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

MUHAMMAD ARSLAN MANZOOR, Mohamed Bin Zayed University of Artificial Intelligence, UAE
SARAH ALBARRI, Mohamed bin Zayed University of Artificial IntelligenceMasdar, UAE
ZITING XIAN, Sun Yat-sen University, China
ZAIQIAO MENG, University of Glasgow, UK
PRESLAV NAKOV, Mohamed bin Zayed University of Artificial Intelligence, UAE
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: • **Computer systems organization → Neural networks**; • **Computing methodologies → Natural language processing**.

Additional Key Words and Phrases: Multimodality, Representation Learning, Pretrained Models, Multimodal Methods

ACM Reference Format:
Muhammad Arslan Manzoor, Sarah AlBarri, Ziting Xian, Zaiqiao Meng, Preslav Nakov, and Shangsong Liang. 2023. Multimodality Representation Learning: A Survey on Evolution, Pretraining and Its Applications. In . ACM, New York, NY, USA, 29 pages. https://doi.org/XXXXXXX.XXXXXXX

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 subpart of multimodal systems, utilize

[*] Shangsong Liang is the corresponding author of the paper.

Manuscript submitted to ACM

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 in a representation space.

In recent years, numerous task-specific deep learning approaches uplifted the performance for diverse multimodal tasks [8]. More recently, the implementation of pretraining and finetuning for Natural Language Processing (NLP) and Computer Vision (CV) got maximum attention due to semantically rich representation and large-scaled publicly available models [9]. We review the evolution of deep multimodal learning approaches and discuss the types and the objectives of pretraining required to make the backbone robust for diverse downstream tasks. Most pretraining approaches are based on transformers that brought up the idea of unifying architectures to deal with all modalities for all downstream tasks [10]. The methodology of several recent pretrained and unifying architectures, their performance on benchmarks, applications, and evaluation on downstream tasks are comprehensively presented in this survey.

Last year, several surveys have been published that discussed vision language pretraining [11, 12]. Comparatively, we cover the architectural detail of vision, language, and audio pretrained models exhibited in recent work [13]. Besides discussing the pretraining types, we reviewed the generic and multimodal versions of pretraining objectives. Furthermore, we summarize the recent unifying architectures (generic models) that eliminate finetuning for the different downstream tasks, which ultimately reduce the time and the computational complexity. Contrary to recent surveys, we focus more on the NLP applications enhanced by the visual and the audio modalities, e.g., sentiment analysis, document understanding, fake news detection, retrieval, translation, and other reasoning applications. 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 the internet yearly.
The contributions of this survey are as follows:

- We present a comprehensive survey of multimodal Representation Learning techniques that bridge the gap between linguistic, vision, and audio inputs in an effective way.
- The evolution of task-specific and transformer-based pretrained architectures that address multimodality.
- The pretraining types, advanced pretraining objectives for multimodal learning, detailed architectural discussion, and comparisons are elaborated.
- The development of unified architectures to address multiple modalities for all downstream tasks are reviewed.
- We develop the taxonomy of deep multimodal architectures and complex multimodal applications.
- The dataset section depicts the comprehensive information of all benchmarks used to pretrain, finetune and evaluate the multimodal methods that provide the ready-to-use details to beginners.
- Finally, leading challenges, open gaps in the field, and possible future forecasting are expressed.

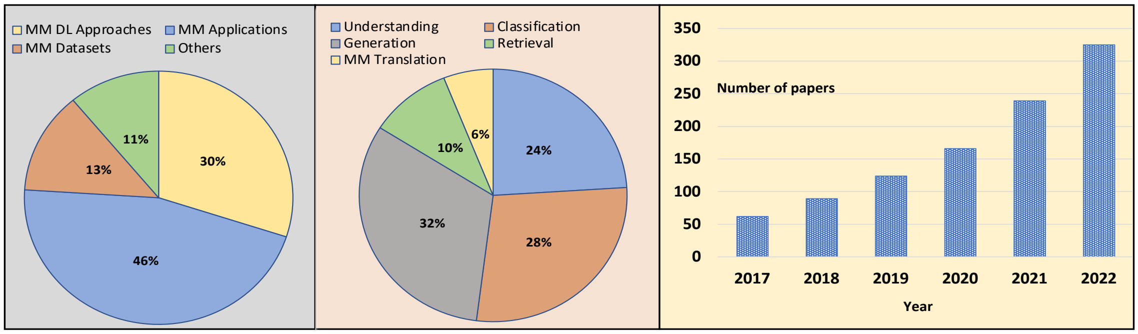

Fig. 1. The pie chart on the left represents the percentage of papers for each section included in this survey. The pie chart in the center represents the percentage of papers for each multimodal application. The rightmost figure expresses the growth of deep learning-based multimodal papers in the last six years.

2 BACKGROUND - THE ADVANCE EVOLUTION OF MULTIMODAL ARCHITECTURES

Researchers got inspiration for multimodal research from audio-visual speech recognition (AVSR) [14], where well-aligned integration of vision and hearing was a fundamental requirement. Results motivated the research community to extend this idea to other tasks. Multimodal techniques evolved over time and were used for multiple modalities such as speech, vision, and emotion recognition simultaneously in the past [2]. The most recent category contains data from multimedia resources that can be more complex than two or three modalities. The multimodal resources encouraged the researchers to solve the problem of data-enhanced multimodal tasks more effectively than the unimodal framework.

The representative cross-modal application is image captioning, which generates the text description of a given image. Document understanding [15] extracts layout, visual features, and tabular and structural information from the documents. Sentiment Analysis [16] classifies the emotion in reviews, tweets and social media posts containing video, images, emojis and other emotional expressions. The detection task includes identifying the specific event, discovering trends, finding the emergency and detecting fake news on the web or social media [17]. Information retrieval also needs multimodal processing, especially in electronic health records, where patient data is in different modalities [18].

The first challenge that algorithms face while dealing with multimodal data is understanding the encoding of different modalities, which is called the heterogeneity gap in learning representation [19]. It can be reduced by associating the similar semantics of different modalities correlatedly. Fusing features from different modalities boosted the performance on cross-media tasks [20]. Recent research found that deep learning outperforms classical approaches in understanding and processing representation due to strong learning abilities [21]. The critical phenomenon is to learn representation hierarchically as general purpose learning without structural or additional information. Mentioned reasons contributed to the fact that deep learning is the most compatible approach for multimodal representation learning.

Early task-specific multimodal methods follow the pipeline approach and are usually trained to address only one task as shown in Figure 2. For example, Vision-Language (VL) task-specific methods have a visual encoder based on CNN [22, 23] that extracts visual features and a text encoder based Bag-of-Words [24], sequence models [25] or transformers [26] to encode textual features. Subsequently, Multimodal Fusion (explored extensively for task-specific models) builds interactions between features of different modalities to produce a common representation [27]. These Multimodal Fusion methods evolve as (*i*) "simple fusion" that fuses features through concatenation or element-wise sum or product [28], (*ii*) "inter-modality attention" [29] captures feature alignment of different modalities and constructs more informative

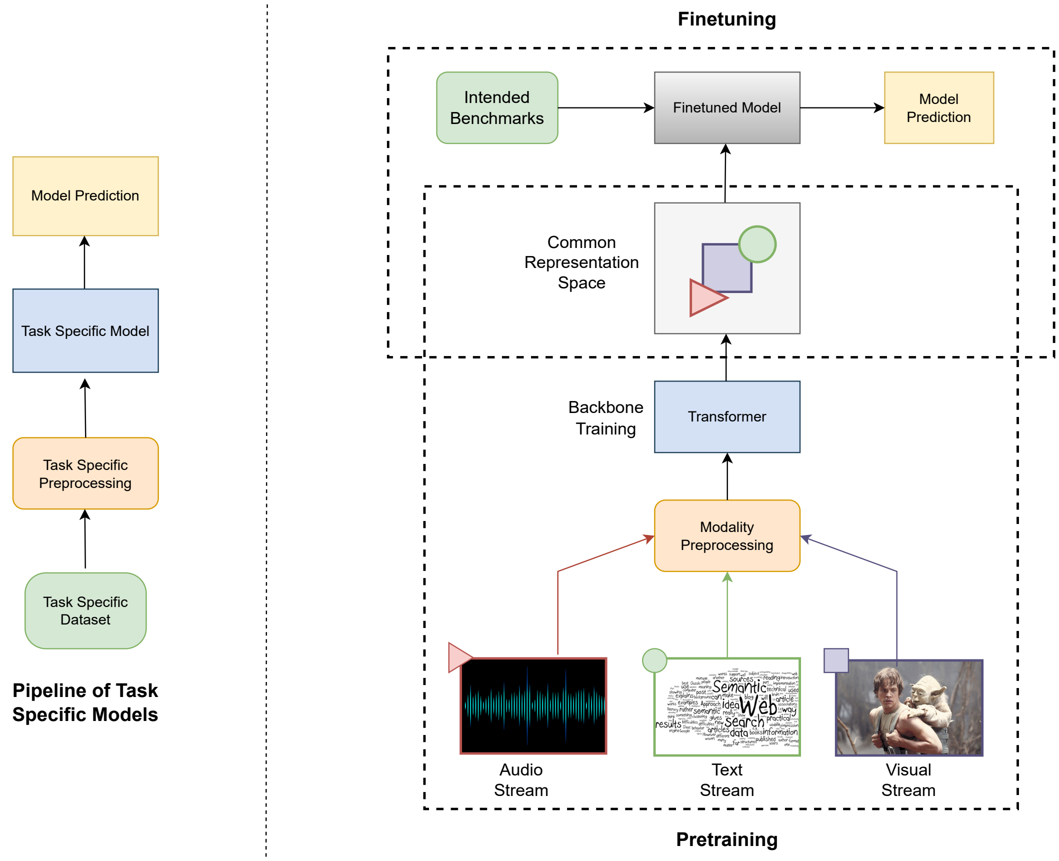

Fig. 2. The figure exhibit the difference in the pipeline of Pretraining-Finetuning architecture and the task-specific architecture. Different modes of input in **Pretraining** show that multimodal data is passed to the model. The **Preprocessing** module formats each modal input according to the model. The **Transformer** serves as the training backbone (could be any variant) that generates the **common representation space**. **Finetuning** is based on the model for any intended (relevant) benchmark using common representation space for any **downstream task**. Contrary to the Pretraining-Finetuning pipeline, the task-specific method on the left is trained on the task-specific dataset from the beginning, using any backbone (Encoder-Decoder, Attention-based, or RL based).

joint representations, (*iii*) "intra-modality attention" [30] develops relational reasoning by constructing a graphical representation, (*iv*) "transformer" relies on inter-modality and intra-modality by attending other modalities and related regions of the same modality [31]. Therefore, transformers are considered the backbone of the universal multimodal pretraining approaches as they learn and correlate deep interactions among different modalities [32, 33], unlike the dual encoder (pipeline) architecture that exploit cosine similarity [34].

In the last decade, several articles discussed the challenges, performance, and application of multimodal approaches in CV and NLP. Various studies focused on conventional multimodal topics i.e., representation, correlation, fusion, translation and co-learning such as [2, 6]. Some authors unfolded general applications related to NLP, and CV [35], and few authors addressed a single application using multimodal architecture [16]. Most of the recently published work targeted challenges, pretraining objectives, and applications with a CV perspective [7, 12, 13]. Unlike, we build up this study that covers the evolution of deep learning based multimodal task-specific methods, pretraining tasks to train

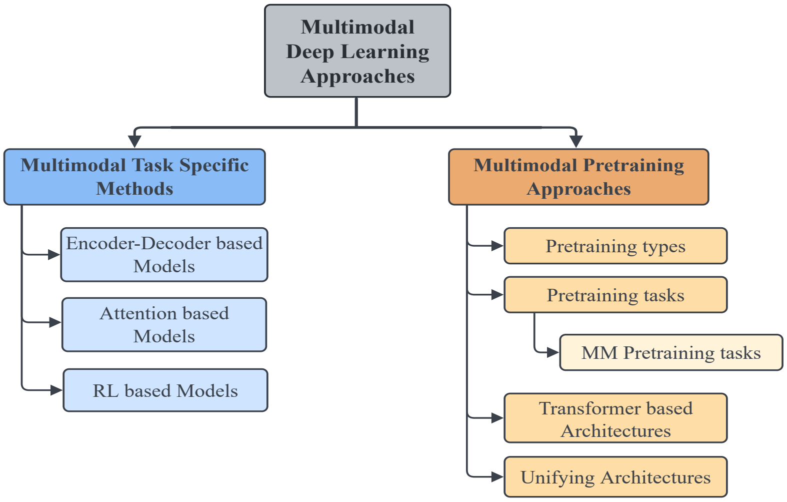
Multimodal Deep Learning Approaches
Multimodal Task Specific Methods
Encoder-Decoder based Models
Attention based Models
RL based Models
Multimodal Pretraining Approaches
Pretraining types
Pretraining tasks
MM Pretraining tasks
Transformer based Architectures
Unifying Architectures

Fig. 3. Taxonomy of deep learning based Multimodal Approaches in section 3.

advanced transformer-based pretraining approaches, recently released unifying architecture, discussion on comparative analysis, multimodal applications, and benchmarks. Finally, future forecasting is presented for peers.

3 MULTIMODAL DEEP LEARNING APPROACHES

This section presents the numerous variants of multimodal architectures mainly categorized as task-specif architecture and Pretraining-Finetuning architectures (pipeline shown in Fig 2). Figure 3 exhibits the taxonomy of section 3. Subsection 3.1 represents the acronyms of the tasks mentioned in this study. Subsection 3.2 comprehensively summarizes the task-specific methods as the baselines of advanced multimodal approaches that transformed into large-scale pretrained approaches in recent years. SubSection 3.3 demonstrates the pretraining process, types, objectives and SOTA frameworks trained on the multimodal dataset to perform enhanced NLP and cross-modal tasks. Furthermore, unifying architectures that achieved attention more recently are detailed at the end. Subsection 3.4 provides a comparative discussion on results produced by SOTA approaches on variety of multimodal tasks.

3.1 Acronyms

This section contains all in the forthcoming tables. **AEC**: Audio Event Classification, **Auton. Driving**: Autonomous Driving, **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, **IE**: Information Extraction, **IR**: Image Retrieval, **IS**: Indoor Segmentation, **ITR**: Image to Text Retrieval, **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 Reasonng, **VE**: Visual Entailment, **VG**: Visual Generation, **VLP**: Vision Language Pretraining, **VQA**: Visual Question Answering, **VU**: Visual Understanding

3.2 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.

3.2.1 Encoder-Decoder Based Models. Encoder-decoder models maps the input to output using two basic building blocks; an encoder and a decoder. The encoder transforms input data into a semantically rich representation known as latent representation. The decoder predicts the output using the input latent representation. The most used encoder-decoder frameworks are cascaded CNN-RNN and RNN-RNN or variations of these architectures. The work of Hu et al. [36], which uses text and image modalities, proposed Cascaded Recurrent Neural Network (CRNN). A frontend and a backend network are cascaded to learn image-text interactions from forward and backward directions. The encoder part is a VGG16 network, a CNN designed for classification and localization tasks, which extracts the deep semantic context of images. The decoder part is Stacked Gated Recurrent Unit (SGRU). Summaira et al. [37] proposed a Globally Enhanced Transformation network (GET). The encoder section extracts global image features via Faster-RCNN [38] and ResNet-101 architectures. The global feature embedding outputted by the encoder guide the decoder in the generation process. GET architecture outperforms the SOTA architectures on MS-COCO dataset [39] with a BLUE score of 81.5 compared to CRNN architecture, whose BLUE score is 69.1.

Autoencoders [40] is a case of encoder-decoder architecture that exploits unsupervised learning. The latent representation produced by the encoder should entail all the required information to obtain the original data points by the decoder [40]. The decoder approximately reconstructs the input by using the latent representation and minimizing the reconstruction loss. In [41], a multimodal architecture is based on autoencoder. The outputs of two parallel autoencoders, which process audio and video modalities, are combined in the latent representation layer. The combined latent representation space is then used to train a conventional encoder-decoder architecture.

3.2.2 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. [42] 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 [42]. Another attention based variant is explored in attention-guided Multimodal Correlation (AMC) [43]. 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.

3.2.3 Reinforcement Learning-Based Models. The interaction between an environment-aware agent and environment through trial and error is the learning phenomenon in reinforcement learning models. The agent explores and interacts with environment, obtains a reward and punishment. High rewarding actions are *reinforced* by the algorithm, and exploration of similar actions is encouraged. Reinforcement learning does not require labeled data; instead, it depends on the balance between exploitation of previously gained knowledge by past actions and exploration of a new set of actions. The concept of reinforcement learning is applied to deep learning architecture in [44]. A Hierarchical-based Reinforcement Learning (HCL) framework employs agents such as a high-level manager, a low-level worker, and an internal critic. The manager model sets goals and subgoals for the worker model, which attempts to fulfill the manager goals by performing primitive actions at each time step. Finally, the internal critic assesses the goal completion process. This architecture incorporates and acts upon context from the environment and completed goals. HCL framework outperforms baseline pipelines for transferring a video modality into text.

The categorization of deep learning methods as encoder-decoder-based, attention-based and reinforcement learning-based is generic and simple. These are baseline categories in which various architectures belong to multiple categories, such as encoder-decoder and attention mechanism. 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) [45, 46]. Even though task-specific architectures perform well, these approaches requires training for each task which is time consuming and computationally expensive. Therefore, many researchers have explored the pretraining framework to avoid retraining repetitively for different tasks.

3.3 Multimodal Pretraining Methods

This section covers a well-detailed discussion of the pretraining framework including pretraining types, objectives, and state-of-the-art methods. Earlier, Self-Supervised Learning (SSL) boosted pretraining for language tasks [26] which was later used for vision tasks and produced efficient results [47]. Pretraining on massive labeled or unlabeled datasets and finetuning on task-specific datasets has become a modern paradigm for various domains [48].

3.3.1 Types of Pretraining. Kalyan et al. [49] 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 [50], RoBERTa [51] and BERT[52]. (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 [53], HateBERT [54] and infoXLM [55] are the examples of CPT. (3) Simultaneous Pretraining (SPT) used in [56]. It consumes domain-specific and general text simultaneously while training from scratch. (4) Task Adaptive Pretraining (TAPT) used in [57] demands a small amount of data. (5) Knowledge Inherited Pretraining (KIPT) [58] uses the Knowledge distillation technique.

3.3.2 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 challenging enough to make the model robust by exploiting the training signal, which include: (1) CLM (Casual Language Modeling) is unidirectional and uses context to predict the next word employed by GPT-1 [59] for the first time. (2) MLM (Masked Language Model) enhanced the representation by considering the context in both directions and consuming only 15% of the tokens applied by [52] for the first time. (3) RTD (Replaced Token Detection) mitigates the pretraining challenges of MLM (less supervisory signal) by classifying the replaced token as original or not [50]. (4) STD (Shuffled Token Detection) reduces the discrepancy between pretraining and finetuning by exploiting discriminative tasks (detection of shuffled tokens) [60].

Other pretraining tasks include RTS (Random Token Substitution), NSP (Next Sentence Prediction), SLM (Swapped Language Models), SOP (Sentence Order Prediction), and some others.

3.2.2.1 Multimodal Pretraining Tasks. These include: (1) Cross-Modal Masked Language Model, which is similar to

Table 1. Transformer-based pretraining approaches, benchmarks, tasks and architectural settings.

Model	Benchmarks	Pretraining Tasks	Architectural Comparison
VLP [65]	VQA2.0, COCO Captions, Flickr30k	IC, VQA	12-layer, , 768-hidden, 12-heads, 110M parameters
VirTex [66]	COCO Captions	Image Classification, IC	Visual backbone (ResNet-50), Textual head (two unidirectional Transformers)
ViLBERT [9]	VQA, VCR, RefCOCO, Image Retrieval	VQA, VCR, GRE CIR	BERT-Base textual encoder, Faster R-CNN (ResNet 101 visual backbone)
ERNIE-ViL [67]	VCR, RefCOCO+, VQA, IR-Flickr30K, QR-Flickr30K	SGP, VCR , VQA, GRE, IR, TR	BERT-Base textual encoder, Faster R-CNN (ResNet 101 visual backbone)
OSCAR [68]	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 [69]	SST-2. QNLI, QQP, MNLI,SQuAD v1.1, SQuAD v2.0, SWAG Avg.	GLUE, SQuAD, SWAG, GLUE	Backbones trained: last 4 layers of BERT for text, ResNeXt-101-32x8d for vision
AV-HuBERT [3]	LRS3, VoxCeleb2	ASR	Hybrid ResNet-ransformer architecture

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 [12]. (2) Cross-Modal Masked Region Prediction: Similar to previous objective, Masked Region Prediction (MRP) masks RoI with zeros. The contextual visual features and text inference predicts the masks. (3) Image-Text Matching: 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: 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.

Pretrained approaches have been used extensively to perform exceptionally well for NLP downstream tasks. The researchers proposed pretrained models including BERT [52], ALBERT [61], RoBERTa [51], and T5 [62], 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 [63], [64]. The following writings present different pretraining architectures designed recently to exploit multimodal data for unimodal and multimodal tasks. The objective is to provide summarized, quality information to the reader about existing pretraining approaches in data modalities, architecture, dataset, and downstream tasks.

3.3.3 Transformer-Based Architecture. Recent work on multimodal pretraining has inspired the research community to progress in multimodal tasks. The combined pretraining-based advanced approaches on image-text at a large scale are developing rapidly to outperform task-specific architectures. In ViLBERT [9], the author Pretrained BERT using textual and visual inputs in different streams that maintain relationships with co-attention layers. After pretraining on the Captions dataset, the model is transferred to established language-vision tasks, including VQA, VCR, and caption-based image retrieval. Significant improvements are achieved for VQA and RefCOCO+.

The VLP (Vision-Language Pretraining) in a unified manner is proposed by [65] that outperformed the previous SOTA for image captioning or Visual Question Answering tasks. The proposed model follows the architecture of a shared encoder-decoder and is pretrained on a large scale on many image-text pairs in an unsupervised way. The model is evaluated on VQA 2.0, COCO Captions, and Flickr30k [70] Captions. Li et al. [68] proposed a pretraining approach in which the object tags are used as anchors to enhance the alignment learning with paired text.

In recent research [67], structured knowledge is extracted from scene graphs that assist in learning joint representation as semantically connected features of text and vision. Scene graph prediction tasks (including Object Prediction, Attribute Prediction, and Relationship Prediction) are constructed by the author in the pretraining phase. The proposed model can learn joint representation in semantically aligned form. The model's effectiveness is evaluated on five cross-modal tasks, and VCR is at the top with a 3.7% improvement.

Word2Vec [71], GPT [59], ELMO [72], BERT, and other SSL based frameworks perform good natural language understanding but does not consider grounding information from visual world as motivated by Bender et al. [73] and Bisk et al. [74]. The Vokenization procedure for pretraining designed by Tan et al. [69] takes language tokens and corresponding visual information (vokens) related to tokens as input supervision. The author 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.

Desai et al. [66] proposed VirTex, a data-efficient alternative to supervised pretraining that uses captions as a supervisory signal for vision tasks. They conducted joint pretraining of Convolution and Transformer from scratch based on language and vision that aims to generate captions for images. The visual backbone comprised of ResNet (50) is used to compute visual features of input images, which then pass to the textual head to generate captions for photos. After training, the textual authority is discarded as the focus is to apply learned features for visual downstream tasks.

Moreover, Shi et al. [3] proposed The Audio-Visual Hidden Unit BERT (AV-HuBERT), trained from scratch on benchmark LRS3 that 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 [75] to outperform the models that are pretrained on a single modality. Table 1 draws the comparative picture discussed approaches, on the basis of benchmarks, tasks and architectural settings.

3.3.4 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.* [76] presented the unified model for text-only and vision-only tasks. The author 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 [77] 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 author's second concept for the WRA task to improve the alignment between images and words.

Hu et al. [78] 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 [63] is used for

Table 2. The unifying architectures with the benchmarks, tasks and architectural setting adopted by the authors.

Model	Benchmarks	Tasks	Architectural Comparison
UniT [78]	COCO, QQP, VQAv2 QNLI, MNLI, SST-2, SNKI-VE	VQA, OD, VE	201M parameters, comprised of ResNet-50 + BERT
ViT-BERT [76]	MNLI, QQP, QNLI, SST-2, RTE Cifar, ImageNet, Flowers, Pet	Vision-only, text-only	12 layer transformer, 768-hidden, 3072 MLP
UNITER [77]	VQA, Flicker30k, NLVR, Ref-COCO	IR, TR, VQA, RE	12-24 layers, 768- 1024 hidden 12-16heads, 86-303M parameters
VATT [79]	AudioSet, HowTo UCF101, HMDB51, MSR-VTT	VAR, AEC, TVR	Transformer-based variants consuming between 155M to 415M Param
OFA [80]	SST-2, RTE, MRPC, QQP, QNLI, MNLI, SNLI-VE	NLU, NLG, Image classification	Transformer-based variants consuming between 33M to 940M Param

visual features encoding and BERT [52] performs the textual feature encoding. Contrastive learning is employed on multimodal data by Akbari et al. [79] to train a transformer encoder that processes audio, text, and video simultaneously.

Wang et al. [80] 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 [59] and BART [81], BPE (Byte-Pair Encoding) is used to divide a sequence of words into sub-word sequences and embed them into features.

3.4 Discussion

Most multimodal approaches used a transformer as a backbone.These architectures outperformed CNNs on most vision tasks as self-attention focuses on 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 mentioned in this survey.

The by VLP [65] considered as the first SOTA performance on vision-language joint task specifically generation and understanding. Likewise, VirTex showed SOTA performance on multimodal tasks [66]. ViLBERT [9] marked SOTA on VQA and RefCOCO+ with huge margin. AV-HuBERT [3] 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 [67] achieved SOTA on 5 cross-modal downstream tasks and ranked at first on the VCR leaderboard with an absolute improvement of 3.7%. OSCAR [68] created new SOTA on six well-established vision-language understanding and generation tasks. Vokenizer [69] 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 2 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 [76] 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 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% avg for both tasks (consume fewer parameters than UniT). The UniT [78] 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 [77] model achieved comparable performance across all the benchmarks but consumed 303M parameters. UNITER-base model also performed well except VQA with 86M parameters. VATT [79] 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 [80] surpassed the UNITER, UNIMO, and other SOTA on VQA and SNLI-VE. Furthermore, 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.

4 MULTIMODAL APPLICATIONS

This section presents the categorical detail of multimodal applications enhanced by deep learning architectures as shown in Figure 4. The multimodal tasks are divided into main categories: understanding, classification, retrieval, and generation. 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 (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 [15]. Contrary to the classical information extraction methods, Visually Rich Documents Understanding considers the visual layout along with the text. LayoutLMv2 [82] 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.* [83] proposed StrucTexT, which outperformed LayoutLMv2 by pretraining a

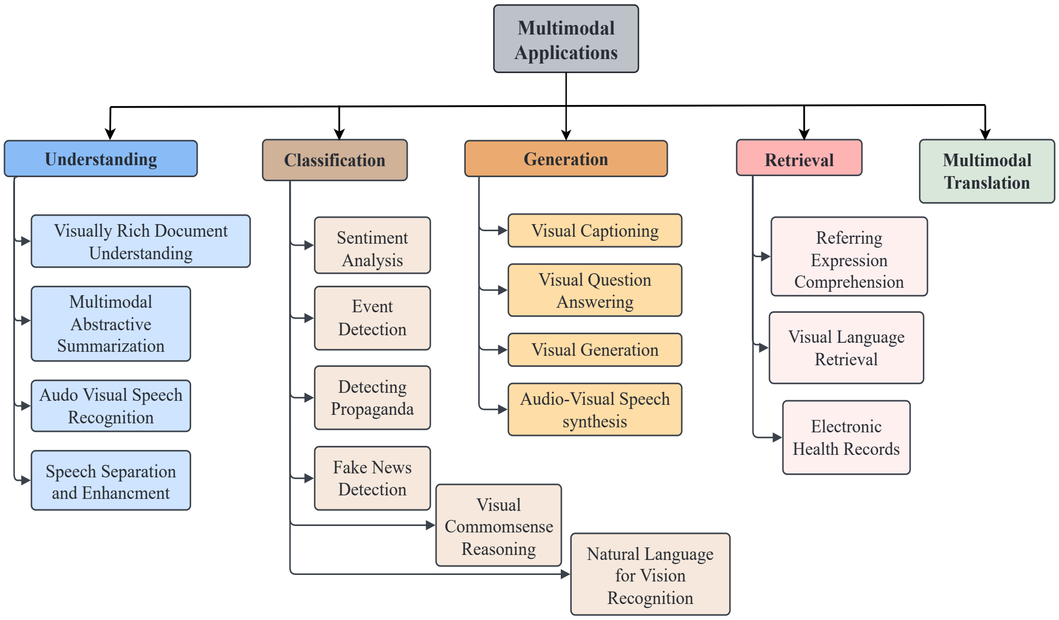
Multimodal Applications
Understanding
Classification
Generation
Retrieval
Multimodal Translation
Visually Rich Document Understanding
Multimodal Abstractive Summarization
Audo Visual Speech Recognition
Speech Separation and Enhancment
Sentiment Analysis
Event Detection
Detecting Propaganda
Fake News Detection
Visual Commomsense Reasoning
Natural Language for Vision Recognition
Visual Captioning
Visual Question Answering
Visual Generation
Audio-Visual Speech synthesis
Referring Expression Comprehension
Visual Language Retrieval
Electronic Health Records

Fig. 4. Taxonomy of the multimodal applications in Section 4.

Transformer in a self-supervised manner, with less parameters, achieved high scores. The publicly available benchmark datasets for downstream tasks are SROIE [84] and FUNSD [85]. LayoutLMv2 achieved 97% and 84% F1 score on SROIE and FUNSD, respectively. StrucTexT had 96.88% and 85.68% F1 on SROIE and FUNSD, respectively.

Gu et al. [86] released an improved version XYLayoutLM of LayoutLMv2 [82]. 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 [87]) 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 [82] 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 [88]. Palaskar et al. [89] 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 [90] 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 [91] dataset, which contains instructional videos with two to three sentences of 2000 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 [92], HuBERT [75] and De-CoAR2.0 [93] 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 Learn audio-visual speech representation by taking favor of lip-reading and ASR. The proposed model achieved 32.5% WER on LRS3 [94] benchmark by using 30 hours of labeled data from 433 hours and achieving 26.9% outperformed the former SOTA that trained on 31K hours [95]. 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 [96]. Previously, these tasks were handled with statistical and mathematical criteria under signal processing [97]. 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.

Audi-Visual Sentence Extraction (AV-SE) and Audi-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) [98]. Sound Source separation is achieved by Zhou et al. [99] using the Appearance Attention Module that leverages categorical information on of single frame video. The author 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 [100]. 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 extract and classify 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 [101], [102], and [103]. However, the sentiment analysis task moved from text modality to another form of modality due to social media and the internet. Chen *et al.* [104] 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 [105]. A massive amount of data is generated that significantly impacts the lives, property, and psychology of humans [106]. Event detection can be mapped to other realistic scenarios, such as Emergency Management, Disaster Detection, and Topic Detection [107]. 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 [108], CrisisNLP and CrisisLex. For traffic event detection, Chen *et al.* [109] created a multimodal dataset by integrating the traffic-related filtered tweets with sensor data. The author achieved 84%, 83%, 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 [110], sentence level and fragment level [111]. 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 [112]. A multimodal architecture can detect fake news by looking for mismatches between the modalities [113].

Wang *et al.* [17] proposed fine-grained multimodal fusion networks (FMFN) to detect fake news. First, CNN extracted visual features, and RoBERTa [51] 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 achived 88% accuracy on the Weibo Dataset [112] 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. [114] 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 [24]. Song et al. [115] 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 [77, 116, 117] considered NLVR as 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) [118] 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 [39] and Flickr30k [70]. The baseline and conventional visual captioning pipeline consist of an image encoder and a text decoder. Recent work leverages BERT capabilities, such as in [119]. BERT models [52] 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. [120] 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. [121] improved CNN and proposed SCA-CNN based on images that CNN can extract: Spatial, Channe-wise, and Muli-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. [122] 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 [123]. 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 [124], attention mechanism [125], compositional model [126] and knowledge base model[127]. There exist review papers targeting VQA as a research field of its own such as in [123] and [5]. In addition to surveys, benchmarks such as [128] by Carnegie Mellon University (CMU) and [129] are available. Benchmark datasets are VAQ v1.0 [5], VAQ-X [130], and VAQ-CP [131].

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. [132] 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 dataset used in this paper are: Caltech-UCSD Birds[133] dataset, Oxford-102 Flowers dataset, and MS COCO [39] dataset. Xu et al. [134] proposed an Attentional Generative Adversarial Network (AttnGAN) that allows attention-driven, multistage refinement for a fine-grained text-to-image generation. AttnGAN consists of two components: the attentional generative network and DAMSM(Deep Attentional Multimodal Similarity Model). The attentional generative network draws different image subregions by focusing on the words most relevant to the subregion being drawn. DAMSM provides an additional fine-grained image-to-text matching loss for training the generator. A comprehensive study is carried out and shows that AttnGAN is significantly better than previous state-of-the-art GAN models. This method is evaluated on CUB[133] and COCO [39] datasets.

4.3.4 Audio-Visual Speech Synthesis. Prajwal et al. [96] presented the model synthesizing speech from lip movement in the presence of noise. The author released the benchmark Lip2Wav Dataset, as the previous work [135] was evaluated on small and limited vocabulary-based datasets [94, 136, 137]. The designed dataset enables the model to synthesize speech from unconstrained lip movements.

High-quality speech is generated [138] using a Tacotron [139] inspired decoder that produce melspectrogram from text inputs. It is conditioned on the face embeddings encoded in the previous representation, and outperformed previous work [140–142] on all objective metrics for lip-to-speech work especially on TIMIT [137] 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 [143], 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 [144] 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 [145]. 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 MRI scans, or signals such as electrocardiogram (ECG) [145]. Information retrieval system 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 [18]. SDPRL is build and tested using the benchmark dataset MIMIC-III.

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. Caglan et al. [146] 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 [23] is used for visual features extraction. Multimodal Attention for Neural Machine Translation (NMT) inspired by [147] achieved the context vector from textual and visual features. Multi30K and flicker30 were used for training. MMT performed better than NMT.

Furthermore, Su et al. [148] proposed the approach that exploits the interaction of semantic representation of different modalities. This work is inspired by QA and NER domains that effectively consider the interaction and produce significant improvement. The author proposed two attention-based variations of models. The first network utilizes a bi-directional attention mechanism to learn the semantic representation from text and visual features by interaction. An attention matrix was used to extract the fine-grained alignments between different modalities that help the semantic representation learning process. The second variation consists of a co-attention mechanism that produces a context vector from textual representation at the first stage and is later used as a supervisor to generate an image context vector. The image context vector is used in the final re-attention process to update the text context vector with more interactive information. Proposed models are evaluated on the Multi30K dataset expanded from Flick30K and famous for Multimodal NMT. Other variations were produced by pretraining on a back-translated dataset of the same benchmark. The experimental results show that pretrained and without pretraining NMTbi-att and NMTcoatt Networks surpass the previous baselines pretrained and without pretraining networks without pretraining approaches, respectively.

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 3. The MS COCO dataset [39] dataset 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 [70] is collected from Flickr, together with 5 image descriptions provided by human annotators. The Flickr8k [149]is smallest version, also collected from Flickr, that leads to train model easier and faster. FUNSD [85] is a small dataset comprising 199 real, fully annotated, scanned forms. FUNSD contains 31485 words, 9707 semantic entities, and 5304 relations and aims to extract and structure forms' textual content. SROIE [84] consists of a dataset with 1000 whole scanned receipt images and annotations for the competition on scanned receipts OCR and key information

Table 3. The Benchmarks' Information for the tasks in this study. The Task column lists the tasks in the corresponding papers.

Dataset	Modality	Samples (Size)	Tasks	Data Category
MS COCO [39]	text + image	328k	Image Recog., IC	Multiple
Flickr30K [70]	text + image	31k	IR, CMR, IC	Multiple
Flickr8K [149]	text + image	8k	IR, CMR, IC	Multiple
FUNSD [85]	text + image	199	TD, OCR, EL.	Scanned forms
SROIE[84]	text + image	1k	OCR, IE.	Scanned images
CMU-MOSEI [150]	text + video	23,453	MMSA	YouTube videos
VQA [24]	text + image	250k	VQA	Multiple
CrisisMMD [108]	text + image	16,097	IC, ED	Disasters tweets
MIMIC [151]	MML data	53,423	IR	Patients Information
Fashion-200K [152]	text + image	200k	IC, IR	Clothes
NYU Depth v1 [153] v2 [154]	RGB + Depth	4GB + 90GB	IS,IC	Indoor Scenes
SKIG [56]	text + image	2160	GR	Hand gestures
GoodNews [155]	text + image	466,000	IC	News
MSR-VTT [156]	text + video	200k	VU	Commercial videos
MSVD-QA [157]	text + video	2089	VR, VQA, VC	Multiple
TGIF-QA [158]	text + image (gif)	103,919	VQA	Multiple
EQA-v1 [159]	text + 3D env.	9,000	EQA, VQA	Multiple
VideoNavQA [160]	text + 3D env.	28	EQA, VQA	Multiple
TDIUC [161]	text + image	1,654,167	VQA	Multiple
nuScenes [162]	image + sensor data	1.4M	IR, Auton. Driving	Driving Information
CUB-200 [133]	text + image	11,788	VG	Birds
Oxford-102 Flowers [163]	text + image	4,080 26,316	VG	Flowers
VCR [114]	text + image	263k	VU	Moive Scenes
How2 [91]	audio + text + video	2000 hours	MML	Instructional Videos
Lip2Wav [96]	audio + video	120 hours	AVSS	Talking Face Videos
MUSIC [100]	audio + video	685	SSS	Videos of Music
MOSI [164]	audio + text + video	3702	Multimodal analysis	YouTube videos
GRID [165]	audio + video	33,000	VSR	Speech

extraction (SROIE). CMU Multimodal Opinion Sentiment and Emotion Intensity (CMU-MOSEI) dataset [150] is the largest dataset of multimodal sentiment analysis and emotion recognition to date. CMU-MOSEI was built in 2018 and contained 23,453 annotated video segments from 1000 distinct online YouTube and 250 topics.

VQA (Visual Question Answering) [24] is a large dataset containing more than 250,000 images with at least 3 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 [108] 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, etc. The last type is "Damage Severity Assessment", which describes the severity of the damage.

MIMIC (Medical Information Mart for Intensive Care) [151] 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 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 [152] contains 200K fashion images, and each image comes with a compact attribute-like product description. MIT-Stata Center [166] 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 [153] 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 [154] includes 464 different indoor scenes and 26 scene types.

The Shefeld Kinect Gesture (SKIG) [56] is a gesture dataset containing 2160 hand gesture sequences. These hand gestures can be classified into 10 categories: circle (clockwise), triangle (anti-clockwise), up-down, right-left, wave, "Z", cross, come-here, turn-around, and pat. The sequences are recorded under 3 different backgrounds and 2 illumination conditions, which provide diversity. GoodNews [155] 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 [156] 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 2089 short video clips (usually less than 10 seconds). The MSVD dataset contains different language descriptions, such as English, Hindi, Romanian, Slovene, etc. MSVD-QA is a benchmark for video retrieval, visual question answering, and video captioning.

TGIF-QA [158] 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 [159] 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 [160] is also a dataset used to study the EQA task. VideoNavQA contains 28 questions belonging to 8 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) [161] 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 [162] 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 4 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, etc. 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. Caltech-UCSD Birds (CUB-200) [133] contains 200 bird species and 11,788 images. Each image in Caltech-UCSD Birds is annotated with a bounding box, a rough bird segmentation, and a set of attribute labels. Oxford-102 Flowers contains 102 bird categories, consisting of 40 and 258 images. The images have large scale, pose and light variations.

Visual Commonsense Reasoning (VCR)[114] 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[91] 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[96] dataset contains 120 hours of talking face videos across 5 speakers. This dataset has about 20 hours of natural speech per speaker and vocabulary sizes of over 5000 words for each. MUSIC (Multimodal Sources of Instrument Combinations) dataset[100] 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[164](Multimodal Opinionlevel Sentiment Intensity) dataset contains 93 videos and 3702 video segments. The dataset provides not only sentiment annotations, but also manual gesture annotations. GRID[165] 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 [167]. Even though the results are impressive on downstream tasks, still could not surpass task-specific models for some of the tasks [168]. 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.

Pretraining Strategy. Developing an optimal set of pretraining objectives is worth exploring for multimodal tasks. Pretext tasksdirectly 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 Wang et al.[32] 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 [169], M3P [170], and MURAL [171] exploit multilingual features by adding a separate encoder for multilingual or translated data. CCLM [172] based on ALBEF [173] 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 [174].

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 ii) – reduce the gap between the representation learning of pretraining tasks and downstream tasks [175]. 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, task level but according to Zhuge et al. [176], 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. [11] 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 [13] 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 [177], 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 [34, 178].

Including Audio Stream. Recent Multimodal surveys [12, 13] 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 [179]. 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. Recent advances and trends in multimodal deep learning: A review. *CoRR*, abs/2105.11087, 2021.

[2] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443, 2018.

[3] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net, 2022.

[4] Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. Detecting propaganda techniques in memes. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 6603–6617. Association for Computational Linguistics, 2021.

[5] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. VQA: visual question answering. *CoRR*, abs/1505.00468, 2015.

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

[7] Khaled Bayoudh, Raja Knani, Fayçal Hamdaoui, and Mtibaa Abdellatif. A survey on deep multimodal learning for computer vision: advances, trends, applications, and datasets. *The Visual Computer*, 38, 08 2022.

[8] Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. *CoRR*, abs/1908.03557, 2019.

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

[11] Feilong Chen, Duzhen Zhang, Minglun Han, Xiuyi Chen, Jing Shi, Shuang Xu, and Bo Xu. VLP: A survey on vision-language pre-training. *CoRR*, abs/2202.09061, 2022.

[12] Yifan Du, Zikang Liu, Junyi Li, and Wayne Xin Zhao. A survey of vision-language pre-trained models, 2022.

[13] Siqu Long, Feiqi Cao, Soyeon Caren Han, and Haiqin Yang. Vision-and-language pretrained models: A survey. In Luc De Raedt, editor, *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 5530–5537. ijcai.org, 2022.

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

[15] Jinbae Im, Moonki Kim, Hoyeop Lee, Hyunsouk Cho, and Sehee Chung. 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)*, pages 388–403, Online, August 2021. Association for Computational Linguistics.

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

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

[18] Xianli Zhang, Buyue Qian, Yang Li, Yang Liu, Xi Chen, Chong Guan, and Chen Li. 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)*, pages 585–593. SIAM, 2021.

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

[20] Amirhossein Habibian, Thomas Mensink, and Cees GM Snoek. Video2vec embeddings recognize events when examples are scarce. *IEEE transactions on pattern analysis and machine intelligence*, 39(10):2089–2103, 2016.

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

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

[23] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[24] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

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

[26] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

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

[28] Allan Jabri, Armand Joulin, and Laurens van der Maaten. Revisiting visual question answering baselines. In *European conference on computer vision*, pages 727–739. Springer, 2016.

[29] Duy-Kien Nguyen and Takayuki Okatani. 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*, pages 6087–6096, 2018.

[30] Ronghang Hu, Anna Rohrbach, Trevor Darrell, and Kate Saenko. Language-conditioned graph networks for relational reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 10294–10303, 2019.

[31] Peng Gao, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven CH Hoi, Xiaogang Wang, and Hongsheng Li. 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*, pages 6639–6648, 2019.

[32] Wenhui Wang, Hangbo Bao, Li Dong, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. *CoRR*, abs/2111.02358, 2021.

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

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

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

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

[37] Jiayi Ji, Yunpeng Luo, Xiaoshuai Sun, Fuhai Chen, Gen Luo, Yongjian Wu, Yue Gao, and Rongrong Ji. Improving image captioning by leveraging intra- and inter-layer global representation in transformer network. *CoRR*, abs/2012.07061, 2020.

[38] Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *CoRR*, abs/1506.01497, 2015.

[39] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312, 2014.

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

[41] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th International Conference on Machine Learning*, ICML '08, page 1096–1103, New York, NY, USA, 2008. Association for Computing Machinery.

[42] Weitao Jiang, Xiying Li, Haifeng Hu, Qiang Lu, and Bohong Liu. Multi-gate attention network for image captioning. *IEEE Access*, 9:69700–69709, 2021.

[43] Kan Chen, Trung Bui, Fang Chen, Zhaowen Wang, and Ram Nevatia. AMC: attention guided multi-modal correlation learning for image search. *CoRR*, abs/1704.00763, 2017.

[44] Xin Wang, Wenhu Chen, Jiawei Wu, Yuan-Fang Wang, and William Yang Wang. Video captioning via hierarchical reinforcement learning. *CoRR*, abs/1711.11135, 2017.

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

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

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

[48] Junyang Lin, Rui Men, An Yang, Chang Zhou, Yichang Zhang, Peng Wang, Jingren Zhou, Jie Tang, and Hongxia Yang. 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*, KDD '21, page 3251–3261, New York, NY, USA, 2021. Association for Computing Machinery.

[49] Katikapalli Subramanyam Kalyan, Ajit Rajasekharan, and Sivanesan Sangeetha. Ammu: A survey of transformer-based biomedical pretrained language models. *J. of Biomedical Informatics*, 126(C), feb 2022.

[50] Kevin Clark, Minh-Thang Luong, Quoc V. Le, and Christopher D. Manning. ELECTRA: pre-training text encoders as discriminators rather than generators. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[51] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[52] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

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

[54] Tommaso Caselli, Valerio Basile, Jelena Mitrovic, and Michael Granitzer. Hatebert: Retraining BERT for abusive language detection in english. *CoRR*, abs/2010.12472, 2020.

[55] Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. Infoxlm: An information-theoretic framework for cross-lingual language model pre-training. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tür, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3576–3588. Association for Computational Linguistics, 2021.

[56] Shoya Wada, Toshihiro Takeda, Shiro Manabe, Shozo Konishi, Jun Kamohara, and Yasushi Matsumura. A pre-training technique to localize medical BERT and enhance biobert. *CoRR*, abs/2005.07202, 2020.

[57] Suchin Gururangan, Ana Marasovic, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A. Smith. Don't stop pretraining: Adapt language models to domains and tasks. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 8342–8360. Association for Computational Linguistics, 2020.

[58] Yujia Qin, Yankai Lin, Jing Yi, Jiajie Zhang, Xu Han, Zhengyan Zhang, Yusheng Su, Zhiyuan Liu, Peng Li, Maosong Sun, and Jie Zhou. Knowledge inheritance for pre-trained language models. In Marine Carpuat, Marie-Catherine de Marneffe, and Iván Vladimir Meza Ruíz, editors, *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2022, Seattle, WA, United States, July 10-15, 2022*, pages 3921–3937. Association for Computational Linguistics, 2022.

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

[60] Subhadarshi Panda, Anjali Agrawal, Jeewon Ha, and Benjamin Bloch. 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*, pages 88–93, 2021.

[61] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[62] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[63] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *European conference on computer vision*, pages 213–229. Springer, 2020.

[64] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable DETR: deformable transformers for end-to-end object detection. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021.

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

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

[67] Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3208–3216. AAAI Press, 2021.

[68] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *European Conference on Computer Vision*, pages 121–137. Springer, 2020.

[69] Hao Tan and Mohit Bansal. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 2066–2080. Association for Computational Linguistics, 2020.

[70] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. *CoRR*, abs/1505.04870, 2015.

[71] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

[72] Matt Gardner, Joel Grus, Mark Neumann, Oyvind Tafjord, Pradeep Dasigi, Nelson F. Liu, Matthew E. Peters, Michael Schmitz, and Luke Zettlemoyer. Allennlp: A deep semantic natural language processing platform. *CoRR*, abs/1803.07640, 2018.

[73] Emily M Bender and Alexander Koller. 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*, pages 5185–5198, 2020.

[74] 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. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 8718–8735. Association for Computational Linguistics, 2020.

[75] Wei-Ning Hsu, Yao-Hung Hubert Tsai, Benjamin Bolte, Ruslan Salakhutdinov, and Abdelrahman Mohamed. 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)*, pages 6533–6537. IEEE, 2021.

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

[77] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020.

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

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

[80] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. OFA: unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 23318–23340. PMLR, 2022.

[81] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020.

[82] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Wanxiang Che, Min Zhang, and Lidong Zhou. Layoutlmv2: Multi-modal pre-training for visually-rich document understanding. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 2579–2591. Association for Computational Linguistics, 2021.

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

[84] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and CV Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520. IEEE, 2019.

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

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

[87] Xiangxiang Chu, Bo Zhang, Zhi Tian, Xiaolin Wei, and Huaxia Xia. Do we really need explicit position encodings for vision transformers? *CoRR*, abs/2102.10882, 2021.

[88] Nayu Liu, Xian Sun, Hongfeng Yu, Wenkai Zhang, and Guangluan Xu. 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)*, pages 1834–1845, 2020.

[89] Shruti Palaskar, Jindrich Libovický, Spandana Gella, and Florian Metze. Multimodal abstractive summarization for how2 videos. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6587–6596. Association for Computational Linguistics, 2019.

[90] Tiezheng Yu, Wenliang Dai, Zihan Liu, and Pascale Fung. Vision guided generative pre-trained language models for multimodal abstractive summarization. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3995–4007. Association for Computational Linguistics, 2021.

[91] Ramon Sanabria, Ozan Caglayan, Shruti Palaskar, Desmond Elliott, Loïc Barrault, Lucia Specia, and Florian Metze. How2: A large-scale dataset for multimodal language understanding. *CoRR*, abs/1811.00347, 2018.

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

[93] Shaoshi Ling and Yuzong Liu. Decoar 2.0: Deep contextualized acoustic representations with vector quantization. *CoRR*, abs/2012.06659, 2020.

[94] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. LRS3-TED: a large-scale dataset for visual speech recognition. *CoRR*, abs/1809.00496, 2018.

[95] Takaki Makino, Hank Liao, Yannis Assael, Brendan Shillingford, Basilio Garcia, Otavio Braga, and Olivier Siohan. Recurrent neural network transducer for audio-visual speech recognition. In *2019 IEEE automatic speech recognition and understanding workshop (ASRU)*, pages 905–912. IEEE, 2019.

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

[97] Robert Rehr and Timo Gerkmann. 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):357–366, 2017.

[98] Timothy F. Cootes, Gareth J. Edwards, and Christopher J. Taylor. Active appearance models. *IEEE Transactions on pattern analysis and machine intelligence*, 23(6):681–685, 2001.

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

[100] Hang Zhao, Chuang Gan, Andrew Rouditchenko, Carl Vondrick, Josh McDermott, and Antonio Torralba. The sound of pixels. In *Proceedings of the European conference on computer vision (ECCV)*, pages 570–586, 2018.

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

[102] José Camacho-Collados and Mohammad Taher Pilehvar. On the role of text preprocessing in neural network architectures: An evaluation study on text categorization and sentiment analysis. In Tal Linzen, Grzegorz Chrupala, and Afra Alishahi, editors, *Proceedings of the Workshop: Analyzing and Interpreting Neural Networks for NLP, BlackboxNLP@EMNLP 2018, Brussels, Belgium, November 1, 2018*, pages 40–46. Association for Computational Linguistics, 2018.

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

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

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

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

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

[108] Firoj Alam, Ferda Ofli, and Muhammad Imran. Crisismmd: Multimodal twitter datasets from natural disasters. In *Twelfth international AAAI conference on web and social media*, 2018.

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

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

[111] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 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)*, pages 5636–5646, 2019.

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

[113] Xinyi Zhou, Jindi Wu, and Reza Zafarani. SAFE: similarity-aware multi-modal fake news detection. In Hady W. Lauw, Raymond Chi-Wing Wong, Alexandros Ntoulas, Ee-Peng Lim, See-Kiong Ng, and Sinno Jialin Pan, editors, *Advances in Knowledge Discovery and Data Mining - 24th Pacific-Asia Conference, PAKDD 2020, Singapore, May 11-14, 2020, Proceedings, Part II*, volume 12085 of *Lecture Notes in Computer Science*, pages 354–367. Springer, 2020.

[114] Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. From recognition to cognition: Visual commonsense reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6720–6731, 2019.

[115] Dandan Song, Siyi Ma, Zhanchen Sun, Sicheng Yang, and Lejian Liao. Kvl-bert: Knowledge enhanced visual-and-linguistic bert for visual commonsense reasoning. *Knowledge-Based Systems*, 230:107408, 2021.

[116] Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 5099–5110. Association for Computational Linguistics, 2019.

[117] Zhicheng Huang, Zhaoyang Zeng, Bei Liu, Dongmei Fu, and Jianlong Fu. Pixel-bert: Aligning image pixels with text by deep multi-modal transformers. *CoRR*, abs/2004.00849, 2020.

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

[119] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and VQA. *CoRR*, abs/1909.11059, 2019.

[120] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[121] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. 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*, pages 5659–5667,

2017.

[122] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-critical sequence training for image captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7008–7024, 2017.

[123] Qi Wu, Damien Teney, Peng Wang, Chunhua Shen, Anthony R. Dick, and Anton van den Hengel. Visual question answering: A survey of methods and datasets. *CoRR*, abs/1607.05910, 2016.

[124] Christel Chappuis, Sylvain Lobry, Benjamin Kellenberger, Bertrand Le Saux, and Devis Tuia. How to find a good image-text embedding for remote sensing visual question answering? *CoRR*, abs/2109.11848, 2021.

[125] Tanzila Rahman, Shih-Han Chou, Leonid Sigal, and Giuseppe Carenini. An improved attention for visual question answering. *CoRR*, abs/2011.02164, 2020.

[126] Sanjay Subramanian, Sameer Singh, and Matt Gardner. Analyzing compositionality in visual question answering. In *ViGIL@NeurIPS*, 2019.

[127] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. *CoRR*, abs/1906.00067, 2019.

[128] Paul Pu Liang, Yiwei Lyu, Xiang Fan, Zetian Wu, Yun Cheng, Jason Wu, Leslie Chen, Peter Wu, Michelle A. Lee, Yuke Zhu, Ruslan Salakhutdinov, and Louis-Philippe Morency. Multibench: Multiscale benchmarks for multimodal representation learning. *CoRR*, abs/2107.07502, 2021.

[129] Xingjian Shi, Jonas Mueller, Nick Erickson, Mu Li, and Alexander J. Smola. Benchmarking multimodal automl for tabular data with text fields. *CoRR*, abs/2111.02705, 2021.

[130] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Attentive explanations: Justifying decisions and pointing to the evidence. *CoRR*, abs/1612.04757, 2016.

[131] Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. Don't just assume; look and answer: Overcoming priors for visual question answering. *CoRR*, abs/1712.00377, 2017.

[132] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[133] Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona, and Serge Belongie. The caltech-ucsd birds-200-2011 dataset. *California Institute of Technology*, 2011.

[134] Tao Xu, Pengchuan Zhang, Qiuyuan Huang, Han Zhang, Zhe Gan, Xiaolei Huang, and Xiaodong He. Attngan: Fine-grained text to image generation with attentional generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1316–1324, 2018.

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

[136] Triantafyllos Afouras, Joon Son Chung, and Andrew Zisserman. Deep lip reading: A comparison of models and an online application. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 3514–3518. ISCA, 2018.

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

[138] Wei Ping, Kainan Peng, Andrew Gibiansky, Sercan Ömer Arik, Ajay Kannan, Sharan Narang, Jonathan Raiman, and John Miller. Deep voice 3: Scaling text-to-speech with convolutional sequence learning. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.

[139] Jonathan Shen, Ruoming Pang, Ron J Weiss, Mike Schuster, Navdeep Jaitly, Zongheng Yang, Zhifeng Chen, Yu Zhang, Yuxuan Wang, Rj Skerrv-Ryan, et al. Natural tts synthesis by conditioning wavenet on mel spectrogram predictions. In *2018 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 4779–4783. IEEE, 2018.

[140] Ariel Ephrat and Shmuel Peleg. Vid2speech: speech reconstruction from silent video. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5095–5099. IEEE, 2017.

[141] Hassan Akbari, Himani Arora, Liangliang Cao, and Nima Mesgarani. Lip2audspec: Speech reconstruction from silent lip movements video. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2516–2520. IEEE, 2018.

[142] Konstantinos Vougioukas, Pingchuan Ma, Stavros Petridis, and Maja Pantic. Video-driven speech reconstruction using generative adversarial networks. In Gernot Kubin and Zdravko Kacic, editors, *Interspeech 2019, 20th Annual Conference of the International Speech Communication Association, Graz, Austria, 15-19 September 2019*, pages 4125–4129. ISCA, 2019.

[143] Yanyuan Qiao, Chaorui Deng, and Qi Wu. Referring expression comprehension: A survey of methods and datasets. *IEEE Transactions on Multimedia*, 23:4426–4440, 2021.

[144] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations. In *International Conference on Learning Representations*, 2020.

[145] Wei-Hung Weng and Peter Szolovits. Representation learning for electronic health records. *CoRR*, abs/1909.09248, 2019.

[146] Ozan Caglayan, Pranava Madhyastha, Lucia Specia, and Loïc Barrault. Probing the need for visual context in multimodal machine translation. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4159–4170. Association for Computational Linguistics, 2019.

[147] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

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

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

[150] Amir Zadeh and Paul Pu. 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)*, 2018.

[151] Alistair EW 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. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9, 2016.

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

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

[154] Pushmeet Kohli Nathan Silberman, Derek Hoiem and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *ECCV*, 2012.

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

[156] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016.

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

[158] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766, 2017.

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

[160] Catalina Cangea, Eugene Belilovsky, Pietro Liò, and Aaron C. Courville. Videonavqa: Bridging the gap between visual and embodied question answering. In *Visually Grounded Interaction and Language (ViGIL), NeurIPS 2019 Workshop, Vancouver, Canada, December 13, 2019*, 2019.

[161] Kushal Kafle and Christopher Kanan. An analysis of visual question answering algorithms. In *ICCV*, 2017.

[162] Holger Caesar, Varun Bankiti, Alex H. Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020.

[163] Maria-Elena Nilsback and Andrew Zisserman. Automated flower classification over a large number of classes. In *Indian Conference on Computer Vision, Graphics and Image Processing*, Dec 2008.

[164] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Mosi: multimodal corpus of sentiment intensity and subjectivity analysis in online opinion videos. *arXiv preprint arXiv:1606.06259*, 2016.

[165] Yuanhang Zhang, Shuang Yang, Jingyun Xiao, Shiguang Shan, and Xilin Chen. 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)*, pages 356–363. IEEE, 2020.

[166] Maurice Fallon, Hordur Johannsson, Michael Kaess, and John J Leonard. The mit stata center dataset. *The International Journal of Robotics Research*, 32(14):1695–1699, 2013.

[167] Alexei Baevski, Wei-Ning Hsu, Qiantong Xu, Arun Babu, Jiatao Gu, and Michael Auli. data2vec: A general framework for self-supervised learning in speech, vision and language. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato, editors, *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA*, volume 162 of *Proceedings of Machine Learning Research*, pages 1298–1312. PMLR, 2022.

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

[169] Mingyang Zhou, Luowei Zhou, Shuohang Wang, Yu Cheng, Linjie Li, Zhou Yu, and Jingjing Liu. Uc2: Universal cross-lingual cross-modal vision-and-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4155–4165, 2021.

[170] Minheng Ni, Haoyang Huang, Lin Su, Edward Cui, Taroon Bharti, Lijuan Wang, Dongdong Zhang, and Nan Duan. M3p: Learning universal representations via multitask multilingual multimodal pre-training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3977–3986, 2021.

[171] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *arXiv preprint arXiv:2109.05125*, 2021.

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

[173] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021.

[174] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

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

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

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

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

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