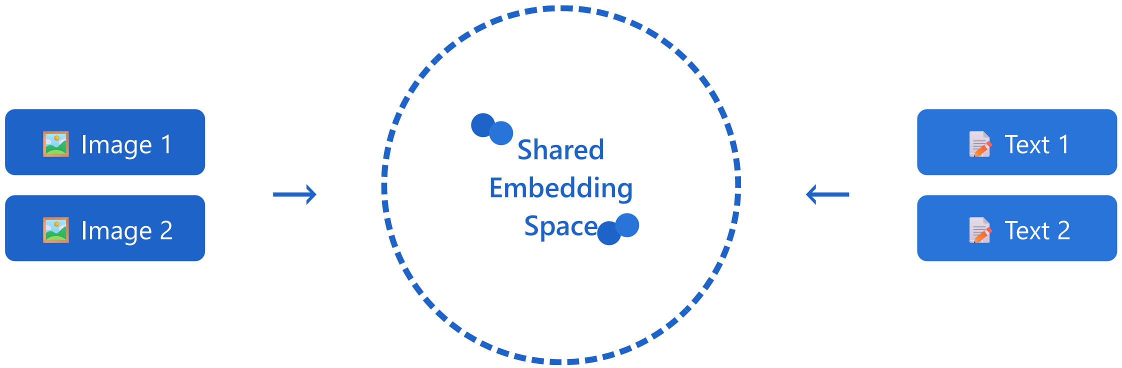
Cross-Modal Alignment: Inter-Modal Alignment

By placing different modalities at **semantically close positions** in a **shared embedding space**, mutual understanding and retrieval become possible


🖼️ Image 1

🖼️ Image 2

→

Shared Embedding Space

←

📝 Text 1

📝 Text 2

Projection Layers

Linear/non-linear transformations that project each modality into a shared space of the same dimension

- Dimension matching with linear layers
- Non-linear transformation with MLP
- Placement on unit sphere via L2 normalization

Contrastive Loss

Improves alignment quality by learning to bring matching pairs closer and push non-matching pairs apart

- Using InfoNCE Loss
- Applying temperature scaling
- Hard negative mining

Triplet Loss

Cross-Attention

Learning relative distances with Anchor, Positive, and Negative samples

- Same patient data: Positive
- Different patient data: Negative
- Margin-based distance optimization

Enhancing inter-modal interaction through Transformer-based attention

- Query-Key-Value mechanism
- Information exchange between modalities
- Dynamic weight learning

Zero-shot Classification

Text→Image Retrieval

Image→Text Retrieval