ViT-Lens: Towards Omni-modal Representations

Weixian Lei, Yixiao Ge, Kun Yi, Jianfeng Zhang, Difei Gao, Dylan Sun, Yuying Ge, Ying Shan, Mike Zheng Shou

CVPR 2024

Presenters: Zoe Fowler and Ghazal Kaviani

Georgia Tech

Outline

- Problem Statement
- Related Works
- Approach
- Experiments & Results
- Limitations, Societal Implications
- Summary of Strengths, Weaknesses, Relationship to Other Papers

Problem Statement

Omni-modal representation learning by perceiving novel modalities

Human perception and physical world is inherently multi-modal

Available data for building AI models:

- Rich resource data:
  - Text , Image , Video , Audio
- Low resource data
  - Tactile, EMG, 3D point cloud, ...

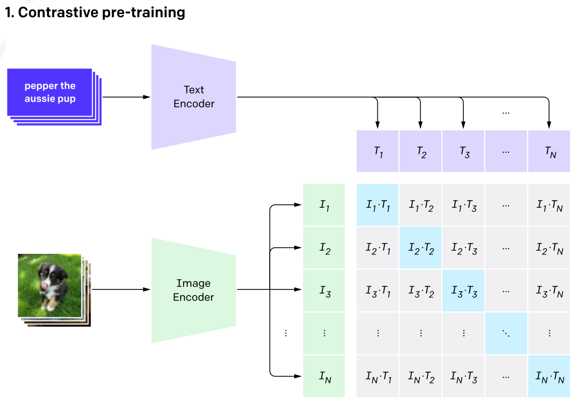
How to integrate low resource data with large scale models?

- Modality specific lens to project any-modal signals to an intermediate embedding space

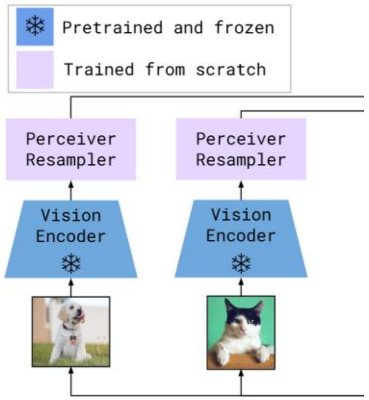
Related Works

Vision Language models (VLMs)

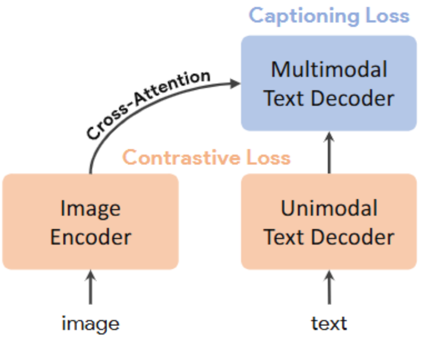
Embed visual and textual representations into a shared space



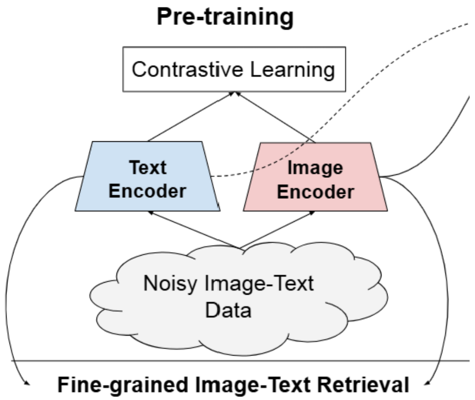
CLIP



Flamingo



CoCa



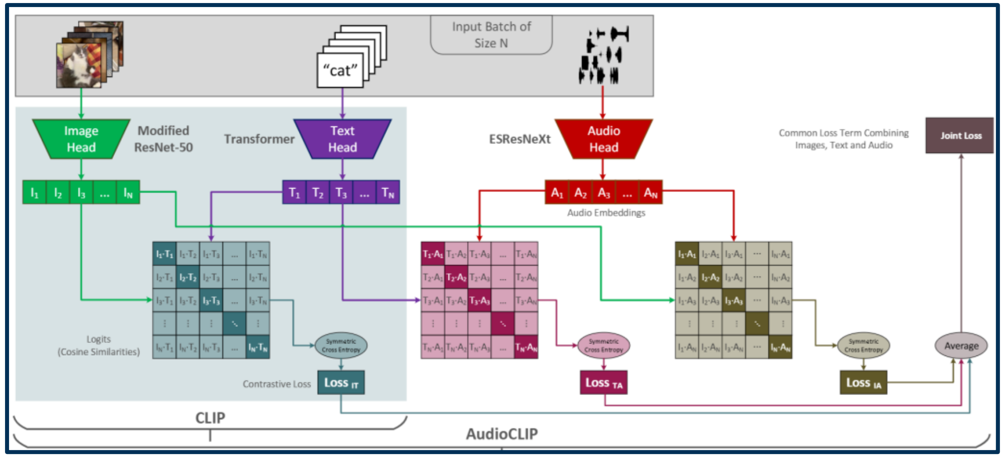
Align

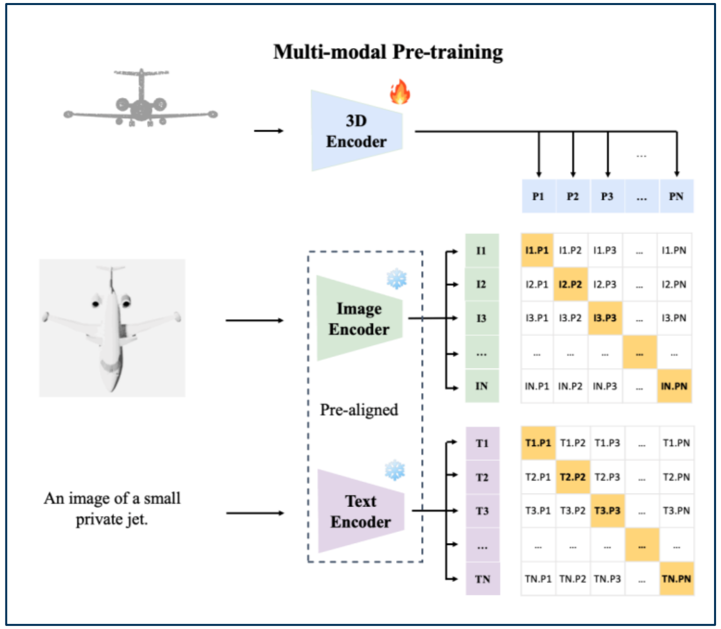
Related Works

Multimodal Foundation Models(MFMs)

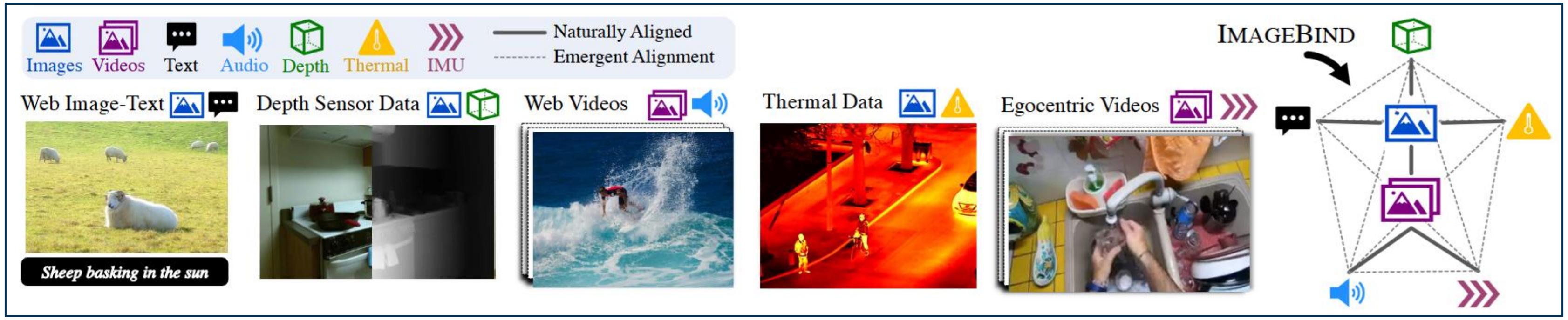
Joint training across multiple modalities

AudioCLIP





ULIP

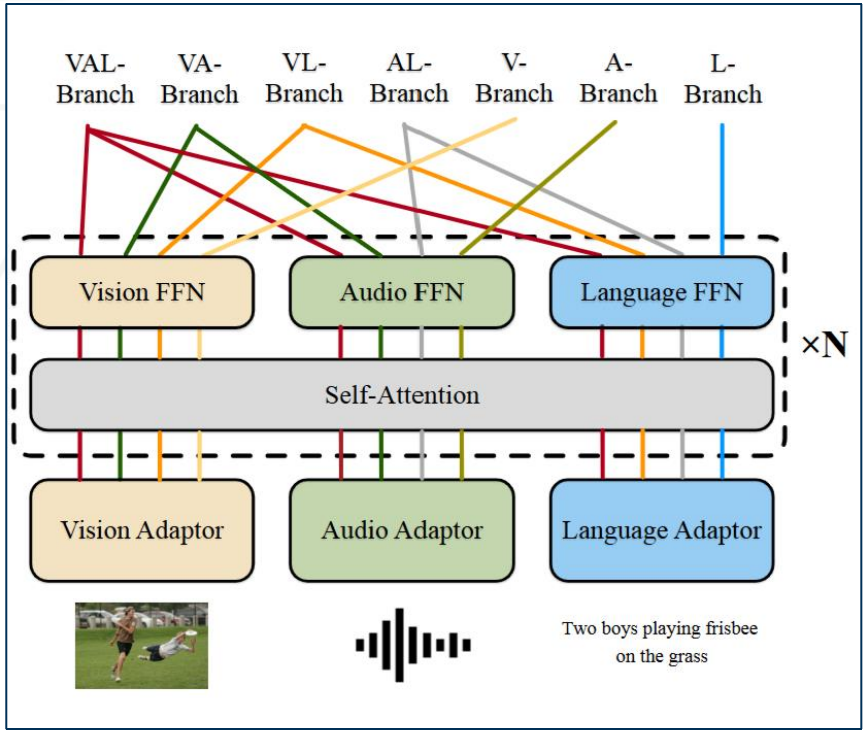


IMAGEBIND

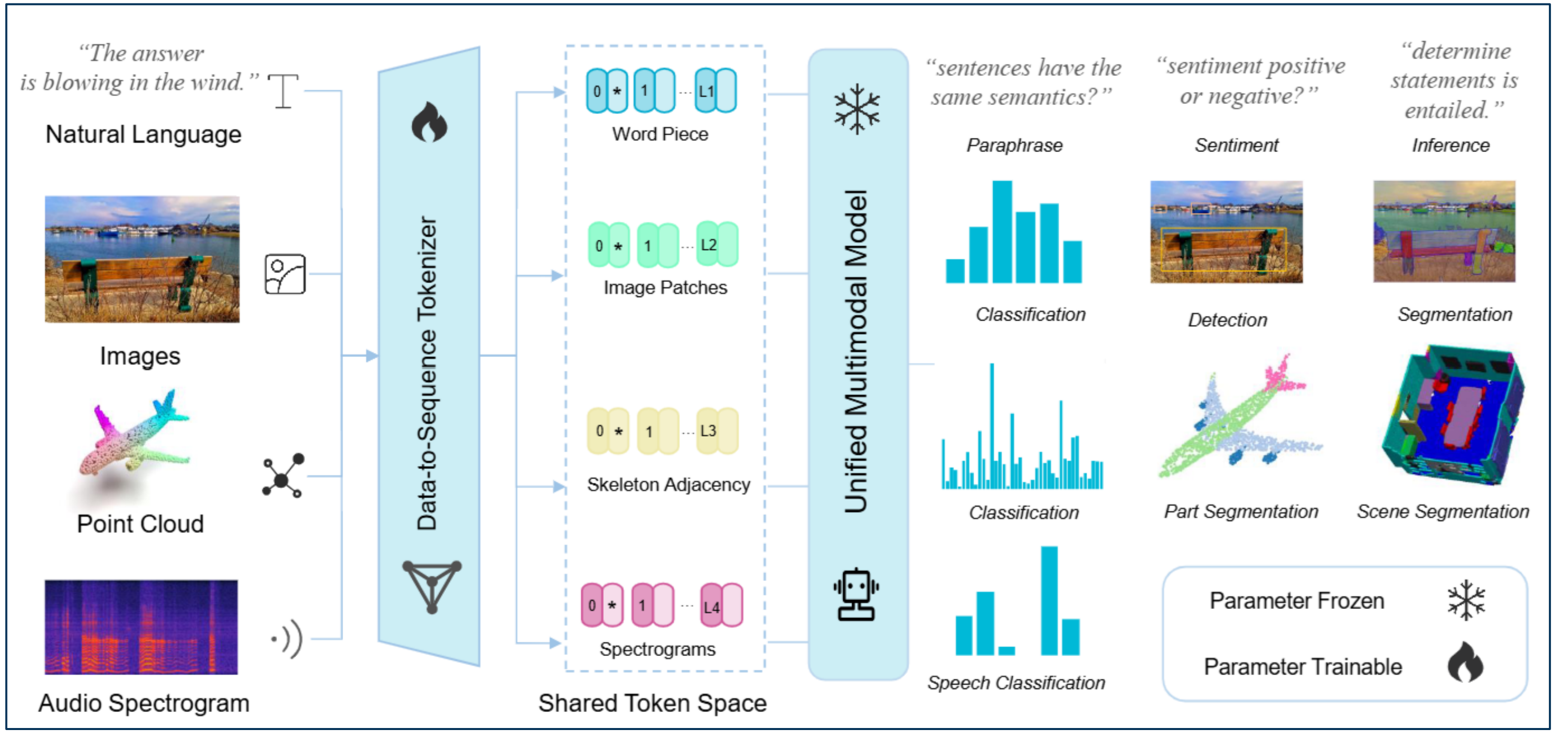
Related Works

Multimodal Foundation Models(MFMs)

Unified Encoder



ONE-PEACE



Meta-Transformer

Related Works

ONE-PEACE

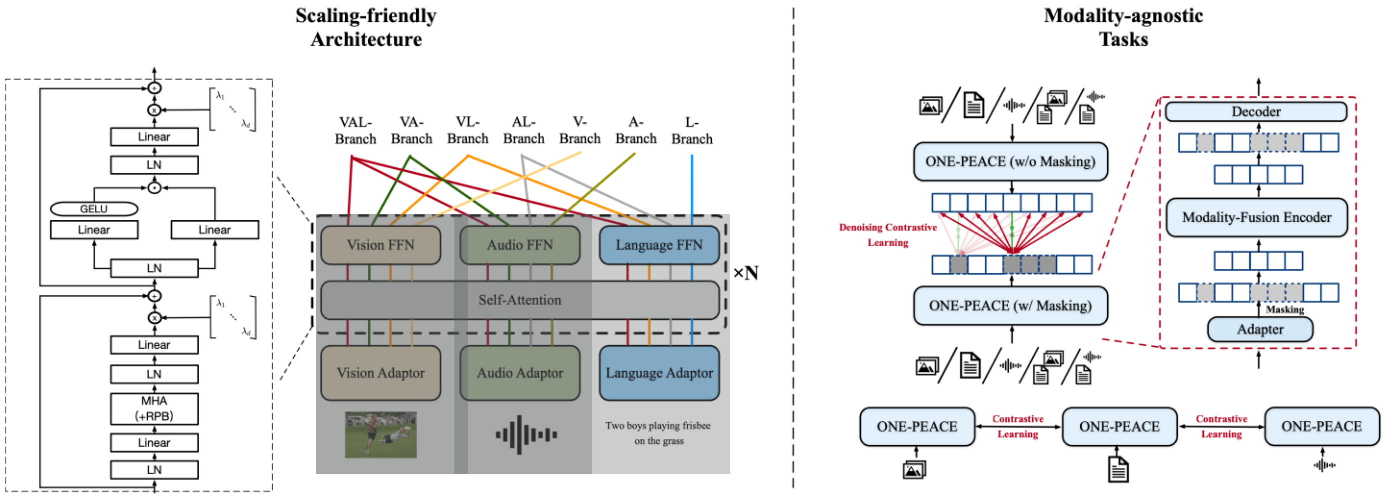
- Modality Adaptors:
  - A module converting different raw signals into unified features
  - Adapters do not interact with each other
  - The backbone can change (CNN, RNN, Transformer)
- Modality Fusion encoder:
  - A transformer block with shared self attention layer and modality-specific feedforward network.
- Sharing separated architecture enables sub-modality branches

- Pretrain Tasks:
  - Cross-Modal Contrastive Learning: aligns the semantic spaces of different modalities

$$\mathcal{L}_{CL} = -\frac{1}{2N}\sum_{i=1}^{N}\left(\log\frac{\exp(\boldsymbol{s}_i^1 \boldsymbol{s}_i^2/\sigma)}{\sum_{j=1}^{N}\exp(\boldsymbol{s}_i^1 \boldsymbol{s}_j^2/\sigma)} + \log\frac{\exp(\boldsymbol{s}_i^1 \boldsymbol{s}_i^2/\sigma)}{\sum_{j=1}^{N}\exp(\boldsymbol{s}_j^1 \boldsymbol{s}_i^2/\sigma)}\right)$$

  - Intra-Modal Denoising Contrastive Learning: emphasis on the learning of fine-grained details within modalities

$$\mathcal{L}_{DCL} = -\frac{1}{N\hat{N}}\sum_{i=1}^{N}\sum_{j=1}^{\hat{N}}\log\frac{\exp(\hat{\boldsymbol{h}}_{ij}\cdot \mathrm{sg}(\boldsymbol{h}_{ij})/\tau)}{\sum_{m=1}^{N}\sum_{n=1}^{\hat{N}}\exp(\hat{\boldsymbol{h}}_{ij}\cdot \mathrm{sg}(\boldsymbol{h}_{mn})/\tau)}$$

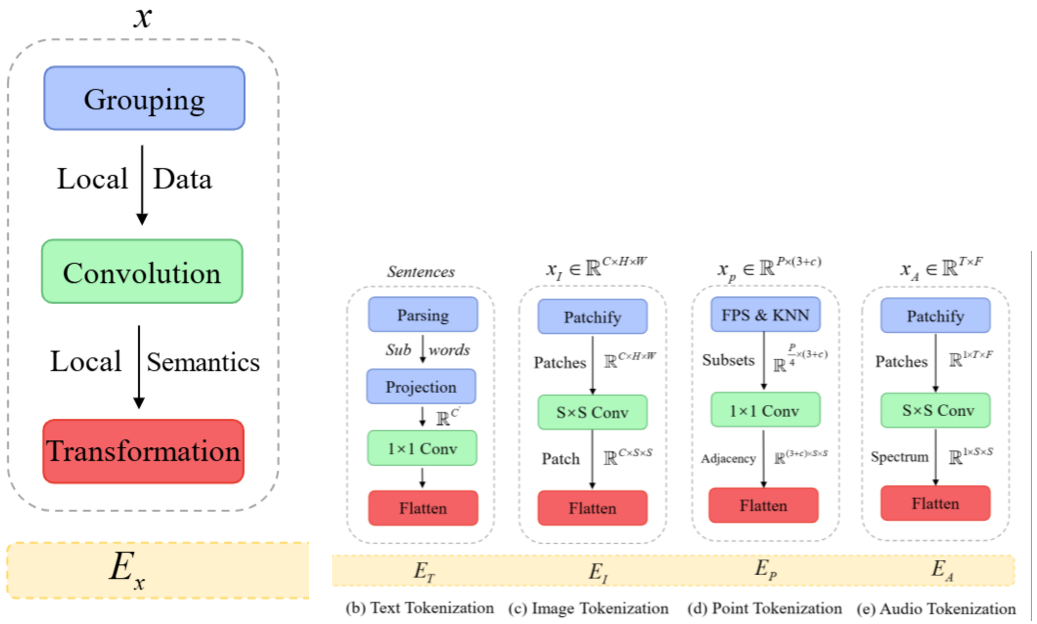


Related Works

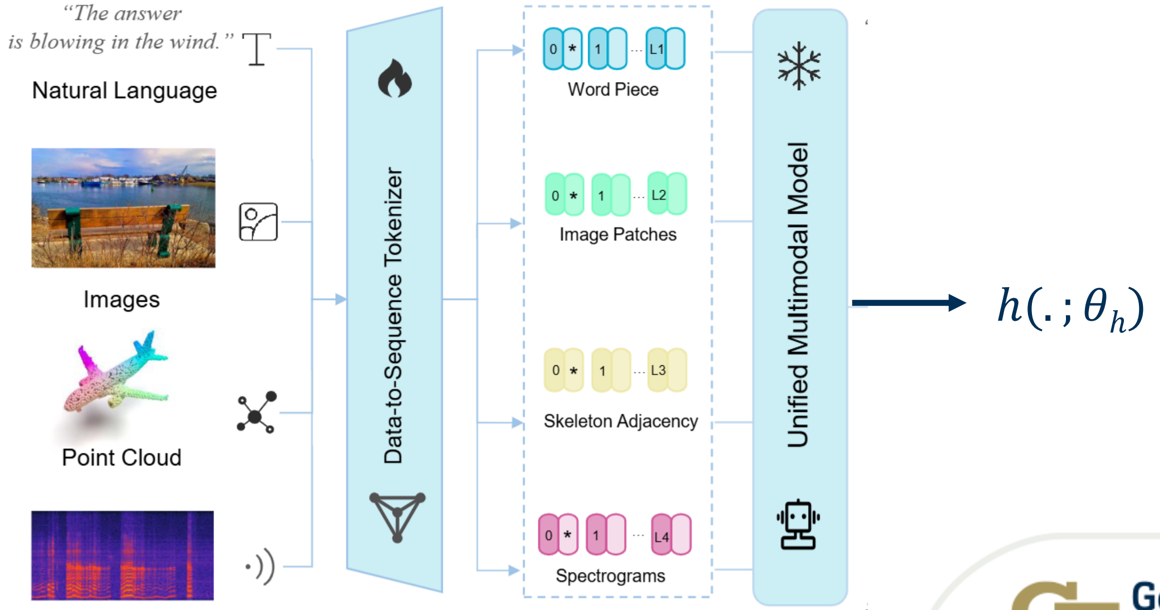
Meta-Transformer

- Tokenizer:
  - Data to Sequence tokenizer via meta scheme
- Modality-agnostic encode:
  - A ViT pretrained on LAION-2B dataset with contrastive learning
- Task-specific head $h(.;\theta_h)$:
  - Consist of MLPs
- Objective function:

$$\hat{\boldsymbol{y}} = \mathcal{F}(\boldsymbol{x};\theta^*) = h \circ g \circ f(\boldsymbol{x}), \quad \theta^* = \arg\min_{\theta} \mathcal{L}(\hat{y}, y)$$



(a) Meta Scheme (b) Text Tokenization (c) Image Tokenization (d) Point Tokenization (e) Audio Tokenization



Georgia Tech

Approach

Leveraging pre-trained large-scale model for low-resource data modalities

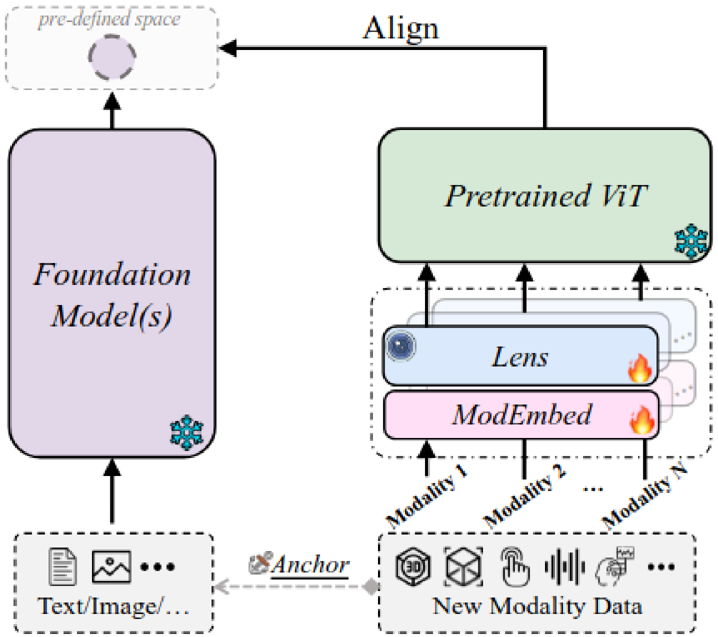
1. Modality Embedding Module:
   - Raw data modality into token embeddings
   - Module is modality-specific.

   E.g. Audio to spectrogram, tactile to spatial map

Image/Depth	patches of size P × P
Video	2 frame clip, patches of size 2×P×P
Point cloud	Point-bert [1] approach
Audio	Convert an audio with duration of t to spectrogram dimension of 128 × 100t
EEG	truncated into a common length of 512 and embed using Conv1D
Tactile	treated similarly to RGB images

P = 16 for VIT-LENS-B
P = 14 for VIT-LENS-L

Construct local patches by sampling 512 sub-clouds, each comprising 32 points. via Farthest Point Sampling (FPS) and the k-Nearest Neighbors (kNN) algorithm.

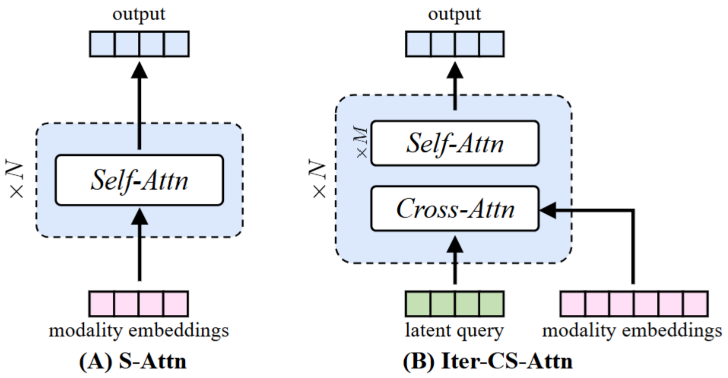


[1] Xumin Yu, Lulu Tang, Yongming Rao, Tiejun Huang, Jie Zhou, and Jiwen Lu. Point-bert: Pre-training 3d point cloud transformers with masked point modeling. In CVPR, 2022.

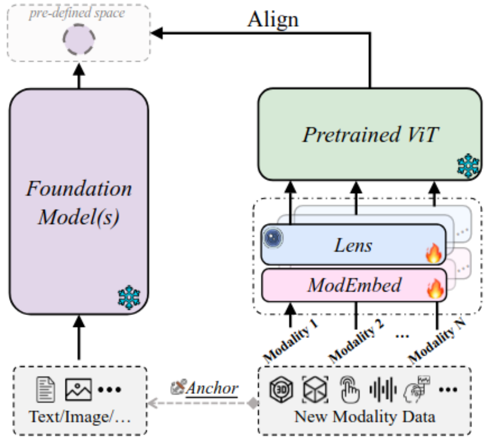
Approach

Leveraging pre-trained large-scale model for low-resource data modalities

2. Modality-*lens*:
    - Convert a sensory data from *ModEmbed* into a format that the pretrained encoder (ViT) can interpret
    - **Self-Attention Blocks (A)** for image-like data
    - **Iterative Cross-Self-Attention Blocks (B)** for lengthy sequential data



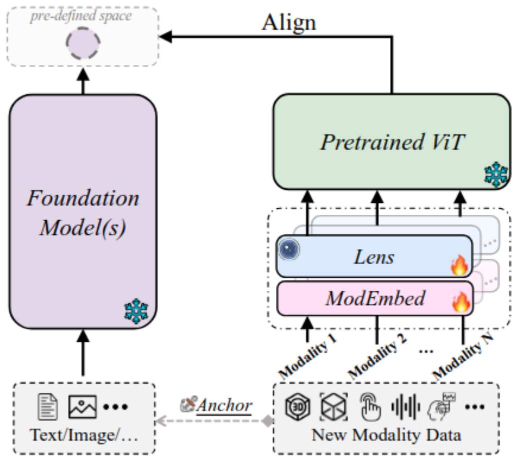
	3D Point Cloud	*Depth*	*Audio*	*Tactile*	*EEG*
ModEmbed ▶	Mini PointNet	PatchEmbed	PatchEmbed	PatchEmbed	Conv1D
Lens Config ▶	Iter-CS-Attn	S-Attn	Iter-CS-Attn	S-Attn	Iter-CS-Attn
	$N=4, M=1$	$N=4$	$N=2, M=3$	$N=4$	$N=1, M=1$
	✓tie weights	CLIP-ViT `Block.1-4` Init	-	CLIP-ViT `Block.1-4` Init	-



Approach

Leveraging pre-trained large-scale model for low-resource data modalities

3. Unified Modality Encoder:
   - Frozen pre-trained ViT model
   - Embeds each modality to a unified feature space



	3D Point Cloud	Depth	Audio	Tactile	EEG
Pretrained ViT Config ▶	CLIP-ViT Block.1-12	CLIP-ViT Block.5-12	CLIP-ViT Block.1-12	CLIP-ViT Block.5-12	CLIP-ViT Block.1-12

Approach

Leveraging pre-trained large-scale model for low-resource data modalities

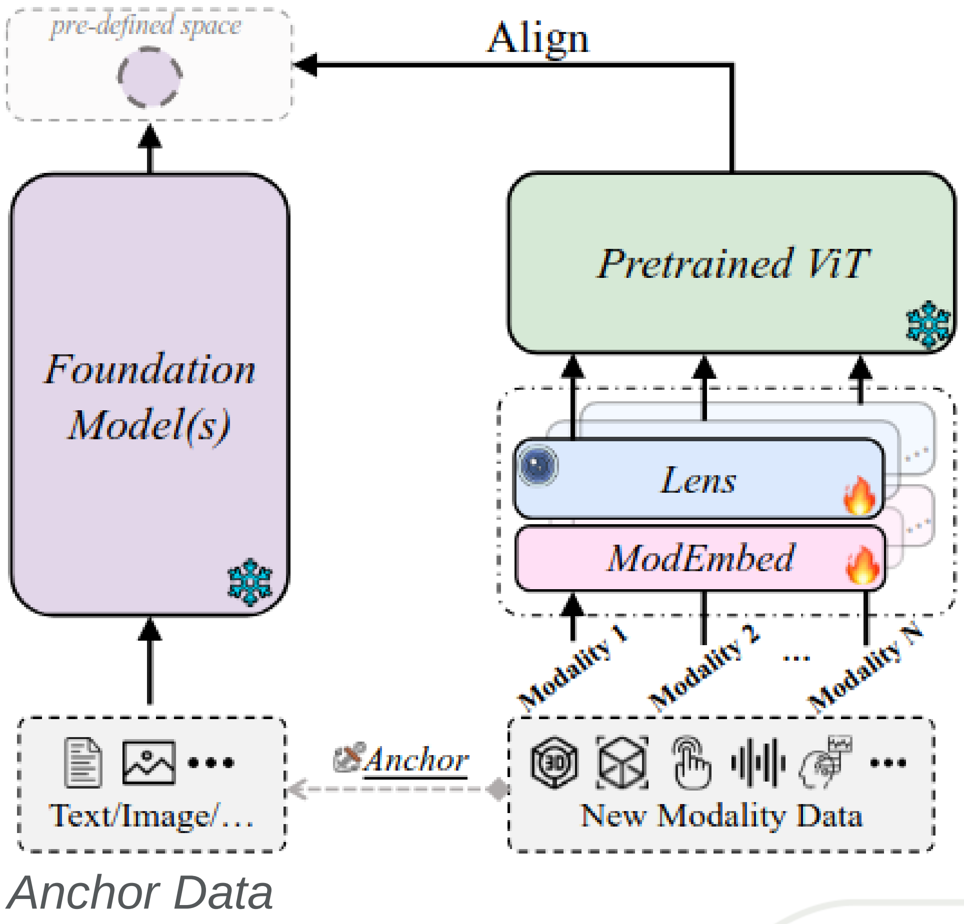
4. Alignment with Anchor Modality:

- Maximizing the similarity between the representations of the new modality and its corresponding anchor modality while minimizing the similarity with unrelated anchor data

*E.g.: **EEG signal paired with a text description**, and ViT-Lens learns to bring the EEG embedding close to the text embedding*

Training Objective:

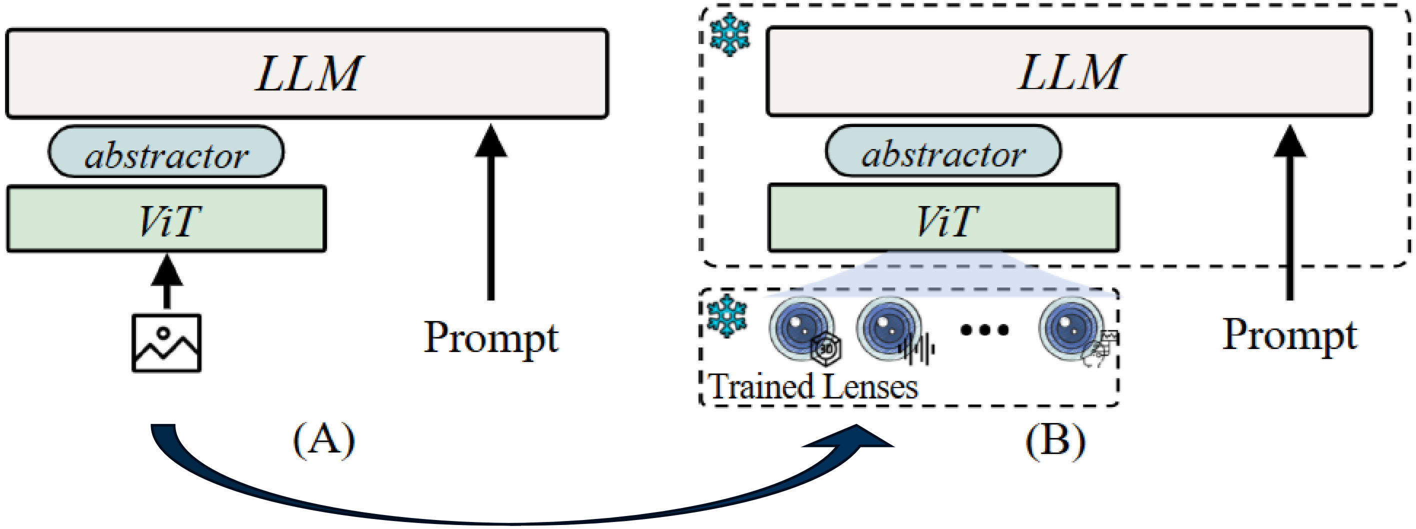
$$\mathcal{L} = -\frac{1}{2B|\mathcal{A}|}\sum_{i=1}^{B}\sum_{k=1}^{|\mathcal{A}|}\left(\log\frac{\exp(h_i^X \cdot h_i^{A_k}/\tau)}{\sum_j \exp(h_i^X \cdot h_j^{A_k}/\tau)} + \log\frac{\exp(h_i^{A_k} \cdot h_i^X/\tau)}{\sum_j \exp(h_i^{A_k} \cdot h_j^X/\tau)}\right),$$



Georgia Tech

Integrating ViT-lens into other MFMs

ViT-lens module can be plugged in place of the image-encoder only foundation model **without further instruct-following training.**

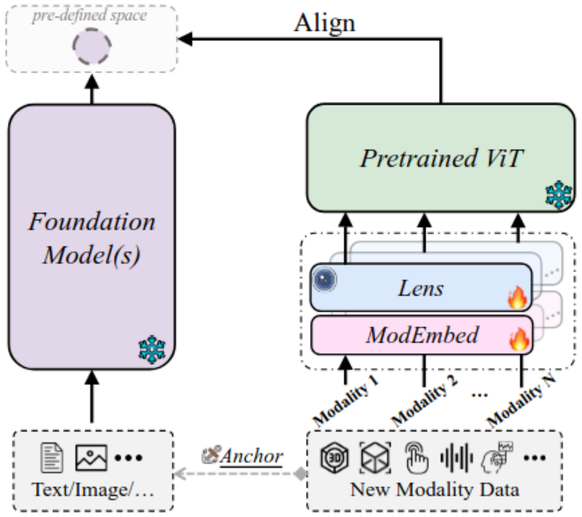


(A) A multimodal foundation model with image encoder

(B) Well trained Lenses for different modalities

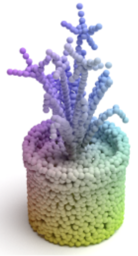
MFM integration Steps

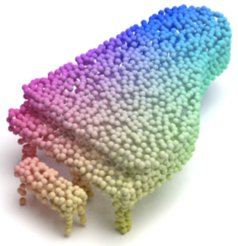
1. Select an MFM (e.g., InstructBLIP or SEED)
2. Take the trained Encoder (CLIP-ViT) as modality encoder for ViT-lens
3. Tune the *ModEmbed* and *lens* parameter (fine-tuning step) for alignment
4. Take the tuned *ModEmbed* and *lens* module add it directly to the pretrained MFM (inference step)



Handle inputs of various modality without specific instruction following

MFM integration with a LLM

Visual input example, Plant:	
	
User	What is this object?
InstructBLIP w/ ViT-Lens	Potted plant.
User	Describe it in detail.
InstructBLIP w/ ViT-Lens	In the image, there is a potted plant sitting in a ceramic pot on a table. The plant appears to be a bamboo-like species, and it is positioned in such a way that its leaves appear to be growing out of the top of the pot. The plant's placement on the table suggests that it is meant to be a decorative element, adding a touch of greenery to the space. The ceramic pot has a diameter of around 10 centimeters, making it a suitable size for holding a small plant like this one. Overall, the image showcases a simple yet elegant arrangement of a potted plant in a ceramic pot on a table.

Visual input example, Piano:	
	
User	What is this object?
InstructBLIP w/ ViT-Lens	Piano.
User	What is the function of this?
InstructBLIP w/ ViT-Lens	Pianos are used to play music.
User	Describe it in detail.
InstructBLIP w/ ViT-Lens	The object in the image is an upright piano, which is a type of keyboard instrument known for its distinct sound and versatility. The piano is placed on a wooden surface, possibly a floor or a stage, and it appears to be quite large in size. Additionally, there is a chair placed next to the piano, suggesting that it may be used for seating during performances or lessons. The piano is positioned in the center of the image, making it the focal point of the scene. Overall, the combination of the piano and the chair creates an inviting atmosphere for music enthusiasts or those interested in learning how to play the piano.

Enabling a LLM to handle 3D points

Experiments: Datasets used for Evaluation

Dataset	Task	#cls	Metric	#test
ModelNet40(MN40) [90]	3D shape cls	40	Acc	2,468
Objaverse-LVIS(O-LVIS) [17]	3D shape cls	1,156	Acc	46,832
ScanObjectNN(SONN) [85]	3D shape cls	15	Acc	581
SUN Depth-only(SUN-D) [78]	Scene cls	19	Acc	4,660
NYU-v2 Depth-only(NYU-D) [61]	Scene cls	10	Acc	654
Audioset Audio-only(AS-A) [29]	Audio cls	527	mAP	17,132[1]
ESC 5-folds(ESC) [70]	Audio cls	50	Acc	2,000
Clotho(Clotho) [22]	Retrieval	-	Recall	1,046
AudioCaps(ACaps) [44]	Retrieval	-	Recall	813[1]
VGGSound(VGGS) [9]	Audio cls	309	Acc	15,434[1]
Touch-and-go(TAG-M) [94]	Material cls	20	Acc	29,879
Touch-and-go(TAG-H/S) [94]	Hard/Soft cls	2	Acc	29,879
Touch-and-go(TAG-R/S) [94]	Rough/Smooth cls	2	Acc	8,085
ImageNet-EEG(IN-EEG) [79]	Visual Concept cls	40	Acc	1,997

Datasets include:
- 3D point cloud
- Depth
- Audio
- Tactile
- EEG

Datasets (More Details)

3D point cloud

- ULIP-ShapeNet Triplets
  - Anchor = image+text
- ULIP2-Objaberse Triplets
- OpenShape Triplets
- ModelNet40
  - Anchor = text
- ScanObjectNN
- Objaverse-LVIS

RGBD (Depth)

- Anchor = image
- SUN-RGBD
- NYU-Depth v2

Audio

- Audioset
  - Anchor = text
- ESC 5-folds
- Clotho
- AudioCaps
- VGGSound

Tactile and EEG

- touch-and-go
  - Anchor = image
- ImageNet-EEG
  - Anchor = image and/or text

Experiments: Zero-shot 3D Classification

- Experimental Setup
  - Authors use triplets to train ViT-Lens
    - (point cloud, image, text)
- 2 Setups
  - First setup: **pretrain on ULIP-ShapeNet or ULIP2-Objaverse**
    - Evaluate on ModelNet40
  - Second setup: **train on OpenShape-Triplets**
    - Evaluate on Objaverse-LVIS, ModelNet40, ScanObjectNN

Experiments: Zero-shot 3D Classification

	Top1	Top5
Trained on ULIP-ShapeNet [92]		
ULIP-PointNet++(ssg) [92]	55.7	75.7
ULIP-PointNet++(msg) [92]	58.4	78.2
ULIP-PointMLP [92]	61.5	80.7
ULIP-PointBERT [92]	60.4	84.0
ViT-Lens$_B$	65.4	92.7
ViT-Lens$_L$	**70.6**	**94.4**
Trained on ULIP2-Objaverse [93]		
ULIP2-PointNeXt [93]	49.0	79.7
ULIP2-PointBERT [93]	70.2	87.0
ViT-Lens$_B$	74.8	93.8
ViT-Lens$_L$	**80.6**	**95.8**

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

ViT-B/16

ViT-L/14

Setup 1

	Objaverse-LVIS			ModelNet40			ScanObjectNN		
	Top1	Top3	Top5	Top1	Top3	Top5	Top1	Top3	Top5
2D inference, no 3D training									
PointCLIP [101]	1.9	4.1	5.8	19.3	28.6	34.8	10.5	20.8	30.6
PointCLIP v2 [106]	4.7	9.5	12.9	63.6	77.9	85.0	42.1	63.3	74.5
Trained on OpenShape-Triplets (No LVIS) [54]									
ULIP-PointBERT [92]	21.4	38.1	46.0	71.4	84.4	89.2	46.0	66.1	76.4
OpenShape-SparseConv [54]	37.0	58.4	66.9	82.6	95.0	97.5	54.9	76.8	87.0
OpenShape-PointBERT [54]	39.1	60.8	68.9	85.3	96.2	97.4	47.2	72.4	84.7
ViT-Lens$_G$	**50.1**	**71.3**	**78.1**	**86.8**	**96.8**	**97.8**	**59.8**	**79.3**	**87.7**
Trained on OpenShape-Triplets [54]									
ULIP-PointBERT [92]	26.8	44.8	52.6	75.1	88.1	93.2	51.6	72.5	82.3
OpenShape-SparseConv [54]	43.4	64.8	72.4	83.4	95.6	97.8	56.7	78.9	88.6
OpenShape-PointBERT [54]	46.8	69.1	77.0	84.4	96.5	98.0	52.2	79.7	88.7
ViT-Lens$_G$	**52.0**	**73.3**	**79.9**	**87.6**	**96.6**	**98.4**	**60.1**	**81.0**	**90.3**

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

Setup 2

ViT-bigG/14

Experiments: Audio Classification and Retrieval

	anchor	AudioSet mAP	VGGSound$^\diamond$ Top1	ESC$^\diamond$ Top1	Clotho$^\diamond$ R@1	Clotho$^\diamond$ R@10	AudioCaps$^\diamond$ R@1	AudioCaps$^\diamond$ R@10
AVFIC [60]	-	-	-	-	3.0	17.5	8.7	37.7
ImageBind-H [32]	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 [38]	I+T	25.9	-	69.4	-	-	-	-
ViT-Lens$_L$	I+T	**26.7**	**31.7**	**75.9**	**8.1**	**31.2**	**14.4**	**54.9**
Prev. ZS SOTA	-	-	29.1/46.2* [89]	91.8 [87]	6.0	28.4 [32]	9.3	42.3 [32]

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

- Pretrained on Audioset dataset (accompanied by image+text as anchor)
- ViT-Lens, when anchored using image and text, has strong performance compared to other baselines

Experiments: Audio and Video Retrieval

- MSR-VTT benchmark

- Utilizes both audio and video modalities
  - They follow ImageBind's method to combine audio and video modalities

	modality	R@1	R@5	R@10
MIL-NCE [57]	V	8.6	16.9	25.8
SupportSet [67]	V	10.4	22.2	30.0
AVFIC [60]	A+V	19.4	39.5	50.3
ImageBind-H [32]	A+V	36.8	61.8	70.0
ViT-Lens_L	A+V	**37.6**	**63.2**	**72.6**
Zero-shot SOTA [10]	V	49.3	68.3	73.9

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

Experiments: Depth-only Scene Classification

- Two benchmarks
  - NYU-D
  - SUN-D
- Pretraining data from Sun RGD-D dataset (paired image and scene labels as anchor)
- Does pretty well compared to the supervised setting!

	anchor	NYU-D	SUN-D
Text Paired [32]	T*	41.9	25.4
ImageBind-H [32]	I	54.0	35.1
ViT-Lens$_L$	I	64.2	37.4
ViT-Lens$_L$	I+T	**68.5**	**52.2**
Supervised SOTA [31]	-	76.7	64.9

Table 5. Depth-only scene classification on NYU-D and SUN-D. *[32] rendered depth as grayscale images for direct testing. The supervised SOTA [31] used RGBD as input and extra training data.

Georgia Tech

Experiments: Tactile Classification Tasks

- Tactile classification tasks
  - Material
  - Hard/soft
  - Rough/smooth
- Train using Touch-and-go train-material split (anchor is paired frame and material label text)
  - test H/S and test R/S are zero-shot classification results
- Linear probing
  - Model is fine-tuned using corresponding train set for a given task

	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	**78.2**
ViT-Lens$_L$	I+T	**65.8**	**74.7**	63.8
Linear Probing				
CMC [82, 94]	I	54.7	77.3	79.4
ViT-Lens$_B$	I	**63.0**	**92.0**	**85.1**

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

Experiments: EEG Visual Concept Classification

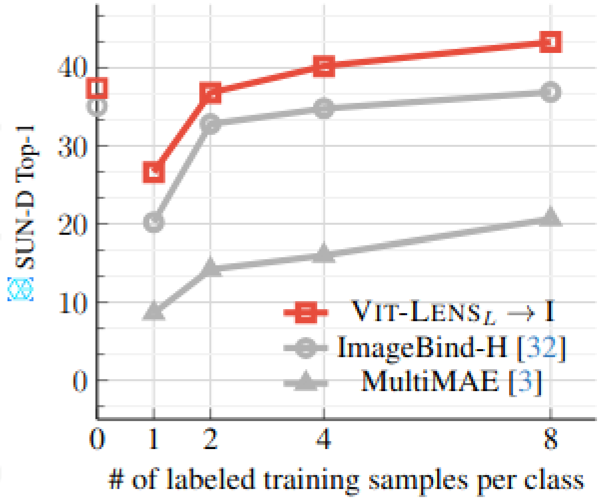
- Trained on ImageNet-EEG (anchor is corresponding ImageNet image and text label)
- Image and text anchor once again provides best performance

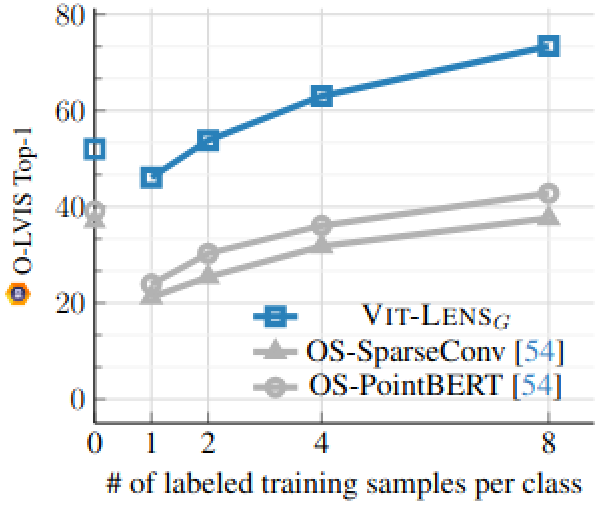
	anchor	Val	Test
ImageBind-B*	I	17.3	18.4
DreamDiffusion-L# [4]	I	20.4	19.2
VIT-LENS$_B$	I	24.6	25.3
VIT-LENS$_L$	I	29.3	29.2
VIT-LENS$_L$	I+T	**41.8**	**42.7**

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.

Georgia Tech

Experiments: Few-shot Linear Probing





- Left: Using ViT-Lens L image depth encoder
- Right: ViT-Lens G 3D encoder
- Good few-shot capabilities

Experiments: Integration with MFMs

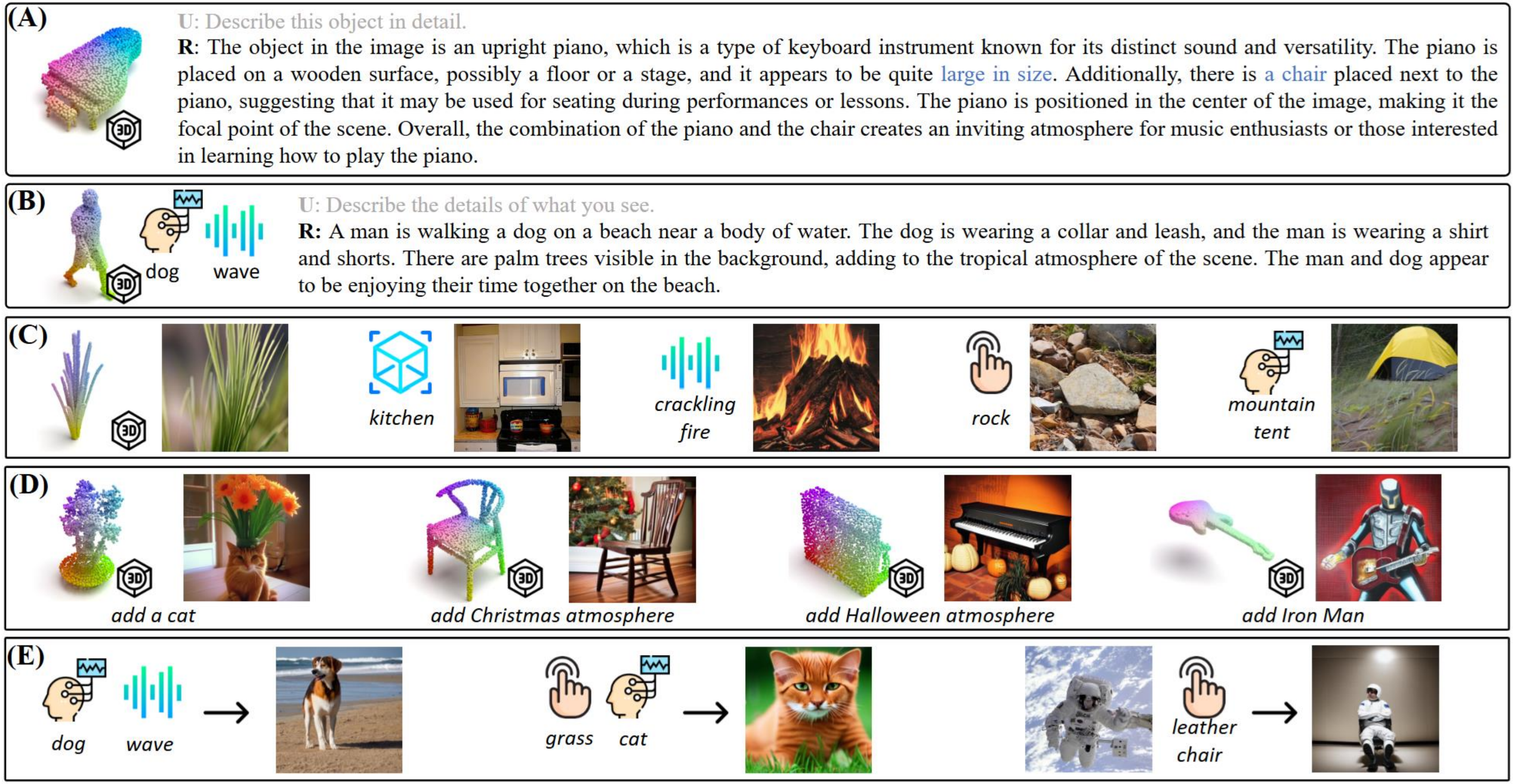
- 2 MFMs selected
  - Instruct-BLIP and SEED
    - EVA01-g14 CLIP-ViT as visual encoder

- Instruct-BLIP
  - Framework for instruction-tuning
    - Complex visual reasoning and image descriptions

- SEED
  - Multimodal comprehension and image generation

- *Can these capabilities of MFMs be extended to novel modalities by integrating ViT-Lens?*

Qualitative examples for plugging VIT-LENS into MFMs



(A-B) Integrate with InstructBLIP

(C-E)Integrate with SEED

Georgia Tech

Qualitative examples for plugging VIT-LENS into MFMs

Ablation Study: Scaling ViT-Lens

- Performed experiments on scaling up ViT-Lens
- Tested on SUN-D dataset (depth) and ModelNet40 (point cloud)
- In conclusion, ViT-Lens is amenable to this scaled up scenario

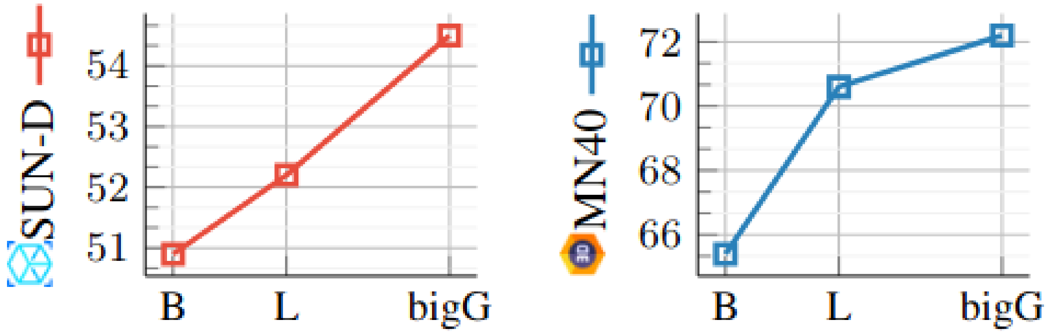
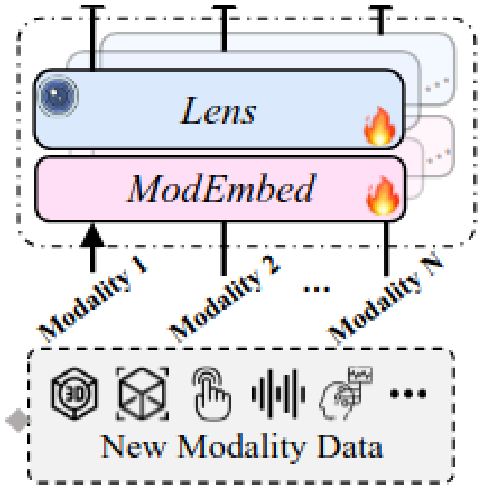


Figure 7. **Scaling the ViT-Lens on depth and 3D point cloud.** B: ViT-Lens$_B$, L: ViT-Lens$_L$, bigG: ViT-Lens$_G$.

Limitations

- Error and Bias Propagation:
  - The pre-trained encoder (ViT here) explicitly passes its biases to this multimodal foundation model
- Assumption that the ViT is a general token learner is suboptimal!
  - ViTs are designed for image data, which has specific spatial and structural properties (like 2D grid patterns)
  - Not every modality shares similar structure and properties!
- Generalization of this model is still bound by the low-resource data modalities.
- Convergence to larger-scale modalities during training for lenses and modality embedders.



Comparison to Other Methods

- ImageBind
  - Trains joint embedding, showing that only image-paired data is enough to get this embedding
  - Separate encoders for modalities

- Unified-IO 2
  - Processes all modalities with single encoder-decoder transformer
  - Combine modality tokens dynamic packing

- Emphasis on this shared embedding space
- More similar to Unified-IO 2 in how modality data is processed

Summary of Strengths, Weaknesses

- Strengths:
  - Computation cost is bound by 1 encoder
    - Compared with multiple encoder or mixture of expert methods
  - Modular and Adaptable
    - Modality embeddings can be replaced
  - Integration with other multimodal foundation model
- Weaknesses:
  - Suboptimal unified encoder
  - Bias propagation by using an image-trained model
  - Scaling the model is bound by low-resource modality