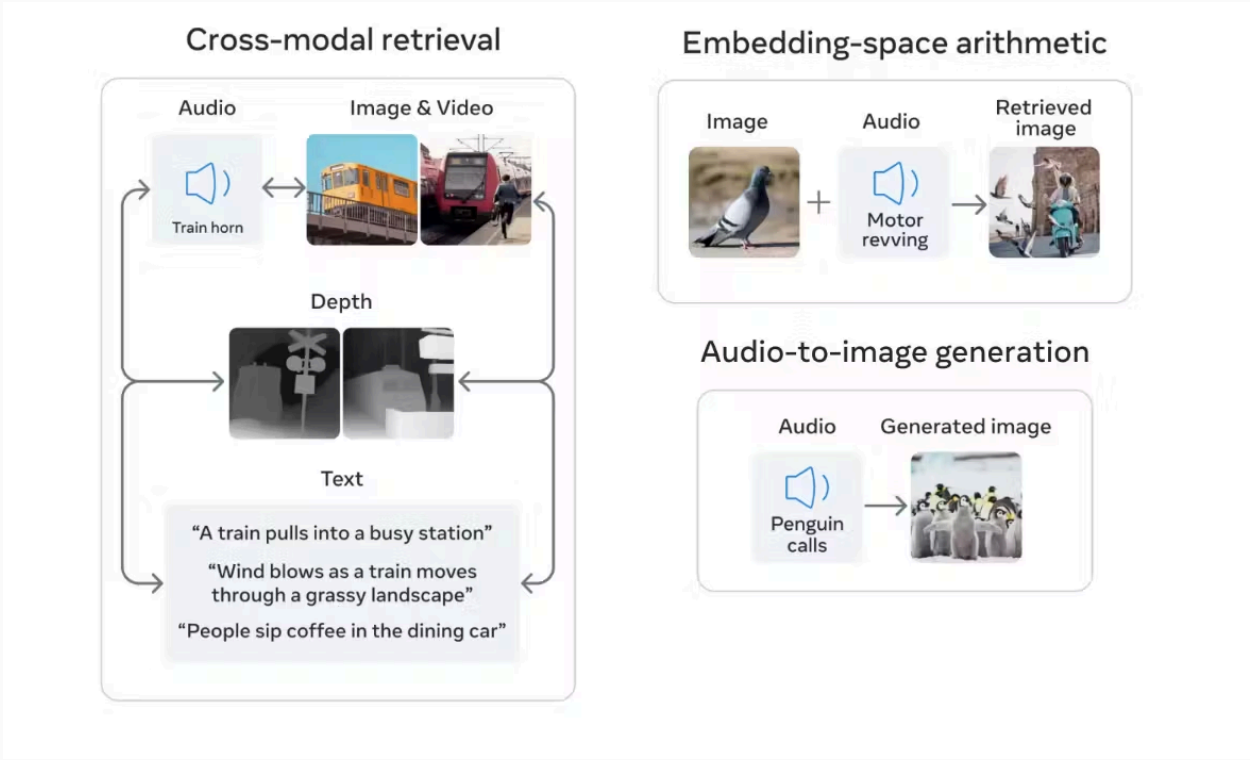
ImageBind MultiJoint Embedding Model from Meta Explained

May 10, 2023 | 5 mins

In the ever-evolving landscape of artificial intelligence, Meta has once again raised the bar with its open-source model, ImageBind, pushing the

boundaries of what's possible and bringing us closer to human-like learning.

Innovation is at the heart of Meta's mission, and their latest offering, ImageBind, is a testament to that commitment. While generative AI models like Midjourney, Stable Diffusion, and DALL-E 2 have made significant strides in pairing words with images, ImageBind goes a step further, casting a net encompassing a broader range of sensory data.



Source

ImageBind marks the inception of a framework that could generate complex virtual environments from as simple an input as a text prompt, image, or audio recording. For example, imagine the possibility of creating a realistic virtual representation of a bustling city or a tranquil forest, all from mere words or sounds.

The uniqueness of ImageBind lies in its ability to integrate six types of data: visual data (both image and video), thermal data (infrared images), text, audio, depth information, and, intriguingly, movement readings from an inertial measuring unit (IMU). This integration of multiple data types into a single embedding space is a concept that will only fuel the ongoing boom in generative AI.

The model capitalizes on a broad range of image-paired data to establish a unified representation space. Unlike traditional models, ImageBind does not require all modalities to appear concurrently within the same datasets. Instead, it takes advantage of the inherent linking nature of images, demonstrating that aligning the embedding of each modality with image embeddings gives rise to an emergent cross-modal alignment.

While ImageBind is presently a research project, it's a powerful indicator of the future of multimodal models. It also underscores Meta's commitment to sharing AI research at a time when many of its rivals, like OpenAI and Google, maintain a veil of secrecy.

In this explainer, we will cover the following:

- What is multimodal learning
- What is an embedding
- ImageBind Architecture
- ImageBind Performance
- Use cases of ImageBind

Meta's History of Releasing Open-Source AI Tools

Meta has been on an incredible run of successful launches over the past two months.

Segment Anything Model

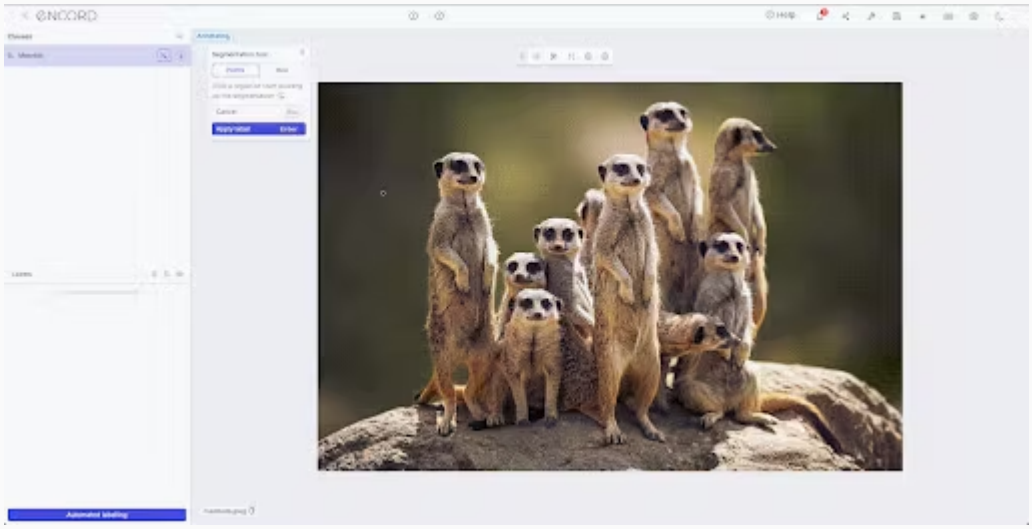
MetaAI's **Segment Anything Model (SAM)** changed image segmentation for the future by applying foundation models traditionally used in natural language processing.

Source

SAM uses prompt engineering to adapt to a variety of segmentation problems. The model enables users to select an object to segment by interacting with prompting using bounding boxes, key points, grids, or text.

SAM can produce multiple valid masks when the object to segment is uncertain and can automatically identify and mask all objects in an image.

Most remarkably, **integrated with a labeling platform**, SAM can provide real-time segmentation masks once the image embeddings are precomputed, which for normal-size images is a matter of seconds.

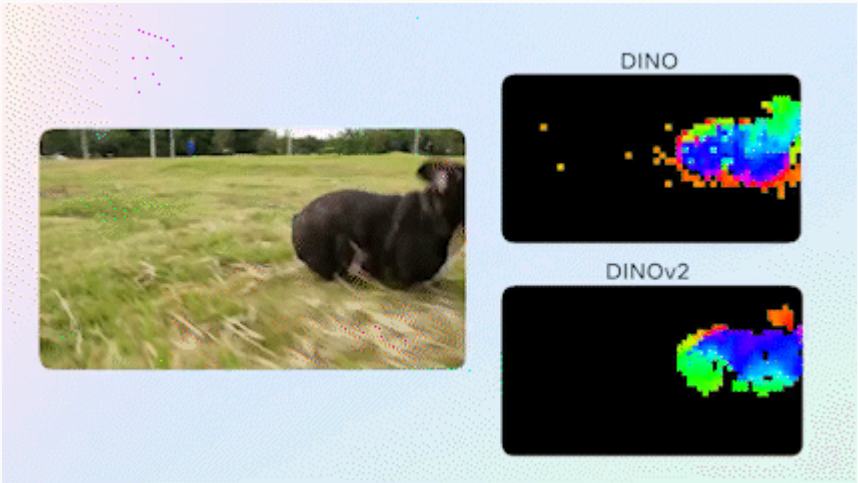


SAM has shown great potential in reducing labeling costs, providing a much-awaited solution for AI-assisted labeling. Whether it's for medical applications, geospatial analysis, or autonomous vehicles, SAM is set to transform the field of computer vision drastically.

> *Check out the full explainer if you would like to know more about **Segment Anything Model**.*

DINOv2

DINOv2 is an advanced self-supervised learning technique designed to learn visual representations from images without using labeled data, which is a significant advantage over supervised learning models that rely on large amounts of labeled data for training.



DINO can be used as a powerful feature extractor for tasks like image classification or object detection. The process generally involves two

stages: pretraining and fine-tuning.

- **Pretraining:** During this stage, the DINO model is pre-trained on a large dataset of unlabeled images. The objective is to learn useful visual representations using self-supervised learning. Once the model is trained, the weights are saved for use in the next stage.
- **Fine-tuning:** In this stage, the pre-trained DINO model is fine-tuned on a task-specific dataset, which usually contains labeled data. For image classification or object detection tasks, you can either use the DINO model as a backbone or as a feature extractor, followed by task-specific layers (like fully connected layers for classification or bounding box regression layers for object detection).

The challenges with SSL remain in designing practical tasks, handling domain shifts, and understanding model interpretability and robustness. However, DINOv2 overcomes these challenges using techniques such as self-DIstillation with NO labels (DINO), which uses SSL and knowledge distillation methods to transfer knowledge from larger models to smaller ones.

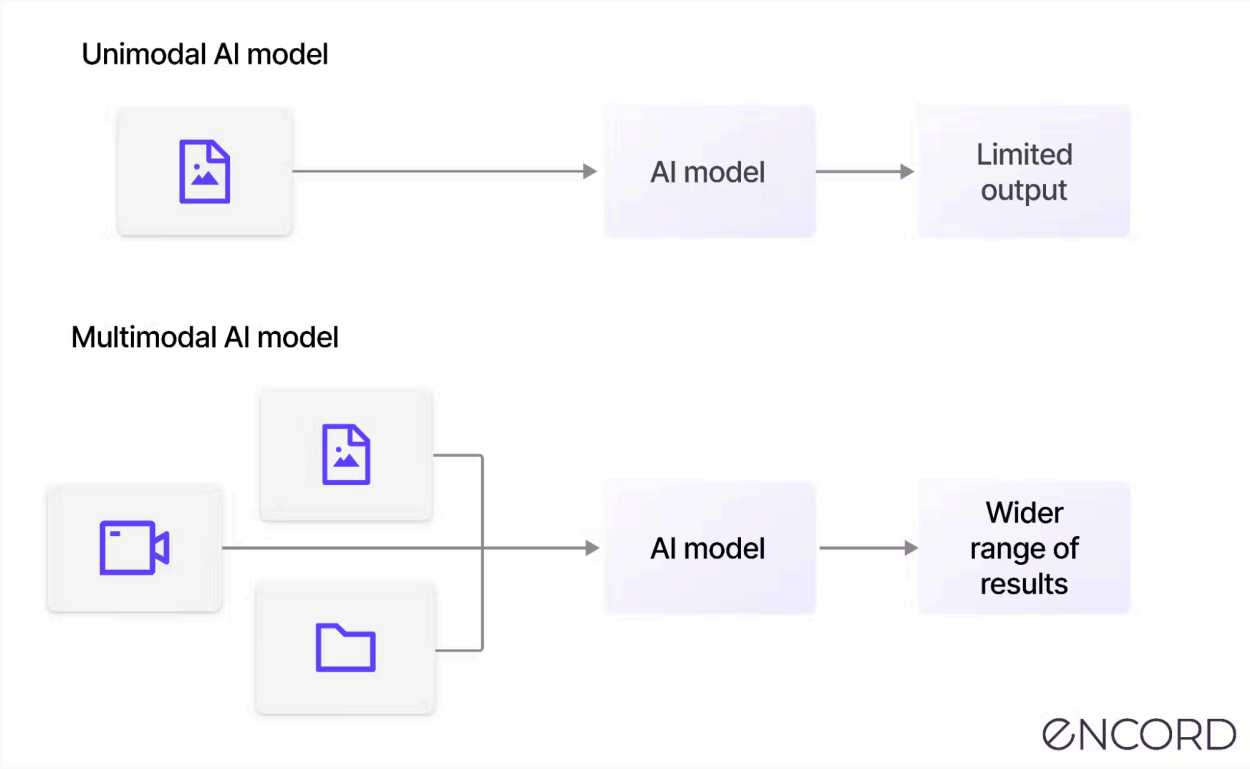
> *You can read the **detailed post on DINOv2** to find out more about it.*

What is Multimodal Learning?

Multimodal learning involves processing and integrating information from multiple modalities, such as images, text, audio, video, and other forms of data.

It combines different sources of information to gain a deeper understanding of a particular concept or phenomenon.

In contrast to unimodal learning, where the focus is on a single modality (e.g.,text-only or image-only), multimedia learning leverages the complementary nature of multiple modalities to improve learning outcomes.

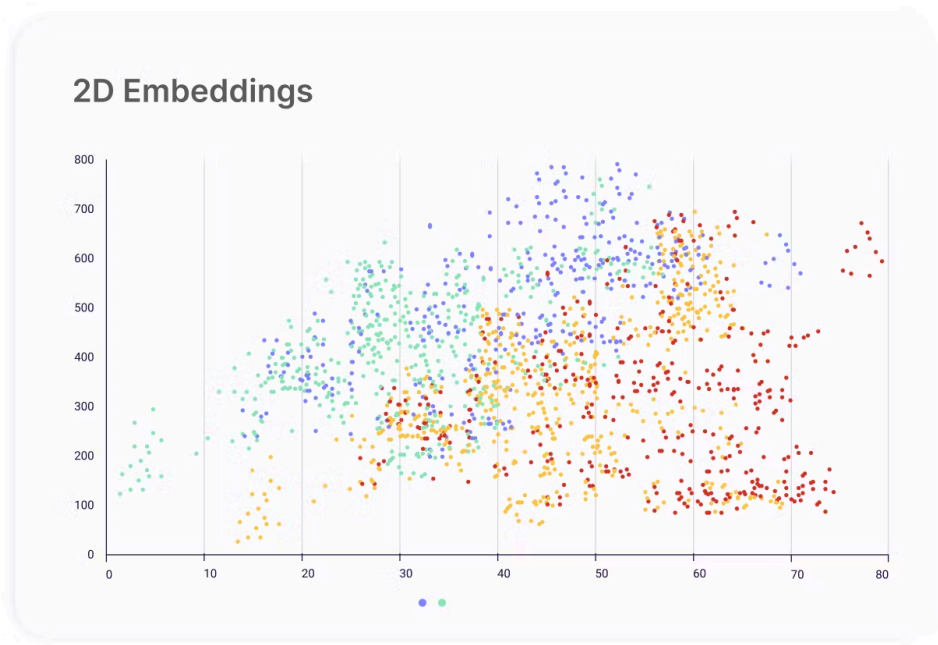


Multimodal learning aims to enable machine learning algorithms to learn from and make sense of complex data from different sources. It allows artificial intelligence to analyze different kinds of data holistically, as humans do.

What is an Embedding?

An embedding is a lower-dimensional representation of high-dimensional vectors, simplifying the processing of significant inputs like sparse vectors representing data. The objective of extracting embeddings is to capture the semantics of the input data by representing them in much lower dimensional space so that semantically similar samples will be close to each other.

Embeddings address the “curse of dimensionality” problem in machine learning, where the input space is too large and sparse to process efficiently by traditional machine learning algorithms. By mapping the high-dimensional input data into a lower-dimensional embedding space, we can reduce the dimensionality of the data and make it easier to learn patterns and relationships between inputs.



Embeddings are especially useful where the input space is typically very high-dimensional and sparse, like text data. For text data, each word is represented by a one-hot vector which is an embedding. By learning embeddings for words, we can capture the semantic meaning of the words and represent them in a much more compact and informative way.

Embeddings are valuable in machine learning since they can be learned from large volumes of data and employed across models.

What is ImageBind?

ImageBind is a brand-new approach to learning a joint embedding space across six modalities. The model has been developed by Meta AI's FAIR Lab and was released on the 9th of May, 2023, on **GitHub**, where you can also find the ImageBind code.

The advent of ImageBind marks a significant shift in machine learning and AI, as it pushes the boundaries of multimodal learning.

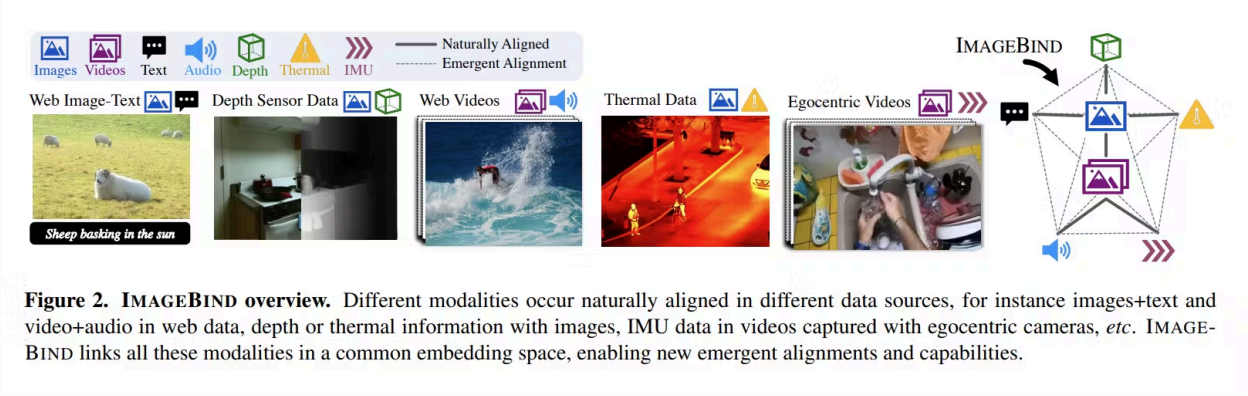


Figure 2. IMAGEBIND overview. Different modalities occur naturally aligned in different data sources, for instance images+text and video+audio in web data, depth or thermal information with images, IMU data in videos captured with egocentric cameras, *etc*. IMAGEBIND links all these modalities in a common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

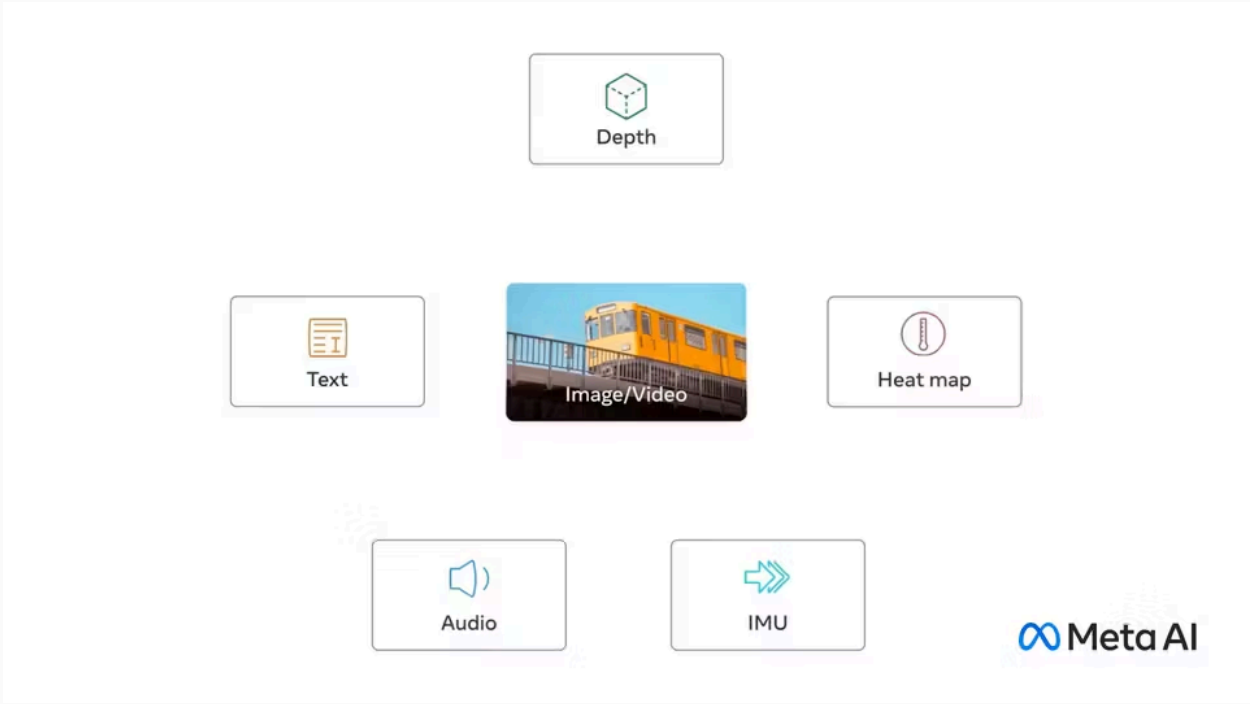
Source

By integrating and comprehending information from multiple modalities, ImageBind paves the way for more advanced AI systems that can process and analyze data more humanistically.

The Modalities Integrated in ImageBind

ImageBind is designed to handle six distinct modalities, allowing it to learn and process information more comprehensively and holistically. These modalities include:

- **Text:** Written content or descriptions that convey meaning, context, or specific details about a subject.
- **Image/Video:** Visual data that captures scenes, objects, and events, providing rich contextual information and forming connections between different elements within the data.
- **Audio:** Sound data that offers additional context to visual or textual information, such as the noise made by an object or the soundscape of a particular environment.
- **Depth (3D):** Three-dimensional data that provides information about the spatial relationships between objects, enabling a better understanding of their position and size about each other.
- **Thermal (heatmap):** Data that captures the temperature variations of objects and their surroundings, giving insight into the heat signatures of different elements within a scene.
- **IMU**: Sensor data that records motion and position, allowing AI systems to understand the movements and dynamics of objects in a given environment.



Source

By incorporating these six modalities, ImageBind can create a unified representation space that enables you to learn and analyze data across various forms of information.

This improves the model's understanding of the world around it and allows it to make better predictions and generate more accurate results based on the data it processes.

ImageBind Architecture

The framework of ImageBind is speculative as the Meta team has not released it; therefore, the framework may still be subject to changes. Here, the architecture discussed is based on the information presented in the research paper published by the team.

> 💡 Encord team will update the blog post once the architecture has been released.

The ImageBind framework uses a separate encoder for image, text, audio, thermal image, depth image, and IMU modalities. A modality-specific linear projection head is added to each encoder to obtain a fixed

dimensional embedding. This embedding is normalized and used in the **InfoNCE loss**.

The architecture of ImageBind consists of three main components:

- A modality-specific encoder
- Cross-model attention module
- A joint embedding space

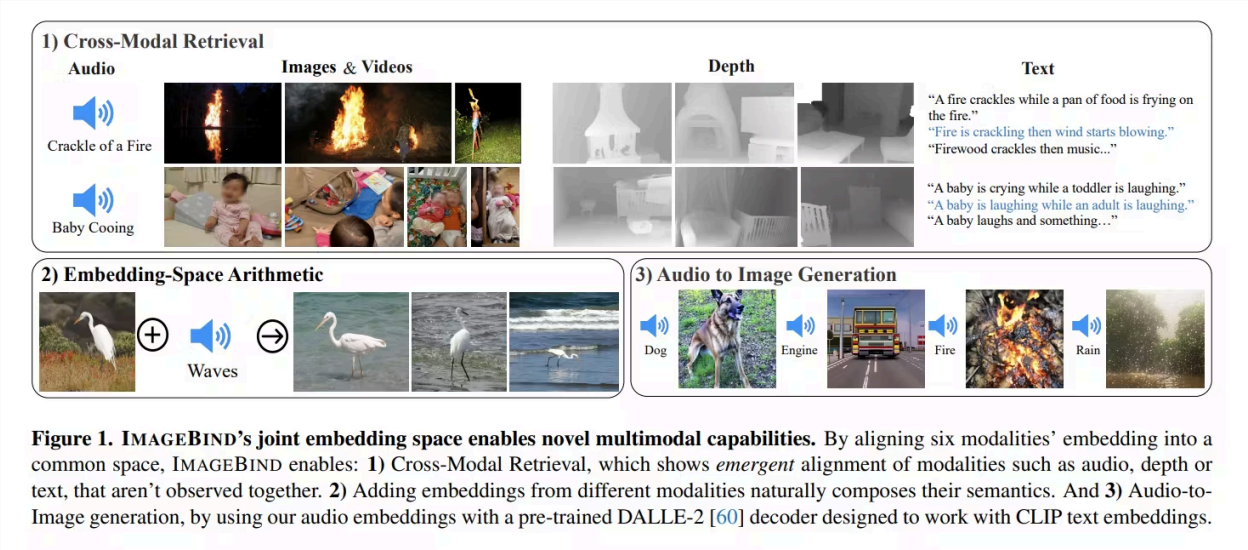


Figure 1. IMAGEBIND's joint embedding space enables novel multimodal capabilities. By aligning six modalities' embedding into a common space, IMAGEBIND enables: **1)** Cross-Modal Retrieval, which shows *emergent* alignment of modalities such as audio, depth or text, that aren't observed together. **2)** Adding embeddings from different modalities naturally composes their semantics. And **3)** Audio-to-Image generation, by using our audio embeddings with a pre-trained DALLE-2 [60] decoder designed to work with CLIP text embeddings.

Example of ImageBind framework for multi-modal tasks.Source

Modality-Specific Encoder

The first component involves training modality-specific encoders for each data type. Next, the encoders convert the raw data into a joint embedding space, where the model can learn the relationships between the different modalities.

The modality encoders use Transformer architecture. The encoders are trained using standard backpropagation with a loss function that encourages the embedding vectors from different modalities to be close to each other if they are related and far from each other if unrelated.

- **For images and videos**, it uses the **Vision Transformer (ViT)**. For video inputs, 2-frame video clips were sampled over a 2-second duration.
- **The audio inputs** are transformed into 2D **mel-spectrograms** using the method outlined in **AST: Audio Spectrogram Transformer**, which

involves converting a 2-second audio sample at 26kHz. As the mel-spectrogram is a 2D signal similar to an image, a ViT model is used to process it.

- **For texts**, a recurrent neural network (RNN) or transformer is used as the encoder. The transformer takes the raw text as input and produces a sequence of hidden states, which are then aggregated to produce the audio embedding vector.
- **The thermal and depth inputs** are treated as 1-channel images where ViT-B and ViT-S encoders are used, respectively.

Cross-Modal Attention Module

The second component, the cross-modal attention module, consists of three main sub-components:

- A modality-specific attention module,
- A cross-modal attention fusion module, and
- A cross-modal attention module

The modality-specific attention module takes the embedding vectors for each modality as input. It produces a set of attention weights that indicate the relative importance of different elements within each modality. This allows the model to focus on specific aspects of each modality relevant to the task.

The cross-modal attention fusion module takes the attention weights from each modality and combines them to produce a single set of attention weights that determine how much importance should be placed on each modality when performing the task, By selectively attending to different modalities based on their significance in the current task, the model can effectively capture the complex relationships and interactions between different data types.

The cross-modal attention module is trained end-to-end with the rest of the model using backpropagation and a task-specific loss function. By jointly learning the modality-specific attention weights and cross-modal attention fusion weights, the model can effectively integrate information from multiple modalities to improve performance on various multi-modal machine learning tasks.

Joint Embedding

The third component is a joint embedding space where all modalities are represented in a single vector space. The embedding vectors are mapped into a common joint embedding space using a shared projection layer, which is also learned during training. This process ensures that embedding vectors from different modalities are located in the same space, where they can be directly compared and combined.

The joint embedding space aims to capture the complex relationships and interactions between the different modalities. For example, related images and text should be located close to each other, while unrelated images and texts should be located far apart.

Using a joint embedding space that enables the direct comparison and combination of different modalities, ImageBind can effectively integrate information from multiple modalities to improve performance on various multi-modal machine learning tasks.

ImageBind Training Data

ImageBind is a novel approach to multimodal learning that capitalizes on images' inherent "binding" properties to connect different sensory experiences.

It's trained using image-paired data (image, X), meaning each image is associated with one of the five other types of data (X): text, audio, depth, IMU, or thermal data. The image and text encoder models are not updated during the ImageBind training whereas the encoders for other modalities are updated.

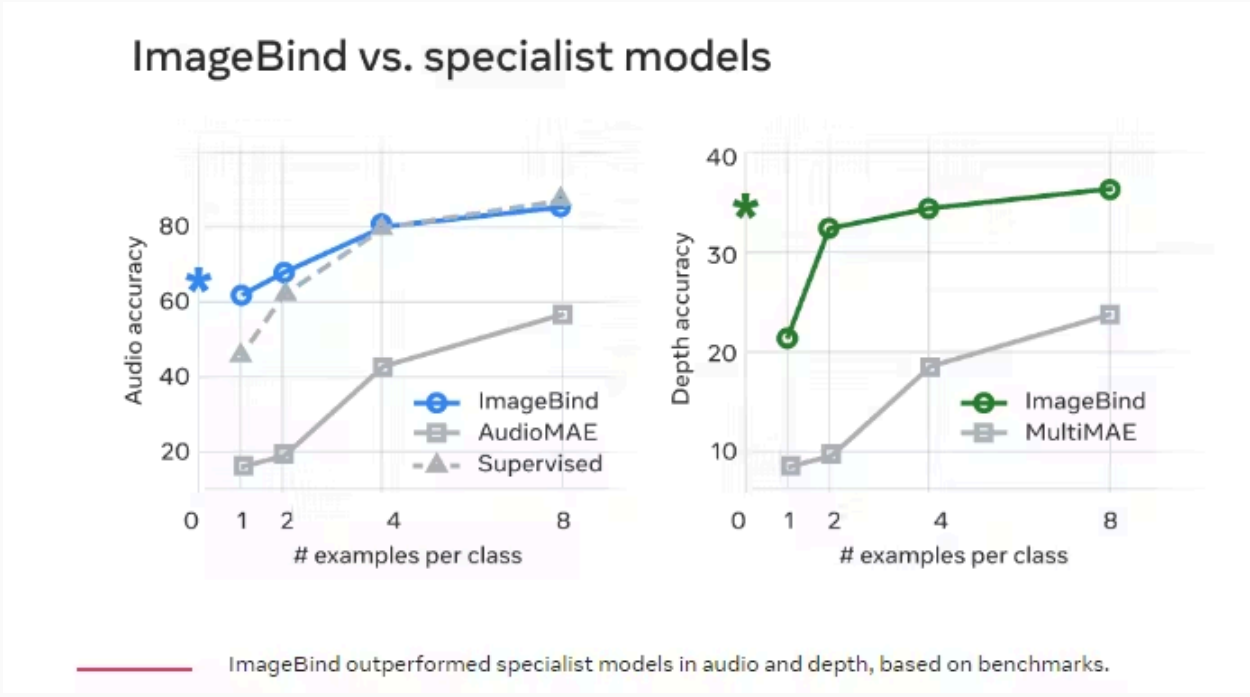
- **OpenCLIP ViT-H Encoder**: This encoder is utilized for initializing and freezing the image and text encoders. The ViT-H encoder is a part of the **OpenCLIP** model, a robust vision-language model that provides rich image and text representations.
- **Audio Embeddings**: ImageBind uses the **Audioset** dataset for training audio embeddings. Audioset is a comprehensive collection of audio event annotations and recordings, offering the model a wide array of sounds to learn from.

- **Depth Embeddings**: The **SUN RGB-D dataset** is used to train depth embeddings. This dataset includes images annotated with depth information, allowing the model to understand spatial relationships within the image.
- **IMU Data:** The **Ego4D dataset** is used for IMU data. This dataset provides IMU readings, which are instrumental in understanding movement and orientation related to the images.
- **Thermal Embeddings**: The **LLVIP dataset** is used for training thermal embeddings. This dataset provides thermal imaging data, adding another layer of information to the model's understanding of the images.

ImageBind Performance

The performance of the ImageBind model is benchmarked to several state-of-the-art approaches. In addition, it is compared against prior work in zero-shot retrieval and classification tasks.

ImageBIND achieves better zero-shot text-to-audio retrieval and classification performance without any text pairing for audio during training. For example, on the Clotho dataset, ImageBIND performs double the performance of **AVFIC** and achieves comparable audio classification performance on ESC compared to the supervised **AudioCLIP** model. On the AudioSet dataset, it can generate high-quality images from audio inputs using a pre-trained **DALLE-2** decoder.



Left: *The performance of ImageBIND exceeds that of the self-supervised AudioMAE model and even surpasses a supervised AudioMAE model by up to 4 shot learning, demonstrating impressive generalization capabilities.* ***Right:*** *ImageBIND outperforms the MultiMAE model, which is trained with images, depth, and semantic segmentation masks across all few-shot settings on few-shot depth classification.* *Source*

Is ImageBind Open Source?

Sadly, ImageBind's code and model weights are released under **CC-BY-NC 4.0 license**. This means it can only be used for research purposes, and all commercial use cases are strictly forbidden.

Future Potential of Multimodal Learning with ImageBind

With its ability to combine information from six different modalities, ImageBind has the potential to create exciting new AI applications,

particularly for creators and the AI research community.

How ImageBind Opens New Avenues

ImageBind's multimodal capabilities are poised to unlock a world of creative possibilities. Seamlessly integrating various data forms empowers creators to:

- **Generate rich media content**: ImageBind's ability to bind multiple modalities allows creators to generate more immersive and contextually relevant content. For instance, imagine creating images or videos based on audio input, such as generating visuals that match the sounds of a bustling market, a blissful rainforest, or a busy street.
- **Enhance content with cross-modal retrieval:** Creators can easily search for and incorporate relevant content from different modalities to enhance their work. For example, a filmmaker could use ImageBind to find the perfect audio clip to match a specific visual scene, streamlining the creative process.
- **Combining embeddings of different modalities:** The joint embedding space allows us to compose two embeddings: e.g., the image of fruits on a table + the sound of chirping birds, and retrieve an image that contains both these concepts, i.e., fruits on trees with birds. A wide range of compositional tasks will likely be made possible by emergent compositionality, which allows semantic content from various modalities to be combined.
- **Develop immersive experiences**: ImageBind's ability to process and understand data from various sensors, such as depth and IMU, opens the door for the development of virtual and augmented reality experiences that are more realistic and engaging.

Other future use cases in more traditional industries are:

- **Autonomous Vehicles:** With its ability to understand depth and motion data, ImageBind could play a crucial role in developing autonomous vehicles, helping them to perceive and interpret their surroundings more effectively.
- **Healthcare and Medical Imaging:** ImageBind could be applied to process and understand various types of medical data (visual, auditory, PDF, etc.) to assist in diagnosis, treatment planning, and patient monitoring.

- **Smart Homes and IoT:** ImageBind could enhance the functionality of smart home devices by enabling them to process and understand various forms of sensory data, leading to more intuitive and effective automation.
- **Environmental Monitoring:** ImageBind could be used in drones or other monitoring devices to analyze various environmental data and detect changes or anomalies, aiding in tasks like wildlife tracking, climate monitoring, or disaster response.
- **Security and Surveillance**: By processing and understanding visual, thermal, and motion data, ImageBind could improve the effectiveness of security systems, enabling them to detect and respond to threats more accurately and efficiently.

The Future of Multimodal Learning

ImageBind represents a significant leap forward in multimodal learning. This has several implications for the future of AI and multimodal learning:

- **Expanding modalities:** As researchers continue to explore and integrate additional modalities, such as touch, speech, smell, and even brain signals, models like ImageBind could play a crucial role in developing richer, more human-centric AI systems.
- **Reducing data requirements:** ImageBind demonstrates that it is possible to learn a joint embedding space across multiple modalities without needing extensive paired data, potentially reducing the data required for training and making AI systems more efficient.
- **Interdisciplinary applications**: ImageBind's success in multimodal learning can inspire new interdisciplinary applications, such as combining AI with neuroscience, linguistics, and cognitive science, further enhancing our understanding of human intelligence and cognition.

As the field of multimodal learning advances, ImageBind is poised to play a pivotal role in shaping the future of AI and unlocking new possibilities for creators and researchers alike.

Conclusion

ImageBind, the first model to bind information from six modalities, is undoubtedly a game-changer in artificial intelligence and multimodal learning.

Its ability to create a single shared representation space across multiple forms of data is a significant step toward machines that can analyze data as holistically as humans. This is an exciting prospect for the AI research community and creators who can leverage these capabilities for richer, more immersive content in the future.

Moreover, ImageBind provides a blueprint for future open-source models, showing that creating a joint embedding space across multiple modalities is possible using specific image-paired data. This could lead to more efficient, powerful models that can learn and adapt in previously unimaginable ways.

However, the model remains under a non-commercial license, and as such, we will have to wait and see how this model can be incorporated into commercial applications. It is doubtful that multiple fully open-source similar models will be available before the end of 2023.

Automate your data curation, annotation and label validation workflows.

★★★★★ 4.8/5

Explore Our Product Suite

WRITTEN BY

Nikolaj Buhl

View more posts →

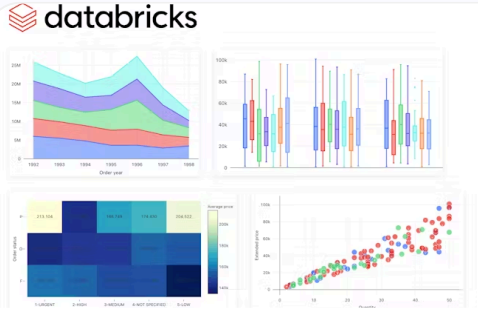
PREVIOUS BLOG

Top 8 Video Annotation Tools for

NEXT BLOG

How to Create Workflows in Encord

Related blogs



MACHINE LEARNING

Visualizations in Databricks

MACHINE LEARNING

Microsoft MORA: Multi-Agent Video Generation...

With data becoming a pillar stone of a company's growth strategy, the...

What is Mora? Mora is a multi-agent framework designed for generalist...

MAR 26 2024 8 M

Pa
En
MA

Software To Help You Turn Your Data Into AI

Forget fragmented workflows, annotation tools, and Notebooks for building AI applications. Encord Data Engine accelerates every step of taking your model into production.

Get started

Automation
API & Python SDK

Quality Assessment

Encord Active

Healthcare & Medical

Insurance

Life Sciences & Biotech

Logistics

Manufacturing

Media, Gaming & Entertainment

Retail & E-commerce

Sports

Technology & Software

Documentation
Glossary

Blog

Press

Pricing

Security

Subscribe

Get occasional product updates and tutorials to your inbox.

Your work email