Multimodal Learning With Transformers: A Survey

Peng Xu, Xiatian Zhu, and David A. Clifton

(Survey Paper)

***Abstract*—Transformer is a promising neural network learner, and has achieved great success in various machine learning tasks. Thanks to the recent prevalence of multimodal applications and Big Data, Transformer-based multimodal learning has become a hot topic in AI research. This paper presents a comprehensive survey of Transformer techniques oriented at multimodal data. The main contents of this survey include: (1) a background of multimodal learning, Transformer ecosystem, and the multimodal Big Data era, (2) a systematic review of *Vanilla* Transformer, Vision Transformer, and multimodal Transformers, from a geometrically topological perspective, (3) a review of multimodal Transformer applications, via two important paradigms, i.e., for multimodal pretraining and for specific multimodal tasks, (4) a summary of the common challenges and designs shared by the multimodal Transformer models and applications, and (5) a discussion of open problems and potential research directions for the community.**

***Index Terms*—Multimodal learning, transformer, introductory, taxonomy, deep learning, machine learning.**

I. Introduction

THE initial inspiration of Artificial Intelligence (AI) is to imitate human perception, e.g., seeing, hearing, touching, smelling. In general, a modality is often associated with a specific sensor that creates a unique communication channel, such as vision and language [1]. In humans, a fundamental mechanism in our sensory perception is the ability to leverage multiple modalities of perception data collectively in order to engage ourselves properly with the world under dynamic unconstrained circumstances, with each modality serving as a distinct information source characterized by different statistical properties. For example, an image gives the visual appearance of an "elephants playing in water" scene via thousands of pixels, whilst the corresponding text describes this moment with a sentence using discrete words. Fundamentally, a multimodal AI system needs to ingest, interpret, and reason about multimodal information sources to realize similar human level perception abilities. Multimodal learning (MML) is a general approach to building AI models that can extract and relate information from multimodal data [1].

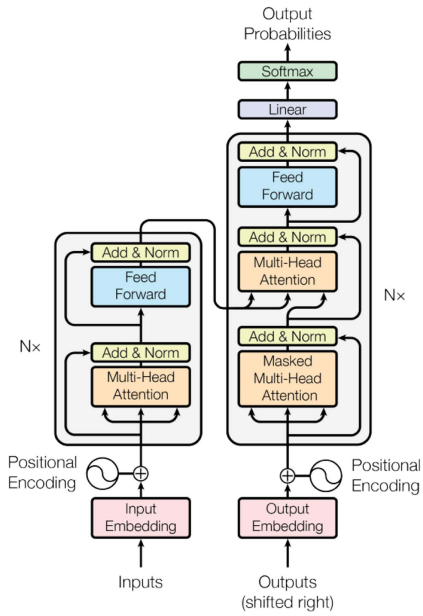


Fig. 1. Overview of Transformer [2].

This survey focuses on multimodal learning with Transformers [2] (as demonstrated in Fig. 1), inspired by their intrinsic advantages and scalability in modelling different modalities (e.g., language, visual, auditory) and tasks (e.g., language translation, image recognition, speech recognition) with fewer modality-specific architectural assumptions (e.g., translation invariance and local grid attention bias in vision) [3]. Concretely, the input to a Transformer could encompass one or multiple sequences of tokens, and each sequence's attribute (e.g., the modality label, the sequential order), naturally allowing for MML without architectural modification [4]. Further, learning per-modal specificity and inter-modal correlation can be simply realized by controlling the input pattern of self-attention. Critically, there is a recent surge of research attempts and activities across distinct disciplines exploring the Transformer architectures, resulting in a large number of novel MML methods being developed in recent years, along with significant and diverse advances in various areas [4], [5], [6], [7], [8]. This calls for a timely review

Manuscript received 30 January 2023; revised 17 April 2023; accepted 26 April 2023. Date of publication 11 May 2023; date of current version 5 September 2023. This work was supported in part by RAEng Research Chair, and NIHR Research Professorship, in part by NIHR Oxford Biomedical Research Centre, in part by InnoHK Centre for Cerebro-cardiovascular Health Engineering, and in part by Oxford Pandemic Sciences Institute. Recommended for acceptance by T. Hassner. *(Corresponding author: David A. Clifton.)*

Peng Xu is with Tsinghua University, Beijing 100084, China (e-mail: peng_xu@tsinghua.edu.cn).

Xiatian Zhu is with the University of Surrey, GU2 7XH Guildford, U.K. (e-mail: eddy.zhuxt@gmail.com).

David A. Clifton is with the University of Oxford, OX1 4BH Oxford, U.K., and also with the Oxford Suzhou Centre for Advanced Research, Suzhou 215123, China (e-mail: davidc@robots.ox.ac.uk).

This article has supplementary downloadable material available at https://doi.org/10.1109/TPAMI.2023.3275156, provided by the authors.

Digital Object Identifier 10.1109/TPAMI.2023.3275156

and summary of representative methods to enable researchers to understand the global picture of the MML field across related disciplines and more importantly to capture a holistic structured picture of current achievements as well as major challenges.

Taxonomy: For better readability and reachability from and across different disciplines, we adopt a two-tier structured taxonomy based on the application and challenge dimensions respectively. This has several benefits: (1) Researchers with expertise in specific applications can find those applications appropriate to their own research domain before connecting to other related domains. (2) Similar model designs and architectures developed in different domains can be summarized in an abstract, formula-driven perspective so that the mathematical ideas of various models formed in different applications can be correlated and contrasted on common ground, crossing domain-specific restrictions. Crucially, our taxonomy offers an interesting stereo-view of individual works with the insights in both application specificity and formulation generality. It is hoped that this can help to break down domain boundaries and foster more effective idea communication and exchange across modalities. By using the prompt modelling strategy [9], [10] as a basis for investigation, we also include the classical classification problem (e.g., image classification) – usually regarded as a single modality learning application in conventional MML surveys [1], [11], [12] – as a special MML application. This has the potential to significantly enrich MML, as the classification problem is an AI topic amongst the most extensive studies in the literature [13].

Scope: This survey will discuss the multimodality specific designs of Transformer architecture including, but not limited to, the following modalities: RGB image [5], depth image [14], multispectral image [15], video [7], audio/speech/music [14], [16], [17], table [18], scene graph/layout [19], [20], [21], [22], pose skeleton [23], SQL [24], [25], recipe [26], programming language [27], sign language [28], [29], [30], point cloud [31], symbolic knowledge (graph) [32], [33], multimodal knowledge graph [34], sketch drawing [35], [36], [37], [38], 3D object/scene [39], [40], [41], document [42], [43], programming code [44] and Abstract Syntax Tree (AST) – a kind of graph [45], optical flow [46], medical knowledge (e.g., diagnosis code ontology [47]). Note that this survey will not discuss the multimodal papers where Transformer is used simply as the feature extractor without multimodal designs.

Related Surveys: We relate this paper to existing surveys of the two specific dimensions MML and Transformers. There exist a few MML surveys [1], [11], [12]. In particular, [1] proposed a structured, acknowledged taxonomy by five challenges, which we also adopt as part of our structure. Unlike [1], [11], and [12], which review general machine learning models, we instead focus on Transformer architectures and their self-attention mechanisms. Several surveys dedicated to Transformers have been recently introduced, with a range of emphases including general Transformers [48], efficient designs [49], visualization [50], computer vision tasks [51], [52], [53], [54], medical imaging [55], video tasks [56], and vision language pretraining [57]. While [51], [53], [54], [55] consider MML, their reviews are somewhat limited in the scope, taxonomy, and coverage. To our knowledge, only a few surveys on video-language pretraining (VLP) [57], [58], [59] are relevant to MML. However, VLP is only a subdomain of MML. In this survey, we focus solely on the intersection of multimodal learning and Transformers.

Features: To our knowledge, this paper is the first comprehensive review of the state of Transformer based multimodal machine learning. The major features of this survey include

(1) We highlight that Transformers have the advantage that they can work in a modality-agnostic way. Thus, they are compatible with various modalities (and combinations of modalities). To support this view, we, for the first time, offer an understanding of the intrinsic traits of Transformers in a multimodal context from a geometrically topological perspective. We suggest that self-attention be treated as a graph style modelling, which models the input sequence (both uni-modal and multimodal) as a fully-connected graph. Specifically, self-attention models the embedding of arbitrary tokens from an arbitrary modality as a graph node.

(2) We discuss the key components of Transformers in a multimodal context as mathematically as possible.

(3) Based on Transformers, cross-modal interactions (e.g., fusion, alignment) are essentially processed by self-attention and its variants. In this paper, we extract the mathematical essence and formulations of Transformer based MML practices, from the perspective of self-attention designs.

Contributions: Having presented our review of the landscape of multimodal learning, Transformer ecosystem, and multimodal Big Data era in Section II, we summarize our main contributions as the follows.

1) In Section III, we present a systematic reviewing of *Vanilla* Transformer, Vision Transformer, and multimodal Transformers, from a geometrically topological perspective.
2) We contribute a taxonomy for Transformer based MML from two complementary perspectives, i.e., application based and challenge based. In Section IV, we provide a review of multimodal Transformer applications, via two important paradigms, i.e., for multimodal pretraining and for specific multimodal tasks. In Section V, we summarize the common challenges and designs shared by the various multimodal Transformer models and applications.
3) In Section VI, we discuss current bottlenecks, existing problems, and potential research directions for Transformer based MML.

II. Background

A. Multimodal Learning (MML)

MML [1], [60], [61] has been an important research area in recent decades; an early multimodal application – audio-visual speech recognition was studied in 1980s [62]. MML is key to human societies. The world we humans live in is a multimodal environment, thus both our observations and behaviours are multimodal [63]. For instance, an AI navigation robot needs multimodal sensors to perceive the real-world environment [64], [65], [66], e.g., camera, LiDAR, radar, ultrasonic, GNSS, HD Map, odometer. Furthermore, human behaviours, emotions, events, actions, and humour are multimodal, thus various human-centred MML tasks are widely studied, including multimodal

emotion recognition [67], multimodal event representation [68], understanding multimodal humor [69], face-body-voice based video person-clustering [70], *etc*.

Thanks to the development of the internet and a wide variety of intelligent devices in recent years, increasing amounts of multimodal data are being transmitted over the internet, thus an increasing number of multimodal application scenarios are emerging. In modern life, we can see various multimodal applications, including commercial services (e.g., e-commerce/commodity retrieval [71], vision-and-language navigation (VLN) [72], [73], [74], [75], [76]), communication (e.g., lip reading [77], sign language translation [28], [29]), human-computer interaction [78], healthcare AI [79], [80], surveillance AI [81], *etc*.

Moreover, in the era of Deep Learning, deep neural networks greatly promote the development of MML, and Transformers [2] are a highly competitive architecture family, bringing new challenges and opportunities to MML. In particular, the recent success of large language models and their multimodal derivatives [82], [83], [84], [85], [86] further demonstrates the potential of Transformers in multimodal foundation models.

B. Transformers: A Brief History and Milestones

Transformers are emerging as promising learners. *Vanilla* Transformer [2] benefits from a self-attention mechanism, and is a breakthrough model for sequence-specific representation learning that was originally proposed for NLP, achieving the state-of-the-art on various NLP tasks. Following the great success of *Vanilla* Transformer, a lot of derivative models have been proposed, e.g., BERT [4], BART [87], GPT [88], Longformer [43], Transformer-XL [89], XLNet [90].

Transformers currently stand at the dominant position in NLP domains, and this motivates researchers try to apply Transformers to other modalities, such as visual domains. In early attempts for visual domain, the general pipeline is "CNN features + standard Transformer encoder", and researchers achieved BERT-style pretraining, via preprocessing raw images by resizing to a low resolution and reshaping into a 1D sequence [91].

Vision Transformer (ViT) [5] is a seminal work that contributes an end-to-end solution by applying the encoder of Transformer to images. Both ViT and its variants have been widely applied to various computer vision tasks, including low-level tasks [92], recognition [93], detection [94], segmentation [95], *etc*, and also work well for both supervised [93] and self-supervised [96], [97], [98] visual learning. Moreover, some recently-released works provide further theoretical understanding for ViT, e.g., its internal representation robustness [99], the continuous behaviour of its latent representation propagation [100], [101].

Motivated by the great success of Transformer, VideoBERT [7] is a breakthrough work that is the first work to extend Transformer to the multimodal tasks. VideoBERT demonstrates the great potential of Transformer in multimodal context. Following VideoBERT, a lot of Transformer based multimodal pretraining models (e.g., ViLBERT [102], LXMERT [103], VisualBERT [104], VL-BERT [105], UNITER [106], CBT [107], Unicoder-VL [108], B2T2 [109], VLP [110], 12-in-1 [111], Oscar [112], Pixel-BERT [113], ActBERT [114], ImageBERT [115], HERO [116], UniVL [117]) have become research topics of increasing interest in the field of machine learning.

In 2021, CLIP [9] was proposed. It is a new milestone that uses multimodal pretraining to convert classification as a retrieval task that enables the pretrained models to tackle zero-shot recognition. Thus, CLIP is a successful practice that makes full use of large-scale multimodal pretraining to enable zero-shot learning. Recently, the idea of CLIP is further studied, e.g., CLIP pretrained model based zero-shot semantic segmentation [118], ALIGN [119], CLIP-TD [120], ALBEF [121], and CoCa [122].

C. Multimodal Big Data

In the past decade, with the rapid development of internet applications such as social media and online retail, massive multimodal datasets have been proposed, e.g., Conceptual Captions [123], COCO [124], VQA [125], Visual Genome [126], SBU Captions [127], Cooking312K [7], LAIT [115], e-SNLI-VE [128], ARCH [129], Adversarial VQA [130], OTT-QA [18], MULTIMODALQA (MMQA) [131], VALUE [132], Fashion IQ [133], LRS2-BBC [134], ActivityNet [135], VisDial [136].

Some emergent new trends among the recently released multimodal datasets are:

1) Data scales are larger. Various recently released datasets are million-scale, e.g., Product1M [137], Conceptual 12M [138], RUC-CAS-WenLan [139] (30 M), HowToVQA69M [140], HowTo100M [141], ALT200M [142], LAION-400M [143].

2) More modalities. In addition to the general modalities of vision, text, and audio, further diverse modalities are emerging, e.g., Pano-AVQA [144] – the first large-scale spatial and audio-visual question answering dataset on 360° videos, YouTube-360 (YT-360) [145] (360° videos), AIST++ [146] (a new multimodal dataset of 3D dance motion and music), Artemis [147] (affective language for visual arts). In particular, MultiBench [148] provides a dataset including 10 modalities.

3) More scenarios. In addition to common caption and QA datasets, more applications and scenarios have been studied, e.g., CIRR [149] (real-life images), Product1M [137], Bed and Breakfast (BnB) [150] (vision-and-language navigation), M3A [151] (financial dataset), X-World [152] (autonomous drive).

4) Tasks are more difficult. Beyond the straightforward tasks, more abstract multimodal tasks are proposed, e.g., MultiMET [153] (a multimodal dataset for metaphor understanding), Hateful Memes [154] (hate speech in multimodal memes).

5) Instructional videos have become increasingly popular, e.g., cooking video YouCookII [155]. Aligning a sequence of instructions to a video of someone carrying out a task is an example of a powerful pretraining pretext task [7], [156]. Pretext tasks are pre-designed problems to force the models to learn representation by solving them.

Similar to other deep neural network architectures, Transformers are also data hungry. Therefore, their high-capacity models and multimodal Big Data basis co-create the prosperity

of the Transformer based multimodal machine learning. For instance, Big Data bring zero-shot learning capability to VLP Transformer models.

III. Transformers

In this section, we use mathematical formulations to review the key techniques of *Vanilla* Transformer [2], Vision Transformer [5], and multimodal Transformers,[1] including tokenized inputs, self-attention, multi-head attention, basic Transformer layers/blocks, *etc*. We highlight that *Vanilla* Transformers can be understood from a geometrically topological perspective [157], because due to the self-attention mechanism, given each tokenized input from any modalities, *Vanilla* self-attention (Transformer) can model it as a fully-connected graph in topological geometry space [158]. Compared with other deep networks (for instance, CNN is restricted in the aligned grid spaces/matrices), Transformers intrinsically have a more general and flexible modelling space. This is a notable advantage of Transformers for multimodal tasks. Sections III-A, III-B, and III-C will review the key designs of *Vanilla* Transformer, Vision Transformer, and multimodal Transformers, respectively.

A. Vanilla *Transformer*

Vanilla Transformer has an encoder-decoder structure and is the origin of the Transformer-based research field. It takes tokenized input (see Section III-A1). Both its encoder and decoder are stacked by the Transformer layers/blocks, as demonstrated in Fig. 1. Each block has two sub-layers, i.e., a multi-head self-attention (MHSA) layer (see Section III-A2) and a position-wise fully-connected feed-forward network (FFN) (see Section III-A3). To help the back propagation of the gradient, both MHSA and FFN use Residual Connection [159] (given an input x, the residual connection of any mapping $f(\cdot)$ is defined as $x \leftarrow f(x) + x$), followed by normalization layer. Thus, assuming that the input tensor is $\mathbf{Z}$, the output of MHSA and FFN sub-layers can be formulated as:

$$\mathbf{Z} \leftarrow N(sublayer(\mathbf{Z}) + \mathbf{Z}), \tag{1}$$

where $sublayer(\cdot)$ is the mapping implemented by the sub-layer itself and $N(\cdot)$ denotes normalization, e.g., $BN(\cdot)$ [160], $LN(\cdot)$ [161].

Discussion: There is an important unsolved problem that is post-normalization versus pre-normalization. The original *Vanilla* Transformer uses post-normalization for each MHSA and FFN sub-layer. However, if we consider this from the mathematical perspective, pre-normalization makes more sense [162]. This is similar to the basic principle of the theory of matrix, that normalization should be performed before projection, e.g., Gram–Schmidt process.[2] This problem should be studied further by both theoretical research and experimental validation.

[1] In this survey, "multimodal Transformer" means "Transformer in multimodal learning context".

[2] https://en.wikipedia.org/wiki/Gram%E2%80%93Schmidt_process

1) Input Tokenization: Tokenization Vanilla: Transformer was originally proposed for machine translation as a sequence-to-sequence model, thus it is straightforward to take the vocabulary sequences as input. As mentioned previously, the original self-attention can model an arbitrary input as a fully-connected graph, independently of modalities. Specifically, both *Vanilla* and variant Transformers take in the tokenized sequences, where each token can be regarded as a node of the graph.

Special/Customized Tokens: In Transformers, various special/customized tokens can be semantically defined as placeholders in the token sequences, e.g., mask token `[MASK]` [4]. Some common special tokens are summarized in appendix, available in the online supplemental material. Special tokens can be used in both uni-modal and multimodal Transformers.

Position Embedding: Position embeddings are added to the token embeddings to retain positional information [4]. *Vanilla* Transformer uses sine and cosine functions to produce position embedding. To date, various implementations of position embedding have been proposed. The concrete solutions are outside the focus of this survey.

Discussion: The main advantages of input tokenization include the following:

1) Tokenization is a more general approach from a geometrically topological perspective, achieved by minimizing constraints caused by different modalities. In general, every modality has intrinsic constraints on modelling. For instance, sentences have sequential structures that are well-suited by RNN, and photos are restricted in aligned grid matrices that CNN works well for. Tokenization helps Transformers inherently to process different modalities universally via irregular sparse structures. Thus even *Vanilla* Transformer can encode multimodal inputs flexibly by just concatenation, weighted summation, even without any multimodal tailor-made modifications.
2) Tokenization is a more flexible approach to organize the input information via concatenation/stack, weighted summation, *etc*. *Vanilla* Transformer injects temporal information to the token embedding by summing position embedding. For instance, when use Transformer to model free-hand sketch drawing [163], each input token can integrate various drawing stroke patterns, e.g., stroke coordinates, stroke ordering, pen state (start/end).
3) Tokenization is compatible with the task-specific customized tokens, e.g., `[MASK]` token [4] for Masked Language Modelling, `[CLASS]` token [5] for classification.

Discussion: How to understand position embedding to Transformers is an open problem. It can be understood as a kind of implicit coordinate basis of feature space, to provide temporal or spatial information to the Transformer. For cloud point [164] and sketch drawing stroke [163], their token element is already a coordinate, meaning that position embedding is optional, not necessary. Furthermore, position embedding can be regarded as a kind of general additional information. In other words, from a mathematical point of view, any additional information can be added, such as detail of the manner of position embedding, e.g., the pen state of sketch drawing stroke [163], cameras and viewpoints in surveillance [165]. There is a comprehensive

survey [166] discussing the position information in Transformers. For both sentence structures (sequential) and general graph structures (sparse, arbitrary, and irregular), position embeddings help Transformers to learn or encode the underlying structures. Considered from the mathematical perspective of self-attention, i.e., scaled dot-product attention, attentions are invariant to the positions of words (in text) or nodes (in graphs), if position embedding information is missing. Thus, in most cases, position embedding is necessary for Transformers.

2) Self-Attention and Multi-Head Self-Attention: The core component of *Vanilla* Transformer is the Self-Attention (SA) operation [2] that is also termed "Scaled Dot-Product Attention". Assume that $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots] \in \mathbb{R}^{N \times d}$ is an input sequence of N elements/tokens, and an optional preprocessing is positional encoding by point-wise summation $\mathbf{Z} \leftarrow \mathbf{X} \oplus PositionEmbedding$ or concatenation $\mathbf{Z} \leftarrow concat(\mathbf{X}, PositionEmbedding)$.

Self-Attention (SA): After preprocessing, embedding $\mathbf{Z}$ will go through three projection matrices ($\mathbf{W}^Q \in \mathbb{R}^{d \times d_q}$, $\mathbf{W}^K \in \mathbb{R}^{d \times d_k}$, and $\mathbf{W}^V \in \mathbb{R}^{d \times d_v}$, $d_q = d_k$) to generate three embeddings $\mathbf{Q}$ (Query), $\mathbf{K}$ (Key), and $\mathbf{V}$ (Value):

$$\mathbf{Q} = \mathbf{Z}\mathbf{W}^Q, \mathbf{K} = \mathbf{Z}\mathbf{W}^K, \mathbf{V} = \mathbf{Z}\mathbf{W}^V. \tag{2}$$

The output of self-attention is defined as

$$\mathbf{Z} = SA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_q}}\right)\mathbf{V}. \tag{3}$$

Given an input sequence, self-attention allows each element to attend to all the other elements, so that self-attention encodes the input as a fully-connected graph. Therefore, the encoder of *Vanilla* Transformer can be regarded as a fully-connected GNN encoder, and the Transformer family has the non-local ability of global perception, similar to the Non-Local Network [167].

Masked Self-Attention (MSA): In practice, modification of self-attention is needed to help the decoder of Transformer to learn contextual dependence, to prevent positions from attending to subsequent positions, as

$$\mathbf{Z} = MSA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = Softmax\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_q}} \odot \mathbf{M}\right)\mathbf{V}, \tag{4}$$

where $\mathbf{M}$ is a masking matrix. For instance, in GPT [88], an upper triangular mask to enable look-ahead attention where each token can only look at the past tokens. Masking can be used in both encoder [163], [168] and decoder of Transformer, and has flexible implementations, e.g., 0-1 hard mask [163], soft mask [168].

In both uni-modal and multimodal practices, specific masks are designed based on domain knowledge and prior knowledge. Essentially, MSA is used to inject additional knowledge to Transformer models, e.g., [24], [163], [169], [170].

Multi-Head Self-Attention (MHSA): In practice, multiple self-attention sub-layers can be stacked in parallel and their concatenated outputs are fused by a projection matrix $\mathbf{W}$, to form a structure named Multi-Head Self-Attention:

$$\mathbf{Z} = MHSA(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = concat(\mathbf{Z}_1, \ldots, \mathbf{Z}_H)\mathbf{W}, \tag{5}$$

where each head $\mathbf{Z}_h = SA(\mathbf{Q}_h, \mathbf{K}_h\mathbf{V}_h)$ and $h \in [1, H]$, and $\mathbf{W}$ is a linear projection matrix. The idea of MHSA is a kind of ensemble. MHSA helps the model to jointly attend to information from multiple representation sub-spaces.

3) Feed-Forward Network: The output of the multi-head attention sub-layer will go through the position-wise Feed-Forward Network (FFN) that consists of successive linear layers with non-linear activation. For instance, a two-layer FFN can be formulated as

$$FFN(\mathbf{Z}) = \sigma(\mathbf{Z}\mathbf{W}_1 + \mathbf{b}_1)\mathbf{W}_2 + \mathbf{b}_2, \tag{6}$$

where $\mathbf{W}_1$, $\mathbf{b}_1$, $\mathbf{W}_2$, and $\mathbf{b}_2$ denote the weights and biases of the two linear transformations, while $\sigma(\cdot)$ is non-linear activation, e.g., ReLU$(\cdot)$ [171], $GELU(\cdot)$ [172]. In some Transformer literature, FFN is also termed Multi-Layer Perceptron (MLP).

B. Vision Transformer

Vision Transformer (ViT) [5] has an image-specific input pipeline in which the input image must be split into fixed-size (e.g., 16×16, 32×32) patches. After going through the linearly embedded layer and adding the position embeddings, all the patch-wise sequences will be encoded by a standard Transformer encoder. Given an image $\mathbf{X} \in \mathbb{R}^{H \times W \times C}$ (H height, W width, C channels), ViT needs to reshape $\mathbf{X}$ into a sequence of flattened 2D patches: $\mathbf{x}_p \in \mathbb{R}^{\mathbf{N} \times (\mathbf{P^2} \cdot \mathbf{C})}$, where $(P \times P)$ is the patch resolution and $N = HW/P^2$. To perform classification, a standard approach is to prepend an extra learnable embedding "classification token" `[CLASS]` to the sequence of embedded patches:

$$\mathbf{Z} \leftarrow concat(\texttt{[CLASS]}, \mathbf{X}\mathbf{W}), \tag{7}$$

where $\mathbf{W}$ denotes the projection.

C. Multimodal Transformers

Recently, a large number of Transformers have been studied extensively for various multimodal tasks, and shown to be compatible with various modalities in both discriminative and generative tasks.

In this section, we will review the key techniques/designs of the existing multimodal Transformer models, from the perspectives of multimodal input (Section III-C1), self-attention variants (Section III-C2), and network architectures (Section III-C3).

1) Multimodal Input: The Transformer family is a general architecture that can be formulated as a type of general graph neural network. Specifically, self-attention can process each input as a fully-connected graph, by attending to the global (non-local) patterns. Therefore, this intrinsic trait helps Transformers can work in a modality agnostic pipeline that is compatible with various modalities by treating the embedding of each token as a node of the graph.

Tokenization and Embedding Processing: Given an input from an arbitrary modality, users only need to perform two main steps, (1) tokenize the input, and (2) select an embedding space to represent the tokens, before inputting the data into Transformers. In practice, both the tokenizing input and selecting embedding for the token are vital for Transformers but highly

TABLE I
TOKENIZATION AND TOKEN EMBEDDING COMPARISON FOR MULTI-MODAL INPUTS FOR TRANSFORMERS

Modalities	Tokenization	Token Embeddings	References
RGB	RoI	CNN embedding	ViLBERT [102], LXMERT [103],
RGB	patch	linear projection	ViT [5]
video	clip of sampled frames	3D CNN embedding	VideoBERT [7], CBT [107], ActBERT [114]
video	sampled frame	2D CNN embedding	[173]
video	voxel of sampled frames	linear projection	VATT [174]
video	patch of sampled frame	linear projection	MBT [175]
360° video	clip of sampled frames	3D CNN embedding	AVSA [145]
audio	frame (mel-spectrogram)	CNN embedding	[173], AVSA [145]
audio	waveform segment	linear projection	VATT [174]
audio	spectrogram patch	linear projection	MBT [175]
speech/audio	waveform segment	1D-CNN (TCN) embedding	[176], FaceFormer [39]
speech	frame (mel-spectrogram)	linear projection and gated CNN embedding	Meta-StyleSpeech [177]
speech	frame (log-Mel filterbanks)	linear projection	AV-HuBERT [178]
speech/audio	frame (log power spectrum)	linear projection	VSET [179]
speech	frame (log-Mel filterbanks)	2D-CNN embedding	FAT-MLM [180]
music	frame (35-dim music feature)	linear projection	FACT [146]
text	word	learned embedding	*Vanilla* Transformer [2]
text	word	GNN embedding	MGNNS [181]
SQL database schema	table node, column node	GNN embedding	SpeechSQLNet [25]
textual question-graph	word node	GNN embedding	SADGA [24]
sketch	key point of stroke	linear projection and learnable embedding	Multi-Graph Transformer [163]
sketch	patch of picture	linear projection	RVT [182]
table	cell	learned embedding	[18]
3D point cloud	point	non-linear projection	Point Cloud Transformer [164]
source code	code	learned embedding	GraphCodeBERT [44]
data flow of source code	variable	learned embedding	GraphCodeBERT [44]
pose	key point	GCN embedding	TriBERT [183]
electronic health records (EHRs)	ICD code	GNN embedding	G-BERT [47]
electronic health records (EHRs)	ICD code	learned embedding	Med-BERT [184]
Gigapixel Whole Slide Images	patch	CNN embedding	MCAT [185]

"ICD": international classification of diseases.

flexible, with many alternatives. For instance, given an image, the solution of tokenizing and embedding is not unique. Users can choose or design tokenization at multiple granularity levels – coarse-grained versus fine-grained. e.g., use ROIs (obtained by an object detector) and CNN features as tokens and token embeddings [102], use patches and linear projection as tokens and token embeddings [5], or use graph node (obtained by object detector and graph generator) and GNN features as tokens and token embeddings [181]. Given a tokenization plan, the subsequent embedding approaches can be diverse. For example, for video input, a common tokenization is to treat the non-overlapping windows (down-sampled) over the video as tokens, and their embeddings can then be extracted by various 3D CNNs, e.g., VideoBERT [7], CBT [107], and UniVL [117] use S3D [186], ActBERT uses ResNet-3D [187].

Table I summarizes some common practices of multi-modal inputs for Transformers, including RGB, video, audio/speech/music, text, graph, *etc*.

Discussion: When considered from the perspective of geometric topology, each of the modalities listed in Table I can be regarded as a graph. An RGB image is essentially a neat grid graph in the pixel space. Both video and audio are clip/segment based graphs over a complex space involving temporal and semantic patterns. Both 2D and 3D drawing sketches [78], [163] are a kind of sparse graph if we consider their key points along the drawing strokes. Similar to sketches, the human pose also is a kind of graph. 3D point cloud is a graph in which each coordinate is a node. Other abstract modalities also can be interpreted as graphs, e.g., source code [44], data flow of source code [44], table [18], SQL database schema [25], text question graph [24], and electronic health records (EHRs) [184].

Token Embedding Fusion: In practice, Transformers allow each token position to contain multiple embeddings. This is essentially a kind of early-fusion of embeddings, for both uni-modal and multimodal Transformer models. (This will be discussed further in subsequent sections.) The most common fusion is the token-wise summing of the multiple embeddings, e.g., a specific token embedding $\oplus$ position embedding. Similar to the flexible tokenization, token embedding fusion is also flexible and widely applied to both uni-modal and multimodal Transformer applications. In [81], token-wise weighted summing is used to perform early-fusion of RGB and grey-scale images for multimodal surveillance AI. In particular, token embedding fusion has an important role in multimodal Transformer applications as various embeddings can be fused by token-wise operators, e.g., in VisualBERT [104] and Unicoder-VL [108], segment embeddings are token-wise added to indicate which modality (vision or language) each token is from, VL-BERT [105] injects global visual context to linguistic domain by "linguistic token embedding $\oplus$ full image visual feature embedding", InterBERT [188] adds location information for ROI by "ROI embedding $\oplus$ location embedding", in ImageBERT [115], five kinds of embeddings are fused "image embedding $\oplus$ position embedding $\oplus$ linguistic embedding $\oplus$ segment embedding $\oplus$ sequence position embedding".

2) Self-Attention Variants in Multimodal Context: In multimodal Transformers, cross-modal interactions (e.g., fusion, alignment) are essentially processed by self-attention and its variants. Thus, in this section, we will review the main multimodal modelling practices of Transformers, from a perspective of self-attention designs, including (1) early summation (token-wise, weighted), (2) early concatenation, (3) hierarchical

TABLE II
SELF-ATTENTION VARIANTS FOR MULTI-MODAL INTERACTION/FUSION

Self-Attention	Definitions	Streams	Formulations	Complexities	References
Early Summation	token-wise sum before Tfs	1	$\mathbf{Z} \leftarrow Tf(\alpha\mathbf{Z}_{(\mathrm{A})} \oplus \beta\mathbf{Z}_{(\mathrm{B})})$	$\mathcal{O}(N^2_{(\mathrm{A})})$	[46], [81]
Early Concat.	token sequence concat. before Tfs	1	$\mathbf{Z} \leftarrow Tf(\mathcal{C}(\mathbf{Z}_{(\mathrm{A})}, \mathbf{Z}_{(\mathrm{B})}))$	$\mathcal{O}((N_{(\mathrm{A})} + N_{(\mathrm{B})})^2)$	[7], [44], [178], [180]
Hierarchical Att.	2-stream Tfs followed by concat.	$2 \rightarrow 1$	$\mathbf{Z} \leftarrow Tf_3(\mathcal{C}(Tf_1(\mathbf{Z}_{(\mathrm{A})}), Tf_2(\mathbf{Z}_{(\mathrm{B})})))$	$\mathcal{O}((N_{(\mathrm{A})} + N_{(\mathrm{B})})^2)$	[146],
Hierarchical Att.	early concat. followed by 2-stream Tfs	$1 \rightarrow 2$	$\begin{cases} \mathcal{C}(\mathbf{Z}_{(\mathrm{A})}, \mathbf{Z}_{(\mathrm{B})}) \leftarrow Tf_1(\mathcal{C}(\mathbf{Z}_{(\mathrm{A})}, \mathbf{Z}_{(\mathrm{B})})), \\ \mathbf{Z}_{(\mathrm{A})} \leftarrow Tf_2(\mathbf{Z}_{(\mathrm{A})}), \\ \mathbf{Z}_{(\mathrm{B})} \leftarrow Tf_3(\mathbf{Z}_{(\mathrm{B})}). \end{cases}$	$\mathcal{O}((N_{(\mathrm{A})} + N_{(\mathrm{B})})^2)$	[188]
Cross-Attention	exchange query	2	$\begin{cases} \mathbf{Z}_{(\mathrm{A})} \leftarrow MHSA(\mathbf{Q}_{\mathrm{B}}, \mathbf{K}_{\mathrm{A}}, \mathbf{V}_{\mathrm{A}}) \\ \mathbf{Z}_{(\mathrm{B})} \leftarrow MHSA(\mathbf{Q}_{\mathrm{A}}, \mathbf{K}_{\mathrm{B}}, \mathbf{V}_{\mathrm{B}}) \end{cases}$	$\mathcal{O}(N^2_{(\mathrm{A})})$	[102], [144]
Cross-Att. to Con.	2-stream cross-att. followed by concat.	$2 \rightarrow 1$	$\begin{cases} \mathbf{Z}_{(\mathrm{A})} \leftarrow MHSA(\mathbf{Q}_{\mathrm{B}}, \mathbf{K}_{\mathrm{A}}, \mathbf{V}_{\mathrm{A}}) \\ \mathbf{Z}_{(\mathrm{B})} \leftarrow MHSA(\mathbf{Q}_{\mathrm{A}}, \mathbf{K}_{\mathrm{B}}, \mathbf{V}_{\mathrm{B}}) \\ \mathbf{Z} \leftarrow Tf(\mathcal{C}(\mathbf{Z}_{(\mathrm{A})}, \mathbf{Z}_{(\mathrm{B})})) \end{cases}$	$\mathcal{O}((N_{(\mathrm{A})} + N_{(\mathrm{B})})^2)$	[69] [137], [189]

α and β denote weightings. "ATT.": attention; "CONCAT."/"CON.": concatenation; "TFS": transformer layers. $N_{(\mathrm{A})}$ and $N_{(\mathrm{B})}$ denote the token sequence lengths of two modalities.

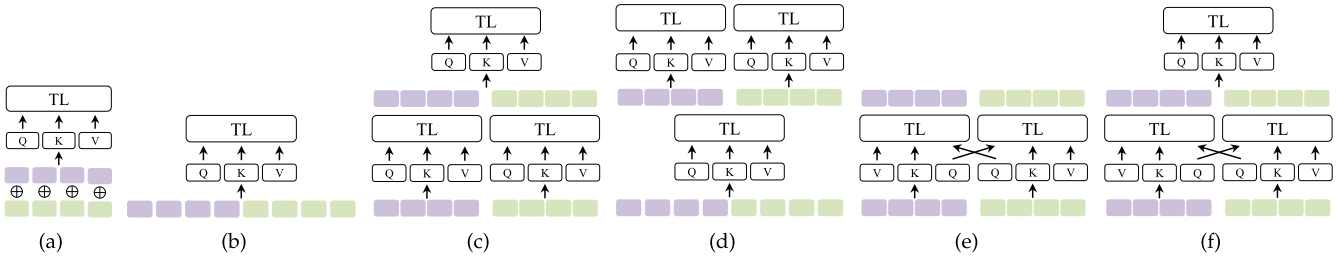


Fig. 2. Transformer-based cross-modal interactions: (a) Early Summation, (b) Early Concatenation, (c) Hierarchical Attention (multi-stream to one-stream), (d) Hierarchical Attention (one-stream to multi-stream), (e) Cross-Attention, and (f) Cross-Attention to Concatenation. "Q": Query embedding; "K": Key embedding; "V": Value embedding. "TL": Transformer Layer. Best viewed in colour.

attention (multi-stream to one-stream), (4) hierarchical attention (one-stream to multi-stream), (5) cross-attention, and (6) cross-attention to concatenation. See Table II and Fig. 2.

For brevity, we will state and compare the mathematical formulations in two-modality cases. Please note that all discussed self-attention and its variants are such flexible that can be extended to multiple modality cases. Specifically, the following formulations are modality-, tokenization-, and embedding- agnostic, as self-attention models the embedding of arbitrary token from arbitrary modality as a node of a graph.

Given inputs $\mathbf{X}_{\mathrm{A}}$ and $\mathbf{X}_{\mathrm{B}}$ from two arbitrary modalities, $\mathbf{Z}_{(\mathrm{A})}$ and $\mathbf{Z}_{(\mathrm{B})}$ denote their respective token embeddings. Let $\mathbf{Z}$ denoting the token embedding (sequence) produced by the multimodal interactions. $Tf(\cdot)$ stands for the processing of Transformer layers/blocks.

(1) Early Summation: In practice, early summation [46], [81] is a simple and effective multimodal interaction, where the token embeddings from multiple modalities can be weighted summed at each token position and then processed by Transformer layers:

$$\mathbf{Z} \leftarrow Tf(\alpha\mathbf{Z}_{(\mathrm{A})} \oplus \beta\mathbf{Z}_{(\mathrm{B})}) = MHSA(\mathbf{Q}_{(\mathrm{AB})}, \mathbf{K}_{(\mathrm{AB})}, \mathbf{V}_{(\mathrm{AB})}), \tag{8}$$

where $\oplus$ is element-wise sum, and α and β are weightings. Concretely, $\mathbf{Q}_{(\mathrm{AB})} = (\alpha\mathbf{Z}_{(\mathrm{A})} \oplus \beta\mathbf{Z}_{(\mathrm{B})})\mathbf{W}^Q_{(\mathrm{AB})}$, $\mathbf{K}_{(\mathrm{AB})} = (\alpha\mathbf{Z}_{(\mathrm{A})} \oplus \beta\mathbf{Z}_{(\mathrm{B})})\mathbf{W}^K_{(\mathrm{AB})}$, and $\mathbf{V}_{(\mathrm{AB})} = (\alpha\mathbf{Z}_{(\mathrm{A})} \oplus \beta\mathbf{Z}_{(\mathrm{B})})\mathbf{W}^V_{(\mathrm{AB})}$. Its main advantage is that it does not increase computational complexity. However, its main disadvantage is due to the manually set weightings. As discussed in Sections III-A1 and III-C1, summing position embedding is intrinsically a case of early summation.

(2) Early Concatenation: Another straightforward solution is early concatenation [7], [44], [178], [180] that the token embedding sequences from multiple modalities are concatenated and input into Transformer layers as

$$\mathbf{Z} \leftarrow Tf(\mathcal{C}(\mathbf{Z}_{(\mathrm{A})}, \mathbf{Z}_{(\mathrm{B})})). \tag{9}$$

Thus, all the multimodal token positions can be attended as a whole sequence, such that the positions of each modality can be encoded well by conditioning the context of other modalities. VideoBERT [7] is the one of the first multimodal Transformer works, where video and text are fused via early concatenation that can encode the global multimodal context well [188]. However, the longer sequence after concatenation will increase computational complexity. Early concatenation is also termed "all-attention" or "Co-Transformer" [137].

(3) Hierarchical Attention (multi-stream to one-stream): Transformer layers can be combined hierarchically to attend to the cross-modal interactions. A common practice is that multimodal inputs are encoded by independent Transformer streams and their outputs are concatenated and fused by another Transformer [146]:

$$\mathbf{Z} \leftarrow Tf_3(\mathcal{C}(Tf_1(\mathbf{Z}_{(\mathrm{A})}), Tf_2(\mathbf{Z}_{(\mathrm{B})}))). \tag{10}$$

This kind of hierarchical attention is an implementation of late interaction/fusion, and can be treated as a special case of early concatenation.

(4) Hierarchical Attention (one-stream to multi-stream): InterBERT [188] is another good practice of hierarchical attention where concatenated multimodal inputs are encoded by a shared single-stream Transformer that is followed by two separate Transformer streams. This flow can be formulated as

$$\begin{cases} \mathcal{C}(\mathbf{Z}_{(\mathrm{A})}, \mathbf{Z}_{(\mathrm{B})}) \leftarrow Tf_1(\mathcal{C}(\mathbf{Z}_{(\mathrm{A})}, \mathbf{Z}_{(\mathrm{B})})), \\ \mathbf{Z}_{(\mathrm{A})} \leftarrow Tf_2(\mathbf{Z}_{(\mathrm{A})}), \\ \mathbf{Z}_{(\mathrm{B})} \leftarrow Tf_3(\mathbf{Z}_{(\mathrm{B})}). \end{cases} \tag{11}$$

This method perceives the cross-modal interactions and meanwhile preserves the independence of uni-modal representation.

(5) Cross-Attention: For two-stream Transformers, if the **Q** (Query) embeddings are exchanged/swapped in a cross-stream manner, the cross-modal interactions can also be perceived. This method is termed cross-attention or co-attention [190], which was first proposed in VilBERT [102]:

$$\begin{cases} \mathbf{Z}_{(\mathrm{A})} \leftarrow MHSA(\mathbf{Q}_{\mathrm{B}}, \mathbf{K}_{\mathrm{A}}, \mathbf{V}_{\mathrm{A}}), \\ \mathbf{Z}_{(\mathrm{B})} \leftarrow MHSA(\mathbf{Q}_{\mathrm{A}}, \mathbf{K}_{\mathrm{B}}, \mathbf{V}_{\mathrm{B}}). \end{cases} \tag{12}$$

Cross-attention attends to each modality conditioned on the other and does not cause higher computational complexity, however if considered for each modality, this method fails to perform cross-modal attention globally and thus loses the whole context. As discussed in [188], two-stream cross-attention can learn cross-modal interaction, whereas there is no self-attention to the self-context inside each modality.

(6) Cross-Attention to Concatenation: The two streams of cross-attention [102] can be further concatenated and processed by another Transformer to model the global context. This kind of hierarchically cross-modal interaction is also widely studied [137], [189], and alleviates the drawback of cross-attention.

$$\begin{cases} \mathbf{Z}_{(\mathrm{A})} \leftarrow MHSA(\mathbf{Q}_{\mathrm{B}}, \mathbf{K}_{\mathrm{A}}, \mathbf{V}_{\mathrm{A}}), \\ \mathbf{Z}_{(\mathrm{B})} \leftarrow MHSA(\mathbf{Q}_{\mathrm{A}}, \mathbf{K}_{\mathrm{B}}, \mathbf{V}_{\mathrm{B}}), \\ \mathbf{Z} \leftarrow Tf(\mathcal{C}(\mathbf{Z}_{(\mathrm{A})}, \mathbf{Z}_{(\mathrm{B})})). \end{cases} \tag{13}$$

Discussion: All these aforementioned self-attention variants for multimodal interactions are modality-generic, and can be applied in flexible strategies and for multi-granular tasks. Specifically, these interactions can be flexibly combined and nested. For instance, multiple cross-attention streams are used in hierarchical attention (one-stream to multi-stream) that in a two-stream decoupled model [191] Tf_2 and Tf_3 of (11) are implemented by cross-attention defined in (12). Moreover, they can be extended to multiple (≥ 3) modalities. TriBERT [183] is a tri-modal cross-attention (co-attention) for vision, pose, and audio, where given a Query embedding, its Key and Value embeddings are the concatenation from the other modalities. Cross-attention to concatenation is applied to three modalities (i.e., language, video, and audio) in [189].

3) Network Architectures: Essentially, various multimodal Transformers work due to their internal multimodal attentions that are the aforementioned self-attention variants. Meanwhile, as illustrated in Fig. 2, these attentions determine the external network structures of the multimodal Transformers where they are embedded.

In general, if we consider from the angle of network structures, (1) early summation and early concatenation work in single-stream, (2) cross-attention work in multi-streams, (3) hierarchical attention and cross-attention to concatenation work in hybrid-streams. Thus, multimodal Transformers can be divided into single-stream (e.g., Uniter [106], Visualbert [104], Vl-bert [105], Unified VLP [110]), multi-stream (e.g., ViLBERT [102], Lxmert [103], ActBERT [114]), hybrid-stream (e.g., InterBERT [188]), *etc*.

From the perspective of timing of interaction, these multimodal attentions fall into three categories, i.e., early interaction: early summation, early concatenation, and hierarchical attention (one-stream to multi-stream), late interaction: hierarchical attention (multi-stream to one-stream), or throughout interaction: cross-attention, cross-attention to concatenation.

As demonstrated in Fig. 2 in [192], the multimodal Transformer models have another architecture taxonomy based on the computational size of the components.

IV. Application Scenarios

In this section we survey multimodal Transformers based on the application scenarios. We consider two important paradigms: (1) Transformers for multimodal pretraining (Section IV-A, including both task-agnostic (Section IV-A1) and task-specific (Section IV-A2) multimodal pretraining), and (2) Transformers for specific multimodal tasks (Section IV-B).

A. Transformers for Multimodal Pretraining

Inspired by the great success of Transformer based pretraining in NLP community, Transformers are also widely studied for multimodal pretraining as the various large-scale multimodal corpora is emerging. Recent work has demonstrated that if pretrained on large scale multimodal corpora Transformer based models [7], [102], [103], [104], [105], [106], [110] clearly outperform other competitors in a wide range of multimodal down-stream tasks, and moreover achieve the zero-shot generalization ability. These superiorities have led Transformer-based multimodal pretraining to become a hot topic, which has two main directions, i.e., general pretraining for agnostic down-stream tasks (Section IV-A1), goal-oriented pretraining for specific down-stream tasks (Section IV-A2).

We focus on these key points: (1) What trends are emerging? (2) Where/how do the cross-modal interactions take place during pretraining? (3) How to sort out and understand the pretraining pretext objectives? How can they drive Transformers to learn the cross-modal interactions?

1) Task-Agnostic Multimodal Pretraining: Recently Transformer-oriented pretraining has been widely studied involving diverse modality combinations, e.g., video-text [7], [107], [117], image-text [102], [103], [104], [193], [194], [195], acoustic-text [180].

Among existing work, the following main trends are emerging:

(1) Vision-language pretraining (VLP) is a major research problem in this field. VLP is including both "image + language" and "video + language", also termed visual-linguistic pretraining. A great deal of excellent work has been proposed, e.g., VideoBERT [7], ViLBERT [102], LXMERT [103], VisualBERT [104], VL-BERT [105], UNITER [106], CBT [107],

Unicoder-VL [108], B2T2 [109], VLP [110], 12-in-1 [111], Oscar [112], Pixel-BERT [113], ActBERT [114], ImageBERT [115], HERO [116], UniVL [117], SemVLP [196].

(2) Speech can be used as text. Thanks to recent advances in automatic speech recognition (ASR) techniques, in a multimodal context, speech can be converted to text by the off-the-shelf speech recognition tools. For instance, VideoBERT [7] and CBT [107] make full use of speech rather than low-level sounds as a source of cross-modal supervision, by extracting high-level semantic text.

(3) Overly dependent on the well-aligned multimodal data. A majority of Transformer-based multimodal pretraining works in a self-supervised manner, however, it is overly dependent on the well-aligned multimodal sample pairs/tuples. For instance, large amount of image-language pretraining Transformer models are pretrained on large-scale image-text pairs, e.g., VisualBERT [104], VL-BERT [105], ViLBERT [102], LXMERT [103], UNITER [106]. For another example, the instructional videos (e.g., cooking) [3] are widely used as the pretraining corpora, e.g., HowToVQA69M [140], HowTo100M [141], as in general, their visual clues/content and the spoken words have a higher probability to align with each other, if compared with other videos. However, using cross-modal alignment as cross-modal supervision is costly for large-scale applications. Thus, how to use the weakly-aligned or even unpaired/unaligned multimodal data as the pretraining corpora is still understudied. Some recent attempts [137], [199] study the use of weakly-aligned cross-modal supervision to train Transformers to learn the cross-modal interactions.

(4) Most of the existing pretext tasks transfer well across modalities. For instance, Masked Language Modelling (MLM) in the text domain has been applied to audio and image, e.g., Masked Acoustic Modelling [180], [200], Masked Image Region Prediction [190], while both Sentence Ordering Modelling (SOM) [201] in text domain and Frame Ordering Modelling (FOM) [116] in video domain share the same idea. We will further discuss the pretext tasks for multimodal Transformer pretraining in the follows.

(5) Model structures are mainly in three categories. Essentially, in multimodal pretraining scenarios, Transformer models work based on those self-attention variants that are discussed in Section III-C2. Thus, if considered from the perspective of model structures, the existing Transformers for multimodal pretraining are also mainly in three categories, i.e., single-stream, multi-stream, hybrid-stream.

(6) Cross-modal interactions can perform within various components/levels in the pretraining pipelines. For Transformer based multimodal pretraining, the key is to drive the Transformer (encoder w/, w/o decoder) to learn the cross-modal interactions. In the existing Transformer-based multimodal pretraining practices, the cross-modal interactions are flexible, which can perform within various components/levels in the pretraining pipelines. In general, Transformer-based multimodal pretraining pipelines have three key components, from bottom to top, i.e., tokenization, Transformer representation, objective supervision.

[3]Note that instructional videos also have weakly aligned cases [197], [198].

For not only the multimodal pretraining but also the specific multimodal tasks, the cross-modal interactions can perform within arbitrary component(s) of the three. As discussed in Section III-C2, because self-attention models the embedding of an arbitrary token from an arbitrary modality as a node of a graph, the existing pretraining pipelines can, in general, be transferred independently across modalities, unless considered with modality-specific objectives.

Discussion: Vision Language Pretraining (VLP) follows two general pipelines: two-stage (need object detector, e.g., Faster R-CNN [202]) (e.g., LXMERT [103], ViLBert [102], VL-Bert [105], UNITER [106]) and end-to-end (e.g., Pixel-Bert [113], SOHO [203], KD-VLP [204], Simvlm [199]). Two-stage pipelines have a main advantage – object-aware perceiving, by using the supervised pre-trained visual detectors, however these are based on a strong assumption that the visual representations can be fixed.

Discussion: How to look for more corpora that intrinsically have well-aligned cross-modal supervision, such as instructional videos, is still an open problem. However, weakly-aligned cross-modal samples are popular in the real-life scenarios, for instance, enormous weakly aligned multimodal data samples are emerging in e-commerce [137], due to fine-grained categories, complex combinations, and fuzzy correspondence. Well labelled/aligned cross-modal datasets are very costly in collecting and annotating; how to use weakly-aligned or even unaligned corpora crawled from the web is a promising question. Some recently successful practice [9], [199], [205] used weakly aligned image-text pairs to perform pretraining, and achieve both competitive performance and zero-shot learning capability for image classification, image-text retrieval, and open-ended visual question answering, *etc*. Because these practices in weak supervision make full use of large-scale pretraining corpora, they yield greater promise of zero-shot generalization.

Pretext Tasks: In Transformer based multimodal pretraining, the pretraining tasks/objectives are also termed pretext tasks/objectives. To date, various pretext tasks have been studied, e.g., masked language modelling (MLM) [137], masked image region prediction/classification (also termed masked object classification (MOC)) [137], [190], masked region regression (MRR) [115], visual-linguistic matching (VLM) (e.g., image–text matching (ITM) [188], image text matching (ITM), phrase-region alignment (PRA) [204], word-region alignment (WRA) [106], video-subtitle matching (VSM) [116]), masked frame modelling (MFM) [116], frame order modelling (FOM) [116], next sentence prediction (NSP) [4], [102], [190], masked sentence generation (MSG) [191], masked group modelling (MGM) [188], prefix language modelling (PrefixLM) [199], video conditioned masked language model [117], text conditioned masked frame model [117], visual translation language modelling (VTLM) [206], and image-conditioned masked language modelling (also termed image-attended masked language modelling) [207]. These down-stream task -agnostic pretext pretraining is optional, and the down-stream task objectives can be trained directly, which will be discussed in Section IV-A2. Table III provides the common and representative pretext tasks for Transformer based multimodal pretraining.

TABLE III
PRETEXT TASK COMPARISON OF MULTI-MODAL PRETRAINING TRANSFORMER MODELS (FOR AGNOSTIC DOWN-STREAM TASKS)

Types (Motivations)	Tasks	C-M Loss	Con. Loss	References
Masking	Masked Language Modelling (MLM)		✓	[7], [137]
	Image-Conditioned Masked Language Modelling (IMLM)		✓	[206], [207] [208]
	Text-Conditioned Masked Region Prediction		✓	[206]
	Masked Acoustic Modelling		✓	[180], [200]
	Masked Image Region Regression		✓	[115]
	Masked Image Region Prediction		✓	[190]
	Masked Frame Modelling (MFM)		✓	[116]
	Masked Sentence Generation (MSG)		✓	[191]
	Video Conditioned Masked Language Model		✓	[117]
	Text Conditioned Masked Frame Model		✓	[117]
Describing	Image-conditioned Denoising Autoencoding (IDA)		✓	[208]
	Text-conditioned Image Feature Generation (TIFG)		✓	[208]
	Prefix Language Modelling (PrefixLM)	✓		[199]
Matching	Image-Text Matching (ITM)	✓		[188] [102], [103], [104], [106], [209],
	Phrase-Region Alignment (PRA)	✓		[204]
	Word-Region Alignment (WRA)	✓		[106], [192]
	Video-Subtitle Matching (VSM)	✓		[116]
	Next Sentence Prediction (NSP)		✓	[4], [102], [190]
Ordering	Sentence Ordering Modelling (SOM)		✓	[201]
	Frame Ordering Modelling (FOM)		✓	[116]

"C-M LOSS": cross-modal loss; "CON. LOSS": loss conditioned on other modality/modalities.

In practice, pretext tasks can be combined, and some representative cases are summarized in Table III of [57], Table II of [58].

The pretext tasks have multiple taxonomies:

(1) Supervision: The common multimodal pretraining Transformers use well-aligned, weakly-aligned, and even unaligned multimodal sample pairs/tuples, to work in supervised, weakly-supervised, and unsupervised manners, respectively. Meanwhile, if we consider the definitions of their pretext tasks/objectives from supervision, the pretexts can be sorted into unsupervised/self-supervised (e.g., masked language modelling (MLM) [7], [137]) and supervised (e.g., image-text matching (ITM) [102], [103], [104], [106], [188], [209]), *etc*. Nowadays, self-supervised attempts are the majority.

(2) Modality: Considering the mathematical formulations, some pretexts are defined on single modality, e.g., masked language modelling [7], masked acoustic modelling [200], masked region regression (MRR) [115], while other pretexts are defined on multiple modalities, e.g., image-conditioned masked language modelling (IMLM) [208], image-text matching (ITM) [188], video-subtitle matching (VSM) [116]. Thus, from this mathematical view, the pretext tasks can be divided into two categories, i.e., uni-modal and multimodal.

However, this classification is not really accurate. It should be highlighted that in multimodal pretraining Transformer models, even if the pretext objective formulations only include uni-modal elements, pretexts can still involve other modalities, essentially conditioned on the clues from other modalities, by (a) prepositive token level interactions and/or Transformer level interactions, (b) co-training with other pretexts that involve other modalities. For instance, VL-BERT [105] uses two dual pretext tasks, i.e., masked language modelling and masked RoI classification.

(3) Motivation: If consider their motivations, the pretext tasks include masking, describing, matching, ordering, *etc*.

Some recent surveys focus on VLP and compare the existing VLP Transformer models from the angles of domain (image-text or video-text), vision feature extraction, language feature extraction, architecture (single- or dual- stream), decoder (w/, w/o), pretext tasks/objectives, pretraining datasets, and down-stream tasks, e.g., Table III of [57], Table II of [58]. Different from these views, in this survey, we would propose our comparisons from some new perspectives. Specifically: *(1)* The core of Transformer ecosystem is self-attention, thus we would compare the existing multimodal pretraining Transformer models from the angles of how and when the self-attention or its variants perform cross-modal interactions. *(2)* Considering from a geometrically topological perspective, self-attention helps Transformers intrinsically work in a modality agnostic pipeline that is compatible with various modalities by taking in the embedding of each token as a node of graph, thus we would highlight that the existing VLP can be applied to other modalities, beyond visual and linguistic domains. *(3)* We suggest to treat the Transformer-based multimodal pretraining pipelines having three key components, from bottom to top, i.e., tokenization, Transformer representation, objective supervision.

Discussion: In spite of the recent advances, multimodal pretraining Transformer methods still have some obvious bottlenecks. For instance, as discussed by [208] in VLP field, while the BERT-style cross-modal pretraining models produce excellent results on various down-stream vision-language tasks, they fail to be applied to generative tasks directly. As discussed in [208], both VideoBERT [7] and CBT [107] have to train a separate video-to-text decoder for video captioning. This is a significant gap between the pretraining models designed for discriminative and generative tasks, as the main reason is discriminative task oriented pretraining models do not involve the decoders of Transformer. Therefore, how to design more unified pipelines that can work for both discriminative and generative down-stream tasks is also an open problem to be solved. Again for instance, common multimodal pretraining models often underperform for fine-grained/instance-level tasks as discussed by [137].

Discussion: As discussed in [208], the masked language and region modelling as pre-training task have a main advantage that the Transformer encoder learned from these supervisions can encode both vision and language patterns based on bidirectional context and it is naturally fit for the semantic understanding tasks, e.g., VQA, image-text retrieval.

Discussion: How to boost the performance for multimodal pretraining Transformers is an open problem. Some practices demonstrate that multi-task training (by adding auxiliary loss) [111], [137] and adversarial training [210] improve multimodal pretraining Transformers to further boost the performance. Meanwhile, overly compound pretraining objectives potentially upgrade the challenge of balancing among different loss terms, thus complicate the training optimization [199]. Moreover, the difficulty of the pretexts is also worth discussing. In general, if aim to learn more explicit object concepts, more complex pretext losses will be used [204]. However, for pretexts, whether more complexity is better remains a question.

2) Task-Specific Multimodal Pretraining: In practices of multimodal Transformers, the aforementioned down-stream task -agnostic pretraining is optional, not necessary, and down-stream task specific pretraining is also widely studied [150], [190], [208], [211]. The main reasons include: (1) Limited by the existing technique, it is extremely difficult to design a set of highly universal network architectures, pretext tasks, and corpora that work for all the various down-stream applications. (2) There are non-negligible gaps among various down-stream applications, e.g., task logic, data form, making it difficult to transfer from pretraining to down-stream applications.

Therefore, a large number of down-stream tasks still need tailor-made pretraining to improve the performance. Guhur et al. [150] propose in-domain pretraining for vision-and-language navigation, as the general VLP focuses on learning vision-language correlations, not designed for sequential decision making as required in embodied VLN. Murahari et al. [190] present a visual dialogue oriented approach to leverage pretraining on general vision-language datasets. XGPT [208] is tailor-made for image captioning, to overcome the limitation that BERT-based cross-modal pre-trained models fail to be applied to generative tasks directly. ERNIE-ViLG [211] is designed for bidirectional image-text generation with Transformers.

Special modalities have their own unique domain knowledge that can be used to design the specific pretrain pretexts. GraphCodeBERT [44] uses two structure-aware pretext tasks (i.e., predict where a variable is identified from, data flow edge prediction between variables) for programming source code. To learn from the spatial cues in 360° video, Morgado et al. [145] propose to perform contrastive audio-visual spatial alignment of 360° video and spatial audio. Med-BERT [184] is a contextualized embedding model pretrained on a structured electronic health record dataset of two million patients. Kaleido-BERT [212] is a VLP Transformer model tailor-made for the fashion domain.

B. Transformers for Specific Multimodal Tasks

Recent work has demonstrated that Transformer models can encode various multimodal inputs in both classical and novel discriminative applications, e.g., RGB & optical flow [46], RGB & depth [213], RGB & point cloud [214], RGB & LiDAR [215], [216], textual description & point cloud [31], acoustic & text [180], audio & visual observation for Audio-Visual Navigation [76], speech query & schema of SQL database [25], text question/query & the schema SQL database [24], audio & tags [217], multimodal representation for video [218], [219], text query & video [220], audio & video for audio visual speech enhancement (AVSE) [179], audio & video for Audio-Visual Video Parsing [173], audio & video for audio-visual speech recognition [134], video & text for Referring Video Object Segmentation (RVOS) [221], source code & comment & data flow [44], image & text for retrieval [222].

Meanwhile, Transformers also contribute to various multimodal generative tasks, including single-modality to single-modality (e.g., raw audio to 3D mesh sequence [39], RGB to 3D scene [40], single image to 3D human texture estimation [223], RGB to scene graph [19], [224], [225], [226], graph to graph [33], knowledge graph to text [227], video to scene graph [228], video to caption [229], [230], [231], [232], image to caption [233], [234], [235], [236], [237], text to speech [238], text to image [205], [239], text to shape [240], RGB to 3D human pose and mesh [41], music to dance [241]), multimodality to single modality (e.g., image & text to scene graph [242], Video Dialogue (text & audio & visual to text) [243], Mono Audio & Depth to Binaural Audio [14], music piece & seed 3D motion to long-range future 3D motions [146], X-raying image & question to answer [244], video & text & audio to text [245]), and multimodality to multimodality (e.g., [246]).

V. Challenges and Designs

Complementing the application scenario taxonomy discussed in Section IV, we further survey prior work from the perspective of technical challenges. We discuss seven challenges of Transformer based multimodal learning, including fusion, alignment, transferability, efficiency, robustness, universalness, and interpretability. This further extends the taxonomy introduced in [1] to tackle the higher diversity and wider scopes of existing Transformer based MML works in recent years.

A. Fusion

In general, MML Transformers fuse information across multiple modalities primarily at three levels: input (i.e., early fusion), intermediate representation (i.e., middle fusion), and prediction (i.e., late fusion). Common early fusion based MML Transformer models [7], [104], [108] are also known as one-stream architecture, allowing the adoption of the merits of BERT due to minimal architectural modification. The main difference between these one-stream models is the usage of problem-specific modalities with variant masking techniques. With attention operation, a noticeable fusion scheme is introduced based on a notion of bottleneck tokens [175]. It applies for both early and middle fusion by simply choosing to-be-fused layers. We note that the simple prediction-based late fusion [247], [248] is less adopted in MML Transformers. This makes sense considering the motivations of learning stronger multimodal contextual

representations and great advance of computing power. For enhancing and interpreting the fusion of MML, probing the interaction and measuring the fusion between modalities [249] would be an interesting direction to explore.

B. Alignment

Cross-modal alignment is the key to a number of real-world multimodal applications. Transformer based cross-modal alignment has been studied for various tasks, e.g., speaker localization in multi-speaker videos [250], speech translation [180], text-to-speech alignment [251], text-to-video retrieval [252], [253], [254], and visual grounding of natural language [255], [256], [257], [258], [259]. Recently, Transformer based alignment [9], [119], [260], [261], [262] has led to a surge of leveraging large quantities of web data (e.g., image-text pairs) for vision and language tasks.

A representative practice is to map two modalities into a common representation space with contrastive learning over paired samples. The models based on this idea are often enormous in size and expensive to optimize from millions or billions of training data. Consequently, successive works mostly exploit pretrained models for tackling various down-stream tasks [120], [263], [264], [265], [266]. These alignment models have the ability of zero-shot transfer particularly for image classification via prompt engineering [267]. This novel perspective is mind-blowing, given that image classification is conventionally regarded as a unimodal learning problem and zero-shot classification remains an unsolved challenge despite extensive research [268]. This has been studied for more challenging and fine-grained tasks (e.g., object detection [269], visual question answering [103], [106], [112], [263], and instance retrieval [222], [263]) by imposing region (semantic parts such as objects) level alignment. Fine-grained alignment will however incur more computational costs from explicit region detection and how to eliminate this whilst keeping the region-level learning capability becomes a challenge. Several ideas introduced recently include random sampling [113], learning concept dictionary [203], uniform masking [270], patch projection [192], joint learning of a region detector [271], and representation aligning before mask prediction [263].

C. Transferability

Transferability is a major challenge for Transformer based multimodal learning, involving the question of how to transfer models across different datasets and applications.

Data augmentation and adversarial perturbation strategies help multimodal Transformers to improve the generalization ability. VILLA [210] is a two-stage strategy (task-agnostic adversarial pretraining, followed by task-specific adversarial finetuning) that improves VLP Transformers.

In practice, the distribution gap between training data and practical data is noticeable. For instance, supervised data samples (well-labelled, well-aligned) are costly in practical applications, thus how to transfer the supervised multimodal Transformers pretrained on well-aligned cross-modal pairs/tuples to the weakly aligned test bed is challenging [137]. CLIP [9] is an inspiring solution that transfers knowledge across modalities by learning a shared multimodal embedding space, enabling zero-shot transfer of the model to down-stream tasks. The main inspiration that CLIP presents the community is that the pretrained multimodal (image and text) knowledge can be transferred to down-stream zero-shot image prediction by using a prompt template "`A photo of a {label}.`" to bridge the distribution gap between training and test datasets.

Over-fitting is a major obstacle to transfer. Multimodal Transformers can be overly fitted to the dataset biases during training, due to the large modelling capability. Some recent practices exploit how to transfer the oracle model trained on noiseless dataset to real dataset. For instance, Kervadec et al. [272], [273] explore how transferable reasoning patterns are in VQA, and demonstrate that for LXMERT [103]/BERT-like reasoning patterns can be partially transferred from an ideal dataset to a real dataset.

Cross-task gap is another major obstacle to transfer [208], [274], due to the different reasoning and input-output workflows, e.g., how to use multimodal datasets to finetune the language pretrained model is difficult [274]. In real applications, multimodal pretrained Transformers sometimes need to handle the uni-modal data at inference stage due to the issue of missing modalities. One solution is using knowledge distillation, e.g., distilling from multimodal to uni-modal attention in Transformers [275], distilling from multiple uni-modal Transformer teachers to a shared Transformer encoder [276]. There is a huge gap across discriminative and generative multimodal tasks. As discussed in [208], the BERT-like encoder-only multimodal Transformers (e.g., VideoBERT [7], CBT [107]) need separately to train decoders for generation tasks. This could create a pretrain-finetune discrepancy detrimental to the generality. Recently, more and more attempts study this issue further, e.g., GilBERT [222] is a generative VLP models for a discriminative task, i.e., image-text retrieval.

Cross-lingual gap also should be considered for the transferability of Transformer based multimodal learning, e.g., universal cross-lingual generalization from English to non-English multimodal contexts [206], [277].

D. Efficiency

Multimodal Transformers suffer from two major efficiency issues: (1) Due to the large model parameter capacity, they are data hungry and thus dependent on huge scale training datasets. (2) They are limited by the time and memory complexities that grow quadratically with the input sequence length, which are caused by the self-attention. In multimodal contexts, calculation explosion will become worse due to jointly high dimension representations. These two bottlenecks are interdependent and should be considered together.

To improve the training and/or inferring efficiency for multimodal Transformers, recent efforts have attempted to find various solutions, to use fewer training data and/or parameters. The main ideas can be summarized as the follows.

1) Knowledge distillation. Distill the knowledge from the trained larger Transformers to smaller Transformers [93].

Miech et al. [278] conduct distillation from a slower model (early concatenation based Transformers, $\mathcal{O}((N_{(\mathrm{A})}+N_{(\mathrm{B})})^2)$) to a faster one (independently dual branch Transformers, $\mathcal{O}(N_{(\mathrm{A})}^2)$).

2) Simplifying and compressing model. Remove the components to simplify the pipelines. Taking the VLP Transformer models as an example, two-stage pipeline is costly as they need object detector. One simplifying is processing the visual input in convolution-free manner, e.g., E2E-VLP [271], ViLT [192]. DropToken [174] reduces the training complexity via random dropping a portion of the video and audio tokens from input sequence during training. DropToken can be treated as an implementation of dropout or adversarial training. Weight-sharing is also a common practice for simplifying multimodal Transformer models. Wen et al. [279] present a weight-sharing Transformer on top of the visual and textual encoders to align text and image. Lee et al. [280] propose a novel parameter sharing scheme based on low-rank approximation.
3) Asymmetrical network structures. Assign different model capacities and computational size properly for different modalities, to save parameters. See Fig. 2 in [192].
4) Improving utilization of training samples. Liu et al. [281] train a simplified LXMERT by making full use of fewer samples at different granularities. Li et al. [282] use fewer data to train CLIP by fully mining the potential self-supervised signals of (a) self-supervision within each modality, (b) multi-view supervision across modalities, and (c) nearest-neighbour supervision from other similar pairs.
5) Compressing and pruning model. Search the optimal sub-structures/sub-networks of multimodal Transformers, e.g., playing Lottery Tickets with the VLP Transformer models [283], adaptively freezing some layers during training [284].
6) Optimizing the complexity of self-attention. Transformers cost time and memory that grows quadratically with the input sequence length [285]. One potential solution is optimizing the $\mathcal{O}(N^2)$ complexity, e.g., Child et al. [286] present sparse factorizations of the attention matrix to reduce the quadratical complexity to $\mathcal{O}(n\sqrt{n})$, Transformer-LS [287] is an efficient Transformer for both language and vision long sequence, with linear computational and memory complexity.
7) Optimizing the complexity of self-attention based multimodal interaction/fusion. Nagrani et al. [175] propose Fusion via Attention Bottlenecks (FSN, fusion bottleneck) to improve the early concatenation based multimodal interaction. FSN passes on the messages through a small number of bottleneck latents, thus requiring the model to purify the most necessary information from each modality for cross-modal sharing. This strategy uses the fusion bottleneck as a bridge, and not only improves fusion performance, but also reduces computational cost.
8) Optimizing other strategies. Use optimal strategies to perform the common Transformer based multimodal interactions. Given the quadratic complexity of self-attention, using early concatenation based multimodal interaction to synchronously fuse the inputs from multiple modalities/views is costly. Yan et al. [288] present an efficient solution that sequentially fuses information between all pairs of two adjacent views in ascending order of sequence length. This is intrinsically a greedy strategy.

E. Robustness

Multimodal Transformers pretrained on large-scale corpora achieve the state-of-the-art for various multimodal applications, while their robustness is still unclear and understudied. This at least involves two key challenges, i.e., how to theoretically analyse the robustness, how to improve the robustness.

Although that recent attempts [99], [182], [289], [290] study and evaluate how the Transformer components/sub-layers contribute to the robustness, the main bottleneck is that the community lacks theoretical tools to analyse the Transformer family. Recently, the common practices to analyse robustness are mainly based on experiment evaluations [291], e.g., cross-dataset evaluations, perturbation-based evaluations. Thus, some multimodal datasets [130], [292] are proposed for evaluating the robustness.

Recent attempts mainly use two straightforward methods to improve the robustness for multimodal Transformer models: (1) augmentation and adversarial learning based strategies [293], [294], (2) fine-grained loss functions [295]. For instance: VILLA [210] is a generic adversarial training framework that can be applied to various multimodal Transformers. Akula et al. [292] empirically demonstrate that ViLBERT fails to exploit linguistic structure, and they propose two methods to improve the robustness of ViLBERT, one based on contrastive learning and the other based on multi-task learning.

F. Universalness

Due to the highly diversity of tasks and modalities of multimodal learning, universalness is an important problem for multimodal Transformer models. A large amount of recent attempts [117], [296], [297], [298] study how to use as unified as possible pipelines to handle various modalities and multimodal tasks. Ideally, the unified multimodal Transformers can be compatible with various data (e.g., aligned and unaligned, uni-modal and multimodal) and tasks (e.g., supervised and unsupervised, uni-modal and multimodal, discriminative and generative), and meanwhile have either few-shot or even zero-shot generalization ability. Thus, the current solutions for universalness goal for multimodal Transformers are preliminary probes.

The currently unifying-oriented attempts mainly include:

1) Unifying the pipelines for both uni-modal and multimodal inputs/tasks. As discussed Section V-C, in practical scenarios, multimodal Transformers need to handle uni-modal data due to the issue of missing modalities. Distilling multimodal knowledge into small models that are adaptable to uni-modal data and tasks is a successful practice [275], [276].
2) Unifying the pipelines for both multimodal understanding and generation. In general, for multimodal Transformer pipelines, understanding and discriminative tasks require Transformer encoders only, while generation/generative

tasks require both Transformer encoders and decoders. Existing attempts use multi-task learning to combine the understanding and generation workflows, where two kinds of workflows are jointly trained by multi-task loss functions. From the perspective of model structures, typical solutions include: (a) encoder + decoder, e.g., E2E-VLP [271]. (b) separate encoders + cross encoder + decoder, e.g., UniVL [117], CBT [107]. (c) single unified/combined encoder-decoder, e.g., VLP [110]. (d) two-stream decoupled design [191].

3) Unifying and converting the tasks themselves, e.g., CLIP [9] converts zero-shot recognition to retrieval, thus reduces the costs of modifying the model.

However, the aforementioned practices suffer some obvious challenges and bottlenecks, at least including:

1) Due to modality and task gaps, universal models should consider the trade-off between universalness and cost. Unifying the pipelines of different modalities and tasks generally cause larger or more complicated model configuration, whereas for a specific modality or task, some components are redundant.

2) Multi-task loss functions increase the complexity of training. How to co-train multiple objectives properly and effectively is challenging, due to that different objectives generally should be optimized in different strategies.

G. Interpretability

Why and how Transformers perform so well in multimodal learning has been investigated [106], [299], [300], [301], [302], [303], [304], [305], [306]. These attempts mainly use probing task and ablation study. Cao et al. [299] design a set of probing tasks on UNITER [106] and LXMERT [103], to evaluate what patterns are learned in pretraining. Hendricks et al. [301] probe the image–language Transformers by fine-grained image–sentence pairs, and find that verb understanding is harder than subject or object understanding. Chen et al. [106] examine the optimal combination of pretraining tasks via ablation study, to compare how different pretexts contribute to the Transformers. Despite these attempts, the interpretability of multimodal Transformers is still under-studied to date.

VI. Discussion and Outlook

Designing the universal MML models to excel across all the unimodal and multimodal down-stream tasks with different characteristics simultaneously [115], [299] is a non-trivial challenge. For instance, two-stream architectures [9], [263] are typically preferred over one-stream ones for cross-modal retrieval-like tasks in efficiency, since the representation of each modality can be pre-computed beforehand and reused repeatedly. That being said, how to design task-agnostic MML architectures is still an open challenge, in addition to other design choices such as pretext and objective loss functions. Furthermore, a clear gap remains between the state-of-the-art and this ultimate goal. In general, existing multimodal Transformer models [9], [199], [263] are superior only for specific MML tasks, as they are designed specifically for only a subset of specific tasks [137], [142], [212], [249], [260], [261], [265], [266]. Encouragingly, several recent studies towards universal modality learning in terms of modality-agnostic network design [3] and more task-generic architecture design [307], [308], [309] have been introduced, and it is hoped this will spark further investigation. To that end, instead of exhaustively exploring the vast model design space, seeking in-depth understanding and interpretation of a MML model's behaviour might be insightful for superior algorithm design, even though the interactions and synergy across different modalities are intrinsically complex and even potentially inconsistent over tasks [249].

For more fine-grained MML, it is widely acknowledged that discovering the latent semantic alignments across modalities is critical. An intuitive strategy is to leverage semantic parts (e.g., objects) pre-extracted by an off-the-shelf detector for MML [103], [104], [105], [106], [112], [204], [310]. This, however, is not only complex and error-prone, but computationally costly [207]. Several remedies introduced recently include random sampling [113], learning concept dictionary [203], jointly learning a region detector [271], and representation aligning before mask prediction [263]. Given the scale of MML training data, exploring this direction needs exhaustive computational costs, and it is supposed that industrial research teams with rich resources are more likely to afford. Ideally, a favourable MML method would leave fine-grained semantic alignment across modalities to emerge on its own, which is worthy of careful investigation in the future.

As the learning scale expands exponentially, the training data become inevitably noisy and heterogeneous [9], [199], [263]. It has been recently shown that properly tackling the noise issue is useful [263], [309]. Another related facet is training strategy, e.g., how many stages of training is superior over the common one-stage policy [115]. Further, the quadratic complexity with Transformers becomes more acute for multimodal data due to longer input. Despite extensive research on efficient variants [49], dedicated efficiency study for MML is still underestimated even empirically and call for more investigation.

Identifying the strengths of Transformers for multimodal machine learning is a big open problem. The following main points can be summarized from the literature: *(1)* Transformers can encode implicit knowledge [32]. *(2)* The multi-head brings multiple modelling sub-spaces that can further enhance the expressive ability of the model. Ideally, multiple heads after training are good and different. This is essentially a good practice of ensemble learning. *(3)* Transformers intrinsically have a nature of global aggregation that perceives the non-local patterns. *(4)* Thanks to the large model capacity, Transformer models handle the challenging domain gaps and shifts (e.g., linguistic and visual) better via effective pretraining on large-scale corpora [294]. *(5)* Transformers can represent the inputs as graphs, which are intrinsically compatible with more modalities, e.g., table and SQL. *(6)* For modelling series and sequence patterns (e.g., time-series), Transformers have better training and inference efficiency against RNN-based models, thanks to their parallel computation in training and/or inference. Transformers are inherently permutation invariant for processing a sequence of points, e.g., well-suited for point cloud learning [164]. *(7)* Tokenization makes Transformers flexible to organize multimodal inputs, as discussed in Section III-A1.

VII. Conclusion

This survey focuses on multimodal machine learning with Transformers. We reviewed the landscape by introducing the Transformer designs and training in the multimodal contexts. We summarized the key challenges and solutions for this emerging and exciting field. Moreover, we discussed open problems and potential research directions. We hope that this survey gives a helpful and detailed overview for new researchers and practitioners, provides a convenient reference for relevant experts (e.g., multimodal machine learning researchers, Transformer network designers), and encourages future progress.

References

[1] T. Baltrušaitis, C. Ahuja, and L.-P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.

[2] A. Vaswani et al., "Attention is all you need," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2017, pp. 5998–6008.

[3] A. Jaegle, F. Gimeno, A. Brock, A. Zisserman, O. Vinyals, and J. Carreira, "Perceiver: General perception with iterative attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4651–4664.

[4] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

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

[6] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 213–229.

[7] C. Sun, A. Myers, C. Vondrick, K. Murphy, and C. Schmid, "VideoBERT: A joint model for video and language representation learning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 7463–7472.

[8] J. Chen et al., "Speech-T: Transducer for text to speech and beyond," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 6621–6633.

[9] A. Radford et al., "Learning transferable visual models from natural language supervision," 2021, *arXiv:2103.00020*.

[10] M. Li et al., "CLIP-event: Connecting text and images with event structures," 2022, *arXiv:2201.05078*.

[11] C. Zhang, Z. Yang, X. He, and L. Deng, "Multimodal intelligence: Representation learning, information fusion, and applications," *IEEE J. Sel. Topics Signal Process.*, vol. 14, no. 3, pp. 478–493, Mar. 2020.

[12] A. Rahate, R. Walambe, S. Ramanna, and K. Kotecha, "Multimodal co-learning: Challenges, applications with datasets, recent advances and future directions," *Inf. Fusion*, vol. 81, pp. 203–239, 2022.

[13] T. Hastie, R. Tibshirani, J. H. Friedman, and J. H. Friedman, *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, vol. 2. Berlin, Germany: Springer, 2009.

[14] K. K. Parida, S. Srivastava, and G. Sharma, "Beyond mono to binaural: Generating binaural audio from mono audio with depth and cross modal attention," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2151–2160.

[15] F. Qingyun, H. Dapeng, and W. Zhaokui, "Cross-modality fusion transformer for multispectral object detection," 2021, *arXiv:2111.00273*.

[16] A. Baevski, Y. Zhou, A. Mohamed, and M. Auli, "wav2vec 2.0: A framework for self-supervised learning of speech representations," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2020, Art. no. 1044.

[17] A. Nagrani, C. Sun, D. Ross, R. Sukthankar, C. Schmid, and A. Zisserman, "Speech2Action: Cross-modal supervision for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10314–10323.

[18] W. Chen, M.-W. Chang, E. Schlinger, W. Wang, and W. W. Cohen, "Open question answering over tables and text," 2020, *arXiv:2010.10439*.

[19] Y. Guo et al., "From general to specific: Informative scene graph generation via balance adjustment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 16363–16372.

[20] K. Gupta, J. Lazarow, A. Achille, L. Davis, V. Mahadevan, and A. Shrivastava, "LayoutTransformer: Layout generation and completion with self-attention," 2020, *arXiv:2006.14615*.

[21] C.-F. Yang, W.-C. Fan, F.-E. Yang, and Y.-C. F. Wang, "LayoutTransformer: Scene layout generation with conceptual and spatial diversity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3731–3740.

[22] R. Li, S. Zhang, and X. He, "SGTR: End-to-end scene graph generation with transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19464–19474.

[23] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12868–12878.

[24] R. Cai, J. Yuan, B. Xu, and Z. Hao, "SADGA: Structure-aware dual graph aggregation network for text-to-SQL," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 7664–7676.

[25] Y. Song, R. C.-W. Wong, X. Zhao, and D. Jiang, "Speech-to-SQL: Towards speech-driven SQL query generation from natural language question," 2022, *arXiv:2201.01209*.

[26] A. Salvador, E. Gundogdu, L. Bazzani, and M. Donoser, "Revamping cross-modal recipe retrieval with hierarchical transformers and self-supervised learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15470–15479.

[27] Z. Zhao et al., "ProTo: Program-guided transformer for program-guided tasks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 17021–17036.

[28] H. Zhou, W. Zhou, W. Qi, J. Pu, and H. Li, "Improving sign language translation with monolingual data by sign back-translation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1316–1325.

[29] G. Varol, L. Momeni, S. Albanie, T. Afouras, and A. Zisserman, "Read and attend: Temporal localisation in sign language videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16852–16861.

[30] H. Bull, T. Afouras, G. Varol, S. Albanie, L. Momeni, and A. Zisserman, "Aligning subtitles in sign language videos," 2021, *arXiv:2105.02877*.

[31] L. Zhao, D. Cai, L. Sheng, and D. Xu, "3DVG-transformer: Relation modeling for visual grounding on point clouds," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2908–2917.

[32] K. Marino, X. Chen, D. Parikh, A. Gupta, and M. Rohrbach, "KRISP: Integrating implicit and symbolic knowledge for open-domain knowledge-based VQA," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 14106–14116.

[33] P. Ammanabrolu and M. O. Riedl, "Learning knowledge graph-based world models of textual environments," 2021, *arXiv:2106.09608*.

[34] X. Zhu et al., "Multi-modal knowledge graph construction and application: A survey," 2022, *arXiv:2202.05786*.

[35] P. Xu et al., "SketchMate: Deep hashing for million-scale human sketch retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8090–8098.

[36] P. Xu, Z. Song, Q. Yin, Y.-Z. Song, and L. Wang, "Deep self-supervised representation learning for free-hand sketch," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 4, pp. 1503–1513, Apr. 2021.

[37] P. Xu et al., "Fine-grained instance-level sketch-based video retrieval," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 1995–2007, May 2021.

[38] Y. Vinker et al., "CLIPasso: Semantically-aware object sketching," 2022, *arXiv:2202.05822*.

[39] Y. Fan, Z. Lin, J. Saito, W. Wang, and T. Komura, "FaceFormer: Speech-driven 3D facial animation with transformers," 2021, *arXiv:2112.05329*.

[40] D. Shin, Z. Ren, E. B. Sudderth, and C. C. Fowlkes, "3D scene reconstruction with multi-layer depth and epipolar transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2172–2182.

[41] K. Lin, L. Wang, and Z. Liu, "End-to-end human pose and mesh reconstruction with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1954–1963.

[42] Y. Xu et al., "LayoutLMv2: Multi-modal pre-training for visually-rich document understanding," 2020, *arXiv:2012.14740*.

[43] I. Beltagy, M. E. Peters, and A. Cohan, "Longformer: The long-document transformer," 2020, *arXiv:2004.05150*.

[44] D. Guo et al., "GraphCodeBERT: Pre-training code representations with data flow," 2020, *arXiv:2009.08366*.

[45] D. Zügner, T. Kirschstein, M. Catasta, J. Leskovec, and S. Günnemann, "Language-agnostic representation learning of source code from structure and context," 2021, *arXiv:2103.11318*.

[46] K. Gavrilyuk, R. Sanford, M. Javan, and C. G. Snoek, "Actor-transformers for group activity recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 836–845.

[47] J. Shang, T. Ma, C. Xiao, and J. Sun, "Pre-training of graph augmented transformers for medication recommendation," 2019, *arXiv:1906.00346*.

[48] T. Lin, Y. Wang, X. Liu, and X. Qiu, "A survey of transformers," 2021, *arXiv:2106.04554*.

[49] Y. Tay, M. Dehghani, D. Bahri, and D. Metzler, "Efficient transformers: A survey," 2020, *arXiv:2009.06732*.

[50] A. M. Braşoveanu and R. Andonie, "Visualizing transformers for NLP: A brief survey," in *Proc. Int. Conf. Inf. Visualisation*, 2020, pp. 270–279.

[51] S. Khan, M. Naseer, M. Hayat, S. W. Zamir, F. S. Khan, and M. Shah, "Transformers in vision: A survey," 2021, *arXiv:2101.01169*.

[52] Y. Liu et al., "A survey of visual transformers," 2021, *arXiv:2111.06091*.

[53] K. Han et al., "A survey on vision transformer," 2020, *arXiv:2012.12556*.

[54] Y. Xu et al., "Transformers in computational visual media: A survey," *Comput. Vis. Media*, vol. 8, pp. 33–62, 2022.

[55] F. Shamshad et al., "Transformers in medical imaging: A survey," 2022, *arXiv:2201.09873*.

[56] J. Selva, A. S. Johansen, S. Escalera, K. Nasrollahi, T. B. Moeslund, and A. Clapés, "Video transformers: A survey," 2022, *arXiv:2201.05991*.

[57] L. Ruan and Q. Jin, "Survey: Transformer based video-language pre-training," 2021, *arXiv:2109.09920*.

[58] F. Chen et al., "VLP: A survey on vision-language pre-training," 2022, *arXiv:2202.09061*.

[59] F. Li et al., "Vision-language intelligence: Tasks, representation learning, and large models," 2022, *arXiv:2203.01922*.

[60] L. Wu, S. L. Oviatt, and P. R. Cohen, "Multimodal integration-a statistical view," *IEEE Trans. Multimedia*, vol. 1, no. 4, pp. 334–341, Dec. 1999.

[61] W. Guo, J. Wang, and S. Wang, "Deep multimodal representation learning: A survey," *IEEE Access*, vol. 7, pp. 63373–63394, 2019.

[62] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 65–71, Nov. 1989.

[63] A. A. Lazarus et al., *Multimodal Behavior Therapy*. Berlin, Germany: Springer, 1976.

[64] D. Feng et al., "Deep multi-modal object detection and semantic segmentation for autonomous driving: Datasets, methods, and challenges," *IEEE Trans. Intell. Transp. Syst.*, vol. 22, no. 3, pp. 1341–1360, Mar. 2021.

[65] Y. Liu, J. Zhang, L. Fang, Q. Jiang, and B. Zhou, "Multimodal motion prediction with stacked transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7573–7582.

[66] A. Moudgil, A. Majumdar, H. Agrawal, S. Lee, and D. Batra, "SOAT: A scene-and object-aware transformer for vision-and-language navigation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 7357–7367.

[67] F. Lv, X. Chen, Y. Huang, L. Duan, and G. Lin, "Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 2554–2562.

[68] R. Zellers et al., "MERLOT: Multimodal neural script knowledge models," 2021, *arXiv:2106.02636*.

[69] M. K. Hasan et al., "Humor knowledge enriched transformer for understanding multimodal humor," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 12972–12980.

[70] A. Brown, V. Kalogeiton, and A. Zisserman, "Face, body, voice: Video person-clustering with multiple modalities," 2021, *arXiv:2105.09939*.

[71] L. Yu et al., "CommerceMM: Large-scale commerce multimodal representation learning with omni retrieval," 2022, *arXiv:2202.07247*.

[72] K. Chen, J. K. Chen, J. Chuang, M. Vázquez, and S. Savarese, "Topological planning with transformers for vision-and-language navigation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11271–11281.

[73] Y. Hong, Q. Wu, Y. Qi, C. Rodriguez-Opazo, and S. Gould, "VLN BERT: A recurrent vision-and-language bert for navigation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 1643–1653.

[74] J. Zhang et al., "Curriculum learning for vision-and-language navigation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 13328–13339.

[75] Y. Qi, Z. Pan, Y. Hong, M.-H. Yang, A. van den Hengel, and Q. Wu, "The road to know-where: An object-and-room informed sequential BERT for indoor vision-language navigation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1635–1644.

[76] C. Chen, Z. Al-Halah, and K. Grauman, "Semantic audio-visual navigation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 15511–15520.

[77] S. Ren, Y. Du, J. Lv, G. Han, and S. He, "Learning from the master: Distilling cross-modal advanced knowledge for lip reading," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 13320–13328.

[78] P. Xu, T. M. Hospedales, Q. Yin, Y.-Z. Song, T. Xiang, and L. Wang, "Deep learning for free-hand sketch: A survey," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 285–312, Jan. 2023.

[79] Y. Li et al., "BEHRT: Transformer for electronic health records," *Sci. Rep.*, vol. 10, 2020, Art. no. 7155.

[80] Y. Li, H. Wang, and Y. Luo, "A comparison of pre-trained vision-and-language models for multimodal representation learning across medical images and reports," in *Proc. IEEE Int. Conf. Bioinf. Biomed.*, 2020, pp. 1999–2004.

[81] P. Xu and X. Zhu, "DeepChange: A large long-term person re-identification benchmark with clothes change," 2021, *arXiv:2105.14685*.

[82] M. Tsimpoukelli, J. Menick, S. Cabi, S. Eslami, O. Vinyals, and F. Hill, "Multimodal few-shot learning with frozen language models," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 200–212.

[83] Y.-L. Sung, J. Cho, and M. Bansal, "VL-ADAPTER: Parameter-efficient transfer learning for vision-and-language tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 5217–5227.

[84] J.-B. Alayrac et al., "Flamingo: A visual language model for few-shot learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 23716–23736.

[85] W. Wang et al., "Image as a foreign language: BEiT pretraining for all vision and vision-language tasks," 2022, *arXiv:2208.10442*.

[86] X. Chen et al., "PaLi: A jointly-scaled multilingual language-image model," 2022, *arXiv:2209.06794*.

[87] M. Lewis et al., "BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," 2019, *arXiv:1910.13461*.

[88] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," 2018. [Online]. Available: https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf

[89] Z. Dai, Z. Yang, Y. Yang, J. Carbonell, Q. V. Le, and R. Salakhutdinov, "Transformer-XL: Attentive language models beyond a fixed-length context," 2019, *arXiv:1901.02860*.

[90] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2019, Art. no. 517.

[91] M. Chen et al., "Generative pretraining from pixels," in *Proc. Int. Conf. Mach. Learn.*, 2020, pp. 1691–1703.

[92] H. Chen et al., "Pre-trained image processing transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12294–12305.

[93] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 10347–10357.

[94] J. Beal, E. Kim, E. Tzeng, D. H. Park, A. Zhai, and D. Kislyuk, "Toward transformer-based object detection," 2020, *arXiv:2012.09958*.

[95] Z. Liu et al., "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[96] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised vision transformers," 2021, *arXiv:2104.02057*.

[97] M. Caron et al., "Emerging properties in self-supervised vision transformers," 2021, *arXiv:2104.14294*.

[98] H. Bao, L. Dong, and F. Wei, "BEiT: BERT pre-training of image transformers," 2021, *arXiv:2106.08254*.

[99] S. Paul and P.-Y. Chen, "Vision transformers are robust learners," 2021, *arXiv:2105.07581*.

[100] M. Raghu, T. Unterthiner, S. Kornblith, C. Zhang, and A. Dosovitskiy, "Do vision transformers see like convolutional neural networks?," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 12116–12128.

[101] S. Cao, P. Xu, and D. A. Clifton, "How to understand masked autoencoders," 2022, *arXiv:2202.03670*.

[102] J. Lu, D. Batra, D. Parikh, and S. Lee, "ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," 2019, *arXiv:1908.02265*.

[103] H. Tan and M. Bansal, "LXMERT: Learning cross-modality encoder representations from transformers," 2019, *arXiv:1908.07490*.

[104] L. H. Li, M. Yatskar, D. Yin, C.-J. Hsieh, and K.-W. Chang, "VisualBERT: A simple and performant baseline for vision and language," 2019, *arXiv:1908.03557*.

[105] W. Su et al., "VL-BERT: Pre-training of generic visual-linguistic representations," 2019, *arXiv:1908.08530*.

[106] Y.-C. Chen et al., "UNITER: Universal image-text representation learning," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 104–120.

[107] C. Sun, F. Baradel, K. Murphy, and C. Schmid, "Learning video representations using contrastive bidirectional transformer," 2019, *arXiv:1906.05743*.

[108] G. Li, N. Duan, Y. Fang, M. Gong, and D. Jiang, "Unicoder-VL: A universal encoder for vision and language by cross-modal pre-training," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 11336–11344.

[109] C. Alberti, J. Ling, M. Collins, and D. Reitter, "Fusion of detected objects in text for visual question answering," 2019, *arXiv:1908.05054*.

[110] L. Zhou, H. Palangi, L. Zhang, H. Hu, J. Corso, and J. Gao, "Unified vision-language pre-training for image captioning and VQA," in *Proc. AAAI Conf. Artif. Intell.*, 2020, pp. 13041–13049.

[111] J. Lu, V. Goswami, M. Rohrbach, D. Parikh, and S. Lee, "12-in-1: Multi-task vision and language representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10434–10443.

[112] X. Li et al., "Oscar: Object-semantics aligned pre-training for vision-language tasks," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 121–137.

[113] Z. Huang, Z. Zeng, B. Liu, D. Fu, and J. Fu, "Pixel-BERT: Aligning image pixels with text by deep multi-modal transformers," 2020, *arXiv:2004.00849*.

[114] L. Zhu and Y. Yang, "ActBERT: Learning global-local video-text representations," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 8743–8752.

[115] D. Qi, L. Su, J. Song, E. Cui, T. Bharti, and A. Sacheti, "ImageBERT: Cross-modal pre-training with large-scale weak-supervised image-text data," 2020, *arXiv:2001.07966*.

[116] L. Li, Y.-C. Chen, Y. Cheng, Z. Gan, L. Yu, and J. Liu, "HERO: Hierarchical encoder for video+ language omni-representation pre-training," 2020, *arXiv:2005.00200*.

[117] H. Luo et al., "UniVL: A unified video and language pre-training model for multimodal understanding and generation," 2020, *arXiv:2002.06353*.

[118] M. Xu et al., "A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model," 2021, *arXiv:2112.14757*.

[119] C. Jia et al., "Scaling up visual and vision-language representation learning with noisy text supervision," 2021, *arXiv:2102.05918*.

[120] Z. Wang et al., "CLIP-TD: CLIP targeted distillation for vision-language tasks," 2022, *arXiv:2201.05729*.

[121] J. Li, R. Selvaraju, A. Gotmare, S. Joty, C. Xiong, and S. C. H. Hoi, "Align before fuse: Vision and language representation learning with momentum distillation," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 9694–9705.

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

[123] P. Sharma, N. Ding, S. Goodman, and R. Soricut, "Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning," in *Proc. Conf. Assoc. Comput. Linguistics*, 2018, pp. 2556–2565.

[124] T.-Y. Lin et al., "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[125] S. Antol et al., "VQA: Visual question answering," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2425–2433.

[126] R. Krishna et al., "Visual genome: Connecting language and vision using crowdsourced dense image annotations," *Int. J. Comput. Vis.*, vol. 123, pp. 32–73, 2017.

[127] V. Ordonez, G. Kulkarni, and T. Berg, "Im2Text: Describing images using 1 million captioned photographs," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2011, pp. 1143–1151.

[128] M. Kayser et al., "e-ViL: A dataset and benchmark for natural language explanations in vision-language tasks," 2021, *arXiv:2105.03761*.

[129] J. Gamper and N. Rajpoot, "Multiple instance captioning: Learning representations from histopathology textbooks and articles," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 16544–16554.

[130] L. Li, J. Lei, Z. Gan, and J. Liu, "Adversarial VQA: A new benchmark for evaluating the robustness of VQA models," 2021, *arXiv:2106.00245*.

[131] A. Talmor et al., "MultiModalQA: Complex question answering over text, tables and images," 2021, *arXiv:2104.06039*.

[132] L. Li et al., "VALUE: A multi-task benchmark for video-and-language understanding evaluation," 2021, *arXiv:2106.04632*.

[133] H. Wu et al., "Fashion IQ: A new dataset towards retrieving images by natural language feedback," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11302–11312.

[134] T. Afouras, J. S. Chung, A. Senior, O. Vinyals, and A. Zisserman, "Deep audio-visual speech recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 12, pp. 8717–8727, Dec. 2022.

[135] R. Krishna, K. Hata, F. Ren, L. Fei-Fei, and J. C. Niebles, "Dense-captioning events in videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 706–715.

[136] A. Das et al., "Visual dialog," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 1080–1089.

[137] X. Zhan et al., "Product1M: Towards weakly supervised instance-level product retrieval via cross-modal pretraining," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 11762–11771.

[138] S. Changpinyo, P. Sharma, N. Ding, and R. Soricut, "Conceptual 12M: Pushing web-scale image-text pre-training to recognize long-tail visual concepts," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3557–3567.

[139] Y. Huo et al., "WenLan: Bridging vision and language by large-scale multi-modal pre-training," 2021, *arXiv:2103.06561*.

[140] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "Just ask: Learning to answer questions from millions of narrated videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1666–1677.

[141] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic, "HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 2630–2640.

[142] X. Hu et al., "Scaling up vision-language pre-training for image captioning," 2021, *arXiv:2111.12233*.

[143] C. Schuhmann et al., "LAION-400M: Open dataset of CLIP-filtered 400 million image-text pairs," 2021, *arXiv:2111.02114*.

[144] H. Yun, Y. Yu, W. Yang, K. Lee, and G. Kim, "Pano-AVQA: Grounded audio-visual question answering on 360° videos," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2011–2021.

[145] P. Morgado, Y. Li, and N. Vasconcelos, "Learning representations from audio-visual spatial alignment," 2020, *arXiv:2011.01819*.

[146] R. Li, S. Yang, D. A. Ross, and A. Kanazawa, "AI choreographer: Music conditioned 3D dance generation with AIST++," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 13381–13392.

[147] P. Achlioptas, M. Ovsjanikov, K. Haydarov, M. Elhoseiny, and L. J. Guibas, "ArtEmis: Affective language for visual art," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 11564–11574.

[148] P. P. Liang et al., "MultiBench: Multiscale benchmarks for multimodal representation learning," 2021, *arXiv:2107.07502*.

[149] Z. Liu, C. Rodriguez-Opazo, D. Teney, and S. Gould, "Image retrieval on real-life images with pre-trained vision-and-language models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2105–2114.

[150] P.-L. Guhur, M. Tapaswi, S. Chen, I. Laptev, and C. Schmid, "AirBERT: In-domain pretraining for vision-and-language navigation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1614–1623.

[151] R. Sawhney, M. Goyal, P. Goel, P. Mathur, and R. Shah, "Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 6751–6762.

[152] J. Zhang, M. Zheng, M. Boyd, and E. Ohn-Bar, "X-world: Accessibility, vision, and autonomy meet," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 9742–9751.

[153] D. Zhang, M. Zhang, H. Zhang, L. Yang, and H. Lin, "MultiMET: A multimodal dataset for metaphor understanding," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 3214–3225.

[154] D. Kiela et al., "The hateful memes challenge: Detecting hate speech in multimodal memes," 2020, *arXiv:2005.04790*.

[155] L. Zhou, C. Xu, and J. J. Corso, "Towards automatic learning of procedures from web instructional videos," in *Proc. AAAI Conf. Artif. Intell.*, 2018, pp. 7590–7598.

[156] J. Malmaud, J. Huang, V. Rathod, N. Johnston, A. Rabinovich, and K. Murphy, "What's Cookin'? Interpreting cooking videos using text, speech and vision," 2015, *arXiv:1503.01558*.

[157] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković, "Geometric deep learning: Grids, groups, graphs, geodesics, and gauges," 2021, *arXiv:2104.13478*.

[158] V. P. Dwivedi and X. Bresson, "A generalization of transformer networks to graphs," 2020, *arXiv:2012.09699*.

[159] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 770–778.

[160] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[161] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," 2016, *arXiv:1607.06450*.

[162] R. Xiong et al., "On layer normalization in the transformer architecture," in *Proc. Int. Conf. Mach. Learn.*, 2020, Art. no. 975.

[163] P. Xu, C. K. Joshi, and X. Bresson, "Multigraph transformer for free-hand sketch recognition," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 33, no. 10, pp. 5150–5161, Oct. 2022.

[164] M.-H. Guo, J.-X. Cai, Z.-N. Liu, T.-J. Mu, R. R. Martin, and S.-M. Hu, "PCT: Point cloud transformer," *Comput. Vis. Media*, vol. 7, pp. 187–199, 2021.

[165] S. He, H. Luo, P. Wang, F. Wang, H. Li, and W. Jiang, "TransReID: Transformer-based object re-identification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 14993–15002.

[166] P. Dufter, M. Schmitt, and H. Schütze, "Position information in transformers: An overview," 2021, *arXiv:2102.11090*.

[167] X. Wang, R. Girshick, A. Gupta, and K. He, "Non-local neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 7794–7803.

[168] L. Zhou, Y. Zhou, J. J. Corso, R. Socher, and C. Xiong, "End-to-end dense video captioning with masked transformer," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 8739–8748.

[169] B. Wang, R. Shin, X. Liu, O. Polozov, and M. Richardson, "RAT-SQL: Relation-aware schema encoding and linking for Text-to-SQL parsers," 2019, *arXiv:1911.04942*.

[170] Z. Wang et al., "SGEITL: Scene graph enhanced image-text learning for visual commonsense reasoning," 2021, *arXiv:2112.08587*.

[171] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Statist.*, 2011, pp. 315–323.

[172] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[173] Y.-B. Lin, H.-Y. Tseng, H.-Y. Lee, Y.-Y. Lin, and M.-H. Yang, "Exploring cross-video and cross-modality signals for weakly-supervised audio-visual video parsing," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 11449–11461.

[174] H. Akbari et al., "VATT: Transformers for multimodal self-supervised learning from raw video, audio and text," 2021, *arXiv:2104.11178*.

[175] A. Nagrani, S. Yang, A. Arnab, A. Jansen, C. Schmid, and C. Sun, "Attention bottlenecks for multimodal fusion," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 14200–14213.

[176] P.-A. Duquenne, H. Gong, and H. Schwenk, "Multimodal and multilingual embeddings for large-scale speech mining," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 15748–15761.

[177] D. Min, D. B. Lee, E. Yang, and S. J. Hwang, "Meta-StyleSpeech: Multi-speaker adaptive text-to-speech generation," 2021, *arXiv:2106.03153*.

[178] B. Shi, W.-N. Hsu, K. Lakhotia, and A. Mohamed, "Learning audio-visual speech representation by masked multimodal cluster prediction," 2022, *arXiv:2201.02184*.

[179] K. Ramesh, C. Xing, W. Wang, D. Wang, and X. Chen, "Vset: A multimodal transformer for visual speech enhancement," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 6658–6662.

[180] R. Zheng, J. Chen, M. Ma, and L. Huang, "Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation," 2021, *arXiv:2102.05766*.

[181] X. Yang, S. Feng, Y. Zhang, and D. Wang, "Multimodal sentiment detection based on multi-channel graph neural networks," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 328–339.

[182] X. Mao et al., "Towards robust vision transformer," 2021, *arXiv:2105.07926*.

[183] T. Rahman, M. Yang, and L. Sigal, "TriBERT: Human-centric audio-visual representation learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 9774–9787.

[184] L. Rasmy, Y. Xiang, Z. Xie, C. Tao, and D. Zhi, "Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction," *NPJ Digit. Med.*, vol. 4, 2021, Art. no. 86.

[185] R. J. Chen et al., "Multimodal co-attention transformer for survival prediction in gigapixel whole slide images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 3995–4005.

[186] S. Xie, C. Sun, J. Huang, Z. Tu, and K. Murphy, "Rethinking spatiotemporal feature learning for video understanding," 2017, *arXiv:1712.04851*.

[187] D. Tran, H. Wang, L. Torresani, J. Ray, Y. LeCun, and M. Paluri, "A closer look at spatiotemporal convolutions for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 6450–6459.

[188] J. Lin, A. Yang, Y. Zhang, J. Liu, J. Zhou, and H. Yang, "InterBERT: Vision-and-language interaction for multi-modal pretraining," 2020, *arXiv:2003.13198*.

[189] Y.-H. H. Tsai, S. Bai, P. P. Liang, J. Z. Kolter, L.-P. Morency, and R. Salakhutdinov, "Multimodal transformer for unaligned multimodal language sequences," in *Proc. Conf. Assoc. Comput. Linguistics*, 2019, pp. 6558–6569.

[190] V. Murahari, D. Batra, D. Parikh, and A. Das, "Large-scale pretraining for visual dialog: A simple state-of-the-art baseline," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 336–352.

[191] Y. Li, Y. Pan, T. Yao, J. Chen, and T. Mei, "Scheduled sampling in vision-language pretraining with decoupled encoder-decoder network," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 8518–8526.

[192] W. Kim, B. Son, and I. Kim, "ViLT: Vision-and-language transformer without convolution or region supervision," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 5583–5594.

[193] J. Yang et al., "Vision-language pre-training with triple contrastive learning," 2022, *arXiv:2202.10401*.

[194] L. H. Li et al., "Grounded language-image pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 10955–10965.

[195] H. Zhang et al., "GLIPv2: Unifying localization and vision-language understanding," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 36067–36080.

[196] C. Li et al., "SemVLP: Vision-language pre-training by aligning semantics at multiple levels," 2021, *arXiv:2103.07829*.

[197] A. Miech, J.-B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman, "End-to-end learning of visual representations from uncurated instructional videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 9876–9886.

[198] T. Han, W. Xie, and A. Zisserman, "Temporal alignment networks for long-term video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 2896–2906.

[199] Z. Wang, J. Yu, A. W. Yu, Z. Dai, Y. Tsvetkov, and Y. Cao, "SimVLM: Simple visual language model pretraining with weak supervision," 2021, *arXiv:2108.10904*.

[200] J. Chen, M. Ma, R. Zheng, and L. Huang, "MAM: Masked acoustic modeling for end-to-end speech-to-text translation," 2020, *arXiv:2010.11445*.

[201] M. Golestani, S. Z. Razavi, Z. Borhanifard, F. Tahmasebian, and H. Faili, "Using BERT encoding and sentence-level language model for sentence ordering," in *Proc. 24th Int. Conf. Text Speech Dialogue*, 2021, pp. 318–330.

[202] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 91–99.

[203] Z. Huang, Z. Zeng, Y. Huang, B. Liu, D. Fu, and J. Fu, "Seeing out of the box: End-to-end pre-training for vision-language representation learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12971–12980.

[204] Y. Liu, C. Wu, S.-Y. Tseng, V. Lal, X. He, and N. Duan, "KD-VLP: Improving end-to-end vision-and-language pretraining with object knowledge distillation," 2021, *arXiv:2109.10504*.

[205] A. Ramesh et al., "Zero-shot text-to-image generation," 2021, *arXiv:2102.12092*.

[206] M. Zhou et al., "UC^2: Universal cross-lingual cross-modal vision-and-language pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4153–4163.

[207] W. Hao, C. Li, X. Li, L. Carin, and J. Gao, "Towards learning a generic agent for vision-and-language navigation via pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 13134–13143.

[208] Q. Xia et al., "XGPT: Cross-modal generative pre-training for image captioning," in *Proc. 10th CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, 2021, pp. 786–797.

[209] L. Yao et al., "FILIP: Fine-grained interactive language-image pre-training," 2021, *arXiv:2111.07783*.

[210] Z. Gan, Y.-C. Chen, L. Li, C. Zhu, Y. Cheng, and J. Liu, "Large-scale adversarial training for vision-and-language representation learning," 2020, *arXiv:2006.06195*.

[211] H. Zhang et al., "ERNIE-ViLG: Unified generative pre-training for bidirectional vision-language generation," 2021, *arXiv:2112.15283*.

[212] M. Zhuge et al., "Kaleido-BERT: Vision-language pre-training on fashion domain," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12642–12652.

[213] Y. Wang, X. Chen, L. Cao, W. Huang, F. Sun, and Y. Wang, "Multimodal token fusion for vision transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12176–12185.

[214] Y. Wang et al., "Bridged transformer for vision and point cloud 3D object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 12104–12113.

[215] A. Prakash, K. Chitta, and A. Geiger, "Multi-modal fusion transformer for end-to-end autonomous driving," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7073–7083.

[216] X. Bai et al., "TransFusion: Robust LiDAR-camera fusion for 3D object detection with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 1080–1089.

[217] X. Favory, K. Drossos, T. Virtanen, and X. Serra, "Learning contextual tag embeddings for cross-modal alignment of audio and tags," in *Proc. Int. Conf. Acoust. Speech Signal Process.*, 2021, pp. 596–600.

[218] V. Gabeur, C. Sun, K. Alahari, and C. Schmid, "Multi-modal transformer for video retrieval," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 214–229.

[219] N. Shvetsova et al., "Everything at once - multi-modal fusion transformer for video retrieval," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 19988–19997.

[220] Z. Wang, Y. Wu, K. Narasimhan, and O. Russakovsky, "Multi-query video retrieval," 2022, *arXiv:2201.03639*.

[221] A. Botach, E. Zheltonozhskii, and C. Baskin, "End-to-end referring video object segmentation with multimodal transformers," 2021, *arXiv:2111.14821*.

[222] W. Hong, K. Ji, J. Liu, J. Wang, J. Chen, and W. Chu, "GilBERT: Generative vision-language pre-training for image-text retrieval," in *Proc. 44th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2021, pp. 1379–1388.

[223] X. Xu and C. C. Loy, "3D human texture estimation from a single image with transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 13829–13838.

[224] X. Lin, C. Ding, J. Zeng, and D. Tao, "GPS-Net: Graph property sensing network for scene graph generation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3743–3752.

[225] W. Wang, R. Wang, and X. Chen, "Topic scene graph generation by attention distillation from caption," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 15880–15890.

[226] Y. Lu et al., "Context-aware scene graph generation with Seq2Seq transformers," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 15911–15921.

[227] P. Ke et al., "JointGT: Graph-text joint representation learning for text generation from knowledge graphs," 2021, *arXiv:2106.10502*.

[228] Y. Teng, L. Wang, Z. Li, and G. Wu, "Target adaptive context aggregation for video scene graph generation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 13668–13677.

[229] M. Chen, Y. Li, Z. Zhang, and S. Huang, "TVT: Two-view transformer network for video captioning," in *Proc. 10th Asian Conf. Mach. Learn.*, 2018, pp. 847–862.

[230] K. Lin et al., "SwinBERT: End-to-end transformers with sparse attention for video captioning," 2021, *arXiv:2111.13196*.

[231] C. Deng, S. Chen, D. Chen, Y. He, and Q. Wu, "Sketch, ground, and refine: Top-down dense video captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 234–243.

[232] T. Wang, R. Zhang, Z. Lu, F. Zheng, R. Cheng, and P. Luo, "End-to-end dense video captioning with parallel decoding," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 6827–6837.

[233] L. Huang, W. Wang, J. Chen, and X.-Y. Wei, "Attention on attention for image captioning," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2019, pp. 4633–4642.

[234] Y. Pan, T. Yao, Y. Li, and T. Mei, "X-linear attention networks for image captioning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10968–10977.

[235] X. Yang, H. Zhang, G. Qi, and J. Cai, "Causal attention for vision-language tasks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9842–9852.

[236] Y. Luo et al., "Dual-level collaborative transformer for image captioning," 2021, *arXiv:2101.06462*.

[237] G. Xu, S. Niu, M. Tan, Y. Luo, Q. Du, and Q. Wu, "Towards accurate text-based image captioning with content diversity exploration," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 12632–12641.

[238] N. Li, S. Liu, Y. Liu, S. Zhao, and M. Liu, "Neural speech synthesis with transformer network," in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6706–6713.

[239] M. Ding et al., "CogView: Mastering text-to-image generation via transformers," 2021, *arXiv:2105.13290*.

[240] A. Sanghi, H. Chu, J. G. Lambourne, Y. Wang, C.-Y. Cheng, and M. Fumero, "CLIP-forge: Towards zero-shot text-to-shape generation," 2021, *arXiv:2110.02624*.

[241] R. Huang, H. Hu, W. Wu, K. Sawada, M. Zhang, and D. Jiang, "Dance revolution: Long-term dance generation with music via curriculum learning," 2020, *arXiv:2006.06119*.

[242] Y. Zhong, J. Shi, J. Yang, C. Xu, and Y. Li, "Learning to generate scene graph from natural language supervision," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1803–1814.

[243] S. Geng et al., "Dynamic graph representation learning for video dialog via multi-modal shuffled transformers," in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 1415–1423.

[244] T. Jaunet, C. Kervadec, R. Vuillemot, G. Antipov, M. Baccouche, and C. Wolf, "VisQA: X-raying vision and language reasoning in transformers," *IEEE Trans. Vis. Comput. Graph.*, vol. 28, no. 1, pp. 976–986, Jan. 2022.

[245] X. Lin, G. Bertasius, J. Wang, S.-F. Chang, D. Parikh, and L. Torresani, "VX2TEXT: End-to-end learning of video-based text generation from multimodal inputs," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7001–7011.

[246] J. Lin et al., "M6: Multi-modality-to-multi-modality multitask mega-transformer for unified pretraining," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discov. Data Mining*, 2021, pp. 3251–3261.

[247] T. Chen and R. R. Rao, "Audio-visual integration in multimodal communication," *Proc. IEEE*, vol. 86, no. 5, pp. 837–852, May 1998.

[248] A. Owens and A. A. Efros, "Audio-visual scene analysis with self-supervised multisensory features," in *Proc. Eur. Conf. Comput. Vis.*, 2018, pp. 639–658.

[249] H. Xue et al., "Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 4514–4528.

[250] T.-D. Truong et al., "The right to talk: An audio-visual transformer approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1085–1094.

[251] M. Chen et al., "MultiSpeech: Multi-speaker text to speech with transformer," 2020, *arXiv:2006.04664*.

[252] S. Ging, M. Zolfaghari, H. Pirsiavash, and T. Brox, "COOT: Cooperative hierarchical transformer for video-text representation learning," 2020, *arXiv:2011.00597*.

[253] M. Patrick et al., "Support-set bottlenecks for video-text representation learning," 2020, *arXiv:2010.02824*.

[254] V. Gabeur, A. Nagrani, C. Sun, K. Alahari, and C. Schmid, "Masking modalities for cross-modal video retrieval," in *Proc. Winter Conf. Appl. Comput. Vis.*, 2022, pp. 2111–2120.

[255] A. Sadhu, K. Chen, and R. Nevatia, "Video object grounding using semantic roles in language description," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 10414–10424.

[256] Y. Zhang, M. Choi, K. Han, and Z. Liu, "Explainable semantic space by grounding language to vision with cross-modal contrastive learning," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 18513–18526.

[257] Y.-W. Chen, Y.-H. Tsai, and M.-H. Yang, "End-to-end multi-modal video temporal grounding," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 28442–28453.

[258] S. Chen and B. Li, "Multi-modal dynamic graph transformer for visual grounding," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 15513–15522.

[259] A. Yang, A. Miech, J. Sivic, I. Laptev, and C. Schmid, "TubeDETR: Spatio-temporal video grounding with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 16421–16432.

[260] H. Xu et al., "VideoCLIP: Contrastive pre-training for zero-shot video-text understanding," 2021, *arXiv:2109.14084*.

[261] J. Lei et al., "Less is more: CLIPBERT for video-and-language learning via sparse sampling," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7327–7337.

[262] J. Yang, Y. Bisk, and J. Gao, "TACo: Token-aware cascade contrastive learning for video-text alignment," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 11542–11552.

[263] D. Li, J. Li, H. Li, J. C. Niebles, and S. C. Hoi, "Align and prompt: Video-and-language pre-training with entity prompts," 2021, *arXiv:2112.09583*.

[264] H. Luo et al., "CLIP4Clip: An empirical study of CLIP for end to end video clip retrieval," 2021, *arXiv:2104.08860*.

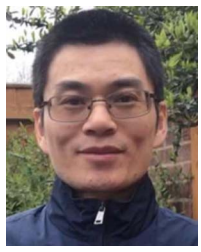
[265] H. Fang, P. Xiong, L. Xu, and Y. Chen, "CLIP2Video: Mastering video-text retrieval via image CLIP," 2021, *arXiv:2106.11097*.

[266] M. Narasimhan, A. Rohrbach, and T. Darrell, "CLIP-It! Language-guided video summarization," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 13988–14000.

[267] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, 2023, Art. no. 195.

[268] Y. Xian, C. H. Lampert, B. Schiele, and Z. Akata, "Zero-shot learning—A comprehensive evaluation of the good, the bad and the ugly," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 9, pp. 2251–2265, Sep. 2019.
[269] X. Gu, T.-Y. Lin, W. Kuo, and Y. Cui, "Open-vocabulary object detection via vision and language knowledge distillation," 2021, *arXiv:2104.13921*.
[270] J. Cho, J. Lu, D. Schwenk, H. Hajishirzi, and A. Kembhavi, "X-LXMERT: Paint, caption and answer questions with multi-modal transformers," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 8785–8805.
[271] H. Xu et al., "E2E-VLP: End-to-end vision-language pre-training enhanced by visual learning," 2021, *arXiv:2106.01804*.
[272] C. Kervadec, C. Wolf, G. Antipov, M. Baccouche, and M. Nadri, "Supervising the transfer of reasoning patterns in VQA," 2021, *arXiv:2106.05597*.
[273] C. Kervadec, T. Jaunet, G. Antipov, M. Baccouche, R. Vuillemot, and C. Wolf, "How transferable are reasoning patterns in VQA?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 4205–4214.
[274] W. Rahman et al., "Integrating multimodal information in large pretrained transformers," in *Proc. Conf. Assoc. Comput. Linguistics*, 2020, pp. 2359–2369.
[275] D. Agarwal, T. Agrawal, L. M. Ferrari, and F. Bremond, "From multimodal to unimodal attention in transformers using knowledge distillation," in *Proc. 17th IEEE Int. Conf. Adv. Video Signal Based Surveill.*, 2021, pp. 1–8.
[276] Q. Li et al., "Towards a unified foundation model: Jointly pre-training transformers on unpaired images and text," 2021, *arXiv:2112.07074*.
[277] M. Ni et al., "M3P: Learning universal representations via multitask multilingual multimodal pre-training," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 3976–3985.
[278] A. Miech, J.-B. Alayrac, I. Laptev, J. Sivic, and A. Zisserman, "Thinking fast and slow: Efficient text-to-visual retrieval with transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 9821–9831.
[279] K. Wen, J. Xia, Y. Huang, L. Li, J. Xu, and J. Shao, "COOKIE: Contrastive cross-modal knowledge sharing pre-training for vision-language representation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 2188–2197.
[280] S. Lee, Y. Yu, G. Kim, T. Breuel, J. Kautz, and Y. Song, "Parameter efficient multimodal transformers for video representation learning," 2020, *arXiv:2012.04124*.
[281] T. Liu, F. Feng, and X. Wang, "Multi-stage pre-training over simplified multimodal pre-training models," 2021, *arXiv:2107.14596*.
[282] Y. Li et al., "Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm," 2021, *arXiv:2110.05208*.
[283] Z. Gan et al., "Playing lottery tickets with vision and language," 2021, *arXiv:2104.11832*.
[284] C. He, S. Li, M. Soltanolkotabi, and S. Avestimehr, "PipeTransformer: Automated elastic pipelining for distributed training of large-scale models," in *Proc. Int. Conf. Mach. Learn.*, 2021, pp. 4150–4159.
[285] T. Dao, D. Fu, S. Ermon, A. Rudra, and C. Ré, "FlashAttention: Fast and memory-efficient exact attention with IO-awareness," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2022, pp. 16344–16359.
[286] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.
[287] C. Zhu et al., "Long-short transformer: Efficient transformers for language and vision," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 17723–17736.
[288] S. Yan et al., "Multiview transformers for video recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 3323–3333.
[289] W. Wang and Z. Tu, "Rethinking the value of transformer components," in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 6019–6029.
[290] M. Ma, J. Ren, L. Zhao, D. Testuggine, and X. Peng, "Are multimodal transformers robust to missing modality?," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 18156–18165.
[291] A. Akula, V. Jampani, S. Changpinyo, and S.-C. Zhu, "Robust visual reasoning via language guided neural module networks," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, 2021, pp. 11041–11053.
[292] A. R. Akula, S. Gella, Y. Al-Onaizan, S.-C. Zhu, and S. Reddy, "Words aren't enough, their order matters: On the robustness of grounding visual referring expressions," 2020, *arXiv:2005.01655*.
[293] L. Li, Z. Gan, and J. Liu, "A closer look at the robustness of vision-and-language pre-trained models," 2020, *arXiv:2012.08673*.
[294] M. Zhang, T. Maidment, A. Diab, A. Kovashka, and R. Hwa, "Domain-robust VQA with diverse datasets and methods but no target labels," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 7042–7052.
[295] Y. Kant, A. Moudgil, D. Batra, D. Parikh, and H. Agrawal, "Contrast and classify: Training robust VQA models," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2021, pp. 1584–1593.
[296] S. Pramanik, P. Agrawal, and A. Hussain, "OmniNet: A unified architecture for multi-modal multi-task learning," 2019, *arXiv:1907.07804*.
[297] P. Wang et al., "Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," 2022, *arXiv:2202.03052*.
[298] R. Girdhar, M. Singh, N. Ravi, L. van der Maaten, A. Joulin, and I. Misra, "Omnivore: A single model for many visual modalities," 2022, *arXiv:2201.08377*.
[299] J. Cao, Z. Gan, Y. Cheng, L. Yu, Y.-C. Chen, and J. Liu, "Behind the scene: Revealing the secrets of pre-trained vision-and-language models," in *Proc. Eur. Conf. Comput. Vis.*, 2020, pp. 565–580.
[300] L. A. Hendricks, J. Mellor, R. Schneider, J.-B. Alayrac, and A. Nematzadeh, "Decoupling the role of data, attention, and losses in multimodal transformers," *Trans. Assoc. Comput. Linguistics*, vol. 9, pp. 570–585, 2021.
[301] L. A. Hendricks and A. Nematzadeh, "Probing image-language transformers for verb understanding," 2021, *arXiv:2106.09141*.
[302] S. Frank, E. Bugliarello, and D. Elliott, "Vision-and-language or vision-for-language? On cross-modal influence in multimodal transformers," 2021, *arXiv:2109.04448*.
[303] H. Chefer, S. Gur, and L. Wolf, "Generic attention-model explainability for interpreting bi-modal and encoder-decoder transformers," 2021, *arXiv:2103.15679*.
[304] L. Parcalabescu, M. Cafagna, L. Muradjan, A. Frank, I. Calixto, and A. Gatt, "VALSE: A task-independent benchmark for vision and language models centered on linguistic phenomena," 2021, *arXiv:2112.07566*.
[305] T. Zhao et al., "VL-CheckList: Evaluating pre-trained vision-language models with objects, attributes and relations," 2022, *arXiv:2207.00221*.
[306] E. Aflalo et al., "VL-InterpreT: An interactive visualization tool for interpreting vision-language transformers," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2022, pp. 21374–21383.
[307] N. Mu, A. Kirillov, D. Wagner, and S. Xie, "SLIP: Self-supervision meets language-image pre-training," 2021, *arXiv:2112.12750*.
[308] H. Xu et al., "VLM: Task-agnostic video-language model pre-training for video understanding," 2021, *arXiv:2105.09996*.
[309] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," 2022, *arXiv:2201.12086*.
[310] P. Zhang et al., "VinVL: Revisiting visual representations in vision-language models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2021, pp. 5575–5584.

Peng Xu is a lecturer with the Department of Electronic Engineering, Tsinghua University. Previously, he was a postdoctoral research assistant with the Department of Engineering Science, University of Oxford.



Xiatian Zhu is a senior lecturer with the Surrey Institute for People-Centred Artificial Intelligence, and Centre for Vision, Speech and Signal Processing (CVSSP), Faculty of Engineering and Physical Sciences, University of Surrey.

David A. Clifton is a professor of clinical machine learning and leads the Computational Health Informatics (CHI) Lab, Department of Engineering Science, University of Oxford.