The Role of Contrastive Learning in Multimodal Large Language Models (MLLMs)

Aditi Sarker

Department of Computer Science

Wayne State University

hq1351@wayne.edu

Date: 03-07-2025

Ph.D. Advisor: Professor Prashant Khanduri

Abstract

Multimodal Large Language Models (MLLMs) have advanced AI by allowing systems to process numerous data types such as text, images, audio, video, and so on. These models enhance human-computer interactions, supporting applications such as image captioning, visual question answering, and multimodal assistants. However, ensuring seamless alignment between different modalities remains a challenge. Contrastive learning, a self-supervised technique, effectively bridges this gap by aligning related data samples while distinguishing unrelated ones, improving representation quality, zero-shot learning, and cross-modal alignment. Leading MLLMs, including CLIP, BLIP-2, MM1, and OpenFlamingo, leverage contrastive learning to enhance multimodal task performance. This survey provides an overview of contrastive learning in MLLMs, examining its principles, implementation in leading models, and key advantages. We discuss computational costs, data requirements, and challenges while highlighting future research directions for improving its efficiency.

Literature survey. Winter 2025 PhD qualifying exam.

Contents

1 Introduction

Multimodal Large Language Models (MLLMs) have emerged as a crucial advancement in artificial intelligence, enabling AI systems to process and integrate information across multiple input modalities [1]. These models facilitate various applications, including image captioning, video understanding, speech-to-text conversion, and multimodal reasoning, marking a transformative shift from unimodal language models that rely solely on textual data [2]. One of the key challenges in MLLMs is ensuring effective alignment between different modalities due to their diverse structures and distributions [3].

Contrastive learning, a self-supervised learning approach, has appeared as an effective solution to this difficulty, efficiently discriminating between comparable and dissimilar data representations. This technique aligns multimodal data in a unified embedding space, enabling MLLMs to achieve superior performance in various multimodal tasks [4, 5, 6]. This approach has revolutionized the way MLLMs understand and generate content across multiple data types, leading to substantial improvements in various tasks [5, 7]. At its core, multimodal contrastive learning aims to generate a unified embedding space where representations of several modalities are aligned [6]. This is typically achieved by training the model to maximize the similarity between paired inputs (e.g., an image and its corresponding caption) while minimizing the similarity between unrelated inputs [5, 6]. The capacity of contrastive learning to acquire meaningful representations without the need for explicit labels is one of its main benefits in multimodal environments [6]. Instead, it leverages the natural correspondence between distinct modalities, such as images and their descriptions, to guide the learning process [5] and studies have shown that incorporating contrastive learning alongside next-token prediction during pretraining can boost MLLMs performance by approximately 2% across various multimodal evaluations without additional compute or training [7].

Current research has illustrated the effectiveness of contrastive learning in multimodal models. For instance, the Contrastive Language-Image Pre-training (CLIP) [1] model has shown remarkable zero-shot capabilities by learning to align visual and textual representations [6]. This model has set new benchmarks in distinct multimodal tasks and has been widely adopted in the field. Building upon the success of CLIP, the Language-Image Mixture of Experts (LIMoE) [8] model employs contrastive learning to achieve state-of-the-art performance in multimodal tasks. LIMoE outperforms dense models of equivalent computational cost, highlighting the efficiency and effectiveness of contrastive learning approaches in multimodal settings. Furthermore, the work on Multimodal Mixup Contrastive Learning (M3CoL) [9] has shown promising results in capturing nuanced shared relations inherent

in multimodal data. This approach goes beyond simple pairwise associations, enabling MLLMs to understand more complex interactions between different modalities.

As the field of multimodal AI continues to evolve, contrastive learning remains a crucial component in developing more sophisticated and capable MLLMs [5, 7]. Its ability to create meaningful cross-modal representations paves the way for advanced applications in areas such as visual question answering, image captioning, and multimodal content generation.

A notable challenge in multimodal AI is the issue of hallucination, where models generate fabricated or incorrect information that does not align with visual inputs. The LLAVA [10] model (Language and Vision Alignment with Attention) is particularly susceptible to hallucination due to factors such as the modality gap between textual and visual representations, over-reliance on language priors, and a lack of fine-grained visual understanding. This often leads the model to produce outputs grounded more in language context rather than visual content, resulting in responses that fail to accurately reflect the associated image. Additionally, inadequate multimodal alignment during training and limited use of hard negative samples further exacerbate the risk of hallucination in LLAVA.

Recent advancements in this field include the introduction of Hallucination Augmented Contrastive Learning (HACL), which addresses the issue of hallucinations in MLLMs by using hallucinative text as hard negative samples in the contrastive learning process [5]. This method has demonstrated significant advancements in decreasing hallucination occurrences and enhancing performance across multiple benchmarks. Another notable development is the Img-Diff dataset [7], which leverages insights from contrastive learning and image difference captioning to improve MLLM's ability to recognize fine-grained images. This method requires models to recognize both matching and unique components in similar images, leading to comprehensive improvements in performance scores over state-of-the-art models in numerous image difference and Visual Question Answering tasks [7]. The ongoing research in this area promises to further enhance the capabilities of MLLMs, pushing the boundaries of what is achievable in multimodal artificial intelligence [11].

This review dives into the integration of contrastive learning in MLLMs, addressing its key ideas, pretraining tactics, loss functions, and data augmentation methodologies. Additionally, it presents case studies that highlight the impact of contrastive learning on MMLLM performance, along with the challenges and future research directions in this domain. By understanding the role of contrastive learning in MLLMs, we can unlock new possibilities for AI systems that exhibit advanced reasoning, adaptability, and generalization capabilities across diverse data sources.

The structure of this review is organized into five main sections. **Section 2** provides an in-depth discussion of the fundamentals and techniques of contrastive learning, covering key methodologies such as loss functions, pair sampling strategies, and well-known models like SimCLR, MoCo, and CLIP. **Section 3** explores Multimodal Large Language Models (MLLMs) and training strategies, detailing general architectures, challenges in multimodal integration, and the role of contrastive learning in overcoming these challenges. **Section 4** delves into Contrastive Learning in MLLMs, analyzing its applications, challenges, and evaluation methods, with case studies on models like CLIP, ALIGN, and Flamingo. **Section 5** outlines future directions, discussing ongoing and proposed advancements in contrastive learning for multimodal AI. Finally, the paper concludes with insights into prospective research opportunities and the evolving landscape of MLLMs, emphasizing scalability, efficiency, and ethical considerations.

2 Fundamentals of Contrastive Learning

Contrastive learning is an effective machine learning technique that aims to learn meaningful representations by contrasting pairs of data points that are either similar (positive pairs) or dissimilar (negative pairs) [4]. The core concept is to bring similar instances closer in the learned embedding space while separating dissimilar ones. Key concepts include:

- Embedding Space: A low-dimensional representation where data points are mapped such that the geometry reflects similarity relationships.
- Positive Pairs: Similar data pairs are frequently produced by applying several augmentations to a single data point.
- Negative Pairs: Data pairs considered dissimilar, typically consisting of different data points within a batch.
- Encoder: A neural network that maps input data into the embedding space.

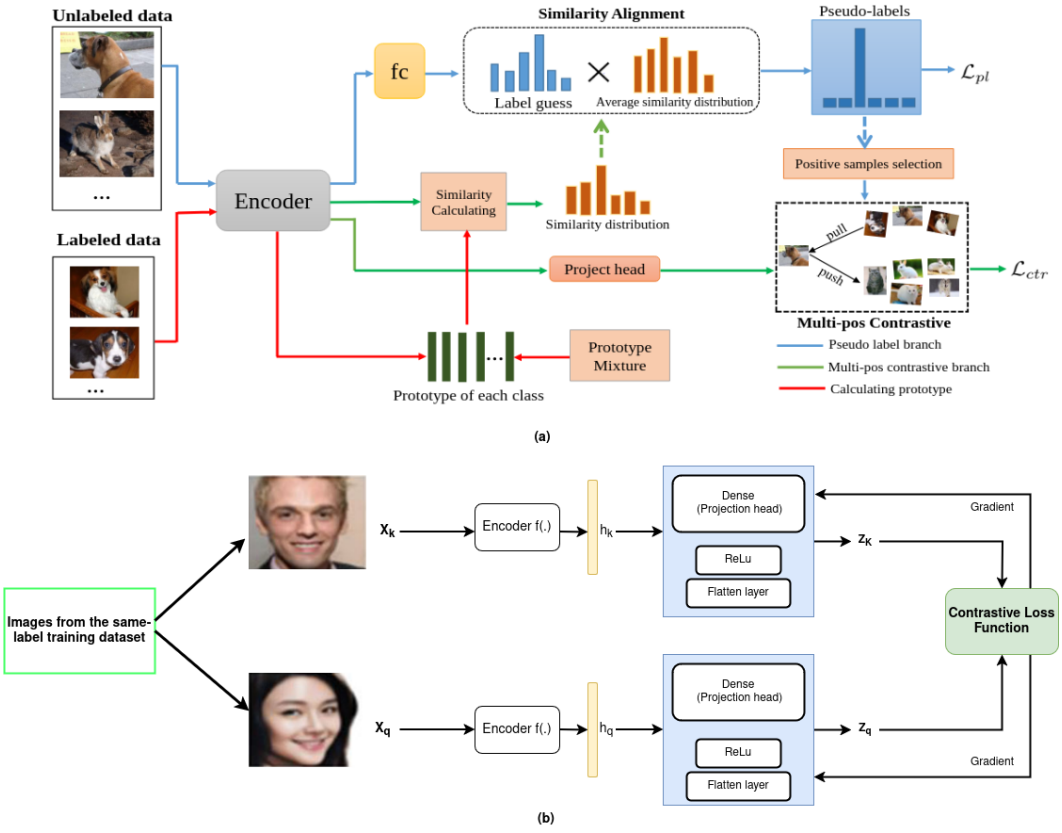


Figure 1: Overview of Contrastive Learning (a) Self-supervised contrastive learning[12] (b) supervised contrastive learning [13].

Contrastive learning can be implemented in two main ways: Supervised Contrastive Learning (SCL) [13] and Self-Supervised Contrastive Learning (SSCL) [12], differing primarily in how they define positive and negative pairs. Figure 1 depicts the architecture of Supervised and Self-Supervised Contrastive Learning. Supervised Contrastive Learning (SCL) employs data with labels to explicitly identify positive and negative pairs, gathering together instances of the identical class while pushing apart distinct classes. It is particularly effective for classification problems. A notable example is CLIP, which learns visual representations by linking images with text. In contrast, Self-Supervised Contrastive Learning (SSCL) eliminates the need for labels by generating positive pairs through augmentations of the same instance and negative pairs from different samples. This approach enables models to learn high-level semantic representations, as seen in SimCLR, MoCo, BYOL. While SCL relies on explicit supervision, SSCL learns directly from data structure, making it more scalable for large, unlabeled datasets.

2.1 Positive and Negative Pair Sampling

In contrastive learning, selecting positive and negative pairs is crucial for effective representation learning.

Positive Pair Sampling: Usually, data augmentation methods like Gaussian blur, color jittering, and random cropping are used to produce positive pairs. These transformations help the model learn invariant representations by reinforcing robustness to minor variations. Recent work, such as "Rethinking Positive Pairs in Contrastive Learning," suggests learning from arbitrary pairs rather than relying strictly on augmentation-based methods [14].

Negative Pair Sampling: Negative pairs contrast a given positive sample with unrelated samples. Effective negative sampling strategies include:

- Random sampling: Randomly selects negative samples from the dataset, ensuring diversity but often leading to "easy negatives" that provide limited learning benefits.
- In-Batch Negatives: Uses other samples within the same mini-batch as negatives, reducing computational cost while enhancing contrastive learning effectiveness. However, it may introduce false negatives if similar-class samples are mistakenly treated as negatives.
- Hard Negative Mining: Prioritizes negatives that are semantically similar to the anchor while belonging to a different class, making the task more challenging and improving

generalization. This method requires similarity computations and ranking, increasing computational cost. To balance difficulty and reliability, a principled hard negative sampling strategy has been proposed to optimize contrastive learning [15].

A key challenge in negative sampling is mitigating false negatives—cases where semantically similar samples are incorrectly labeled as negatives. Methods such as "Contrastive Learning with Negative Sampling Correction" address this by treating generated negatives as unlabeled data and applying corrective mechanisms [16]. By balancing diverse negative sampling strategies, contrastive learning models achieve more informative feature representations, enhancing performance across multimodal tasks.

2.2 Contrastive Learning Loss Functions

In contrastive learning, selecting a loss function is critical for efficiently learning representations by differentiating between similar and dissimilar data points. The basic goal in the embedding space is to reduce the distance between positive pairings (similar samples) while increasing the distance between negative pairs (dissimilar samples).

Contrastive Loss

Contrastive Loss [17] operates on pairs of samples, bringing together comparable embeddings and pushing apart dissimilar ones. It is frequently specified with a buffer to keep the model from collapsing the embedding space. The loss function is defined as follows:

$$L = (1 - Y) \cdot \frac{1}{2} D^2 + Y \cdot \frac{1}{2} \left(\max(0, m - D) \right)^2, \tag{1}$$

where Y is a binary label indicating whether the pair is similar (0) or dissimilar (1). D represents the Euclidean distance between the embeddings of the pair and m is the margin parameter that defines the minimum distance between dissimilar pairs.

Triplet Loss

Triplet Loss [18] extends the concept by considering triplets of samples: an anchor, a positive sample (same class as the anchor), and a negative sample (different class). The goal is to ensure that the anchor is closer to the positive than to the negative by at least a margin. The triplet loss function is defined as:

$$L = \sum_{i}^{N} \left[\|f(x_i^a) - f(x_i^p)\|_2^2 - \|f(x_i^a) - f(x_i^n)\|_2^2 + \alpha \right]_+ \tag{2}$$

where $f(x)$ denotes the embedding function,x_i^a, x_i^p, x_i^n are the anchor, positive, and negative samples, respectively. α is the margin parameter.

InfoNCE Loss

InfoNCE (Information Noise-Contrastive Estimation) loss is the most popular technique used in contrastive learning frameworks [19]. It encourages the model to assign higher similarity scores to positive pairs compared to negative pairs within a batch. The InfoNCE loss function is defined as:

$$L = -\sum_{i=1}^{N} \log \frac{\exp\left(\text{sim}(h_i, h_i^+)/\tau\right)}{\sum_{j=1}^{N} \exp\left(\text{sim}(h_i, h_j^-)/\tau\right)} \tag{3}$$

where h_i and h_i^+ are the embeddings of the anchor and positive samples. h_j^- are the embeddings of negative samples. sim denotes a similarity function, such as cosine similarity, and τ is a temperature parameter that scales the logits. The InfoNCE loss aims to maximize the similarity between positive pairs (e.g., an image and its corresponding caption) while minimizing the similarity between negative pairs (unrelated images and captions). This helps the model develop more robust and meaningful representations across several modalities by creating a uniform embedding space where similar samples are closer together.

2.3 Contrastive Learning Models (SimCLR, MoCo, CLIP)

Contrastive learning has resulted in the generation of several prominent models, each employing unique strategies for sampling positive and negative pairs to learn effective representations. Below is a detailed overview of prominent models, and the summary of contrastive learning models is depicted in table 1.

SimCLR (Simple Framework for Contrastive Learning of Visual Representations): SimCLR, proposed by [4], is a self-supervised learning framework that leverages contrastive learning to learn visual representations without labeled data. The key idea behind SimCLR is to apply a series of data augmentations such as random cropping, color jittering, and Gaussian blur to the same image to generate positive pairs, ensuring the model learns invariant features to these transformations. Negative pairs are formed by comparing different images within the same mini-batch, known as

in-batch negatives, necessitating large batch sizes to maximize the diversity of negative examples. To improve representation quality, SimCLR introduces a nonlinear projection head that maps representations to a contrastive loss space, helping the model learn more discriminative features. The study demonstrated that larger batch sizes and stronger augmentations significantly enhance the effectiveness of contrastive learning. Figure 2, depicts the architecture of SimCLR.

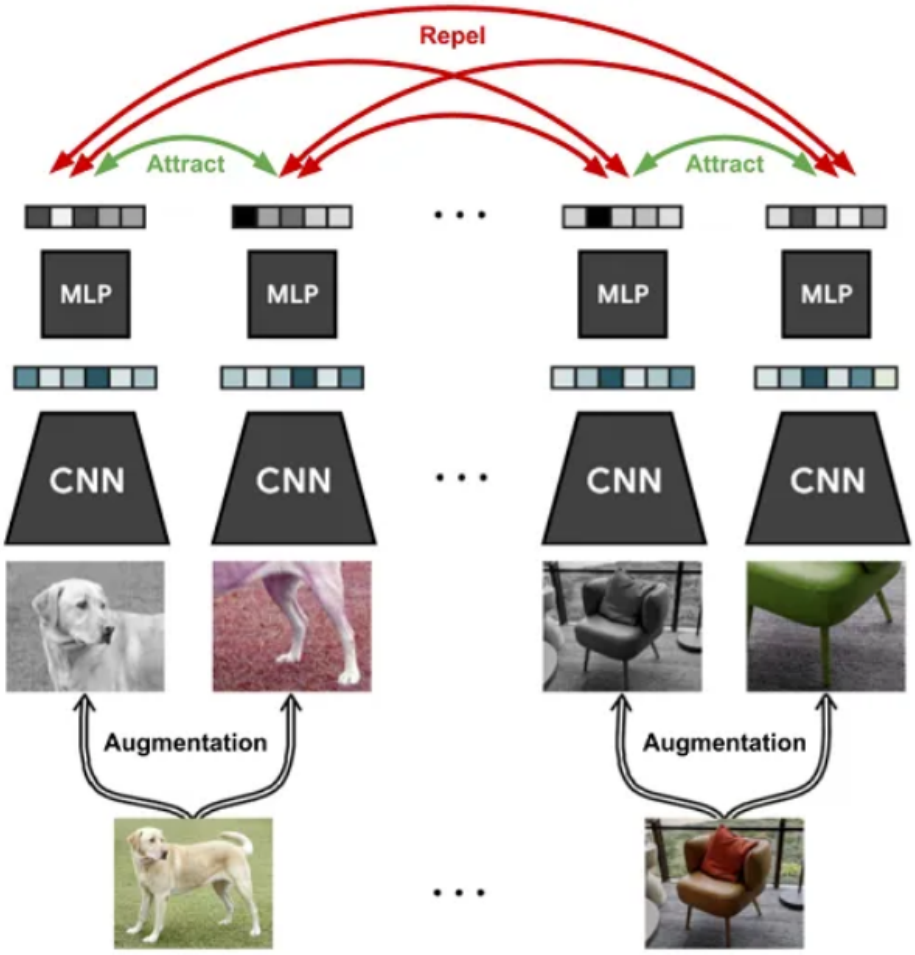


Figure 2: SimCLR: A simple framework for contrastive learning of visual representations [4].

MoCo (Momentum Contrast): MoCo , introduced by [20], is designed to overcome the limitation of large batch sizes required in SimCLR by maintaining a dynamic queue of negative samples. Like SimCLR, MoCo creates positive pairs using data augmentations applied to the same image but differs in its approach to negative pair sampling. Instead of relying solely on in-batch negatives, MoCo maintains a large queue of negative examples and updates it dynamically using a momentum encoder. This design allows MoCo to utilize a large and consistent set of negatives over time without requiring large batch sizes, making it more memory efficient. The momentum encoder ensures stable feature

representations, reducing inconsistency between different training steps. These innovations make MoCo a scalable and efficient approach for self-supervised learning. Figure 3 depicts the architecture of MoCo.

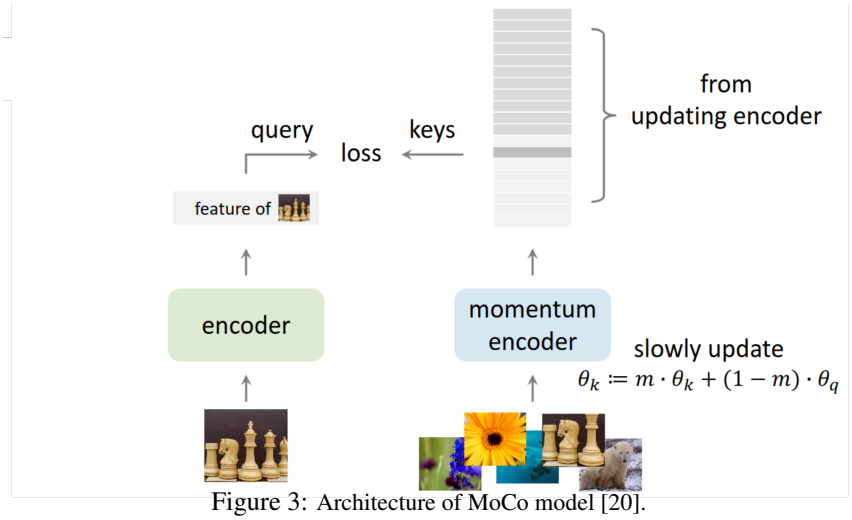


Figure 3: Architecture of MoCo model [20].

CLIP (Contrastive Language-Image Pre-training):

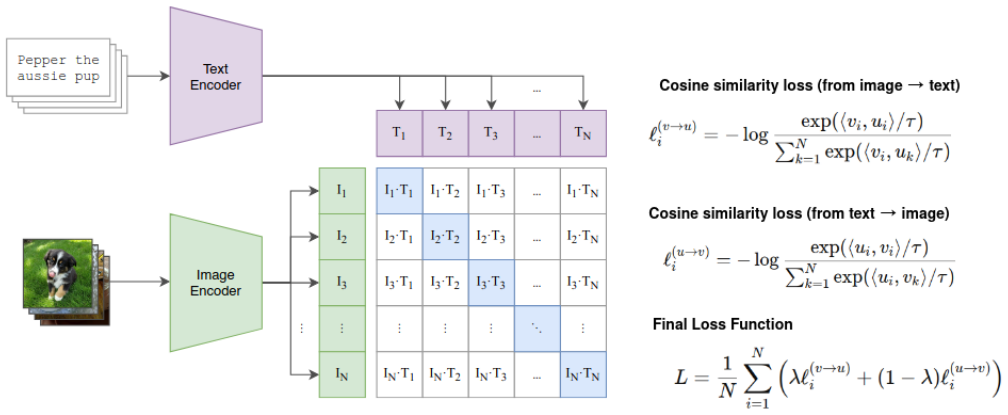


Figure 4: Architecture of CLIP model [1].

Figure 4, depicts the architecture of CLIP, a model developed by OpenAI [1] that extends contrastive learning beyond visual representations to align images and text within a shared embedding space. Unlike SimCLR and MoCo, which focus solely on image-based contrastive learning, CLIP learns to associate images with their corresponding textual descriptions. Positive pairs consist of an image and its correct caption, while negative pairs are created by pairing an image with mismatched captions. This contrastive training approach enables CLIP to develop a strong understanding of visual and textual relationships, facilitating zero-shot learning, where the model can perform new tasks without

explicit training. By training on a massive dataset of image-caption pairs, CLIP achieves impressive generalization across diverse vision-language tasks.

Model	Modality	Negative Sampling Strategy	Zero-Shot Learning
SimCLR	Image	In-batch negatives, large batch size	×
MoCo	Image	Momentum encoder, dynamic queue	×
CLIP	Image + Text	Contrastive image-text pairs	✓

Table 1: Summary of Contrastive Learning Models.

Table 1, summarizes the contrastive learning models, highlighting that among SimCLR, MoCo, and CLIP, CLIP is the most essential for Multimodal Large Language Models (MLLMs) due to its ability to establish a shared embedding space between images and text, whereas the other models focus exclusively on visual representations. CLIP's contrastive learning on large-scale image-text datasets allows for zero-shot classification, cross-modal retrieval, and multimodal reasoning, making it foundational for MMLLM architectures [3]. Inspired by CLIP, Netflix adopted similar methods to enhance video-text search, while CLAP extended the approach to text-audio learning for improved speech and sound-based applications [21]. Expanding on this, Meta AI's ImageBind introduced a unified embedding across six modalities—images, text, audio, depth, thermal, and IMU data—demonstrating CLIP's influence on multimodal AI [22]. This evolution has shaped advanced MLLMs like GPT-4V, PaLM-E [23], and Flamingo [24], confirming CLIP as the foundation of multimodal AI.

3 Multimodal Large Language Models and Training Strategies

Multimodal Large Language Models (MLLMs) are artificial intelligence models designed to process and integrate multiple input modalities, such as text, images, audio, and video, to perform complex reasoning and generation tasks [25]. Unlike traditional language models that rely solely on textual data, MLLMs extend their capabilities by learning from diverse input sources, enabling tasks such as image-text understanding, video-text comprehension (e.g., scene recognition and video summarization), and audio-text processing (e.g., speech recognition and emotion analysis). Additionally, these models facilitate multimodal generation, where they can generate textual content from visual or auditory inputs. The fundamental principle behind MLLMs is to unify different modalities within a common representational space, enabling seamless reasoning and decision-making across multiple data types. This unification allows MLLMs to provide richer contextual understanding and improved performance on tasks requiring cross-modal comprehension.

3.1 General Architectures of MLLMs

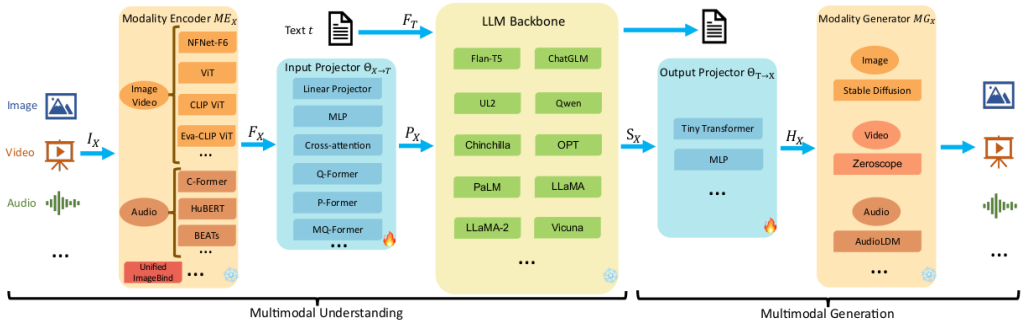
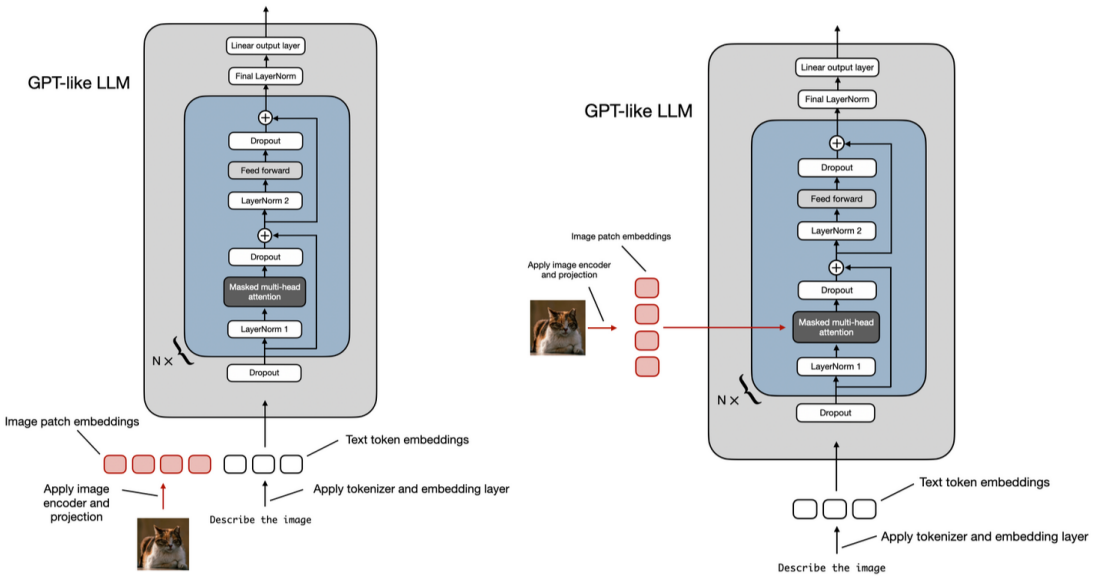


Figure 5: The general model architecture of MLLMs and the implementation choices for each component [26]

Figure 7, shows the general architecture of MLLMs, it is a highly intricate system that integrates various specialized components to enable comprehensive multimodal understanding and generation. It begins with **modality encoders**, such as NFNet-F6 [27], CLIP ViT [1], VideoMAE [28], and HuBERT [29], which extract high-level feature representations from images, videos, and audio. These encoders capture semantic, spatial, temporal, and phonetic structures, ensuring rich feature extraction for different modalities. The extracted representations are then aligned through an **input projector**, employing MLPs, linear projections, and cross-attention mechanisms like Q-Former [30], P-Former, and MQ-Former [31], which transform modality-specific embeddings into a unified representation.

Contrastive learning plays a critical role in multiple stages of MLLM processing by ensuring effective alignment between different modalities. In modality encoders, contrastive loss functions help map images, audio, and text into a shared embedding space (e.g., CLIP aligns images and text using contrastive learning). In the input projector, contrastive learning helps enforce structural similarity across modalities, ensuring effective fusion before passing embeddings to the LLM backbone. Additionally, contrastive pretraining enhances zero-shot and few-shot learning, allowing models like OpenFlamingo and ALIGN to perform retrieval, reasoning, and generation tasks more effectively. By integrating contrastive learning, MLLMs achieve better multimodal representation learning, improved generalization, and enhanced cross-modal understanding.

The LLM Backbone, composed of transformer-based models such as GPT-4 [32], LLaMA-2 [33], and PaLM [34], processes these multimodal embeddings, leveraging extensive datasets for reasoning and content generation. Following this, an **output projector** applies lightweight transformers, MLPs, or attention-based refinements [35] to adjust the outputs for specific tasks. Finally, the Modality Generator, consisting of Stable Diffusion [36], and AudioLDM [37], transforms refined representations into high-quality outputs such as images, videos, or audio. This modular pipeline enables MLLMs to perform advanced multimodal reasoning, making them effective for diverse applications in AI-driven content creation and interactive systems. There are two basic techniques to developing multimodal LLMs, as shown in figure 6.



(a) Unified Embedding Decoder Architecture (b) Cross-Modality Attention Architecture

Figure 6: The two main approaches to developing multimodal LLM architectures

Unified Embedding Decoder Architecture: This method, like other LLM systems like GPT-2 or Llama 3.2, has a single decoder model. After concatenation, the LLM may process both text and image input tokens concurrently because images are transformed into tokens with the same embedding size as the original text tokens.

Cross-modality Attention Architecture: This technique integrates text and image embeddings directly within the attention layer via a cross-attention mechanism [38].

These architectural approaches aim to effectively combine different modalities, allowing MLLMs to process and generate content across various data types [25].

3.2 Challenges in Multimodal Integration

Multimodal integration enables AI systems to process various data types, improving performance in applications like healthcare and autonomous systems [39]. However, several challenges hinder seamless integration.

- Heterogeneity of Modalities: Different data modalities vary in structure and representation, making integration complex [40]. For instance, images use pixel matrices, while text relies on sequential embeddings. Methods like contrastive learning (CLIP) and cross-attention architectures (Flamingo) help in feature alignment across modalities [1].

- Data Synchronization and Alignment: Aligning multimodal data streams is crucial, as errors in synchronization lead to poor predictions [41]. Real-time applications like autonomous driving require precise fusion of camera and LiDAR data.Techniques like neural architecture search (AutoML) and temporal transformers (TimeSformer) enhance synchronization [39].

- Dimensionality and Computational Complexity: Multimodal data is often high-dimensional, increasing computational costs [42]. Efficient architectures and optimization techniques are necessary.Efficient architectures like MiniGPT-4 and parameter-efficient tuning (LoRA, PEFT) reduce computational overhead [43].

- Feature Fusion Strategies: Early, late, and hybrid fusion techniques must effectively combine multimodal features without redundancy [44]. Poor fusion can degrade model accuracy.Tensor fusion networks (TFN) and transformer-based architectures (MM-T5) improve multimodal representation [45].

- Data Imbalance and Missing Modalities: Many real-world datasets suffer from missing or imbalanced modalities, requiring imputation or attention-based techniques [46]. Self-supervised learning (SimCLR) and weakly supervised pretraining (GPT-4V caption generation) help address data limitations [4].

- Interpretability and Explainability: Understanding how multimodal models make decisions remains challenging [47]. Explainability is critical for trust in domains like healthcare and finance. Bias calibration techniques (MM-Guard) and fairness-aware training (RLHF) mitigate interpretability issues [48].

Addressing these challenges requires advancements in representation learning, data alignment, and interpretability methods [4]. Progress in these areas will enable robust multimodal AI applications.

3.3 Role of Contrastive Learning in Overcoming the Challenges

Contrastive learning helps address modality misalignment by enforcing feature similarity across different modalities while distinguishing unrelated samples [1]. By training models to associate similar instances while differentiating distinct ones, contrastive learning enhances the effectiveness of multimodal embeddings. Models like CLIP and ALIGN create joint embeddings for image-text pairs, improving zero-shot classification and retrieval tasks [3]. This ensures semantically related representations remain close in the latent space, leading to better transfer learning performance and robust domain adaptation, crucial for dynamic environments like autonomous driving and medical diagnosis [4].

By leveraging self-supervised contrastive learning, multimodal AI models extract meaningful patterns without requiring extensive labeled data, thereby reducing dependency on supervised datasets [47]. This improves data efficiency, allowing models to generalize better across unseen domains and strengthening the robustness of multimodal AI systems.

Multimodal Large Language Models (MLLMs) benefit from contrastive learning as it enables better feature alignment across diverse data types (text, images, audio) without relying on extensive annotations [1]. This is particularly useful for large-scale multimodal datasets, which often contain noisy, imbalanced, or missing data. By enforcing invariant representations across modalities, contrastive learning ensures stable performance in complex multimodal scenarios.

A significant advantage of contrastive learning is its ability to enhance zero-shot and few-shot learning. Models like CLIP and ALIGN perform well across unseen data distributions without requiring extensive fine-tuning [3]. Additionally, contrastive learning bridges modality gaps, aligning structurally different data types (e.g., text and images) into a common representational space [29]. This enhances cross-modal understanding, essential for applications like multimodal question answering and interactive AI systems.

Furthermore, contrastive learning enables few-shot learning, improving low-resource language processing, medical imaging analysis, and other data-limited applications [47]. By reducing reliance on expensive labeled datasets, contrastive learning promotes efficient training strategies, making it a vital component of next-generation AI research. Overall, contrastive learning enhances adaptability, efficiency, and robustness in multimodal LLMs, ensuring better generalization and performance across diverse multimodal tasks.

3.4 Pretraining vs. Fine-Tuning

MLLMs typically employ two major training strategies: pretraining and fine-tuning, each serving a distinct purpose. Pretraining is conducted on large-scale multimodal datasets and follows self-supervised paradigms such as contrastive learning (CLIP), masked auto-encoding [28], and retrieval-based learning [30], allowing models to develop generalized multimodal representations. However, fine-tuning adapts pretrained models to specific downstream tasks, such as image captioning, video summarization, or speech synthesis, using supervised learning [10], instruction tuning [30], or reinforcement learning [49]. While pretraining enhances generalization across domains, fine-tuning ensures higher accuracy for domain-specific applications. Parameter-efficient fine-tuning (LoRA, PEFT) reduces computational overhead by tuning only specific layers, making MLLMs more adaptable to new tasks [26].

Contrastive Pretraining Strategies: Contrastive pretraining in MLLMs follows various strategy refinements to improve alignment and data efficiency. Vision-language pretraining (CLIP, ALIGN) relies on contrastive loss functions, ensuring tight coupling between images and their textual descriptions [1]. Audio-language pretraining follows similar techniques, where HuBERT [29] and BEATs use contrastive objectives to align waveform embeddings with transcriptions, improving speech representation learning. Video-language pretraining is even more challenging, requiring long-term alignment, as in VideoChatGPT [1] and LLaMA-VID [10], which employ frame-based

and sequential contrastive objectives to improve video-text interactions. These strategies significantly reduce modality gaps and improve zero-shot generalization.

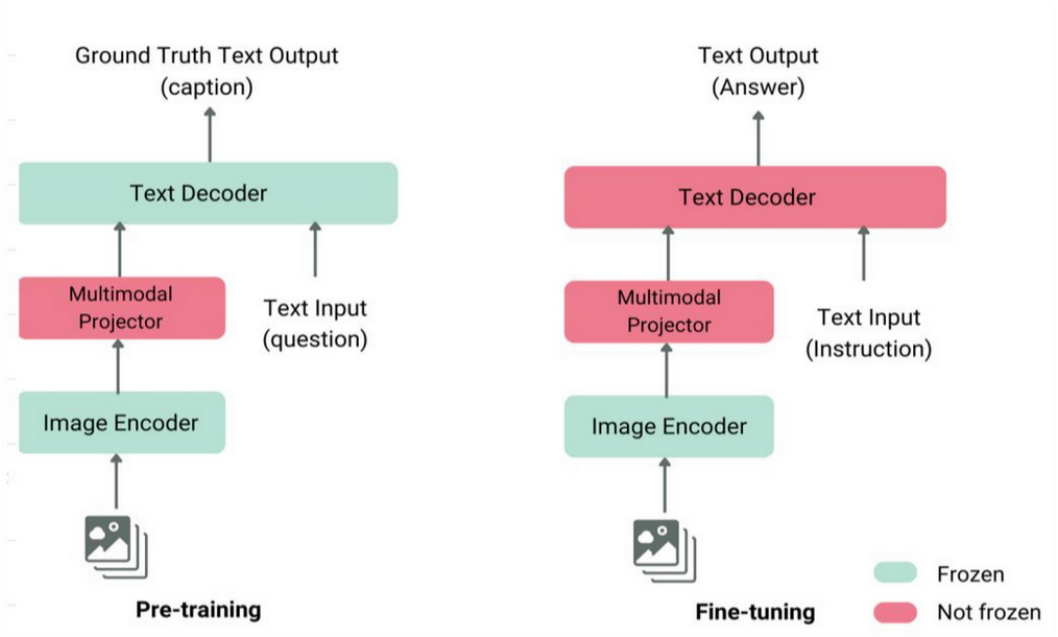


Figure 7: Pre-training vs. Fine-tuning in a Multimodal Model.

3.5 Loss Functions and Regularization Techniques

Loss functions play a crucial role in training Multimodal Large Language Models (MLLMs) by aligning different modalities, reconstructing missing data, and optimizing classification tasks. The primary loss functions in MLLMs include contrastive loss for modality alignment, masked modeling loss for feature learning, reconstruction loss for self-supervised training, and cross-entropy loss for classification tasks. These losses ensure that multimodal embeddings are effectively learned and integrated across text, vision, audio, and video domains. The sections below provide detailed descriptions of each loss function, including mathematical formulations.

Contrastive Loss for Modality Alignment: Contrastive learning is a widely used strategy in vision-language models (CLIP, ALIGN, BLIP-2) and speech-text models (HuBERT, BEATs), ensuring that paired multimodal representations remain close while unpaired representations are pushed apart. This method enables zero-shot learning, multimodal retrieval, and cross-modal reasoning.The contrastive loss function (InfoNCE - Noise Contrastive Estimation) is defined in equation 3.

Masked Modeling Loss: Masked modeling loss is used in self-supervised pretraining, where the model learns multimodal representations by predicting missing information. This technique is highly

effective for text, vision, and multimodal transformers. It follows a similar principle to BERT's Masked Language Modeling (MLM) but extends to images and videos. The masked modeling loss function is:

$$L_{\text{masked}} = \mathbb{E}_{(x,y)\sim D} \left[-\log P(y_{\text{masked}} \mid x) \right] \tag{4}$$

Where x represents the observed input (e.g., an image, video, or text), y_{masked} represents the masked tokens that the model must reconstruct. The model predicts masked tokens, which improves the quality of feature representation. Masked modeling loss is used in VideoMAE [28] that masks video frames and trains the model to reconstruct them. PaLI-X [4] masks both text and vision features, forcing the model to understand multimodal dependencies. While masked modeling improves representation learning, it is not as effective as contrastive loss for modality alignment since it does not explicitly enforce cross-modal similarity.

Reconstruction Loss (For feature learning in autoencoders): Reconstruction loss is widely used in autoencoder-based MLLMs such as HuBERT, BEATs, and DALL-E, where the model learns self-supervised embeddings by compressing multimodal inputs and reconstructing them. This method works especially effectively with speech-text and vision-language models. The Mean Squared Error (MSE)-based reconstruction loss is given as:

$$L_{\text{reconstruction}} = \sum_{i} \|x_i - \hat{x}_i\|^2 \tag{5}$$

Where x_i is the original input (e.g., an image patch or audio signal), $\hat{x}_i$ is the reconstructed output from the model and $\|\cdot\|^2$ represents the Mean Squared Error (MSE) distance measure. Reconstruction loss is used in HuBERT [29] that learns speech representations by reconstructing masked waveforms, and BEATs uses self-supervised learning to improve audio-text alignment. Although reconstruction loss helps models learn data structures, it does not explicitly improve cross-modal interactions, making it less effective than contrastive loss for multimodal tasks.

Cross-Entropy Loss (For classification tasks): Cross-entropy loss is a common technique used in classification-based MLLMs to ensure that the model gives high probabilities to right labels while minimizing wrong predictions. It is particularly effective in image-text categorization, speech recognition, and text production jobs. The cross-entropy loss function is as follows:

$$L_{\text{CE}} = -\sum_{i} y_i \log P(y_i \mid x) \tag{6}$$

Where y_i represents the true class label, $P(y_i \mid x)$ is the predicted probability assigned by the model.

Cross-entropy loss is applied in, GPT-4, PaLM for text-based multimodal classification, Flamingo, Kosmos-2 for vision-language alignment through classification-based multimodal queries and LLaVA [37] for vision-language answering models.While cross-entropy loss is critical for classification, it is not as effective as contrastive loss for aligning multimodal embeddings.

4 Contrastive Learning in MLLMs: Applications, Challenges, and Evaluation

As previously stated, contrastive learning has emerged as an effective technique for training Multimodal Large Language Models (MLLMs), allowing for strong cross-modal representation learning, retrieval, and generation tasks. This section delves into its primary uses, followed by the obstacles, evaluation methodologies, and real-world implementations. Table 2 provides an overview of contrastive learning models in MLLMs.

4.1 Applications of Contrastive Learning in MLLMs

4.1.1 Image-Text Understanding

Contrastive learning plays a fundamental role in vision-language alignment, enabling image-text retrieval, zero-shot classification, and multimodal captioning. CLIP [1] pioneered this approach by jointly training an image encoder (ViT/ResNet) and a text encoder (Transformer) using a contrastive loss function (see figure 3). The goal was to increase similarity between paired image-text embeddings and reduce similarity between mismatched pairs. BLIP [30] extended CLIP by introducing a query-based transformer (Q-Former), refining multimodal embeddings for captioning and image-grounded reasoning. These models enable image-to-text and text-to-image retrieval, improving tasks like visual question answering (VQA) and scene understanding. Figure 8 shows a summary of the CLIP model technique.

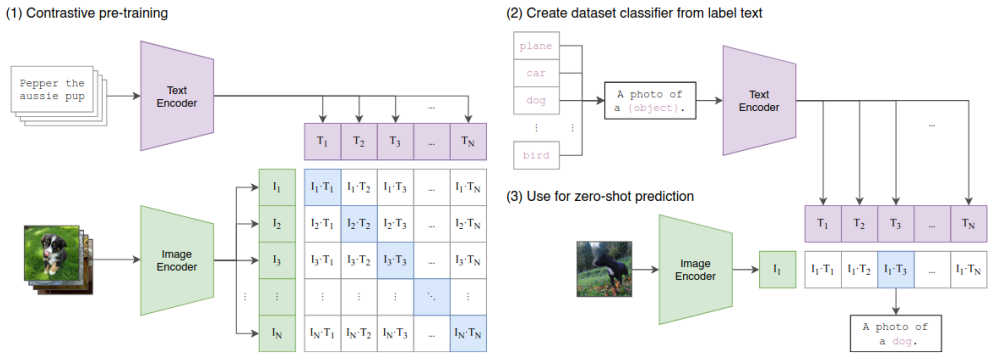


Figure 8: architecture of CLIP model. [1].

4.1.2 Audio-Visual Language Models

Contrastive learning is also crucial for audio-visual language understanding, aligning speech, sound, and vision. AudioCLIP [50] extends CLIP's architecture by introducing an audio encoder alongside

vision and text encoders, enabling audio-image-text retrieval and sound classification. AV-HuBERT [51] focuses on speech-video alignment, learning self-supervised audiovisual representations from unlabeled video speech. These models improve speech-to-text generation, lip-reading, and multi-modal speech recognition, significantly benefiting assistive AI applications. These models enable real-time multi-modal search for applications like Google Lens and visual content retrieval. The overview of the AudioCLIP model depicted in figure 9.

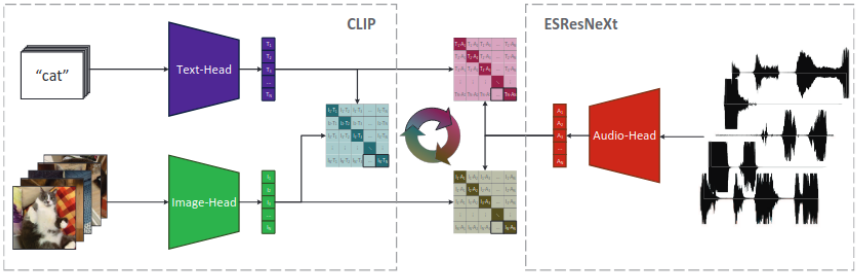


Figure 9: Overview of the proposed AudioCLIP model. [50].

4.1.3 Cross-Modal Retrieval Systems

Cross-modal retrieval involves searching across various modalities (e.g., retrieving images from text queries). ALIGN [3] trained image-text embeddings on a massive dataset (1.8 billion image-text pairs), significantly improving image search and captioning tasks. Similarly, Florence [52] introduced a unified transformer model to optimize multimodal retrieval using contrastive learning. These models enable real-time multi-modal search for applications like Google Lens and visual content retrieval. The summary of the ALIGN model approach depicted in figure 10.

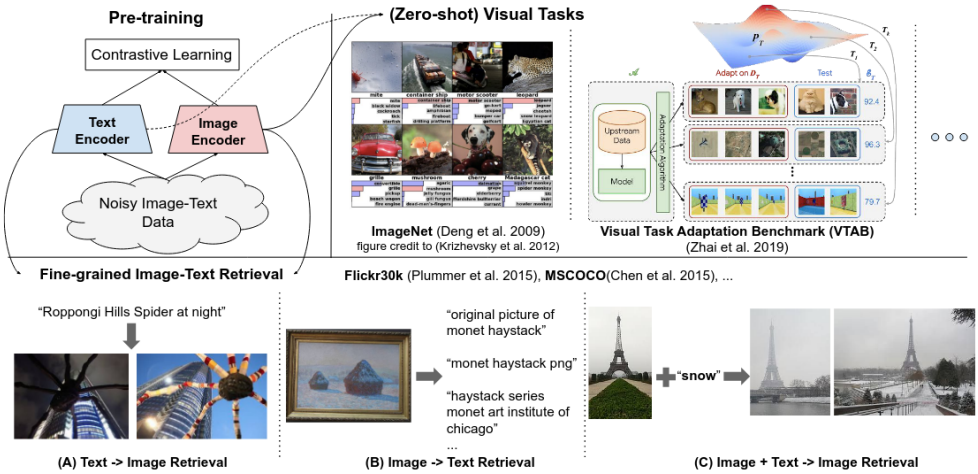


Figure 10: A summary of ALIGN method [3].

4.1.4 Zero-shot and Few-shot Learning

Contrastive learning enables MLLMs to generalize to unseen tasks without additional training. Flamingo [24] introduced gated cross-attention layers, allowing few-shot multimodal learning. CoCa [53], further improved zero-shot generation by using contrastive pretraining followed by autoregressive fine-tuning.

$$L_{\text{zero-shot}} = -\mathbb{E}\left[\log \frac{P(T|I)}{\sum_j P(T|I_j)}\right] \tag{7}$$

where $P(T|I)$ is the probability of generating the correct text T given an image I. These models excel in generalization, multimodal translation, and low-resource settings. The figure 11 provides an overview of Flamingo architecture.

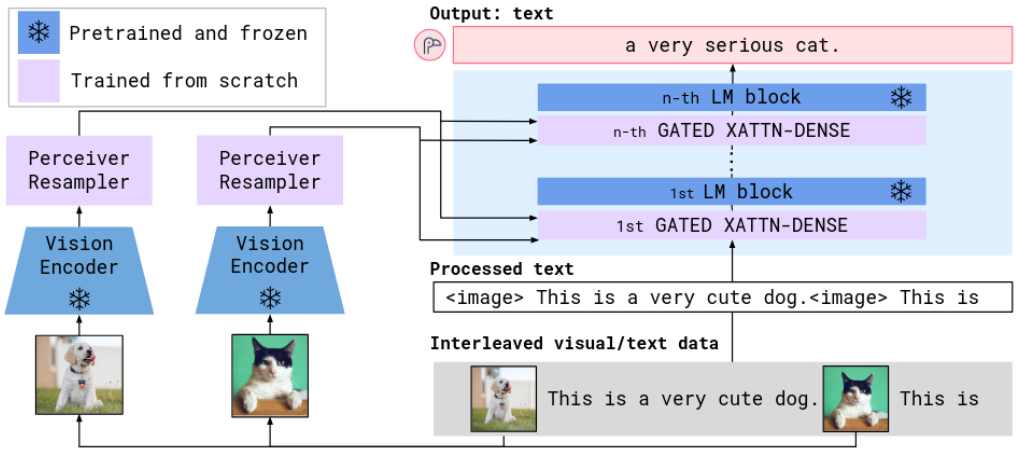


Figure 11: Flamingo architecture overview. [24].

4.1.5 Multimodal Sentiment Analysis

Contrastive learning is widely used in sentiment analysis, where text, facial expressions, and speech are analyzed together. MMIM [54] and MISA [55] optimize multimodal embeddings, allowing better emotion recognition.

$$L_{\text{sentiment}} = -\sum_i \log P(y_i|X_i) \tag{8}$$

where y_i is the sentiment label, and X_i is the multimodal input. These models are crucial for customer service automation and psychological analysis.

Model	Application	Task	Dataset Used	Advantage	Disadvantage
CLIP [1]	Image-Text Understanding	Image-text retrieval, zero-shot classification	LAION-400M, Conceptual Captions	Strong alignment; effective zero-shot classification	Requires large-scale data; computationally expensive
BLIP [30]	Image-Text Understanding	Multimodal captioning, image grounding	Conceptual Captions, COCO	Refines multimodal embeddings for captioning	Higher complexity; requires fine-tuning
AudioCLIP [50]	Audio-Visual Learning	Audio-image-text retrieval, sound classification	AudioSet, ESC-50	Enhances audio-text-visual retrieval	Limited to pre-defined sound categories
AV-HuBERT [51]	Audio-Visual Learning	Speech-video alignment, self-supervised learning	LRS3, VoxCeleb	Self-supervised audiovisual representations	Requires large-scale video datasets
ALIGN [3]	Cross-Modal Retrieval	Image-text retrieval, captioning	JFT-300M, WebImageText	Trained on a massive dataset	High computational cost
Florence [52]	Cross-Modal Retrieval	Multimodal retrieval, image understanding	COCO, OpenImages	Optimized for multimodal retrieval	Lacks task-specific fine-tuning
Flamingo [24]	Zero-shot Learning	Few-shot multimodal learning	WebLI, LAION	Few-shot multimodal learning	Requires high computational resources
CoCa [53]	Zero-shot Learning	Zero-shot generation, autoregressive fine-tuning	JFT-300M, WebLI	Improves generalization with autoregressive fine-tuning	High latency due to fine-tuning
MMIM [54]	Sentiment Analysis	Multimodal sentiment analysis	CMU-MOSEI	Optimized multimodal embeddings	Struggles with unseen emotions
MISA [55]	Sentiment Analysis	Emotion recognition	CMU-MOSI, CMU-MOSEI	Enhances emotion recognition	Susceptible to bias in sentiment classification

Table 2: Summary of Contrastive Learning Models in MLLMs

4.1.6 Evaluation Metrics for Measuring Contrastive Learning

Evaluating the effectiveness of contrastive learning models requires well-defined metrics that assess various aspects such as retrieval accuracy, consistency in learned representations, and generalization to unseen data. Below, we discuss three key metrics commonly used in evaluating contrastive learning models: Recall@K, Normalized Mutual Information (NMI), and Zero-shot Accuracy.

Recall@K: Recall@K [1, 3, 4] is a ranking-based metric that measures the retrieval performance of a model. Given a query sample, the model ranks all possible matches based on similarity scores. Recall@K calculates the fraction of times the correct match appears within the top K retrieved results. The mathematical formulation for recall@K is:

$$\textit{Recall@}K = \frac{1}{N}\sum_{i=1}^{N}\mathbf{1}(\text{GT}_i \in R_K(q_i)) \tag{9}$$

where N is the total number of queries, GT_i is the ground-truth match for the i-th query, $R_K(q_i)$ is the top-K retrieved candidates for query q_i, $\mathbb{1}(\cdot)$ is an indicator function that equals 1 if GT_i is found in the retrieved set, and 0 otherwise. Recall@K is widely used in evaluating image-text retrieval and multimodal matching models, where the goal is to retrieve the correct text (or image) for a given image (or text).

Normalized Mutual Information (NMI): NMI [56, 20, 57] measures the quality of clustering by evaluating how well the predicted cluster assignments align with the ground-truth labels. It is particularly useful for assessing the semantic consistency of learned embeddings in contrastive learning. The Normalized Mutual Information is defined as:

$$\text{NMI}(Y, C) = \frac{2 \times I(Y; C)}{H(Y) + H(C)} \tag{10}$$

where $I(Y; C)$ is the mutual information between Y and C, $H(Y)$ and $H(C)$ are the entropies of Y and C, respectively. NMI is normalized between 0 and 1, where 1 indicates perfect alignment between clusters and labels, and 0 indicates no mutual information. NMI is used in self-supervised learning to evaluate the alignment of representations in unsupervised clustering tasks, such as image or text categorization.

Zero-shot Accuracy: Zero-shot [1, 3, 2] accuracy evaluates the model's ability to generalize to unseen classes or tasks without further fine-tuning. It measures how well the learned representations transfer to new data distributions. The zero-shot accuracy is defined as:

$$\text{Zero-shot Accuracy} = \frac{\sum_{i=1}^{N} \mathbf{1}(\hat{y}_i = y_i)}{N} \tag{11}$$

where $\hat{y}_i$ is the predicted class label, y_i is the ground-truth label, N is the total number of test samples. Zero-shot accuracy is widely used in vision-language models (e.g., CLIP, ALIGN) to measure their ability to classify images based on textual descriptions without explicit training on those categories.

4.2 Challenges in Contrastive Learning

4.2.1 Computational Complexity

Contrastive learning is computationally intensive due to large batch sizes, high-dimensional embeddings, and significant GPU memory requirements. Methods like SimCLR [4] and MoCo [20] exhibit

quadratic complexity $O(N^2 d)$, necessitating large, diverse negative samples. High-dimensional embeddings (128–1024) further escalate memory consumption, slowing similarity computations and matrix operations [1, 2]. These demands require high-end hardware (e.g., TPU pods, NVIDIA A100 GPUs). Techniques like momentum contrast (MoCo) and dimensionality reduction have been proposed to optimize memory use [56], yet real-time application feasibility remains a challenge.

4.2.2 Bias and Fairness Considerations

Contrastive learning models, including CLIP [1], are prone to biases (gender, racial, cultural) stemming from dataset imbalances, contrastive sampling strategies, and multimodal interactions. Large-scale datasets like LAION-400M [58] reinforce stereotypes, associating women with domestic roles and men with professional settings [59], favoring Western perspectives while underrepresenting non-Western imagery [60]. Addressing bias necessitates interventions in dataset curation, fairness-aware training, and demographic balancing [61, 62, 63]. LAION-5B employs explicit filtering to enhance representation diversity.

Contrastive sampling can amplify biases when negative samples disproportionately link specific demographics with negative associations. Demographically fair sampling ensures uniform pair selection, reducing skewed relationships. Counterfactual data augmentation (CDA) [64] and debasing objectives [65] help mitigate biased learning. Bias auditing techniques—embedding fairness metrics, stereotype association tests [66], and intersectional analysis [67]—further promote equitable model behavior.

4.2.3 Data Privacy Concerns

MLLMs using contrastive learning rely on vast datasets containing sensitive and proprietary information across text, images, and audio. Privacy threats include data leakage, membership inference attacks, and cross-modal privacy leaks, posing risks in healthcare, finance, and biometrics [68]. Datasets often contain personally identifiable information (PII) and copyrighted content, raising ethical and legal concerns under regulations like GDPR, CCPA, and HIPAA [69, 70].

Risks [71] are reduced by privacy-preserving strategies including homomorphic encryption, secure multi-party computation, federated learning, and differential privacy. Federated learning decentralizes training, preventing central data breaches [72], while differential privacy adds noise to limit re-identification risks [73]. SMPC and homomorphic encryption enable collaborative model training

without exposing private datasets [74]. Recent research explores privacy-aware embeddings and adversarial defenses to counter model inversion attacks, where attackers attempt to reconstruct original data from embeddings [75]. Future directions emphasize synthetic data generation, decentralized self-supervised learning, and secure model auditing frameworks for enhanced privacy protection [76, 77].

4.2.4 Hallucination

Hallucination in contrastive learning-based MLLMs arises due to modality misalignment, over-reliance on language priors, dataset bias, and limitations in contrastive learning techniques. Models like CLIP and BLIP-2 may generate misleading outputs by prioritizing learned statistical correlations over actual image-text relationships, particularly when trained on noisy, web-scraped datasets that contain mislabeled or biased data [1, 30, 78]. The contrastive learning process, which relies on negative sampling, can further exacerbate these hallucinations by reinforcing spurious associations instead of factual grounding [30, 78]. To mitigate these issues, Hallucination-Augmented Contrastive Learning (HACL) incorporates synthetic hard negatives to improve robustness against false associations, while self-consistency training ensures model reliability by selecting the most consistent response across multiple outputs [30]. Additionally, cross-modal verification enforces alignment between generated content and reference modalities, preventing incorrect associations [1]. Lastly, curated pretraining datasets that prioritize high-quality, human-annotated samples over noisy web data have been shown to significantly mitigate hallucinations and enhance model reliability in multimodal applications [30, 78].

4.3 Real-world Implementations

4.3.1 Case Studies: OpenAI's DALL-E, Google's PaLM-E, Meta's ImageBind

Significant progress has been achieve in MLLMs, which allow AI systems to process and produce data from a variety of modalities, including text, photos, and videos. Leading AI research organizations—including OpenAI, Google, and Meta—have developed advanced multimodal AI systems that leverage large-scale pretraining. Below is an in-depth case study of OpenAI's DALL-E, Google's PaLM-E, and Meta's ImageBind, focusing on their architectures, features, privacy implications, and research contributions.

OpenAI's DALL-E: DALLE [79, 80] is a powerful image-generation model developed by OpenAI that transforms textual descriptions into high-quality images. Built on an autoregressive transformer-based architecture, the model can synthesize realistic and diverse images from natural language prompts. It supports various creative capabilities, including style transfer, scene composition, and fine-grained image attribute control, all through text conditioning.

Despite its impressive capabilities, DALL-E raises critical privacy and ethical concerns. The ability to generate highly realistic synthetic images increases risks such as deepfake creation, misinformation dissemination, and fabrication of convincing yet fictional visual content. Furthermore, DALL-E is trained on large-scale scraped datasets, leading to copyright issues and ethical dilemmas regarding the unauthorized use of images. These challenges highlight the necessity for responsible AI deployment, watermarking techniques, and regulatory measures to mitigate potential misuse.

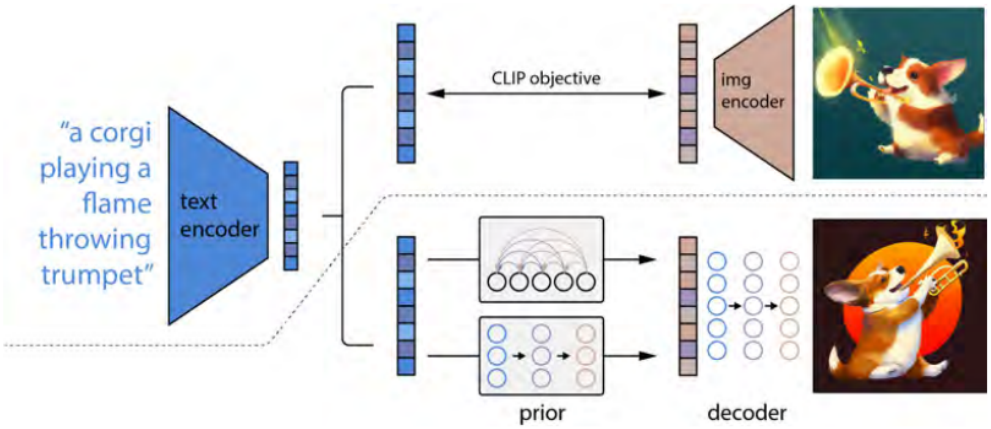


Figure 12: Flamingo architecture overview. [24].

Google's PaLM-E: PaLM-E [23] is a multimodal, embodied AI model developed by Google Research. It extends the Pathways Language Model (PaLM) by integrating natural language processing with sensory and visual data, enabling robots to interpret and act on real-world environments. Unlike traditional language models that process only text, PaLM-E incorporates images, video frames, and other perceptual data to enhance robotic control, visual question answering, and instruction-following capabilities.

PaLM-E is built on Google's Pathways AI architecture, which allows it to generalize across various tasks without requiring task-specific fine-tuning. This makes it a scalable and adaptable solution for embodied AI applications. However, its reliance on large-scale multimodal datasets introduces concerns about privacy, data bias, and potential misuse in surveillance and automation. Ensuring

responsible AI development requires transparency in data collection, bias mitigation strategies, and strict ethical guidelines to prevent the potential exploitation of AI-powered robotics.

Meta's ImageBind: ImageBind [22], developed by Meta AI, is a multimodal representation learning model that unifies six different data types—text, image, audio, depth, thermal, and inertial measurement unit (IMU) signals—within a shared embedding space. Unlike conventional multimodal models that require explicit cross-modal supervision, ImageBind exploits natural co-occurrences in large-scale datasets to establish relationships between diverse modalities. Extending the principles of CLIP (Contrastive Language–Image Pretraining), ImageBind enables seamless cross-modal retrieval, multimodal reasoning, and zero-shot learning.

This method greatly improves AI's comprehension and processing of contextual data from a variety of sensory inputs. Applications of ImageBind include image-audio search, enhanced scene understanding in AR/VR environments, and security surveillance using integrated sensory data. However, the inclusion of multiple sensory inputs raises ethical concerns regarding privacy risks, potential surveillance misuse, and dataset biases. These issues underscore the importance of responsible AI governance, security measures, and regulatory oversight to ensure ethical deployment.

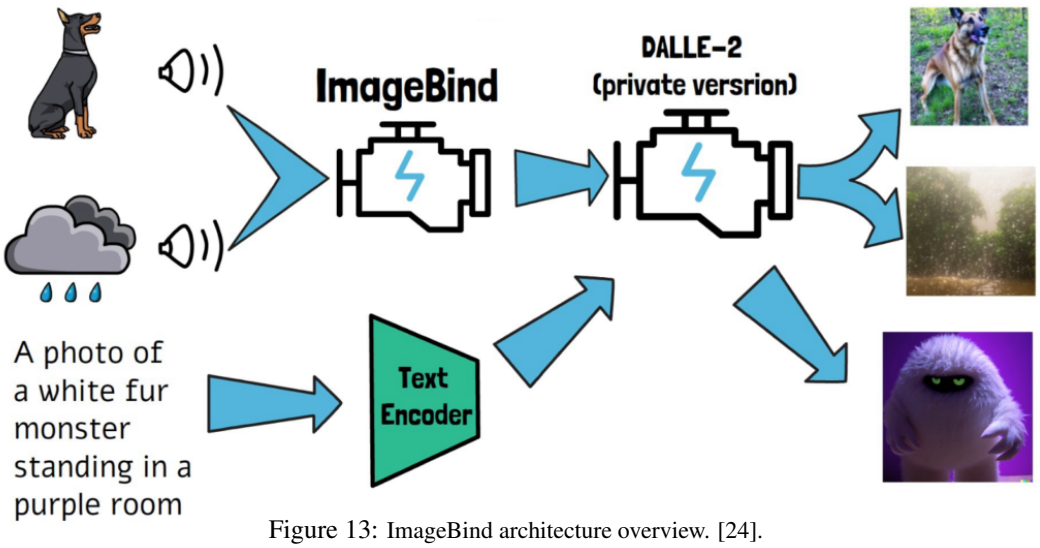
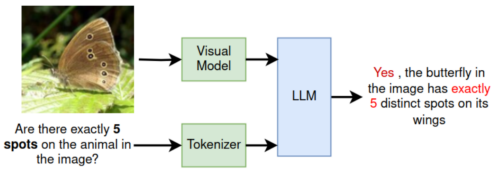


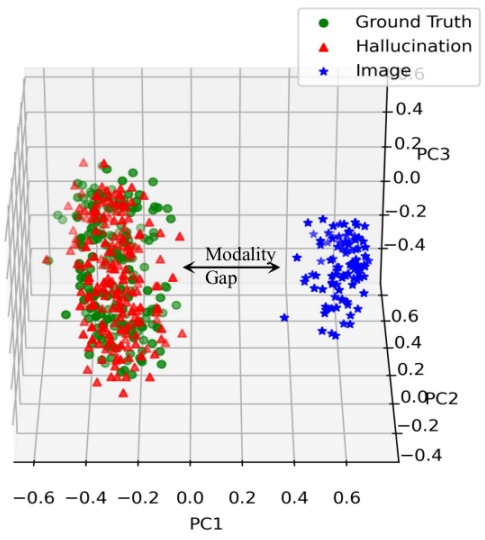
Figure 13: ImageBind architecture overview. [24].

5 Future Directions

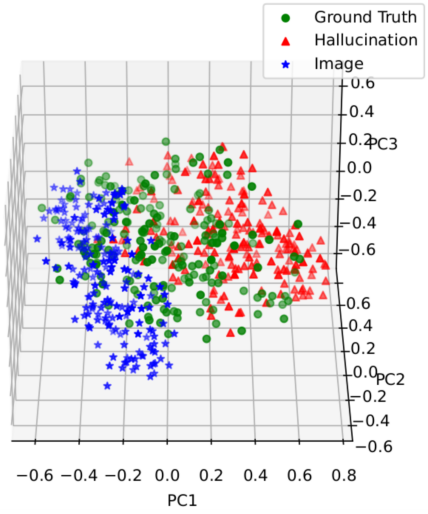
In contrast to standard LLMs, which rely primarily on textual input, multimodal Large Language Models (MLLMs) are meant to process and analyze several data modalities such as text, images, and video. However, they commonly induce hallucinations, providing false or inaccurate information that does not correlate with visual input, as seen in figure 14a. A key challenge contributing to this issue is the modality gap between textual and visual representations, which results in misalignment. Furthermore, hallucinative and non-hallucinative texts often become entangled, making it difficult to differentiate between accurate and erroneous outputs, as depicted in figure 14b.



(a) Example of hallucination



(b) Token Representations w/o Contrastive Learning



(c) Token Representations w/ Contrastive Learning

Figure 14: Effect of Contrastive Learning on Multimodal Representations. Contrastive learning reduces the modality gap, improving alignment between hallucinated outputs, ground truth, and image representations.

Several factors contribute to hallucinations in MLLMs, including representation misalignment, where visual features fail to align with the textual space, and lack of semantic differentiation, making it difficult to distinguish correct from incorrect textual outputs. Noisy, biased, or incomplete multimodal datasets further exacerbate the problem, while over-reliance on language priors causes models to

prioritize textual context over actual visual content, leading to incorrect descriptions. Additionally, inefficient cross-modal learning in projection-based approaches limits the integration of visual cues into language models, increasing hallucination occurrences. As shown in figure 14c, contrastive learning mitigates these issues by improving vision-language alignment and reducing hallucinations.

5.1 Ongoing Project

Problem Statement :

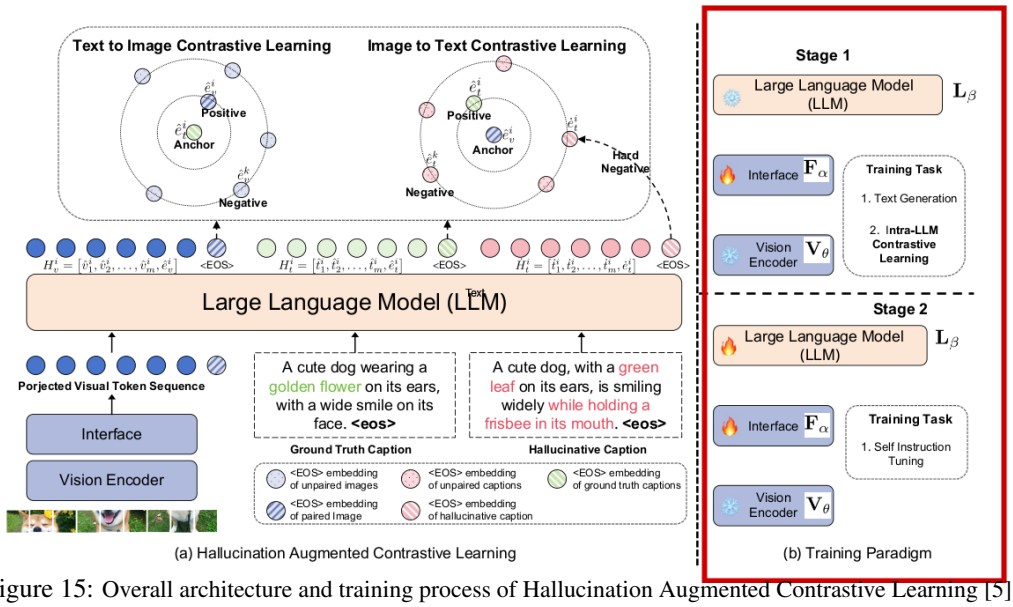


Figure 15: Overall architecture and training process of Hallucination Augmented Contrastive Learning [5].

Multimodal Large Language Models (MLLMs) such as LLaVA [10], Flamingo [24], and Hallucination Augmented Contrastive Learning (HACL) [5] are designed to process both visual and textual inputs for tasks like image captioning, visual question answering, multimodal reasoning. A key challenge in these models is achieving effective alignment between vision and text representations while ensuring task-specific performance. To address this, existing methods adopt a two-stage training paradigm, where contrastive learning is first used to align visual and textual features, followed by fine-tuning for downstream tasks. Specifically, in HACL, contrastive learning optimizes a learnable interface F_α, while the vision encoder V_θ and large language model L_β remain frozen as illustrated in figure 15. This process is governed by the objective

$$L = \min_{\alpha,\beta} [L_T(\beta, \alpha) + \lambda L_{CL}(E)] \tag{12}$$

where L_{CL} ensures cross-modal alignment, and L_T fine-tunes the model for specific tasks. However, this approach introduces a fundamental trade-off: it attempts to optimize both embedding learning and task-specific adaptation using a single objective, which often leads to compromising either embedding quality or task performance. Intuitively, we want the model to learn the best embeddings while also solving the task optimally, but the current formulation forces a balance between the two, limiting overall effectiveness. This challenge motivates our hierarchical formulation, where we decouple embedding learning from task-specific optimization, allowing embeddings to be continuously refined throughout training instead of remaining static after contrastive pretraining. This strategy improves representation alignment, minimizes hallucinations, and enhances generalization, making the model more adaptive and robust in multimodal learning.

In the integrated learning approach, embedding learning and fine-tuning are performed together, combining contrastive learning and task-specific adaptation within a unified framework. During pretraining, vision-text representations are aligned using contrastive loss L_{CL}, but in LLAVA [10], the absence of task-specific supervision results in suboptimal embeddings, requiring extensive pretraining data for better generalization. However, in HACL [5], task-specific supervision is incorporated in pretraining via contrastive learning. Fine-tuning L_T then applies task learning, but in LLAVA, it does not update the learned embeddings, making adaptation challenging and limiting flexibility in downstream applications. Since model parameters (α^*, β^*) remain fixed in LLAVA during fine-tuning, their ability to refine representations for specific tasks is restricted, contributing to a modality gap where textual and visual features are not continuously aligned. However, in HACL, embeddings are updated during both pretraining and fine-tuning, mitigating this gap. This rigidity often leads to hallucinations in LLAVA, as the model fails to correctly associate visual inputs with textual descriptions, generating incorrect outputs. In contrast, HACL refines embeddings using contrastive learning with hallucinated samples, improving alignment. Additionally, in LLAVA, the lack of continuous representation refinement results in inefficient adaptation to new multimodal tasks, since embeddings learned in pretraining are not dynamically optimized for real-world applications. The main obstacle with this strategy is that while embedding learning and task learning are coupled, they do not interact dynamically throughout training, preventing the model from adapting embeddings in response to fine-tuning objectives. As a result, LLAVA struggles with poor generalization, while HACL alleviates this limitation by continuously aligning textual and visual features. Overcoming these limitations requires a more adaptive learning strategy,

where contrastive alignment and task-specific learning evolve simultaneously, ensuring continuous refinement of multimodal representations.

Proposed Method: Hierarchical Fine-Tuning:

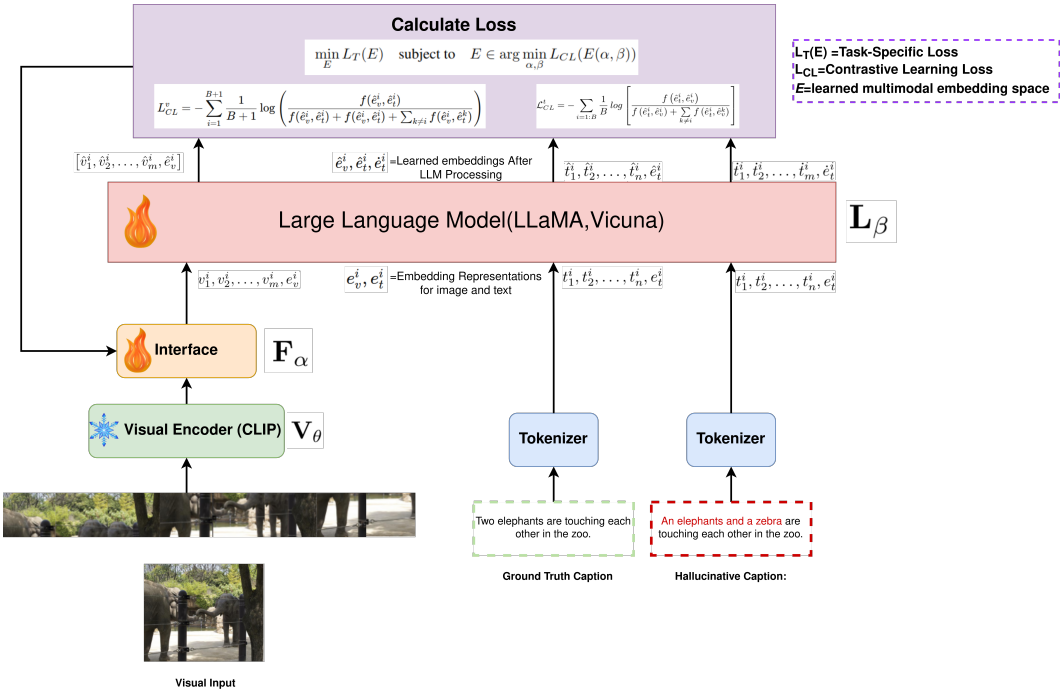


Figure 16: Overview of proposed hierarchical fine-tuning approach

Our hierarchical fine-tuning approach enables decoupled embedding and parameter learning, allowing continuous refinement of embeddings during fine-tuning rather than relying solely on pretraining for representation alignment, which can lead to persistent hallucinations. Unlike traditional methods where embeddings remain frozen, our approach ensures they dynamically adapt throughout fine-tuning, leading to improved multimodal representation alignment. By integrating contrastive learning with task-specific optimization, our method reduces hallucinations and enhances vision-text alignment, ensuring more reliable outputs. The overview of the proposed hierarchical fine-tuning approach is depicted in figure 16 and the overall loss function is formulated as :

$$\min_E L_T(E) \quad \text{subject to} \quad E \in \arg\min_{\alpha,\beta} L_{CL}(E(\alpha,\beta)) \tag{13}$$

where E is the learned multimodal embedding space and α, β are model parameters for contrastive learning.

The total loss is :

$$L_{Total} = \min_{E=\arg\min_{\alpha,\beta} L_{CL}(E(\alpha,\beta))} L_T(E) \tag{14}$$

where inner optimization L_{CL} refines embeddings using contrastive learning to minimize modality misalignment, and outer optimization L_T updates embeddings dynamically for better generalization across multimodal tasks. This joint optimization strategy allows for more adaptive learning, improving performance while reducing errors in vision-language models.

5.2 Future Project

We also plan to enhance our proposed architecture by incorporating Direct Preference Optimization (DPO) [81] for improved optimization. Additionally, instead of processing visual and text embeddings separately, we aim to integrate both embeddings directly into the LLM [82], reducing complexity and improving multimodal representation learning.

Timeline:

As shown in figure 17, my research for the rest of the Ph.D. will concentrate on several core challenges and proceed within a structured time frame. Initially, I will concentrate on implementing the joint optimization of α and β(March – July 2025) for better embedding and efficient training. Next, I will research and implement the integration of Direct Preference Optimization (DPO) (August – December 2025) to enhance alignment in Vision-Language Models (VLMs) by refining contrastive learning techniques. Following this, I will focus on simplifying multimodal representation learning (January – September 2026) to strengthen the integration between textual and visual elements, ensuring improved multimodal understanding. Finally, I will explore adversarial vulnerabilities and attacks on multimodality (October 2026 onward) to enhance the robustness and interpretability of VLMs. This research trajectory aims to bridge fairness, alignment, and explainability in multimodal frameworks, contributing to the broader developments in interpretable and secure NLP systems.

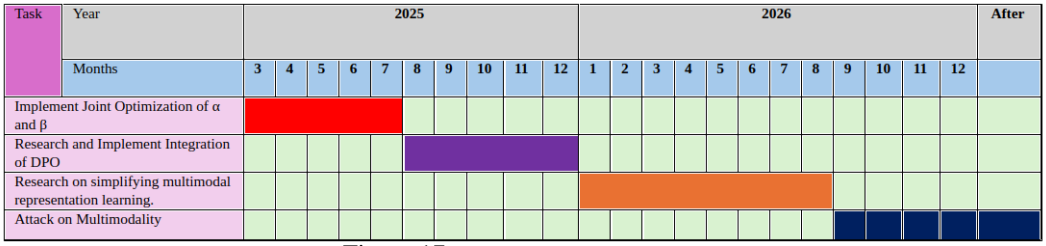


Figure 17: Ph.D. Research Timeline.

Conclusion

Multimodal Large Language Models (MLLMs) have revolutionized artificial intelligence by enabling seamless integration of diverse modalities, including text, images, and audio. The role of contrastive learning in MLLMs is pivotal in improving cross-modal alignment, enhancing zero-shot learning, and optimizing data representation. Through models like CLIP, BLIP-2, and OpenFlamingo, contrastive learning has demonstrated significant advantages in improving multimodal understanding, retrieval, and reasoning tasks. Despite these advances, challenges remain, particularly in handling modality imbalance, computational costs, and interpretability. Future research should concentrate on designing effective contrastive learning techniques that reduce reliance on extensive labeled datasets while ensuring fairness, scalability, and robustness. Additionally, novel architectures that integrate self-supervised learning with contrastive objectives could further improve the generalization capabilities of MLLMs. By addressing these challenges, MLLMs can continue to push the boundaries of AI, enabling more adaptive, interpretable, and human-aligned multimodal systems.

Disclaimer

Portions of this report were paraphrased using OpenAI's ChatGPT-4 to assist in correcting and rephrasing the content of the original writing. The use of ChatGPT-4 was aimed at enhancing the clarity and conciseness of the presented information, not to generate creative content.

References

[1] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark *et al.*, "Learning transferable visual models from natural language supervision," in *International conference on machine learning*. PMLR, 2021, pp. 8748–8763.

[2] A. Dosovitskiy, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv:2010.11929*, 2020.

[3] C. Jia, Y. Yang, Y. Xia, Y.-T. Chen, Z. Parekh, H. Pham, Q. Le, Y.-H. Sung, Z. Li, and T. Duerig, "Scaling up visual and vision-language representation learning with noisy text supervision," in *International conference on machine learning*. PMLR, 2021, pp. 4904–4916.

[4] T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, "A simple framework for contrastive learning of visual representations," in *International conference on machine learning*. PMLR, 2020, pp. 1597–1607.

[5] C. Jiang, H. Xu, M. Dong, J. Chen, W. Ye, M. Yan, Q. Ye, J. Zhang, F. Huang, and S. Zhang, "Hallucination augmented contrastive learning for multimodal large language model," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 27 036–27 046.

[6] R. Nakada, H. I. Gulluk, Z. Deng, W. Ji, J. Zou, and L. Zhang, "Understanding multimodal contrastive learning and incorporating unpaired data," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2023, pp. 4348–4380.

[7] Q. Jiao, D. Chen, Y. Huang, Y. Li, and Y. Shen, "Img-diff: Contrastive data synthesis for multimodal large language models," *arXiv preprint arXiv:2408.04594*, 2024.

[8] B. Mustafa, C. Riquelme, J. Puigcerver, R. Jenatton, and N. Houlsby, "Multimodal contrastive learning with limoe: the language-image mixture of experts," *Advances in Neural Information Processing Systems*, vol. 35, pp. 9564–9576, 2022.

[9] R. Kumar, R. Singhal, P. P. Kulkarni, D. Mehta, and K. S. Jadhav, "M3col: Harnessing shared relations via multimodal mixup contrastive learning for multimodal classification," in *UniReps: 2nd Edition of the Workshop on Unifying Representations in Neural Models*, 2024. [Online]. Available: https://openreview.net/forum?id=vOr0OX8nFD

[10] H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," *Advances in neural information processing systems*, vol. 36, 2024.

[11] B. Dufumier, J. Castillo-Navarro, D. Tuia, and J.-P. Thiran, "What to align in multimodal contrastive learning?" *arXiv preprint arXiv:2409.07402*, 2024.

[12] Y. Zhang, X. Zhang, J. Li, R. C. Qiu, H. Xu, and Q. Tian, "Semi-supervised contrastive learning with similarity co-calibration," *IEEE Transactions on Multimedia*, vol. 25, pp. 1749–1759, 2022.

[13] K. Vasudeva, A. Dubey, and S. Chandran, "Scl-fexr: supervised contrastive learning approach for facial expression recognition," *Multimedia Tools and Applications*, vol. 82, no. 20, pp. 31 351–31 371, 2023.

[14] J. Wu, S. Mo, Z. Feng, S. Atito, J. Kitler, and M. Awais, "Rethinking positive pairs in contrastive learning," *arXiv preprint arXiv:2410.18200*, 2024.

[15] J. Robinson, C.-Y. Chuang, S. Sra, and S. Jegelka, "Contrastive learning with hard negative samples," *arXiv preprint arXiv:2010.04592*, 2020.

[16] L. Wang, C. Du, P. Zhao, C. Luo, Z. Zhu, B. Qiao, W. Zhang, Q. Lin, S. Rajmohan, D. Zhang *et al.*, "Contrastive learning with negative sampling correction," *arXiv preprint arXiv:2401.08690*, 2024.

[17] R. Hadsell, S. Chopra, and Y. LeCun, "Dimensionality reduction by learning an invariant mapping," in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, vol. 2, 2006, pp. 1735–1742.

[18] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[19] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.

[20] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 9729–9738.

[21] B. Elizalde, S. Deshmukh, M. Al Ismail, and H. Wang, "Clap learning audio concepts from natural language supervision," in *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2023, pp. 1–5.

[22] R. Girdhar, A. El-Nouby, Z. Liu, M. Singh, K. Alwala, A. Joulin, and I. Misra, "Imagebind: one embedding space to bind them all. arxiv," *Preprint posted online on May*, vol. 9, 2023.

[23] D. Driess, F. Xia, M. S. Sajjadi, C. Lynch, A. Chowdhery, B. Ichter, A. Wahid, J. Tompson, Q. Vuong, T. Yu *et al.*, "Palm-e: An embodied multimodal language model," *arXiv preprint arXiv:2303.03378*, 2023.

[24] J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds *et al.*, "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23 716–23 736, 2022.

[25] C. X. Liang, P. Tian, C. H. Yin, Y. Yua, W. An-Hou, L. Ming, T. Wang, Z. Bi, and M. Liu, "A comprehensive survey and guide to multimodal large language models in vision-language tasks," *arXiv preprint arXiv:2411.06284*, 2024.

[26] D. Zhang, Y. Yu, J. Dong, C. Li, D. Su, C. Chu, and D. Yu, "Mm-llms: Recent advances in multimodal large language models," *arXiv preprint arXiv:2401.13601*, 2024.

[27] A. Brock, S. De, S. L. Smith, and K. Simonyan, "High-performance large-scale image recognition without normalization," in *International conference on machine learning*. PMLR, 2021, pp. 1059–1071.

[28] Z. Tong, Y. Song, J. Wang, and L. Wang, "Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training," *Advances in neural information processing systems*, vol. 35, pp. 10 078–10 093, 2022.

[29] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.

[30] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," in *International conference on machine learning*. PMLR, 2023, pp. 19 730–19 742.

[31] J. Lu, R. Gan, D. Zhang, X. Wu, Z. Wu, R. Sun, J. Zhang, P. Zhang, and Y. Song, "Lyrics: Boosting fine-grained language-vision alignment and comprehension via semantic-aware visual objects," *arXiv preprint arXiv:2312.05278*, 2023.

[32] J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat *et al.*, "Gpt-4 technical report," *arXiv preprint arXiv:2303.08774*, 2023.

[33] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar *et al.*, "Llama: Open and efficient foundation language models," *arXiv preprint arXiv:2302.13971*, 2023.

[34] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, S. Gehrmann *et al.*, "Palm: Scaling language modeling with pathways," *Journal of Machine Learning Research*, vol. 24, no. 240, pp. 1–113, 2023.

[35] M. Tsimpoukelli, J. L. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," *Advances in Neural Information Processing Systems*, vol. 34, pp. 200–212, 2021.

[36] R. Campos, M. A. Haider, S. Tipirneni, and S. Werleman, "High-resolution image synthesis with latent diffusion models (stable diffusion)," 2022.

[37] H. Liu, Z. Chen, Y. Yuan, X. Mei, X. Liu, D. Mandic, W. Wang, and M. D. Plumbley, "Audioldm: Text-to-audio generation with latent diffusion models," *arXiv preprint arXiv:2301.12503*, 2023.

[38] X. Wei, T. Zhang, Y. Li, Y. Zhang, and F. Wu, "Multi-modality cross attention network for image and sentence matching," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2020, pp. 10 941–10 950.

[39] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE transactions on pattern analysis and machine intelligence*, vol. 41, no. 2, pp. 423–443, 2018.

[40] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, A. Y. Ng *et al.*, "Multimodal deep learning." in *ICML*, vol. 11, 2011, pp. 689–696.

[41] Y.-H. H. Tsai, P. P. Liang, A. Zadeh, L.-P. Morency, and R. Salakhutdinov, "Learning factorized multimodal representations," *arXiv preprint arXiv:1806.06176*, 2018.

[42] N. Srivastava and R. R. Salakhutdinov, "Multimodal learning with deep boltzmann machines," in *Advances in Neural Information Processing Systems*, F. Pereira, C. Burges, L. Bottou, and K. Weinberger, Eds., vol. 25. Curran Associates,

Inc., 2012. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2012/file/af21d0c97db2e27e13572cbf59eb343d-Paper.pdf

[43] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," *arXiv preprint arXiv:2106.09685*, 2021.

[44] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[45] L. Xue, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020.

[46] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?" in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 18 177–18 186.

[47] Y. Goyal, Z. Wu, J. Ernst, D. Batra, D. Parikh, and S. Lee, "Counterfactual visual explanations," in *International Conference on Machine Learning*. PMLR, 2019, pp. 2376–2384.

[48] S. Venkatasubbu and G. Krishnamoorthy, "Ethical considerations in ai addressing bias and fairness in machine learning models," *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, vol. 1, no. 1, pp. 130–138, 2022.

[49] T. Yu, Y. Yao, H. Zhang, T. He, Y. Han, G. Cui, J. Hu, Z. Liu, H.-T. Zheng, M. Sun *et al.*, "Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 13 807–13 816.

[50] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "Audioclip: Extending clip to image, text and audio," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 976–980.

[51] B. Shi, A. Mohamed, and W.-N. Hsu, "Learning lip-based audio-visual speaker embeddings with av-hubert," *arXiv preprint arXiv:2205.07180*, 2022.

[52] L. Yuan, D. Chen, Y.-L. Chen, N. Codella, X. Dai, J. Gao, H. Hu, X. Huang, B. Li, C. Li *et al.*, "Florence: A new foundation model for computer vision," *arXiv preprint arXiv:2111.11432*, 2021.

[53] J. Yu, Z. Wang, V. Vasudevan, L. Yeung, M. Seyedhosseini, and Y. Wu, "Coca: Contrastive captioners are image-text foundation models," *arXiv preprint arXiv:2205.01917*, 2022.

[54] W. Han, H. Chen, and S. Poria, "Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis," *arXiv preprint arXiv:2109.00412*, 2021.

[55] D. Hazarika, R. Zimmermann, and S. Poria, "Misa: Modality-invariant and-specific representations for multimodal sentiment analysis," in *Proceedings of the 28th ACM international conference on multimedia*, 2020, pp. 1122–1131.

[56] M. Caron, I. Misra, J. Mairal, P. Goyal, P. Bojanowski, and A. Joulin, "Unsupervised learning of visual features by contrasting cluster assignments," *Advances in neural information processing systems*, vol. 33, pp. 9912–9924, 2020.

[57] J.-B. Grill, F. Strub, F. Altché, C. Tallec, P. Richemond, E. Buchatskaya, C. Doersch, B. Avila Pires, Z. Guo, M. Gheshlaghi Azar *et al.*, "Bootstrap your own latent-a new approach to self-supervised learning," *Advances in neural information processing systems*, vol. 33, pp. 21 271–21 284, 2020.

[58] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman *et al.*, "Laion-5b: An open large-scale dataset for training next generation image-text models," *Advances in Neural Information Processing Systems*, vol. 35, pp. 25 278–25 294, 2022.

[59] K. Hamidieh, H. Zhang, W. Gerych, T. Hartvigsen, and M. Ghassemi, "Identifying implicit social biases in vision-language models," in *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, vol. 7, 2024, pp. 547–561.

[60] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," *arXiv preprint arXiv:2110.01963*, 2021.

[61] J. Dodge, M. Sap, A. Marasović, W. Agnew, G. Ilharco, D. Groeneveld, M. Mitchell, and M. Gardner, "Documenting large webtext corpora: A case study on the colossal clean crawled corpus," *arXiv preprint arXiv:2104.08758*, 2021.

[62] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," *arXiv preprint arXiv:1707.09457*, 2017.

[63] A. Birhane and V. U. Prabhu, "Large image datasets: A pyrrhic win for computer vision?" in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*. IEEE, 2021, pp. 1536–1546.

[64] D. Kaushik, E. Hovy, and Z. C. Lipton, "Learning the difference that makes a difference with counterfactually-augmented data," *arXiv preprint arXiv:1909.12434*, 2019.

[65] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd innovations in theoretical computer science conference*, 2012, pp. 214–226.

[66] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? debiasing word embeddings," *Advances in neural information processing systems*, vol. 29, 2016.

[67] J. Buolamwini and T. Gebru, "Gender shades: Intersectional accuracy disparities in commercial gender classification," in *Conference on fairness, accountability and transparency*. PMLR, 2018, pp. 77–91.

[68] V. Nagarajan and J. Z. Kolter, "Uniform convergence may be unable to explain generalization in deep learning," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[69] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language models are few-shot learners," *Advances in neural information processing systems*, vol. 33, pp. 1877–1901, 2020.

[70] G. GDPR, "General data protection regulation," *Regulation (EU)*, vol. 679, 2016.

[71] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial intelligence and statistics*. PMLR, 2017, pp. 1273–1282.

[72] P. Kairouz, H. B. McMahan, B. Avent, A. Bellet, M. Bennis, A. N. Bhagoji, K. Bonawitz, Z. Charles, G. Cormode, R. Cummings *et al.*, "Advances and open problems in federated learning," *Foundations and trends® in machine learning*, vol. 14, no. 1–2, pp. 1–210, 2021.

[73] C. Dwork, A. Roth *et al.*, "The algorithmic foundations of differential privacy," *Foundations and Trends® in Theoretical Computer Science*, vol. 9, no. 3–4, pp. 211–407, 2014.

[74] K. Bonawitz, V. Ivanov, B. Kreuter, A. Marcedone, H. B. McMahan, S. Patel, D. Ramage, A. Segal, and K. Seth, "Practical secure aggregation for federated learning on user-held data," *arXiv preprint arXiv:1611.04482*, 2016.

[75] N. Carlini, M. Jagielski, and I. Mironov, "Cryptanalytic extraction of neural network models," in *Annual international cryptology conference*. Springer, 2020, pp. 189–218.

[76] S. A. Osia, A. S. Shamsabadi, S. Sajadmanesh, A. Taheri, K. Katevas, H. R. Rabiee, N. D. Lane, and H. Haddadi, "A hybrid deep learning architecture for privacy-preserving mobile analytics," *IEEE Internet of Things Journal*, vol. 7, no. 5, pp. 4505–4518, 2020.

[77] N. Papernot, S. Song, I. Mironov, A. Raghunathan, K. Talwar, and Ú. Erlingsson, "Scalable private learning with pate," *arXiv preprint arXiv:1802.08908*, 2018.

[78] D. Zhu, J. Chen, K. Haydarov, X. Shen, W. Zhang, and M. Elhoseiny, "Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions," *arXiv preprint arXiv:2303.06594*, 2023.

[79] A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever, "Zero-shot text-to-image generation," in *International conference on machine learning*. Pmlr, 2021, pp. 8821–8831.

[80] A. Ramesh, P. Dhariwal, A. Nichol, C. Chu, and M. Chen, "Hierarchical text-conditional image generation with clip latents," *arXiv preprint arXiv:2204.06125*, vol. 1, no. 2, p. 3, 2022.

[81] J. Fu, S. Huangfu, H. Fei, X. Shen, B. Hooi, X. Qiu, and S.-K. Ng, "Chip: Cross-modal hierarchical direct preference optimization for multimodal llms," *arXiv preprint arXiv:2501.16629*, 2025.

[82] F. Ma, Y. Zhou, H. Li, Z. He, S. Wu, F. Rao, Y. Zhang, and X. Sun, "Ee-mllm: A data-efficient and compute-efficient multimodal large language model," *arXiv preprint arXiv:2408.11795*, 2024.