Bridging Modalities - 01

Bridging multimodal representation gaps to improve cross-modal alignment in AI systems.

Authors

Yann Baglin-Bunod

Michael Sheroubi

Methodology: MLP Heads for Modality Alignment

Multimodal models embed diverse data into a shared latent space, but modality gaps may impact alignment. Training these models is costly, so we first review the literature to assess the significance of this gap and potential solutions. Our focus is on understanding existing approaches and exploring feasible post-processing methods to improve cross-modal alignment.

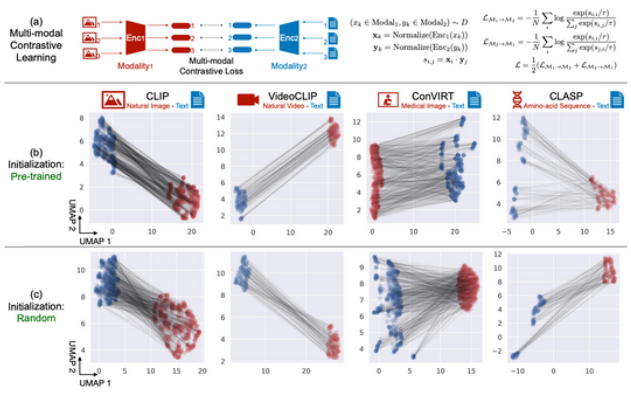
Objective

Our objective is to assess the impact of modality gaps in multimodal models by reviewing existing literature on cross-modal embeddings. Given the high computational cost of training these models, we explore potential post-processing methods, focusing on the feasibility of using MLP heads to refine alignment and improve the interaction between audio, image, and text embeddings.

Academic Context: Understanding Modality Gaps

Mind the Gap

Contrastive learning in multimodal models like CLIP maintains a modality gap, where embeddings cluster separately. This study identifies structural causes like the cone effect, where neural networks form narrow modality-specific regions. The authors show that contrastive loss fails to reduce this gap and that temperature scaling influences modality separation.



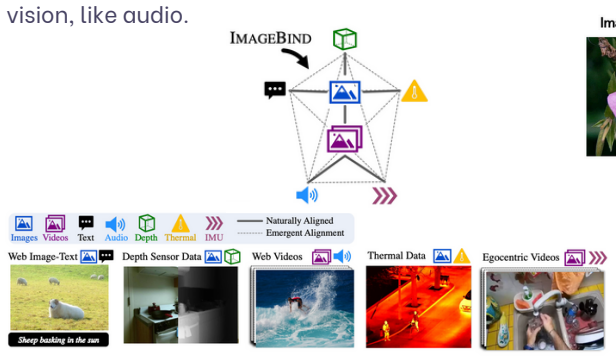
The Platonic Represenation Hypothesis

This hypothesis suggests that foundational multimodal models, despite being trained differently, converge toward a universal latent space. The study analyzes trends across vision, language, and audio models, proposing that a single optimal representation may exist. If true, this could indicate that post-processing methods, such as MLP transformations, might help refine embeddings toward this universal space.

Multimodal Models

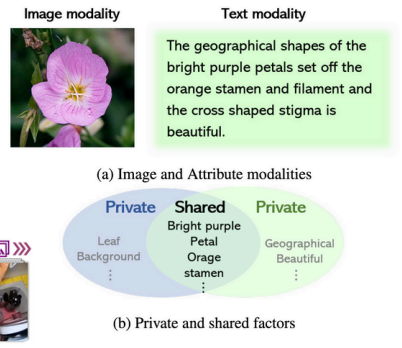
ImageBind

ImageBind extends CLIP's contrastive learning framework to six modalities (image, text, audio, depth, thermal, IMU). By leveraging image-paired data, it aligns all modalities into a shared space without direct pairwise training. However, residual modality gaps remain, especially for modalities less tightly linked to vision, like audio.



DMVAE

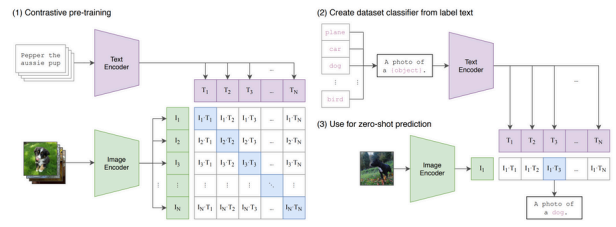
DMVAE proposes a disentangled latent space where embeddings are divided into shared (cross-modal) and private (modality-specific) representations. This structure allows each modality to retain unique characteristics while still aligning with others. However, it requires paired multimodal data, making it less flexible for datasets with missing modalities.



Multimodal Models

CLIP

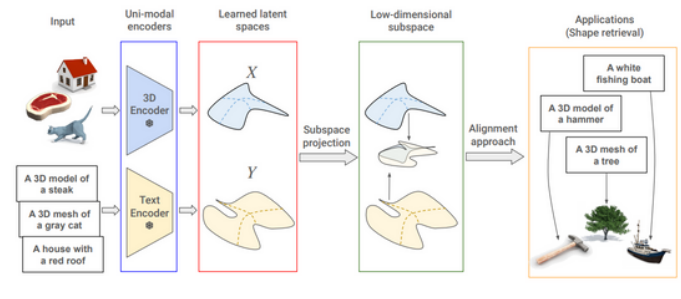
CLIP introduced contrastive learning for image-text alignment, achieving impressive zero-shot capabilities. However, it struggles with hierarchical semantics, where embeddings fail to capture structured relationships across modalities. Its reliance on paired image-text data also limits generalization to unseen modalities.



Post-processing Methods

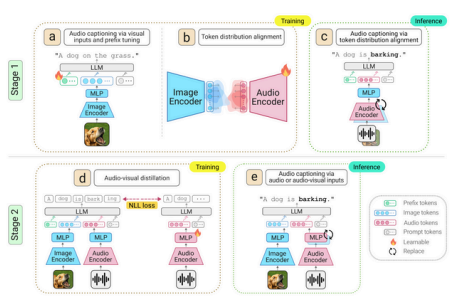
Escaping Plato's Cave

This work investigates 3D-text embedding alignment, showing that naïve feature alignment fails when embeddings are trained independently. The authors use Canonical Correlation Analysis (CCA) and Procrustes Alignment to project embeddings into a shared subspace, improving cross-modal retrieval without requiring joint pretraining. Their method demonstrates that post-hoc transformations can effectively refine multimodal alignment in a computationally efficient way.



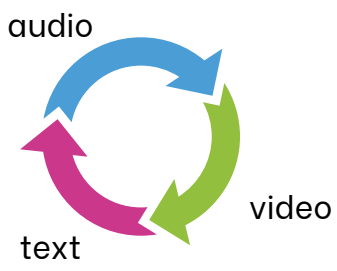
An Eye for an Ear

This work aligns audio and image embeddings using a two-stage learning approach. By training a small transformation network, it enables zero-shot audio captioning. However, results are dataset-dependent, and the method does not generalize well to more than two modalities.



Proposed Methodology

We propose using MLP heads as a lightweight post-ImageBind processing method to enable circular transformation between modalities (audio → image → text → audio). This approach refines cross-modal alignment, enhances retrieval consistency, and offers a low-cost adaptation strategy without retraining the entire model. If successful, it suggests that modality gaps can be bridged with simple, learnable transformations



Literaeture References

CLIP (Radford et al., 2021) – Contrastive learning for image-text alignment.
ImageBind (Girdhar et al., 2023) – Unified multimodal embeddings.
DMVAE (Lee & Pavlovic, 2021) – Disentangled shared-private latent spaces.
Mind the Gap (Liang et al., 2022) – Structural modality gaps in contrastive learning.
Escaping Plato's Cave (Hadgi et al., 2025) – Aligning 3D and text embeddings via subspace projection.
An Eye for an Ear (Malard et al., 2024) – Zero-shot audio captioning via audiovisual alignment.
The Platonic Representation Hypothesis (Huh et al., 2024) – Investigating universal multimodal representations.