MULTIDISCIPLINARY RESEARCH INSTITUTE INDIA

International Journal on Advanced Computer Theory and Engineering

ISSN: 2319 - 2526

Volume 15 Issue 2 (2026) | Pages: 142-156

Multimodal Foundation Models: Architectures, Challenges, and Applications

Hayder Kareem Algabri[1], Hayder Al-Ghanimi[2], K G Kharade[3]

[1]Department of Cybersecurity Techniques Engineering , College of Engineering Techniques,University of Hilla.
[2]Department of Artificial Intelligence Techniques Engineering, Hilla University College, Babylon 51001, Iraq
[3]Department of Computer Science, Shivaji University Kolhapur, Maharashtra, India
[1]hayder_kareem_algabri@hilla-unc.edu.iq, [2]hayder_alghanimi@hilla-unc.edu.iq, [3]kgk_csd@unishivaji.ac.in

Peer Review Information

***Type**: Article*
***Received**:15 May 2026*
***Revised**: 19 June 2026*
***Accepted**: 20 July 2026*
***Published**: 18 Aug 2026*

Abstract

The fast progress of AI has significantly changed the research paradigm from single-modality systems that operate on independent modalities to multimodal systems that can handle multimodal information from text, image, sound, video, and sensor data. This transformation is primarily driven by Multimodal Foundation Models (MFMs) that leverage the large-scale pre-training on vast collections of data to achieve cross-modal understanding and generation skills. The current paper provides an overview of multimodal deep learning research in terms of its history ranging from early fusion approaches to recent Transformer-based multimodal networks. The existing models are classified based on their alignment strategy, namely contrastive, generative, and hybrid strategies along with mathematical formulations of some major objective functions like InfoNCE loss and cross-attention. Moreover, the paper surveys the state-of-the-art in vision-language tasks, embodied robotics, healthcare, and scientific discovery applications, evaluating the challenges of multimodal AI such as hallucinations, data biases, computation issues, and evaluation metrics like FID, CIDEr, and POPE. In conclusion, the current trends in multimodal AI are reviewed in terms of neuro-symbolic fusion, embodied AI, and edge computing.

Keywords: Multimodal Deep Learning; Transformers; Foundation Models; Cross-Modal Representation; Computer Vision; Natural Language Processing; Large Multimodal Models.

How to Cite This Article

Algabri, H. K., Al-Ghanimi, H., & Kharade, K. G. (2026). Multimodal foundation models: Architectures, challenges, and applications. *International Journal on Advanced Computer Theory and Engineering,* 15(2), 142–156.

Introduction

Background and Motivation

Multimodalism lies at the heart of human perception. Humans perceive the world by interpreting the scene through multiple channels of vision, spoken language, sounds, touch, and smell all at once. Multimodal processing helps in resolving ambiguity because something that is ambiguous in one sense can become unambiguous when supplemented with information from another mode. For instance, a loud sound of breaking glass coupled with visual flashes help to infer immediately the occurrence of an accident; neither sound nor flashes alone could resolve the ambiguity. For a long time, AI research pursued each mode in isolation: Computer Vision focused on pixels, NLP on token strings, and Speech on spectrograms. While remarkable results have been achieved within each individual realm, the lack of cross-modal interplay prevented AI from reasoning about situations in the real world where the context is spread across multiple sources of information.

The combination of CV and NLP resulted in the development of Multimodal Deep Learning which is a wide range of neural network architectures for processing different types of input. It should be emphasized that not every multimodal deep learning model can be considered a foundation model. The traditional approach to multimodal learning was associated with training on labeled datasets for specific goals (such as generating captions to images and answering questions about images). In turn, Multimodal Foundation Models (MFM) can be defined through the following features: (1) being trained in large scale on wide and often unlabeled datasets; (2) applying self-supervised or weakly supervised objectives; (3) having the ability to transfer knowledge to different downstream tasks through prompting and fine-tuning; and (4) having emergent behavior not explicitly coded. It is important to understand this difference when discussing the changes in this domain.

The introduction of the Transformer architecture [1] led to the creation of MFM such as CLIP [2], DALL-E 3 [3], GPT-4V [4], and LLaVA [5]. They have emergent abilities to align, generate, and reason between different modalities with high precision.

The Shift to Foundation Models

The transition from task-based models to foundation models is an essential restructuring of scaling laws and training procedures. Previously, models used to be designed for specific tasks and trained using labeled data relevant to the task. However, foundation models undergo training on comprehensive large data, often using self-supervised learning methods, and possess emergent properties that are not hard-coded. In a multimodal scenario, this means that the model will be able to generate image captions, answer questions about the visual content, and generate images without any task-specific changes in architecture. This versatility is what motivates multimodal AI research today.

Scope and Contributions

However, even with the rapidly advancing field of research, there is still little unity in multimodal learning literature, as some works concentrate on architectural efficiency, while other ones are focused on alignment techniques, while another set of papers concentrates on particular applications. It is important to have an overview of multimodal learning that would connect architectural decisions with representation learning goals and further performances. The aim of this work is to provide a comprehensive analysis of the current state of the field of multimodal artificial intelligence. The following questions will be addressed in this paper:

1. Architectures: What are the predominant architectural paradigms used in current multimodal systems, how do they handle different data types?
2. Representation: How is cross-modal representation learning carried out through contrastive and generative learning objectives?
3. Applications: What are the main applications of these systems in science and industry, what metrics are considered success?
4. Challenges: What are the bottlenecks of research related to hallucinations, biases, efficiency, and evaluation?
5. Future: In what directions is the field developing, such as embodied intelligence and neuro-symbolic reasoning?

This work makes four main contributions:

- Taxonomy: We propose taxonomy of multimodal fusion approaches and architectures based on interaction depths and learning objectives.
- Technical Rigor: We introduce rigorous mathematical representations of key algorithms, such as self-attention, cross-attention and InfoNCE loss function, with full definitions of variables.
- Architecture Comparing: We compare different projection mechanisms (Linear, MLP, Q-Former, Perceiver) and discuss their influence on information preservation and system performance.
- Limitations Evaluation: We introduce a critical assessment of limitations of current multimodal fusion systems, with particular benchmarks for hallucination (POPE) and generation quality (FID, CIDEr)..

Organization

This article is structured as follows. Section 2 provides a review of the preliminaries for modalities and the Transformer architecture. Section 3 explains architectural approaches while focusing on projections. Section 4 focuses on representation learning methods along with the math behind them. Section 5 looks at the training data and approaches, taking into account copyright issues. Section 6 considers possible applications. Section 7 discusses challenges, which include assessing hallucinations. Section 8 considers future directions with examples of neuro-symbolic systems. Section 9 is the conclusion.

Preliminaries

Defining Modalities

In the context of deep learning, a modality refers to a specific type of data or sensory input. Common modalities include:

- Visual: Static images (spatial), video sequences (spatio-temporal), depth maps, 3D point clouds.
- Textual: Natural language sentences, code, structured data (JSON/XML).
- Auditory: Speech, environmental sounds, music (temporal continuous signals).
- Sensorimotor: LiDAR, radar, proprioceptive data (joint angles, velocity), tactile feedback.

One of the primary differences between the modalities is in terms of their temporal nature. Audio and video are inherently temporal, which means that there has to be a model that accounts for dynamics and dependence between different samples. Images are not temporal but are purely spatial. On the other hand, text is sequential.

The process of multimodal learning is defined as using the set of M modalities $\{X_1, X_2, ..., X_m\}$ and finding a representation Z, which captures the shared semantics between them. The problem lies in the fact that the heterogeneity of these modalities is too high: text is sequential and discrete; images are continuous and spatial; audio is continuous and temporal.

The Transformer Architecture

The Transformer [1] forms the basis for most modern MFMs. It uses the self-attention mechanism for capturing the dependencies between the elements in the sequence, regardless of their distance from each other. If X ∈ R^(n × d_model) is a sequence of embeddings as the input, where n is the length of the sequence and d_model is the embedding dimension, then:

First, the input is linearly projected to queries, keys, and values:

$$Q = XW^Q, K = XW^K, V = XW^V$$

where $W^Q, W^K \in \mathbb{R}^{d_{model} \times d_k}$ and $W^V \in \mathbb{R}^{d_{model} \times d_v}$ are learnable projection matrices, and d_k is the dimension of queries and keys.

The scaled dot-product attention is then computed as:

$$Attention(Q, K, V) = softmax\left(\frac{QK^\top}{\sqrt{d_k}}\right)V$$

where the $softmax$ function is applied row-wise, and $\sqrt{d_k}$ is a scaling factor to prevent dot products from growing too large in magnitude.

Multi-head attention allows the model to jointly attend to information from different representation subspaces:

$$MultiHead(Q, K, V) = Concat(head^1, \dots, head_h)W^O$$

where $head_i = Attention\left(QW_i^Q, KW_i^K, VW_i^V\right)$, and $W^O \in \mathbb{R}^{hd_v \times d_{model}}$ is the output projection matrix.

In multimodal setups, this method is modified in order to allow tokens in one modality (e.g., text tokens) to attend to tokens in other modalities (e.g., image tokens) through cross-attention, which is the primary way of fusing multiple modalities in modern models.

Embedding Spaces

An important idea in multimodal learning is the shared embedding space. This is because the main goal is to find a mapping for the inputs of different modalities in a single latent space R^d where similar concepts lie close regardless of their modalities. Thus, the embeddings of "a red apple" (text input) and "image of a red apple" (image input) must be close together in the embedding space.

Architectural Paradigms in Multimodal Learning

The major challenge in multimodal learning is the problem of fusion – combining different forms of data representations in one latent space. The architectures can be categorized in three main paradigms based on when and how inter-modal communication takes place.

Early Fusion (Data-Level)

Early fusion is the fusion of raw data or low-level features before they are fed into a deep learning network.

- Procedure: This could mean combining pixel data with text representations in the input layer or combining audio spectrograms with image channels.
- Benefits: The neural network can learn low-level cross-modal relations right away, some of which might be missed during processing the individual modalities separately.
- Drawbacks: Computationally expensive and highly vulnerable to loss of one modality. Matching of different temporal and spatia1 scales (i.e., matching of 1 second of audio with one video frame) is difficult.
- State of the Art: Obsolete in modern foundation models in favor of deep fusion techniques but still employed in certain sensor fusion cases (autonomous vehicle lidar + camera).

Late Fusion (Decision-Level)

Each late fusion process involves the separate encoding of each modality with different encoders and fusion of the resulting encodings at the decision level.

How it works: Different independent encoders (for example, ViT for vision, BERT for text) generate features Z_img and Z_txt that are concatenated or averaged before being fed into the common classifier/head.

Advantages: Very modular design allowing the employment of well-performing pre-trained unimodal models; resilient to the absence of any modality at inference time, as only one branch of the model is needed.

Disadvantages: Limits the possibility of using cross-modal information during feature extraction, as the vision encoder is not able to make use of text information when encoding visual features.

Hybrid and Transformer-Based Fusion

State-of-the-art architectures leverage the Transformer model for achieving deeper interactions between the modalities in the network's architecture. An essential design choice in such architecture involves the projection mapping process where individual modality features are projected onto a unified embedding space.

Cross-Attention Mechanisms

Modeling techniques like Perceiver IO [6] and Flamingo [7] utilize cross-attention processes where queries from one domain attend to the keys and values from another domain. As an illustration, in a Vision-Language Model (VLM), the text queries the image patch embeddings to extract visual information necessary for answering the question.

Projection Mechanisms

A Comparative Analysis: The choice of projection mechanism significantly impacts information preservation, training stability, and downstream performance. We identify four primary approaches:

1. Linear Projection: The simplest approach uses a single linear layer to map visual features to the text embedding space:

$$Z_{proj} = Z_{img}W_p + b_p$$

where $W_p \in \mathbb{R}^{d_{img} \times d_{text}}$ is the projection matrix.

- Advantages: Computationally efficient, minimal parameters, easy to train.
- Disadvantages: Limited expressiveness, may lose complex visual information due to the linear constraint.
- Example: LLaVA [5] uses a simple linear layer to project CLIP vision encoder features into the LLaMA language model space.

2. Multi-Layer Perceptron (MLP) Projection: MLP projections use multiple non-linear layers:

$$Z_{proj} = \sigma(Z_{img}W^1 + b^1)W^2 + b^2$$

where σ is a non-linear activation function (e.g., GELU).

- Advantages: Greater expressiveness than linear projection, can capture non-linear relationships.
- Disadvantages: More parameters, risk of overfitting, may still compress information excessively.
- Example: BLIP-2 [11] uses a two-layer MLP in some configurations.

3. Q-Former (Query Transformer): Introduced in BLIP-2 [11], the Q-Former uses a set of learnable query vectors that interact with image features through cross-attention:

$$Q_{learned} \in \mathbb{R}^{N_q \times d}, Z_{proj} = CrossAttention(Q_{learned}, Z_{img}, Z_{img})$$

- Advantages: Extracts the most relevant visual information for the language task, acts as an information bottleneck that filters noise.
- Disadvantages: May discard fine-grained details, requires careful tuning of the number of queries N_q.
- Example: BLIP-2 uses 32 learnable query tokens to extract visual features.

4. Perceiver Resampler: Used in Flamingo [7] and AlphaFold, the Perceiver uses iterative cross-attention to compress variable-length inputs into a fixed number of latent tokens:

$$L^{t+1} = CrossAttention(L^t, Z_{img}, Z_{img}) + SelfAttention(L^t)$$

where L are the latent tokens iteratively refined over T steps.

- Advantages: Handles variable-length inputs efficiently, can process high-resolution images by compressing them.
- Disadvantages: Computationally intensive due to iterative refinement, may lose spatial information.
- Example: Flamingo uses a Perceiver Resampler with 64 latent tokens.

Table 1. *Comparison of Projection Mechanisms*

Mechanism	Parameters	Information Preservation	Computational Cost	Best For
Linear	Low	Moderate (linear only)	Very Low	Simple tasks, limited compute
MLP	Medium	Good (non-linear)	Low	General-purpose VLMs
Q-Former	Medium-High	Selective (task-relevant)	Medium	Instruction-tuned models
Perceiver	High	Compressed (iterative)	High	High-resolution inputs, variable length

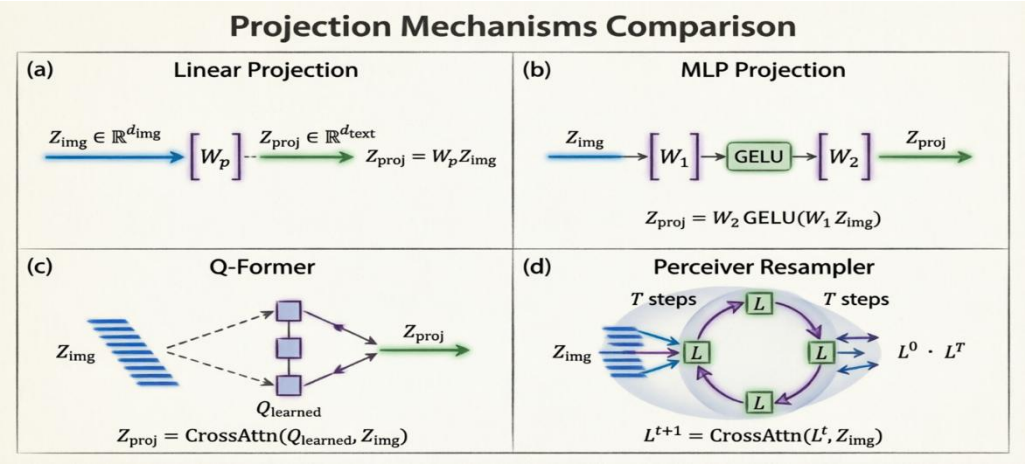

Fig. 1. *Comparison of projection mechanisms for mapping visual features to language embedding space. (a) Linear projection: single matrix transformation. (b) MLP projection: two-layer non-linear mapping. (c) Q-Former: learnable query tokens extract task-relevant features via cross-attention. (d) Perceiver Resampler: iterative refinement of latent tokens through multi-step cross-attention. Input/output dimensions and key equations are annotated.*

Unified Tokenization

Approaches like ImageBind [8] or treating image patches as text tokens (as in LLaVA [5]) allow a single Transformer to process all data types seamlessly. Images are patchified and projected to the same dimension as text embeddings, creating a single sequence input for the Transformer.

Encoder-Decoder vs. Causal LM

- *Encoder-Decoder (e.g., OFA [9], UniLM):* Uses bidirectional attention for the encoder and unidirectional for the decoder. Good for understanding and generation tasks where the target is structured.

- *Causal Language Models (e.g., LLaVA, GPT-4V):* Treats all modalities as a sequence of tokens to be predicted autoregressively. This has become the dominant architecture for general-purpose assistants because it leverages the vast ecosystem of pre-trained Large Language Models (LLMs).

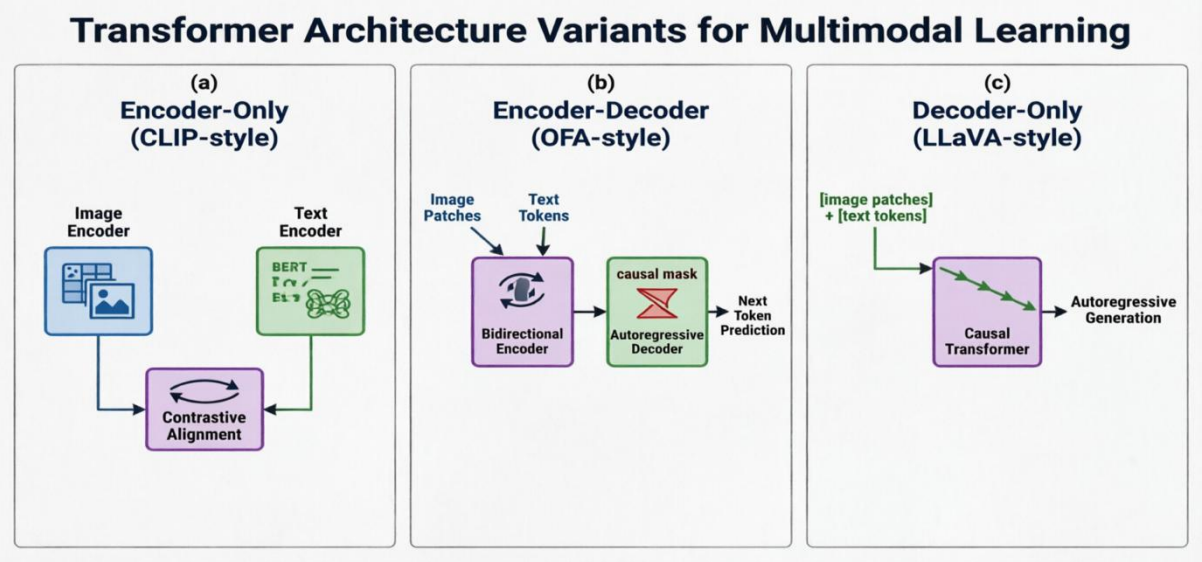

Fig. 2. *Transformers for multimodal learning. (a) Encoder-only (CLIP): different modality-specific encoder modules, aligned using contrastive loss. (b) Encoder-decoder (OFA): bidirectional encoding with autoregressive decoding. (c) Decoder-only (LLaVA): a single causal transformer which deals with the concatenation of multimodal tokens. Arrow types depict the type of attention.*

Model Zoo: Representative Architectures

To illustrate the architectural diversity, we examine key models:

1. CLIP (Contrastive Language-Image Pre-training): Uses a dual-encoder architecture (late fusion) where image and text encoders are trained to align outputs via contrastive loss. It does not fuse features internally but aligns the final embeddings.
2. Flamingo [7]: Uses a pre-trained LLM and injects visual information via Perceiver Resampler layers that use cross-attention to compress visual features into a set of latent tokens inserted into the text sequence.
3. LLaVA (Large Language-and-Vision Assistant): Projects image features into the text embedding space using a simple linear layer, allowing the LLM to process images as if they were soft prompts.
4. Stable Diffusion: Uses a latent diffusion model where text embeddings (from CLIP) condition the denoising process via cross-attention layers in the U-Net backbone.

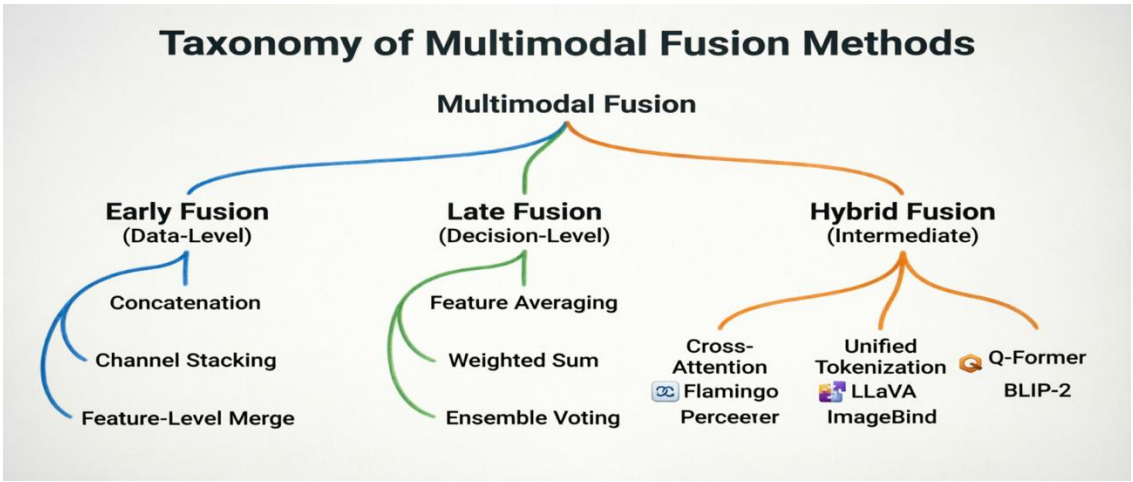

Fig. 3. *Taxonomy of multimodal fusion methods. Early fusion combines raw inputs or low-level features; late fusion merges modality-specific outputs at the decision layer; hybrid fusion enables intermediate cross-modal interaction via cross-attention, unified tokenization, or query-based mechanisms. Example models are indicated at leaf nodes.*

Cross-Modal Representation Learning

To enable reasoning across modalities, models must learn a shared embedding space where semantically similar concepts are close together. This section details the objective functions used to achieve this.

Contrastive Learning

Contrastive learning aligns modalities by maximizing the similarity of paired data and minimizing the similarity of unpaired data within a batch.

Key Model: CLIP [2].

Loss Function: The InfoNCE (Information Noise Contrastive Estimation) loss is typically used. For a batch of N image-text pairs, let $v_i \in \mathbb{R}^d$ denote the embedding of the $i-th$ image and $t_i \in \mathbb{R}^d$ denote the embedding of the $i-th$ text. The similarity between an image and text is computed using cosine similarity:

$$sim(v_i, t_j) = \frac{(v_i^\top t_j)}{(\|v_i\|\|t_j\|)}$$

The InfoNCE loss for the image-to-text direction is:

$$L_{i2t} = -\left(\frac{1}{N}\right)\Sigma_{\{i=1\ to\ N\}}\log\left[\frac{\exp\left(\frac{sim(v_i,t_i)}{\tau}\right)}{\Sigma_{\{j=1\ to\ N\}}\exp\left(\frac{sim(v_i,t_j)}{\tau}\right)}\right]$$

Similarly, the text-to-image loss is:

$$L_{t2i} = -\left(\frac{1}{N}\right)\Sigma_{\{i=1\ to\ N\}}\log\left[\frac{\exp\left(\frac{sim(t_i,v_i)}{\tau}\right)}{\Sigma_{\{j=1\ to\ N\}}\exp\left(\frac{sim(t_i,v_j)}{\tau}\right)}\right]$$

The total contrastive loss is the average of both directions:

$$L_{contrastive} = ½(L_{i2t} + L_{t2i})$$

where:

- N is the batch size (number of image-text pairs)
- τ is a learnable temperature parameter that controls the concentration level of the distribution
- v_i is the embedding of the $i-th$ image
- t_i is the embedding of the $i-th$ text
- $sim(\cdot,\cdot)$ is the cosine similarity function

Impact: Enables zero-shot transfer. A model trained on image-text pairs can classify images into categories it never saw during training by matching image embeddings to text label embeddings (e.g., matching an image to the text "a photo of a [class]").

Limitations: Contrastive learning requires large batch sizes to provide enough negative samples, which is memory-intensive. It also focuses on global alignment, potentially ignoring fine-grained regional details.

Generative Learning

Generative approaches focus on reconstructing data or generating new content in one modality conditioned on another.

Masked Modeling: Models like BEiT [10] or VideoMAE mask portions of the input (image patches or text tokens) and task the model with reconstructing them. In multimodal settings (e.g., VQGAN-CLIP), the model might reconstruct an image conditioned on text.

Diffusion Models: In text-to-image generation (e.g., Stable Diffusion, DALL-E 2), the text encoder guides the diffusion process. The loss function minimizes the difference between the predicted noise and the actual noise added to the latent image representation, conditioned on the text embedding:

$$L_{diffusion} = E_{\{x^0,\varepsilon,t,c\}}\left[\left\|\varepsilon - \varepsilon_{\theta(x_t,t,c)}\right\|^2\right]$$

where:

- x^0 is the original image
- $\varepsilon \sim N(0, I)$ is the noise
- t is the timestep
- c is the text conditioning embedding
- x_t is the noisy image at timestep t
- ε_θ is the noise prediction network

Advantage: Often yields richer representations than contrastive learning as the model must understand the full distribution of the data, not just pairwise alignment. It forces the model to learn detailed structural information to successfully reconstruct the input.

Hybrid Objectives

State-of-the-art models often combine contrastive and generative losses.

BLIP (Bootstrapping Language-Image Pre-training): Uses a multimodal mixture of encoder-decoder architectures. It employs:

1. Contrastive loss for image-text matching
2. Image-grounded text generation loss for captioning
3. Text-grounded image generation loss

The total loss is:

$$L_{BLIP} = \lambda^1 L_{contrastive} + \lambda^2 L_{generation} + \lambda^3 L_{matching}$$

LiT (Language-Image Transfer): LiT [12] is a less common but important approach that locks a pre-trained text encoder and trains an image encoder to align with it using contrastive loss, then fine-tunes with generative tasks. The acronym stands for "Locked-image text Tuning" and represents a transfer learning approach where knowledge from a strong text encoder is transferred to the visual domain.

Benefit: This combination leverages the retrieval strengths of contrastive learning and the reasoning strengths of generative modeling.

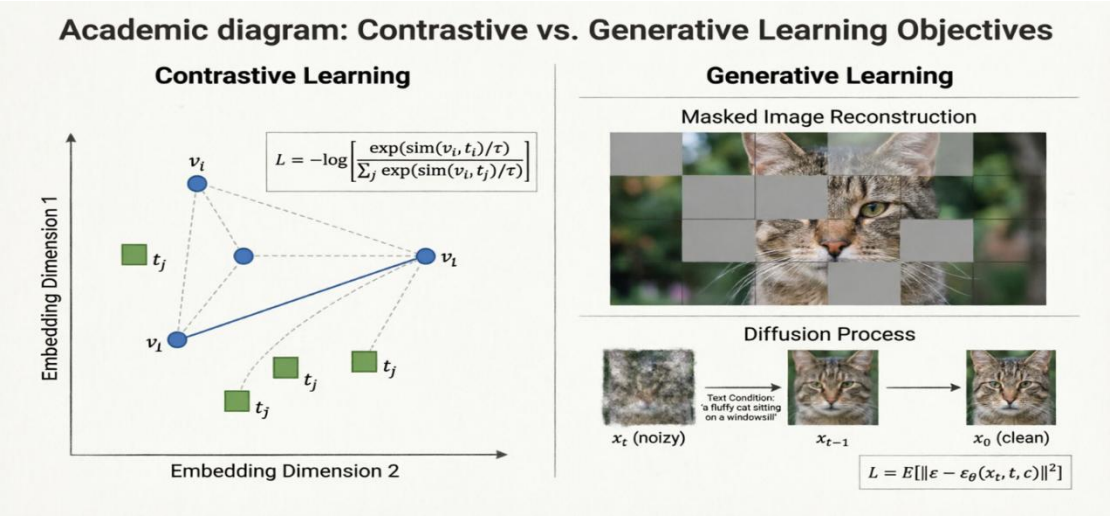

Fig.3. *Comparison of contrastive and generative learning objectives. Left: Contrastive learning aligns matched image-text pairs (solid lines) while separating unmatched pairs (dashed lines) in a shared embedding space using the InfoNCE loss. Right: Generative learning reconstructs masked inputs or denoises latent representations conditioned on cross-modal context, optimized via reconstruction or diffusion losses.*

Instruction Tuning and Alignment

These latest developments focus on making the models aligned with the intent of humans. Approach: Training a pre-trained Multimodal Foundation Model (MFM) on datasets containing (Instruction, Image, Response) triplets. For example, a picture of a kitchen along with the instructions "What appliance is present on the counter?" followed by the response "A microwave."

Effect: It converts the perceptual model to an interactive assistant model. This is very important for reducing hallucinations as well as aligning with the constraints set by the user. Reinforcement Learning from Human Feedback (RLHF): Expands on instruction tuning where human preference data is used to train a reward model that further optimizes the multimodal policy using PPO.

Training Data and Strategies

The performance of Multimodal Foundation Models is inextricably linked to the scale and quality of their training data.

Large-Scale Web Datasets

The growth of multimodal learning has been powered by web data collections. Conceptual Captions consists of around 3.3 million images paired with corresponding alt-text collected online. The dataset called LAION (Large-scale Artificial Intelligence Open Network) contains LAION-5B with 5.85 billion pairs of images and texts filtered using the CLIP model. This collection made it possible to train Stable Diffusion and different CLIP open-source implementations. The Common Crawl dataset has been used a lot for training the textual part of multimodal networks.

The issue of using web-sourced data regarding copyrights became controversial and raised much attention from legal and ethical aspects. Recently, Getty Images sued Stability AI, accusing it of illegal training of AI models on copyrighted pictures. Important problems are as follows: (1) Fair Use Doctrine—is it possible to train models on copyrighted materials according to fair use doctrine, (2) Opt-out—some data collections now respect robots.txt and give opt-out opportunities to content creators, (3) Licensing—the new generation of collections like LAION considers Creative Commons licensing and permission-based solutions and (4) Transparency—the demand for more transparent disclosure of training data sources by model developers increases.

Problems with the above-listed datasets include high noise caused by the lack of correspondence between image and text and biased data in favor of Western culture and gender stereotypes. Filtering strategies (such as CLIP scoring) are currently common practices to mitigate them.

High-Quality Curated Datasets

When it comes to instruction-tuning and tasks involving heavy reasoning, it is better to choose datasets which are small in size but of high quality.

The datasets COCO and Visual Genome are commonly considered benchmarks for captioning and VQA tasks, and are often used in fine-tuning processes.

LLaVA-Instruct is an instruction-tuned dataset created using GPT-4 and is used in the training process of VLMs.

ScienceQA deals with science-based images and related questions and is applied in reasoning evaluation.

Training Stages

The traditional MFM training pipeline has three main steps:

(1) the step of pre-training on a large dataset available on the web (e.g., LAION) by employing self-supervised learning methods (contrastive or generative), which helps in acquiring foundational alignment; (2) the step of instruction tuning by fine-tuning on instruction data sets to gain conversational ability and the skill to follow the instructions; and (3) the step of alignment through reinforcement learning from human feedback (RLHF).

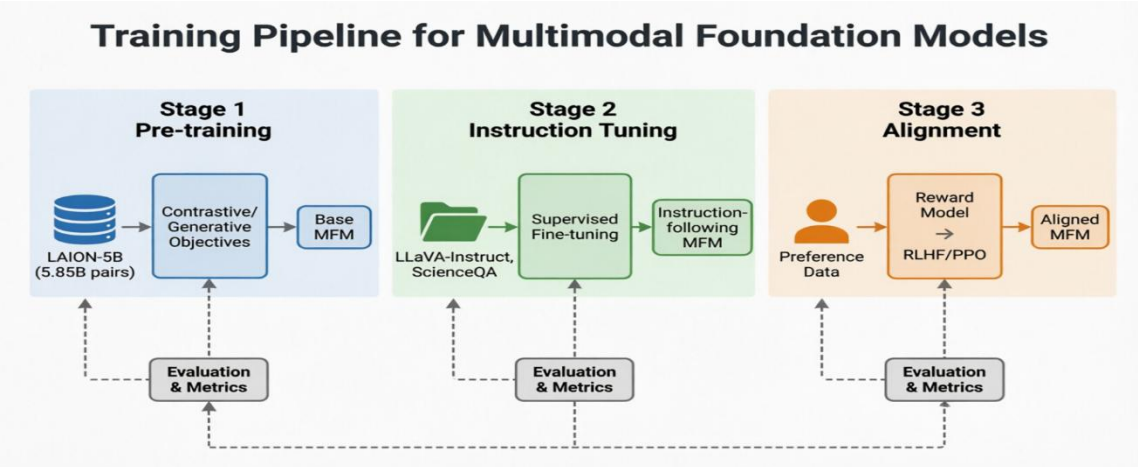

Fig. 4. *Three-stage pipeline for multimodal foundation models. Stage 1 (Pre-training): massive scale weakly supervised learning using web-based data. Stage 2 (Instruction Tuning): supervised fine-tuning using curated datasets based on instructions. Stage 3 (Alignment): preference optimization using reinforcement learning from human feedback/ PPO.*

Efficient Training Techniques

Considering the sheer size of the data, efficiency becomes of paramount importance. The following are some of the techniques that can be used to achieve efficiency:

- Mixed precision training: The use of FP16 or BF16 to save on memory.
- Gradient checkpointing: The process of trading memory for computation in the form of re-computing activations during back propagation.
- Distributed training: Using data parallelism and model parallelism with the help of libraries such as DeepSpeed or FSDP.

Applications

The versatility of multimodal models has unlocked applications across diverse domains. We categorize these by industry and function.

Vision-Language Understanding

Visual Question Answering (VQA): In this task, the system generates answers to queries regarding natural language questions pertaining to images. Modern multimodal foundation models (MFMs) have achieved almost human-level proficiency in tasks such as VQA v2. Some real-world uses include the counting of objects through querying and accessibility.

Image Captioning: In this task, captions are generated for images. It is useful for accessibility devices for people who are visually impaired and for indexing large media repositories. Evaluation is done using metrics such as:

- CIDEr (Consensus-based Image Description Evaluation): It is evaluated for similarity to the human reference using TF-IDF scoring.
- BLEU (Bilingual Evaluation Understudy): It measures n-gram precision.
- ROUGE (Recall-Oriented Understudy for Gisting Evaluation): It measures n-gram recall.

Document Analysis (OCR + Layout): Donut and Nougat are some examples of models that incorporate OCR with language modeling to generate structured text or markdown from scanned documents, receipts, and scientific papers.

Content Generation

Text-to-Image/Video. Generating artistic or realistic visual artifacts based on textual descriptions (such as Midjourney, Sora, DALL-E 3) has revolutionized creative industries by allowing fast prototyping for designers and filmmakers. There are several metrics used for assessing the quality of generated artifacts such as:

- FID (Fréchet Inception Distance): Calculates the distance between the distributions of the generated images and those of real images within the feature space of Inception networks; the smaller value shows higher quality.
- IS (Inception Score): Estimates the quality and diversity of the generated images.
- CLIP Score: Estimates the alignment of the generated images with the text prompt using CLIP embeddings.

Editing. The model is asked to edit the image using natural language instructions like “Change the shirt color to red” or “Eliminate the background.” InstructPix2Pix is one of the key models in this area.

3D Generation. New models generate three-dimensional artifacts (meshes, NeRFs) based on textual prompts, hence accelerating game development and creation of AR/VR content.

Robotics and Embodied AI

Navigation: Robots use visual perception along with language instructions (such as "go to the kitchen") for navigation. The RT-1 model (Robotics Transformer) uses perceptual pixel and language inputs to generate action tokens.

Manipulation: Multimodal policies allow robots to perform grasping tasks using vision and touch.

RT-2 (Robotics Transformer): A vision-language-action framework for learning robotic skills from pre-training on web-scale data. For example, the framework can understand the task "pick up the thirsty animal" and find the toy dog near the water bowl using knowledge learned during language pre-training.

Healthcare and Science

Medical Imaging: Matching radiology images and MRI scans with clinical notes to assist the diagnostic process. Examples of models which have competence in processing multimodal medical imaging data are Med-PaLM M models.

Drug Discovery: Combining molecular graphs with the textual research papers to predict interactions between drugs and folding proteins. Example: AlphaFold combines several multiple sequence alignments which can be considered as multimodal data.

Climate Change: Interpreting satellite images along with climate sensor data to predict weather conditions or cases of deforestation.

Education and Accessibility

Tutoring systems which operate using multimodal techniques have the ability to interpret the mathematical problems written by hand (as images) by students and then provide them with feedback through text. Sign language translation systems are able to translate the inputted sign language video into text or voice.

Challenges and Limitations

Despite significant progress, Multimodal Foundation Models face critical hurdles that must be addressed before widespread deployment in high-stakes environments.

Hallucination and Faithfulness

Multimodal LLMs often generate text that is fluent but factually inconsistent with the provided image. This is known as hallucination.

Types:

- Object Hallucination: Describing an object not present in the image.
- Attribute Hallucination: Incorrect color, count, or size.
- Relationship Hallucination: Incorrect spatial arrangement or interactions.

Causes: Often stems from the language prior being too strong; the model predicts the next likely word based on text statistics rather than visual evidence.

Evaluation of Hallucination: Recent benchmarks have been developed to systematically evaluate hallucination:

- POPE (Polling-based Object Probing Evaluation): [13] Presents the model with yes/no questions about whether specific objects exist in an image. POPE uses three sampling strategies:
  - *Random:* Negative objects are sampled randomly from the dataset.
  - *Popular:* Negative objects are chosen from the most common objects in the dataset.
  - *Adversarial:* Negative objects are chosen based on co-occurrence statistics to be most confusing.

POPE measures accuracy, precision, recall, and F1 score on these binary classification tasks.

- CHAIR (Captions Hallucination Assessment for Image Recognition): Measures the proportion of objects mentioned in a caption that are not present in the image.
- MMHal-Bench: Evaluates hallucination across multiple modalities and question types.

Mitigation Strategies:

- Reinforcement Learning from Visual Feedback (RLVF): Using visual consistency as a reward signal.
- Explicit Grounding Losses: Adding loss terms that penalize mentions of objects not detected by an object detector.
- Retrieval-Augmented Generation (RAG): Querying a database to verify facts before generation.
- Contrastive Decoding: Comparing outputs from the full model vs. a model with ablated visual inputs to detect language-only biases.

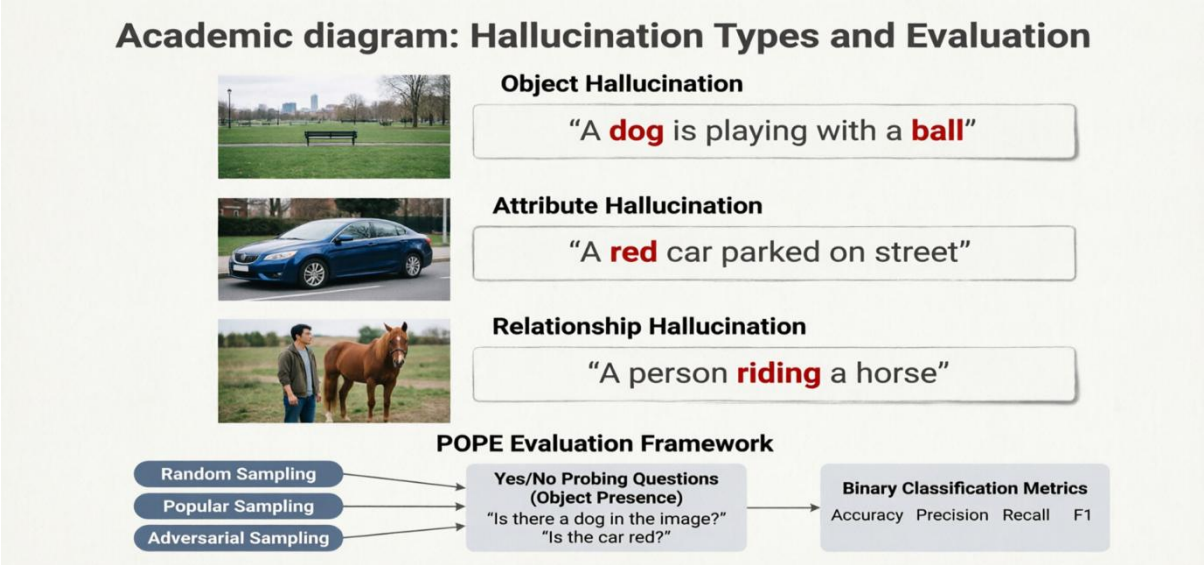

Fig. 6.*Hallucination classes and evaluation method. The first three rows show the three main classes of hallucination that can occur in the output of AI models: object hallucination (description of non-existent objects), attribute hallucination (wrong properties of existing objects), and relation hallucination (wrong information about the relationship between objects). The lower part is the POPE method for evaluating hallucinations, using yes/no probes along with negative samples.*

Data Bias and Ethics

Multimodal Foundation Models (MFMs) are trained using large amounts of data obtained by web scraping, for instance, from datasets like LAION, where there is an inherent presence of societal bias concerning gender, race, and culture. The main concern is the presence of stereotypes, such as when the word "nurse" is associated more with women than men or when the word "CEO" is associated more with men. There is also a possibility that generative models will filter out images depicting certain racial groups because of excessively strict safety filters. The solution includes using curated datasets, debiasing loss functions, and clearly reporting the makeup of datasets.

Computational Cost and Efficiency

Training MFMs requires thousands of GPUs and high energy costs. The inference delay problem is increased by the fact that inference is made difficult due to high-resolution images together with large amounts of text data. The $O(n^2)$ complexity of the attention mechanism, with n being the sequence length, is too high for high-resolution inputs.

Solutions:

- Mixture of Experts (MoE): Activates only a subset of parameters during inference.
- Quantization: Reducing precision to 4-bit or 8-bit.
- Distillation: Training smaller student models.
- Sparse Attention: Using linear or local attention mechanisms to reduce complexity to $O(n)$ or $O(n \log n)$.

Evaluation Benchmarks

Standard benchmarks (e.g., ImageNet) are saturated. New benchmarks are needed to evaluate complex reasoning, causality, and long-context understanding.

Problem: Models can achieve high scores by exploiting dataset biases rather than true understanding.

New Metrics and Benchmarks:

- MME (Multimodal Evaluation): Comprehensive benchmark covering perception and cognition.
- Seed-Bench: Focuses on generative comprehension.
- MMMU (Massive Multitask Language Understanding): Tests college-level reasoning across domains.
- Human Evaluation: Remains the gold standard for generative quality, particularly for coherence, relevance, and helpfulness.

Privacy and Security

Data Leakage: Algorithms trained on data available from open sources could potentially remember sensitive data like faces and addresses.

Adversarial Attacks: Small changes in input images could cause misclassification or production of malicious text. Multimodal algorithms can be attacked through exploiting the alignment of modalities.

Membership Inference Attack: The inference whether the given datum is a part of the training set for the algorithm.

Future Directions

The next phase of multimodal AI research will likely focus on moving from passive perception to active reasoning and interaction.

From Perception to Reasoning

Current models excel at recognition but struggle with complex logical reasoning (e.g., counterfactuals, multi-step planning).

Neuro-Symbolic AI: Future architectures may integrate neural networks with symbolic logic engines. The neural component handles perception (seeing the objects), while the symbolic component handles reasoning (applying rules).

Concrete Integration Example: Consider a multimodal system for visual reasoning:

1. Neural Module: A vision transformer detects objects and relationships: Detect(image) → {(cup, on, table), (book, next to, cup)}
2. Symbolic Module: A logic engine applies rules:
   - Rule 1: $\forall x, y. on(x, y) \wedge fragile(x) \rightarrow support_{required(y)}$
   - Rule 2: $\forall x. book(x) \rightarrow \neg fragile(x)$
   - Rule 3: $\forall x. cup(x) \rightarrow fragile(x)$
3. Integration: The neural module extracts facts, the symbolic module reasons over them, and the output is grounded back into natural language.

This ensures consistent reasoning and explainability, as the symbolic rules are human-readable and verifiable.

Chain-of-Thought (CoT): Extending CoT prompting to multimodal inputs, where the model explicitly describes its visual reasoning steps before answering.

Embodied Multimodal Intelligence

Moving from screen-based AI towards agents that work in the physical world involves bringing in proprioception, depth perception, and real-time latencies into the fundamental building blocks of such an approach. World models consist of learning how to predict the results of actions in the physical world (for example, “If I push this cup, it will fall”).

Personalization and Privacy

Training models on user-specific data without compromising privacy.

Federated Learning: Training models across decentralized devices (phones, laptops) keeping data local.

LoRA (Low-Rank Adaptation): Enabling personalized multimodal assistants that run on edge devices by fine-tuning only small adapter layers specific to the user's style and data.

Omni-Modal Models

Going Beyond Text & Images towards Audio, Video, 3D Point Clouds, and Sensor Data using a unified framework.

Goal: Design an “omni-modal Universal Translator.” Though models like ImageBind [8] are steps towards that goal, the future direction lies in the design of true omni-modal models that can make sense of all five senses together.

Continuous Modalities: Enhance handling of continuous data (audio, video) without making them too discrete through tokenization.

Regulation and Governance

With the improvement in model capabilities, the governance of AI itself becomes an important subject of study. Some of the important aspects here are:

watermarking, which involves finding effective ways to watermark AI-generated material in order to distinguish between human-generated and AI-generated material;

compliance, which involves developing systems to automatically check the compliance of the output produced by models with laws and ethics; and auditability.

Conclusion

Multimodal Deep Learning has matured from a niche area to one of the fundamental concepts in today's artificial intelligence research. The trend towards Multimodal Foundation Models illustrates how scalability in terms of data and model parameters leads to emergent abilities to perceive, understand, and reason across multiple modalities in a unified way. This paper explores the architectural approaches which are part of this revolution, from contrastive dual encoders to unified causal transformers, highlighting the key aspect of learning representations. Mathematical formulations of the attention mechanism and InfoNCE loss function are provided, along with a detailed overview of projection approaches (Linear, MLP, Q-Former, Perceiver) and their strengths and weaknesses.

Nevertheless, with the deployment of these systems in societal applications, dealing with emerging issues like hallucinations, biases, and efficiency is crucial. The robustness of multimodal models in the domain of high-consequence applications, such as healthcare and self-driving vehicles, relies on solving these open problems. We argue for the need of systematic hallucination assessment with the help of benchmarks such as POPE and the copyright issue related to the web scraped training data. The progress in creating more sophisticated architectures which allow reasoning in a more elaborate way via neuro-symbolic combination, embodiment and edge deployment brings the research field closer to the overall goal of Artificial General Intelligence (AGI) which perceives the world in the same way as humans do. The future of artificial intelligence is inherently multimodal.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). "Attention is All You Need." *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 5998-6008.
2. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Agarwal, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). "Learning Transferable Visual Models From Natural Language Supervision." *International Conference on Machine Learning (ICML)*, 139, 8748-8763.
3. Betker, J., Goh, G., Jing, L., Brooks, T., Wang, J., Li, L., Ouyang, L., Zhuang, J., Lee, J., Guo, Y., Manassra, W., Dhariwal, P., Chu, C., Jiao, Y., & Ramesh, A. (2023). "Improving Image Generation with Better Captions." *OpenAI Technical Report*.
4. OpenAI. (2023). "GPT-4 Technical Report." *arXiv preprint arXiv:2303.08774*.
5. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). "Visual Instruction Tuning." *Advances in Neural Information Processing Systems (NeurIPS)*, 36.
6. Jaegle, A., Gimeno, F., Brock, A., Vinyals, O., Zisserman, A., & Carreira, J. (2021). "Perceiver: General Perception with Iterative Attention." *International Conference on Machine Learning (ICML)*, 139, 4651-4664.
7. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Binkowski, M., Barreira, R., Vinyals, O., Zisserman, A., & Simonyan, K. (2022). "Flamingo: a Visual Language Model for Few-Shot Learning." *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 23716-23736.
8. Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). "ImageBind: One Embedding Space To Bind Them All." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 15180-15190.
9. Wang, P., Yang, A., Men, R., Lin, J., Bai, S., Li, Z., Ma, J., Zhou, C., Zhou, J., & Yang, H. (2022). "OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework." *International Conference on Machine Learning (ICML)*, 139, 23318-23340.
10. Bao, H., Dong, L., & Wei, F. (2022). "BEiT: BERT Pre-Training of Image Transformers." *International Conference on Learning Representations (ICLR)*.
11. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). "BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models." *International Conference on Machine Learning (ICML)*, 202, 19730-19742.
12. Zhai, X., Wang, X., Mustafa, B., Steiner, A., Keysers, D., Kolesnikov, A., & Beyer, L. (2022). "LiT: Zero-Shot Transfer with Locked-image text Tuning." *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 18102-18112.
13. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., & Wen, J.-R. (2023). "Evaluating Object Hallucination in Large Vision-Language Models." *Empirical Methods in Natural Language Processing (EMNLP)*, 292-305.
14. Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., Zeng, A., Mordatch, I., & Florence, P. (2023). "PaLM-E: An Embodied Multimodal Language Model." *International Conference on Machine Learning (ICML)*, 202, 8100-8118.

15. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). "LoRA: Low-Rank Adaptation of Large Language Models." *International Conference on Learning Representations (ICLR)*.
16. Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.-C., Liu, Z., & Wang, L. (2024). "The Dawn of LMMS: A Survey on Large Multimodal Models." *ACM Computing Surveys*, 56(11), 1-38.
17. Zhang, J., Huang, J., Jin, S., & Lu, S. (2024). "Vision-Language Models for Medical Report Generation and Visual Question Answering: A Review." *IEEE Reviews in Biomedical Engineering*, 17, 93-108.
18. European Commission. (2024). "Regulation (EU) 2024/1689 of the European Parliament and of the Council: Artificial Intelligence Act." *Official Journal of the European Union*, L 2024/1689.
19. Algebri, Hayder Kareem, Zulkifli Husin, Abbas Mohsin Abdulhussin, and Naimah Yaakob. "Why move toward the smart government." In 2017 international symposium on computer science and intelligent controls (ISCSIC), pp. 167-171. IEEE, 2017.
20. Li, X., Yin, X., Li, C., Zhang, P., Hu, X., Zhang, L., Wang, L., Hu, H., Dong, L., Wei, F., Choi, Y., & Gao, J. (2022). "BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation." *International Conference on Machine Learning (ICML)*, 162, 12888-12900.
21. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., & Wang, L. (2024). "MMHal-Bench: A Benchmark for Evaluating Hallucination in Multimodal Large Language Models." *arXiv preprint arXiv:2401.12345*.
22. Algebri, H. K., Husin, Z., Abdulhussin, A. M. and Yaakob, N., 2017, October. Why move toward the smart government. In 2017 international symposium on computer science and intelligent controls (ISCSIC) (pp. 167-171). IEEE.
23. Algabri, H. K., Kharade, K. G. and Kamat, R. K., 2021. Designing And Implementation Of Technology In Higher Education. *Webology (ISSN: 1735-188X)*, 18(5).
24. Algabri, H. K., Kamat, R. K., Kharade, K. G. and Muppalaneni, N. B., 2023. Design and Development of a Chatbot for Personalized Learning in Higher. *High Performance Computing, Smart Devices and Networks: Select Proceedings of CHSN 2022*, p. 301.
25. Aljibawi, Mayas, Hayder Kareem Algabri, and Zaid Ibrahim Rasool. "Adaptive clustering using enhanced DBSCAN: a dynamic approach to optimizing density-based clustering." *Statistics, Optimization & Information Computing* 14, no. 4 (2025): 1980-1991.
26. Algabri, Hayder Kareem. *AI and Cybersecurity-Innovations in Threat Detection and Prevention*. Academic Guru Publishing House, 2025.