Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications

MUHAMMAD ARSLAN MANZOOR and SARAH ALBARRI, Mohamed bin Zayed University of Artificial Intelligence, UAE
ZITING XIAN, Sun Yat-sen University, China
ZAIQIAO MENG, University of Glasgow, UK
PRESLAV NAKOV and SHANGSONG LIANG, Mohamed bin Zayed University of Artificial Intelligence, UAE

Multimodality Representation Learning, as a technique of learning to embed information from different modalities and their correlations, has achieved remarkable success on a variety of applications, such as Visual Question Answering (VQA), Natural Language for Visual Reasoning (NLVR), and Vision Language Retrieval (VLR). Among these applications, cross-modal interaction and complementary information from different modalities are crucial for advanced models to perform any multimodal task, e.g., understand, recognize, retrieve, or generate optimally. Researchers have proposed diverse methods to address these tasks. The different variants of transformer-based architectures performed extraordinarily on multiple modalities. This survey presents the comprehensive literature on the evolution and enhancement of deep learning multimodal architectures to deal with textual, visual and audio features for diverse cross-modal and modern multimodal tasks. This study summarizes the (*i*) recent task-specific deep learning methodologies, (*ii*) the pretraining types and multimodal pretraining objectives, (*iii*) from state-of-the-art pretrained multimodal approaches to unifying architectures, and (*iv*) multimodal task categories and possible future improvements that can be devised for better multimodal learning. Moreover, we prepare a dataset section for new researchers that covers most of the benchmarks for pretraining and finetuning. Finally, major challenges, gaps, and potential research topics are explored. A constantly-updated paperlist related to our survey is maintained at https://github.com/marslanm/multimodality-representation-learning.

CCS Concepts: • **Computing methodologies** → **Natural language processing**; **Neural networks;**

Additional Key Words and Phrases: Multimodality, representation learning, pretrained models, multimodal methods, multimodal applications

Authors' addresses: M. A. Manzoor, S. AlBarri, P. Nakov, and S. Liang (Corresponding author), Mohamed bin Zayed University of Artificial Intelligence, Building 1B, Masdar City, Abu Dhabi, P. O. Box 7909, United Arab Emirates; e-mails: muhammad.arslan@mbzuai.ac.ae, sarah.albarri@mbzuai.ac.ae, Preslav.nakov@mbzuai.ac.ae, Shangsong.liang@mbzuai.ac.ae; Z. Xian, Sun Yat-sen University, Univ Town Waihuan E Rd 132, Guangzhou, 510006, China; e-mail: xianzt@mail2.sysu.edu.cn; Z. Meng, University of Glasgow, Glasgow G12 8QQ, United Kingdom; e-mail: zaiqiao.meng@glasgow.ac.uk.

74

1 INTRODUCTION

Multimodal systems utilize two or more input modalities, such as audio, text, images, or video to produce an output modality, which could be different from the inputs. Cross-modal systems, a sub-part of multimodal systems, utilize information from one modality to enhance the performance in the other modality. For example, a multimodal system would use image and textual modalities to assess a situation and perform a task, whereas a cross-modal system would use an image modality to output a textual modality [1, 2]. **Audio-Visual Speech Recognition (AVSR)** [3], detecting propaganda in memes [4], and **Visual Question Answering (VQA)** [5] are examples of the multimodal systems. Multimodal representation learning techniques reduce the heterogeneity gap between different modalities by processing raw heterogeneous data hierarchically [6]. Heterogeneous features from different modalities offer additional semantics in the form of contextual information [6]. Thus, complementary information can be learned through multiple modalities. For example, the visual modality can help speech recognition by providing lip motion [7] in the AVSR. Recent advanced variants of deep learning approaches have addressed classical multimodal challenges (correlation, translation, alignment, fusion) by mapping different modalities into a standard representation space.

In the past, many task-specific deep learning strategies have significantly enhanced the performance for a variety of multimodal tasks [8]. More recently, pre-training and fine-tuning methods for **Natural Language Processing (NLP)** and **Computer Vision (CV)** have garnered significant attention, primarily due to their semantically rich representation and the availability of large-scale public models [9]. This study summarized the pipeline of the task-specific methods for multimodal representation. We discuss pretraining types and pretext tasks required to make the pretrained model robust on the variety of multimodal or cross-modal downstream tasks. We demonstrate that a majority of pretraining methods leverage transformers, leading to the inception of unifying architectures. These models, capable of handling all modalities across diverse downstream tasks, mitigate the need for individual task-specific fine-tuning, thereby reducing both computational complexity and processing time [10]. Additionally, evaluation on downstream tasks, multimodal benchmarks, extensive range of multimodal applications, including NLP tasks that are enriched by the incorporation of visual and audio modalities, e.g., sentiment analysis, document understanding, fake news detection, retrieval, translation, and other reasoning applications are comprehensively presented in this survey. Figure 1 exhibits the categorical percentage of deep learning multimodal papers included in this survey. The bar graph shows the development and the availability of deep learning multimodal approaches on Google Scholar yearly.

2 BACKGROUND - THE ADVANCE EVOLUTION OF MULTIMODAL ARCHITECTURES

The concept of multimodal research was initially inspired by the field of audio-visual speech recognition (AVSR) [11], where well-aligned integration of vision and hearing was a fundamental requirement. Encouraging outcomes from AVSR studies prompted the research community to broaden the application of this approach to other multimodal tasks, e.g., content indexing and retrieval [12], video summarization [13], and shot-boundary detection [14]. In the early 2000s, the AMI Meeting [15] and SEMAINE [16] corpora were generated with the intention of understanding human multimodal behaviors and the dynamics of interpersonal interactions. These foundational datasets not only led to the inception of the **Audio-Visual Emotion Challenge (AVEC)** [17], but also later propelled its expansion into the realm of automatic detection of depression and anxiety [18]. Over time, multimodal methodologies have advanced, enabling simultaneous processing of diverse modalities such as speech, vision, and text [2, 19]. The latest category
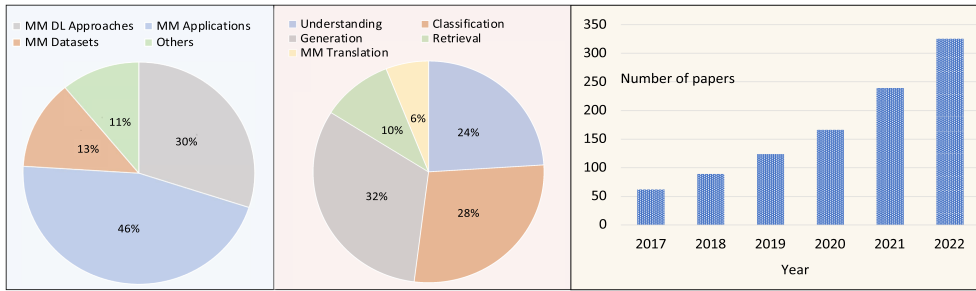
Fig. 1. The pie chart on the left represents the percentage of papers for each section included in this survey. The pie chart in the center represents the percentage of papers for each multimodal application. The rightmost figure expresses the growth of deep learning-based multimodal papers in the last six years on Google Scholar.

encompasses data derived from multimedia sources that extend beyond the complexity of two or three modalities. These complex multimodal approaches motivate researchers to tackle associated challenges through more effective strategies such as pretraining and fine-tuning. These methods have been found to deliver superior performance compared to traditional task-specific frameworks. Advanced approaches are particularly adept at tackling complex multimodal tasks, exhibiting superior performance to previous models.

The primary obstacle for multimodal methods lies in understanding the relationships between various encoding schemes of multiple modalities, a problem commonly identified as the heterogeneity gap in learning representation [20]. The fusion module in multimodal learning systems aids in reducing the heterogeneity gap by correlating the similar semantics across different modalities, a strategy that has been proven to boost performance in a majority of tasks [21]. Recent research has demonstrated that deep learning methods effectively understand and process representations due to their robust learning capabilities [22]. One of the key advantages of deep learning is its ability to hierarchically learn representations as general-purpose learning without needing additional structural information. These factors contribute to making deep learning an extremely compatible approach for multimodal representation learning.

The early task-specific multimodal methods were trained to address only one multimodal task. Figure 2(a) broadly depicts the pipeline of task-specific multimodal methods. For example, in VQA, a visual encoder, usually based on some variant of a **Convolutional Neural Network (CNN)** [23, 24], is used to extract visual features. Meanwhile, textual features are encoded by text encoding blocks that typically build on Bag-of-Words [25], seq2seq models [26], possibly with attention mechanism [27]. Subsequently, the encoded representations are channeled to Multimodal Fusion (explored extensively for multimodal learning), which fuses and builds interactions between features of different modalities to produce a unified representation [28]. These Multimodal Fusion methods have evolved as (*i*) *simple fusion*, which fuses features through concatenation or element-wise sum or product [29], (*ii*) *inter-modality attention*, which aligns features of different modalities and constructs more informative joint representations [30], (*iii*) *intra-modality attention*, which develops relational reasoning by forming a graphical representation within modalities [31], and (*iv*) *transformers*, which leverage the aforementioned techniques by concurrently focusing on other modalities and associated regions within the same modality [32]. Since transformers learn and establish intricate interactions between different modalities [33, 34], they are viewed as the structural foundational block of universal multimodal pretraining methods.

With the availability of large-scale datasets and computing resources, the pretraining paradigm has yielded efficient performance in various domains, including text [35], vision [36], and
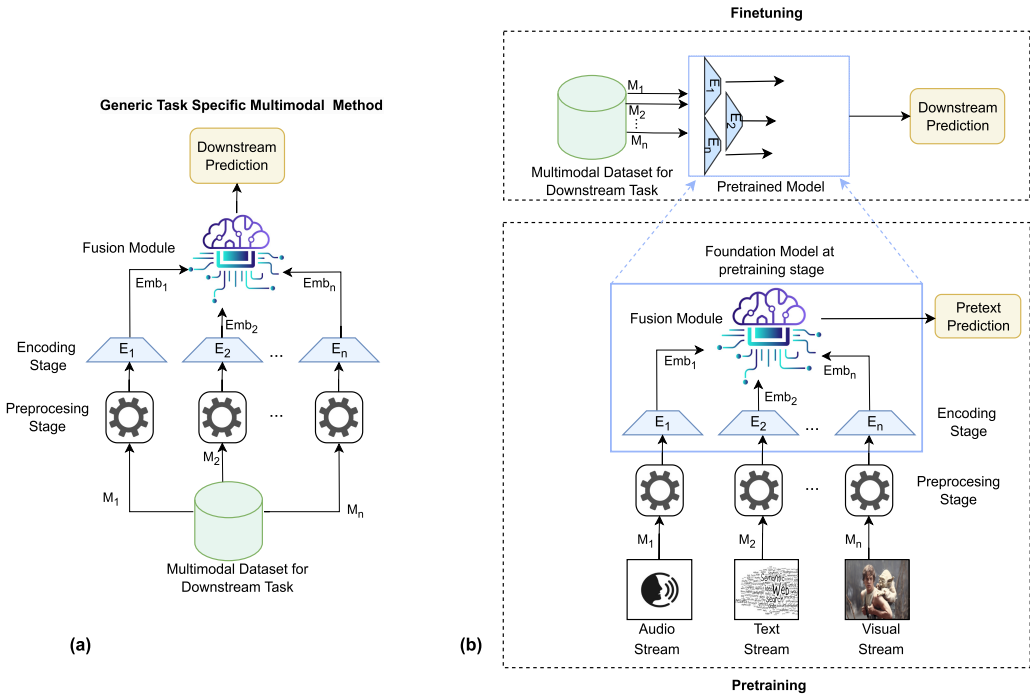
Fig. 2. (a) Task-specific multimodal methods are trained on benchmarks created for specialized downstream tasks. After preprocessing, each encoder (E_i) receives one modality (M_i) data to produce an embedding. The fusion module is responsible for the interaction of features from different modalities and is trained to predict the downstream task. (b) In pretraining, the raw data from the web or dataset, in the form of any modality (M_i), is preprocessed and passed to a specialized encoder (M_i) at the encoding stage. Encoders produce embeddings which are integrated by the fusion module to produce a meaningful unified representation by predicting pretext tasks. The blue-bounded box in the pretraining stage represents the foundation model, which is readily available at the fine-tuning stage as a pretra ined model. A multimodal dataset for the downstream task can leverage this generic pretrained model for downstream prediction.

speech [37]. This has spurred interest in multimodal pretraining, allowing for the processing of raw data (both labeled and unlabeled) using semi-supervised, self-supervised, or unsupervised approaches. As depicted in Figure 2(b), each modality is processed by a dedicated encoder to generate embeddings. These embeddings are then passed to a fusion module specifically trained for multimodal pretext tasks. The generation of contextually rich embeddings through dedicated encoders allows the fusion module to learn more robust associations between different modalities. This, in turn, can lead to more accurate performance in downstream tasks, as the model becomes adept at handling complex multimodal data. In the fine-tuning phase of multimodal learning, the pre-trained model, serving as a flexible backbone, is further trained using a specific multimodal dataset for a target task. Through fine-tuning, the model adjusts its learned representations to better accommodate the relationships and interactions between different modalities present in the task-specific dataset. Consequently, the rich, general-purpose knowledge from pretraining is transformed into task-specific insights, improving the model's performance on the downstream task. Figure 2 clearly illustrates two pivotal approaches in multimodal learning: task-specific methods and the pretraining-fine-tuning paradigm. They are primarily distinguished by the type of input

data: task-specific methods leverage specialized benchmarks, while pretraining can employ any kind of data. Task-specific methods are trained directly for the downstream task, whereas pretraining is for a pretext task and yields a generic pretrained model that can be adapted to a variety of downstream tasks, thus saving computational time and complexity associated with individually training each model.

2.1 Comparison with Previous Surveys

Numerous surveys have extensively investigated multimodality, with a specific focus on multimodal learning techniques, pretraining technologies, model-specific topologies, and task-specific applications. Representation, correlation, fusion, translation and co-learning are essential components of multimodal learning that are surveyed in [2, 6, 7, 38, 39]. In [2], the taxonomy of multimodal learning components is established. Additionally, [7] contributes by offering a perspective on these components from the domain of computer vision. The work of [6] and [39] solely examined representation and fusion, respectively, presenting their mathematical frameworks, architectures, challenges, and prospects. Despite the importance of multimodal learning components in understanding and processing different modalities, these surveys are focused and fail to present a holistic depiction of multimodality.

Vision-language pretraining objectives, strategies, architectures and datasets are extensively studied in [28, 40–42]. The multimodal pretraining survey [41] focused on image-text tasks as it reviews methodologies of encoding of raw image and text, and architectures for the interactions. While downstream tasks are addressed in all pretraining surveys, the level of detail regarding state-of-the-art frameworks is limited, and there is insufficient coverage of diverse applications within the multimodal domain. The studies [1, 43], and [44] inspected multimodal deep learning models, which are presented with respect to modalities involved in [1] and to prominent tasks in [44]. The work of [43] presented methodologies for finding optimal multimodal architectures and cross-modality regularization. Various works assessed multimodal learning from a specific application perspective, such as visual question answering [45], biomedical applications [46], event detection [47], sentiment analysis [48, 49], and object detection [50].

Our work expands on existing research, providing a more comprehensive exploration of task-specific methods and multimodal pretraining developments. We elaborate on the significance of pretraining tasks for large-scale pretrained multimodal models and their robustness across diverse downstream tasks. Additionally, we catalogue pretraining benchmark datasets, covering their sizes, included modalities, and targeted tasks. This survey also investigates the advancements of unifying architectures and presents a comparative analysis of various SOTA models. We extend our exploration to a wide range of applications, including vision, language, audio, healthcare, and NLP-specific tasks. Table 1 provides a comparative view of our work with other surveys, highlighting the inclusion of pretraining concepts, downstream applications, NLP applications, evolutionary perspectives, unifying architectures, datasets, and active repository availability. This broad approach facilitates a holistic view of multimodal learning, guiding future research by shedding light on efficient multimodal handling techniques.

3 MULTIMODAL DEEP LEARNING APPROACHES

This section explores multimodal architectures, divided into task-specific and Pretraining-Finetuning architectures (Figure 2). Section 3.1 reviews task-specific methods, recently evolving into large-scale pretrained approaches. Section 3.2 outlines the pretraining process, including pretext tasks and **state-of-the-art (SOTA)** frameworks for multimodal tasks. Section 3.3 compares results from SOTA approaches, while Section 3.4 lists the acronyms used.

Table 1. Survey Comparison Comparing between Different Survey Work Based on Their Content

Ref.	Year	Pretraining	Applications	NLP focused	Evolution	Unifying Arch.	Datasets	Active repo
[43]	2017	✗	✓	✗	✗	✗	✓	✗
[2]	2018	✗	✓	✓	✓	✓	✗	✗
[6]	2019	✗	✗	✗	✓	✓	✗	✗
[1]	2021	✗	✓	✓	✓	✗	✓	✗
[44]	2021	✓	✓	✓	✗	✗	✓	✗
[41]	2022	✓	✓	✗	✗	✓	✓	✗
[7]	2022	✗	✓	✓	✗	✓	✓	✗
[42]	2022	✓	✓	✓	✗	✗	✓	✗
[28]	2022	✓	✓	✗	✗	✓	✓	✗
[40]	2023	✓	✓	✗	✗	✗	✓	✗
Our Work	2023	✓	✓	✓	✓	✓	✓	✓

The content studied is the presence or absence of pretraining concepts, multimodal applications, NLP focused (textual and audio forms), evolutionary perspective, unifying architectures, benchmark datasets, and active repository.

3.1 Multimodal Task-Specific Methods

Multimodal representation learning improves the robustness of **deep learning (DL)** models as complementary features are present. This section describes task-specific multimodal methods that can be generally classified into encoder-decoder-based, attention-based, and reinforcement learning-based models.

3.1.1 Encoder-Decoder Based Models. Encoder-decoder models, such as cascaded CNN-RNN and RNN-RNN, transform input data to output via semantically rich latent representations. Wu and Hu [51] proposed a **Cascaded Recurrent Neural Network (CRNN)** to learn image-text interactions. This uses a VGG16 network as the encoder and a **Stacked Gated Recurrent Unit (SGRU)** as the decoder. Ji et al. [52] suggested a **Globally Enhanced Transformation (GET)** network, employing Faster-RCNN and ResNet-101 as the encoder. The GET network outperformed the CRNN on the MS-COCO dataset [53] with a higher BLEU score.

Autoencoders [54], a form of encoder-decoder model using unsupervised learning, construct latent representations to reconstruct original data points [54]. In [55], a multimodal architecture is built on autoencoders for processing audio and video modalities, feeding into a conventional encoder-decoder architecture.

3.1.2 Graph Based Models. In multimodal learning, the fusion module, which facilitates the interplay of features across various modalities, has emerged as a critical and central component [56]. Task-specific methods extensively employ graph-based approaches for fusion tasks [28]. **Graph Neural Networks (GNNs)** demonstrate exceptional robustness when handling multimodal data, skillfully compensating for incomplete or noisy inputs in one mode by leveraging information from other modes [57]. In multimodal tasks, GNNs are capable of encapsulating both local and global context through iterative neighborhood aggregation [58]. Graph-based multimodal fusion encoder [59] for **Neural Machine Translation (NMT)** performed superior on Multi30K datasets. The method constructs graphs containing two types of nodes: all words in the sentence are included as textual nodes and the nouns in the sentence, identified by the Stanford Parser, are detected by the visual grounding toolkit as visual nodes. Additionally, two types of edges are employed: intra-modal edges to connect nodes in the same modality, and inter-modal edges to connect corresponding nodes in different modalities, capturing the semantic relationship of multi-modal data. These graphs are processed through multiple graph-based multimodal fusion layers, which iteratively facilitate semantic interactions and learn node representations, ultimately producing an attention-based context vector for the decoder. Likewise, Gao et al. [60] addressed the scene textual

ambiguity in a better way than pre-trained word-embeddings by designing **Multimodal Graph Neural Network (MM-GNN)**. VQA-GNN [61] integrates image-level information and conceptual knowledge into a unified multimodal semantic graph for joint reasoning. The effectiveness of this model is demonstrated by top-ranking performances in the VCR task, outperforming previous models, and notably, it also provides a pioneering cross-domain interpretability for visual and textual knowledge in the visual question answering task.

3.1.3 Attention Based Models. The attention mechanism improves model learning by inspecting information flow, features, and resources. These models help the network to deal with long-time dependencies by allocating attention to vital information and filtering all unnecessary or irrelevant stimuli. A recent work by Jiang et al. [62] employed the attention notion through **Attention Weight Gate (AWG)** module and **Self-Gated (SG)** module to the **Multi Gate Attention Network (MGAN)** pipeline. Moreover, MGAN extracted and utilized intra-object relation in the network [62]. Another attention based variant is explored in **attention-guided Multimodal Correlation (AMC)** [63]. The attention mechanism is applied to the modality rather than a semantic context in a vector, varying the importance of each modality depending on the required query from the system. The application of AMC has applied search logs, where the user inputs an image and text to be searched, and the system determines which modality is more valuable.

3.1.4 Reinforcement Learning-Based Models. Reinforcement learning models learn through trial-and-error interactions between an environment-aware agent and its environment, balancing between exploiting prior knowledge and exploring new actions. The algorithm "reinforces" actions yielding high rewards. Unlike supervised learning, it doesn't need labeled data. This concept is integrated into deep learning in [64], using a **Hierarchical-based Reinforcement Learning (HCL)** framework with roles like a manager, worker, and critic. The manager sets goals for the worker, who performs actions to achieve them, while the critic evaluates goal completion. The HCL outperforms baseline models in transforming video modality into text.

As the complexity of multimodal frameworks increases, further deep learning architectures are modified and enhanced, such as **Generative Adversarial Networks (GAN)** and **Probabilistic Graphical Models (PGM)** [65, 66]. However, the approaches mentioned in this section pose challenges in terms of time and computational resource efficiency due to the need for retraining for each specific task. Recognizing these challenges, many researchers have now shifted their focus to pretraining frameworks. These aim to address and overcome the limitations inherent in task-specific methods, reducing the need for repetitive retraining for different tasks, and thereby presenting a more efficient avenue for future multimodal representation learning.

3.2 Multimodal Pretraining Methods

This section covers a well-detailed discussion of the pretraining framework, with a special emphasis on multimodal pretraining, including pretraining types, pretext tasks, and state-of-the-art methods. Earlier,**Self-Supervised Learning (SSL)** boosted pretraining for language tasks [27] which was later used for vision tasks and produced efficient results [67]. This concept was then extended to multimodal scenarios. Pretraining on massive labeled or unlabeled datasets and fine-tuning on task-specific datasets has become a modern paradigm for various domains [68].

3.2.1 Types of Pretraining. Kalyan et al. [46] proposed different types of pretraining approaches used by researchers to design and train the model on a large scale. These approaches include: (1) **Pretraining From Scratch (PTS)** in which massive unlabeled text is used to train the model from scratch using random initialization for all the layers of the language model, e.g., ELECTRA [69], RoBERTa [70] and BERT [35]. (2) **Continual Pretraining (CPT)**, models are initialized

using existing pretrained models. This approach reduces the required computational resources. The domain-specific task uses this method to alleviate the challenges of short target vocabulary. BioBERT [71], HateBERT [72] and infoXLM [73] are the examples of CPT. (3) **Simultaneous Pretraining (SPT)** used in [74]. It consumes domain-specific and general text simultaneously while training from scratch. (4) **Task Adaptive Pretraining (TAPT)** used in [75] demands a small amount of data. (5) **Knowledge Inherited Pretraining (KIPT)** [76] uses the Knowledge distillation technique.

3.2.2   Pretraining Tasks. In the pretraining phase, predefined or pretext tasks are solved to learn language representation. These tasks are based on self-supervised learning and are challenging enough to make the model robust by exploiting the training signal, which include: (1) **Casual Language Modeling (CLM)** is unidirectional and uses context to predict the next word employed by GPT-1 [77] for the first time. (2) **Masked Language Model (MLM)** enhanced the representation by considering the context in both directions and consuming only 15% of the tokens applied by [35] for the first time. (3) **Replaced Token Detection (RTD)** mitigates the pretraining challenges of MLM (less supervisory signal) by classifying the replaced token as original or not [69]. (4) **Shuffled Token Detection (STD)** reduces the discrepancy between pretraining and finetuning by exploiting discriminative tasks (detection of shuffled tokens) [78].Other pretraining tasks include **RTS (Random Token Substitution), NSP (Next Sentence Prediction)**, **SLM (Swapped Language Models)**, and **SOP (Sentence Order Prediction)**, which established as baseline pretext tasks for the multimodal pretraining.

3.2.2.1 Multimodal Pretraining Tasks. Multimodal pretraining tasks offer the benefits of enhanced representation learning, cross-modal transfer, improved generalization, effective multimodal fusion, and improved performance in a wide range of multimodal applications. These tasks include: (1) **Cross-Modal Masked Language Model (CMML)**, which is similar to MLM in BERT; moreover, pretraining models also consider contextual visual features and unmasked tokens. It is an effective approach to vision language pretraining by assisting the model for better align and regard the relationship between different modalities [41]. (2) Cross-Modal **Masked Region Prediction (MRP)** [79]: Similar to the previous objective, MRP masks RoI with zeros. The contextual visual features and text inference predicts the masks. (3) **Image-Text Matching (ITM)** or **Visual Linguistic Matching (VLM)** [80]: This objective considers the coarse-grained relationship between image and text features by matching and aligning. It measures the relevancy between different modalities at high level. (4) **Cross-Modal Contrastive Loss (CMCL)**: This objective learns universal vision and language features by pushing the relevant image-text pairs close and non-relevant apart. The highest similarity score between embeddings of different modalities help make the decision. The similar alternative multimodal pretraining tasks are **Masked Object Classification (MOC)** [79], **Phrase-Region Alignment (PRA)** [81], **Word Region Alignment (WRA)** [82], **Video-Subtitle Matching (VSM)** [83], **Mask Frame Modeling (MFM)** [83], **Frame Order Modeling (FOM)** [83], **Visual Translation Language Model (VTLM)** [84], and so on.

Pretrained approaches have been used extensively to perform exceptionally well for NLP downstream tasks. The researchers proposed pretrained models including BERT [35], ALBERT [85], RoBERTa [70], and T5 [86], to learn general textual representation and to boost downstream tasks' performance. Similarly, pretrained approaches have been adopted to reduce the time and computation for vision-based downstream tasks [87, 88]. Simultaneously, the pretraining paradigm is making strides in the domain of multimodal data, enhancing performance across a diverse range of tasks, from cross-modal to fully multimodal. The upcoming section unveils recent architectures crafted to exploit this rich multimodal data.

3.2.3 Transformer-Based Architectures. Recent work on large scale pretraining has inspired the research community to progress in multimodal tasks [89]. The combined pretraining-based advanced approaches on image-text pairs at a large scale are developing rapidly which outperform task-specific architectures. For instance, in ViLBERT [9], the authors pretrained BERT on textual and visual inputs in different streams that maintain relationships through co-attention layers. After pretraining on the Conceptual Captions dataset [90], the model is transferred to mainstream language-vision tasks, including VQA, VCR, and caption-based image retrieval and achieved consistent improvement over task-specific architectures. The **Vision-Language Pretraining (VLP)** in a unified manner is proposed by [91] that outperformed the previous SOTA for image captioning and VQA tasks. The proposed model follows the shared encoder-decoder architecture and is pretrained on a large scale image-text pairs in an unsupervised manner. They evaluated the model on VQA 2.0, COCO Captions, and Flickr30k [92] Captions. In line with prior work, Li et al. [93] proposed a pretraining approach in which the object tags are used as anchors to enhance the alignment learning with paired text.

In addition to aforementioned approaches, structured knowledge is extracted from scene graphs in the recent research [95], which assist in learning joint representations as semantically connected features of text and vision. Scene graph prediction tasks which includes **Object Prediction (OP)**, **Attribute Prediction (AP)**, and **Relationship Prediction (RP)** are newly proposed by the authors in the pretraining phase. The proposed model can learn semantically aligned joint representation. The model's efficacy is evaluated across five cross-modal tasks, wherein it leads the VCR task, exhibiting a performance enhancement of 3.7%. Desai and Johnson [94] proposed VirTex, a data-efficient alternative to supervised pretraining that uses captions as a supervisory signal for vision tasks. They conducted joint pretraining of the model which comprises of convolution and transformer backbones from scratch for vision and language modalities which aims to generate captions for images. The visual backbone comprised of ResNet-50 is used to compute visual features of input images, which are then passed to the textual head to generate captions for corresponding images. After training, the textual head is discarded, as the primary aim is to utilize the learned features for subsequent visual tasks. Moreover, Shi et al. [3] proposed **Audio-Visual Hidden Unit BERT (AV-HuBERT)**, trained from scratch on LRS3 benchmark which consists of speech and visual features. The hybrid architecture ResNet-transformer serves as the backbone of the proposed model. AV-HuBERT extracts the phonetic and linguistic features from audio and visual streams simultaneously to produce latent representation. This representation can be used for pretraining baselines, e.g., HuBERT [99] which can outperform models that are pretrained on a single modality. Table 2 draws the comparative picture of discussed approaches, on the basis of benchmarks, tasks and architectural settings, respectively.

Word2Vec [100], GPT [77], ELMO [101], BERT, and other SSL-based frameworks perform good natural language understanding but do not consider grounding information from the visual world as motivated by Bender and Koller [102] and Bisk et al. [103]. The pretraining process designed by Tan and Bansal [96], known as Vokenization, considers both language tokens and their corresponding visual information, termed as "vokens", as input for supervision. The authors utilized the small image captioning dataset to train a vokenizer that can further generate language vokens of large corpora. The results achieved by visually supervised models outperformed SSL-based pretraining approaches on GLUE, SWAG, and SQuAD benchmarks. However, models trained using supervised methods often rely heavily on labeled data, which can limit their generality and usability. Therefore, to overcome these challenges, self-supervised multimodal learning methods have emerged, such as CLIP [89]. **CLIP (Contrastive Language-Image Pre-training)** is pretrained on **WIT (WebImageText)** [89], which contains image-text pairs collected from the web. Unlike standard vision models, CLIP jointly trains an image encoder and a text encoder during the training phase.

Table 2. Transformer-based Pretraining Approaches, Benchmarks, Tasks and Architectural Settings

Model	Benchmarks	Pretraining Tasks	Architectural Comparison
VLP [91]	VQA2.0, COCO Captions, Flickr30k	IC, VQA	12-layer, 768-hidden, 12-heads, 110M param.
VirTex [94]	COCO Captions	Image Classification, IC	Visual backbone: ResNet-50; textual head: two unidirectional Transformers
ViLBERT [9]	VQA, VCR, RefCOCO, IR	VQA, VCR, GRE, CIR	Textual Encoder: BERT-base; Visual backbone: Faster R-CNN (ResNet 101)
ERNIE-ViL [95]	VCR, RefCOCO+, VQA, IR-Flickr30K, QR-Flickr30K	SGP, VCR, VQA, GRE, IR, TR	Textual Encoder: BERT-base; Visual backbone: Faster R-CNN (ResNet 101)
OSCAR [93]	COCO, CC, SBU captions, flicker30k, GQA	VQA, ITR, IC, NOC, GQA, NLVR2	Base: 12-layer, 768-hidd, 12-heads, 110M param.; Large: 24-layer, 1024-hidd, 16-heads, 340M param.
Vokenizer [96]	SST-2. QNLI, QQP, MNLI,SQuAD v1.1, SQuAD v2.0, SWAG Avg.	GLUE, SQuAD, SWAG, GLUE	Textual Encoder: BERT; Vision Encoder: ResNeXt-101-32x8d
AV-HuBERT [3]	LRS3, VoxCeleb2	ASR	Hybrid ResNet-Transformer architecture
CLIP [89]	COCO, Visual Genome, YFCC100M	Image Classification, IC, OCR, AR, Geo-Localization	Textual Encoder: GPT-2, GPT-3; Vision Encoder: ViT-B/32, ViT-B/16, ViT-L/14
BLIP [97]	No-Caps, COCO, Flickr30k	ITR, IC, VQA, VD, NLVR	Textual Encoder: BERT base; Image Encoder: ViT-B/16, ViT-L/16
BLIP-2 [98]	No-Caps, COCO, Flickr30k, VQAv2, OK-VQA, GQA	Instructed Zero-shot Image-to-Text Generation, VQA, IC, ITR	Textual Encoder: Q-former; Vision Encoder: FlanT5, ViT-L/14, ViT-g/14

The performance of CLIP is benchmarked across over 30 vision and vision-language datasets, including COCO [53], Visual Genome [104], and YFCC100M [105], and the like. **BLIP (Bootstrapping Language-Image Pre-training)** [97], a unified vision-language framework, uses knowledge distillation on captions for improved performance. It achieves state-of-the-art results on various tasks, including image-text retrieval, image captioning, VQA, and even in zero-shot performance on text-to-video retrieval and VideoQA. BLIP is extensively evaluated on No-Caps [106], COCO [53], and Flickr30k [92]. BLIP-2 [98], an improved visual-language model based on comprehensive pretraining strategy, achieves state-of-the-art performance across numerous vision-language tasks. BLIP-2 exhibits a wide range of zero-shot image-to-text abilities, which include visual knowledge reasoning, visual common sense reasoning, visual conversation, and personalized image-to-text generation. Like its predecessor, BLIP-2 uses the same evaluation datasets and continues to implement the **Captioning and Filtering (CapFilt)** method.

3.2.4 Unifying Architectures. Unifying architecture is designed to accept different modalities as input and train the model on multiple tasks to lessen the task-specific parameters as generic architecture. Li et al. [107] presented the unified model for text-only and vision-only tasks. The authors put effort into creating a single suitable transformer-based pretrained model that can be finetuned on any modality for any downstream task. The foundation model is a transformer pretrained jointly on unpaired images and text. The pretraining is based on knowledge distillation

from the teacher pretrained model for better joint training and gradient masking for balancing the parameters' updates. The shared transformer that can encode all modalities for the different tasks by employing a task-specific classifier is used. The aspiration is to reduce the task and modality-specific parameters by bringing up a more generic model, and maximum computation occurs in the shared transformer module.

The UNITER [82] is a pretraining approach conducted at a large scale over four benchmark datasets. Joint embedding from both modalities is learned to perform heterogeneous downstream tasks. MLM, **MRM (Masked Region Modeling)**, **ITM (Image-Text Matching)**, and **WRA (Word-Region Alignment)** are employed as a pretraining task. Additionally, the pretraining task achieves global image text alignment using Conditional masking. Optimal Transport is the authors' second concept for the WRA task to improve the alignment between images and words. Hu and Singh [108] proposed a unified transformer for multitasking based on multimodal learning at a time. The proposed architecture uses a specialized encoder for each modality and a shared decoder for every task. DETR [87] is used for visual features encoding and BERT [35] performs the textual feature encoding. Contrastive learning is employed on multimodal data by Akbari et al. [109] to train a transformer encoder that processes audio, text, and video simultaneously.

Wang et al. [110] proposed the **One For All (OFA)** method, which unifies tasks and modalities via a sequence-to-sequence framework based on Unified vocabulary (for all kinds of the modality). OFA represents data of different modalities in a unified space that discretizes images and text to form a unified output vocabulary. They presented three qualities that a unifying model should support to maintain multitasking for any modality: (1) Task-Agnostic: handcraft instruction-based learning is utilized to achieve this property. (2) Modal-Agnostic: single Transformer-based architecture uses globally shared multimodal vocabulary to make it modal agnostic. (3) Task comprehensiveness: pretraining conducted on various unimodal and multimodal tasks to achieve task comprehensiveness. The transformer is the backbone of the encoder-decoder unified network for the pretraining, finetuning, and zero-shot tasks. It considers multimodality and multitasking to make it more generalized for unseen tasks. OFA achieved SOTA on multimodal and outperforms other well-known pretrained for unimodal. As in GPT [77] and BART [111], **BPE (Byte-Pair Encoding)** is used to divide a sequence of words into sub-word sequences and embed them into features.

InstructBLIP [112] a general-purpose vision-language model, has achieved state-of-the-art, zero-shot performance on numerous vision-language tasks. InstructBLIP is a vision-language instruction tuning framework consisting of an image encoder, a **Large Language Model (LLM)**, and a Q-Former [98]. In its evaluation, InstructBLIP employs two LLMs: FlanT5 [113] and Vicuna [114]. The evaluation metric incorporates 11 tasks across 28 datasets, including image captioning, video reasoning, visual conversational QA, knowledge grounded image question answering, video question answering, image captioning reading comprehension, image question generation, image classification, and LLaVa-Instruct-150k, a benchmark that incorporates visual conversation, complex reasoning, and detailed image description.

3.3 Discussion

Most multimodal approaches use a transformer as a backbone. These architectures have outperformed CNNs on most vision tasks as self-attention focuses on global contextual learning at a low-level stage. However, the generalization of large-scale multimodal architectures heavily depends upon the pretraining objective, learning technique, and nature of the benchmarks illustrated in this survey.

VLP [91] is considered as the first SOTA on joint vision-language multi-modal tasks specifically generation and understanding. Likewise, VirTex showed SOTA performance on multimodal tasks [94]. ViLBERT [9] marked SOTA on VQA and RefCOCO+ with a huge margin. AV-HuBERT [3]

Table 3. The Unifying Architectures with the Benchmarks, Tasks and Architectural Setting Adopted by the Authors

Model	Benchmarks	Tasks	Architectural Comparison
UniT [108]	COCO, QQP, VQAv2 QNLI, MNLI, SST-2, SNKI-VE	VQA, OD, VE	201M param., comprised of ResNet-50 + BERT
ViT-BERT [107]	MNLI, QQP, QNLI, SST-2, RTE Cifar, ImageNet, Flowers, Pet	Vision-only, text-only	12 layer transformer, 768-hidden, 3072 MLP
UNITER [82]	VQA, Flicker30k, NLVR, Ref-COCO	IR, TR VQA, RE	12-24 layers, 768- 1024 hidden 12-16heads, 86-303M param.
VATT [109]	AudioSet, HowTo UCF101, HMDB51, MSR-VTT	VAR, AEC, TVR	Transformer-based variants; 155M to 415M param.
OFA [110]	SST-2, RTE, MRPC, QQP, QNLI, MNLI, SNLI-VE	NLU, NLG, Image Classification	Transformer-based variants; 33M to 940M param.
InstructBLIP [98]	No-Caps, Flickr30k, GQA,VSR, IconQA, TextVQA, Visdial, HM, VizWiz, SciQA IMG, MSVD QA, MSRVTT QA, iVQA	IC, VR, ICRC, Video QA, IQG, Image QA,LLaVa-Instruct-150k	Variations of BLIP-2, Image encoder:ViT-g/14,LLMs:FlanT5-XL (3B), FlanT5-XXL (11B), Vicuna-7B and Vicuna-13B

used audio-visual representation on the benchmark (for audio-only speech recognition) and led to 40% relative WER reduction over the state-of-the-art performance. ERNIE-ViL [95] achieved SOTA on five cross-modal downstream tasks and ranked at first on the VCR leaderboard with an absolute improvement of 3.7%. OSCAR [93] created new SOTA on six well-established vision-language understanding and generation tasks. Vokenizer [96] presented visually-supervised language models with consistent improvements on multiple language tasks. Finetuning on a similar benchmark offers outstanding results, though this needs more parameters, time, and resources. Many researchers are addressing these challenges by proposing unifying architectures to deal with multiple modalities with the same backbone for multitasking. Such generic models as summarized in Table 3 can share the parameters for different downstream tasks and avoid extensive pretraining or finetuning for each task. Although unifying architectures are not specialized for a specific task, they show competitive results.

It is observed that pretraining in a correlated manner considering different modalities enhance performance if strong alignment and correspondence among representation are considered during the training phase. The models trained on multiple modalities also express good performance on unimodal tasks as ViT-BERT [107] surpassed the ViT on vision-only tasks. However, ViT-BERT could not surpass BERT on text-only tasks due to the low amount of data for finetuning, specifically on the RTE dataset. Likewise, the results achieved by joint training of baselines give better performance than cross-modal finetuning, which faces the challenge of mismatching between pretraining objectives and downstream tasks. ViT-BERT achieved 83% and 89% average scores on Text-only and Vision-only tasks, respectively, making it a moderate option with an 86% average for both tasks (consume fewer parameters than UniT). The UniT [108] as a generic model achieved promising results with shared parameters on all tasks. However, UniT could not surpass

VisualBERT or BERT, which are trained for specific tasks only. Smaller batch size and higher learning rate cause the problem of lower performance and divergence, respectively.

The UNITER-large [82] model achieved comparable performance across all the benchmarks but consumed 303M parameters. UNITER-base model also performed well except VQA with 86M parameters. VATT [109] outperformed the CNN-based approaches in all metrics for audio event recognition and video action recognition tasks. Additionally, it exhibits competitive results for text-to-video retrieval. The author introduced the DropToken strategy that reduces the computational complexity of training. DropToken is the sampling technique alternative to dimension and resolution reduction to mitigate redundancy challenges. Additionally, Multimodal Contrastive Learning based on Noise Contrastive Estimation and Multiple Instance Learning is used to align video-audio and video-text pairs after projecting in the shared space. OFA [110] surpassed the UNITER, UNIMO, and other SOTA on VQA and SNLI-VE. Notably, it achieved the highest score on RTE and RefCOC0+ by outperforming task-specific and unifying architectures. Specifically, OFA results assert that large-scale multimodal pretrained approaches compete with natural language pretrained SOTA for understanding and generation tasks. Furthermore, OFA outperformed Uni-Perceiver on out-of-domain tasks, e.g., single sentence and sentence pair classification.

InstructBLIP [112] proposed a novel instruction tuning framework towards generalized vision-language models. Their proposed method achieved state-of-the-art performance across an array of benchmarks with strong evaluation protocol. Besides, InstructBLIP has demonstrated its potential as an advanced initial model for finetuning on downstream tasks.

3.4 Acronyms

This section contains all acronyms in the tables. **AEC:** Audio Event Classification, **Auton. Driving:** Autonomous Driving, **AR:** Action Recognition, **AVSS:** Audio Visual Speech Synthesis, **CIR:** Caption to Image Retrieval, **CMR:** Cross-Modal Retrieval, **ED:** Event Detection, **EL:** Entity Labeling, **EQA:** Embodied Question Answering, **GR:** Gesture Recognition, **GRE:** Generation of Referring Expressions, **IC:** Image Captioning, **ICRC:** Image Captioning Reading Comprehension, **IE:** Information Extraction, **IQG:** Image Question Generation, **IR:** Image Retrieval, **IS:** Indoor Segmentation, **ITR:** Image to Text Retrieval, **LaVa-Instruct-150k:** includes visual conversation, complex reasoning, and detailed image description, **MML:** Multimodal Learning, **MMSA:** Multimodal Sentiment Analysis, **NLG:** Natural Language Generation, **NLU:** Natural Language Understanding, **NLVR:** Natural Language for Visual Reasoning, **OCR:** Optical Character Recognition, **OD:** Object Detection, **RE:** Referring Expression, **SSS:** Sound Source Separation, **TD:** Text Detection, **TR:** Text Retrieval, **TVR:** Text-to-Video Retrieval, **VAC:** Video Action Recognition, **VAR:** Visual to Audio Retrieval, **VC:** Video Captioning, **VCR:** Visual Common Sense Reasoning, **VD:** Visual Dialog, **VE:** Visual Entailment, **VG:** Visual Generation, **VLP:** Vision Language Pretraining, **VQA:** Visual Question Answering, **VR:** Video Reasoning, **VU:** Visual Understanding.

4 MULTIMODAL APPLICATIONS

This section presents the categorical detail of multimodal applications enhanced by deep learning architectures as shown in Figure 3. The multimodal tasks are divided into main categories: understanding, classification, retrieval, generation, and translation. The benchmark, evaluation metrics, description, and comparison for the best-performing architectures are discussed for each multimodal application.

4.1 Understanding

Understanding text, speech, and vision is the most important and primary task of AI-based systems. Relevant information extraction, recognition, and identification of entities from different mediums
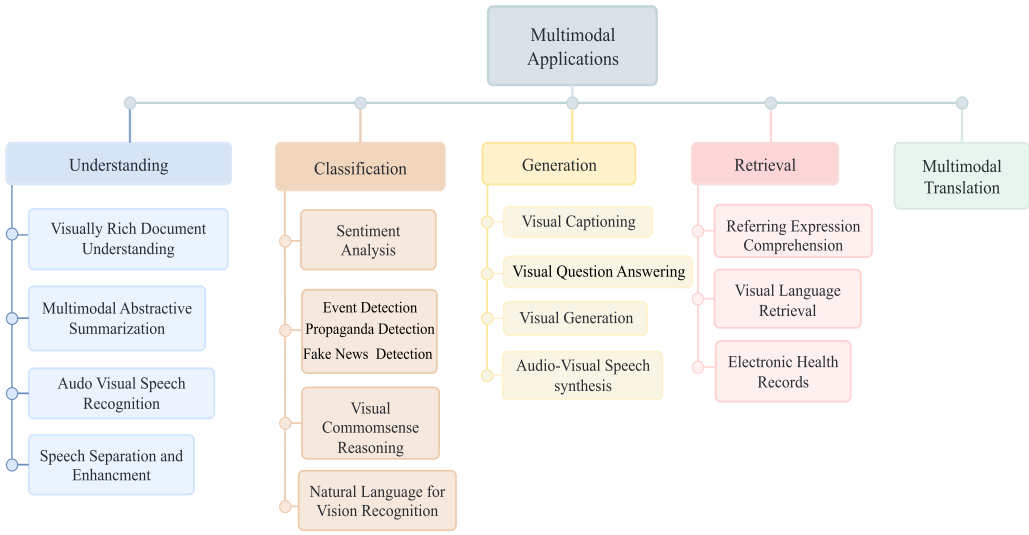
Fig. 3. Taxonomy of the multimodal applications in Section 4.

(text, images, speech) are its subtasks. The information achieved through understanding benefits many downstream tasks directly or indirectly.

4.1.1 Visually Rich Document Understanding. Understanding visually rich documents with structured information is important for different applications [115]. Contrary to the classical information extraction methods, Visually Rich Documents Understanding considers the visual layout along with the text. LayoutLMv2 [116] pretrained a Transformer that learns from different modalities at the pretraining stage by integrating and aligning layout, textual, and visual information. LayoutMv2 used more than 10 billion documents to pretrain a model that outperformed SOTA on document understanding tasks. More recently, Li et al. [117] proposed StrucTexT, which outperformed LayoutLMv2 by pretraining a Transformer in a self-supervised manner, with fewer parameters, achieved high scores. The publicly available benchmark datasets for downstream tasks are SROIE [118] and FUNSD [119]. LayoutLMv2 achieved 97% and 84% F1 scores on SROIE and FUNSD, respectively. StrucTexT had 96.88% and 85.68% F1 on SROIE and FUNSD, respectively.

Gu et al. [120] released an improved version XYLayoutLM of LayoutLMv2 [116]. They proposed Augmented XY Cut to capture reading orders as layout information, which was neglected previously. Furthermore, the variable length of the input sequence is dealt with **DCPE (Dilated Conditional Position Encoding** - inspired by CPE [121]) that creates the 1D and 2D features of textual and visual input, respectively, for Convolution. The model succeeded in surpassing the other approaches on XFUN [116] benchmark for **SER (Semantic Entity Recognition)** and **RE (Relation Extraction)** tasks.

4.1.2 Multimodal Abstractive Summarization. This task takes massive multimodal content (video, images and corresponding text) from the internet and extracts the vital information to generate a summary [122]. Palaskar et al. [123] generated a summary using Multisource seq2seq with Hierarchical attention that integrates information from different modalities. Likewise, a Multistage fusion network was proposed that established the interaction between different source modalities. Recently, an approach was proposed [124] that exploits the visual modality to generate a summary with **Generative pretraining language models (GPLMs)**. The add-on layer based on attention

is inserted in visually-guided GPLMs to maintain the visual incorporation and text generation. The model was evaluated on the How2 [125] dataset, which contains instructional videos with two to three sentences of 2,000 hours. The visually grounded variants of the proposed model outpaced the baselines, achieving the highest ROGUE-1, ROGUE-2, and ROGUE-L on the benchmark.

4.1.3 Audio-Visual Speech Recognition. **Audio-Visual Speech recognition (AVSR)** is one of the earliest multimodal research domains that encouraged the research community to understand speech by using hearing and vision features simultaneously. Visual features and speech play a vital role in understanding speech, especially noise and assisting patients suffering from speech impairment. Most of the recent models wav2vec2.0 [37], HuBERT [99], and De-CoAR2.0 [126] were using audio only for speech recognition. Lip movement provides a supervisory signal by exploiting self-supervision to recognize better.

Recently released, **Audio-Visual Hidden Unit BERT (AV-HuBERT)** [3] is a self-supervised representation learning framework. The mentioned approach learns audio-visual speech representation by taking favor of lip-reading and ASR. The proposed model achieved 32.5% WER on LRS3 [127] benchmark by using 30 hours of labeled data from 433 hours and in achieving 26.9% outperformed the former SOTA that trained on 31K hours [128]. The author also stated that exact representation performed speech recognition on audio-only tasks by reducing 40% WER. The model is exceptional for visual only modality by exploiting contextualized representation of AV-HuBERT.

4.1.4 Speech Separation and Enhancement. **Speech enhancement (SE)** is the process in which speech signals of the target speaker are extracted in an acoustically noisy environment. SE improves speech quality (sounds) and speech intelligibility (linguistic content). The estimation of multiple targets in the speech is known as speech separation, or source separation [129]. Previously, these tasks were handled with statistical and mathematical criteria under signal processing [130]. Currently, the evolution of supervised learning, multimodal methods, and fusion techniques attain the uninfluenced visual features of the speaker accompanying acoustic features in a noisy environment.

Audio-Visual Sentence Extraction (AV-SE) and **Audio-Visual Sentence Separation (AV-SS)** systems count several speakers and trace their faces. Detection and Tracking algorithms are used to achieve high-dimensional visual frames of faces. The dimensionality is reduced with an active appearance model that is based on **principal component analysis (PCA)** [131]. Sound Source separation is achieved by Zhu and Rahtu [132] using the Appearance Attention Module that leverages categorical information on of single frame video. The authors optimized the model by correlating the appearance embedding and feature maps. The sound source is located by the scalar product of embedding and feature maps. The dataset MUSIC contains YouTube videos made up of different musical instruments and has off-screen noise [133]. The model is evaluated on the MUSIC dataset using standardized evaluation matrices, i.e., **Signal to Distortion Ratio (SDR)**, **Signal to Interference Ration (SIR)**, and **Signal to Artifact Ratio (SAR)**. The attention-based approach surpassed the Resnet-18 and Resnet-50 by achieving SDR 10.74, SIR 17.29 and SAR 13.04.

4.2 Classification

Classification is the systematic arrangement of samples into a group or category that follows the established criteria. Several tasks of vision, textual and speech fall under the umbrella of classification. The prominent multimodal classification tasks discussed in this study are sentiment analysis, fake news detection, and event detection.

4.2.1 Sentiment Analysis. Sentiment analysis is the core task in NLP that extracts and classifies reviews, feeling, gestures and behavior toward a specific entity. Sentiment Analysis understands

people's perspectives for efficient decision-making in multiple domains. Textual representation has been analysed widely in previous research such as [134, 135], and [136]. However, the sentiment analysis task moved from text modality to another form of modality due to social media and the internet. Chen and Li [137] designed deep learning architecture to extract sentiment from multimodal complex data by designing two component-based methods consisting of shallow fusion and aggregation parts. The shallow fusion component extracts contextual information from the different domains using the attention mechanism, and the aggregation part attains sentimental word-aware fusion. The proposed architecture outperformed other methods by achieving the highest score on multimodal datasets CMU-MOSI, CMU-MOSEI, and YouTube datasets.

4.2.2 Event Detection. Detection of an event, trend or situation specifically through content available on social media becomes more efficient and generous by considering data from different modalities [138]. A massive amount of data is generated that significantly impacts the lives, property, and psychology of humans [139]. Event detection can be mapped to other realistic scenarios, such as Emergency Management, Disaster Detection, and Topic Detection [47]. Even though the data from different perspectives enhance the performance, the challenges are also increased for methods to deal with redundant and heterogeneous characteristics. No large enough dataset is available for disaster detection to employ deep learning architecture except CrisisMMD [140], CrisisNLP, and CrisisLex. For traffic event detection, Chen et al. [141] created a multimodal dataset by integrating the traffic-related filtered tweets with sensor data. The author achieved 84%, 83%, and 87% F1 score with CNN, RNN, and **mmGAN (multimodal GAN)** models, respectively.

4.2.3 Detecting Propaganda in Memes. Propaganda is the type of communication that affects the psychology of people that leads them to keep specific opinions about any entity or perform some action. Due to the high usage of social media, it has become a societal and political problem. Memes that contain visual and textual content are being used as a significant fraction of the medium on the internet to trigger this issue.

Dimitrov et al. [4] approached this problem as a multimodal and multi-label task by detecting the techniques to promote propaganda in memes. They released a new dataset consisting of 950 memes (containing both textual and visual content) annotated with 22 different propaganda techniques. Previous datasets addressed the propaganda at document level [142], sentence level and fragment level [143]. SOTA models performed the experiments. Experiments were conducted on variations of models that are pretrained on two approaches: (i) unimodally pretrained and (ii) pretrained with the multimodal objective. BERT and ResNet-152 are trained separately in the first variation and fused using **MMBT (Multimodal bitransformers)** with early, middle, or late fusion. ViLBERT and Visual BERT are pretrained in multimodal nature on Conceptual Captions and MS COCO, respectively. The results achieved by ViLBERT and Visual BERT surpassed the baseline and other variations. Results analysis expressed that transformer-based multimodal approaches are practical and produce efficient results.

4.2.4 Fake News Detection. Fake news can use multimedia to cause panic on social media. The text and the image content misinterpret the facts and manipulate human psychology, leading to rapid propagation of fake news [144]. A multimodal architecture can detect fake news by looking for mismatches between the modalities [145].

The problem of detecting fake news for unseen and emerging events is addressed in [146] by leveraging an Adversarial Neural Network that uses an event discriminator to remove event-specific features and preserve shared features among events and modalities. Khattar et al. [147] proposed SpotFake, which utilizes BERT to learn textual features and VGG-19 (pre-trained on ImageNet) to learn visual features. They employ a simple concatenation technique to combine features

obtained from different modalities. Similarly, [147] introduces the **Multimodal Variational Autoencoder (MVAE)**, consisting of an encoder, a decoder, and a fake news detector module. The variational autoencoder utilizes probabilistic latent variable models by optimizing a bound on the marginal likelihood of the observed data. By leveraging the acquired multimodal representations from the bimodal variational autoencoder, the fake news detector classifies multimodal posts as genuine or fake.

Recently, Wang et al. [148] proposed **fine-grained multimodal fusion networks (FMFN)** to detect fake news. First, CNN extracted visual features, and RoBERTa [70] is used to get contextualized embedding of words. Then an attention mechanism is used between visual and textual features to enhance the correlation for fusing features. Finally, a binary classifier is adopted to perform detection on fused features. Their model achieved 88% accuracy on the Weibo dataset [144] by focusing only on tweets that contain text and images.

4.2.5 Visual Commonsense Reasoning. Visual Commonsense Reasoning performs robust Visual understanding by Integrating cognition, grounding, and language reasoning with recognition tasks. Zellers et al. [149] introduced the VCR dataset containing 290k **MCQA (Multiple Choice Question Answer)** problems. The author proposed the Recognition-to-Cognition Networks that consider the grounding and contextual information for cognition level visual understanding. The model surpassed the previous SOTA developed on VQA [25]. Song et al. [150] proposed the Knowledge enhanced Visual and linguistic BERT that utilized the external commonsense Knowledge. This approach outperformed the BERT-based approaches and previous VCR models by marginal improvement.

4.2.6 Natural Language for Vision Recognition. The **NLVR (Natural Language Vision Recognition)** task checks the relationship consistency between the provided one text description and multiple images. Previous pretraining approaches [82, 151, 152] considered NLVR as a binary classifier. Vision Language Navigation is a similar task in which agents traverse the real-world dynamics by following linguistic instructions. Zhu et al. proposed **AuxRN (Auxiliary Reasoning Navigation)** [153] based on self-supervision that Observe Semantic Information from surrounding for Vision-Language Navigation. In other words, the model learns the implicit information from the environment, estimates the navigation, and predicts the next position.

4.3 Generation

In multimodal generation tasks, the data provided by the input modality is processed to generate another modality or paraphrased into the same modality as the output. The tasks concealed by generative category heavily rely on the previously mentioned understanding and classifications tasks. The image captioning task generates the text after classifying the object in the input image. Likewise, the text-to-image generation task understands the input text to propose a target image.

4.3.1 Visual Captioning. Visual captioning is the task of converting an image or video modality into a related text modality. The input is image pixels, and the output is syntactically and semantically meaningful text. Producing captions based on images bridges the gap between low and high semantic features found in image and text modalities. Some SOTA datasets in image captioning are MS-COCO [53] and Flickr30k [92]. The baseline and conventional visual captioning pipeline consist of an image encoder and a text decoder. Recent work leverages BERT capabilities, such as in [154]. BERT models [35] are initially trained using a large textual corpus, then fused textual and visual modalities embedding are finetuned using the right-masking pretraining technique. Vinyals et al. [155] presented an end-to-end system based on a neural network and combined SOFA

sub-networks for vision and language models. The method in this article significantly exceeds the performance of previous methods. In 2017, Chen et al. [156] improved CNN and proposed SCA-CNN based on images that CNN can extract: Spatial, Channel-wise, and Multi-layer. SCA-CNN can incorporate Spatial and Channel-wise attention in a CNN and dynamically modulate the sentence generation context in multi-layer feature maps. Unlike previous studies, Rennie et al. [157] used reinforcement learning to optimize image captioning systems. They proposed that optimizing the CIDEr using **metricSCST (self-critical sequence training)** is highly effective.

4.3.2   Visual Question Answering. **Visual Question Answering (VQA)** attempts to answer linguistic questions by retrieving information from visual cues [45]. VQA combines information from written questions and high-dimensional visual images or videos. Questions could vary from simple true/false to knowledge-based and open-ended questions, whereas visuals could vary from a simple sketch or image to a video. Furthermore, VQA combines functions from NLP and CV fields such as language understanding, relation extracting, attribute and object classifying, counting, knowledge-base, and commonsense reasoning [1]. Baseline VQA maps question text and visual embeddings obtained via recurrent and convolutional neural networks (RNN and CNN), respectively, to a common vector space. Mapping embedded representation to a shared vector space enables VQA to tackle open-ended free-form questions. In the literature, VQA deep learning techniques are classified as joint embedding models [158], attention mechanism [159], compositional model [160], **Graph Neural Networks (GNN)** [60] and knowledge base model [161]. There exist review papers targeting VQA as a research field of its own such as in [45] and [5]. In addition to surveys, benchmarks such as [162] by **Carnegie Mellon University (CMU)** and [163] are available. Benchmark datasets are VAQ v1.0 [5], VAQ-X [164], and VAQ-CP [165].

4.3.3   Visual Generation. Visual Generation, also known as text-to-image generation, typically uses the input text to generate images. Visual generation is the reverse direction of image captioning. In 2016, Reed et al. [166] first proposed deep convolutional **generative adversarial networks (GAN)** to synthesize images based on text descriptions. The training model is based on DC-GAN, but different from traditional GAN, the input of D is added with real image and false text description pairs. The author trained a CNN to predict style using an image generated by generator G. The predicted style can be used in the composition of G. The datasets used in this paper are: Caltech-UCSD Birds [167] dataset, Oxford-102 Flowers dataset, and MS COCO [53] dataset. Xu et al. [168] introduced **Attentional Generative Adversarial Network (AttnGAN)** for fine-grained text-to-image generation. AttnGAN, composed of an attentional generative network and **Deep Attentional Multimodal Similarity Model (DAMSM)**, focuses on relevant words to depict specific image subregions. DAMSM augments generator training with a fine-grained image-to-text matching loss. Comprehensive evaluation shows AttnGAN significantly outperforms previous GAN models. This method is evaluated on CUB [167] and COCO [53] datasets.

4.3.4   Audio-Visual Speech Synthesis. Prajwal et al. [129] presented the model synthesizing speech from lip movement in the presence of noise. The author released the benchmark Lip2Wav Dataset, as the previous work [169] was evaluated on small and limited vocabulary-based datasets [127, 170, 171]. The designed dataset enables the model to synthesize speech from unconstrained lip movements.

High-quality speech is generated [172] using a Tacotron [173] inspired decoder that produce melspectrogram from text inputs. It is conditioned on the face embeddings encoded in the previous representation, and outperformed previous work [174–176] on all objective metrics for lip-to-speech work especially on the TIMIT [171] dataset.

4.4 Retrieval

Retrieval systems take the input from the user as a query and provide the most relevant results by considering the contextual information. The evolution of multimedia-enhanced is in the form of well-established cross-modal tasks.

4.4.1 Referring Expression Comprehension. In **Referring Expression Comprehension (REC)** application, an expression is used to refer to and localize a target object in an image. The referring expressions are specific and detail the object properties and the relationship to its surroundings. Therefore, visual attributes, relationships, and contextual information need to be addressed. In [50], REC supervised DL architectures are categorized into one or two stages. In two-stage pipelines, the first stage generates and lists proposed objects based on the image. The second stage encodes the referring expression and computes a matching score between the proposed objects and encoded referring expressions. In one-stage pipelines, image and language features are concatenated and fed into the model in a single faster step. Some pretrained models for task-agnostic vision and language applications used for REC are VL-BERT [177] and ViLBERT [9].

4.4.2 Visual Language Retrieval. Indexing, query formulation, retrieval and evaluation are the steps required to build an **Information Retrieval (IR)** application. In indexing and query formulation, documents and user interfaced queries are represented by their characteristic features, respectively [178]. The retrieval system then maps both representations to retrieve or extract the required useful information. The performance of the retrieval task is evaluated based on recall and precision. In multimodal information retrieval, the system searches documents with different modalities such as text, images, videos, or physiological signals and images. Employing more than one modality enriches information retrieving processes. One example of multimodal information retrieval is in electronic health records.

4.4.3 Electronic Health Records. Patients health records contain various modalities such as categorical data, text, images such as MRI scans, or signals such as **electrocardiograms (ECG)** [178]. Information retrieval systems can extract information from different modalities to report, present, and/ or predict a patient health status. For instance, **Supervised Deep Patient Representation Learning Framework (SDPRL)** engages different modalities information to learn patient representation [179]. SDPRL is built and tested using the benchmark dataset MIMIC-III. Chen et al. [180] address the problem of **Major Depressive Disorder (MDD)** detection by proposing a novel **Graph Neural Network (GNN)**-based multimodal fusion strategy, coined as "modal-shared modal-specific GNN". This approach accounts for heterogeneity/homogeneity among psychophysiological modalities and inter/intramodal characteristics, and it employs a reconstruction network and attention mechanism to obtain a compact multimodal representation. Similarly, Multimodal Graph Neural Network framework was introduced to predict cancer survival using multimodal data such as gene expression, copy number alteration, and clinical data [181].

4.5 Multimodal Translation (MMT)

Visual modality has been considered marginally beneficial for machine translation due to the absence of sufficient features in the image. Caglayan et al. [182] proposed that visual information benefits **Machine Translation (MT)** when the source sentence lacks linguistic context. The author exploited the Degradation technique in the form of Color Deprivation, Entity Masking and Progressive Masking to degrade the source sentence. Additionally, unrelated images (violation of semantic compatibility) were fed to evaluate the visual sensitivity of the approach. ResNet-50 CNN based encoder [24] is used for visual features extraction. Multimodal Attention for **Neural Machine Translation (NMT)** inspired by [183] achieved the context vector from textual and

visual features. Multi30K and flicker30 were used for training. The authors presented a novel encoder for Neural Machine Translation (NMT) [60] that uses a graph-based approach for combining multiple modalities. This approach takes advantage of detailed semantic connections between these different data types to enhance the learning of combined representations.

Su et al. [184] introduced an approach that exploits the interaction of semantic representations across modalities, drawing inspiration from QA and NER domains. Two attention-based models were proposed: one using a bi-directional attention mechanism to learn from text and visual feature interaction, and the other incorporating a co-attention mechanism that generates context vectors from textual representation and subsequently uses it to update the text context vector with further interactive information. These models were evaluated on the expanded Multi30K dataset, showing superior performance over previous baselines, with or without pretraining.

5 DATASET

This section summarizes the benchmarks for the pretraining, finetuning, and evaluation of multimodal models. As per our knowledge, we cover all the benchmarks containing image, text, video, and audio modalities. The motive is to provide ready-to-use information to the new researchers regarding the nature of benchmarks, the number of samples in each benchmark, the usability for different tasks, and the comparative information as shown in Table 4. The MS COCO dataset [53] is used for image recognition, image detection, image segmentation, and image captioning. For each image in the dataset, five different descriptions are provided. This dataset contains 91 object types that would be easily recognizable. The dataset includes more than 300,000 images and 2.5 million labeled instances.

The Flickr30k [92] is collected from Flickr, together with five image descriptions provided by human annotators. The Flickr8k [185] is the smallest version, also collected from Flickr, that leads to training the model easier and faster. FUNSD [119] is a small dataset comprising 199 real, fully annotated, scanned forms. FUNSD contains 31,485 words, 9,707 semantic entities, and 5,304 relations and aims to extract and structure forms' textual content. SROIE [118] consists of a dataset with 1,000 whole scanned receipt images and annotations for the competition on **scanned receipts OCR and key information extraction (SROIE)**. **CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI)** dataset [186] is the largest dataset of multimodal sentiment analysis and emotion recognition to date. In 2018, the CMU-MOSEI dataset was created, consisting of 23,453 annotated video clips. These clips were sourced from 1,000 unique YouTube videos that covered 250 different topics.

VQA (Visual Question Answering) [25] is a large dataset containing more than 250,000 images with at least three questions per image and 10 ground-truth answers per question. VQA v1 provides 6,141,630 ground-truth answers and 1,842,489 plausible answers, making it harder for the model to answer the questions correctly. **VAQ-CP (Visual Question Answering under Changing Priors)** is the splits of the VQA v1 and VQA v2 datasets. CrisisMMD [140] is a sizeable multimodal dataset of natural disasters collected from Twitter, including earthquakes, hurricanes, wildfires, and floods that happened in the year 2017 across different parts of the world. It has three types of annotations. The first one is "Informative" or "Not Informative", which determines whether it is helpful for humanitarian aid. The second type is "Humanitarian Categories", such as infrastructure and utility damage, vehicle damage, and the like. The last type is "Damage Severity Assessment", which describes the severity of the damage.

MIMIC (Medical Information Mart for Intensive Care) [187] is a publicly available dataset developed by the Laboratory for Computational Physiology that comprises de-identified health data associated with thousands of intensive care unit admissions. Currently, it has three versions: MIMIC-II, MIMIC-III, and MIMIC-IV. MIMIC-III contains information about 53,423 adults admitted

Table 4. The Benchmarks' Information for the Tasks in this Study

Dataset	Modality	Count	Tasks	Data Category
MS COCO [53]	text + image	328k	Image Recog., IC	Multiple
Conceptual Captions [90]	text + image	3.3M	VQA, VCR, IR, IC	Multiple
Flickr30K [92]	text + image	31k	IR, CMR, IC	Multiple
Flickr8K [185]	text + image	8k	IR, CMR, IC	Multiple
FUNSD [119]	text + image	199	TD, OCR, EL	Scanned forms
SROIE [118]	text + image	1k	OCR, IE	Scanned images
CMU-MOSEI [186]	text + video	23,453	MMSA	YouTube videos
VQA [25]	text + image	250k	VQA	Multiple
CrisisMMD [140]	text + image	16,097	IC, ED	Disasters tweets
MIMIC [187]	MML data	53,423	IR	Patients Information
Fashion-200K [188]	text + image	200k	IC, IR	Clothes
NYU Depth v1 [189], v2 [190]	RGB + Depth	4GB, 90GB	IS, IC	Indoor Scenes
SKIG [74]	text + image	2160	GR	Hand gestures
GoodNews [191]	text + image	466,000	IC	News
MSR-VTT [192]	text + video	200k	VU	Commercial videos
MSVD-QA [193]	text + video	2089	VR, VQA, VC	Multiple
TGIF-QA [194]	text + image (gif)	103,919	VQA	Multiple
EQA-v1 [195]	text + 3D env.	9,000	EQA, VQA	Multiple
VideoNavQA [196]	text + 3D env.	28	EQA, VQA	Multiple
TDIUC [197]	text + image	1,654,167	VQA	Multiple
nuScenes [198]	image + sensor data	1.4M	IR, Auton. Driving	Driving Information
CUB-200 [167]	text + image	11,788	VG	Birds
Oxford-102 Flowers [199]	text + image	4,080-26,316	VG	Flowers
VCR [149]	text + image	263k	VU	Movie Scenes
How2 [125]	audio + text + video	2000 hours	MML	Instructional Videos
Lip2Wav [129]	audio + video	120 hours	AVSS	Talking Face Videos
MUSIC [133]	audio + video	685	SSS	Videos of Music
MOSI [200]	audio + text + video	3702	Multimodal analysis	YouTube videos
GRID [201]	audio + video	33,000	VSR	Speech

The Task column lists the tasks in the corresponding papers.

to critical care units from 2001 to 2012, such as the patient's gender, height, and other essential information, such as blood routine, liver function, and other hospital test data, as well as medication information.

Fashion200K [188] contains 200K fashion images, and each image comes with a compact attribute-like product description. MIT-Stata Center [202] is a multimodal dataset containing vision (stereo and RGB-D), laser and proprioceptive data. This dataset comprises over 2.3 TB, 38 h and 42 km. This dataset also includes ground-truth position estimates of the robot at every instance. This is an instrumental dataset for robotic mapping and CV research. The NYU-Depth dataset consists of video sequences recorded by the RGB and Depth cameras from the Microsoft Kinect. NYU-Depth v1 [189] provides about 4GB of labeled data and about 90GB of raw data. NYU-Depth v1 includes 64 different indoor scenes and 7 scene types. NYU-Depth v2 [190] includes 464 different indoor scenes and 26 scene types.

The **Shefeld Kinect Gesture (SKIG)** [74] is a gesture dataset containing 2,160 hand gesture sequences. These hand gestures can be classified into 10 categories: circle (clockwise), triangle (anticlockwise), up-down, right-left, wave, "Z", cross, come-here, turn-around, and pat. The sequences are recorded under three different backgrounds and two illumination conditions, which provide diversity. GoodNews [191] is a large dataset containing 466,000 images with captions, headlines, and text articles. However, different from datasets like MSCOCO or Flicker8k, GoodNews includes a single ground truth caption per image. GoodNews captions written by expert journals have a longer average length than generic captioning datasets, meaning that these captions are more descriptive.

MSR-VTT [192] is a large-scale video description dataset: 10K web video clips with 38.7 hours and 200K clip-sentence pairs. MSR-VTT was created from 257 popular queries from a commercial video search engine. Each clip in MSR-VTT is annotated with approximately 20 natural sentences. This dataset is presented for video understanding. The **Microsoft Research Video Description Corpus (MSVD)** dataset MSVD-QA contains 122K descriptions of 2,089 short video clips (usually less than 10 seconds). The MSVD dataset contains different language descriptions, such as English, Hindi, Romanian, Slovene, and more. MSVD-QA is a benchmark for video retrieval, visual question answering, and video captioning.

TGIF-QA [194] is a large-scale dataset containing 103,919 QA pairs collected from 56,720 animated GIFs. These GIFs are from the TGIF dataset. The TGIF dataset is based on GIFS data as GIFs have a concise format and cohesive storytelling. TGIF-QA can be used for visual question-answering research. **EQA (Embodied Question Answering)** v1.0 [195] is a dataset containing 9,000 questions from 774 environments. The visual questions and answers in this dataset are grounded in House3D. EQA-v1 contains location, color, and place preposition questions.

VideoNavQA [196] is also a dataset used to study the EQA task. VideoNavQA contains 28 questions belonging to eight categories with 70 possible answers. The complexity of the questions in VideoNavQA far exceeds that of similar tasks that use generation methods that extract ground truth information from the video to generate questions. **Task Directed Image Understanding Challenge (TDIUC)** [197] is a dataset containing 167,437 images and 1,654,167 question-answer pairs. TDIUC divides VQA into 12 constituent tasks, which makes it easier to measure and compare the performance of VQA algorithms. The 12 different question types are grouped according to these tasks.

nuScenes [198] is a large-scale public dataset for an autonomous driving dataset with 3D object annotations. It is also a multimodal dataset. nuScenes provides 1.4 million camera images, 1500h of driving data from four cities (Boston, Pittsburgh, Las Vegas and Singapore), sensor data released for 150h (5x LIDAR, 8x camera, IMU, GPS), detailed map information, 1.4M 3D bounding boxes manually annotated for 23 object classes, and so on. nuScenes can be used for intelligent agent research. nuImages is a large-scale autonomous driving dataset with image-level 2D annotations. It has 93k video clips of 6s each, 93k annotated and 1.1M un-annotated images. The Caltech-UCSD Birds dataset (CUB-200) [167] includes images of 200 different bird species: a total of 11,788 images. Every image in this dataset has annotations that include a bounding box around the bird, a basic segmentation of the bird, and labels for various attributes. The Oxford-102 Flowers dataset features images from 102 bird categories, with each category containing between 40 and 258 images. These images display a wide range of sizes, poses, and lighting conditions.

Visual Commonsense Reasoning (VCR) [149] contains over 212K (training), 26K (validation), and 25K (testing) questions, answers, and rationales derived from 110K movie scenes. It is widely used for cognition-level visual understanding. How2 [125] is a multimodal dataset containing instructional videos with English subtitles and crowdsourced Portuguese translations. How2 covers a wide variety of topics across 80,000 clips (about 2,000 hours). Lip2Wav [129] dataset contains 120 hours of talking face videos across five speakers. This dataset has about 20 hours of natural speech per speaker and vocabulary sizes of over 5,000 words for each. **MUSIC (Multimodal Sources of Instrument Combinations)** dataset [133] contains 685 untrimmed videos of musical solos and duets. The dataset spans 11 instrument categories: accordion, acoustic guitar, cello, clarinet, erhu, flute, saxophone, trumpet, tuba, violin and xylophone. **MOSI** [200]**(Multimodal Opinionlevel Sentiment Intensity)** dataset contains 93 videos and 3,702 video segments. The dataset provides not only sentiment annotations, but also manual gesture annotations. GRID [201] consists of video recordings from 54 speakers, with 100 utterances per talker, 33,000 utterances in total.

6 FUTURE FORECASTING

This section describes the techniques researchers have observed from previous studies to adopt and consider for achieving better results and reducing the computational cost. Additionally, these approaches have not been considered or explored thoroughly in recent studies even after expressing efficient performance in relevant research.

Building Unified Models. Previous studies proved the effectiveness of transformers for NLP, CV, and multimodal tasks. Recently, researchers have presented a unified architecture with a single agent transformer serving as the backbone and deal with all modalities for multiple tasks [203]. Even though the results are impressive on downstream tasks, still could not surpass task-specific models for some of the tasks [204]. For zero-shot learning, the performance highly depends on instructions. More optimized instructions can lead to more satisfactory results. Researchers are trying to address the challenge of sensitivity that the model expresses with the slight changes in prompts and parameters.

Leveraging GNNs for multimodal pretraining. Ektefai et al. [205] proposed Multimodal Graph Learning (MGL) to manage multimodal data input and produce a common output representation for a variety of downstream tasks. This methodology initially identifies relevant entities across data modalities, projecting them into a shared namespace, and subsequently combines these different modalities. The process applies a message-passing module to learn node representation based on intra-modality and inter-modality adjacency matrices. Finally, through an aggregation process, it generates node, subgraph, or graph-level representations suitable for downstream tasks.

Pretraining Strategy. Developing an optimal set of pretraining objectives is worth exploring for multimodal tasks. Pretext tasks directly influence the performance for downstream tasks. While selecting pretraining objectives, we need to consider modalities, datasets, network architecture, and downstream tasks. There is a potential to apply efficient stage-wise pretraining for multimodal methods as Bao et al. [33] successfully adopt it for a single modality.

Enhancing Pretraining with Multilingual Features. Multilingual pretraining of multimodal systems is less examined as English-only multimodal benchmarks are experimented heavily. UC2 [84], M3P [206], and MURAL [207] exploit multilingual features by adding a separate encoder for multilingual or translated data. CCLM [208] based on ALBEF [209] proposed an approach that outperformed the SOTA in a zero-shot cross-lingual set-up. In short, the multilingual perspective is essential for enhancing the scaling success of multimodal pretraining frameworks [210].

Prompt Tuning. Prompt tuning received attention after GPT-2 and GPT-3. This technique converts implicit information to a question or descriptive text, which assists in exploiting the textual probability of language models. Prompt learning with a pretraining objective (MLM) can address two challenges of finetuning: (i) reduce the computational complexity by avoiding the different parameters for each downstream task, and (ii) reduce the gap between the representation learning of pretraining tasks and downstream tasks [211]. This technique is not explored for multimodal approaches yet. Technically, prompt tuning may transform the unifying architecture to a more normalized and generalized version.

Masking Stage. The interaction of different modalities faces alignment challenges. Masking is an effective strategy used to address alignment problem by generating text and images. Masking can be applied at different stages, e.g., input level and task level, but according to Zhuge et al. [212], embedding level masking is the most effective stage for aligning the representation of different modalities. Fine-grain representation can be learned if we focus on the optimal alignment among audio, video and text modalities.

Knowledge Infusion. There is a need to fuse external knowledge in multimodal representation for visual, language and audio to make the model well-informed, as Chen et al. [213] utilized illustrative knowledge for vision and language. More intelligent architectures are required that

integrate knowledge base features for multimodal pretraining and downstream tasks. Knowledge Distillation enhanced the performance of ViT-BERT on all tasks by offering effective learning representation compared to other variants that do not exploit it. For SST-2 and IN-1K, accuracy improved 10% and 5%, respectively. Therefore, while designing pretraining objectives and cognitive architectures, commonsense, situational knowledge, hierarchical, structural, or network information must be considered.

Considering Anticipated Evaluation. Deep learning architectures need high computational resources for experiments. Current evaluation methods provide information about the effectiveness of the models after extensive experiments. For large-scale multimodal methods, there must be evaluation strategies to examine the model at an earlier stage [42] to verify whether models are compatible with downstream tasks before paying the computation cost.

The Acceleration and Scaling of Multimodal Architectures. Aside from knowledge distillation [214], pruning and quantization are still unexplored to compress and accelerate the model and improve the cross-modal inference speed. Inspired by the performance of sizeable pretrained language models, efforts are made to achieve success on multimodal tasks through extensive training and complex architectures. However, well-developed benchmarks and models are required to cope with the multimodal at a large scale with higher proficiency [215, 216].

Including Audio Stream. Recent multimodal surveys [41, 42] specifically for pretraining have not considered the audio stream. Audio can be semantically rich like text and can provide emotional and supplementary information about the speaker, including the boundary data in the case of multiple speakers. Pretraining with audio is extremely important for unifying architectures as it makes the model capable of performing downstream tasks that contain the audio stream [217]. However, the challenges of alignment and correspondence increase with multiple modalities.

7 CONCLUSION

This survey covered the role of deep multimodal learning and architectures to effectively deal with advanced multimodal tasks. We started with deep learning-based task-specific multimodal architectures based on encoder-decoder, attention, and reinforcement learning. Then, we covered the advancements in hardware, large-scale computation resources, and pretraining approaches. Pretraining and finetuning alleviate various challenges of multimodal and cross-modal tasks, including multitasking. Therefore, we discussed the types and tasks of pretraining, which make the training procedure effective, and SOTA transformer-based pretraining approaches that produce high impact recently, summarize comparatively. We further showed that the research community focused on pretraining at a large scale to create unified and generic architectures using a transformer as the backbone to perform a more complex task with identical parameters. We covered SOTA multimodal architectures for different multimodal applications and their performance on benchmarks. The applications showed the usefulness and the practicality of multimodal systems. Multimodal possibilities are enormous, yet so far, the capabilities are explored only for the English language. A possible suggested future research trajectory is to construct multimodal datasets in other languages and build multilingual frameworks.

REFERENCES

[1] Jabeen Summaira, Xi Li, Amin Muhammad Shoib, Songyuan Li, and Jabbar Abdul. 2021. Recent advances and trends in multimodal deep learning: A review. *arXiv preprint arXiv:2105.11087* (2021).

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41, 2 (2018), 423–443.

[3] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. 2022. Learning audio-visual speech representation by masked multimodal cluster prediction. In *Proceedings of the International Conference on Learning Representations*.

[4] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting propaganda techniques in memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 6603–6617.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[6] Wenzhong Guo, Jianwen Wang, and Shiping Wang. 2019. Deep multimodal representation learning: A survey. *IEEE Access* 7 (2019), 63373–63394.

[7] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Abdellatif Mtibaa. 2021. A survey on deep multimodal learning for computer vision: Advances, trends, applications, and datasets. *The Visual Computer* (2021), 1–32.

[8] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. 2019. VisualBERT: A simple and performant baseline for vision and language. *arXiv preprint arXiv:1908.03557* (2019).

[9] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in Neural Information Processing Systems* 32 (2019).

[10] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10437–10446.

[11] Harry McGurk and John MacDonald. 1976. Hearing lips and seeing voices. *Nature* 264, 5588 (1976), 746–748.

[12] Pradeep K. Atrey, M. Anwar Hossain, Abdulmotaleb El Saddik, and Mohan S. Kankanhalli. 2010. Multimodal fusion for multimedia analysis: A survey. *Multimedia Systems* 16 (2010), 345–379.

[13] Georgios Evangelopoulos, Athanasia Zlatintsi, Alexandros Potamianos, Petros Maragos, Konstantinos Rapantzikos, Georgios Skoumas, and Yannis Avrithis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia* 15, 7 (2013), 1553–1568.

[14] Rainer W. Lienhart. 1998. Comparison of automatic shot boundary detection algorithms. In *Storage and Retrieval for Image and Video Databases VII*, Vol. 3656. SPIE, 290–301.

[15] Jean Carletta, Simone Ashby, Sebastien Bourban, Mike Flynn, Mael Guillemot, Thomas Hain, Jaroslav Kadlec, Vasilis Karaiskos, Wessel Kraaij, Melissa Kronenthal, Guillaume Lathoud, Mike Lincoln, Masson Agnes Lisowska, Iain McCowan, Wilfried Post, Dennis Reidsma, and Pierre D. Wellner. 2006. The AMI meeting corpus: A pre-announcement. In *Machine Learning for Multimodal Interaction: Second International Workshop (MLMI 2005, Edinburgh, UK, July 11-13, 2005, Revised Selected Papers 2)*, Springer, 28–39.

[16] Gary McKeown, Michel F. Valstar, Roderick Cowie, and Maja Pantic. 2010. The SEMAINE corpus of emotionally coloured character interactions. In *2010 IEEE International Conference on Multimedia and Expo*. IEEE, 1079–1084.

[17] Björn Schuller, Michel Valstar, Florian Eyben, Gary McKeown, Roddy Cowie, and Maja Pantic. 2011. AVEC 2011–the first international audio/visual emotion challenge. In *Affective Computing and Intelligent Interaction: Fourth International Conference, ACII 2011, Memphis, TN, USA, October 9–12, 2011, Proceedings, Part II*. Springer, 415–424.

[18] Michel Valstar, Björn Schuller, Kirsty Smith, Timur Almaev, Florian Eyben, Jarek Krajewski, Roddy Cowie, and Maja Pantic. 2014. AVEC 2014: 3D dimensional affect and depression recognition challenge. In *Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge*. 3–10.

[19] Guanzheng Chen, Fangyu Liu, Zaiqiao Meng, and Shangsong Liang. 2022. Revisiting parameter-efficient tuning: Are we really there yet?. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2612–2626.

[20] Nikhil Rasiwasia, Jose Costa Pereira, Emanuele Coviello, Gabriel Doyle, Gert R. G. Lanckriet, Roger Levy, and Nuno Vasconcelos. 2010. A new approach to cross-modal multimedia retrieval. In *Proceedings of the 18th ACM International Conference on Multimedia*. 251–260.

[21] Amirhossein Habibian, Thomas Mensink, and Cees G. M. Snoek. 2016. Video2vec embeddings recognize events when examples are scarce. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 39, 10 (2016), 2089–2103.

[22] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

[23] Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. *International Conference on Learning Representations* (2015).

[24] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 770–778.

[25] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

[26] Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation* 9, 8 (1997), 1735–1780.

[27] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in Neural Information Processing Systems* 30 (2017).

[28] Zhe Gan, Linjie Li, Chunyuan Li, Lijuan Wang, Zicheng Liu, and Jianfeng Gao. 2022. Vision-language pre-training: Basics, recent advances, and future trends. *Foundations and Trends® in Computer Graphics and Vision* 14, 3-4 (2022), 163–352.

[29] Allan Jabri, Armand Joulin, and Laurens van der Maaten. 2016. Revisiting visual question answering baselines. In *European Conference on Computer Vision.* Springer, 727–739.

[30] Duy-Kien Nguyen and Takayuki Okatani. 2018. Improved fusion of visual and language representations by dense symmetric co-attention for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition.* 6087–6096.

[31] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. 2019. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision.* 10294–10303.

[32] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven C. H. Hoi, Xiaogang Wang, and Hongsheng Li. 2019. Dynamic fusion with intra-and inter-modality attention flow for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 6639–6648.

[33] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, Songhao Piao, and Furu Wei. 2022. VLMo: Unified vision-language pre-training with mixture-of-modality-experts. *Advances in Neural Information Processing Systems* 35 (2022), 32897–32912.

[34] Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. ViLT: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning.* PMLR, 5583–5594.

[35] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the NAACL-HLT.* 4171–4186.

[36] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2017. ImageNet classification with deep convolutional neural networks. *Commun. ACM* 60, 6 (2017), 84–90.

[37] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems* 33 (2020), 12449–12460.

[38] Anil Rahate, Rahee Walambe, Sheela Ramanna, and Ketan Kotecha. 2022. Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions. *Information Fusion* 81 (2022), 203–239.

[39] Jing Gao, Peng Li, Zhikui Chen, and Jianing Zhang. 2020. A survey on deep learning for multimodal data fusion. *Neural Computation* 32, 5 (2020), 829–864.

[40] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. VLP: A survey on vision-language pre-training. *Machine Intelligence Research* 20, 1 (2023), 38–56. DOI : http://dx.doi.org/10.1007/s11633-022-1369-5

[41] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. 2022. A survey of vision-language pre-trained models. In *Proceedings of the Thirty-first International Joint Conference on Artificial Intelligence (IJCAI-22) Survey Track.*

[42] Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. 2022. Vision-and-language pretrained models: A survey. In *Proceedings of the Thirty-first International Joint Conference on Artificial Intelligence, IJCAI 2022*, Luc De Raedt (Ed.). ijcai.org, 5530–5537.

[43] Dhanesh Ramachandram and Graham W. Taylor. 2017. Deep multimodal learning: A survey on recent advances and trends. *IEEE Signal Processing Magazine* 34, 6 (2017), 96–108.

[44] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2021. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Int. Res.* 71 (Sep. 2021), 1183–1317. DOI : http://dx.doi.org/10.1613/jair.1.11688

[45] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony Dick, and Anton van den Hengel. 2017. Visual question answering: A survey of methods and datasets. *Computer Vision and Image Understanding* 163 (2017), 21–40.

[46] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. 2022. AMMU: A survey of transformer-based biomedical pretrained language models. *Journal of Biomedical Informatics* 126 (2022), 103982.

[47] Kejing Xiao, Zhaopeng Qian, and Biao Qin. 2022. A survey of data representation for multi-modality event detection and evolution. *Applied Sciences* 12, 4 (2022), 2204.

[48] Lukas Stappen, Alice Baird, Lea Schumann, and Schuller Bjorn. 2021. The multimodal sentiment analysis in car reviews (muse-car) dataset: Collection, insights and improvements. *IEEE Transactions on Affective Computing* (2021).

[49] Ganesh Chandrasekaran, Tu N. Nguyen, and Jude Hemanth D. 2021. Multimodal sentimental analysis for social media applications: A comprehensive review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 11, 5 (2021), e1415.

[50] Yanyuan Qiao, Chaorui Deng, and Qi Wu. 2021. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia* 23 (2021), 4426–4440. DOI : http://dx.doi.org/10.1109/TMM.2020.3042066

[51] Jie Wu and Haifeng Hu. 2017. Cascade recurrent neural network for image caption generation. *Electronics Letters* 53, 25 (2017), 1642–1643.

[52] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 1655–1663.

[53] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.

[54] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y. Ng. 2011. Multimodal deep learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML'11)*. 689–696.

[55] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. 2008. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*. 1096–1103.

[56] Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. 2022. Multi-modal knowledge graph construction and application: A survey. *IEEE Transactions on Knowledge and Data Engineering* (2022).

[57] Yujia Li, Richard Zemel, Marc Brockschmidt, and Daniel Tarlow. 2016. Gated graph sequence neural networks. In *Proceedings of International Conference on Learning Representations*.

[58] Petar Velickovic, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. 2017. Graph attention networks. *Stat* 1050, 20 (2017), 10–48550.

[59] Yongjing Yin, Fandong Meng, Jinsong Su, Chulun Zhou, Zhengyuan Yang, Jie Zhou, and Jiebo Luo. 2020. A novel graph-based multi-modal fusion encoder for neural machine translation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 3025–3035.

[60] Difei Gao, Ke Li, Ruiping Wang, Shiguang Shan, and Xilin Chen. 2020. Multi-modal graph neural network for joint reasoning on vision and scene text. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12746–12756.

[61] Yanan Wang, Michihiro Yasunaga, Hongyu Ren, Shinya Wada, and Jure Leskovec. 2022. VQA-GNN: Reasoning with multimodal semantic graph for visual question answering. *arXiv preprint arXiv:2205.11501* (2022).

[62] Weitao Jiang, Xiying Li, Haifeng Hu, Qiang Lu, and Bohong Liu. 2021. Multi-gate attention network for image captioning. *IEEE Access* 9 (2021), 69700–69709. DOI : http://dx.doi.org/10.1109/ACCESS.2021.3067607

[63] Kan Chen, Trung Bui, Chen Fang, Zhaowen Wang, and Ram Nevatia. 2017. AMC: Attention guided multi-modal correlation learning for image search. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2644–2652.

[64] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. 2018. Video captioning via hierarchical reinforcement learning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4213–4222.

[65] Jinyuan Fang, Shangsong Liang, Zaiqiao Meng, and Qiang Zhang. 2021. Gaussian process with graph convolutional kernel for relational learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 353–363.

[66] Guanzheng Chen, Jinyuan Fang, Zaiqiao Meng, Qiang Zhang, and Shangsong Liang. 2022. Multi-relational graph representation learning with Bayesian Gaussian process network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 5530–5538.

[67] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. VinVL: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5579–5588.

[68] Junyang Lin, Rui Men, An Yang, Chang Zhou, Yichang Zhang, Peng Wang, Jingren Zhou, Jie Tang, and Hongxia Yang. 2021. M6: Multi-modality-to-multi-modality multitask mega-transformer for unified pretraining. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 3251–3261.

[69] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. 2020. ELECTRA: Pre-training text encoders as discriminators rather than generators. In *Proceedings of the International Conference on Learning Representations*. OpenReview.net.

[70] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692* (2019).

[71] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 36, 4 (2020), 1234–1240.

[72] Tommaso Caselli, Valerio Basile, Jelena Mitrović, and Michael Granitzer. 2021. HateBERT: Retraining BERT for abusive language detection in English. In *Proceedings of the 5th Workshop on Online Abuse and Harms (WOAH'21)*. 17–25.

[73] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, He-Yan Huang, and Ming Zhou. 2021. InfoXLM: An information-theoretic framework for cross-lingual language model pre-training. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3576–3588.

[74] Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. 2020. Pre-training technique to localize medical BERT and enhance biomedical BERT. *arXiv preprint arXiv:2005.07202* (2020).

[75] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 8342–8360.

[76] Yujia Qin, Yankai Lin, Jing Yi, Jiajie Zhang, Xu Han, Zhengyan Zhang, Yusheng Su, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. 2022. Knowledge Inheritance for Pre-trained Language Models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 3921–3937.

[77] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *The University of British Columbia* (2018).

[78] Subhadarshi Panda, Anjali Agrawal, Jeewon Ha, and Benjamin Bloch. 2021. Shuffled-token detection for refining pre-trained RoBERTa. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. 88–93.

[79] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. 2021. Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 11782–11791.

[80] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. 2020. InterBERT: Vision-and-language interaction for multi-modal pretraining. *arXiv preprint arXiv:2003.13198* (2020).

[81] Yongfei Liu, Chenfei Wu, Shao-Yen Tseng, Vasudev Lal, Xuming He, and Nan Duan. 2022. KD-VLP: Improving end-to-end vision-and-language pretraining with object knowledge distillation. In *Findings of the Association for Computational Linguistics: NAACL 2022*. 1589–1600.

[82] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. UNITER: Universal image-text representation learning. In *European Conference on Computer Vision*. Springer, 104–120.

[83] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. HERO: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2046–2065.

[84] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. 2021. UC2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4155–4165.

[85] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the International Conference on Learning Representations*.

[86] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research* 21, 1 (2020), 5485–5551.

[87] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. 2020. End-to-end object detection with transformers. In *European Conference on Computer Vision*. Springer, 213–229.

[88] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. 2021. Deformable DETR: Deformable transformers for end-to-end object detection. In *International Conference on Learning Representations*.

[89] 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. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning, PMLR*, 8748–8763.

[90] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2556–2565.

[91] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13041–13049.

[92] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE International Conference on Computer Vision*. 2641–2649.

[93] 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. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, Springer, 121–137.

[94] Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11162–11173.

[95] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. ERNIE-ViL: Knowledge enhanced vision-language representations through scene graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 35. 3208–3216.

[96] Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 2066–2080.

[97] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (Proceedings of Machine Learning Research)*, Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato (Eds.), Vol. 162. PMLR, 12888–12900. https://proceedings.mlr.press/v162/li22n.html

[98] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597* (2023).

[99] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 2021. HuBERT: How much can a bad teacher benefit ASR pre-training?. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 6533–6537.

[100] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in Neural Information Processing Systems* 26 (2013).

[101] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. 2018. AllenNLP: A deep semantic natural language processing platform. In *Proceedings of Workshop for NLP Open Source Software (NLP-OSS)*. 1–6.

[102] Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 5185–5198.

[103] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph P. Turian. 2020. Experience Grounds Language. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP'20)*. 8718–8735.

[104] 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 Li Fei-Fei. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* 123, (2017), 32–73.

[105] Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. YFCC100M: The new data in multimedia research. *Commun. ACM* 59, 2 (2016), 64–73.

[106] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 8948–8957.

[107] Qing Li, Boqing Gong, Yin Cui, Dan Kondratyuk, Xianzhi Du, Ming-Hsuan Yang, and Matthew Brown. 2021. Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text. *arXiv preprint arXiv:2112.07074* (2021).

[108] Ronghang Hu and Amanpreet Singh. 2021. Unit: Multimodal multitask learning with a unified transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1439–1449.

[109] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. 2021. VATT: Transformers for multimodal self-supervised learning from raw video, audio and text. *Advances in Neural Information Processing Systems* 34 (2021).

[110] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. OFA: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*. PMLR, 23318–23340.

[111] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7871–7880.

[112] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. *arXiv preprint arXiv:2305.06500* (2023).

[113] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai-hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022).

[114] Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yong-hao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An Open-Source Chatbot Impressing GPT-4 with 90%* ChatGPT Quality. (March 2023). https://lmsys.org/blog/2023-03-30-vicuna/

[115] Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 2021. Self-supervised multimodal opinion summarization. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*.

[116] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei Florencio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. 2021. LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. 2579–2591.

[117] Yulin Li, Yuxi Qian, Yuechen Yu, Xiameng Qin, Chengquan Zhang, Yan Liu, Kun Yao, Junyu Han, Jingtuo Liu, and Errui Ding. 2021. StrucTexT: Structured text understanding with multi-modal transformers. In *Proceedings of the 29th ACM International Conference on Multimedia*. 1912–1920.

[118] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. 2019. ICDAR2019 competition on scanned receipt OCR and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*. IEEE, 1516–1520.

[119] Guillaume Jaume, Hazim Kemal Ekenel, and Jean-Philippe Thiran. 2019. FUNSD: A dataset for form understanding in noisy scanned documents. In *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, Vol. 2. IEEE, 1–6.

[120] Zhangxuan Gu, Changhua Meng, Ke Wang, Jun Lan, Weiqiang Wang, Ming Gu, and Liqing Zhang. 2022. XYLayoutLM: Towards layout-aware multimodal networks for visually-rich document understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 4583–4592.

[121] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. 2021. Do we really need explicit position encodings for vision transformers. *arXiv preprint arXiv:2102.10882* 3, 8 (2021).

[122] Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 2020. Multistage fusion with forget gate for multimodal summarization in open-domain videos. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 1834–1845.

[123] Shruti Palaskar, Jindřich Libovický, Spandana Gella, and Florian Metze. 2019. Multimodal abstractive summarization for how2 videos. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 6587–6596.

[124] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Ngan Fung. 2021. Vision guided generative pre-trained language models for multimodal abstractive summarization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*.

[125] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. 2018. How2: A large-scale dataset for multimodal language understanding. In *Conference on Neural Information Processing Systems*.

[126] Shaoshi Ling and Yuzong Liu. 2020. DeCoAR 2.0: Deep contextualized acoustic representations with vector quantization. *arXiv preprint arXiv:2012.06659* (2020).

[127] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. 2018. LRS3-TED: A large-scale dataset for visual speech recognition. *arXiv preprint arXiv:1809.00496* (2018).

[128] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. 2019. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*. IEEE, 905–912.

[129] K. R. Prajwal, Rudrabha Mukhopadhyay, Vinay P. Namboodiri, and C. V. Jawahar. 2020. Learning individual speaking styles for accurate lip to speech synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 13796–13805.

[130] Robert Rehr and Timo Gerkmann. 2017. On the importance of super-Gaussian speech priors for machine-learning based speech enhancement. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 26, 2 (2017), 357–366.

[131] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. 2001. Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 6 (2001), 681–685.

[132] Lingyu Zhu and Esa Rahtu. 2021. Leveraging category information for single-frame visual sound source separation. In *2021 9th European Workshop on Visual Information Processing (EUVIP)*. IEEE, 1–6.

[133] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. 2018. The sound of pixels. In *Proceedings of the European Conference on Computer Vision (ECCV)*. 570–586.

[134] Erin Hea-Jin Kim, Yoo Kyung Jeong, Yuyoung Kim, Keun Young Kang, and Min Song. 2016. Topic-based content and sentiment analysis of Ebola virus on Twitter and in the news. *Journal of Information Science* 42, 6 (2016), 763–781.

[135] Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. 40–46.

[136] Benjamin Wood, Owain Williams, Vijaya Nagarajan, and Gary Sacks. 2021. Market strategies used by processed food manufacturers to increase and consolidate their power: A systematic review and document analysis. *Globalization and Health* 17, 1 (2021), 1–23.

[137] Minping Chen and Xia Li. 2020. SWAFN: Sentimental words aware fusion network for multimodal sentiment analysis. In *Proceedings of the 28th International Conference on Computational Linguistics*. 1067–1077.

[138] Linmei Hu, Bin Zhang, Lei Hou, and Juanzi Li. 2017. Adaptive online event detection in news streams. *Knowledge-Based Systems* 138 (2017), 105–112.

[139] Nilani Algiriyage, Raj Prasanna, Kristin Stock, Emma E. H. Doyle, and David Johnston. 2022. Multi-source multimodal data and deep learning for disaster response: A systematic review. *SN Computer Science* 3, 1 (2022), 1–29.

[140] Firoj Alam, Ferda Ofli, and Muhammad Imran. 2018. CrisisMMD: Multimodal Twitter datasets from natural disasters. In *Twelfth International AAAI Conference on Web and Social Media*.

[141] Qi Chen, Wei Wang, Kaizhu Huang, Suparna De, and Frans Coenen. 2021. Multi-modal generative adversarial networks for traffic event detection in smart cities. *Expert Systems with Applications* 177 (2021), 114939.

[142] Alberto Barrón-Cedeno, Israa Jaradat, Giovanni Da San Martino, and Preslav Nakov. 2019. Proppy: Organizing the news based on their propagandistic content. *Information Processing & Management* 56, 5 (2019), 1849–1864.

[143] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5636–5646.

[144] Zhiwei Jin, Juan Cao, Han Guo, Yongdong Zhang, and Jiebo Luo. 2017. Multimodal fusion with recurrent neural networks for rumor detection on microblogs. In *Proceedings of the 25th ACM International Conference on Multimedia*. 795–816.

[145] Xinyi Zhou, Jindi Wu, and Reza Zafarani. 2020. Similarity-aware multi-modal fake news detection. In *Advances in Knowledge Discovery and Data Mining: 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11–14, 2020, Proceedings, Part II*. Springer, 354–367.

[146] Yaqing Wang, Fenglong Ma, Zhiwei Jin, Ye Yuan, Guangxu Xun, Kishlay Jha, Lu Su, and Jing Gao. 2018. EANN: Event adversarial neural networks for multi-modal fake news detection. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. 849–857.

[147] Dhruv Khattar, Jaipal Singh Goud, Manish Gupta, and Vasudeva Varma. 2019. MVAE: Multimodal variational autoencoder for fake news detection. In *The World Wide Web Conference*. 2915–2921.

[148] Jingzi Wang, Hongyan Mao, and Hongwei Li. 2022. FMFN: Fine-grained multimodal fusion networks for fake news detection. *Applied Sciences* 12, 3 (2022), 1093.

[149] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6720–6731.

[150] Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. 2021. KVL-BERT: Knowledge enhanced visual-and-linguistic BERT for visual commonsense reasoning. *Knowledge-Based Systems* 230 (2021), 107408.

[151] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5100–5111.

[152] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. 2020. Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers. *arXiv preprint arXiv:2004.00849* (2020).

[153] Fengda Zhu, Yi Zhu, Xiaojun Chang, and Xiaodan Liang. 2020. Vision-language navigation with self-supervised auxiliary reasoning tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10012–10022.

[154] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and VQA. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 34. 13041–13049.

[155] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

[156] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 2017. SCA-CNN: Spatial and channel-wise attention in convolutional networks for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5659–5667.

[157] Steven J. Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 7008–7024.

[158] Christel Chappuis, Sylvain Lobry, Benjamin Alexander Kellenberger, Bertrand Le Saux, and Devis Tuia. 2021. How to find a good image-text embedding for remote sensing visual question answering?. In *European Conference on Machine Learning (ECML) Workshops*.

[159] Tanzila Rahman, Shih-Han Chou, Leonid Sigal, and Giuseppe Carenini. 2021. An improved attention for visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 1653–1662.

[160] Sanjay Subramanian, Sameer Singh, and Matt Gardner. 2019. Analyzing compositionality in visual question answering. *Advances in Neural Information Processing Systems* 7 (2019).

[161] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition*. 3195–3204.

[162] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Yufan Chen, Peter Wu, Michelle A Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. MultiBench: Multiscale Benchmarks for Multimodal Representation Learning. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1)*.

[163] Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alex Smola. Benchmarking multimodal AutoML for tabular data with text fields. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

[164] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. 2016. Attentive explanations: Justifying decisions and pointing to the evidence. *arXiv preprint arXiv:1612.04757* (2016).

[165] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4971–4980.

[166] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. 2016. Generative adversarial text to image synthesis. In *International Conference on Machine Learning*. PMLR, 1060–1069.

[167] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. 2011. The Caltech-UCSD Birds-200-2011 dataset. *California Institute of Technology* (2011).

[168] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. 2018. AttnGAN: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1316–1324.

[169] Leyuan Qu, Cornelius Weber, and Stefan Wermter. 2019. LipSound: Neural mel-spectrogram reconstruction for lip reading. In *INTERSPEECH*. 2768–2772.

[170] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. 2018. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 12 (2018), 8717–8727.

[171] Naomi Harte and Eoin Gillen. 2015. TCD-TIMIT: An audio-visual corpus of continuous speech. *IEEE Transactions on Multimedia* 17, 5 (2015), 603–615.

[172] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan O. Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. 2018. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *Proceedings of the International Conference on Learning Representations*.

[173] Jonathan Shen, Ruoming Pang, Ron J. Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, Rif A. Saurous, Yannis Agiomyrgiannakis, and Yonghui Wu. 2018. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'18)*. IEEE, 4779–4783.

[174] Ariel Ephrat and Shmuel Peleg. 2017. Vid2Speech: Speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 5095–5099.

[175] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. 2018. Lip2AudSpec: Speech reconstruction from silent lip movements video. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2516–2520.

[176] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. 2019. Video-driven speech reconstruction using generative adversarial networks. *Proc. Interspeech 2019* (2019), 4125–4129.

[177] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. VL-BERT: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*. https://openreview.net/forum?id=SygXPaEYvH

[178] Wei-Hung Weng and Peter Szolovits. 2019. Representation learning for electronic health records. *arXiv preprint arXiv:1909.09248* (2019).

[179] Xianli Zhang, Buyue Qian, Yang Li, Yang Liu, Xi Chen, Chong Guan, and Chen Li. 2021. Learning robust patient representations from multi-modal electronic health records: A supervised deep learning approach. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 585–593.

[180] Tao Chen, Richang Hong, Yanrong Guo, Shijie Hao, and Bin Hu. 2022. MS^2-GNN: Exploring GNN-based multimodal fusion network for depression detection. *IEEE Transactions on Cybernetics* (2022).

[181] Jianliang Gao, Tengfei Lyu, Fan Xiong, Jianxin Wang, Weimao Ke, and Zhao Li. 2021. Predicting the survival of cancer patients with multimodal graph neural network. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 19, 2 (2021), 699–709.

[182] O. Caglayan, P. Madhyastha, L. Specia, and L. Barrault. 2019. Probing the need for visual context in multimodal machine translation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, 4159–4170.

[183] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations*, Yoshua Bengio and Yann LeCun (Eds.).

[184] Jinsong Su, Jinchang Chen, Hui Jiang, Chulun Zhou, Huan Lin, Yubin Ge, Qingqiang Wu, and Yongxuan Lai. 2021. Multi-modal neural machine translation with deep semantic interactions. *Information Sciences* 554 (2021), 47–60.

[185] Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *Journal of Artificial Intelligence Research* 47 (2013), 853–899.

[186] Amir Zadeh and Paul Pu. 2018. Multimodal language analysis in the wild: CMU-MOSEI dataset and interpretable dynamic fusion graph. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*.

[187] Alistair E. W. Johnson, Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3, 1 (2016), 1–9.

[188] Xintong Han. 2017. Fashion 200K Benchmark. https://github.com/xthan/fashion-200k. (2017). [Online; accessed 2017].

[189] N. Silberman and R. Fergus. 2011. Indoor scene segmentation using a structured light sensor. In *Proceedings of the International Conference on Computer Vision - Workshop on 3D Representation and Recognition*.

[190] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. 2012. Indoor segmentation and support inference from RGBD images. In *Proceedings of the 12th European Conference on Computer Vision-Volume Part V*. 746–760.

[191] Ali Furkan Biten, Lluis Gomez, Marçal Rusinol, and Dimosthenis Karatzas. 2019. Good news, everyone! Context driven entity-aware captioning for news images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12466–12475.

[192] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. MSR-VTT: A large video description dataset for bridging video and language. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5288–5296.

[193] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM International Conference on Multimedia*. 1645–1653.

[194] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2758–2766.

[195] Licheng Yu, Xinlei Chen, Georgia Gkioxari, Mohit Bansal, Tamara L. Berg, and Dhruv Batra. 2019. Multi-target embodied question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6309–6318.

[196] Catalina Cangea, Eugene Belilovsky, Pietro Liò, and Aaron C. Courville. 2019. VideoNavQA: Bridging the gap between visual and embodied question answering. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop*.

[197] Kushal Kafle and Christopher Kanan. 2017. An analysis of visual question answering algorithms. In *Proceedings of the IEEE International Conference on Computer Vision*. 1965–1973.

[198] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. 2020. nuScenes: A multimodal dataset for autonomous driving. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 11621–11631.

[199] Maria-Elena Nilsback and Andrew Zisserman. 2008. Automated flower classification over a large number of classes. In *2008 Sixth Indian Conference on Computer Vision, Graphics & Image Processing*. IEEE, 722–729.

[200] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. 2016. MOSI: Multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259* (2016).

[201] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. 2020. Can we read speech beyond the lips? Rethinking RoI selection for deep visual speech recognition. In *2020 15th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2020)*. IEEE, 356–363.

[202] Maurice Fallon, Hordur Johannsson, Michael Kaess, and John J. Leonard. 2013. The MIT Stata Center Dataset. *The International Journal of Robotics Research* 32, 14 (2013), 1695–1699.

[203] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. 2022. Data2vec: A general framework for self-supervised learning in speech, vision and language. In *International Conference on Machine Learning*. PMLR, 1298–1312.

[204] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. FLAVA: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15638–15650.

[205] Yasha Ektefaie, George Dasoulas, Ayush Noori, Maha Farhat, and Marinka Zitnik. 2023. Multimodal learning with graphs. *Nature Machine Intelligence* (2023), 1–11.

[206] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. 2021. M3P: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3977–3986.

[207] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. 2021. MURAL: Multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125* (2021).

[208] Yan Zeng, Wangchunshu Zhou, Ao Luo, and Xinsong Zhang. 2022. Cross-view language modeling: Towards unified cross-lingual cross-modal pre-training. *arXiv preprint arXiv:2206.00621* (2022).

[209] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems* 34 (2021), 9694–9705.

[210] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. 2022. Pali: A jointly-scaled multilingual language-image model. arXiv preprint arXiv:2209.06794 (2022).

[211] Maria Tsimpoukelli, Jacob L. Menick, Serkan Cabi, S. M. Eslami, Oriol Vinyals, and Felix Hill. 2021. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems* 34 (2021), 200–212.

[212] Mingchen Zhuge, Dehong Gao, Deng-Ping Fan, Linbo Jin, Ben Chen, Haoming Zhou, Minghui Qiu, and Ling Shao. 2021. Kaleido-BERT: Vision-language pre-training on fashion domain. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 12647–12657.

[213] Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. 2023. VLP: A survey on vision-language pre-training. *Machine Intelligence Research* 20, 1 (2023), 38–56.

[214] Zhiyuan Fang, Jianfeng Wang, Xiaowei Hu, Lijuan Wang, Yezhou Yang, and Zicheng Liu. 2021. Compressing visual-linguistic model via knowledge distillation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 1428–1438.

[215] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*. PMLR, 4904–4916.

[216] Nanyi Fei, Zhiwu Lu, Yizhao Gao, Guoxing Yang, Yuqi Huo, Jingyuan Wen, Haoyu Lu, Ruihua Song, Xin Gao, Tao Xiang, Haoran Sun, and Jiling Wen. 2022. Towards artificial general intelligence via a multimodal foundation model. *Nature Communications* 13, 1 (2022), 3094.

[217] Daniel Michelsanti, Zheng-Hua Tan, Shi-Xiong Zhang, Yong Xu, Meng Yu, Dong Yu, and Jesper Jensen. 2021. An overview of deep-learning-based audio-visual speech enhancement and separation. *IEEE/ACM Transactions on Audio, Speech, and Language Processing* 29 (2021), 1368–1396.