

# Multimodal Alignment and Fusion: A Survey

Songtao Li\* · Hao Tang†

Received: date / Accepted: date

**Abstract** This survey provides a comprehensive overview of recent advances in multimodal alignment and fusion within the field of machine learning, driven by the increasing availability and diversity of data modalities such as text, images, audio, and video. Unlike previous surveys that often focus on specific modalities or limited fusion strategies, our work presents a structure-centric and method-driven framework that emphasizes generalizable techniques. We systematically categorize and analyze key approaches to alignment and fusion through both structural perspectives—data-level, feature-level, and output-level fusion—and methodological paradigms—including statistical, kernel-based, graphical, generative, contrastive, attention-based, and large language model (LLM)-based methods, drawing insights from an extensive review of over 260 relevant studies. Furthermore, this survey highlights critical challenges such as cross-modal misalignment, computational bottlenecks, data quality issues, and the modality gap, along with recent efforts to address them. Applications ranging from social media analysis and medical imaging to emotion recognition and embodied AI are explored to illustrate the real-world impact of robust multimodal systems. The insights provided aim to guide future research toward optimizing multimodal learning systems for improved scalability, robustness, and generalizability across diverse domains.

---

Songtao Li (\*work done during the visit at Peking University)  
Sydney Smart Technology College, Northeastern University, Qinhuangdao 066006, China. E-mail: 202219226@stu.neu.edu.cn

Hao Tang (†corresponding author)  
The State Key Laboratory of Multimedia Information Processing, School of Computer Science, Peking University, Beijing 100871, China. E-mail: haotang@pku.edu.cn

**Keywords** Multimodal Alignment, Multimodal Fusion, Multimodality, Machine Learning, Survey

## 1 Introduction

Rapid advancement in technology has led to an exponential increase in the generation of multimodal data, including images, text, audio, and video [15]. This abundance of data presents opportunities and challenges for researchers and practitioners in diverse fields, such as computer vision and natural language processing. Integrating information from multiple modalities can significantly enhance the performance of machine learning models, improving their ability to understand complex scenarios in the real world [20, 93, 94, 253, 12, 49].

At the core of multimodal learning lie two interdependent problems: *alignment* and *fusion*. Alignment aims to establish semantic correspondences across modalities so that their representations occupy a shared space, while fusion merges these aligned features into unified predictions or embeddings. The combination of modalities is generally pursued with two main objectives: (i) Different data modalities can complement each other, thus improving the precision and effectiveness of models for specific tasks [102, 247, 121]. (ii) Some modalities may have limited data availability or may be challenging to collect in large quantities; therefore, training in an LLM-based model can leverage knowledge transfer to achieve satisfactory performance in tasks with sparse data [104, 121].

For example, in social media analysis, combining textual content with related images or videos offers a more comprehensive understanding of user sentiment and behavior [15, 113]. Beyond social networks, multimodal methods have shown promising results in applications

such as automated caption generation for medical images, video summarization, and emotion recognition [55, 123, 33, 223, 259, 214]. Despite these advancements, two major technical challenges remain in effectively leveraging multimodal data: alignment—ensuring semantic consistency across modalities—and fusion—integrating complementary cues to enhance downstream performance.

To illustrate how modern architectures approach these challenges, Figure 1 provides an overview of three typical multimodal model structures: (a) *Two-Tower*, which processes modalities separately and combines embeddings via simple operations; (b) *Two-Leg*, which introduces a dedicated fusion network on top of separate encoders; and (c) *One-Tower*, which jointly encodes all modalities in a unified network.

While prior surveys often categorize fusion either by the stage of integration (e.g., data-level, feature-level, output-level fusion) [15, 18], our work complements rather than replaces existing taxonomies. We retain the traditional classification while introducing a new organization based on the core methods they employ—ranging from statistical and kernel-based techniques to generative, contrastive, attention-based, and LLM-driven frameworks. This approach foregrounds methodological innovations and highlights how each paradigm contributes to deeper and more flexible multimodal integration. Additionally, unlike prior surveys that focus on specific modalities or models—such as vision–language models—and often treats alignment and fusion methods as a part of their survey, our work adopts a structure-centric and method-driven perspective, emphasizing general-purpose alignment and fusion methods [240, 258, 103]. Note that while vision–text research dominates the current literature due to data availability and historical development patterns, the structural and methodological frameworks we present are designed to be transferable across modality combinations.

The organization of this survey is as follows. Section 2 presents an overview of the foundational concepts in multimodal learning, including recent advances in LLMs and vision models, laying the groundwork for discussions on fusion and alignment. Section 3 focuses on why to conduct a survey on alignment and fusion. Section 4 examines multimodal alignment and fusion approaches, presenting a dual perspective that combines structural categorizations—such as data-level, feature-level, and output-level fusion—with classifications based on the core methods and features involved in modern models. The section heavily focuses later classification from traditional strategies to recent advances, including statistical, kernel-based, graphical, generative, contrastive, attention-based and llm-based fusion frameworks, highlighting how these techniques enable deeper inter-modal

integration and more flexible modeling of complex relationships. Section 5 addresses key challenges in multimodal fusion and alignment, including feature alignment, computational efficiency, data quality, and scalability. Finally, section 7 outlines the potential directions for future research and discusses practical implications, with the aim of guiding further innovation in the field.

## 2 PRELIMINARIES

This section provides a brief overview of key topics and concepts to enhance the understanding of our work.

### 2.1 MLLM

Recently, both natural language processing (NLP) and computer vision (CV) have experienced rapid development, especially since the introduction of attention mechanisms and the Transformer [194, 69, 249, 107, 248, 135, 83, 173, 45]. Building on this framework, numerous large language models (LLMs) have emerged, such as OpenAI’s GPT series [149, 24, 138] and Meta’s Llama series [50]. Similarly, in the vision domain, large vision models (LVMs) have been proposed, including Segment Anything [80], DINO [238], and DINOv2 [139].

However, these LLMs struggle to understand visual information and handle other modalities, such as audio or sensor inputs, while LVMs have limitations in reasoning [225]. Given their complementary strengths, LLMs and LVMs are increasingly being combined, leading to the emergence of a new field called multimodal large language models (MLLMs). To extend the strong performance of LLMs in text processing to tasks involving other modalities, significant research efforts have been dedicated to developing large-scale multimodal models.

To extend the strong performance of LLMs in text processing to tasks involving other modalities, significant research efforts have focused on the development of large-scale multimodal models [53]. Kosmos-2 [143] introduces grounding capabilities by linking textual descriptions with visual contexts, allowing more accurate object detection and phrase recognition. PaLM-E [48] further integrates these capabilities into real-world applications, using sensor data for embodied tasks in robotics, such as sequential planning and visual question answering. Additionally, models like ContextDET [234] excel in contextual object detection, overcoming previous limitations in visual–language association by directly linking visual elements to language inputs.

Several models have adopted a hierarchical approach to managing the complexity of multimodal data. For example, SEED-Bench-2 benchmarks hierarchical MLLM

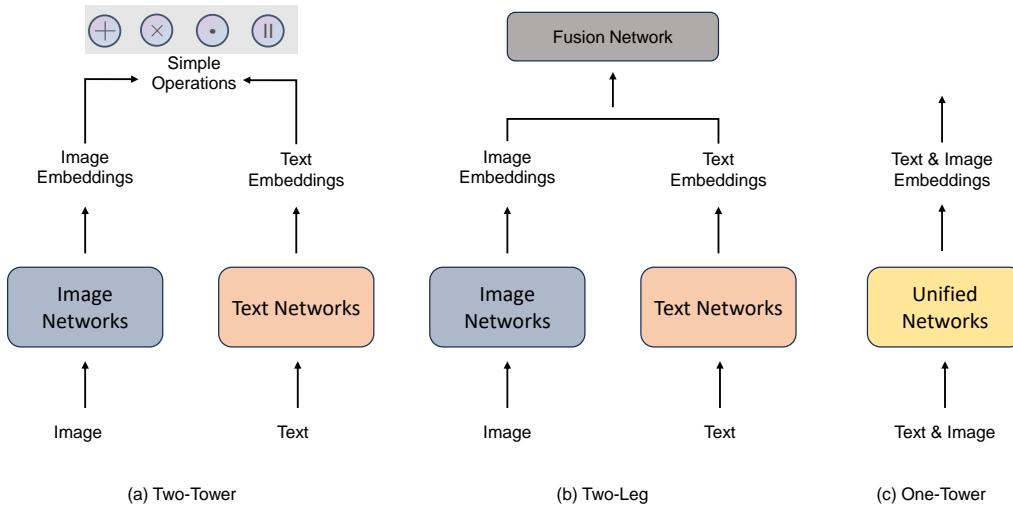


Fig. 1: Overview of multimodal model architectures: (a) Two-Tower [148, 72, 39, 193, 218, 170, 111, 28, 54, 211, 189, 231]: processes images and text separately, combining embeddings through simple operations (add, multiple, dot product and concatenate); (b) Two-Leg [5, 13, 42, 61, 71, 100, 92, 124, 126, 128, 153, 169, 190, 210, 79]: combines separate image and text embeddings using a Fusion Network; (c) One-Tower [17, 96, 95, 40, 30, 259, 209, 201, 14, 4]: utilizes a unified network to jointly embed image and text inputs.

capabilities, providing a structured framework to evaluate and improve model performance in both perception and cognition tasks [87]. Furthermore, X-LLM enhances multimodal alignment by treating each modality as a “foreign language”, allowing a more effective alignment of audio, visual, and textual inputs with large language models [29].

As MLLMs continue to evolve, foundational frameworks such as UnifiedVisionGPT enable the integration of multiple vision models into a unified platform, accelerating advancements in multimodal AI [76]. These frameworks demonstrate the potential of MLLMs to not only leverage vast multimodal datasets but also adapt to a wide range of tasks, representing a significant step toward achieving artificial general intelligence.

## 2.2 Multimodal Dataset

Different modalities offer unique characteristics. For example, images provide visual information, but are susceptible to variations in lighting and viewpoint [255]. The text data are linguistically diverse and may contain ambiguities [225]. Audio data conveys emotional content and other non-verbal cues [15].

Multimodal datasets are foundational for training vision-language models (VLMs) by providing large-scale paired image-text data that enable model learning across various tasks, such as image captioning, text-to-image retrieval, and zero-shot classification. Key datasets include LAION-5B, WIT, and newer specialized datasets

like RS5M, which target specific domains or challenges within multimodal learning. Table 1 summarizes the commonly used datasets and their characteristics.

For example, the LAION-5B dataset contains more than 5 billion CLIP-filtered image-text pairs, enabling researchers to fine-tune models such as CLIP (Contrastive Language-Image Pretraining) and GLIDE, supporting open-domain generation and robust zero-shot classification tasks [157]. The WIT (Wikipedia-based Image Text) dataset, with more than 37 million image-text pairs in 108 languages, is designed to support multilingual and diverse retrieval tasks, focusing on cross-lingual understanding [166]. The RS5M dataset, which consists of 5 million remote sensing image-text pairs, is optimized for domain-specific learning tasks such as semantic localization and vision-language retrieval in geospatial data [254]. Furthermore, fine-grained datasets like ViLLA are tailored to capture complex region-attribute relationships, which are critical for tasks such as object detection in medical or synthetic imagery [192].

### 2.3 Characteristics and Targets

Each modality in multimodal learning presents unique challenges. For example, image data often face issues such as lighting variations, occlusions, and perspective distortions, which can affect a model's ability to recognize objects and scenes under varying conditions [256]. Text data bring complexities due to the variability of natural language, including ambiguity, slang, and pol-

Table 1: Overview of different datasets' characteristics.

Dataset	Size	Modalities	Features
SBU Captions [140]	1M	Image, Text	More unique words than CC-3M but fewer captions.
MS-COCO [108]	1.64M	Image, Text	Created by having crowd workers provide captions for images.
YFCC-100M [183]	100M	Image, Text	Contains 100 million image-text pairs, unclear average match degree between text and image.
Flickr30k [144]	30k	Image, Text	Created by having crowd workers provide captions for approximately 30,000 images.
Visual Genome [84]	5.4M	Image, Text	Includes structured image concepts such as region descriptions, object instances, relationships, etc.
RedCaps [43]	12.01M	Image, Text	Distributed across 350 subreddits with a long-tail distribution. Contains the distribution of visual concepts encountered by humans in everyday life without predefined object class ontologies. Higher linguistic diversity compared to other datasets like CC-3M and SBU.
CC-12M [27]	12.4M	Image, Text	Lower linguistic diversity compared to RedCaps.
WIT [166]	37.6M	Image, Text	Subset of multilingual Wikipedia image-text dataset.
TaiSu [114]	166M	Image, Text	TaiSu is a large-scale, high-quality Chinese cross-modal dataset containing 166 million images and 219 million Chinese captions, designed for vision-language pre-training.
COYO-700M [25]	700M	Image, Text	Collection of 700 million informative image-alt text pairs from HTML documents.
LAION-5B [157]	5.85B	Image, Text	LAION-5B is a publicly available, large-scale dataset containing over 5.8 billion image-text pairs filtered by CLIP, designed for training the next generation of image-text models.
DATACOMP-1B [56]	1.4B	Image, Text	Collected from Common Crawl using simple filtering. Models trained on this dataset achieve higher accuracy using fewer MACs compared to previous results.
RS5M [254]	5M	Image, Text	The RS5M dataset is a large-scale remote sensing image-text paired dataset, containing 5 million remote sensing images alongside corresponding English descriptions.
DEAP [81]	2122 samples	EEG, ECG, GSR	Contains 40 one-minute music video excerpts with continuous affect ratings from 32 participants for emotion recognition.
PAMAP2 [151]	3850505 samples	IMU, Heart Rate	Multi-sensor time-series data for 18 activities with high-resolution physiological and motion data collected from 9 subjects.
MHEALTH [16]	120 samples	ECG, EMG, Motion	Biometric sensor streams capturing high-intensity functional movements for medical activity monitoring from 10 subjects.
CH-SIMS [230]	2,281 videos	Text, Audio, Video	Tri-modal dataset featuring multi-modal annotations across text, audio, and video modalities for sentiment analysis, with higher linguistic diversity compared to audio-only approaches.
MuSe-CaR [168]	40 hours video	Text, Audio, Video	Multimodal dataset for sentiment analysis in car reviews containing 40 hours of user-generated video material with more than 350 reviews, combining spoken language, vocal qualities, and visual cues for comprehensive sentiment understanding.

ysemy, which complicate accurate interpretation and alignment with other modalities [225]. Similarly, audio data is susceptible to background noise, reverberation, and environmental interference, which can distort the intended signal and reduce model accuracy [150].

To address these challenges, specific loss functions are employed in multimodal learning to optimize both representations and alignments. These losses define how features from different modalities should be related or

transformed to achieve meaningful alignment or fusion. Notable examples include:

- **Contrastive Loss (and Variants)**, commonly used in tasks such as image-text matching, aims to pull together semantically similar pairs while pushing apart dissimilar ones in the embedding space. This objective supports better alignment and discrimination across modalities. The basic contrastive loss is defined as:

$$\mathcal{L}_{\text{contrastive}} = (1 - y) \cdot d^2 + y \cdot \max(0, m - d)^2, \quad (1)$$

where  $d$  is the distance between embeddings of a pair,  $y = 1$  indicates a negative pair, and  $m$  is a margin hyperparameter.

Supervised contrastive loss extends this idea by leveraging class labels to construct positive and negative pairs within a batch. It encourages samples from the same class to cluster tightly while separating those from different classes. Its formulation is:

$$\mathcal{L}_{\text{supcon}} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\sum_{j \in \mathcal{P}(i)} \exp(\text{sim}(z_i, z_j)/\tau)}{\sum_{k=1}^K \exp(\text{sim}(z_i, z_k)/\tau)}, \quad (2)$$

where  $\mathcal{P}(i)$  denotes indices of positive samples for anchor  $i$ ,  $\text{sim}(\cdot, \cdot)$  is cosine similarity,  $\tau$  is a temperature parameter, and  $K$  includes all samples in the batch. CLIP [148] shares a similar objective, using contrastive-style learning with image-text pairs as positives and cross-modal negatives.

- **Sigmoid Loss** [235] proposes a simplified and more efficient alternative to the softmax-based contrastive loss used in models like CLIP. Instead of normalizing similarities across the entire batch via softmax, which couples the loss computation to the global batch structure, the sigmoid loss treats each image-text pair independently as a binary classification problem: matching (positive) or non-matching (negative). It applies the sigmoid function to the scaled similarity score (with a learnable temperature  $t$  and bias  $b$ ) and computes a binary cross-entropy loss. Formally, for a batch of size  $n$ , the loss is:

$$\mathcal{L}_{\text{sigmoid}} = -\frac{1}{n} \sum_{i=1}^n \sum_{j=1}^n \log \sigma(z_{ij} \cdot s_{ij}), \quad (3)$$

where  $s_{ij} = t \cdot \text{sim}(z_i^{\text{img}}, z_j^{\text{text}}) + b$ , and

$$z_{ij} = \begin{cases} 1, & \text{if } i = j \text{ (positive pair)} \\ -1, & \text{otherwise (negative pair)} \end{cases}$$

This formulation significantly improves training efficiency and scalability. First, it eliminates the need for costly all-to-all batch synchronization for softmax normalization, enabling a highly memory-efficient "chunked" implementation. Second, it decouples the batch size from the loss definition, leading to superior performance at small batch sizes and allowing stable training at extremely large batch sizes (e.g., one million). Furthermore, the introduction of a learnable bias term  $b$  stabilizes training by counteracting the initial extreme imbalance between positive and negative pairs, ensuring the model starts training from a more reasonable prior. These improvements make language-image pre-training more accessible and efficient, especially with limited computational resources.

- **Cross-Entropy Loss**, a widely used classification loss, calculates the divergence between predicted

and true probability distributions, enabling label-driven learning across modalities. It is fundamental in supervised classification tasks, and variants such as set cross-entropy offer greater flexibility for multimodal tasks by handling multiple target answers [257, 8].

Cross-entropy loss is particularly effective when the goal is to map multimodal inputs into a shared label space. Given predicted logits  $p$  and ground truth  $y$ , it is defined as:

$$\mathcal{L}_{\text{CE}} = -\sum_{c=1}^C y_c \log(p_c), \quad (4)$$

where  $C$  is the number of classes. In multimodal settings, cross-entropy can be applied after fusing modalities or used independently per modality to encourage consistency.

- **Reconstruction Loss**, used in autoencoders and multimodal fusion tasks, aims to reconstruct input data or mask noise, making models more resilient to modality-specific distortions. This type of loss is essential for multimodal tasks requiring robust feature alignment and noise resilience, such as visual-textual and audio-visual fusion [141].

Reconstruction loss serves the purpose of preserving information during modality transformation or fusion. It ensures that no critical semantic content is lost during projection into a shared space. A common form is mean squared error (MSE):

$$\mathcal{L}_{\text{rec}} = \frac{1}{N} \sum_{i=1}^N \|x_i - \hat{x}_i\|^2, \quad (5)$$

where  $x_i$  is the original input and  $\hat{x}_i$  is the reconstructed version. This loss is especially useful in unsupervised or semi-supervised multimodal architectures.

### 3 Why Multimodal Alignment and Fusion

Alignment and fusion are two fundamental concepts in multimodal learning that, while distinct, are deeply interconnected and often mutually reinforce [3, 15]. Alignment involves ensuring that the different modalities are properly matched and synchronized, making the information they convey coherent and suitable for integration. Fusion, on the other hand, refers to the combination of information from different modalities to create a unified representation that captures the essence of the data in a comprehensive way [15, 184, 158]. Furthermore, many recent methods find it challenging to fusion without an alignment process [97].

### 3.1 Enhancing Comprehensiveness and Robustness

Alignment ensures that data from different sources are synchronized in terms of time, space, or context, enabling a meaningful combination. Without proper alignment, the fusion process can result in misinterpretations or loss of crucial information [18].

Once alignment is achieved, fusion utilizes the aligned data to produce a more robust and comprehensive representation [97, 180]. By integrating multiple perspectives, fusion mitigates the weaknesses of individual modalities, leading to improved accuracy and reliability.

### 3.2 Addressing Data Sparsity and Imbalance

In many real-world applications, data from certain modalities may be scarce or difficult to obtain. Alignment helps to synchronize the available data, even if limited, to ensure that it can be used effectively [165, 197].

The fusion then enables the transfer of knowledge between modalities, allowing the model to leverage the strengths of one modality to compensate for the weaknesses of another. This is particularly beneficial in scenarios where one modality has abundant data, while another is limited.

### 3.3 Improving Model Generalization and Adaptability

Alignment ensures that the relationships between different modalities are well understood and accurately modeled, which is crucial for the model's ability to generalize across various contexts and applications [15, 18].

Fusion improves the model's adaptability by creating a unified representation that captures the nuances of the data more effectively. This unified representation can be more easily adapted to new tasks or environments, enhancing the overall flexibility of the model [15, 18].

### 3.4 Enabling Advanced Applications

Alignment and fusion together enable advanced applications such as cross-modal retrieval, where information from one modality (e.g., text) is used to search for relevant information in another modality (e.g., images) [198]. These processes are also crucial for tasks like emotion recognition [137], where combining visual and auditory cues provides a more accurate understanding of human emotions compared to using either modality alone.

## 4 Multimodal Alignment and Fusion

Multimodal data involves the integration of various types of information, such as images, text, and audio, which can be processed by machine learning models to improve performance across numerous tasks [104, 15, 18, 228]. In this context, multimodal alignment and fusion are essential techniques that aim to effectively combine information from different modalities. While early approaches often processed modalities separately with only basic integration, recent methods have evolved to better capture semantic correspondences and interactions among modalities.

At a high level, these processes involve establishing meaningful relationships between heterogeneous data sources, either explicitly or implicitly, to construct representations that reflect shared semantics [133, 146]. Explicit strategies may rely on similarity matrices to directly measure correspondences, while implicit approaches often operate in latent spaces, learning joint representations through intermediate steps such as translation or prediction [15, 120]. These mechanisms are not strictly confined to one task or another but instead contribute to the broader objective of integrating diverse signals into a coherent model.

Fusion methods, in particular, play a critical role in how modalities interact within the architecture. Traditional classifications distinguish between early fusion, which combines data at the feature level [163], late fusion, which merges outputs at the decision level [130], and hybrid approaches that integrate both strategies [255]. However, as technology evolves, the boundaries between these categories have become increasingly blurred. Modern architectures—such as CLIP [148]—utilize dual encoders with relatively shallow interaction mechanisms, which are effective for retrieval-based tasks [155, 182], but fall short in more complex scenarios requiring nuanced understanding [236, 179].

For tasks like visual question answering and reasoning, deeper integration is essential to capture the interdependencies between modalities, going beyond simple concatenation or independent encoding [167, 247]. This has led to the development of advanced fusion techniques that operate simultaneously at multiple levels of abstraction, challenging traditional stage-based categorizations. As a result, there is a growing need for a classification framework based on the core characteristics of current fusion technologies, rather than rigid temporal distinctions. Notably, attention-based mechanisms have emerged as powerful tools in this space, warranting separate treatment due to their unique contributions and rapid evolution in recent years [1]. In the following parts of this section, this paper presents

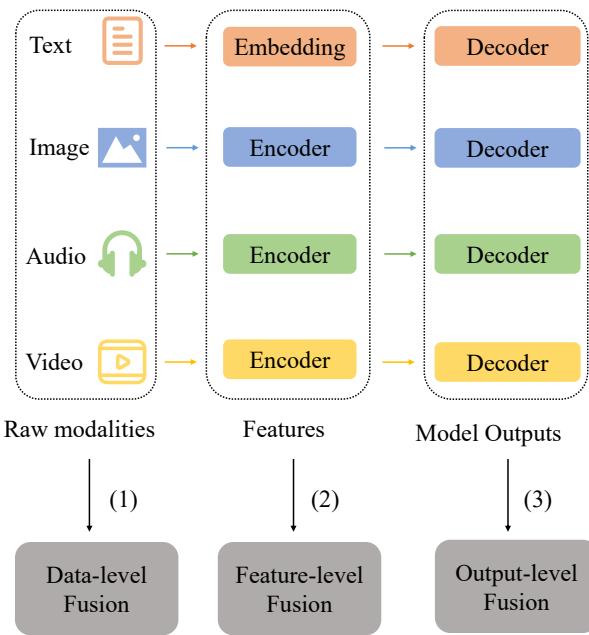


Fig. 2: Three types of fusion with structural perspective: (1) Data-level Fusion: directly combines raw data from multiple modalities; (2) Feature-level Fusion: integrates encoded features from each modality; (3) Output-level Fusion: fuses outputs from individual modality decoders to produce a final result.

two classification approaches: one follows the traditional structural categorization, while the other is based on the core methods and features involved in the models.

#### 4.1 Structural Perspectives: A Three-level Taxonomy

From structural perspectives, a multimodal model typically involves an encoder that captures essential features from the input data and compresses them into a compact form, while the decoder reconstructs the output from this compressed representation [13].

In this architecture, the system is primarily composed of two major components: the encoder and the decoder. The encoder typically functions as a high-level feature extractor, transforming the input data into a latent space of significant features [13, 190]. In other words, the encoding process preserves important semantic information while reducing redundancy. Once the encoding step is complete, the decoder generates a corresponding “reconstructed” output based on the latent representation [13, 92]. In tasks like semantic segmentation, the decoder’s output is usually a semantic label map that matches the size of the input.

In this section, we review models based on architecture. The encoder-decoder framework is an intuitive approach in which an encoder first extracts features, and then these more expressive representations are used to learn the correlations, enabling interactions between different modalities and integrating features from diverse sources. Increasingly, researchers are exploring hybrid ways to integrate features from different modalities to better reveal the relationships among them. To provide a summary, detailed information on representative models is presented in Table 2.

This encoder-decoder architecture typically takes three forms: (1) Data-level fusion, where raw data from different modalities is concatenated and fed into a shared encoder; (2) Feature-level fusion, where features are extracted separately from each modality, possibly including intermediate layers, and then combined before being input into the decoder; and (3) Output-level fusion, where outputs of individual modality-specific models are concatenated after processing. Figure 2 illustrates these three types of encoder-decoder fusion structures. Feature-level fusion is often the most effective, as it considers the relationships between different modalities, enabling deeper integration rather than a superficial combination.

##### 4.1.1 Data-level Methods

In this method, data from each modality or processed data from each modality’s unique preprocessing steps are combined at the input level [42]. After this integration, the unified input from all modalities is passed through a single encoder to extract higher-level features. Essentially, data from different modalities is merged at the input stage, and a single encoder is used to extract comprehensive features from the multimodal information.

Recent research has focused on data-level fusion to improve object detection and perception in autonomous vehicles. Studies have explored fusing camera and LiDAR data at the early stages of neural network architectures, demonstrating enhanced 3D object detection accuracy, particularly for cyclists in sparse point clouds [153]. A YOLO-based framework that jointly processes raw camera and LiDAR data showed a 5% improvement in vehicle detection compared to traditional decision-level fusion [42]. Additionally, an open hardware and software platform for low-level sensor fusion, specifically leveraging raw radar data, has been developed to facilitate research in this area [169]. These studies highlight the potential of raw-data-level fusion to exploit inter-sensor synergies and improve overall system performance.

Table 2: Summary of models from structural perspective.

<b>Model</b>	<b>Year</b>	<b>Category</b>	<b>Modality</b>	<b>Explanation</b>
TFN [233]	2017	Output-level	Text, Image	First to use tensor fusion for high-order modality interactions at decision level. Enables end-to-end learning without intermediate fusion, capturing complex multimodal dynamics.
MFAS [188]	2018	Hybrid	Text, Image	Combines feature- and output-level fusion via factorized attention and modality gating. Dynamically weights modalities, outperforming static fusion strategies.
ViLBERT [116]	2019	Feature-level	Text, Image	Employs a two-stream architecture with co-attentional transformer layers for separate vision and language processing. Innovates by pretraining on masked multimodal modeling and alignment prediction, enabling deep cross-modal interaction while preserving modality-specific processing depths.
UNITER [35]	2020	Feature-level	Text, Image	Early unified Transformer for vision-language pretraining. Pioneered MLM, MRM, and ITM tasks, setting the standard for subsequent VLP models.
Perceiver [70]	2021	Data-level	Images, Point Clouds, Audio, Video	Uses asymmetric cross-attention to fuse raw inputs into a latent bottleneck. Innovates by enabling a single Transformer to handle arbitrary modalities, avoiding modality-specific architectures.
VX2TEXT [109]	2021	Data-level	Video, Audio, Text	Employs learnable tokenizers to convert all modalities into embeddings for unified Transformer encoding. First to unify diverse modalities end-to-end for generation, eliminating preprocessing disparities.
CLIP [148]	2021	Feature-level	Text, Image	Aligns image and text via contrastive learning on large-scale data. Innovates with scalable pretraining and strong zero-shot transfer, surpassing fine-tuning-dependent models.
BLIP [96]	2022	Feature-level	Text, Image	Enhances CLIP with generative captioning for data refinement. Jointly optimizes understanding and generation, improving performance on both retrieval and captioning tasks.
FLAVA [162]	2022	Feature-level	Text, Image	Uses three separate encoders with multimodal fusion and joint unimodal/multimodal pretraining. Unique in supporting comprehensive unimodal and multimodal tasks within one model.
ImageBind [58]	2023	Feature-level	Text, Image, Audio, Depth, Thermal, IMU	Aligns six modalities in joint embedding space using image as binding medium via contrastive learning. Achieves "emergent alignment" for unseen modality pairs without direct training, enabling zero-shot cross-modal retrieval and classification.
ProVLA [68]	2023	Output-level	Text, Image	Uses two-stage Transformer with hard negative mining for progressive fusion. Improves retrieval accuracy through iterative cross-modal refinement.
TextBind [91]	2024	Hybrid	Text, Image	Combines feature-level fusion (Q-Former mapping) and output-level fusion (LM to Stable Diffusion). Supports interleaved multimodal inputs/outputs for comprehensive instruction following with understanding and generation capabilities.

#### 4.1.2 Feature-level Methods

The concept behind this fusion technique is to combine data from multiple levels of abstraction, allowing features extracted at different layers of hierarchical deep networks to be utilized, ultimately enhancing model performance. Many applications have implemented this fusion strategy [171, 156, 100, 124, 126].

Feature-level fusion has emerged as a powerful approach in various computer vision tasks. It involves combining features at different levels of abstraction to improve performance. For instance, in gender classification, a two-level hierarchy that fused local patches proved effective [156]. For salient object detection, a network that hierarchically fused features from different

VGG levels preserved both semantic and edge information [100]. In multimodal affective computing, a “divide, conquer, and combine” strategy explored both local and global interactions, achieving state-of-the-art performance [124]. For adaptive visual tracking, a hierarchical model fusion framework was developed to update object models hierarchically, guiding the search in parameter space and reducing computational complexity [126]. These approaches demonstrate the versatility of hierarchical feature fusion across various domains, showcasing its ability to capture both fine-grained and high-level information for improved performance in complex visual tasks.

#### 4.1.3 Output-level Methods

Output-level fusion is a technique that improves accuracy in various applications by integrating the outputs from multiple models. For example, in landmine detection using ground penetrating radar (GPR), Missaoui et al. [128] demonstrated that fusing Edge Histogram Descriptors and Gabor Wavelets through a multi-stream Continuous hidden markov model (HMM) outperformed individual features and equal-weight combinations.

In multimodal object detection, Guo and Zhang [61] applied fusion methods such as averaging, weighting, cascading, and stacking to combine the results from models processing images, speech, and video, thereby improving performance in complex environments. For facial action unit (AU) detection, Jaiswal et al. [71] found that output-level fusion using artificial neural networks (ANNs) was more effective than simple feature-level approaches.

Additionally, for physical systems involving multi-fidelity computer models, Allaire and Willcox [5] developed a fusion methodology that uses model inadequacy information and synthetic data, resulting in better estimates compared to individual models. In quality control and predictive maintenance, a novel output-level fusion approach outperformed traditional methods, reducing prediction variance and increasing accuracy [210]. These studies demonstrate the effectiveness of output-level fusion across various domains.

## 4.2 Methodological Approaches: Classification Based on Core Techniques

#### 4.2.1 Statistical methods

In the early stages, “alignment” often referred to the process of mapping vectors from different modalities into a shared vector space, while “fusion” typically involved a simple summation of the aligned modality vectors,

followed by feeding the summed result into a neural network to obtain the final fused representation. Such alignment methods frequently relied on statistical techniques such as dynamic time warping (DTW) [191, 85] and canonical correlation analysis (CCA) [67].

DTW measures the similarity between two sequences by finding an optimal match through time warping, which involves inserting frames to align the sequences [191]. However, the original DTW formulation requires a predefined similarity metric, so it has been extended with CCA, introduced by Harold Hotelling in 1936 [67], to project two different spaces into a common space through linear transformations. The goal of CCA is to maximize the correlation between the two spaces by optimizing the projection. CCA facilitates both alignment (through DTW) and joint learning of the mapping between modalities in an unsupervised manner, as seen in multimodal applications such as video-text and video-audio alignment. Figure 3 visualizes the CCA method. Specifically, the objective function of CCA can be expressed as:

$$\max \rho = \text{corr}(u^T X, v^T Y), \quad (6)$$

where:

- $X$  and  $Y$  are the data matrices from two different spaces;
- $u$  and  $v$  are the linear transformation vectors (or canonical vectors) that project  $X$  and  $Y$  into the common space;
- $\rho$  is the correlation coefficient between the projections  $u^T X$  and  $v^T Y$ ;
- The goal is to find  $u$  and  $v$  that maximize the correlation  $\rho$  between the projected data.

However, CCA can only capture linear relationships between two modalities, limiting its applicability in complex scenarios involving non-linear relationships. To address this limitation, kernel canonical correlation analysis (KCCA) was introduced to handle non-linear dependencies by mapping the original data into a higher-dimensional feature space using kernel methods [11, 65]. Extensions such as multi-label KCCA and deep canonical correlation analysis (DCCA) further improved upon the original CCA method [2, 127, 11, 65, 6].

Additionally, Verma and Jawahar demonstrated that multimodal retrieval could be achieved using support vector machines (SVMs) [195]. Furthermore, methods such as linear mapping between feature modalities for image alignment have been developed to address multimodal alignment through complex spatial transformations [74].

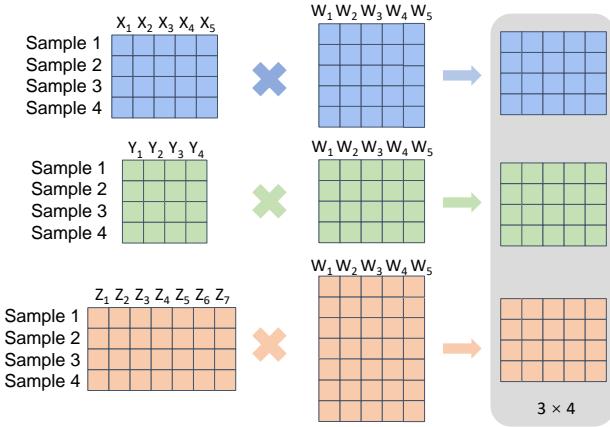


Fig. 3: Canonical Correlation Analysis (CCA), a classic alignment method, aligns different sample matrices with varying feature dimensions using a shared weight matrix to produce a unified representation.  $X$ ,  $Y$  and  $Z$  are the data matrices from three different spaces.

#### 4.2.2 Kernel-based Methods

Kernel-based techniques have gained prominence across various domains for their ability to handle nonlinear relationships and effectively integrate heterogeneous data sources. These methods leverage the kernel trick to map data into higher-dimensional spaces, enabling improved feature representation and analysis [7, 131]. By selecting appropriate kernel functions, such as polynomial kernels or radial basis function kernels, these methods can achieve computational efficiency while maintaining model complexity and accuracy.

Kernel cross-modal factor analysis has been introduced as a novel approach for multimodal fusion, particularly for bimodal emotion recognition [208]. This technique identifies optimal transformations to represent coupled patterns between different feature subsets. In drug discovery, integrating multiple data sources through kernel functions within SVMs enhances drug-protein interaction predictions [207]. For audio-visual voice activity detection, kernel-based fusion with optimized bandwidth selection outperforms traditional approaches in noisy environments [47]. In multimedia semantic indexing, kernel-based normalized early fusion and contextual late fusion schemes demonstrate improvements over standard fusion methods [10]. For drug repositioning, kernel-based data fusion effectively integrates heterogeneous information sources, outperforming rank-based fusion and providing a unique solution for identifying new therapeutic applications of existing drugs [7].

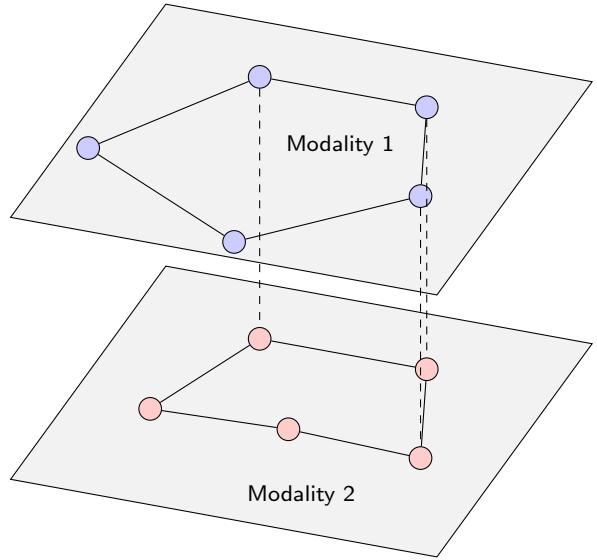


Fig. 4: In graph-based alignment, different data modalities can form graphs with distinct meanings, where the interpretation of edges and nodes may vary. For example, in [82], the interpretation of vertices and edges depends on the type of biological networks being compared.

Through the use of the kernel trick, these methods achieve computational efficiency and improve prediction accuracy by better representing patterns. However, challenges exist, including difficulty in selecting the right kernel and tuning parameters, potential scalability issues with large datasets, reduced interpretability due to higher-dimensional projections, and the risk of overfitting if not properly regularized.

#### 4.2.3 Graphical Model-Based Methods

The integration of graph structures allows for better modeling of complex relationships between different modalities, enabling more accurate and efficient processing of multimodal data. Such methods are commonly applied in aligning images with text or images with signals. For instance, certain models enable few-shot in-context imitation learning by aligning graph representations of objects, allowing robots to perform tasks on new objects without prior training [196]. The GraphAlignment algorithm, based on an explicit evolutionary model, demonstrates robust performance in identifying homologous vertices and resolving paralogs, outperforming alternatives in specific scenarios [82]. Figure 4 illustrates how graphs are used in alignment.

A significant challenge in these tasks is aligning implicit information across modalities, where multimodal signals do not always correspond directly to one another. Graph-based models have proven effective in addressing this challenge by representing complex relationships be-

tween modalities as graphs, where nodes represent data elements (e.g., words, objects, or frames) and edges represent relationships (e.g., semantic, spatial, or temporal) between them.

Recent studies have explored various aspects of multimodal alignment using graph structures. For instance, Tang et al. [178] introduced a graph-based multimodal sequential embedding approach to improve sign language translation. By embedding multimodal data into a unified graph structure, their model better captures complex relationships.

Another application is in sentiment analysis, where implicit multimodal alignment plays a crucial role. Yang et al. [224] proposed a multimodal graph-based alignment model (MGAM) that jointly models explicit aspects (e.g., objects, sentiment) and implicit multimodal interactions (e.g., image-text relations).

In the domain of embodied AI, Song et al. [164] explored how scene-driven knowledge graphs can be constructed to model implicit relationships in complex multimodal tasks. Their work integrates both textual and visual information into a knowledge graph, where multimodal semantics are aligned through graph-based reasoning. Aligning implicit cues, such as spatial and temporal relationships between objects in a scene, is crucial for improving decision-making and interaction in embodied AI systems.

For named entity recognition (NER), Zhang et al. [252] proposed a token-wise graph-based approach that incorporates implicit visual information from images associated with text. This method leverages spatial relations in the visual domain to improve the identification of named entities, which are often ambiguous when using isolated textual data.

In tasks such as image captioning and visual question answering (VQA), scene graphs also play a crucial role. Xiong et al. [215] introduced a scene graph-based model for semantic alignment across modalities. By representing objects and their relationships as nodes and edges in a graph, the model improves the alignment of visual and textual modalities.

Besides, graphical models provide a powerful approach for representing and fusing multimodal data, effectively capturing complex relationships between different modalities [261]. These models are particularly useful for handling incomplete multimodal data. For example, the heterogeneous graph-based multimodal Fusion (HGMF) method [31] constructs a heterogeneous hypernode graph to model and fuse incomplete multimodal data. HGMF leverages hypernode graphs to accommodate diverse data combinations without requiring data imputation, enabling robust representations

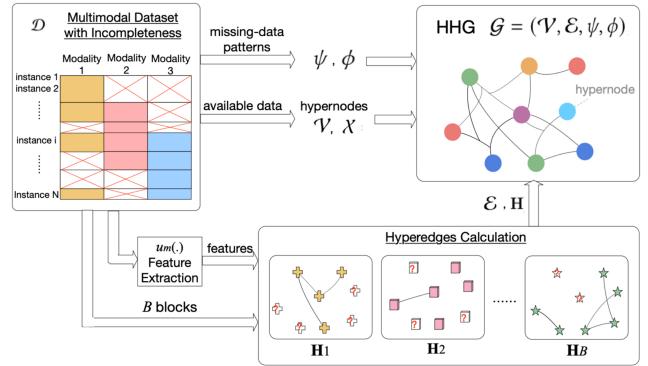


Fig. 5: Illustration from [31], demonstrating how graph models can effectively fuse modalities, even when some data is missing.

across various modalities [31]. Figure 5 illustrates the construction of hypernodes in [31].

Graphical fusion methods are increasingly used to combine data from multiple modalities for various applications, such as Alzheimer's disease (AD) diagnosis and target tracking [19, 160]. For example, in AD diagnosis, heterogeneous graph-based models integrate neuroimaging modalities like MRI and PET, capturing complex brain network structures to improve prediction accuracy [161]. In recommendation systems, heterogeneous graphs enable the effective integration of text, image, and social media data, enhancing the quality of recommendations by capturing multimodal relationships [216]. However, traditional linear combination approaches for multimodal fusion face limitations in capturing complementary information and are often sensitive to modality weights [186].

To address these issues, researchers have developed nonlinear graph fusion techniques that efficiently exploit multimodal complementarity [186, 187]. These techniques, such as early fusion operators in heterogeneous graphs, outperform linear approaches by capturing inter-modal interactions and have demonstrated improvements in one-class learning and multimodal classification tasks [63]. For instance, nonlinear fusion methods have shown enhanced classification accuracy for AD and its prodromal stage, mild cognitive impairment (MCI) [187].

Recent advancements include adversarial representation learning and graph fusion networks, which aim to learn modality-invariant embedding spaces and explore multi-stage interactions between modalities [125]. These approaches have demonstrated state-of-the-art performance in multimodal fusion tasks and provide improved visualization of fusion results [19, 125].

In summary, graph-based methods provide a powerful framework for representing diverse data types and

capturing complex, high-order interactions across modalities, making them highly effective for applications in medical diagnosis, social recommendation, and sentiment analysis. With ongoing advancements, graph-based methods hold great promise for handling incomplete, heterogeneous data and driving innovation in AI-powered multimodal applications. However, this flexibility also presents significant challenges. The sparsity and dynamic nature of graph structures complicate optimization. Unlike matrices or vectors, graphs have irregular unstructured connections, leading to high computational complexity and memory constraints. These issues persist even with advanced hardware platforms. Additionally, graph neural networks (GNNs) are particularly sensitive to hyperparameters. Choices related to network architecture, graph sampling, and loss function optimization directly impact performance, increasing the difficulty of GNN design and practical deployment.

#### 4.2.4 Generative Methods

Generative methods have shown remarkable promise in learning cross-modal relationships by synthesizing and aligning high-dimensional data from different modalities. Generative adversarial networks (GANs) remain a foundational model in this domain [175, 172, 174, 176], offering effective solutions for complex mappings between modalities. For example, DMF-GAN integrates multi-head attention and recurrent semantic fusion networks to achieve fine-grained text-to-image synthesis, significantly improving semantic alignment between modalities [221]. Similarly, GAN-based frameworks have been used for multimodal MRI synthesis, where a single generator learns unified mappings across image modalities [41].

Variational autoencoders (VAEs) also play a key role in multimodal alignment. By projecting data into shared latent spaces, VAEs enable the fusion of semantic information across modalities. This technique has proven effective in compositional tasks like image-text representation learning [213], and cross-modal quantization using VAEs has further demonstrated success in aligning text and image representations [86].

A significant recent development in generative multimodal modeling is the adoption of diffusion models. These models offer a robust alternative to GANs and VAEs, especially in terms of stability, mode diversity, and representation fidelity [26]. Diffusion-driven fusion techniques have enabled tasks such as face image generation from both visual prompts and text, showing strong cross-modal alignment by integrating latent representations from both GANs and diffusion models [77]. Furthermore, semi-supervised methods like diffusion transport

alignment (DTA) leverage diffusion processes for manifold alignment using minimal supervision, effectively capturing geometric similarities across modalities [51].

More broadly, diffusion models have been successfully applied in multimodal generative frameworks like Stable Diffusion and DALL-E, enabling synthesis tasks such as image-to-audio and text-to-video generation by iteratively denoising representations conditioned on multimodal inputs [21, 26]. These approaches not only enhance the quality and coherence of generated outputs but also support flexible alignment by embedding conditional semantics at multiple stages of the generation process.

In summary, the evolution from GANs and VAEs to diffusion models marks a paradigm shift in generative multimodal fusion and alignment, offering better performance, interpretability, and multimodal coherence.

#### 4.2.5 Contrastive Methods

Contrastive learning has become a cornerstone in multimodal alignment and fusion due to its ability to bring semantically related modalities closer in a shared embedding space. One of the most influential contrastive architectures is CLIP [148], which aligns text and image embeddings through large-scale pretraining on paired image-text data. CLIP and its variants have set new benchmarks in vision-language tasks, and inspired a wide range of fusion strategies.

At its core, CLIP learns aligned representations by training two encoders—one for images and one for text—so that the embeddings of matching image-text pairs are pulled closer together, while those of mismatched pairs are pushed apart. This is achieved by contrasting each sample against a batch of negatives, encouraging the model to capture semantic correspondences without requiring explicit annotations. The result is a powerful zero-shot transfer capability, where the model can generalize to unseen categories simply by encoding their textual descriptions [148].

The original CLIP model aligns global representations of images and text using a simple contrastive loss, achieving strong performance on zero-shot classification and retrieval tasks [148]. Recent advances have extended CLIP’s architecture to improve representation granularity and domain adaptability. For instance, HiCLIP enhances CLIP by incorporating hierarchy-aware attention in both visual and textual branches to better model fine-grained semantic relationships [57]. Similarly, Set-CLIP reformulates alignment as a manifold matching problem and introduces semantic density loss to improve contrastive alignment even in low-alignment settings without paired data [165].

Other works, such as ComKD-CLIP, use contrastive distillation to transfer alignment knowledge from large CLIP models into smaller networks using image-text fusion attention mechanisms [36]. SyCoCa further augments contrastive captioners by introducing bidirectional attention flows and text-guided masked image modeling to unify global and local cross-modal alignments [122]. Furthermore, models like Domain-Aligned CLIP explore few-shot adaptation through intra- and inter-modal contrastive learning without full finetuning [59].

The influence of CLIP-based models extends beyond standard vision-language domains. Applications such as image-guided editing [212] and 3D representation alignment [88] demonstrate the adaptability of contrastive alignment for diverse multimodal scenarios.

Together, these advances highlight how CLIP and its extensions form a central framework for contrastive-based multimodal alignment, driving both theoretical insight and practical performance gains.

#### 4.2.6 Attention-based Methods

Before the widespread adoption of attention mechanisms, earlier methods such as OSCAR [99], UNITER [35], VILA [106], and VinVL [244] relied on object detectors to extract modality features, followed by relatively shallow fusion strategies. These pipelines lacked dynamic alignment and adaptive weighting capabilities, which are now addressed by attention-based models. Later models such as CLIP [148] significantly advanced image-text representation learning through contrastive pretraining.

However, the cross-modal interaction in CLIP was limited to a dot product between global embeddings, lacking fine-grained alignment or deep fusion at the token level [79]. This shallow attention interaction hindered the model’s ability to fully capture complex semantic relationships across modalities, motivating the development of more integrated fusion mechanisms.

To address this limitation, methods focusing on deeper inter-modal interactions were developed, often employing Transformer encoders or other complex architectures to achieve higher-level modality integration [15]. The introduction of the Vision Transformer (ViT) marked a significant shift in multimodal learning.

ViLT [79] demonstrates the feasibility of performing multimodal tasks without convolutional networks or region supervision, using Transformers exclusively for feature extraction and processing. However, the simplistic structure of ViLT led to performance issues, particularly when compared to methods that emphasized deeper inter-modal interactions and fusion [97, 15, 222]. ViLT lagged behind these methods in many tasks, possi-

bly due to dataset bias or the inherent need for stronger visual capabilities [97]. Generally, visual models need to be larger than text models to achieve better results, and the performance degradation was not primarily caused by the lightweight visual embedding strategy.

Subsequent works, such as ALBEF [97], introduced more sophisticated model designs. ALBEF emphasized aligning image and text representations before their fusion using a contrastive loss. By employing momentum distillation, it generated pseudo-labels to mitigate challenges posed by noisy datasets. Following this, BLIP [96] adopted a bootstrapping mechanism, using initially generated captions from the model to filter out dataset noise, thereby improving the quality of subsequent training.

Attention-based mechanisms gained prominence with the introduction of the Transformer architecture [194] and following works [44, 260, 60, 241], where the attention function takes queries ( $Q$ ), keys ( $K$ ), and values ( $V$ ) and computes the relevance of each key to a given query. The scaled dot-product attention is given by:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V. \quad (7)$$

This operation dynamically weighs each input feature according to its contextual importance, enabling the model to capture long-range dependencies across sequences or modalities.

The use of attention mechanisms enables decoders to focus selectively on specific subcomponents of the source input. This contrasts with traditional encoder-decoder models that treat all source features as a single representation. Attention modules guide decoders to highlight task-relevant regions—such as specific image patches, words in a sentence, or audio frames—during output generation. For instance, in image captioning, attention allows decoders to attend to relevant visual regions when generating each word, instead of encoding the entire image as a static vector [75].

Considering the inherent connection between attention score and similarity score, attention mechanisms are widely applied for alignment, i.e., learning correspondences between semantically similar elements across modalities. For example, the Att-Sinkhorn method uses the Sinkhorn distance in conjunction with attention to model cross-modal optimal transport, aligning features from different distributions [120]. The AbFTNet model emphasizes “alignment-before-fusion” by first synchronizing modality-specific features via a Transformer-based mechanism and then integrating them through a cross-modal aggregation module [136]. In the domain of knowledge representation, DSEA applies a dynamic self-attention network to evaluate the importance of

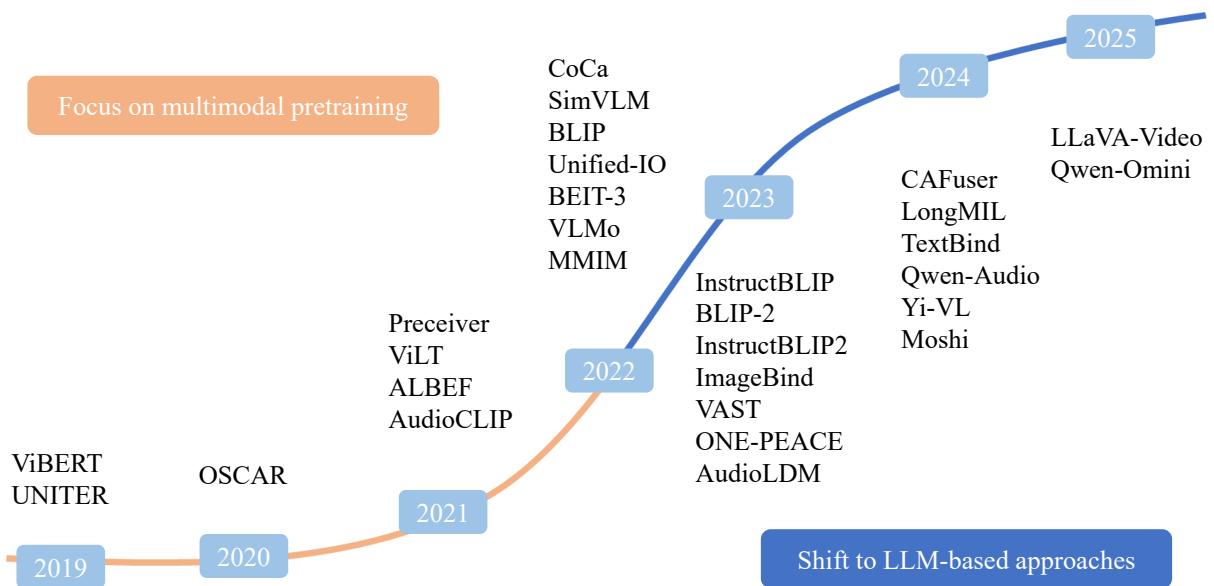


Fig. 6: Timeline of multimodal attention-based models from 2019 to 2025. From 2019 to 2022, researches heavily focused on integrating the transformer architecture for multimodal pretraining. From 2023 to 2025, the core concern has been shifted to how to reuse the knowledges of LLMs. 2019: ViLBERT [116], UNITER [35]; 2020: OSCAR [99]; 2021: ViLT [79], ALBEF [97], Perceiver [70], AudioCLIP [62]; 2022: Unified-IO [117], BEIT-3 [204], BLIP [96], VLMo [17], CoCa [227], MMIM [64], SimVLM [209]; 2023: ImageBind [58], VAST [34], ONE-PEACE [202], AudioLDM [112], InstructBLIP [40], BLIP-2 [95], InstructBLIP2 [30]; 2024: TextBind [91], CAFuser [23], LongMIL [89], Qwen-Audio [38], Yi-VL [226], Moshi [52]; 2025: LLaVA-Video [251], Qwen-Omini [217].

structural and attribute information, improving entity alignment across multimodal knowledge graphs [145].

Attention-based fusion integrates multimodal information by learning how much to attend to each modality. This selective integration helps manage noise, modality imbalance, and complementary cues. For instance, BMAN applies multi-head attention to align and fuse audio-text features with learnable weights, improving sentiment prediction across unaligned datasets [245]. ProVLA employs a cross-attention fusion encoder that progressively aligns vision and language representations for compositional image retrieval [68].

Attention-based fusion is particularly effective in multimodal tasks because it supports flexible integration and handles modality-specific uncertainties [177, 200]. However, this flexibility comes at the cost of increased computational complexity and often demands large-scale annotated datasets.

Recent advances such as ProVLA [68] and AbFTNet [136] take Transformer-based fusion further by introducing progressive alignment and structured attention mechanisms prior to fusion. ProVLA employs a two-stage alignment-fusion paradigm, leveraging cross-attention for robust semantic integration and using momentum-based hard negative mining to enhance alignment robustness.

Similarly, AbFTNet introduces a CAP (Cross-modal Aggregation Promoting) module, aligning unimodal features through self-attention before cross-modal integration, thus addressing modality-specific information disparities.

Fusion mechanisms like TokenFusion [129] also explore token-level replacement and residual alignment to balance modality contributions and avoid attention dilution. These designs allow Transformers to retain unimodal strengths while gaining inter-modal awareness through informed token substitution and dynamic fusion.

CoCa [227] combined contrastive loss with captioning loss, achieving remarkable performance. In particular, CoCa excelled not only in multimodal tasks, but also performed well on single-modal tasks such as ImageNet classification. BEIT-3 [204] further advanced multimodal learning with the implementation of Multiway Transformers, enabling the simultaneous processing of images, text, and image-text pairs. By applying masked data modeling to these inputs, BEIT-3 achieved state-of-the-art performance across various visual and vision-language tasks.

Figure 6 illustrates the relationships among major works related to attention mechanisms and transform-

ers and Table 3 provides a summary of representative models.

#### 4.2.7 LLM-based Methods

As mentioned in Section 4.2.6 and Figure 6, the trend has shifted to LLM-based models from 2022. Figure 7 illustrates a common scenario of LLM-based method. After the encoder extracts features from each modality, a connector maps these features into the text space, where they are processed together by the LLM. Previously, this connector was often a simple MLP, but it can now be a more complex attention mechanism. Recently, researchers have proposed various architectures and techniques aimed at enhancing cross-modal capabilities. They embed adapters into frozen LLMs to facilitate interactions between modalities [118]. Figure 8 shows the typical structure of this approach. The key difference from previous methods is that adapters are embedded directly into the LLMs, allowing for end-to-end training with alignment included. For example, the Qwen-VL series models [14] advanced cross-modal learning through the design of visual receptors, input-output interfaces, and a multi-stage training pipeline, achieving notable performance in image and text understanding, localization, and text reading. In video understanding, the ViLA network [205] introduced a learnable text-guided Frame-Prompter and a cross-modal distillation module (QFormer-Distiller) optimized for key frame selection, improving both accuracy and efficiency in video-language alignment. Additionally, CogVLM [203] incorporated visual expertise into pretrained language models using Transformers. In emotion recognition tasks, COLD Fusion added an uncertainty-aware component for multimodal emotion recognition [181].

Various pre-training strategies have been developed to facilitate multimodal fusion. For example, BLIP-2 [95] introduced a bootstrapping approach that used frozen image encoders and large language models for vision-language pre-training, reducing the number of parameters while enhancing zero-shot learning performance. Similarly, the VAST model [34] explored a comprehensive multimodal setup involving vision, audio, subtitles, and text, constructing a large-scale dataset and training a foundational model capable of perceiving and processing all these modalities. Furthermore, the ONE-PEACE model [202] employed a modular adapter design and shared self-attention layers to provide a flexible and scalable architecture that could be extended to more modalities. The research by Zhang et al. [239] used Transformers for end-to-end anatomical and functional image fusion, leveraging self-attention to incorporate global contextual information.

Despite these advances, the field still faces several challenges. One of the main challenges is data bias, where inherent biases in training datasets limit model performance. Another concern is maintaining consistency across modalities to ensure coherent information integration without loss or inconsistency. Additionally, as models grow in scale, there is an increasing demand for computational resources, necessitating more efficient algorithms and hardware support. Table 4 summarizes some state-of-the-art (SOTA) or popular LLM-based models.

In conclusion, multimodal fusion remains a dynamic and evolving area of research, driven by advances in attention-based mechanisms and model architectures. Although significant progress has been made in developing models that effectively integrate information from multiple modalities, ongoing challenges such as data bias, modality consistency, and computational demands persist. Continued exploration of new theoretical frameworks and technical solutions is necessary to achieve more intelligent and adaptable multimodal systems, advancing artificial intelligence technologies, and providing powerful tools for practical applications.

## 5 Challenges in Multimodal Alignment and Fusion

### 5.1 Multimodal Misalignment and Modality Gap

Multimodal misalignment and modality gap are two critical challenges in multimodal representation learning that significantly affect model performance. Misalignment refers to the mismatch between different modalities, such as images and their corresponding textual descriptions, which can arise due to noisy or incorrect annotations [185]. For instance, Ma et al. [121] identified modality misalignment as a significant barrier to transferring knowledge across different modalities, emphasizing that pre-trained models frequently struggle with knowledge transfer when there is a substantial semantic gap between modalities.

The modality gap, on the other hand, describes the disparity in embedding distributions of different modalities within a shared space, leading to suboptimal cross-modal interactions [105]. The modality gap in multimodal contrastive representation learning arises due to a combination of factors. First, deep neural networks inherently create a “cone effect”, where embeddings from a single modality are restricted to a narrow region of the embedding space. This geometric bias is amplified by nonlinear activation functions and network depth. Second, different random initializations for separate encoders in multimodal models result in distinct cones

Table 3: Summary of attention-based multimodal models.

Model	Year	Modality	Training Modules
ViLBERT [116]	2019	Vision, Text	Dual-stream co-attentional transformer, pre-trained on Conceptual Captions, for task-agnostic visiolinguistic representation in VQA, VCR, referring expressions, and image retrieval.
UNITER [35]	2019	Vision, Text	MLM, MRM, ITM, and WRA via Optimal Transport for fine-grained alignment.
Oscar [99]	2020	Vision, Text	Object-Semantics Aligned Pre-training using detected object tags as anchor points.
ViLT [79]	2021	Vision, Text	Masked Language Modeling; Image-Text Matching; Word-Patch Alignment.
ALBEF [97]	2021	Vision, Text	Image-Text Contrastive Loss; Image-Text Matching Loss; Masked-Language-Modeling Loss.
Perceiver [70]	2021	Arbitrary Modalities	General-purpose architecture handling sequences of varying modalities.
AudioCLIP [62]	2021	Audio, Vision, Text	Contrastive learning extended from CLIP to audio domain.
Unified-IO [117]	2022	Vision, Text	Object Segmentation; Visual Question Answering; Depth Estimation; Object Localization.
BEIT-3 [204]	2022	Vision, Text	Masked "language" modeling for images, texts, and image-text pairs.
BLIP [96]	2022	Vision, Text	Image-Text Contrastive Loss; Image-Text Matching Loss; Language Modeling Loss.
VLMo [17]	2022	Vision, Text	Image-Text Contrastive Learning; Masked Language Modeling; Image-Text Matching.
CoCa [227]	2022	Vision, Text	Captioning Loss; Contrastive Loss.
MMIM [64]	2022	Audio, Text, Non-Verbal Context	Multi-task learning with unobserved multimodal context for sentiment analysis.
ImageBind [58]	2023	Vision, Text, Audio, Depth, Thermal, IMU	Zero-shot alignment across 6 modalities via image-centric binding.
VAST [34]	2023	Vision, Text	OM-VCC; OM-VCM; OM-VCG
ONE-PEACE [202]	2023	Vision, Audio, Text	Masked Contrastive Learning; supports multi-modal fusion through self-attention layers.
Coupled Mamba [98]	2024	Vision, Text, Audio	Enhanced multimodal fusion with State Space Models (SSMs) for non-LLM architectures.
CAFuser [23]	2024	Vision, Text	Image-Text Contrastive Loss
LongMIL [89]	2024	Medical Images, Text	Local-Global Hybrid Transformer architecture for long-context multiple instance learning; Self-supervised contrastive learning for medical image-text alignment.

in the shared embedding space, causing a separation between modalities even at initialization. Finally, during training, contrastive learning preserves this gap. The contrastive loss, especially at low temperatures, maintains a repulsive structure in the optimization landscape, preventing the gap from closing. Together, these factors explain the persistent separation between modalities observed in multimodal models like CLIP [105].

Many approaches seek to mitigate the modality gap through various architectural and training innovations. For instance, noise-injected embeddings, such as those used in CapDec, improve generalization by perturbing CLIP embeddings to reduce overfitting to specific modality characteristics and enhance alignment in low-resource settings [9]. Meanwhile, VT-CLIP improves modality alignment by incorporating visual-guided text generation to highlight image regions that correspond to key

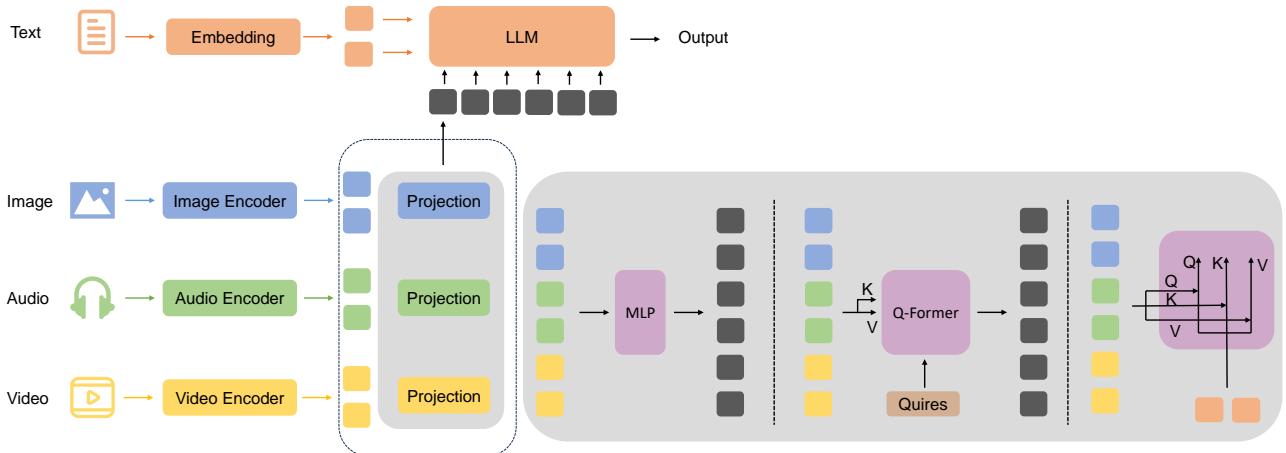


Fig. 7: This pipeline demonstrates multimodal fusion using a large language model (LLM). Text inputs are embedded and processed by the LLM, while image, audio, and video inputs are encoded, projected into a shared embedding space, and passed through modules such as an MLP and Q-Former. The Q-Former uses attention mechanisms (queries, keys, and values) to align multimodal features before generating a final output through the LLM.

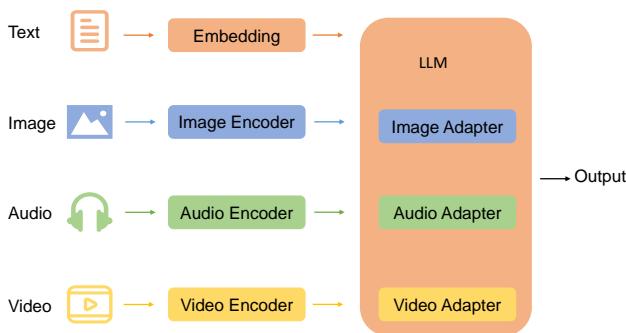


Fig. 8: Adapters embedded in LLM. Each modality (text, image, audio, and video) is processed by its respective encoder and a modality-specific adapter. These adapters then feed the encoded features into an LLM, which generates the final output.

linguistic cues [147]. Finite discrete tokens (FDT) aim to resolve the granularity gap between visual patches and textual tokens by embedding both into a unified semantic space [37]. Modality knowledge alignment (MoNA) introduces a meta-learning paradigm to learn target data transformations that reduce modality knowledge discrepancies prior to transfer [121]. Recently, Tong et al. [185] investigate the gap between the visual embedding space of CLIP and vision-only self-supervised learning, proposing a mixture of features (MoF) approach to enhance visual grounding capabilities.

To advance this field, future research should focus on developing more sophisticated embedding space modeling techniques, dynamically adjusting modality gaps during training, and improving data quality through

better annotation practices. Additionally, enhancing the compositional understanding of VLMs by incorporating syntactic structure and word order sensitivity—such as generating targeted negative captions through linguistic element swapping—could further improve model capabilities [232]. These directions aim to bridge existing gaps in multimodal learning and push models toward a more human-like comprehension of multimodal input.

## 5.2 Computational Efficiency Challenge

Early multimodal models faced significant computational demands due to their reliance on object detectors, particularly during inference. The introduction of vision transformers (ViTs) enabled patch-based visual representations instead of bounding boxes, reducing computational complexity. However, simply tokenizing textual and visual features remains insufficient for effective multimodal processing. Recent work has proposed several efficient fusion mechanisms to mitigate this challenge.

TokenFusion, for instance, dynamically replaces less informative tokens with fused inter-modal features to reduce token redundancy in transformer architectures [129]. Attention bottlenecks allow selective modality interaction through shared tokens, enabling low-rank communication between modalities with minimal overhead [132]. Prompt-based multimodal fusion (PMF) introduces deep-layer prompts that interact across pretrained unimodal Transformers, achieving comparable performance to full fine-tuning [101]. Other notable methods include sparse fusion transformers (SFT), which sparsify unimodal tokens prior to fusion [46], and dynamic multimodal fusion (DynMM), which adaptively determines forward com-

Table 4: Summary of LLM-based multimodal models with diverse modalities.

Model	Year	Modality	Used LLM	Training Modules
MiniGPT-4 [259]	2023	Text, Image	Vicuna	Two-stage training: Stage 1: Freeze visual feature extractor, train projection layer to align visual features with Vicuna; Stage 2: Instruction finetuning on dialogue data
Qwen-VL [14]	2023	Text, Image	Qwen-7B	Stage 1: Image caption generation; Stage 2: Multi-task pretraining; Stage 3: Supervised finetuning
BLIP-2 [95]	2023	Text, Image	OPT, FlanT5	Stage 1: Vision-Language Representation Learning; Stage 2: Vision-to-Language Generation Learning
LLaVA [113]	2023	Text, Image	GPT-3, GPT-3.5, LLaMA	Visual Instruction Tuning
LaVIN [118]	2023	Text, Image	LLaMA	Fine-tuning with MoE adapter
MiniGPT-v2 [32]	2023	Text, Image	Vicuna (7B/13B)	Multitask learning
InstructBLIP [40]	2023	Text, Image	Vicuna (7B/13B)	Visual Instruction Tuning
InternLM-XComposer [242]	2023	Text, Image	InternLM-Chat-7B	Pre-training, Multi-task Training, Instruction Fine-tuning
Macaw-LLM [119]	2023	Text, Image, Audio, 3D	LLaMA	Multimodal language modeling with unified representation
3D-MMLM [66]	2024	Text, Image, 3D	LLaMA-3	3D understanding, point cloud processing, cross-modal alignment
Qwen2-VL [201]	2024	Text, Image	Qwen-2	Visual Instruction Tuning
Moshi [52]	2024	Text, Audio	LLaMA-2	Text-to-speech, speech-to-text, audio understanding
MM-LLMs [237]	2024	Text, Image, Video, Audio, 3D	Various	Handling different modalities where X can be image, video, audio, 3D, etc.
AudioPaLM [154]	2023	Text, Audio	PaLM-2	Speech recognition, speech synthesis, multilingual audio understanding
Qwen-Audio [38]	2024	Text, Audio	Qwen-1.5	Audio instruction tuning, audio-text alignment
Yi-VL [226]	2024	Text, Image	Yi-Chat	Three-stage training: 1. Train ViT and projection module; 2. Increase image resolution and retrain; 3. Train entire model
InternLM-XComposer-2.5 [243]	2024	Text, Image	InternLM2-7B	Pre-training, Multi-task Training, Instruction Fine-tuning
CogVLM [203]	2024	Text, Image	LLaMA-2	Pre-training + Supervised Fine-tuning on vision-language tasks
ViLA [205]	2024	Text, Image	Supports Frozen and Finetuned (LoRA) usage of LLM	Distillation loss; Visual Question Answering loss
LLaVA-Video [251]	2025	Text, Image, Video	LLaMA-2	Video instruction tuning, video question answering
Qwen2.5-Omni [217]	2025	Text, Image, Audio, Video, 3D	Qwen-2.5	Video and audio branches with cross-attention for multimodal understanding

putation paths using data-dependent gating [219]. Low-rank tensor fusion approaches have also been explored, leveraging compact representations to avoid exponential parameter growth [115]. Recently, GeminiFusion proposed a pixel-wise fusion method with linear complexity by combining intra- and inter-modal attention, controlled by layer-adaptive noise modulation [73].

Despite these advances, fusion remains the computational bottleneck in large-scale multimodal models. Future research should focus on developing adaptive,

scalable, and resource-efficient fusion architectures to meet the growing demand of real-world multimodal tasks.

### 5.3 Data Quality and Availability

One of the primary obstacles is the scarcity of large-scale, high-quality multimodal datasets, which are essential for training robust multimodal language models

(MLLMs) [142, 22]. Complex and unbiased data that accurately reflect the richness of reality are necessary to train these models effectively. This challenge is particularly pronounced in specialized fields, such as nuclear medicine, where access to sufficient clinical data for model refinement is limited [22]. Furthermore, the need for task-specific and domain-specific datasets adds to the complexity of data collection and integration processes [142].

Large-scale multimodal datasets obtained from the Internet, such as image-caption pairs, often contain mismatches or irrelevant content between images and their corresponding texts. This issue arises mainly because these image-text pairs are optimized for search engines rather than for precise multimodal alignment. Consequently, models trained on such noisy data may struggle to generalize effectively. Tong et al. [185] identifies specific instances where advanced systems like GPT-4V struggle with VQA due to inaccurate visual grounding. To address this problem, several approaches have been proposed to improve data quality.

Kim et al. [78] introduced a novel methodology called hyperbolic entailment filtering (HYPE), which goes beyond traditional CLIP-based filtering by incorporating both alignment and specificity metrics. HYPE leverages hyperbolic embeddings and entailment cones to filter out samples with underspecified or meaningless semantics, ensuring better cross-modal alignment and modality-wise meaningfulness.

Other notable efforts include Nguyen et al. [134], who tackled noise in web-scraped datasets by using synthetic captions generated through image captioning models. By integrating synthetic descriptions with the original captions, they achieved improvements in data utility across multiple benchmark tasks, demonstrating that improved caption quality can significantly benefit model performance. Similarly, CapsFusion [229] introduced a framework that leverages large language models to refine synthetic and natural captions in multimodal datasets, thus improving caption quality and sample efficiency for large-scale models. Furthermore, the LAION-5B dataset [157] provides a large collection of CLIP-filtered image-text pairs, showing that combining high data volume with effective filtering can enhance the robustness and zero-shot capabilities of vision language models.

Despite these improvements, challenges remain in scalable data filtering and maintaining diversity. For example, DataComp [56] has shown that even with effective filtering, achieving high-quality and diverse representation in large multimodal datasets is complex. It requires ongoing innovation in data pruning and quality assessment to ensure that models trained on these datasets generalize effectively across domains. HYPE's

ability to consider both cross-modal alignment and intra-modal specificity offers a promising direction for addressing these limitations, especially in large-scale settings.

In summary, while synthetic captioning and large-scale filtering methods have improved the quality of multimodal datasets, further advances in scalable filtering techniques and diversity retention are needed to fully address the challenges associated with web-scraped multimodal datasets.

#### 5.4 Scale of Training Datasets Challenge

Another significant challenge in multimodal learning is acquiring sufficiently large and high-quality datasets for model training, particularly for combining vision and language tasks. There is a pressing need for extensive and reliable datasets that can be used to train models effectively across a variety of tasks. For instance, the introduction of the LAION-5B dataset, comprising billions of CLIP-filtered image-text pairs, has provided a scalable, open-source dataset that supports training and fine-tuning large-scale vision-language models, helping democratize access to high-quality data [157]. Similarly, the WIT dataset enables multimodal, multilingual learning by offering a curated, entity-rich dataset sourced from Wikipedia, featuring a high degree of concept and language diversity, which has proven beneficial for downstream retrieval tasks [166].

Although these datasets represent substantial progress, scalability and data quality remain challenging. For example, [199] proposes compressing vision-language pre-training (VLP) datasets to retain essential information while reducing redundancy and misalignment, resulting in a smaller but higher-quality training set. Additionally, scaling techniques like sparse mixture of experts (MoE) [159] aim to improve the efficiency of large models by training specialized sub-models within a unified framework, balancing compute costs and performance. While these innovations are steps toward addressing data scale and quality challenges, efficient access to diverse and large datasets for multimodal learning remains a difficulty for the research community.

#### 5.5 Ethical Bias

Ethical bias represents a critical and under-addressed challenge in multimodal alignment and fusion. These systems often inherit and even amplify biases present in individual modalities, such as language, vision, or speech. Studies have demonstrated that fusion processes may introduce new forms of unfairness not present in unimodal inputs, due to disparate representations and

unequal weighting of modalities [220, 152]. For example, [220] revealed that different modalities in personality assessment contribute uniquely to bias, and simple fusion can worsen demographic disparities. Similarly, [250] developed a theoretical framework showing that multimodal networks often suffer from unimodal bias, where a dominant modality suppresses others, leading to skewed outputs. Ethical alignment frameworks, such as RoBERTa-based classifiers trained on human ethical feedback, have been proposed to better detect and mitigate bias in multimodal responses [152]. Moreover, novel alignment datasets specifically designed to reduce gender bias in language models are also being explored to guide fairer training processes [246]. These findings underscore the need for standardized bias auditing protocols and ethically-aware training pipelines when developing multimodal systems.

## 6 Discussion and Future Directions

Multimodal alignment and fusion techniques have made significant strides in enabling complex reasoning capabilities across tasks such as visual question answering, spatial localization, and semantic composition. Recent developments demonstrate that the effectiveness of these techniques is highly task-dependent, with distinct strategies outperforming in compositional versus spatial reasoning scenarios.

For spatial reasoning, recent research underscores the continued value of incorporating explicit spatial alignment even in deep learning architectures. Wang et al. demonstrated that spatially aligning medical images before feeding them into deep fusion networks leads to notable improvements in diagnostic accuracy. Their addition of a spatial transformer network (STN) module further enhanced this effect, providing adaptive alignment during training [206]. Similarly, the MulFS-CAP framework bypasses traditional registration by jointly learning alignment and fusion in a unified network, which proves especially effective for unregistered infrared-visible image pairs [90]. These results suggest that both explicit and implicit spatial alignment strategies are essential for spatially sensitive fusion tasks.

In contrast, compositional reasoning benefits more from methods emphasizing feature-level integration and cross-modality contextual alignment. ST-Align introduces a foundation model for spatial transcriptomics that aligns image and gene data across spatial scales, enabling nuanced cross-level semantic understanding [110]. Meanwhile, Set-CLIP addresses the low-alignment data challenge by learning modality alignment through distribution-level semantic constraints, demonstrating

that strong performance can be achieved even in semi-supervised settings [165]. This highlights the importance of representation space regularization and contrastive learning in achieving effective compositional fusion.

From these studies, several **generalizable lessons** emerge:

- *Implicit fusion architectures* with built-in alignment (e.g., STN, shared encoders) are increasingly favored for practical deployment due to reduced preprocessing demands.
- *Cross-level fusion frameworks*, such as those combining local and global spatial cues, significantly enhance performance in hierarchical reasoning tasks.
- *Modality-specific feature preservation*, via dictionaries or attention-based fusion, allows models to better retain complementary information during integration.
- *Distribution-aware alignment strategies* improve robustness in data-scarce or weakly aligned settings.

Future directions include developing unified multimodal architectures that balance task-specific performance with generalization. More interpretable fusion methods are also critical, especially in clinical or scientific domains. Finally, the field would benefit from standardized multimodal benchmarks that isolate spatial and compositional reasoning tasks, enabling more consistent evaluation of fusion techniques.

Overall, effective multimodal fusion requires a deliberate choice of alignment and integration strategies, tailored to the reasoning demands of the target application. Spatial tasks benefit from precise alignment techniques, while compositional reasoning favors context-aware and semantically structured fusion methods.

## 7 Conclusion

Multimodal alignment and fusion are fundamental to advancing artificial intelligence systems capable of understanding and interacting with complex real-world data. Our survey has provided a comprehensive review of over 200 studies, categorizing existing techniques from both structural and methodological perspectives. Despite notable progress—particularly in contrastive learning, attention-based models, and LLM-driven architectures—achieving robust and scalable integration across modalities remains a significant challenge.

Current approaches continue to grapple with critical issues, including modality misalignment, inconsistent data quality, and high computational overhead. Additionally, the modality gap and the scarcity of large-scale, high-quality datasets hinder the full realization of effective multimodal learning. Although many techniques

have advanced the field, limitations persist in aligning heterogeneous signals and managing data noise.

To bridge these gaps, future research should prioritize the development of adaptive, noise-resilient frameworks, capable of dynamically adjusting to diverse modalities while maintaining interpretability. Promising directions include efficient token-level fusion mechanisms, cross-modal graph reasoning, and alignment-aware training objectives that explicitly mitigate the modality gap. Furthermore, continued innovation in dataset construction, annotation quality, and filtering strategies—such as hyperbolic entailment filtering and synthetic captioning—will be essential for improving model robustness.

By addressing these challenges, the next generation of multimodal systems can become more versatile, efficient, and generalizable, paving the way for broader deployment in domains such as healthcare, autonomous systems, and human-computer interaction.

## References

- Kenan E. Ak, Gary Lee, Yan Xu, and Mingwei Shen. Leveraging efficient training and feature fusion in transformers for multimodal classification. In *2023 IEEE International Conference on Image Processing (ICIP)*, pages 1420–1424, 2023.
- S. Akaho. A kernel method for canonical correlation analysis. In *Proceedings of the International Meeting on Psychometric Society*, 2001.
- A. Akhmerov, A. Vasilev, and A.V. Vasileva. Research of spatial alignment techniques for multimodal image fusion. In *Proceedings of the SPIE*, volume 11059, pages 1105916 – 1105916–9, 2019.
- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: A visual language model for few-shot learning. In *NeurIPS*, 2022.
- Douglas L. Allaire and Karen E. Willcox. Fusing information from multifidelity computer models of physical systems. *2012 15th International Conference on Information Fusion*, pages 2458–2465, 2012.
- Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. Deep canonical correlation analysis. In *Proceedings of the 30th International Conference on Machine Learning*, 2013.
- Adam Arany, Bence Bolgár, Balázs Balogh, Péter Antal, and Péter Mátyus. Multi-aspect candidates for repositioning: data fusion methods using heterogeneous information sources. *Current medicinal chemistry*, 20 1:95–107, 2012.
- Masataro Asai. Set cross entropy: Likelihood-based permutation invariant loss function for probability distributions, 2018.
- Association for Computational Linguistics 2022, Amir Globerson, Ron Mokady, and David Nukrai. Text-only training for image captioning using noise-injected clip, 2022.
- S. Ayache, Georges Quénod, and Jérôme Gensel. Classifier fusion for svm-based multimedia semantic indexing. In *European Conference on Information Retrieval*, 2007.
- F. R. Bach and M. I. Jordan. Kernel independent component analysis. *Journal of Machine Learning Research*, 3:1–48, 2002.
- Roman Bachmann, Oğuzhan Fatih Kar, David Mizrahi, Ali Garjani, Mingfei Gao, David Griffiths, Jiaming Hu, Afshin Dehghan, and Amir Zamir. 4m-21: An any-to-any vision model for tens of tasks and modalities, 2024.
- Vijay Badrinarayanan, Alex Kendall, and Roberto Cipolla. Segnet: A deep convolutional encoder-decoder architecture for image segmentation. *IEEE TPAMI*, 2015.
- Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.
- Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE TPAMI*, 41(2), 2018.
- Oresti Banos, Rafael Garcia, Juan A. Holgado-Terriza, Miguel Damas, Hector Pomares, Ignacio Rojas, Alejandro Saez, and Claudia Villalonga. mhealthdroid: A novel framework for agile development of mobile health applications. In Leandro Pecchia, Liming Luke Chen, Chris Nugent, and José Bravo, editors, *Ambient Assisted Living and Daily Activities*, pages 91–98, Cham, 2014. Springer International Publishing.
- Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhrojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts, 2022.
- Arnab Barua, Mobyen Uddin Ahmed, and Shahina Begum. A systematic literature review on multimodal machine learning: Applications, challenges, gaps and future directions. *IEEE Access*, 11:14804–14831, 2023.
- Erik Blasch, Georgiy M. Levchuk, Gennady Staskevich, Dustin Burke, and Alex Aved. Visualization of graphical information fusion results. In *Defense + Security Symposium*, 2014.
- K. Boehm, P. Khosravi, R. Vanguri, Jianjiong Gao, and S. Shah. Harnessing multimodal data integration to advance precision oncology. *Nature Reviews Cancer*, 22:114–126, 2021.
- Fouad Bousetouane. Generative ai for vision: A comprehensive study of frameworks and applications, 2025.
- Tyler J. Bradshaw, Xin Tie, Joshua Warner, Junjie Hu, Quanzheng Li, and Xiang Li. Large language models and large multimodal models in medical imaging: A primer for physicians. *Journal of Nuclear Medicine*, 2025.
- Tim Broedermann, Christos Sakaridis, Yuqian Fu, and Luc Van Gool. Condition-aware multimodal fusion for robust semantic perception of driving scenes, 2024.
- Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, and et al. Language models are few-shot learners.
- Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. Coyo-700m: Image-text pair dataset, 2022.
- Yongnian Cao, Xuechun Yang, and Rui Sun. Generative ai models: Theoretical foundations and algorithmic practices. *Journal of Industrial Engineering and Applied Science*, 3(1):1–9, 2025.

27. Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.
28. D. Chen, J. Chen, L. Yang, and F. Shang. Mix-tower: Light visual question answering framework based on exclusive self-attention mechanism. *Neurocomputing*, 2024.
29. Feilong Chen, Minglun Han, Haozhi Zhao, Qingyang Zhang, Jing Shi, Shuang Xu, and Bo Xu. X-llm: Bootstrapping advanced large language models by treating multi-modalities as foreign languages, 2023.
30. H. Chen and T. Xu. Instructblip 2: Extending vision-language models with fine-grained instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2023.
31. Jiayi Chen and Aidong Zhang. Hgmf: Heterogeneous graph-based fusion for multimodal data with incompleteness. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2020.
32. Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning, 2023.
33. K. Chen and Y. Sun. Llava-med: Medical image understanding with large language models. *IEEE TNNLS*, 2023.
34. Sihan Chen, Handong Li, Qunbo Wang, Zijia Zhao, Mingzhen Sun, Xinxin Zhu, and Jing Liu. Vast: A vision-audio-subtitle-text omni-modality foundation model and dataset, 2023.
35. Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2020.
36. Yifan Chen, Xiaozhen Qiao, Zhe Sun, and Xuelong Li. Comkd-clip: Comprehensive knowledge distillation for contrastive language-image pre-training model, 2024.
37. Yuxiao Chen, Jianbo Yuan, Yu Tian, Shijie Geng, Xinyu Li, Ding Zhou, Dimitris N. Metaxas, and Hongxia Yang. Revisiting multimodal representation in contrastive learning: From patch and token embeddings to finite discrete tokens. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15095–15104, 2023.
38. Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023.
39. Y. Cui, S. Liang, and YY Zhang. Multimodal representation learning for tourism recommendation with two-tower architecture. *PLoS One*, 2024.
40. Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning, 2023.
41. X. Dai, Y. Lei, Y. Fu, W. Curran, T. Liu, H. Mao, and Xiaofeng Yang. Multimodal mri synthesis using unified generative adversarial networks. *Medical Physics*, 2020.
42. Gokulesh Danapal, Giovanni A. Santos, João Paulo C. L. da Costa, Bruno J. G. Praciano, and Gabriel P. M. Pinheiro. Sensor fusion of camera and lidar raw data for vehicle detection. In *2020 Workshop on Communication Networks and Power Systems (WCNPS)*, pages 1–6, 2020.
43. Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: web-curated image-text data created by the people, for the people, 2021.
44. Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.
45. Lei Ding, Dong Lin, Shaofu Lin, Jing Zhang, Xiaojie Cui, Yuebin Wang, Hao Tang, and Lorenzo Bruzzone. Looking outside the window: Wide-context transformer for the semantic segmentation of high-resolution remote sensing images. *IEEE TGRS*, 60:1–13, 2022.
46. Yi Ding, Alex Rich, Mason Wang, Noah Stier, Matthew Turk, Pradeep Sen, and Tobias Höllerer. Sparse fusion for multimodal transformers, 2021.
47. David Dov, Ronen Talmon, and Israel Cohen. Kernel-based sensor fusion with application to audio-visual voice activity detection. *IEEE Transactions on Signal Processing*, 64:6406–6416, 2016.
48. Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. Palm-e: An embodied multimodal language model, 2023.
49. Bin Duan, Hao Tang, Wei Wang, Ziliang Zong, Guowei Yang, and Yan Yan. Audio-visual event localization via recursive fusion by joint co-attention. In *WACV*, 2021.
50. Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and et al. The llama 3 herd of models.
51. Andres F. Duque, Guy Wolf, and Kevin R. Moon. Difusion transport alignment, 2022.
52. Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue, 2024.
53. Han et al. Onellm: One framework to align all modalities with language. In *CVPR*, 2024.
54. N. Fei, Z. Lu, Y. Gao, G. Yang, Y. Huo, J. Wen, and H. Lu. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications*, 2022.
55. Valentin Gabeur, Chen Sun, Kartek Alahari, and Cordelia Schmid. Multi-modal transformer for video retrieval. In *ECCV*, pages 214–229, 2020.
56. Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets, 2023.
57. Shijie Geng, Jianbo Yuan, Yu Tian, Yuxiao Chen, and Yongfeng Zhang. Hiclip: Contrastive language-image pretraining with hierarchy-aware attention, 2023.
58. Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and

- Ishan Misra. Imagebind: One embedding space to bind them all, 2023.
59. Muhammad Waleed Gondal, Jochen Gast, Inigo Alonso Ruiz, Richard Droste, Tommaso Macri, Suren Kumar, and Luitpold Staudigl. Domain aligned clip for few-shot classification. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5721–5730, January 2024.
  60. Quentin Grail, Julien Perez, and Eric Gaussier. Globalizing BERT-based transformer architectures for long document summarization. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1792–1810, Online, April 2021. Association for Computational Linguistics.
  61. Chen Guo and Li Zhang. A model-level fusion-based multi-modal object detection and recognition method. In *2023 7th Asian Conference on Artificial Intelligence Technology (ACAIT)*, pages 34–38, 2023.
  62. Andrey Guzhov, Federico Raue, Jörn Hees, and Andreas Dengel. Audioclip: Extending clip to image, text and audio. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 976–980, 2022.
  63. M. Gölo, Marcelo Isaias De Moraes, R. Goularte, and R. Marcacini. On the use of early fusion operators on heterogeneous graph neural networks for one-class learning. In *Proceedings of the 29th Brazilian Symposium on Multimedia and the Web*, 2023.
  64. Wei Han, Hui Chen, and Soujanya Poria. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9180–9192, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.
  65. D. R. Hardoon, S. Szedmak, and J. Shawe-Taylor. Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664, 2004.
  66. Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 3d-llm: Injecting the 3d world into large language models, 2023.
  67. Harold Hotelling. Relations between two sets of variates. *Biometrika*, 28(3-4):321–377, December 1936.
  68. Zhizhang Hu, Xinliang Zhu, Son Tran, René Vidal, and Arnab Dhua. Provla: Compositional image search with progressive vision-language alignment and multimodal fusion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 2772–2777, October 2023.
  69. Saidul Islam, Hanae Elmekki, Ahmed Elsebai, Jamal Bentahar, Najat Drawel, Gaith Rjoub, and Witold Pedrycz. A comprehensive survey on applications of transformers for deep learning tasks.
  70. Andrew Jaegle, Felix Gimeno, Andrew Brock, Andrew Zisserman, Oriol Vinyals, and Joao Carreira. Perceiver: General perception with iterative attention, 2021.
  71. Shashank Jaiswal, Brais Martínez, and Michel F. Valstar. Learning to combine local models for facial action unit detection. *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, 06:1–6, 2015.
  72. Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yunhsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision, 2021.
  73. Ding Jia, Jianyuan Guo, Kai Han, Han Wu, Chao Zhang, Chang Xu, and Xinghao Chen. Geminifusion: Efficient pixel-wise multimodal fusion for vision transformer, 2024.
  74. Yanyun Jiang, Yuanjie Zheng, Sujuan Hou, Yuchou Chang, and J. Gee. Multimodal image alignment via linear mapping between feature modalities. *Journal of Healthcare Engineering*, 2017.
  75. Andrej Karpathy and Fei-Fei Li. Deep visual-semantic alignments for generating image descriptions. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3128–3137, 2015.
  76. Chris Kelly, Luhui Hu, Cindy Yang, Yu Tian, Deshun Yang, Bang Yang, Zaoshan Huang, Zihao Li, and Yuexian Zou. Unifiedvisiongpt: Streamlining vision-oriented ai through generalized multimodal framework, 2023.
  77. Jihyun Kim, Changjae Oh, Hoseok Do, Soohyun Kim, and Kwanghoon Sohn. Diffusion-driven gan inversion for multi-modal face image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10403–10412, June 2024.
  78. Wonjae Kim, Sanghyuk Chun, Taekyung Kim, Dongyoon Han, and Sangdoo Yun. Hype: Hyperbolic entailment filtering for underspecified images and texts. In *European Conference on Computer Vision*, pages 247–265. Springer, 2024.
  79. Wonjae Kim, Bokyung Son, and Ildoo Kim. ViLT: Vision-and-language transformer without convolution or region supervision.
  80. Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything, 2023.
  81. Sander Koelstra, Christian Muhl, Mohammad Soleymani, Jong-Seok Lee, Ashkan Yazdani, Touradj Ebrahimi, Thierry Pun, Anton Nijholt, and Ioannis Patras. Deap: A database for emotion analysis ;using physiological signals. *IEEE Transactions on Affective Computing*, 3(1):18–31, 2012.
  82. M. Kolář, J. Meier, V. Mustonen, et al. Graphalignment: Bayesian pairwise alignment of biological networks. *BMC Syst Biol*, 6:144, 2012.
  83. Zhenglun Kong, Dongkuan Xu, Zhengang Li, Peiyan Dong, Hao Tang, Yanzhi Wang, and Subhabrata Mukherjee. Autovit: Achieving real-time vision transformers on mobile via latency-aware coarse-to-fine search. *Springer IJCV*, pages 1–17, 2025.
  84. Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Fei-Fei Li. Visual genome: Connecting language and vision using crowdsourced dense image annotations, 2016.
  85. Joseph B. Kruskal. An overview of sequence comparison: Time warps, string edits, and macromolecules. *SIAM Rev.*, 25(2):201–237, April 1983.
  86. Hyungyu Lee, Sungjin Park, and E. Choi. Unconditional image-text pair generation with multimodal cross quantizer. *ArXiv*, 2022.
  87. Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. Seed-bench-2: Benchmarking multimodal large language models, 2023.

88. Haoyuan Li, Yanpeng Zhou, Yihan Zeng, Hang Xu, and Xiaodan Liang. Gs-clip: Gaussian splatting for contrastive language-image-3d pretraining from real-world data, 2024.
89. Honglin Li, Yunlong Zhang, Pingyi Chen, Zhongyi Shui, Chenglu Zhu, and Lin Yang. Rethinking transformer for long contextual histopathology whole slide image analysis. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 101498–101528. Curran Associates, Inc., 2024.
90. Huafeng Li, Zengyi Yang, Yafei Zhang, Wei Jia, Zhengtao Yu, and Yu Liu. Mulfs-cap: Multimodal fusion-supervised cross-modality alignment perception for unregistered infrared-visible image fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(5):3673–3690, 2025.
91. Huayang Li, Siheng Li, Deng Cai, Longyue Wang, Lemao Liu, Taro Watanabe, Yujiu Yang, and Shuming Shi. TextBind: Multi-turn interleaved multimodal instruction-following in the wild. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9053–9076, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
92. Hui Li and Xiaojun Wu. Densefuse: A fusion approach to infrared and visible images. *IEEE TIP*, 2018.
93. Jiale Li, Hang Dai, and Yong Ding. Self-distillation for robust LiDAR semantic segmentation in autonomous driving. In *ECCV*, pages 659–676, 2022.
94. Jiale Li, Hang Dai, Hao Han, and Yong Ding. Mseg3d: Multi-modal 3d semantic segmentation for autonomous driving. In *CVPR*, pages 21694–21704, 2023.
95. Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.
96. Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022.
97. Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705, 2021.
98. Wenbing Li, Hang Zhou, Junqing Yu, Zikai Song, and Wei Yang. Coupled mamba: Enhanced multi-modal fusion with coupled state space model, 2024.
99. Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks, 2020.
100. Xuelong Li, Dawei Song, and Yongsheng Dong. Hierarchical feature fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:9165–9175, 2020.
101. Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. Efficient multimodal fusion via interactive prompting. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2604–2613, 2023.
102. Yinheng Li, Han Ding, and Hang Chen. Data processing techniques for modern multimodal models. *arXiv preprint arXiv:2407.19180*, 2024.
103. Zongxia Li, Xiyang Wu, Hongyang Du, Fuxiao Liu, Huy Nghiêm, and Guangyao Shi. A survey of state of the art large vision language models: Alignment, benchmark, evaluations and challenges, 2025.
104. Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. Foundations & trends in multimodal machine learning: Principles, challenges, and open questions. *ACM Comput. Surv.*, 56(10), jun 2024.
105. Victor Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. *Advances in Neural Information Processing Systems*, 35:17612–17625, 2022.
106. Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. Vila: On pre-training for visual language models, 2024.
107. Tianyang Lin, Yuxin Wang, Xiangyang Liu, and Xipeng Qiu. A survey of transformers. *AI Open*, 3:111–132, 2022.
108. Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015.
109. Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end learning of video-based text generation from multimodal inputs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7001–7011, 2021.
110. Yuxiang Lin, Ling Luo, Ying Chen, Xushi Zhang, Zihui Wang, Wenxian Yang, Mengsha Tong, and Rongshan Yu. St-align: A multimodal foundation model for image-gene alignment in spatial transcriptomics, 2024.
111. C. Liu, H. Liu, H. Chen, and W. Du. Touchformer: A transformer-based two-tower architecture for tactile temporal signal classification. *IEEE Transactions on Multimedia*, 2023.
112. Hao Liu, Zehua Chen, Yi Yuan, Xinhao Mei, Kubo Liu, Danilo Mandic, Wenwu Wang, and Mark D. Plumley. Audioldm: Text-to-audio generation with latent diffusion models, 2023.
113. Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.
114. Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 16705–16717. Curran Associates, Inc., 2022.
115. Zhun Liu, Ying Shen, Varun Bharadhwaj Lakshminarasimhan, Paul Pu Liang, AmirAli Bagher Zadeh, and Louis-Philippe Morency. Efficient low-rank multimodal fusion with modality-specific factors. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2247–2256, Melbourne, Australia, July 2018. Association for Computational Linguistics.
116. Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019.
117. Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. Unified-io: A uni-

- fied model for vision, language, and multi-modal tasks, 2022.
118. Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 29615–29627. Curran Associates, Inc., 2023.
  119. Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration, 2023.
  120. Qianxia Ma, Ming Zhang, Yan Tang, and Zhen Huang. Att-sinkhorn: Multimodal alignment with sinkhorn-based deep attention architecture. In *2023 28th International Conference on Automation and Computing (ICAC)*, 2023.
  121. Wenzuan Ma, Shuang Li, Lincan Cai, and Jingxuan Kang. Learning modality knowledge alignment for cross-modality transfer. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 33777–33793. PMLR, 21–27 Jul 2024.
  122. Ziping Ma, Furon Xu, Jian Liu, Ming Yang, and Qingpei Guo. Sycoca: Symmetrizing contrastive captioners with attentive masking for multimodal alignment, 2024.
  123. M. Mahmud, M.S. Kaiser, A. Hussain, and Stefano Vascanelli. Applications of deep learning and reinforcement learning to biological data. *IEEE Transactions on Neural Networks and Learning Systems*, 29:2063–2079, 2017.
  124. Sijie Mai, Haifeng Hu, and Songlong Xing. Divide, conquer and combine: Hierarchical feature fusion network with local and global perspectives for multimodal affective computing. In *Annual Meeting of the Association for Computational Linguistics*, 2019.
  125. Sijie Mai, Haifeng Hu, and Songlong Xing. Modality to modality translation: An adversarial representation learning and graph fusion network for multimodal fusion, 2020.
  126. Alexandros Makris, Dimitrios I. Kosmopoulos, Stavros J. Perantonis, and Sergios Theodoridis. A hierarchical feature fusion framework for adaptive visual tracking. *Image Vis. Comput.*, 29:594–606, 2011.
  127. T. Melzer, M. Reiter, and H. Bischof. Nonlinear feature extraction using generalized canonical correlation analysis. In *Proceedings of the International Conference on Artificial Neural Networks (ICANN)*, 2001.
  128. Oualid Missaoui, Hichem Frigui, and Paul D. Gader. Model level fusion of edge histogram descriptors and gabor wavelets for landmine detection with ground penetrating radar. *2010 IEEE International Geoscience and Remote Sensing Symposium*, pages 3378–3381, 2010.
  129. Veronica Grazia Morelli, Mirko Paolo Barbato, Flavio Piccoli, and Paolo Napoletano. Multimodal fusion methods with vision transformers for remote sensing semantic segmentation. In *2023 13th Workshop on Hyperspectral Imaging and Signal Processing: Evolution in Remote Sensing (WHISPERS)*, pages 1–5, 2023.
  130. Emilie Morvant, Amaury Habrard, and Stéphane Ayache. Majority vote of diverse classifiers for late fusion. In *Structural, Syntactic, and Statistical Pattern Recognition*, pages 153–162. Springer, 2014.
  131. Alberto Muñoz and Javier González. Functional learning of kernels for information fusion purposes. In *Iberoamerican Congress on Pattern Recognition*, 2008.
  132. Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion, 2022.
  133. Huda Nassar and David Gleich. Multimodal network alignment. *ArXiv*, 2017.
  134. Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Se-woong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning, 2023.
  135. Jianyuan Ni, Hao Tang, Syed Tousiful Haque, Yan Yan, and Anne HH Ngu. A survey on multimodal wearable sensor-based human action recognition. *arXiv preprint arXiv:2404.15349*, 2024.
  136. Meng Ning, Fan Zhou, Wei Wang, Shaoqiang Wang, Peiying Zhang, and Jian Wang. Abftnet: An efficient transformer network with alignment before fusion for multimodal automatic modulation recognition. *Electronics*, 13(18), 2024.
  137. Fatemeh Noroozi, Marina Marjanovic, Angelina Njegus, Sergio Escalera, and Gholamreza Anbarjafari. Audio-visual emotion recognition in video clips. *IEEE Transactions on Affective Computing*, 10(1):60–75, 2019.
  138. OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, and et al. GPT-4 technical report.
  139. Maxime Oquab, Timothée Darcret, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, and et al. Dinov2: Learning robust visual features without supervision, 2024.
  140. Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2text: Describing images using 1 million captioned photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011.
  141. Sanjeel Parekh, Slim Essid, Alexey Ozerov, Ngoc Q. K. Duong, Patrick Pérez, and Gaël Richard. Weakly supervised representation learning for audio-visual scene analysis. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:416–428, 2020.
  142. Priyaranjan Pattnayak, Hitesh Laxmichand Patel, Bhargava Kumar, Amit Agarwal, Ishan Banerjee, Srikant Panda, and Tejaswini Kumar. Survey of large multimodal model datasets, application categories and taxonomy, 2024.
  143. Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shao-han Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world, 2023.
  144. Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models, 2016.
  145. Ye Qian and Li Pan. Leveraging multimodal features for knowledge graph entity alignment based on dynamic self-attention networks. *Expert Systems with Applications*, 228:120363, 2023.
  146. Jiahao Qin, Yitao Xu, Zihong Luo, Chengzhi Liu, Zong Lu, and Xiaojun Zhang. Alternative telescopic displacement: An efficient multimodal alignment method. *ArXiv*, 2023.
  147. Longtian Qiu, Renrui Zhang, Ziyu Guo, Ziyao Zeng, Zilu Guo, Yafeng Li, and Guangnan Zhang. Vt-clip: Enhancing vision-language models with visual-guided texts, 2023.
  148. Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen

- Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, pages 8748–8763, 2021.
149. Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2019.
150. Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion*, 81:203–239, May 2022.
151. Attila Reiss and Didier Stricker. Introducing a new benchmarked dataset for activity monitoring. In *2012 16th International Symposium on Wearable Computers*, pages 108–109, 2012.
152. Alexis Roger, Esma Aïmeur, and Irina Rish. Towards ethical multimodal systems, 2024.
153. András Rövid and Viktor Remeli. Towards raw sensor fusion in 3d object detection. *2019 IEEE 17th World Symposium on Applied Machine Intelligence and Informatics (SAMI)*, pages 293–298, 2019.
154. Paul K. Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quirly, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, Hannah Muckenhirm, Dirk Padfield, James Qin, Danny Rozenberg, Tara Sainath, Johan Schalkwyk, Matt Sharifi, Michelle Tadmor Ramanovich, Marco Tagliasacchi, Alexandru Tudor, Mihajlo Velimirović, Damien Vincent, Jiahui Yu, Yongqiang Wang, Vicky Zayats, Neil Zeghidour, Yu Zhang, Zhishuai Zhang, Lukas Zilka, and Christian Frank. Audiopalm: A large language model that can speak and listen, 2023.
155. Md Imran Sarker and M Milanova. Deep learning-based multimodal image retrieval combining image and text. In *2022 International Conference on Computational Science and Computational Intelligence (CSCI)*, pages 1543–1546, 2022.
156. F. Scalzo, George Bebis, Mircea Nicolescu, Leandro A. Loss, and A. Tavakkoli. Feature fusion hierarchies for gender classification. *2008 19th International Conference on Pattern Recognition*, pages 1–4, 2008.
157. Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022.
158. Shiv Shankar, Laure Thompson, and Madalina Fiterau. Progressive fusion for multimodal integration, 2022.
159. Sheng Shen, Zhewei Yao, Chunyuan Li, Trevor Darrell, Kurt Keutzer, and Yuxiong He. Scaling vision-language models with sparse mixture of experts, 2023.
160. Daqian Shi, Xiaolei Diao, Lida Shi, Hao Tang, Yang Chi, Chuntao Li, and Hao Xu. Charformer: A glyph fusion based attentive framework for high-precision character image denoising. In *ACM MM*, 2022.
161. Gen Shi, Yifan Zhu, Wenjin Liu, Quanming Yao, and Xuesong Li. Heterogeneous graph-based multimodal brain network learning, 2022.
162. Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. Flava: A foundational language and vision alignment model, 2022.
163. Cees G. M. Snoek, Marcel Worring, and Arnold W. M. Smeulders. Early versus late fusion in semantic video analysis. *ACM MM*, 2005.
164. Yiming Song, Zhen Li, and Wei Song. Scene-driven multimodal knowledge graph construction for embodied ai. *IEEE Transactions on Robotics*, 39(1):45–60, 2023.
165. Zijia Song, Zelin Zang, Yelin Wang, Guozheng Yang, Kaicheng Yu, Wanyu Chen, Miaoyu Wang, and Stan Z. Li. Set-clip: Exploring aligned semantic from low-alignment multimodal data through a distribution view, 2024.
166. Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR ’21. ACM, July 2021.
167. Nitish Srivastava and Ruslan Salakhutdinov. Multimodal learning with deep boltzmann machines. *Journal of Machine Learning Research*, 15:2949–2980, 2012.
168. Lukas Stappen, Alice Baird, Lea Schumann, and Björn Schuller. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements, 2021.
169. Josef Steinbaeck, Christian Steger, Gerald Holweg, and Norbert Druml. Design of a low-level radar and time-of-flight sensor fusion framework. *2018 21st Euromicro Conference on Digital System Design (DSD)*, pages 268–275, 2018.
170. L. Su, F. Yan, J. Zhu, X. Xiao, and H. Duan. Beyond two-tower matching: Learning sparse retrievable cross-interactions for recommendation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2023.
171. Hao Tang, Hong Liu, Wei Xiao, and Nicu Sebe. Fast and robust dynamic hand gesture recognition via key frames extraction and feature fusion. *Elsevier Neurocomputing*, 331:424–433, 2019.
172. Hao Tang, Hong Liu, Dan Xu, Philip HS Torr, and Nicu Sebe. Attentiongan: Unpaired image-to-image translation using attention-guided generative adversarial networks. *IEEE TNNLS*, 34(4):1972–1987, 2021.
173. Hao Tang, Ling Shao, Nicu Sebe, and Luc Van Gool. Graph transformer gans with graph masked modeling for architectural layout generation. *IEEE TPAMI*, 46(6):4298–4313, 2024.
174. Hao Tang, Ling Shao, Nicu Sebe, and Luc Van Gool. Enhanced multi-scale cross-attention for person image generation. *IEEE TPAMI*, 2025.
175. Hao Tang, Dan Xu, Nicu Sebe, Yanzhi Wang, Jason J Corso, and Yan Yan. Multi-channel attention selection gan with cascaded semantic guidance for cross-view image translation. In *CVPR*, 2019.
176. Hao Tang, Dan Xu, Yan Yan, Philip HS Torr, and Nicu Sebe. Local class-specific and global image-level generative adversarial networks for semantic-guided scene generation. In *CVPR*, 2020.
177. Jiajia Tang, Dongjun Liu, Xuanyu Jin, Yong Peng, Qianchuan Zhao, Yu Ding, and Wanpeng Kong. Bafn: Bi-direction attention based fusion network for multimodal sentiment analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:1966–1978, 2023.
178. Shengeng Tang, Dixin Guo, Rui Hong, and Min Wang. Graph-based multimodal sequential embedding for sign language translation. *IEEE Transactions on Multimedia*, 23:1056–1067, 2021.
179. Wei Tang, Fazhi He, Yefeng Liu, and Ying Duan. Matr: Multimodal medical image fusion via multiscale adaptive transformer. *IEEE Transactions on Image Processing*, 31:5134–5149, 2022.

180. Ming Tao, Hao Tang, Fei Wu, Xiao-Yuan Jing, Bing-Kun Bao, and Changsheng Xu. Df-gan: A simple and effective baseline for text-to-image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 16515–16525, 2022.
181. Mani Kumar Tellamekala, Shahin Amiriparian, Björn W. Schuller, Elisabeth André, Timo Giesbrecht, and Michel Valstar. COLD fusion: Calibrated and ordinal latent distribution fusion for uncertainty-aware multimodal emotion recognition. *IEEE TPAMI*, 46(2):805–822, 2024.
182. T. M. Thai, A. T. Vo, Hao K. Tieu, Linh Bui, and T. Nguyen. Uit-saviors at medvqa-gi 2023: Improving multimodal learning with image enhancement for gastrointestinal visual question answering, 2023.
183. Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: the new data in multimedia research. *Communications of the ACM*, 59(2):64–73, January 2016.
184. Haiman Tian, Yudong Tao, Samira Pouyanfar, Shu-Ching Chen, and Mei-Ling Shyu. Multimodal deep representation learning for video classification. *World Wide Web*, 22:1325–1341, 2019.
185. Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9568–9578, June 2024.
186. Tong Tong, Katherine R. Gray, Qinquan Gao, Liang Chen, and Daniel Rueckert. Nonlinear graph fusion for multi-modal classification of alzheimer’s disease. In *Machine Learning for Multimodal Interaction*, 2015.
187. Tong Tong, Katherine R. Gray, Qinquan Gao, Liang Chen, and Daniel Rueckert. Multi-modal classification of alzheimer’s disease using nonlinear graph fusion. *Pattern Recognit.*, 63:171–181, 2017.
188. Yao-Hung Hubert Tsai, Paul Pu Liang, Amir Zadeh, Louis-Philippe Morency, and Ruslan Salakhutdinov. Learning factorized multimodal representations, 2019.
189. J. Tu, X. Liu, R. Lin, Z. and Hong, and M. Wang. Differentiable cross-modal hashing via multimodal transformers. In *Proceedings of the 30th ACM International Conference on Multimedia*, 2022.
190. Tatsumi Uezato, Danfeng Hong, Naoto Yokoya, and Wei He. Guided deep decoder: Unsupervised image pair fusion. In *ECCV*, 2020.
191. Unknown. Dynamic time warping. In *Information Retrieval for Music and Motion*. Springer, Berlin, Heidelberg, 2007.
192. Maya Varma, Jean-Benoit Delbrouck, Sarah Hooper, Akshay Chaudhari, and Curtis Langlotz. Villa: Fine-grained vision-language representation learning from real-world data, 2023.
193. Yannis Vasilakis, Rachel Bittner, and Johan Pauwels. I can listen but cannot read: An evaluation of two-tower multimodal systems for instrument recognition, 2024.
194. Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need.
195. Y. Verma and C. V. Jawahar. Im2text and text2im: Associating images and texts for cross-modal retrieval. In *Proceedings of the British Machine Vision Conference (BMVC)*, page 2, 2014.
196. Vitalis Vosylius and Edward Johns. Few-shot in-context imitation learning via implicit graph alignment. In *Proceedings of The 7th Conference on Robot Learning*, volume 229 of *Proceedings of Machine Learning Research*, pages 3194–3213, 2023.
197. Noël Vouitsis, Zhaoyan Liu, Satya Krishna Gorti, Valentin Villecroze, Jesse C. Cresswell, Guangwei Yu, Gabriel Loaiza-Ganem, and Maksims Volkovs. Data-efficient multimodal fusion on a single gpu, 2024.
198. Yongquan Wan, Wenhui Wang, Guobing Zou, and Bofeng Zhang. Cross-modal feature alignment and fusion for composed image retrieval. In *CVPRW*, pages 8384–8388, 2024.
199. Alex Jinpeng Wang, Kevin Qinghong Lin, David Junhao Zhang, Stan Weixian Lei, and Mike Zheng Shou. Too large; data reduction for vision-language pre-training, 2023.
200. Jinping Wang and Xiaojun Tan. Mutually beneficial transformer for multimodal data fusion. *IEEE Transactions on Circuits and Systems for Video Technology*, 33:7466–7479, 2023.
201. Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhi-hao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution, 2024.
202. Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. One-peace: Exploring one general representation model toward unlimited modalities, 2023.
203. Weihua Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. CogVLM: Visual expert for pretrained language models.
204. Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhajit Som, and Furu Wei. Image as a foreign language: BEiT pretraining for all vision and vision-language tasks.
205. Xijun Wang, Junbang Liang, Chun-Kai Wang, Kenan Deng, Yu Lou, Ming Lin, and Shan Yang. Vila: Efficient video-language alignment for video question answering, 2024.
206. Xingyue Wang, Kuang Shu, Haowei Kuang, Shiwei Luo, Richu Jin, and Jiang Liu. The role of spatial alignment in multimodal medical image fusion using deep learning for diagnostic problems. In *Proceedings of the 2021 International Conference on Intelligent Medicine and Health, ICIMH ’21*, page 40–46, New York, NY, USA, 2022. Association for Computing Machinery.
207. Yong-Cui Wang, Chunhua Zhang, Naiyang Deng, and Yong Wang. Kernel-based data fusion improves the drug-protein interaction prediction. *Computational biology and chemistry*, 35 6:353–62, 2011.
208. Yongjin Wang, Ling Guan, and Anastasios N. Venetsanopoulos. Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition. *IEEE Transactions on Multimedia*, 14:597–607, 2012.
209. Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. Simvlm: Simple visual language model pretraining with weak supervision, 2022.
210. Yupeng Wei, Dazhong Wu, and Janis P. Terpenny. Decision-level data fusion in quality control and predictive maintenance. *IEEE Transactions on Automation Science and Engineering*, 18:184–194, 2021.

211. Haoyang Wen, Honglei Zhuang, Hamed Zamani, Alexander Hauptmann, and Michael Bendersky. Multimodal reranking for knowledge-intensive visual question answering, 2024.
212. Fei Wu, Yongheng Ma, Hao Jin, Xiao-Yuan Jing, and Guo-Ping Jiang. Mfeclip: Clip with mapping-fusion embedding for text-guided image editing. *IEEE Signal Processing Letters*, 31:116–120, 2024.
213. Mike Wu and Noah Goodman. Multimodal generative models for compositional representation learning, 2019.
214. Qinhua Xie and Hao Tang. TTTFusion: A Test-Time Training-Based Strategy for Multimodal Medical Image Fusion in Surgical Robots. In *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2025.
215. Wei Xiong, Yifan Zhang, and Wei Li. Scene graph-based semantic alignment for multimodal tasks. *IEEE Transactions on Multimedia*, 22(5):1231–1243, 2020.
216. Yu Xiong, Daling Wang, Yifei Zhang, Shi Feng, and Guoren Wang. Multimodal data fusion in text-image heterogeneous graph for social media recommendation. In *International Conference on Neural Information Processing*, pages 96–99, 2014.
217. Jin Xu, Zhifang Guo, Jincheng He, Hangrui Hu, Ting He, Shuai Bai, Kegui Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report, 2025.
218. X. Xu, C. Wu, S. Rosenman, V. Lal, and W. Che. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2023.
219. Zihui Xue and Radu Marculescu. Dynamic multimodal fusion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 2575–2584, June 2023.
220. Shen Yan, Di Huang, and Mohammad Soleymani. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 361–369, 2020.
221. Bing Yang, Xueqin Xiang, Wangzeng Kong, Jianhai Zhang, and Yong Peng. Dmf-gan: Deep multimodal fusion generative adversarial networks for text-to-image synthesis. *IEEE Transactions on Multimedia*, 26:6956–6967, 2024.
222. Guanglei Yang, Enrico Fini, Dan Xu, Paolo Rota, Mingli Ding, Hao Tang, Xavier Alameda-Pineda, and Elisa Ricci. Continual attentive fusion for incremental learning in semantic segmentation. *IEEE TMM*, 25:3841–3854, 2022.
223. H. Yang and S. Li. Videochat: Conversational agents in video understanding. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.
224. Haoyan Yang, Yifan Wu, Zhenyu Si, Yijun Zhao, Jinfeng Liu, and Bing Qin. Macsa: A multimodal aspect-category sentiment analysis dataset with multimodal fine-grained aligned annotations. *Multimedia Tools and Applications*, 82:3839–3858, 2023.
225. Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models.
226. Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, Kaidong Yu, Peng Liu, Qiang Liu, Shawn Yue, Senbin Yang, Shiming Yang, Tao Yu, Wen Xie, Wenhao Huang, Xiaohui Hu, Xiaoyi Ren, Xinyao Niu, Pengcheng Nie, Yuchi Xu, Yudong Liu, Yue Wang, Yuxuan Cai, Zhenyu Gu, Zhiyuan Liu, and Zonghong Dai. Yi: Open foundation models by 01.ai, 2024.
227. Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. CoCa: Contrastive captioners are image-text foundation models.
228. Pinrui Yu, Zhenglun Kong, Pu Zhao, Peiyan Dong, Hao Tang, Fei Sun, Xue Lin, and Yanzhi Wang. Q-tempfusion: Quantization-aware temporal multi-sensor fusion on bird’s-eye view representation. In *2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 5489–5499, 2025.
229. Qiyi Yu, Quan Sun, Xiaosong Zhang, Yufeng Cui, Fan Zhang, Yue Cao, Xinlong Wang, and Jingjing Liu. Capsfusion: Rethinking image-text data at scale, 2024.
230. Wenmeng Yu, Hua Xu, Fanyang Meng, Yilin Zhu, Yixiao Ma, Jiele Wu, Jiyun Zou, and Kaicheng Yang. CH-SIMS: A Chinese multimodal sentiment analysis dataset with fine-grained annotation of modality. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetraeault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3718–3727, Online, July 2020. Association for Computational Linguistics.
231. S. Yuan, P. Bhatia, B. Celikkaya, H. Liu, and K. Choi. Towards user friendly medication mapping using entity-boosted two-tower neural network. In *International Workshop on Deep Learning for Human Activity Recognition*, 2021.
232. Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023.
233. Amir Zadeh, Minghai Chen, Soujanya Poria, Erik Cambria, and Louis-Philippe Morency. Tensor fusion network for multimodal sentiment analysis. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1103–1114, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.
234. Yuhang Zang, Wei Li, Jun Han, Kaiyang Zhou, and Chen Change Loy. Contextual object detection with multimodal large language models, 2024.
235. Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training, 2023.
236. Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14:478–493, 2020.
237. Duzhen Zhang, Yahan Yu, Jiahua Dong, Chenxing Li, Dan Su, Chenhui Chu, and Dong Yu. MM-LLMs: Recent advances in MultiModal large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12401–12430, Bangkok, Thailand, August 2024. Association for Computational Linguistics.
238. Hao Zhang, Feng Li, Shilong Liu, Lei Zhang, Hang Su, Jun Zhu, Lionel M. Ni, and Heung-Yeung Shum. Dino: Detr with improved denoising anchor boxes for end-to-end object detection, 2022.
239. Jing Zhang, Aiping Liu, Dan Wang, Yu Liu, Z. Jane Wang, and Xun Chen. Transformer-based end-to-end anatomical and functional image fusion. *IEEE Transactions on Instrumentation and Measurement*, 71:1–11, 2022.

240. Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024.
241. Muhan Zhang. Neural attention: Enhancing qkv calculation in self-attention mechanism with neural networks, 2023.
242. Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Haodong Duan, Songyang Zhang, Shuangrui Ding, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition, 2023.
243. Pan Zhang, Xiaoyi Dong, Yuhang Zang, Yuhang Cao, Rui Qian, Lin Chen, Qipeng Guo, Haodong Duan, Bin Wang, Linke Ouyang, Songyang Zhang, Wenwei Zhang, Yining Li, Yang Gao, Peng Sun, Xinyue Zhang, Wei Li, Jingwen Li, Wenhui Wang, Hang Yan, Conghui He, Xingcheng Zhang, Kai Chen, Jifeng Dai, Yu Qiao, Dahua Lin, and Jiaqi Wang. Internlm-xcomposer-2.5: A versatile large vision language model supporting long-contextual input and output, 2024.
244. Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models, 2021.
245. Rui Zhang, Chengrong Xue, Qingfu Qi, Liyuan Lin, Jing Zhang, and Lun Zhang. Bimodal fusion network with multi-head attention for multimodal sentiment analysis. *Applied Sciences*, 13(3), 2023.
246. Tao Zhang, Ziqian Zeng, Yuxiang Xiao, Huiping Zhuang, Cen Chen, James Foulds, and Shimei Pan. Genderalign: An alignment dataset for mitigating gender bias in large language models, 2024.
247. Weifeng Zhang, Jing Yu, Yuxia Wang, and Wei Wang. Multimodal deep fusion for image question answering. *Knowledge-Based Systems*, 212:106639, 2021.
248. Xiaofeng Zhang, Chen Shen, Xiaosong Yuan, Shaotian Yan, Liang Xie, Wenxiao Wang, Chaochen Gu, Hao Tang, and Jieping Ye. From redundancy to relevance: Enhancing explainability in multimodal large language models. *arXiv preprint arXiv:2406.06579*, 2024.
249. Xiaofeng Zhang, Zishan Xu, Hao Tang, Chaochen Gu, Wei Chen, Shanying Zhu, and Xinping Guan. Enlighten-your-voice: When multimodal meets zero-shot low-light image enhancement. *arXiv:2312.10109*, 2023.
250. Yedi Zhang, Peter E. Latham, and Andrew Saxe. Understanding unimodal bias in multimodal deep linear networks, 2024.
251. Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data, 2025.
252. Zhiwei Zhang, Wenyu Mai, Heng Xiong, and Cheng Wu. A token-wise graph-based framework for multimodal named entity recognition. *IEEE Transactions on Knowledge and Data Engineering*, 33(10):3121–3134, 2021.
253. Zhiyuan Zhang, Licheng Yang, and Zhiyu Xiang. Risurconv: Rotation invariant surface attention-augmented convolutions for 3d point cloud classification and segmentation, 2024.
254. Zilun Zhang, Tiancheng Zhao, Yulong Guo, and Jianwei Yin. Rs5m and georsclip: A large-scale vision- language dataset and a large vision-language model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*, 62:1–23, 2024.
255. Fei Zhao, Chengcui Zhang, and Baocheng Geng. Deep multimodal data fusion. *ACM Computing Surveys*, 56(9):1–36, 2024.
256. Lihong Zhao and Huan Wang. Deep multimodal learning with vision, audio, and text: Challenges and innovations. *IEEE Transactions on Neural Networks and Learning Systems*, 35:1172–1184, 2024.
257. Tao Zhou, Jiuxin Cao, Xueling Zhu, Bo Liu, and Shancang Li. Visual-textual sentiment analysis enhanced by hierarchical cross-modality interaction. *IEEE Systems Journal*, 15:4303–4314, 2021.
258. Yutong Zhou and Nobutaka Shimada. Vision + language applications: A survey. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 826–842, June 2023.
259. Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.
260. Rui Zuo, Guoqing Li, Bongshin Choi, Sourav Bhowmick, Daphne N. yin Mah, and Gary L. Wong. Svp-t: A shape-level variable-position transformer for multivariate time series classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 11497–11505, 2023.
261. Müjdat Çetin, Lei Chen, John W. Fisher III, Alexander T. Ihler, Randolph L. Moses, Martin J. Wainwright, and Alan S. Willsky. Distributed fusion in sensor networks: a graphical models perspective. 2006.