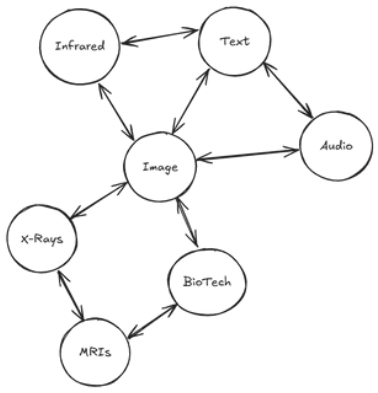
Bridging Modalities - 02

This poster explores the use of MLP heads as a post-processing method to refine modality alignment in ImageBind embeddings. We investigate whether adding small MLP transformations can help map modalities and help cross-modal retrieval without retraining the entire model.

Authors
Yann Baglin-Bunod
Michael Sheroubi

Methodology: MLP Heads for Modality Alignment

We introduce MLP heads as a lightweight post-processing method to refine modality alignment in ImageBind embeddings. Each modality (audio, image, text) has its own MLP head, allowing embeddings to be transformed in a circular manner (e.g., audio → image → text → audio). Unlike full model retraining, this method is computationally efficient and allows for fine-tuned modality adjustments while preserving the original latent space structure.



Experimental Setup:

Our setup is built on top of image bind, where we used the Omega Multimodal dataset containing precomputed embeddings from video, audio, and text modalities. Each embedding vector had 1024 dimensions, across all modalities. We split the dataset into 80% training and 20% testing sets.

We designed three separate MLP transformation heads and trained them simultaneously using a cosine embedding loss, to enable circular transformation between modalities:

- Audio → Text
- Text → Video
- Video → Audio

The Dataset

We use ImageBind-preprocessed embeddings, which provide triplets of audio, video, and text representations in a common space. This dataset allows us to focus on post-processing rather than feature extraction, enabling direct evaluation of alignment improvements. By testing MLP transformations on these embeddings, we analyze how effectively small neural networks can refine cross-modal relationships without full retraining.

Over 1 million hours of footage and 30 million+ 2-minute video clips, covering 50+ scenarios and 15,000+ action phrases.

Link to dataset:
huggingface.co/datasets/omegalabsinc/omega-multimodal

Link to embedding model:
github.com/facebookresearch/ImageBind

Fig 1 - Modality Bridge on Test Data

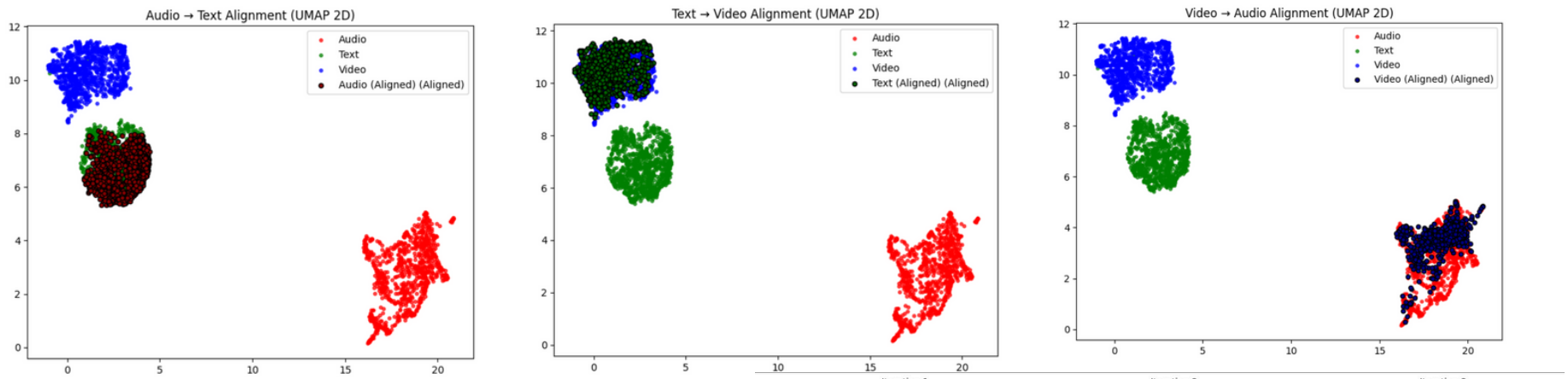


Fig 2 - Losses

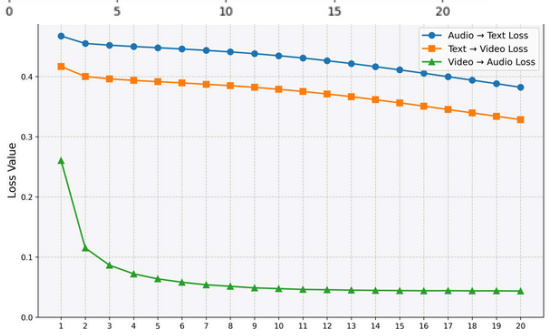
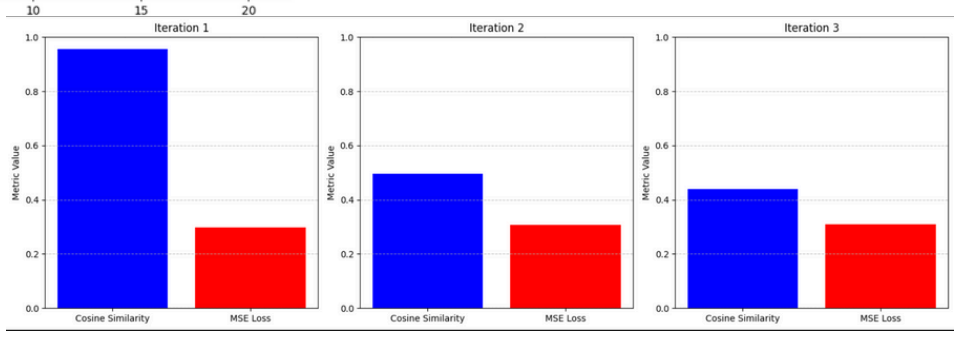


Fig 3 - Embedding Distortion Across MLP Iterations



Future Direction

First, we plan to incorporate more modalities beyond the current audio, text, and video, starting with modalities supported by ImageBind, then extending to more modalities like tabular.

Secondly, we want the MLP heads to be bi-directional, allowing easier graph traversal between modalities. Third, translate embeddings back to their original modality formats.

Finally, use transformer architectures to replace or augment the current MLP heads, as transformers will likely have better performance in capturing relationships between modalities.

Fig 4 - Focus on Video and audio