an image, generate an image based on a text prompt given input from any modality, and seamlessly blend visual concepts from combinations of any modalities into a coherent and plausible image.

#### 4.5. Ablation Study

We conduct ablations studies to investigate the effectiveness of various designs for VIT-LENS in omni-modal learning. We report the main results here and full details are in Supp.

**Lens designs for different modalities.** We study the effect of Lens designs as outlined in Sec. 3.1 for different modalities. We use VIT-LENS-B and set comparable amount of



Test Dataset ▶	MN40	SUN-D	ESC	TAG-M	IN-EEG
S-Attn w/o pt weights	63.8	48.6	70.1	61.8	25.4
S-Attn w/ pt weights	65.4	50.9	70.9	63.6	26.3
Iter-CS-Attn	65.4	47.5	71.2	60.6	35.9

Table 8. **Lens designs** for different modalities. All modalities are aligned to “I+T”. Lens w/ pt weights means tuning corresponding Self-Attn blocks in the pretrained-ViT, and w/o means random initialization. Default setting is marked with color box.

trainable parameters for the two variants. We also examine the effectiveness of initializing S-Attn type Lens with pretrained weights. We train 3D point cloud on ULIP-ShapeNet and follow the main settings for other modalities. The results are shown in Tab. 8. We observe consistent performance enhancement by initializing S-Attn Lens with pretrained weights. For image-like inputs such as depth maps and RGB-based tactile data, the S-Attn design exhibits superiority. Conversely, modalities significantly different from image inputs, like 3D point clouds, audio spectrograms, and EEG, benefit more from Iter-CS-Attn design. Additionally, it reduces the computational overhead by reducing the input length for ViT. Further details are available in the Supp.

**Modality encoder designs and settings.** We investigate the efficacy of integrating a set of pretrained-ViT layers into the modality encoder. We use the same datasets for training and testing as in the Lens design ablation. We compare VIT-LENS-B with an architecture that combines ModEmbed and ViT and employing different settings for the ViT component, as detailed in Tab. 9. Results indicate that simply adding the ModEmbed to a pretrained-ViT cannot fully exploit the the potential of the pretrained-ViT (#2). Training the entire encoder with pretrained weights outperforms training from scratch, highlighting the effectiveness of utilizing the pre-trained weights for learning (#1 vs #3). In comparison to #3, VIT-LENS-B achieves comparable or better performance, especially for the less common modalities. Moreover, our VIT-LENS employs fewer trainable parameters than training the entire encoder and reduces computational overhead for modalities with lengthy inputs. Consequently, by introducing Lens, VIT-LENS effectively and efficiently transfers the capabilities of pretrained-ViT to various modalities.

Test Dataset ▶	MN40	SUN-D	ESC	TAG-M	IN-EEG
#1 M.E. → ViT (scratch)	62.4	46.3	68.8	55.6	20.5
#2 M.E. → ViT (pt, frozen)	50.0	36.8	54.9	24.8	14.2
#3 M.E. → ViT (pt, tune)	67.4	48.2	71.6	59.4	27.2
VIT-LENS-B	65.4	50.9	71.2	63.6	35.9

Table 9. **Encoder designs and settings** for different modalities. All modalities are aligned to “I+T”. M.E. denotes ModEmbed. VIT-LENS-B is the default setting.

**Scaling up foundation model and VIT-LENS.** We explore the effectiveness of scaling up VIT-LENS for feature alignment. We conduct experiments to pretrain for 3D on ULIP-ShapeNet, and depth on SUN-D. While previous works [28, 48] show that scaling to a large encoder(>100M) degrades the performance, we show in Fig. 7 that scaling up

VIT-LENS can improve the 3D and depth representation and enhance performance.

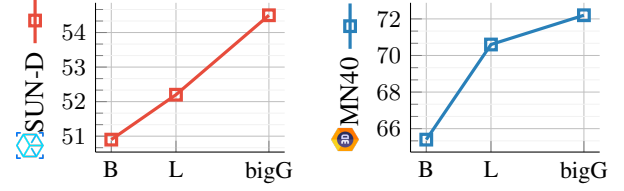


Figure 7. **Scaling the VIT-LENS on depth and 3D point cloud.** B: VIT-LENS-B, L: VIT-LENS-L, bigG: VIT-LENS-G.

**Different pretrained-ViTs for VIT-LENS.** We evaluate different pretrained-ViT variants for omni-modal representation learning. We use CLIP-ViT-bigG/14 as the teacher foundation model and apply different ViTs for the modality encoder. We use the same datasets for training and testing as in the model scaling ablation. Results in Tab. 10 demonstrate that the use of pretrained-ViTs including the self-supervised and CLIP pretrained variants, outperforms training from scratch on both depth and 3D modalities. This indicates that different pretrained-ViTs possess the potential to serve as effective omni-modal learners.

ViT variant ▶	RndInit	DINO [7]	OpenCLIP	OpenCLIP	OpenCLIP
	ViT-B16	ViT-B16	ViT-B16	ViT-L14	ViT-bigG14
SUN-D	48.0	50.9	51.4	53.2	54.5
MN40	66.2	68.5	68.3	71.4	72.2

Table 10. **Different ViT** for modality encoders in VIT-LENS. we train the entire encoder for the baseline RndInit (random initialization), while others follow VIT-LENS training setting.

## 5. Discussion and Limitations

VIT-LENS is a straightforward yet effective method to advance omni-modal representations. It leverages the rich knowledge embedded in pretrained-ViT from extensive data, and eliminates the need for separate modality-specific architectures. We demonstrate the effectiveness of VIT-LENS across various modalities, including 3D point cloud, depth, audio, tactile, and EEG, achieving leading performance in understanding-based tasks. Furthermore, integrating VIT-LENS into existing MFMs unlocks new capabilities, such as any-modality instruction following and any-modality-to-image generation. We anticipate that VIT-LENS will inspire further research and innovation in omni-modal representation learning, fostering the development of more versatile and robust AI systems. However, it is important to acknowledge that VIT-LENS may inherit biases and errors from the pretrained-ViT, and it cannot be readily used for deployment in real-world scenarios.

**Acknowledgements.** This project is supported by the National Research Foundation, Singapore under its NRFF Award NRF-NRFF13-2021-0008.



## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv*, 2022. 2, 3, 4
- [2] Relja Arandjelovic and Andrew Zisserman. Look, listen and learn. In *ICCV*, 2017. 2
- [3] Roman Bachmann, David Mizrahi, Andrei Atanov, and Amir Zamir. Multima: Multi-modal multi-task masked autoencoders. In *ECCV*, 2022. 6
- [4] Yunpeng Bai, Xintao Wang, Yanpei Cao, Yixiao Ge, Chun Yuan, and Ying Shan. Dreamdiffusion: Generating high-quality images from brain eeg signals. *arXiv*, 2023. 1, 2, 5, 6
- [5] Hangbo Bao, Li Dong, and Furu Wei. Beit: Bert pre-training of image transformers. *arXiv*, 2021. 2
- [6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *NeurIPS*, 2020. 2
- [7] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *ICCV*, 2021. 2, 3, 8
- [8] Angel X Chang, Thomas Funkhouser, Leonidas Guibas, Pat Hanrahan, Qixing Huang, Zimo Li, Silvio Savarese, Manolis Savva, Shuran Song, Hao Su, et al. Shapenet: An information-rich 3d model repository. *arXiv*, 2015. 5
- [9] Honglie Chen, Weidi Xie, Andrea Vedaldi, and Andrew Zisserman. Vggsound: A large-scale audio-visual dataset. In *ICASSP*, 2020. 4
- [10] Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset. *arXiv*, 2023. 6
- [11] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. *arXiv*, 2022. 2
- [12] 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. *arXiv*, 2022. 2, 3, 4, 5
- [13] Jasmine Collins, Shubham Goel, Kenan Deng, Achleshwar Luthra, Leon Xu, Erhan Gundogdu, Xi Zhang, Tomas F Yago Vicente, Thomas Dideriksen, Himanshu Arora, et al. Abo: Dataset and benchmarks for real-world 3d object understanding. In *CVPR*, 2022. 5
- [14] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *arXiv*, 2023. 2, 4, 7
- [15] Matt Deitke, Dustin Schwenk, Jordi Salvador, Luca Weihs, Oscar Michel, Eli VanderBilt, Ludwig Schmidt, Kiana Ehsani, Aniruddha Kembhavi, and Ali Farhadi. Objaverse: A universe of annotated 3d objects. In *CVPR*, 2023. 2, 4, 5
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, 2018. 2, 3
- [17] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv*, 2020. 2
- [18] Danny Driess, Fei Xia, Mehdi SM Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, et al. Palm-e: An embodied multimodal language model. *arXiv*, 2023. 2
- [19] Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP*, 2020. 4
- [20] Yuxin Fang, Quan Sun, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva-02: A visual representation for neon genesis. *arXiv*, 2023. 2
- [21] Yuxin Fang, Wen Wang, Binhui Xie, Quan Sun, Ledell Wu, Xinggang Wang, Tiejun Huang, Xinlong Wang, and Yue Cao. Eva: Exploring the limits of masked visual representation learning at scale. In *CVPR*, 2023. 2, 3, 7
- [22] Huan Fu, Rongfei Jia, Lin Gao, Mingming Gong, Binqiang Zhao, Steve Maybank, and Dacheng Tao. 3d-future: 3d furniture shape with texture. In *IJCV*, 2021. 5
- [23] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. Multi-modal graph neural network for joint reasoning on vision and scene text. In *CVPR*, 2020. 2
- [24] Yuying Ge, Yixiao Ge, Ziyun Zeng, Xintao Wang, and Ying Shan. Planting a seed of vision in large language model. *arXiv*, 2023. 2, 3, 4, 7
- [25] Yuying Ge, Sijie Zhao, Ziyun Zeng, Yixiao Ge, Chen Li, Xintao Wang, and Ying Shan. Making llama see and draw with seed tokenizer. *arXiv*, 2023. 2, 3, 4, 7
- [26] 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 *ICASSP*, 2017. 4, 5
- [27] Rohit Girdhar, Mannat Singh, Nikhila Ravi, Laurens van der Maaten, Armand Joulin, and Ishan Misra. Omnivore: A single model for many visual modalities. In *CVPR*, 2022. 2, 6
- [28] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *CVPR*, 2023. 1, 2, 5, 6, 8
- [29] Rohit Girdhar, Alaaeldin El-Nouby, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Omnima: Single model masked pretraining on images and videos. In *CVPR*, 2023. 2
- [30] Yuan Gong, Yu-An Chung, and James Glass. AST: Audio Spectrogram Transformer. In *Interspeech*, 2021. 5

- [31] Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. *arXiv*, 2021. [2](#)
- [32] Ziyu Guo, Renrui Zhang, Xiangyang Zhu, Yiwen Tang, Xianzheng Ma, Jiaming Han, Kexin Chen, Peng Gao, Xianzhi Li, Hongsheng Li, et al. Point-bind & point-llm: Aligning point cloud with multi-modality for 3d understanding, generation, and instruction following. *arXiv*, 2023. [3](#)
- [33] Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP*, 2022. [2](#), [4](#), [6](#)
- [34] Jiaming Han, Renrui Zhang, Wenqi Shao, Peng Gao, Peng Xu, Han Xiao, Kaipeng Zhang, Chris Liu, Song Wen, Ziyu Guo, et al. Imagebind-llm: Multi-modality instruction tuning. *arXiv*, 2023. [3](#)
- [35] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *CVPR*, 2022. [2](#), [3](#), [6](#)
- [36] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *ICML*, 2021. [4](#)
- [37] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *ICML*, 2021. [2](#)
- [38] Micah K. Johnson and Edward H. Adelson. Retrographic sensing for the measurement of surface texture and shape. In *CVPR*, 2009. [5](#)
- [39] Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*, 2019. [4](#)
- [40] Boyi Li, Kilian Q Weinberger, Serge Belongie, Vladlen Koltun, and Rene Ranftl. Language-driven semantic segmentation. In *ICLR*, 2022. [2](#)
- [41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, 2022. [3](#)
- [42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv*, 2023. [3](#), [4](#)
- [43] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv*, 2023. [3](#)
- [44] Valerii Likhoshesterov, Anurag Arnab, Krzysztof Choromanski, Mario Lucic, Yi Tay, Adrian Weller, and Mostafa Dehghani. Polyvit: Co-training vision transformers on images, videos and audio. *arXiv*, 2021. [2](#)
- [45] Kevin Qinghong Lin, Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Z XU, Difei Gao, Rong-Cheng Tu, Wenzhe Zhao, Weijie Kong, et al. Egocentric video-language pretraining. *NeurIPS*, 2022. [2](#)
- [46] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. 2023. [3](#), [4](#)
- [47] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv*, 2023. [2](#), [3](#), [4](#)
- [48] Minghua Liu, Ruoxi Shi, Kaiming Kuang, Yinhao Zhu, Xuanlin Li, Shizhong Han, Hong Cai, Fatih Porikli, and Hao Su. Openshape: Scaling up 3d shape representation towards open-world understanding. *arXiv*, 2023. [1](#), [2](#), [4](#), [5](#), [6](#), [7](#), [8](#)
- [49] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv*, 2019. [2](#), [3](#)
- [50] Xu Ma, Can Qin, Haoxuan You, Haoxi Ran, and Yun Fu. Rethinking network design and local geometry in point cloud: A simple residual mlp framework. *arXiv*, 2022. [5](#)
- [51] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019. [6](#)
- [52] Pedro Morgado, Nuno Vasconcelos, and Ishan Misra. Audio-visual instance discrimination with cross-modal agreement. In *CVPR*, 2021. [2](#)
- [53] Arsha Nagrani, Paul Hongsuck Seo, Bryan Seybold, Anja Hauth, Santiago Manen, Chen Sun, and Cordelia Schmid. Learning audio-video modalities from image captions. In *ECCV*, 2022. [6](#)
- [54] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgb-d images. In *ECCV*, 2012. [4](#)
- [55] OpenAI. Introducing chatgpt. OpenAI Blog, 2021. [2](#), [3](#)
- [56] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv*, 2023. [2](#), [3](#)
- [57] Mandela Patrick, Po-Yao Huang, Yuki Asano, Florian Metze, Alexander G Hauptmann, Joao F. Henriques, and Andrea Vedaldi. Support-set bottlenecks for video-text representation learning. In *ICLR*, 2021. [6](#)
- [58] Songyou Peng, Kyle Genova, Chiyu "Max" Jiang, Andrea Tagliasacchi, Marc Pollefeys, and Thomas Funkhouser. Open-scene: 3d scene understanding with open vocabularies. In *CVPR*, 2023. [2](#)
- [59] Zhiliang Peng, Li Dong, Hangbo Bao, Qixiang Ye, and Furu Wei. Beit v2: Masked image modeling with vector-quantized visual tokenizers. *arXiv*, 2022. [2](#)
- [60] Karol J. Piczak. ESC: Dataset for Environmental Sound Classification. In *ACM MM*, 2015. [4](#)
- [61] Charles Ruizhongtai Qi, Li Yi, Hao Su, and Leonidas J Guibas. Pointnet++: Deep hierarchical feature learning on point sets in a metric space. *NeurIPS*, 2017. [5](#)
- [62] Guocheng Qian, Yuchen Li, Houwen Peng, Jinjie Mai, Hasan Hammoud, Mohamed Elhoseiny, and Bernard Ghanem. Pointnext: Revisiting pointnet++ with improved training and scaling strategies. *NeurIPS*, 2022. [5](#)
- [63] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. *OpenAI blog*, 2018. [2](#), [3](#)
- [64] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 2019. [2](#), [3](#)
- [65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry,

- Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *ICML*, 2021. 2, 3, 4
- [66] Shuran Song, Samuel P Lichtenberg, and Jianxiong Xiao. Sun rgb-d: A rgb-d scene understanding benchmark suite. In *CVPR*, 2015. 4, 5
- [67] Concetto Spampinato, Simone Palazzo, Isaak Kavasidis, Daniela Giordano, Nasim Souly, and Mubarak Shah. Deep learning human mind for automated visual classification. In *CVPR*, 2017. 4, 5
- [68] Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. Pandagpt: One model to instruction-follow them all. *arXiv*, 2023. 3
- [69] Quan Sun, Qiyang Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, Yueze Wang, Hongcheng Gao, Jingjing Liu, Tiejun Huang, and Xinlong Wang. Generative pretraining in multi-modality. *arXiv*, 2023. 3
- [70] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 6
- [71] Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *ECCV*, 2020. 2
- [72] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv*, 2023. 2, 3
- [73] Mikaela Angelina Uy, Quang-Hieu Pham, Binh-Son Hua, Thanh Nguyen, and Sai-Kit Yeung. Revisiting point cloud classification: A new benchmark dataset and classification model on real-world data. In *ICCV*, 2019. 4
- [74] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *NeurIPS*, 2017. 3
- [75] 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*, 2023. 2, 3, 6
- [76] Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. Next-gpt: Any-to-any multimodal llm. *arXiv*, 2023. 3
- [77] Yusong Wu\*, Ke Chen\*, Tianyu Zhang\*, Yuchen Hui\*, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *ICASSP*, 2023. 6
- [78] Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 3d shapenets: A deep representation for volumetric shapes. In *CVPR*, 2015. 4
- [79] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. *CVPR*, 2016. 6
- [80] Le Xue, Mingfei Gao, Chen Xing, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In *CVPR*, 2023. 2, 4, 5
- [81] Le Xue, Ning Yu, Shu Zhang, Junnan Li, Roberto Martín-Martín, Jiajun Wu, Caiming Xiong, Ran Xu, Juan Carlos Niebles, and Silvio Savarese. Ulip-2: Towards scalable multi-modal pre-training for 3d understanding. *arXiv*, 2023. 2, 4, 5
- [82] Fengyu Yang, Chenyang Ma, Jiacheng Zhang, Jing Zhu, Wenzhen Yuan, and Andrew Owens. Touch and go: Learning from human-collected vision and touch. In *NeurIPS*, 2022. 1, 2, 4, 5, 6
- [83] Kim Youwang, Kim Ji-Yeon, and Tae-Hyun Oh. Clip-actor: Text-driven recommendation and stylization for animating human meshes. In *ECCV*, 2022. 2
- [84] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *arXiv*, 2022. 2
- [85] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In *CVPR*, 2022. 5
- [86] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. In *CVPR*, 2022. 2
- [87] Renrui Zhang, Ziyu Guo, Wei Zhang, Kunchang Li, Xupeng Miao, Bin Cui, Yu Qiao, Peng Gao, and Hongsheng Li. Point-clip: Point cloud understanding by clip. In *CVPR*, 2022. 2, 5
- [88] Yiyuan Zhang, Kaixiong Gong, Kaipeng Zhang, Hongsheng Li, Yu Qiao, Wanli Ouyang, and Xiangyu Yue. Meta-transformer: A unified framework for multimodal learning. *arXiv*, 2023. 3
- [89] Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *ECCV*, 2022. 2
- [90] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv*, 2023. 2, 3, 4
- [91] Xiangyang Zhu, Renrui Zhang, Bowei He, Ziyao Zeng, Shanghang Zhang, and Peng Gao. Pointclip v2: Adapting clip for powerful 3d open-world learning. *arXiv*, 2022. 2, 5