

IMAGEBIND: One Embedding Space To Bind Them All

Rohit Girdhar* Alaaeldin El-Nouby* Zhuang Liu Mannat Singh Kalyan Vasudev Alwala Armand Joulin Ishan Misra* FAIR, Meta AI
CVPR 2023

Published at: IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2023.

Presented by Yilun Zhou, Zhenyang Chen, Yunhai Han

Outline

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

Align data across all senses

Image -> Audio

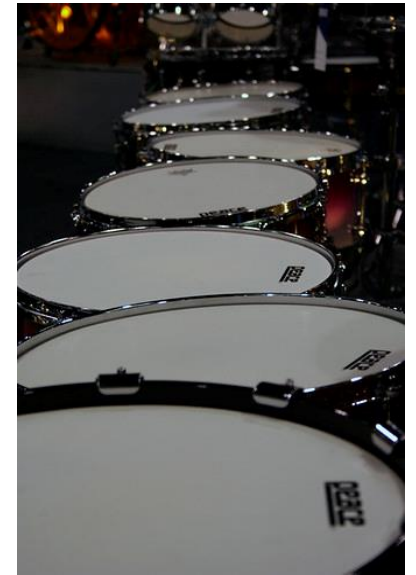


Audio -> Image



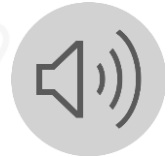
Text -> Audio & Image

"Drums"



Align data across all senses

Audio & Image -> Image



Audio -> Image (**Generation** [1])



3) Audio to Image Generation



But no paired data across all modalities

Text

Image

Sound

Depth

.....

"A dog"

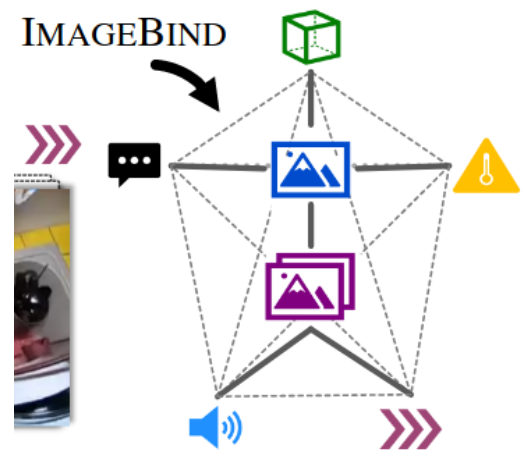
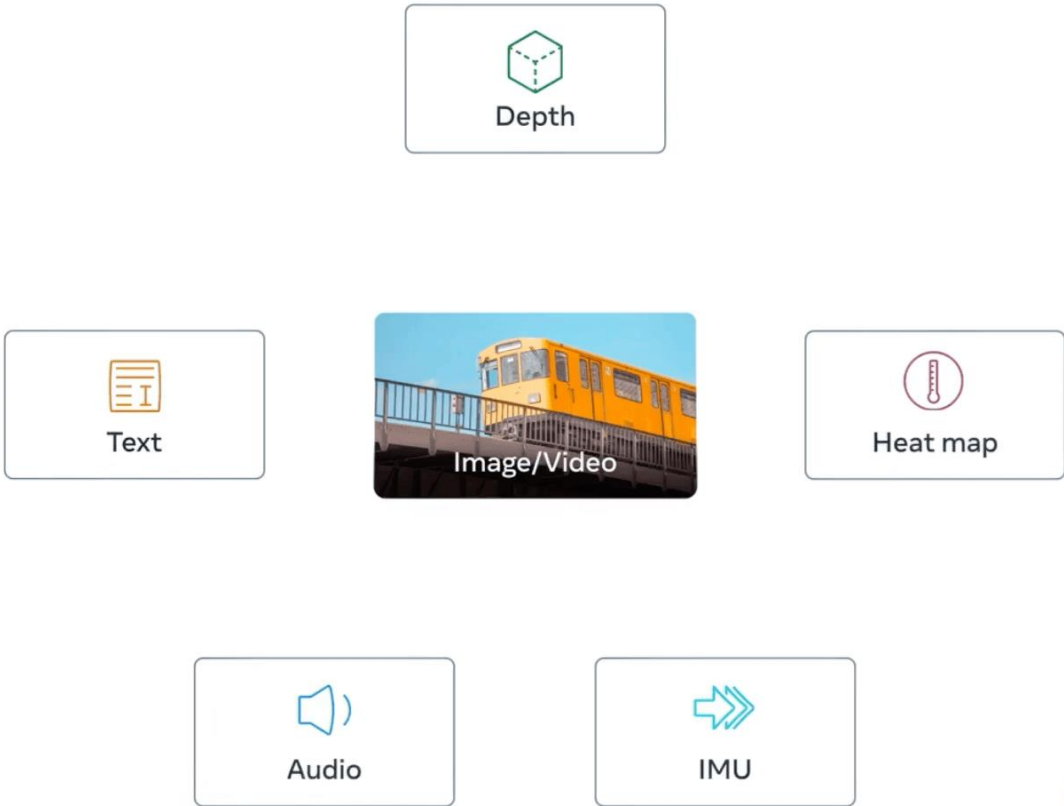


?

?

?

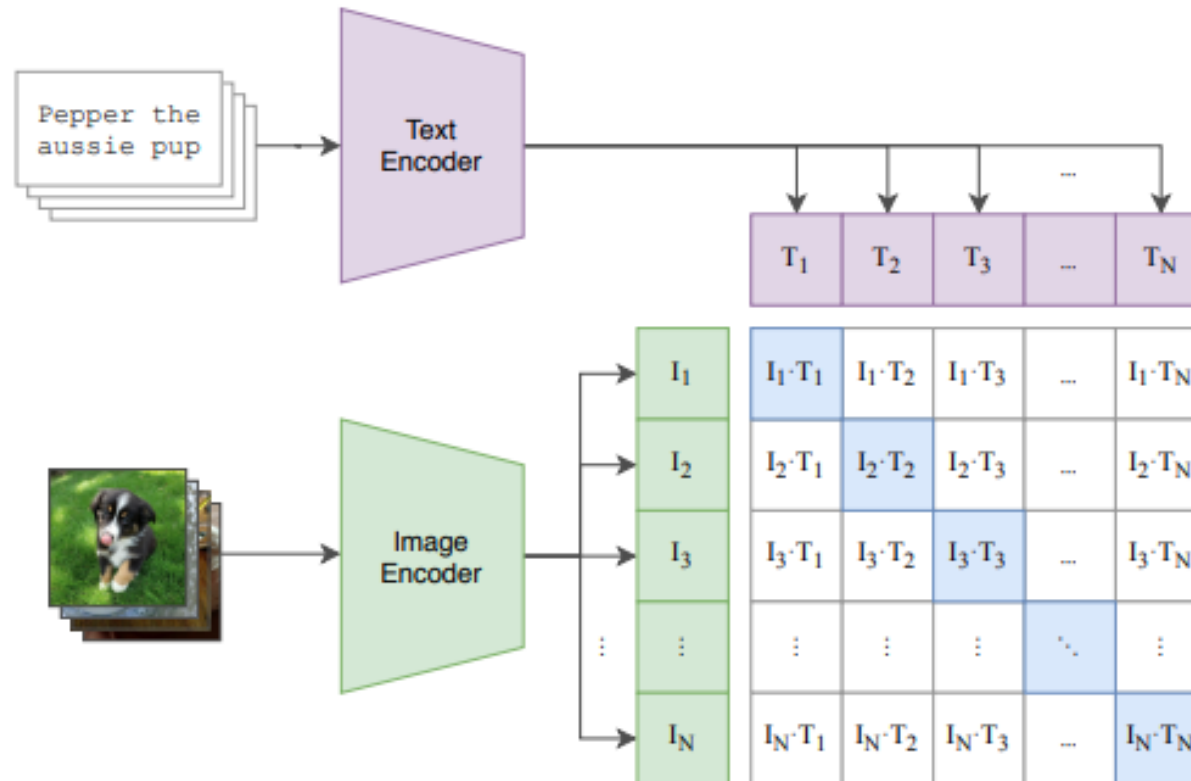
Link all modality data via images



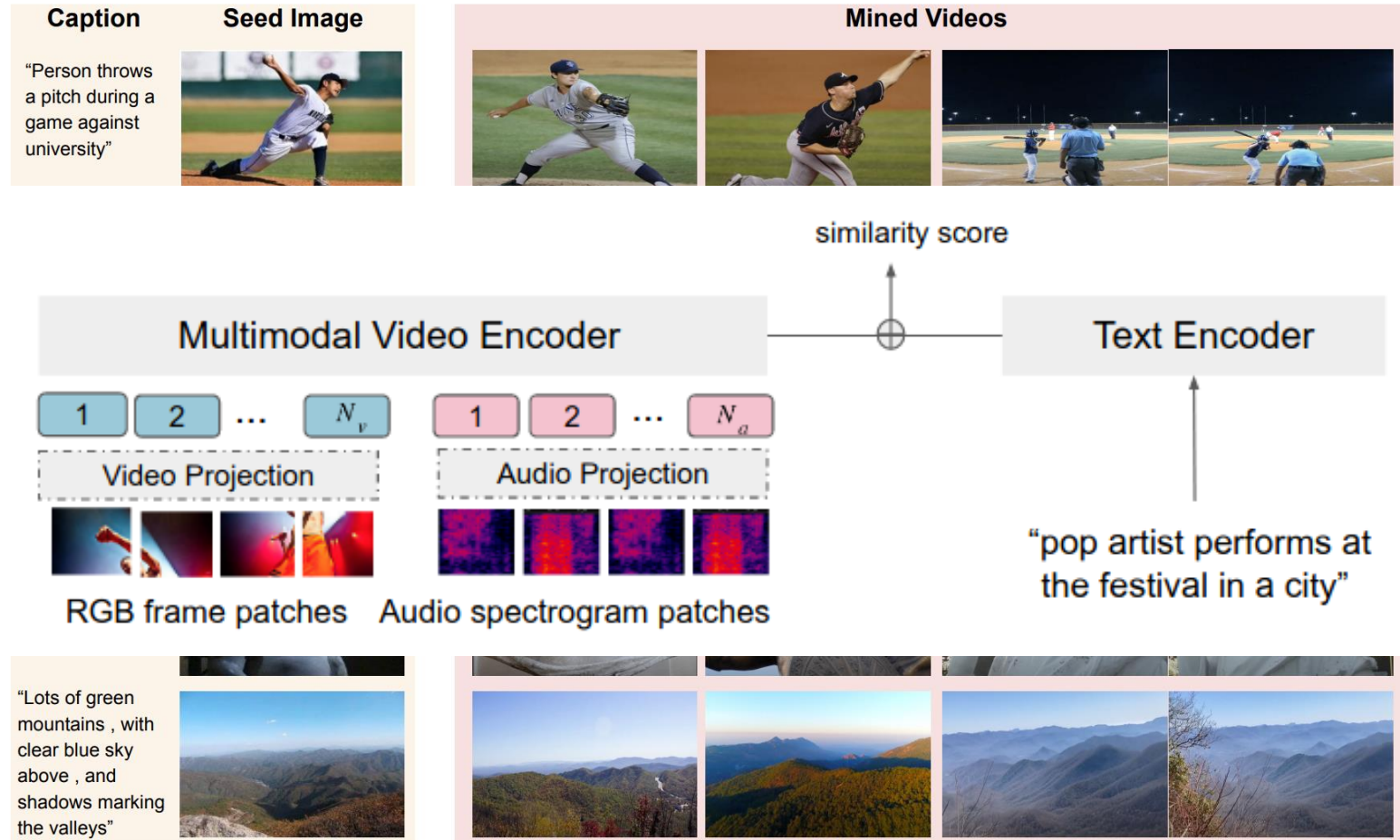
Meta AI

Related Works: CLIP [1] - Image & Text

(1) Contrastive pre-training

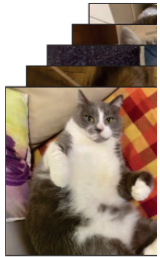


Related Works: Mining ^[2] - Video from seed Images



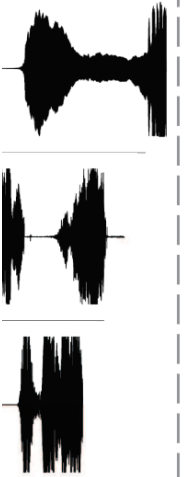
Related Works: AudioCLIP [3] - Audio & Image & Text

“cat”

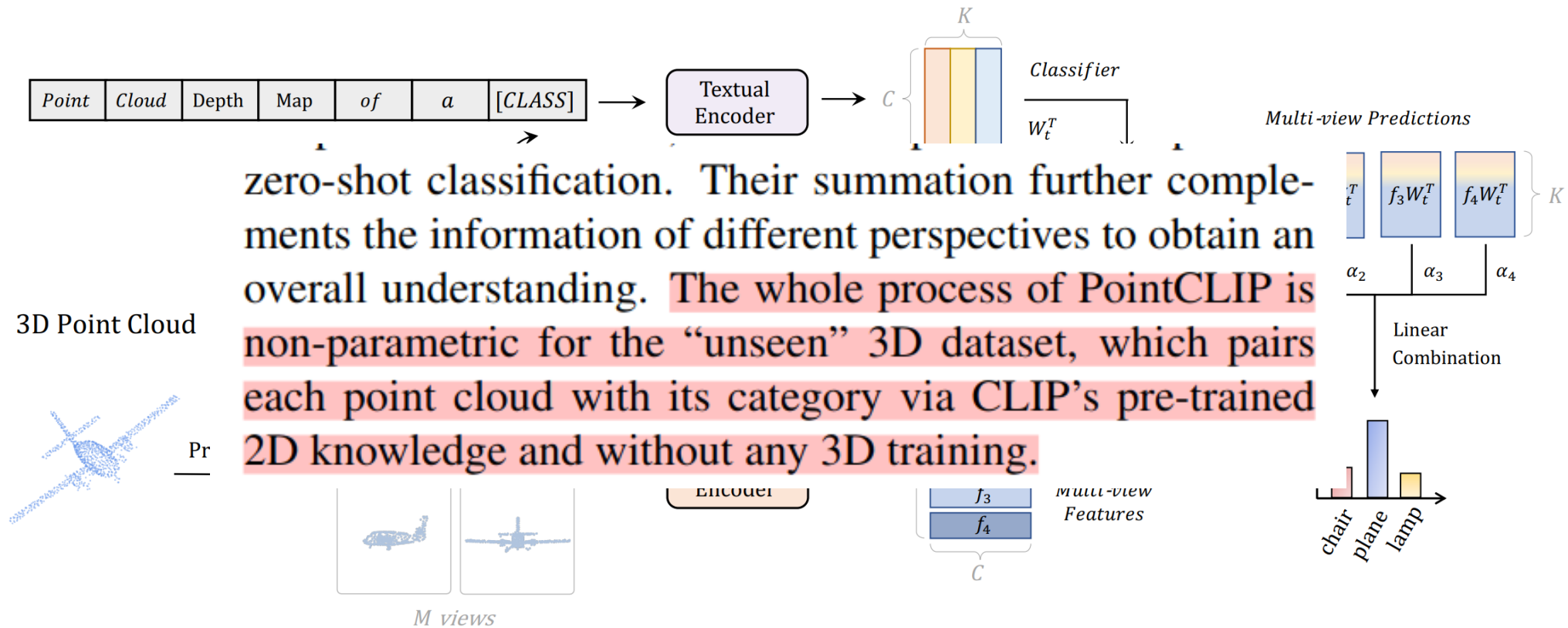


AudioSet: Being proposed in [7], the AudioSet dataset provides a large-scale collection (~ 1.8 M & ~ 20 k evaluation set) of audible data organized into 527 classes in a non-exclusive way. Each sample is a snippet up to 10 s long from a YouTube-video, defined by the corresponding ID and timings.

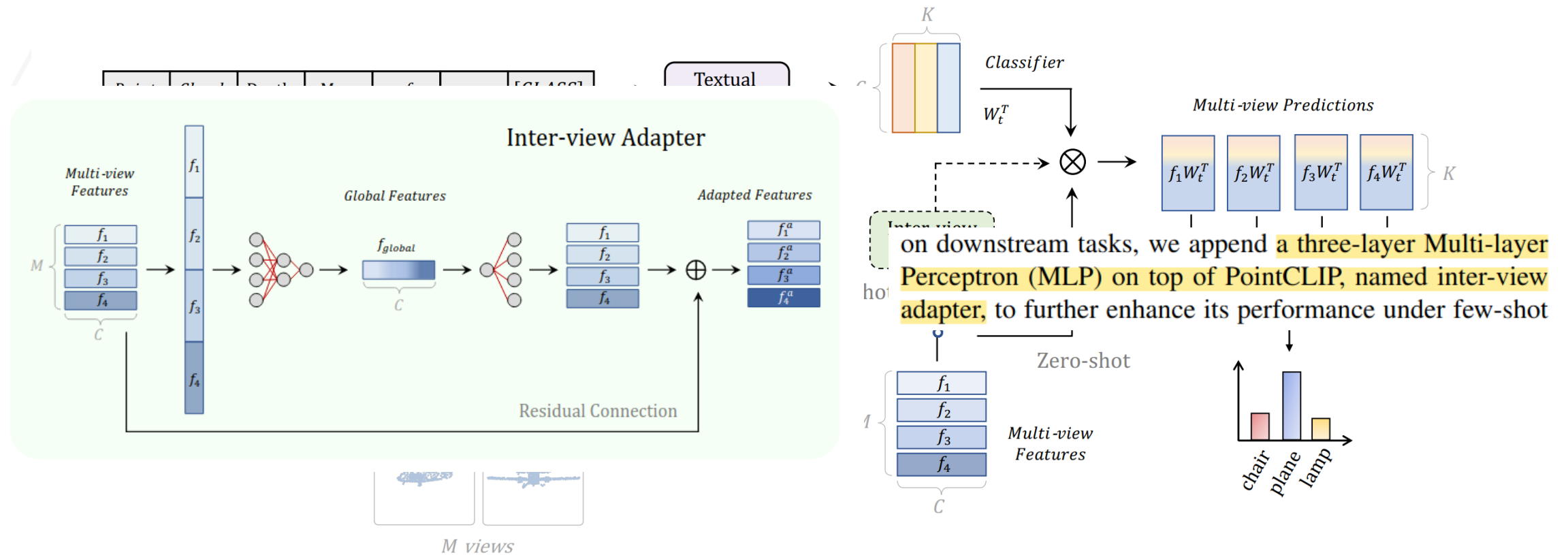
For this work, we acquired video frames in addition to audio tracks. Thus, the AudioSet dataset became the glue between the vanilla CLIP framework and our tri-modal extension on top of it. In particular, audio tracks and the respective class labels were used to perform image-to-audio transfer learning for the ESResNeXt model, and then, the extracted frames in addition to audio and class names served as an input for the hybrid AudioCLIP model.



Related Works: PointCLIP [4] - PL & Image & Text



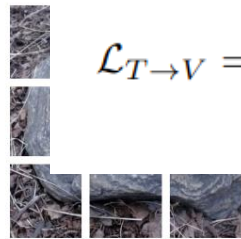
Related Works: PointCLIP [4] - PL & Image & Text



Related Works: Binding Touch [5] - Tactile & Image & Text

We denote Ω_v as the visual image domain and Ω_t as the tactile image domain. Thus, given B visual and touch pairs in a batch, $\{(\mathbf{v}_i, \mathbf{t}_i)\}_{i=1}^B$, where $\mathbf{v}_i : \Omega_v \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ and $\mathbf{t}_i : \Omega_t \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$, we align a tactile embedding $\mathcal{F}_T(\mathbf{t}_i) \in \mathbb{R}^C$ with the pretrained visual embedding $\mathcal{F}_V(\mathbf{v}_i) \in \mathbb{R}^C$ from [35] by maximizing the cosine similarity between corresponding visuo-tactile pairs. We optimize this objective using InfoNCE loss [81] to match touches to correct images:

$$\mathcal{L}_{T \rightarrow V} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\mathcal{F}_T(\mathbf{t}_i) \cdot \mathcal{F}_V(\mathbf{v}_i) / \tau)}{\sum_{j=1}^B \exp(\mathcal{F}_T(\mathbf{t}_i) \cdot \mathcal{F}_V(\mathbf{v}_j) / \tau)}, \quad (1)$$



Image

Related Works: Summary

To scale up the modality, there were two typical methods:

1. Separate encoder for each modality + paired data for all modalities
2. Only visual-text encoders + project other modality data to visual data

While in this work, we demonstrate its ***emergent*** zero-shot generalization across various modalities.

Approach

- The goal is to learn a single joint embedding space for all modalities by **using images** to bind them together.
- I.e. Image as the linkage

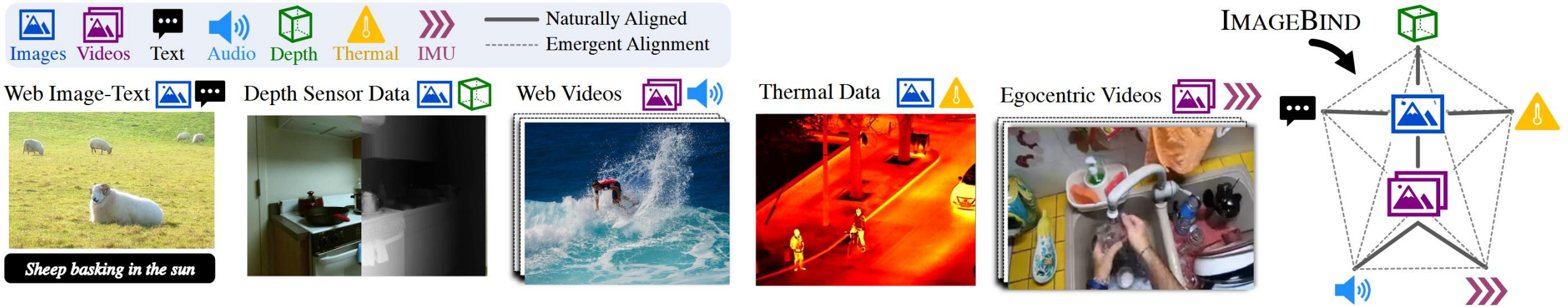


Figure 2. IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc.* IMAGEBIND links all these modalities in a common embedding space, enabling new emergent alignments and capabilities.

Approach

- Modality encoder

$$\mathbf{q}_i = f(\mathbf{I}_i) \text{ and } \mathbf{k}_i = g(\mathbf{M}_i)$$

- Where I represents images, M is another modality
- Dataset: (image, text) pairs and (image, modality) pairs

Approach

- Modality encoder

$$\mathbf{q}_i = f(\mathbf{I}_i) \text{ and } \mathbf{k}_i = g(\mathbf{M}_i)$$

- Loss design

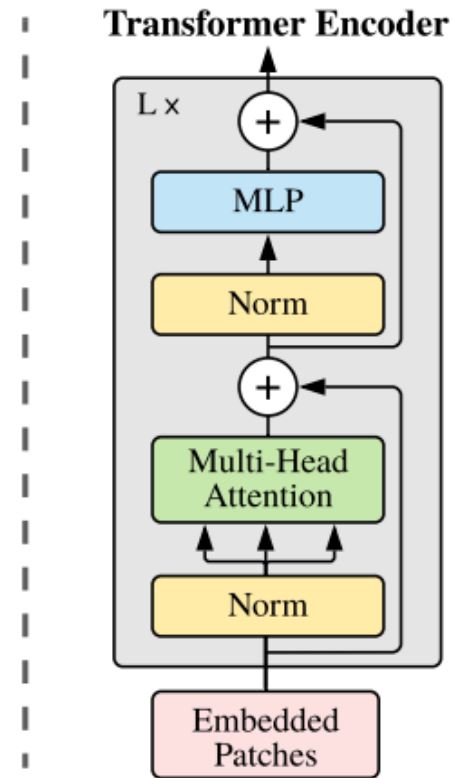
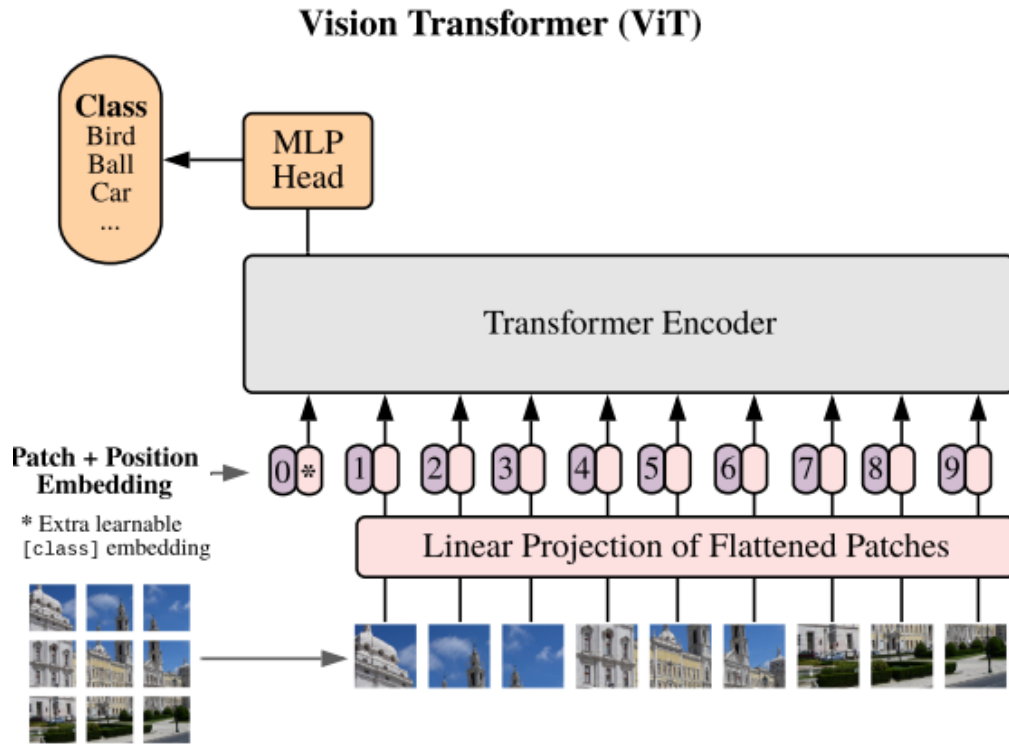
$$L_{\mathcal{I}, \mathcal{M}} = -\log \frac{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau)}{\exp(\mathbf{q}_i^\top \mathbf{k}_i / \tau) + \sum_{j \neq i} \exp(\mathbf{q}_i^\top \mathbf{k}_j / \tau)},$$

- InfoNCE loss to compare embeddings
- Tau as temperature to control concentration
- Symmetric loss $L_{\mathcal{I}, \mathcal{M}} + L_{\mathcal{M}, \mathcal{I}}$.
- j denotes negative examples

Approach

- Image

- Depth and thermal image are treated in the same way, as 1-D image
 - Convert depth into disparity maps for scale invariance.



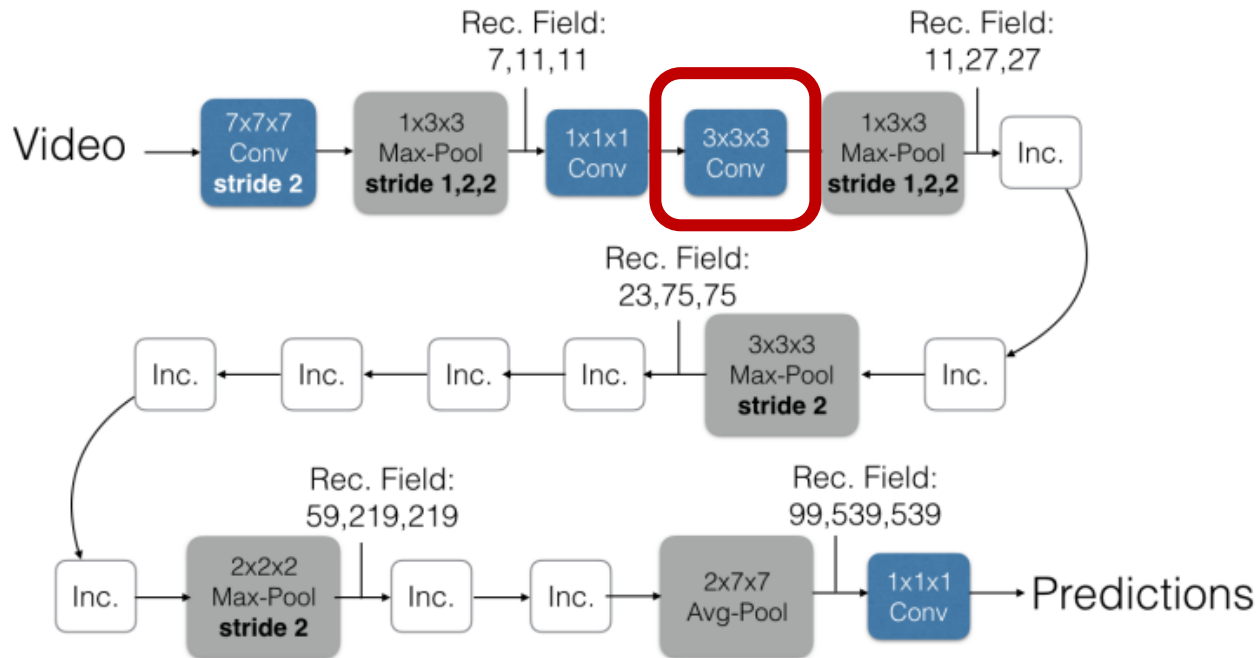
- 17 OmniMAE: Single Model Masked Pretraining on Images and Videos.
- An image is worth 16x16 words: Transformers for image recognition at scale

Approach

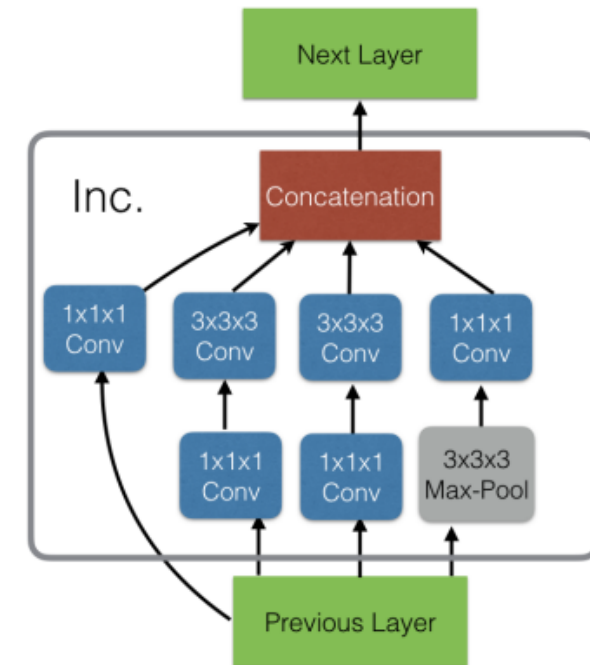
- Videos

- How to address the temporal relations?
 - Use 2 frame video clips sampled from 2 seconds
 - Inflate the patch projection layer (using same encoder as image)

Inflated Inception-V1



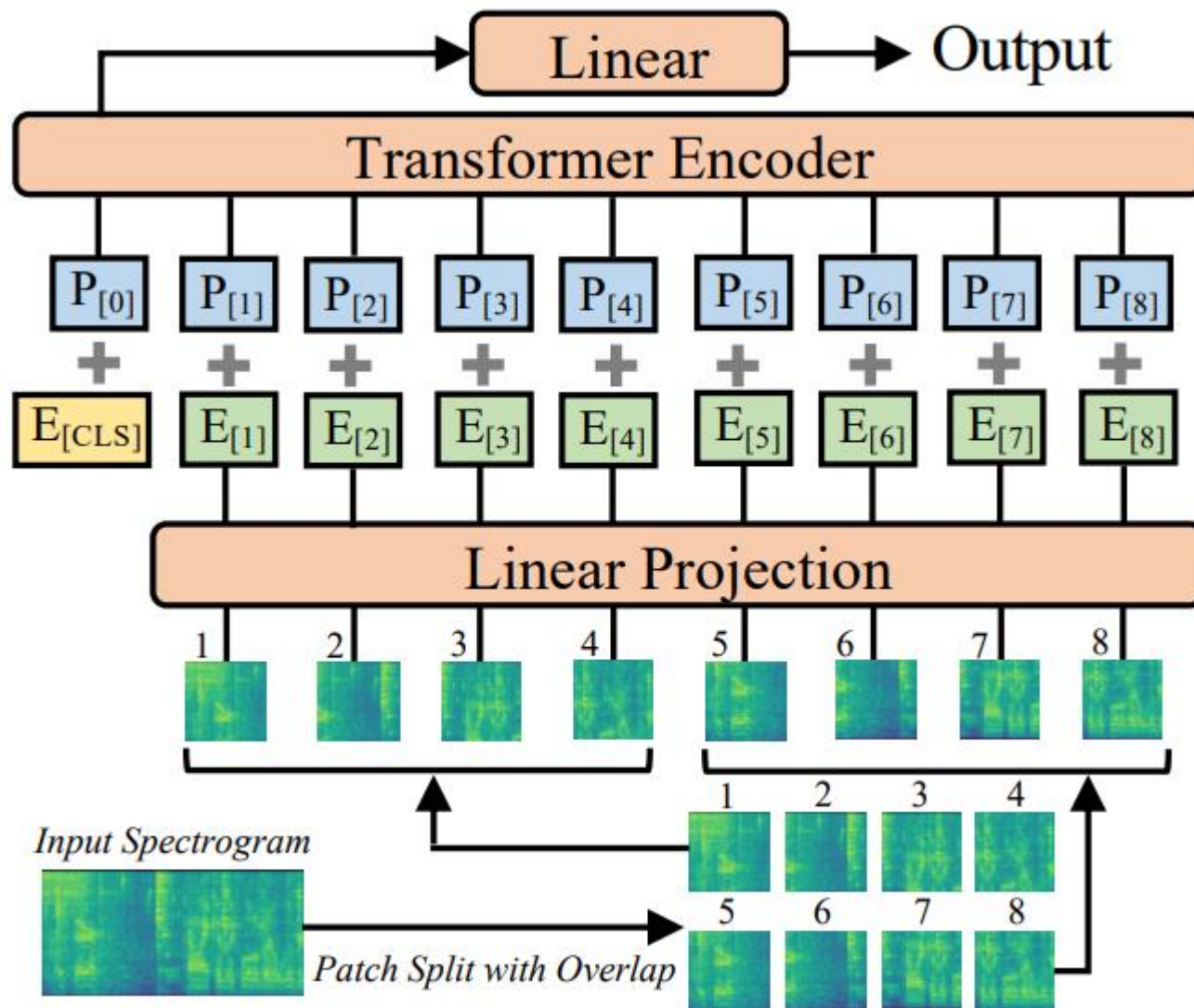
Inception Module (Inc.)



Approach

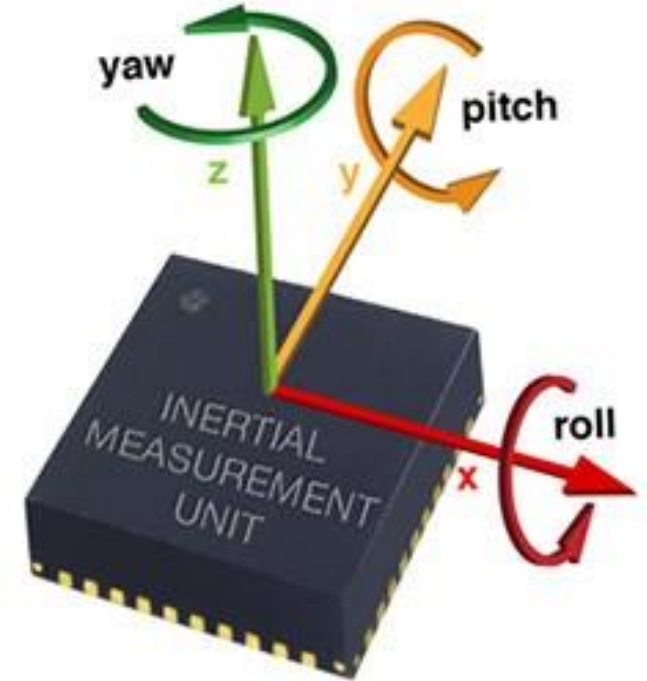
- Audio

- Encoding audio and convert a **2 second audio** sampled at **16kHz** into spectrograms using **128 mel-spectrogram bins**.
- Spectrogram:
 - X-axis time
 - Y-axis frequency
- Design choices:
 - Overlap size
 - Patch size
 - ImageNet pretraining



Approach

- IMU inertial measurement unit
 - Usually include accelerometer and gyroscope. Measure linear acceleration and angular rate
 - X, Y , and Z axes. 5 second clips resulting in 2K time step IMU readings which are projected using a 1D convolution with a kernel size of 8
 - All modality is projected by a linear layer



Downstream Tasks

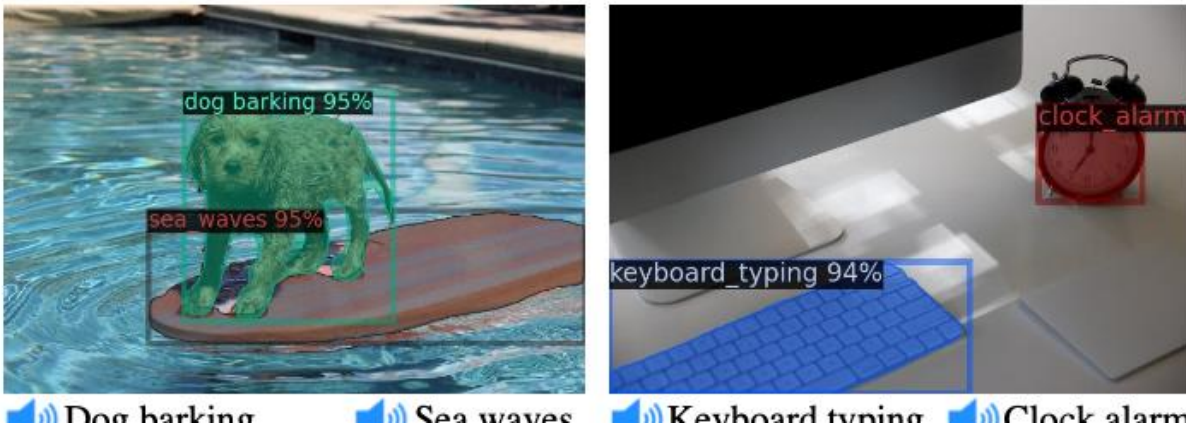
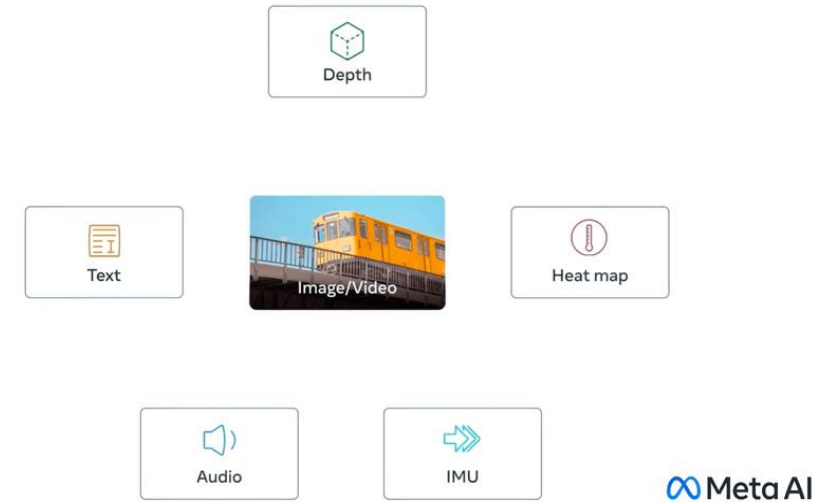
Dataset	Task	#cls	Metric	#test
AudioSet Audio-only (AS-A) [19]	Audio cls.	527	mAP	19048
ESC 5-folds (ESC) [59]	Audio cls.	50	Acc	400
Clotho (Clotho) [17]	Retrieval	-	Recall	1045
AudioCaps (AudioCaps) [37]	Retrieval	-	Recall	796
VGGSound (VGGs) [8]	Audio cls.	309	Acc	14073
SUN Depth-only (SUN-D) [69]	Scene cls.	19	Acc	4660
NYU-v2 Depth-only (NYU-D) [66]	Scene cls.	10	Acc	653
LLVIP (LLVIP) [32]	Person cls.	2	Acc	15809
Ego4D (Ego4D) [23]	Scenario cls.	108	Acc	68865

Table 1. Emergent zero-shot classification datasets for audio, depth, thermal, and Inertial Measurement Unit (IMU) modalities.

- [AudioSet](#)<video, audio, text>: Training & Validation. Contains 2,084,320, **10-second video clips** from **YouTube**, labeled across 632 audio event classes, covering diverse human, animal, musical, and environmental sounds.
- [VGGs](#)<audio, text>: Validation. 200k **10-second video clips**, annotated with 209 sound classes consisting of human actions, sound-emitting objects and human object interactions.
- [Clotho](#)<text, audio>: Features audio clips from Freesound with 2,893 development and 1,045 test clips, each paired with **5 text descriptions**, designed for text-to-audio retrieval tasks.
- [AudioCaps](#)<text, audio>: Training & Validation. Includes over 51,000 YouTube audio clips paired with **human-written captions**, intended for text-to-audio retrieval.
- [SUN-D](#)<image, depth>: Training & Validation Consists of 10,335 depth images across **19 scene classes**, focused on depth-based scene classification.
- [LLVIP](#)<image, thermal>: Contains 15,809 thermal images in low-light conditions. Collected in an outdoor setting using fixed cameras observing street scenes.
- [Ego4D](#) <video, IMU>: Offers 3,000 hours of egocentric video footage, labeled into 108 scenario classes, with multimodal data including IMU readings for scenario classification tasks.

Training Details

- LLVIP: Crop out pedestrian bounding boxes and random background bounding boxes.
 - Person: ['person', 'man', 'woman', 'people']
 - Background ['street', 'road', 'car', 'light', 'tree']
- Ego4D: Scenario classification with 108 unique scenario labels, Activities such as “cooking a meal,” “gardening work outdoors,” “riding a bike,” and “cleaning a room.”




Emergent Zero-Shot vs Zero-Shot (Audio cls.)












- Zero-Shot: Direct paired training – <text, audio>
- Emergent Zero-Shot: No direct paired training - <text, image>, <image, audio>

Training Details

Modality	Data Representation	Encoder	State
Image/video	RGB/2-frame clips	Pretrained ViT-H 630M params	Frozen
Text	-	OpenCLIP 302M params	Frozen
Audio	2D mel-spectograms	ViT-B	Updated
Thermal	1-channel Image	ViT-S	Updated
Depth	1-channel Image	ViT-S	Updated
IMU	6xT tensor		Updated

Zero-shot Classification



	 IN1K	 P365	 K400	 MSR-VTT	 NYU-D	 SUN-D	 AS-A	 VGGs	 ESC	 LLVIP	 Ego4D
Random	0.1	0.27	0.25	0.1	10.0	5.26	0.62	0.32	2.75	50.0	0.9
IMAGEBIND	77.7	45.4	50.0	36.1	54.0	35.1	17.6	27.8	66.9	63.4	25.0
Text Paired	-	-	-	-	41.9*	25.4*	28.4 [†] [27]	-	68.6 [†] [27]	-	-
Absolute SOTA	91.0 [82]	60.7 [67]	89.9 [80]	57.7 [79]	76.7 [21]	64.9 [21]	49.6 [39]	52.5 [36]	97.0 [9]	-	-

- **Random:** Performance without learned associations, showing baseline results with no alignment between modalities.
- **Text Paired:** paired text data for that modality
- **Absolute SOTA:** Uses additional supervision, model ensembles
- Strong performance on non-visual modalities such as audio and IMU
- Overall, IMAGEBIND shows strong emergent zero-shot performance.

Comparison to prior work

	Emergent	Clotho		AudioCaps		ESC
		R@1	R@10	R@1	R@10	Top-1
<i>Uses audio and text supervision</i>						
AudioCLIP [27]	✗	-	-	-	-	68.6
<i>Uses audio and text loss</i>						
AVFIC [51]	✗	3.0	17.5	8.7	37.7	-
<i>No audio and text supervision</i>						
IMAGEBIND	✓	6.0	28.4	9.3	42.3	66.9
<i>Supervised</i>						
AVFIC finetuned [51]	✗	8.4	38.6	-	-	-
ARNLQ [53]	✗	12.6	45.4	24.3	72.1	-

Table 3. Emergent zero-shot audio retrieval and classification.

- IMAGEBIND outperforms AVFIC on audio-text retrieval tasks, achieving **double** the performance on the **Clotho** dataset
- **Matches** performance with AudioCLIP's on **ESC**

	Modality	Emergent	MSR-VTT		
			R@1	R@5	R@10
MIL-NCE [49]	V	✗	8.6	16.9	25.8
SupportSet [57]	V	✗	10.4	22.2	30.0
FIT [5]	V	✗	15.4	33.6	44.1
AVFIC [51]	A+V	✗	19.4	39.5	50.3
IMAGEBIND	A	✓	6.8	18.5	27.2
IMAGEBIND	A+V	✗	36.8	61.8	70.0

Table 4. Zero-shot text based retrieval on MSR-VTT 1K-A.

- IMAGEBIND performs strongly in audio-only retrieval
- Combining audio and video modalities further improves retrieval accuracy.

Few-shot classification

- **Setup:** Evaluated on **ESC (audio)** and **SUN (depth)** datasets, using **k-shot samples** per class where $k = \{1, 2, 4, 8\}$.
- **Training Approach:** Encoder parameters were frozen; a **linear classifier** was trained for few-shot learning.
- **Audio:**
 - **Comparison Model:** AudioMAE (self-supervised and supervised) with **ViT-B audio encoder** for learning audio features.
 - **Performance:** ImageBind outperforms AudioMAE on audio feature learning.
 - **Training Details:** ESC dataset using **AdamW** optimizer (learning rate = 1.6×10^{-3} , weight decay = 0.05) for **50 epochs**.
- **Depth:**
 - **Comparison Model:** MultiMAE with **ViT-B/16** encoder for depth feature learning.
 - **Performance:** ImageBind outperforms MultiMAE on depth features.
 - **Training Details:** SUN dataset using **AdamW** optimizer (learning rate = 10^{-2} , no weight decay) for **60 epochs**.

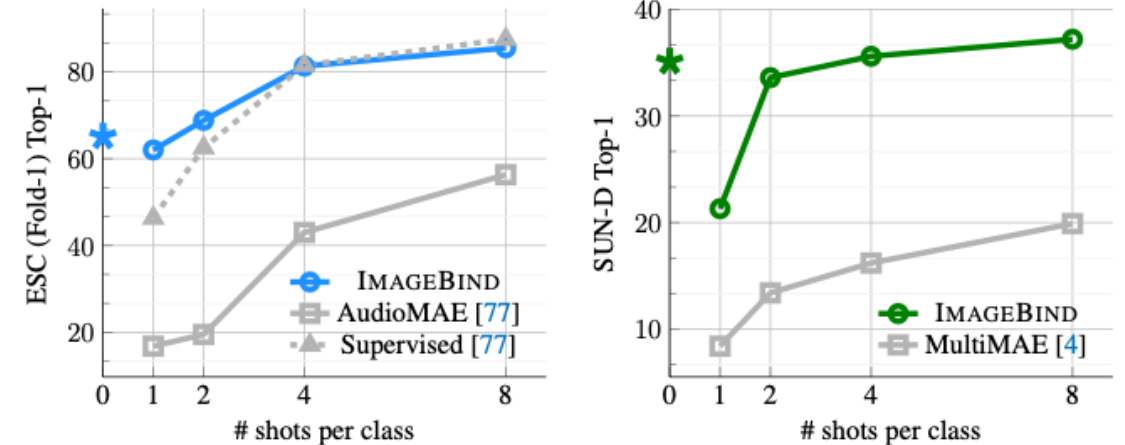


Figure 3. Few-shot classification on audio and depth. We report

Ablation Study – Scaling the Image Encoder

- Experiments with different image encoder sizes (ViT-B, ViT-L, ViT-H) to see the effect on performance.
- Focus on image representation impact, other modality encoders (e.g., depth, audio) are kept at a fixed size.
- **Larger image encoders lead to better emergent zero-shot accuracy**

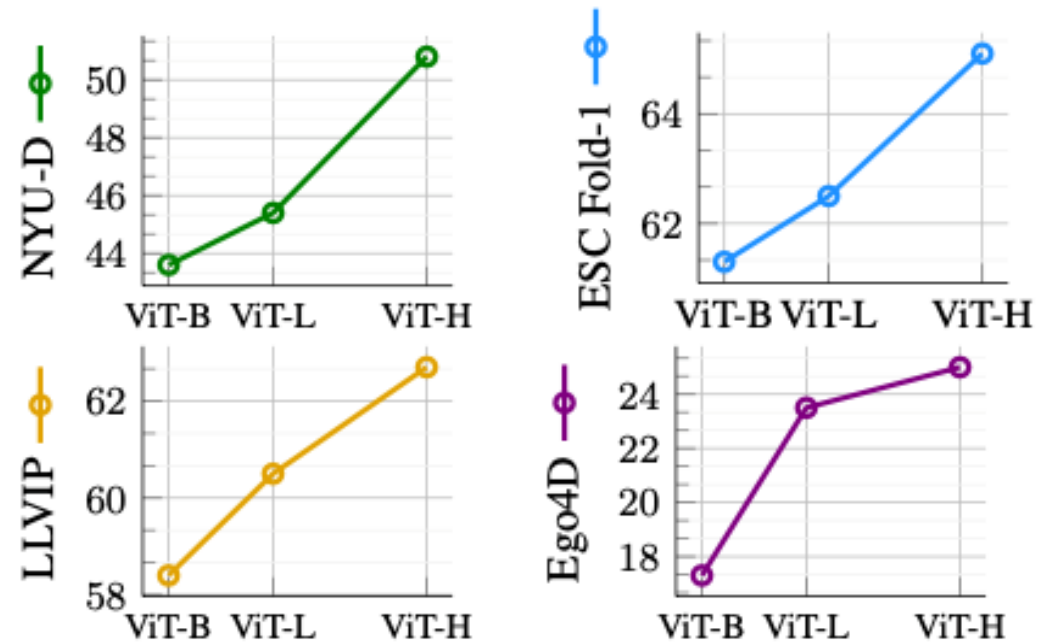


Figure 6. Scaling the image encoder size while keeping the other

Applications of Emergent Capabilities

Without re-training, we can 'upgrade' existing vision models that use CLIP embeddings to use IMAGEBIND embeddings from other modalities such as audio.

- **Multimodal embedding space arithmetic:** add together image and audio embeddings and retrieve the new image
- **Upgrading text-based detectors to audio-based:** replace CLIP-based 'class' (text) embeddings with IMAGEBIND's audio embeddings.
- **Upgrading text-based diffusion models to audio-based:** replace prompt embeddings by ImageBind's embeddings

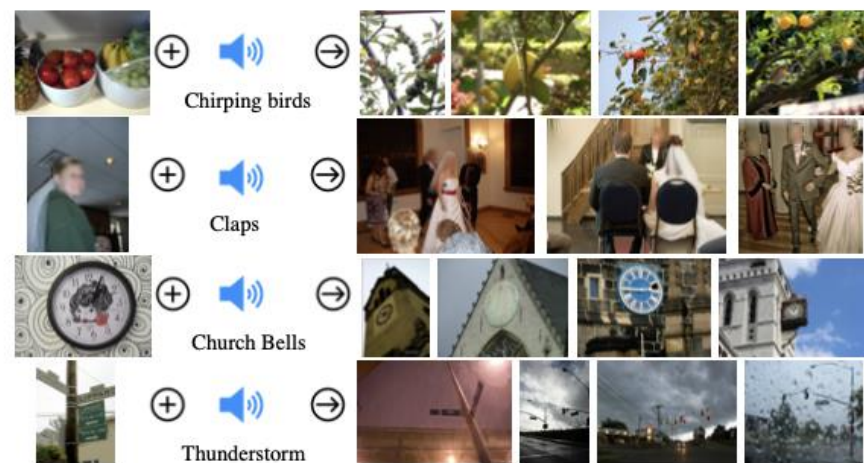


Figure 4. Embedding space arithmetic where we add image

3) Audio to Image Generation

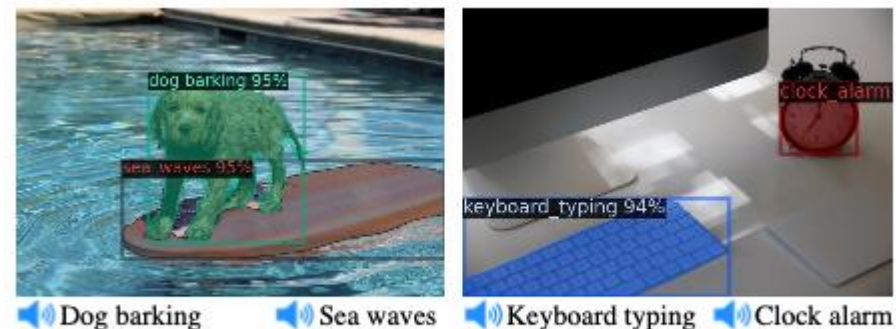
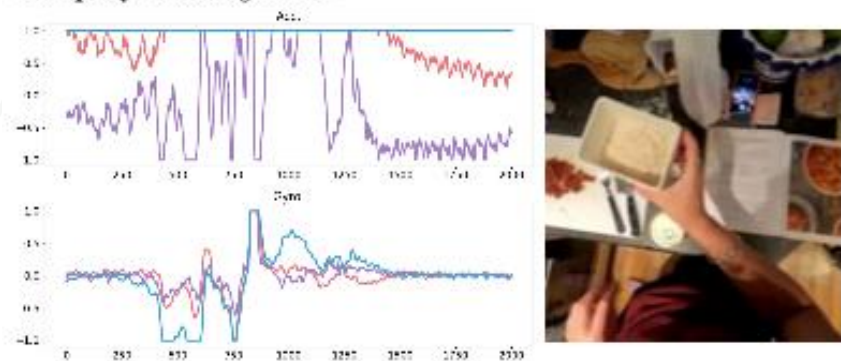


Figure 5. Object detection with audio queries. Simply replacing

Training details

Text query: "Cooking a meal"



Text query: "A person doing gardening work outdoors"

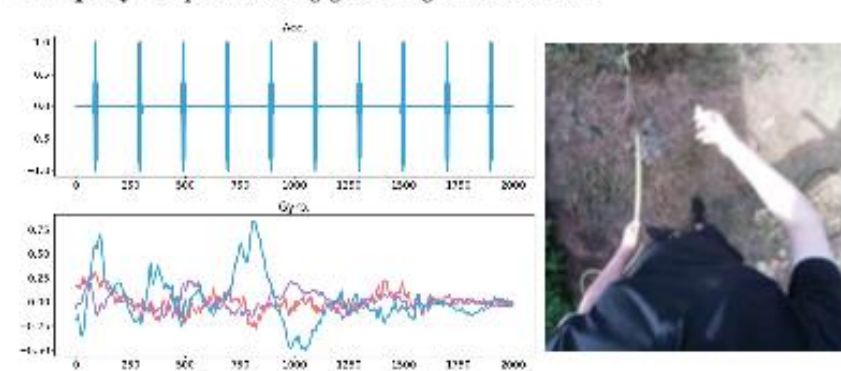


Figure 7. IMU retrievals. Given a text query, we show some MU retrievals and corresponding video frames.

3-channel Accelerometer and Gyroscope recording, matches the text query.

Config	AS	SUN	LLVIP	Ego4D
Vision encoder		ViT-Huge		
embedding dim.	768	384	768	512
number of heads	12	8	12	8
number of layers	12	12	12	6
Optimizer		AdamW		
Optimizer Momentum		$\beta_1 = 0.9, \beta_2 = 0.95$		
Peak learning rate	1.6e-3	1.6e-3	5e-4	5e-4
Weight decay	0.2	0.2	0.05	0.5
Batch size	2048	512	512	512
Gradient clipping	1.0	1.0	5.0	1.0
Warmup epochs		2		
Sample replication	1.25	50	25	1.0
Total epochs	64	64	64	8
Stoch. Depth [29]	0.1	0.0	0.0	0.7
Temperature	0.05	0.2	0.1	0.2
Augmentations:				
RandomResizedCrop				
size	-	224px		-
interpolation	-	Bilinear	Bilinear	-
RandomHorizontalFlip	-	$p = 0.5$	$p = 0.5$	-
RandomEraser	-	$p = 0.25$	$p = 0.25$	-
RandAugment	-	9/0.5	9/0.5	-
Color Jitter	-	0.4	0.4	-
Frequency masking	12	-	-	-

Table 9. Pretraining hyperparameters

Ablation Study – Training Loss and Architecture

- a. Fixed Temperature: **Fixed temperature** in contrastive loss outperforms a learnable one across modalities.
- b. Projection Head: **Linear projection** head is better than MLP for depth and audio embeddings.
- c. Training Duration: **Longer training** boosts zero-shot classification for all modalities.
- d. Image Augmentation: Strong augmentation aids depth classification; Basic augmentation helps audio.
- e.f. Spatial Alignment: **Aligned** image and depth crops improve performance; RandErase^[1] is crucial for depth.
- g.h. Temporal Alignment: **Aligned audio** and video enhance performance; **frequency** augmentation slightly helps audio.

Temp →	Learn	0.05	0.07	0.2	1.0
SUN-D	24.1	27.0	27.3	26.7	28.0
ESC	54.8	56.7	52.4	45.4	24.3

(a) Temperature for loss.

Spatial align →	None	Aligned
SUN-D	16.0	26.7

(e) Spatial alignment of depth.

Epochs →	16	32	64
SUN-D	26.7	27.9	29.9
ESC	56.7	61.3	62.9

(c) Training epochs.

Temporal align →	None	Aligned
ESC	55.7	56.7

(g) Temporal alignment of audio.

Proj head →	Linear	MLP
SUN-D	26.7	26.5
ESC	56.7	51.0

(b) Projection Head.

Data aug →	None	RandErase
SUN-D	24.2	26.7

(f) Depth data aug.

Data aug →	Basic	Strong
SUN-D	25.4	26.7
ESC	56.7	22.6

(d) Data aug for image.

Data aug →	Basic	+Freq mask
ESC	56.5	56.7

(h) Audio data aug.

Ablation Study – Training Loss and Architecture

- **Capacity of the audio and depth encoders:** A smaller Depth encoder improves performance presumably because of the relatively small size of the dataset.
- **Effect of batch size:** batch size can vary across modalities depending on the size of pretraining datasets.
- **ImageBind to evaluate Pretrained Vision Model:**
 - We use image-paired data to align and train text, audio, and depth encoders
 - DINO model is better at emergent zero-shot classification on both depth and audio modalities.

Image Encoder	Audio Encoder (ESC)		Depth Encoder (SUN)	
	ViT-S	ViT-B	ViT-S	ViT-B
ViT-B	52.8	56.7	30.7	26.7
ViT-H	54.8	60.3	33.3	29.5

Table 6. Capacity of the audio and depth encoders and their

Batch size →	512	1k	2k	4k
NYU-D	47.3	46.5	43.0	39.9
ESC	39.4	53.9	56.7	53.9

Table 7. Effect of scaling batch size. We found the optimal batch

	IN1K	VGGS	ESC	SUN-D	NYU-D
DINO [6]	64.4	17.2	44.7	26.8	48.8
DeiT [72]	74.4 [†]	9.6	25.0	25.2	48.0

Table 8. IMAGEBIND as an evaluation tool. We initialize (and

Summary of Strengths, Weaknesses

Strength:

- Multimodal Alignment
- Emergent zero-shot capabilities
- Upgrade existing models without additional training

Weakness:

- Limited benchmark
- Underperform task-specific model

Further Work:

- Enrich Image Alignment Loss: using other alignment data.
- Task-Specific Training: Implement training on targeted downstream tasks, like object detection.
- New Evaluation Benchmarks: better assess emergent zero-shot and cross-modal capabilities.

Zero-shot Capabilities of Binding Touch [5]

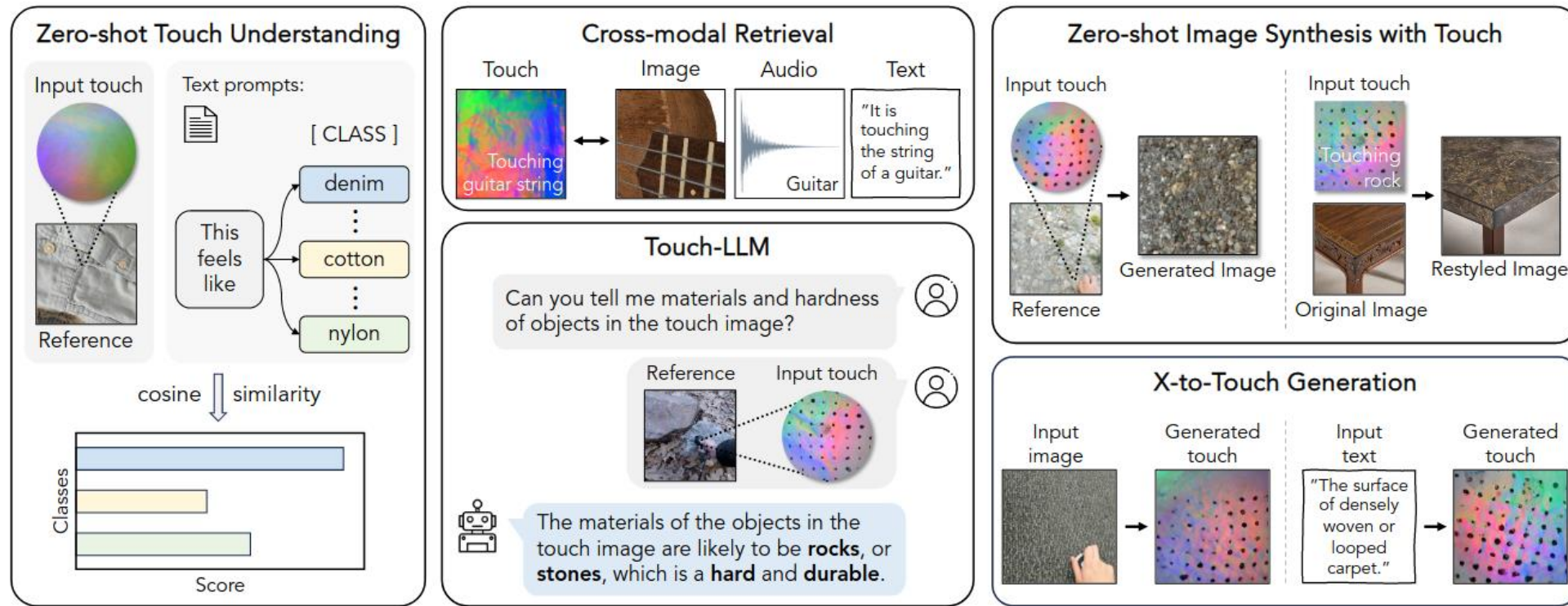


Figure 1. **Putting touch “in touch” with other modalities.** We show that a variety of tactile sensing tasks, ranging from touch image understanding to image synthesis with touch, can be solved zero-shot by aligning touch to pretrained multimodal models, extending previous approaches on work on other modalities [35]. Our learned model can be applied to various vision-based tactile sensors and simulators (*e.g.*, GelSight, DIGIT, Taxim, and Tacto). **For visualization purposes, we show the corresponding visual signal (labeled “reference”) for each touch signal, even though it is not used by the model.**

Sensor Token of Binding Touch [5]

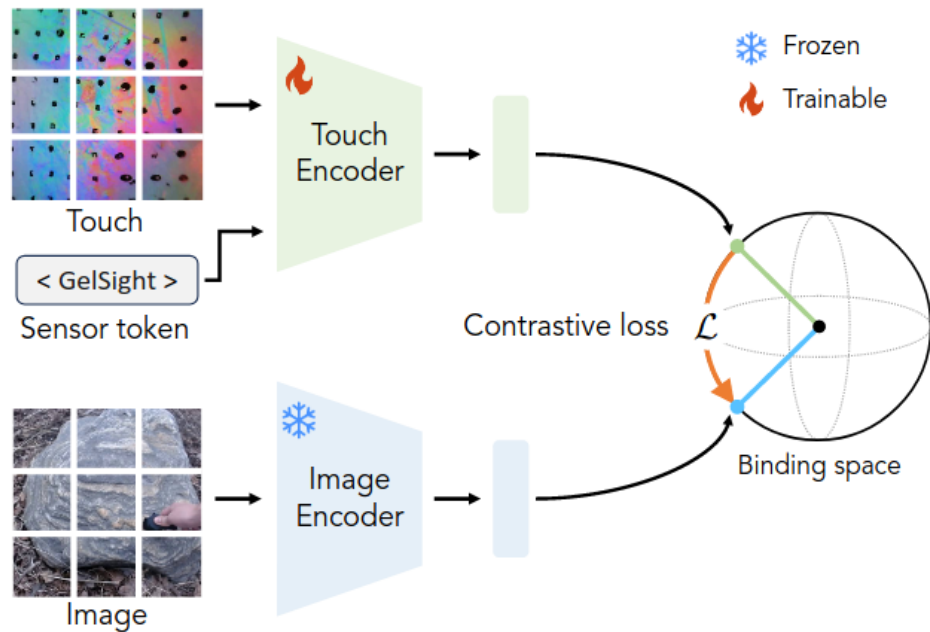


Figure 3. **Method overview.** We align our touch embedding with a pre-trained image embedding derived from large-scale vision language data, using sensor-specific tokens for multi-sensor training.

Specifically, we introduce a set of learnable sensor-specific tokens $\{s_k\}_{k=1}^K$, where $s_k \in R^{L \times D}$, to capture specific details for each sensor, *e.g.*, calibration and background color in touch images, so that the remaining model capacity can be used to learn common knowledge across different type of touch sensors, such as texture and geometry. Here, K represents the number of sensors we train on, L is the number of sensor-specific tokens for each sensor, and D is the token dimension. For the given touch image t_i , and its corresponding tactile sensor tokens s_{t_i} , we append these sensor-specific tokens as prefixes to touch image patch tokens and then encode them with our touch encoder resulting in the final embedding $\mathcal{F}_T(t_i, s_{t_i})$ (Fig. 3). For our contrastive vision-touch pretraining, we optimize:

$$\mathcal{L}_{T \rightarrow V} = -\frac{1}{B} \sum_{i=1}^B \log \frac{\exp(\mathcal{F}_T(t_i, s_{t_i}) \cdot \mathcal{F}_V(v_i) / \tau)}{\sum_{j=1}^B \exp(\mathcal{F}_T(t_i, s_{t_i}) \cdot \mathcal{F}_V(v_j) / \tau)}, \quad (3)$$

In-batch Data Sampling of Binding Touch [5]

In-batch data sampling. We found that batch sampling strategy [18] plays an important role when we train with data, acquired by multiple touch sensors, using contrastive learning. The model will under-perform if we randomly sample from each data source [113] which results in a surplus of easy negatives due to the domain gap between different sensors. Therefore, we design a batch sampling strategy to guarantee that σ percent of training examples in a batch are sampled from the same datasets. Given that our dataset \mathcal{D} is the union over N datasets collected with diverse tactile sensors $\mathcal{D} = \bigcup_{n \in \{1, 2, \dots, N\}} \mathcal{D}_n$, the probability of selecting a given dataset \mathcal{D}_n to sample from is defined as:

$$p_n = \frac{\|\mathcal{D}_n\|}{\sum_{m=1}^N \|\mathcal{D}_m\|}, \quad (4)$$

where $\|\cdot\|$ denotes cardinality. \mathcal{D}_σ denotes the selected dataset from which we perform uniform random sampling to yield $\sigma \cdot B$ examples; the rest $(1 - \sigma) \cdot B$ examples are uniformly sampled from other datasets, i.e., $\mathcal{D} \setminus \mathcal{D}_\sigma$, where

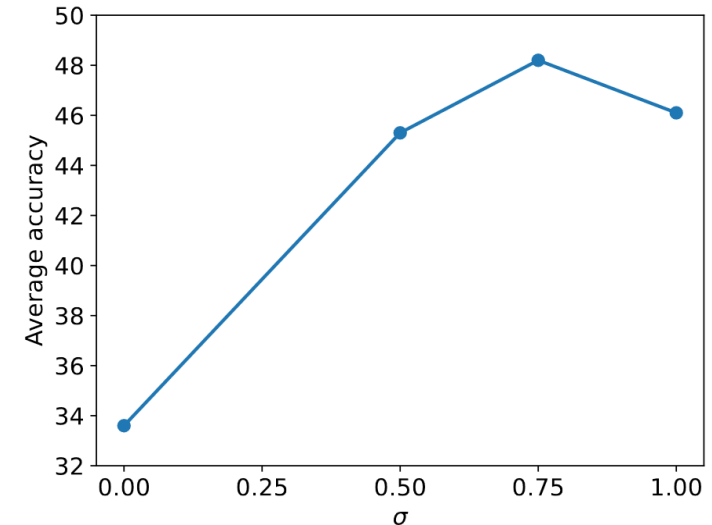


Figure 6. **Effect of σ for in-batch sampling.** We compare the average zero-shot material classification accuracy from six datasets using different σ of 0, 0.5, 0.75, 1.