Large MultiModal Models - ImageBind

Rowel Atienza, PhD
University of the Philippines
github.com/roatienza

2023

Why Foundation Model for MultiModal Data

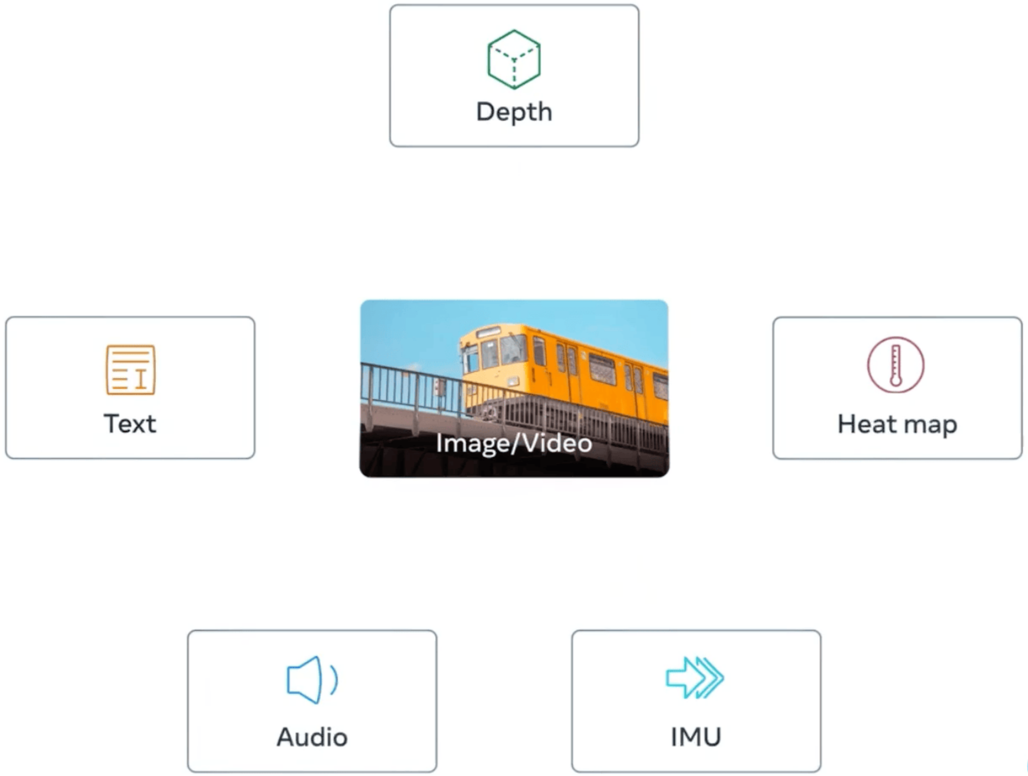
- Broad benefits extending beyond language and vision
- Our world is multimodal
  - Five senses – five sources of data processed in different ways
  - Multimodal data compliment each other (eg sight, sound and smell of a fireplace)

Why ImageBind

- Large scale training in 6 modalities
  - Language, vision, depth map, IMU, sound, thermal
- Demonstrated competitive zero-shot capabilities

Girdhar, Rohit, et al. "ImageBind: One Embedding Space To Bind Them All." *arXiv preprint arXiv:2305.05665* (2023).

ImageBind



ImageBind – Aligning Modalities

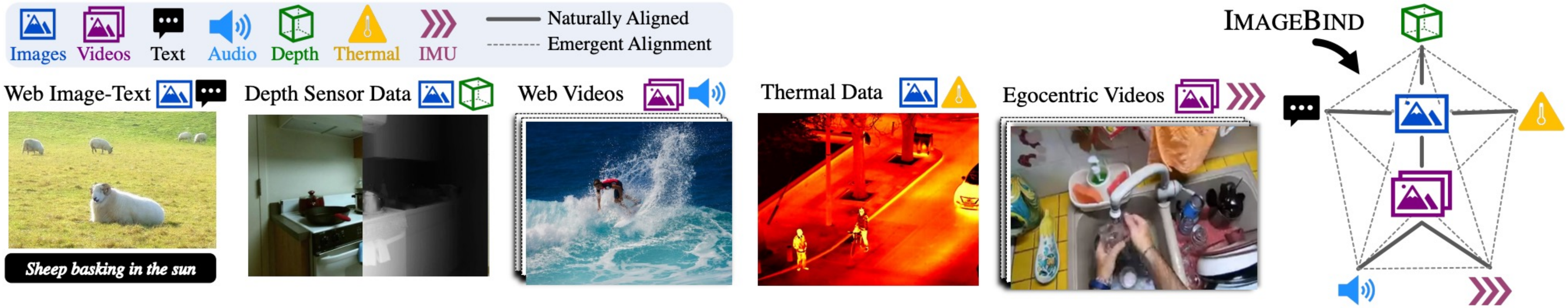


Figure 2. IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc*. IMAGEBIND links all these modalities in a common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

Core Idea

- Train from naturally aligned data
  - Image and {text, thermal, audio, depth, IMU}

- Take advantage of implicit alignment
  - $x \in \{\text{image, text, thermal, audio, depth, IMU}\}$ and $y \in \{\text{image, text, thermal, audio, depth, IMU}\} \neq x$
  - Pick one from x and pick one from y (eg.$x \in depth$ and $y \in \ audio$)

Dataset

- $\{\mathcal{I}, \mathcal{M}\}$ where $\mathcal{I}$ is image and $\mathcal{M}$ is another modality
- Objective is to learn a single joint embedding with only $\mathcal{I}$ as the only common modality using InfoNCE
  - Image embedding : $\boldsymbol{q}_i = f(\boldsymbol{I}_i)$
  - Modality embedding: $\boldsymbol{k}_i = g(\boldsymbol{M}_i)$

$$\mathcal{L}_{\mathcal{I},\mathcal{M}} = -\log \frac{e^{\boldsymbol{q}_i^T \boldsymbol{k}_i / \tau}}{e^{\boldsymbol{q}_i^T \boldsymbol{k}_i / \tau} + \sum_{j \neq i} e^{\boldsymbol{q}_i^T \boldsymbol{k}_j / \tau}}$$

[InfoNCE] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. In NeurIPS, 2018.

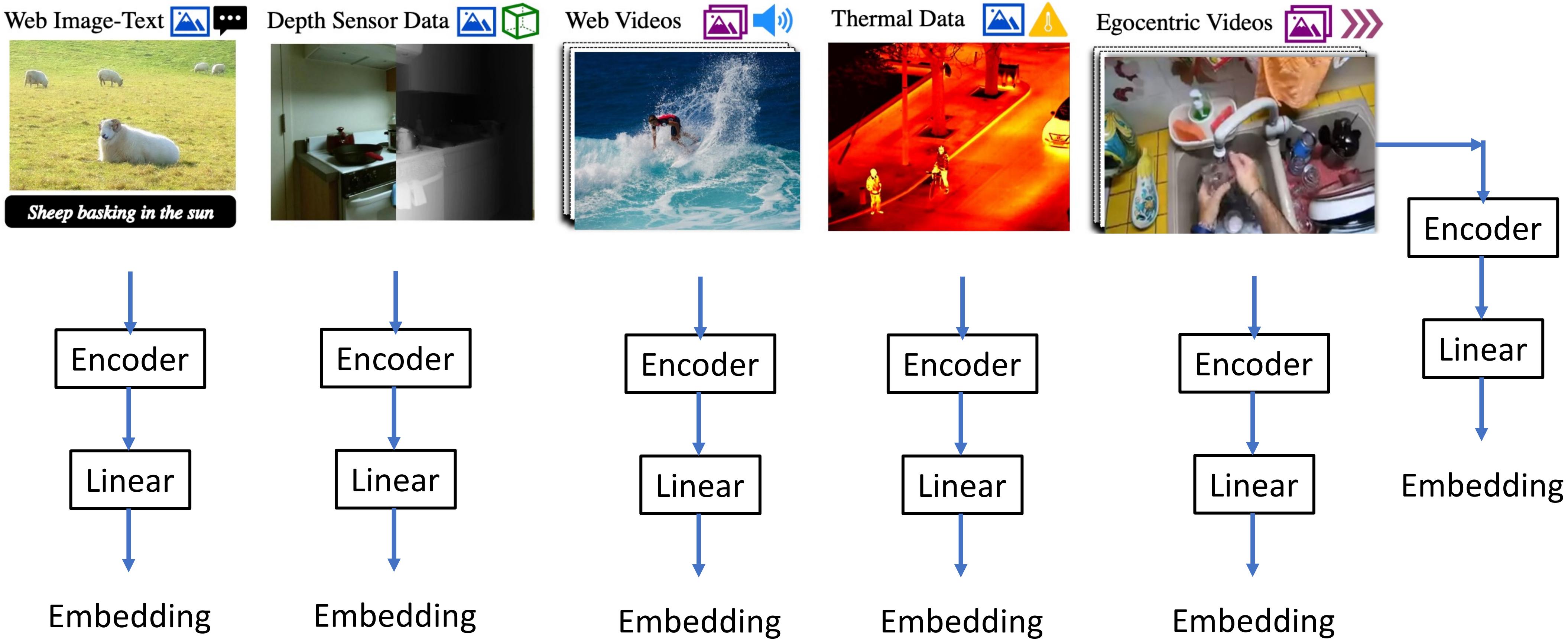
InfoNCE – Core Idea

- Embeddings of related modalities i get closer together
  - Also known as positive pairs (eg $\text{image}: audio = \{I_{dog}, M_{dog}\}$)
- Embeddings of unrelated modalities $i \neq j$ move away from each other
  - Also known as negative pairs (eg $\text{image}: audio = \{I_{dog}, M_{train}\}$)

Emergent Abilities

- Natural alignment: $\{\mathcal{I}, \mathcal{M}_1\}$ and $\{\mathcal{I}, \mathcal{M}_2\}$
- Emergent alignment: $\{\mathcal{M}_1, \mathcal{M}_2\}$

6 Encoders – All transformers



Zero-Shot Emergent Behavior

Zero-shot classification across modalities

	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† [26]	-	68.6† [26]	-	-
Absolute SOTA	91.0 [80]	60.7 [65]	89.9 [78]	57.7 [77]	76.7 [20]	64.9 [20]	49.6 [38]	52.5 [35]	97.0 [9]	-	-

With text prompt

Zero-shot audio retrieval

	Emergent	Clotho		AudioCaps		ESC
		R@1	R@10	R@1	R@10	Top-1
Uses audio and text supervision						
AudioCLIP [26]	✗	–	–	–	–	**68.6**
Uses audio and text loss						
AVFIC [50]	✗	3.0	17.5	8.7	37.7	–
No audio and text supervision						
IMAGEBIND	✓	**6.0**	**28.4**	**9.3**	**42.3**	66.9
Supervised						
AVFIC finetuned [50]	✗	8.4	38.6	–	–	–
ARNLQ [52]	✗	12.6	45.4	24.3	72.1	–

Table 3. Emergent zero-shot audio retrieval and classification. We compare IMAGEBIND to prior work on zero-shot audio retrieval and audio classification. Without using audio-specific supervision, IMAGEBIND outperforms prior methods on zero-shot retrieval and has comparable performance on the classification task. IMAGEBIND's emergent zero-shot performance approaches those of specialist supervised models.

Zero-shot text and image retrieval

	Modality	Emergent	MSR-VTT		
			R@1	R@5	R@10
MIL-NCE [48]	V	✗	8.6	16.9	25.8
SupportSet [56]	V	✗	10.4	22.2	30.0
FIT [5]	V	✗	15.4	33.6	44.1
AVFIC [50]	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. We compare IMAGEBIND's emergent retrieval performance using audio and observe that it performs favorably to methods that use the stronger video modality for retrieval.

Few-shot classification

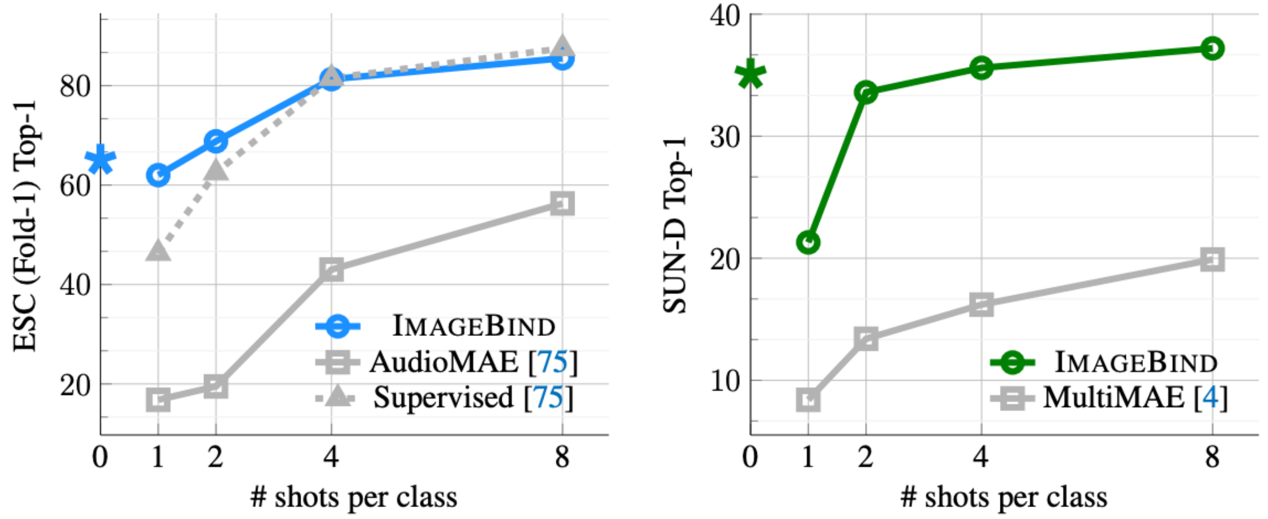


Figure 3. Few-shot classification on audio and depth. We report the emergent zero-shot classification performance on each benchmark (denoted by ⋆). We train linear classifiers on fixed features for the $\geq$ 1-shot case. **(Left)** In all settings, IMAGEBIND outperforms the self-supervised AudioMAE model. IMAGEBIND even outperforms a supervised AudioMAE model upto 4 shot learning showing its strong generalization. **(Right)** We compare with the MultiMAE model trained with images, depth, and semantic segmentation masks. IMAGEBIND outperforms MultiMAE across all few-shot settings on few-shot depth classification.

Embedding space arithmetic

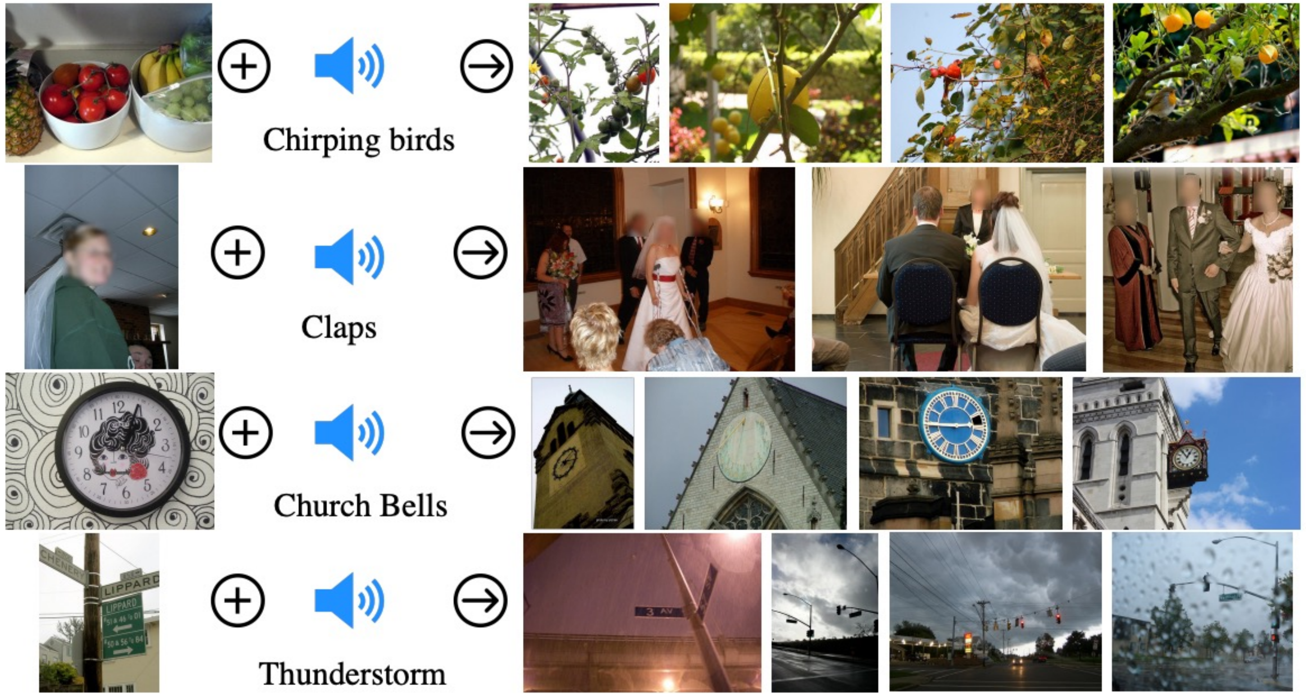


Figure 4. Embedding space arithmetic where we add image and audio embeddings, and use them for image retrieval. The composed embeddings naturally capture semantics from different modalities. Embeddings from an image of fruits + the sound of birds retrieves images of birds surrounded by fruits.

Code demo