Unified Multimodal Learning and Confidence Calibration

Abstract & Problem Statement : Multimodal learning models traditionally suffer from **image-centric bias** and limited text semantics. For example, CLIP and ImageBind align modalities via images as the central reference, leading to unbalanced representations. Text labels in such systems are often just single words (e.g. "dog", "airplane"), which are **weak semantically** and fail to capture rich contextual meaning. Moreover, existing models typically **lack calibrated confidence** in their predictions – they tend to be overconfident in their outputs, with probability scores not reflecting true accuracy. The UNIBIND project addresses these issues by learning a unified, modality-agnostic embedding space with improved semantic grounding and confidence reliability.

Introduction : **Unified multimodal learning** aims to map diverse data (images, audio, video, text, etc.) into a common representation space. Prior multimodal models like CLIP and ImageBind have made progress by training on image–text or image–others pairs, but they inherently center everything on images, resulting in an unbalanced emphasis on visual features. This image-centric design can bias the representation space, causing other modalities (e.g. audio or depth) to be tethered to visual semantics rather than treated equally. Furthermore, using only short text labels (e.g. *"dog"*, *"airplane"*) for supervision provides **weak semantic grounding**, as single words cannot capture the full diversity of each category. These limitations motivate a new approach: **UNIBIND (Unified Multimodal Learning using Contrastive Embeddings)**, which seeks to create one unified and balanced embedding space for up to seven modalities. In addition, UNIBIND recognizes the importance of well-calibrated confidence estimates.

Methodology : UNIBIND employs a three-stage pipeline to learn unified, modality-agnostic representations using large language models (LLMs):

Stage 1: Knowledge Base Construction – LLMs (e.g. GPT-4, BLIP-2) generate ~1000 rich descriptions per class, forming a semantic knowledge base. These serve as anchors to ground multimodal inputs in a shared semantic space.

Stage 2: Unified Representation Learning – Pre-trained modality-specific encoders (e.g. for image, audio, video) are frozen; lightweight adapters are trained using contrastive loss to align modality embeddings with text-based anchors. This ensures balanced, modality-agnostic alignment with 90% fewer trainable parameters.

Stage 3: Embedding Center Localization (ECL) – For each class, the top-k (e.g. 50) LLM-generated descriptions closest to a canonical prompt define the semantic center. Inputs are classified via cosine similarity to these centers.

Confidence Estimation: Calibrated confidence scores are derived using softmax with temperature scaling. Metrics like Expected Calibration Error (ECE) and reliability diagrams ensure confidence reflects accuracy. Entropy is used to measure uncertainty.

Fallback Mechanisms: When confidence is low (e.g. <60%), CLIP-based VQA serves as a zero-shot validator. It answers open or yes/no questions to verify predictions across modalities. For captions, BLIP-2 scores are fused with CLIP similarity to estimate confidence.

Challenges Faced : UNIBIND faced several notable challenges during development. The computational cost was significant, as generating extensive prompt data and training across seven modalities demanded high GPU resources and long runtimes. System integration was also complex combining multiple pre-trained encoders (e.g., for image, audio, video) with LLMs and calibration tools required careful setup and dependency management. Incorporating the audio modality posed additional difficulties due to its inherent noise and variability, making it harder to align with other modalities. Moreover, the model's performance was highly sensitive to the quality and phrasing of LLM-generated prompts; small changes could impact alignment, necessitating prompt regulation to ensure stability.

Results : UNIBIND significantly improved multimodal learning by leveraging rich LLM-generated descriptions. Compared to models using simple prompts, it achieved up to 6.5% higher classification and retrieval accuracy across modalities reaching 83% accuracy versus 77–78% in baselines. Confidence calibration also improved. With temperature scaling, the Expected Calibration Error (ECE) dropped. and reliability diagrams showed better alignment between predicted confidence and actual accuracy.Coupled with fallback VQA and captioning checks, the system provided robust, trustworthy outputs.

Future Enhancements : Future enhancements for UNIBIND focus on improving transparency and scalability. Interpretability will be addressed by incorporating attention maps and feature alignment tools to visualize how different modalities contribute to predictions, enabling better understanding of **cross-modal relationships** (e.g., why a specific audio sample aligns with certain visual features). Additionally, the architecture will be extended to support multi-label classification, allowing the model to recognize and differentiate multiple concepts within a single input by adapting its training and inference strategies to handle overlapping class centers. These advancements aim to make UNIBIND more robust, flexible, and applicable to complex, real-world multimodal tasks.