Multimodal Transformers: A survey on Architecture and existing challenges in Multimodal transformers

Sowmya Jayaram Iyer
Department of Computer Science
Purdue University
jayarami@purdue.edu

Abstract

Transformers have only been gaining more attention in the research community due to their promising success in various tasks across Deep Learning. Developing transformer-based models that can extract and relate information from multimodal data has become increasingly popular due to the prevalence of multimodal applications and big data. This project aims to present a comprehensive survey of Transformer techniques specifically geared towards multimodal data. The expected contributions of this survey include giving: (1) a theoretical review of Multimodal learning, Transformers, Vision Transformers, and Multimodal Transformers, (2) a review of multimodal Transformers through the perspective of two important applications-specific multimodal tasks, (3) to summarize commonalities in challenges faced and designs of existing Transformer models and their applications.

1. Introduction

In order to interact with the world appropriately under dynamic, unrestricted conditions, humans have a fundamental mechanism in our sensory perception that allows us to combine data from multiple perception modalities, with each modality acting as a separate information source with its own set of statistical properties. Generally, a modality is commonly associated with a particular sensor that creates a special communication pathway, such as vision and language [3]. In order to achieve equal human level perceptual skills, a multimodal AI system must fundamentally consume, understand, and reason about multimodal input sources. A broad method for creating AI models that can extract and link information from multimodal input is called multimodal learning (MML).

This report will concentrate on multimodal learning with Transformers, motivated by their inherent advantages and scalability in modeling various modalities (such as language, visual, and audio) and tasks (such as visual question answering, image retrieval, and speech translation and generation), which require fewer architectural assumptions that are modality-specific (such as local grid attention bias in vision and translation invariance). Transformer input might include one or more sequences of tokens and each sequence's attribute (e.g., modality label, sequential order), enabling MML without architectural change [21]. Controlling self-attention input pattern enables learning per-modal specificity and inter-modal correlation. Recent research efforts and activities across disciplines studying Transformer designs have led to a huge number of unique MML approaches and considerable breakthroughs in several domains [7, 21, 23, 73, 94]. This begs for a timely evaluation and summary of representative approaches to help academics comprehend the whole picture of MML across disciplines and capture a comprehensive organized picture of current successes and main problems.

We use a two-tiered taxonomy based on application and problem dimensions for greater readability and accessibility across disciplines. This enables researchers with knowledge in particular applications to locate their own research domain's applications before linking to adjacent areas. Also, similar model designs and architectures generated in different domains may be described in an abstract, formula-driven viewpoint along with mathematical concepts can be associated and compared. This may help break down domain barriers and promote cross-modal idea sharing.

The scope of this survey is limited to discussing the multi-modal specific designs of Transformer architecture to modalities such as: RGB images, RGB-D images, Videos, Audio/Speech, Scene graph, Pose data, Point cloud, Multimodal knowledge graph, 3D object, 3D scene, Healthcare data. The aim is to contribute a taxonomy for Transformer for MML from application based and challenge based perspectives. This survey excludes multimodal publications where Transformer is employed just as a feature extractor. The survey's key features can be summarized as (1)

Highlight transformers which are modality-agnostic, that is, they're compatible with several modalities (and combinations). To support this notion, we give a geometrically topological interpretation of Transformers' inherent multimodal features. Self-attention is suggested to be modeled as a fully-connected graph with uni-modal and multimodal input sequences. Self-attention embeds tokens from any modality as graph nodes.(2) In this research, we extract the mathematical core and formulations of Transformer-based MML methods, from the perspective network architectures eg. uni-stream (UNITER [14]), multi-stream (ViLBERT [28]). Having presented an overview of the current trend with respect to MML and Transformers, our contributions are summarized as:

1. Examine Vision Transformer and multimodal Transformers Architectures.
2. Provide an application-based taxonomy for Transformer-based MML. In Section 4, we discuss multimodal Transformer applications using a major paradigm, namely specialized multimodal tasks. Section 5 summarizes prevalent difficulties and designs for multimodal Transformer models and their applications.
3. Address Transformer-based MML bottlenecks, challenges, and research prospects.

2. Related Work

Surveys in MML such as [3] and [93] exist which give a comprehensive analysis of multimodal learning. This project adopts [3]'s categorisation of task due to its neat structure. These MML surveys speak about general ML models. This project, however, only focuses on Transformers for multimodal data and self-attention mechanisms used in them. Several surveys on Transformers with a wide range of emphasis on various vision tasks have been recently published. While a few talk about MML, they are not in depth and limited in scope and coverage. Papers such as [67] focus specifically on transformers for Video Language Processing which is simply one portion of MML. Though used as reference points, this project aims to widen the scope to an intersection of all MML applications and transformers.

3. Background

3.1. Multi Modal Learning (MML)

MML has been a major study field for decades. Human emotions, behaviors and actions are all multimodal, which have led to various human-centered MML tasks extensively being studied, such as multimodal emotion recognition [55], multimodal event representation [91], *etc.* Due to the internet and a range of intelligent gadgets, more multimodal data has been made available, leading to more multimodal application scenarios. Multimodal applications include autonomous driving [52], communication (sign language translation [6, 88]), surveillance AI [83], healthcare AI [18, 71], *etc.*), vision and language navigation (VLN) ([11, 33, 64]), home navigation [64]. Deep neural networks boost MML development in the age of Deep Learning. Transformers are architecturally competitive and have brought new potential to MML.

3.2. Multi modal Transformers

VideoBERT [73], a ground-breaking effort, was the first to adapt Transformer to multimodal activities inspired by the immense success of Transformers and ViTs. VideoBERT illustrates the tremendous multimodal possibilities of Transformer. In the wake of VideoBERT, a number of Transformer-based multimodal pretraining models (e.g.,UNITER [14], VL-BERT [72], VLP [100], ActBERT [102], and HERO [43]) have emerged as hot research topics in Machine Learning. CLIP [65], proposed in 2021, is a novel benchmark that use multimodal pretraining to transform classification into a retrieval challenge, hence enabling pretrained models to perform zero-shot recognition. Recent research has expanded on the concept of CLIP, such as CLIP pretrained model based zero-shot semantic segmentation [82], and CLIP-TD [81].

MMT4: Multi Modality To Text Transfer Transformer [75], introduced in 2022, is an encoder-decoder model with the T5 (Text To Text Transfer Transformer) base that fuses non-text components with text tokens. The encoder uses relative positional and token type embeddings, while the decoder creates new text matching to various jobs. Another notable model is the ASD-transformer [19], a novel architecture that effectively captures the relationship between audio and visual modalities for ASD by using contemporary transformer and concatenation processes. Multimodal active speaker identification (ASD) techniques categorize each participant in a video clip as speaking or not speaking.

4. Transformer Architecture

In this part, we examine the key approaches of Transformer and multimodal Transformers 1 using mathematical formulations, including tokenized inputs, self-attention, multi-head attention, fundamental Transformer layers/components, etc. Because of the self- attention mechanism, given evey tokenized input from any modality, self-attention (Transformer) can model it as a fully-connected graph. Transformers naturally have a more generic and flexible modeling space than other deep networks (such as CNN, which is constrained to aligned grid spaces/matrices). This is a significant benefit of using Transformers for multimodal tasks. The 4.1 will examine the major Vanilla Transformer designs.

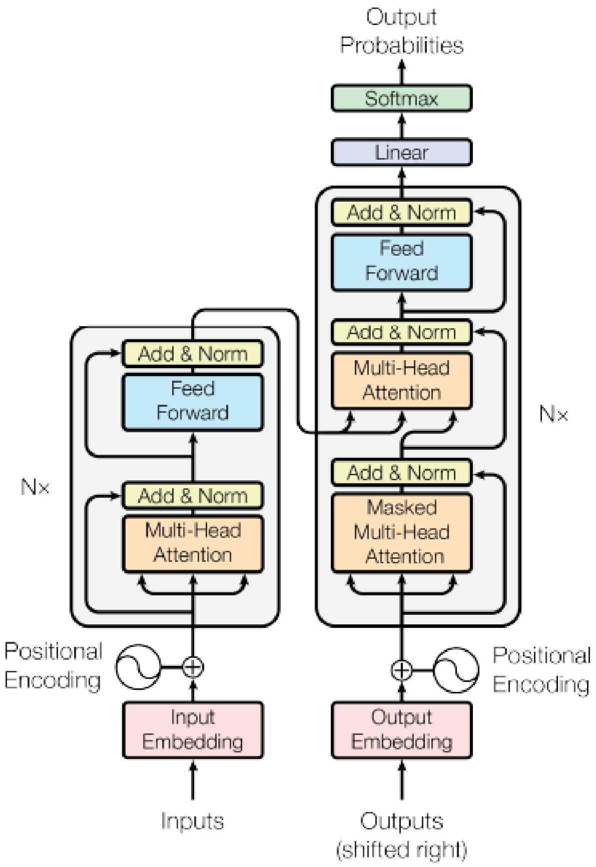


Figure 1. Vanilla Transformer model overview [79]

4.1. Transformer

Vanilla Transformer contains an encoder-decoder structure 1 which takes input as tokens (see Section 4.1.1). As shown in Figure 1, its encoder and decoder are layered by the Transformer layers/blocks. Each block contains two modules: a multi-head self-attention (MHSA) layer (see Section 4.1.2) and a position-wise fully-connected feed-forward network (FN) layer (see Section 4.1.4). Both MHSA and FN employ Residual Connection. Given an input t, residual connection of any mapping $f()$ is defined as $t \rightarrow f(t)+t$. This aids in the back propagation of the gradient. Then comes the normalizing layer. Assuming the input tensor to be T, the output of the MHSA and FN modules may be expressed as follows:

$$T \leftarrow N(M(T)+T) \tag{1}$$

where, N(.) denotes the normalising layer, which can be Batch Normalization $BN(.)$ or Layer Normalisation $LN(.)$ and M(.) denotes the module mapping/operations.

4.1.1 Tokenizing Inputs

Vanilla Transformer was initially proposed for machine translation as a sequence-to-sequence model; therefore, it is simple to input vocabulary sequences. Vanilla Transformer generates position embedding using sine and cosine functions. It may be seen as a kind of implicit coordinate basis of feature space that provides the Transformer with time or spatial information. The token element of cloud point or camera angle is already a coordinate, so position embedding is optional and unnecessary. Moreover, position embedding may be seen as a kind of generic supplementary data. In other words, any extra information may be included, such as a description of the method of position embedding, such as the cameras and views used in surveillance. There is a comprehensive survey [24] discussing the transformers position data.

The key benefits of input tokenization are the following: (1) Tokenization is a more broad technique from a geometrical standpoint, accomplished by reducing restrictions induced by different modalities. (2) Tokenization is a more versatile method for organizing input data via concatenation, labelling, weighted summation, etc. Vanilla Transformer adds temporal information to the token embedding by summing the token's timestamp.

4.1.2 Self-Attention

The fundamental concept behind the working of Transformers is the Self-Attention (SA) mechanism. Let $X = [x_1, x_2, x_3, \ldots] \in \mathcal{R}^{N \times d}$ be the input sequence to the transformer of N elements or tokens. After positional encoding PE is applied, say point-wise, the input is given by $T \leftarrow X \oplus PE$. This pre-processed input is fed into the Self-Attention (SA) module.

The tensor T is passed through three projection matrices to give Q (Query), K (Key) and V (Value) embeddings. The output of the self-attention module is given by:

$$SA(Q,K,V) = Softmax\left(\frac{QK^T}{\sqrt{d_q}}V\right) \tag{2}$$

Given an input sequence, self-attention enable each element to attend to all other elements, encoding the input sequence as a fully-connected graph. Consequently, the encoder of Vanilla Transformer may be viewed as a fully-connected GNN encoder, and the Transformer family has the non-local capability for global perception.

4.1.3 Multi-headed Self-Attention

In the Multi-Head Self-Attention module, multiple self-attention heads are parallelly stacked and their concatenated

Self-Attention	Definitions	Streams	Complexities	References
Early Concatenation	token sequence concatenation + M	1	$\mathcal{O}((N_1+N_2)^2)$	[31,70,73]
Early Summation	token-wise $\oplus$ + M	1	$\mathcal{O}(max(N_1,N_2)^2)$	[29,83]
Hierarchical Attention	2-stream M + concatenation	$2 \rightarrow 1$	$\mathcal{O}((N_1+N_2)^2)$	[44]
Hierarchical Attention	early concatenation + 2-stream M	$1 \rightarrow 2$	$\mathcal{O}((N_1+N_2)^2)$	[48]
Cross-Attention	swap query	2	$\mathcal{O}(max(N_1,N_2)^2)$	[53,90]
Cross-Attention to Concatenation	2-stream M + concatenation	$2 \rightarrow 1$	$\mathcal{O}((N_1+N_2)^2)$	[9,77,92]

Table 1. Self-attention variants for cross-modal fusion. "M": "transformer Layer", N: sequence length

outputs are combined by a projection matrix.

$$\begin{aligned} MHSA(Q,K,V) &= concat(SA1(Q_1,K_1,V_1), \\ &\quad SA2(Q_2,K_2,V_2),\ldots) \\ &= concat(Z_1,Z_2,Z_3,\ldots Z_H) \rightarrow P \\ &= X^* \end{aligned} \tag{3}$$

where H is the number of self-attention heads used and $Z_1, Z_2, \ldots$ are the outputs of each SA head. The concatenation can be done using a projection matrix P. The concatenation is done using a projection matrix. This enables the transformers to pay attention to different sub-spaces.

4.1.4 Feed-forward Network

The output of the multi-head attention module will is passed through a position-wise Feed-forward Network (FN) consisting of stacked linear layers with non-linear activation. A 2-layer Feed Forward Network can be represented as:

$$FN(X^*) = \sigma(X^*W_1 + b_1)W_2 + b_2 \tag{4}$$

where W_1, W_2 are the weights and b_1, b_2 are the biases of the two successive neural layers respectively. $\sigma(.)$ represesents the non-linear activation like ReLu.

4.2. Multimodal Transformers

Lately, Transformers have proven themselves to be compatible with a variety of modalities in both generative and discriminative tasks. This section aims to present a categorization of multimodal transformers based on the key techniques and architectural details of existing multimodal transformers from the lens of SA variants (see Section 4.2.1 and network architecture (see Section 4.2.2).

4.2.1 Based on Self-Attention variants

Cross-modal Interactions in multimodal transformers are primarily processed in the self-attention module. Current works employ various self-attention variants for different use-cases which is discussed in this section. To keep things simple, we'll only describe the arithmetic and compare it in two-modality scenarios. All of the self-attention and their variations, shown in Figure 2 and discussed later in this section, are sufficiently adaptable that they may be used to situations involving more than one modality. The following formulations are independent of the modalities used, the tokenizations used, and the embeddings used since self-attention models the embedding of each token in any modality as a node of a graph.

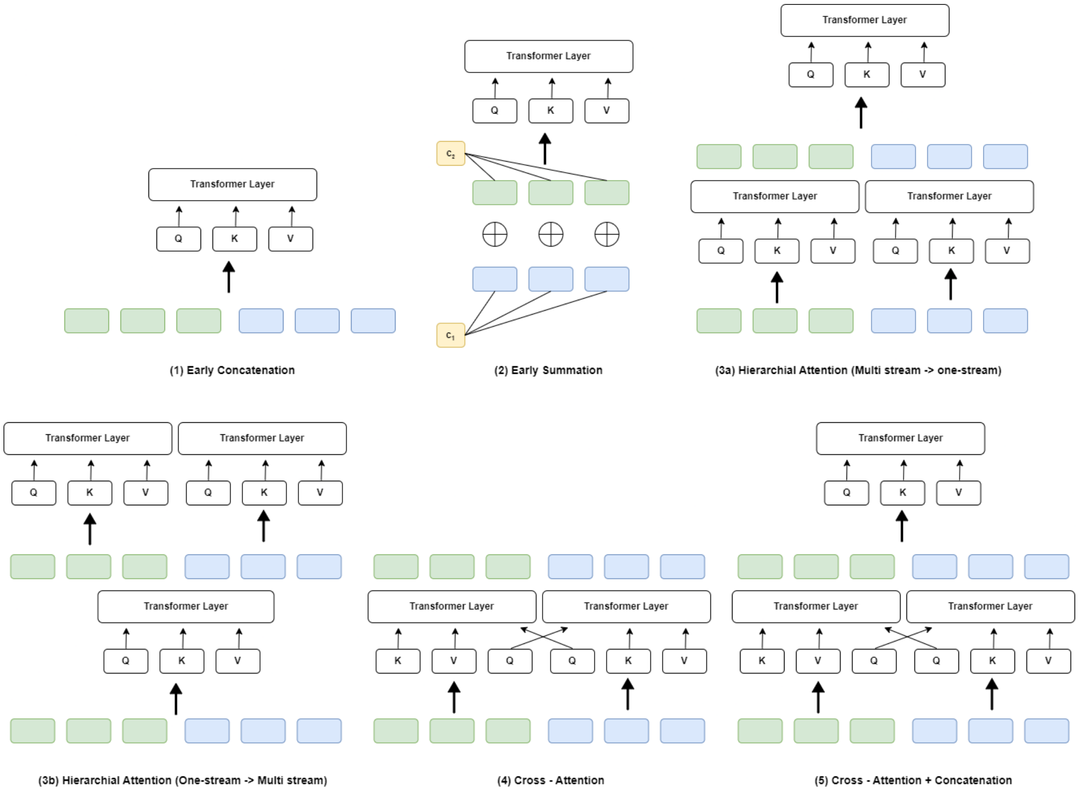
Let the inputs from any two modalities be represented as $X_{(1)}$ and $X_{(2)}$. After token embedding, the inputs become $T_{(1)}$ and $T_{(2)}$. The token sequence given by multimodal interactions become X^*. Let $M(.)$ stand for the module operations of Transformer blocks and or or layers. **(1) Early Concatenation** Also known as Co-Transformer, simply concatenates the tokens from multiple modalities and passes as inputs to Transformer Blocks:

$$\begin{aligned} T &\leftarrow concat(T_{(1)}, T_{(2)}) \\ X^* &= M(Q_T, K_T, V_T) \end{aligned} \tag{5}$$

This method was first employed in VideoBERT [73] for video and text. The positions of each modality, conditioning other modality's context, encodes the global context in multi-modal environment well. However, for very long sequences after concatenation the computational complexity is increased. Thus, it works well for few modalities of shorter sequences of data.

(2) Early Summation The global context is given by performing a weighted summation at each token position for all modalities and then passed to the Transformer Blocks:

$$\begin{aligned} T_{12} &\leftarrow (c_1T_{(1)} \oplus c_2T_{(2)}) \\ \implies Q_{12} &= (c_1T_{(1)} \oplus c_2T_{(2)})W^Q \\ K_{12} &= (c_1T_{(1)} \oplus c_2T_{(2)})W^K \\ V_{12} &= (c_1T_{(1)} \oplus c_2T_{(2)})W^V \\ X^* &= M(Q_{12}, K_{12}, V_{12}) \end{aligned} \tag{6}$$



(1) Early Concatenation

(2) Early Summation

(3a) Hierarchial Attention (Multi stream -> one-stream)

(3b) Hierarchial Attention (One-stream -> Multi stream)

(4) Cross - Attention

(5) Cross - Attention + Concatenation

Figure 2. Self-Attention variants

where M is usually a multihead self attention block and c_1, c_2 are arbitrary weights. This has been employed in works such as [28, 83].This method can be scaled well without increasing memory requirements.

(3) Heirarchial Attention

(3a) Multi-stream → Single-stream The inputs from different modals are encoded by individual Transformer streams and the outputs from each modal is concatenated by a separate transformer (a special case of early concatenation):

$$\begin{aligned} T \leftarrow concat(M_{(1)}(T_{(1)}), M_{(2)}(T_{(2)})) \\ X^* = M_{(3)}(T) \end{aligned} \tag{7}$$

(3b) Single-stream → Multi-stream A shared single stream Transformer is used to encode concatenated multimodal inputs. This step is followed by separate Transformer streams for each modality in order to preserve the distinct properties of a single modal representation.

$$\begin{aligned} X^*_{(1)}, X^*_{(2)} \leftarrow M_{(1)}(concat(T_{(1)}, T_{(2)})) \\ X^*_{(1)} \leftarrow M_{(2)}(T_{(1)}) \\ X^*_{(2)} \leftarrow M_{(3)}(T_{(2)}) \end{aligned} \tag{8}$$

(4) Cross-Attention Also known as co-attention, the cross modal interactions are induced or tend to by switching the Q (Query) tokens alone of every modal.

$$\begin{aligned} X^*_{(1)} \leftarrow MHSA(Q_{(T_2)}, K_{(T_{(2)})}, V_{(T_{(2)})}) \\ X^*_{(2)} \leftarrow MHSA(Q_{(T_2)}, K_{(T_{(1)})}, V_{(T_{(1)})}) \end{aligned} \tag{9}$$

This method was first employed in VilBERT [28] and does not tax the computational complexity. The main disadvantage is that the global context is lost since there is no SA to the individual contexts for each modality.

(5) Cross-Attention + Concatenation The global context missing in (4) is modelled by concatenating the output after cross attention.

$$\begin{aligned} X^*_{(1)} \leftarrow MHSA(Q_{(T_2)}, K_{(T_{(2)})}, V_{(T_{(2)})}) \\ X^*_{(2)} \leftarrow MHSA(Q_{(T_2)}, K_{(T_{(1)})}, V_{(T_{(1)})}) \\ X^* \leftarrow M(concat(X^*_{(1)}, X^*_{(2)})) \end{aligned} \tag{10}$$

Table 1 presents the architectures using the above mentioned self-attention variants and works which use them for reference.

4.2.2 Based on Network Architecture

To sum the previous section up, the internal multimodal attentions of different Transformers, which include the

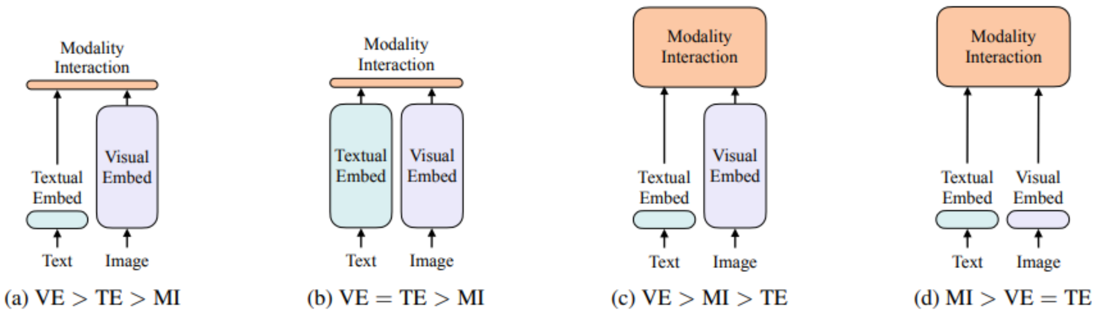


Figure 3. Four categories of vision-and-language models. The height of each rectangle denotes its relative computational size. VE, TE, and MI are short for visual embedder, textual embedder, and modality interaction, respectively [39]

aforementioned self-attention varieties, is what makes them functional. As shown in Figure 2, these considerations establish the external network architectures of the multimodal Transformers.

Based on Fusion strategy Based on when the multimodals fuse or attain global context, SA-variants can be divided into three types: *Before Self-Attention*: (1) early concatenation, (2) early summation, (3b) hierarchical Attention (Single-stream → Multi-stream); *Throughout Self-Attention*: (4) Cross-Attention, (5) Cross-Attention + Concatenation; *After Self-Attention*: (3) hierarchical Attention (Single-stream → Multi-stream).

Based on network structures It can be observed that SA-variants (1) early concatenation and (2) Early summation work in uni-stream, (3) Hierarchical Attention and (5) Cross-Attention + Concatenation both work in hybrid-stremas. and (4) Cross-Attention works in multiple-stream. Using the above classification, Transformers can thus be divided into uni-stream *eg.,* UNITER [14], Vl-BERT [72]; hybrid-streams *eg.,*InterBERT [48] and multi-stream *e.g.,* ViLBERT [53], ActBERT [102], *etc.*.

Based on computational complexity [39] divides vision-and-language models into four types as seen in Figure 3. Figure 3(a) includes visual semantic embedding (VSE) approaches such as VSE++ [25] and SCAN [40]. Image embedders are heavier than text embedders. The similarities between the two modalities are shown with dot products or shallow attention layers. CLIP [82] requires separate, costly transformer embedders for each modality (Figure 3(b)). Image vector and text vector interaction is shallow (dot product). Despite CLIP's zero-shot performance on image-to-text retrieval, the same level on other vision-and-language downstream tasks couldn't be seen. Recent VLP models employ a deep transformer to simulate image and text feature interaction. Convolutional networks extracting and embedding picture characteristics accounts for most of the computation. Modulation-based vision-and-language models (Film [61], MoVie [58]) also fit under Figure 3(c), with visual CNN stems corresponding to visual embedder, RNNs yielding modulation parameters to textual embedder, and modulated CNNs to textual embedder. ViLT [39] is the first Figure 3(d) model with shallow embedding layers of raw pixels computationally lighter than text tokens. This design focuses on simulating modality interactions.

5. Transformers for specific multimodal task

5.1. Discriminative tasks

Transformer models can encode a wide variety of multimodal inputs for use in both traditional and cutting-edge discriminative applications. This includes, but is not limited to, tasks tabulated in Table 2.

5.2. Generative Tasks

Transformers also contribute to a variety of multimodal generating tasks. These tasks include transforming single-modality into single-modality (see Table 3), multimodality into a single modality (see Table 4), and multimodality into multimodality (e.g. [47] which allows for unified pretraining on the data of single modality and multiple modalities and VATT [2]) .

5.3. Challenges

This Section presents an investigation of the preceding work from the standpoint of the technological difficulties. A concise description of the four difficulties associated with learning via Transformer, including Cross-modal interaction, Incomplete modality, transferability, and Efficiency. This significantly expands the taxonomy that was established in [3] in order to deal with the increased variety and bigger ranges of current Transformer-based MML studies that have been completed in recent years.

Modals	Task	Ref.
Test desc. + point cloud	Visual Grounding	[96]
acoustic + text	Bilingual Speech Translaton	[98]
audio + visual observation	Audio-Visual Navigation	[8]
Contextual Tag embeddings	Cross-modal Alignment of Audio and Tags	[26]
appearance + audio + speech	Video-retrieval	ConTra [27]
text query + image	Video-retrieval	AVSeeker [76]
	image-text retrieval	VL-BEiT [4], GilBERT [45]
	Document AI	LayoutLMv3 [35]
audio + video	Audio-Visual Video Parsing	
	Audio-Visual speech enhancement (AVSE)	Vset [66], [15]
	Audio-Visual speech recognition	AV-ASR [69], [1]
video + text	Referring Video Object Segmentation (RVOS)	[13], [5]

Table 2. Multimodal Discriminative Tasks

From	To	Ref.
image	3D human texture	[84]
single-image	geo-localition	TransLocator [63]
RGB	Scene Graph	Relationformer [74], [32, 54, 80]
Video	Caption	SwinBERT [49], [20], VX2TEXT [51]
Image	Caption	M^2 [17], AoA [34]
text	Speech	Dict-TTS [37], TransformerTTS [10], Grad-TTS [62], Glow-TTS [38]
Text	Image	DALLE-URBAN [68], CogView [22]
RGB	3D human pose	GTRS [97], [50]
music	dance	Transflower [78], DanceNet3D [41], DanceFormer [42]

Table 3. Generative tasks: single modality to single modality

From	To	Ref.
image + text	scene graph	[99]
image + Query	Answer	[36]
Aud + Visual + Scene	Dialog	DST [89], AVSD-T [30, 46, 60, 85],

Table 4. Generative tasks: multiple modality to single modality

5.4. Cross-modal interaction/ Fusion strategy

MML Transformers, in general, combine information from many modalities mainly at three typical levels [12]: the level of input (also known as early cross-modal interaction), the level of intermediate representation (often known as intermediate cross-modal interaction), and the level of prediction (i.e., late cross-modal interaction). When the representations of two modalities are fed directly into the standard attention module, it is possible to accomplish intermediate cross-modal interaction with latent adaptation, which ultimately results in late cross-modal interaction of the final bimodality representations [77]. This concept may be developed further via the use of alternating [60] or compounding [87] with unimodal attention, as well as token exchange [9] across modalities. Different modalities begin to integrate as early as the input stage, as shown in [14, 73]. This was inspired by the astounding success of BERT [21]. They are also referred to as one-stream architecture, and they make it possible to incorporate the benefits of BERT with just a small amount of architectural change. The use of problem-specific modalities in conjunction with varied masking strategies stands out as a significant contrast to these onestream approaches. A conspicuous cross-modal interaction strategy that is based on the concept of bottleneck tokens as part of the attention operation. Simply selecting the layers that are going to be fused makes it applicable for both early and intermediate cross-modal interaction. It has to to our attention that the straightforward prediction-based late cross-modal interaction [12, 59] is not used as often in MML Transformers. Taking into account the goals of developing better multimodal contextual representations and the tremendous development in computational power, this makes perfect sense. Investigating the ways in which modalities interact with one another would be a fascinating path to take if one were interested in improving and understanding the cross-modal interaction of MML.

5.5. Incomplete modal data

In [56], an important question- Are Transformer models robust against missing-modal data?; is posed. It should come as no surprise that that Transformer models do suffer a significant setback in the case of incomplete modal data. The robustness will be considerably impacted in a variety of ways depending on the fusion strategy. The optimal strategy for fusing data depends on the datasets being used; there is no one-size-fits-all approach that is effective in all circumstances where there is modal-incompete data present. [56] uses multi-task optimization to optimize Transformer models using both modal-complete and modal-incomplete data simultaneously. In addition to this, a searching technique to get the best possible fusion approach considering various datasets is provided. The authors [95] propose a novel multimodal Medical Transformer (mmFormer) for incomplete multimodal learning, which consists of three main components: the hybrid modality-specific encoders that connect a convolutional encoder and an intra-modal Transformer for both local and global context modeling within every modality; an inter-modal Transformer to construct and align the long-range correlations across modalities for modality-invariant features with global knowledge representation corresponding to tumor region; and a multimodal fusion layer.

5.6. Tranferability

How to transfer models between datasets and applications is a significant challenge for Transformer-based multimodal learning.

Multimodal Transformers increase generalization with data augmentation and adversarial disruption. Certain works use task-agnostic adversarial pretraining and task-specific adversarial fine-tuning to enhance VLP Transformers.

In practice, the training-practice data gap is obvious. In real applications, supervised data samples (well-labeled, well-aligned) are expensive, therefore transferring supervised multimodal Transformers trained on well-aligned cross-modal pairs/tuples to a poorly aligned test bed is hard [92]. CLIP [65] transmits information across modalities by learning a multimodal embedding space, allowing zero-shot model transfer to downstream tasks. CLIP's basic motivation is that pretrained multimodal (image and text) knowledge may be transferred to downstream zero-shot picture prediction by employing a prompt template ”A photo of a label.” to bridge training and test dataset distribution gaps.

Overfitting hinders transfer. Due of their vast modeling capabilities, Multimodal Transformers may overfit dataset biases during training. Some recent methods move oracle models from noiseless to actual data. LXMERT/BERT-like patterns may be transferred from an ideal dataset to a real dataset.

Cross-task gap is another key impediment to transfer, owing to differing reasoning and input-output processes, e.g., using multimodal datasets to fine-tune a language pretrained model is problematic. Due to missing modalities, multimodal pretrained Transformers must occasionally handle uni-modal input during inference. One approach is knowledge distillation, e.g., from multimodal to uni-modal Transformer attention, from numerous uni-modal Transformer instructors to a common Transformer encoder. Discriminative and generative multimodal tasks differ greatly. BERT-like encoder-only multimodal Transformers (e.g., VideoBERT [73]) require to train decoders independently for generating tasks. This might cause a pretrain-finetune discrepancy that hurts generalization. GilBERT [45] VL-BEiT [4] and are generative VLP models for image-text retrieval.

Cross-lingual gap should also be addressed for Transformer-based multimodal learning, e.g., universal cross-lingual generalization from one language to another language multimodal situations [98].

5.7. Efficiency

Multimodal transformers have two efficiency problems: (a) Large model parameter capacities need large training datasets. (b) Self-attention causes temporal and memory difficulties that increase quadratically with input sequence length. High-dimensional representations worsen multimodal computation explosion.

Recent approaches have tried to utilize less training data and/or parameters to increase multimodal Transformer training and/or inference efficiency. Below are few key points:

(1) **Model Compression**. To simplify pipelines, remove components. Two-stage pipelines are expensive since they need an object detector for VLP Transformer models. ViLT [39] handle visual input convolution-free. DropToken [176] decreases training complexity by randomly dropping video and audio tokens. DropToken used in [2] is a form

of dropout or adversarial training. Weight-sharing also simplifies multimodal Transformer models. Other works have used a low-rank parameter-sharing approach.

(2) **Knowledge distillation**. Smaller Transformers distill the knowledge from trained larger ones.

(3) **Improving sample utilization**. Use training samples to train models on fewer samples. In [45], CLIP is trained with less data by mining self-supervised signals from (a) each modality, (b) across modalities, and (c) similar modalities.

(4) **Network asymmetry**. To help to maintain parameters, assign different model capacity and computational size to distinct modalities [39].

(5) **Improving self-attention**. Transformers need quadruple the input sequence length in time and memory. [16] provide sparse factorizations of the attention matrix to minimize O(N2) complexity. Transformer-LS [101] has linear computational and memory complexity, reducing quadratic complexity to O(nn).Optimizing self-attention-based multimodal interaction/fusion. [57] suggest FSN to enhance early concatenation-based multimodal interaction. FSN transmits messages via a limited number of bottleneck latents, forcing the model to purify crossmodal data. This method increases fusion performance and decreases computing cost by using the bottleneck as a bridge. (7) Other strategy optimization. Optimize multimodal Transformer interactions. Given the quadratic complexity of self-attention, early concatenation-based multimodal interaction is expensive. [86] describe an effective, greedy, approach that successively blends information across adjacent views.

6. Results

The project gives a detailed overview of multi-modal transformers to provide a convenient reference for recent works in this domain up to current date updated till as recent as 2022. Also, summarizes key designs in the existing Multimodal Transformers and discusses current challenges of the same. This can be extensively used to identify future research scopes.

References

[1] Triantafyllos Afouras, Joon Son Chung, Andrew Senior, Oriol Vinyals, and Andrew Zisserman. Deep audio-visual speech recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):8717–8727, dec 2022. 7

[2] Hassan Akbari, Liangzhe Yuan, Rui Qian, Wei-Hong Chuang, Shih-Fu Chang, Yin Cui, and Boqing Gong. Vatt: Transformers for multimodal self-supervised learning from raw video, audio and text, 2021. 6, 8

[3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423, 2019. 1, 2, 6

[4] Hangbo Bao, Wenhui Wang, Li Dong, and Furu Wei. Vlbeit: Generative vision-language pretraining, 2022. 7, 8

[5] Adam Botach, Evgenii Zheltonozhskii, and Chaim Baskin. End-to-end referring video object segmentation with multimodal transformers, 2021. 7

[6] Necati Cihan Camgoz, Oscar Koller, Simon Hadfield, and Richard Bowden. Sign language transformers: Joint end-to-end sign language recognition and translation, 2020. 2

[7] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. *CoRR*, abs/2005.12872, 2020. 1

[8] Changan Chen, Ziad Al-Halah, and Kristen Grauman. Semantic audio-visual navigation, 2020. 7

[9] Chun-Fu Chen, Quanfu Fan, and Rameswar Panda. Crossvit: Cross-attention multi-scale vision transformer for image classification, 2021. 4, 7

[10] Mingjian Chen, Xu Tan, Yi Ren, Jin Xu, Hao Sun, Sheng Zhao, Tao Qin, and Tie-Yan Liu. Multispeech: Multi-speaker text to speech with transformer, 2020. 7

[11] Shizhe Chen, Pierre-Louis Guhur, Cordelia Schmid, and Ivan Laptev. History aware multimodal transformer for vision-and-language navigation, 2021. 2

[12] Tsuhan Chen and R.R. Rao. Audio-visual integration in multimodal communication. *Proceedings of the IEEE*, 86(5):837–852, 1998. 7, 8

[13] Weidong Chen, Dexiang Hong, Yuankai Qi, Zhenjun Han, Shuhui Wang, Laiyun Qing, Qingming Huang, and Guorong Li. Multi-attention network for compressed video referring object segmentation. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4416–4425, New York, NY, USA, 2022. Association for Computing Machinery. 7

[14] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning, 2019. 2, 6, 7

[15] I-Chun Chern, Kuo-Hsuan Hung, Yi-Ting Chen, Tassadaq Hussain, Mandar Gogate, Amir Hussain, Yu Tsao, and Jen-Cheng Hou. Audio-visual speech enhancement and separation by leveraging multi-modal self-supervised embeddings, 2022. 7

[16] Rewon Child, Scott Gray, Alec Radford, and Ilya Sutskever. Generating long sequences with sparse transformers, 2019. 9

[17] Marcella Cornia, Matteo Stefanini, Lorenzo Baraldi, and Rita Cucchiara. Meshed-memory transformer for image captioning, 2019. 7

[18] Yin Dai and Yifan Gao. Transmed: Transformers advance multi-modal medical image classification. *CoRR*, abs/2103.05940, 2021. 2

[19] Gourav Datta, Tyler Etchart, Vivek Yadav, Varsha Hedau, Pradeep Natarajan, and Shih-Fu Chang. Asd-transformer: Efficient active speaker detection using self and multimodal transformers. In *ICASSP 2022*, 2022. 2

[20] Chaorui Deng, Shizhe Chen, Da Chen, Yuan He, and Qi Wu. Sketch, ground, and refine: Top-down dense video

captioning. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 234–243, 2021. 7

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. 1, 7

[22] Ming Ding, Zhuoyi Yang, Wenyi Hong, Wendi Zheng, Chang Zhou, Da Yin, Junyang Lin, Xu Zou, Zhou Shao, Hongxia Yang, and Jie Tang. Cogview: Mastering text-to-image generation via transformers, 2021. 7

[23] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. 1

[24] Philipp Dufter, Martin Schmitt, and Hinrich Schütze. Position information in transformers: An overview, 2021. 3

[25] Fartash Faghri, David J. Fleet, Jamie Ryan Kiros, and Sanja Fidler. Vse++: Improving visual-semantic embeddings with hard negatives, 2017. 6

[26] Xavier Favory, Konstantinos Drossos, Tuomas Virtanen, and Xavier Serra. Learning contextual tag embeddings for cross-modal alignment of audio and tags, 2020. 7

[27] Adriano Fragomeni, Michael Wray, and Dima Damen. Contra: (con)text (tra)nsformer for cross-modal video retrieval, 2022. 7

[28] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition, 2020. 2, 5

[29] Kirill Gavrilyuk, Ryan Sanford, Mehrsan Javan, and Cees G. M. Snoek. Actor-transformers for group activity recognition, 2020. 4

[30] Shijie Geng, Peng Gao, Moitreya Chatterjee, Chiori Hori, Jonathan Le Roux, Yongfeng Zhang, Hongsheng Li, and Anoop Cherian. Dynamic graph representation learning for video dialog via multi-modal shuffled transformers, 2020. 7

[31] Daya Guo, Shuo Ren, Shuai Lu, Zhangyin Feng, Duyu Tang, Shujie Liu, Long Zhou, Nan Duan, Alexey Svyatkovskiy, Shengyu Fu, Michele Tufano, Shao Kun Deng, Colin Clement, Dawn Drain, Neel Sundaresan, Jian Yin, Daxin Jiang, and Ming Zhou. Graphcodebert: Pre-training code representations with data flow, 2020. 4

[32] Yuyu Guo, Lianli Gao, Xuanhan Wang, Yuxuan Hu, Xing Xu, Xu Lu, Heng Tao Shen, and Jingkuan Song. From general to specific: Informative scene graph generation via balance adjustment, 2021. 7

[33] Yicong Hong, Qi Wu, Yuankai Qi, Cristian Rodriguez-Opazo, and Stephen Gould. A recurrent vision-and-language bert for navigation, 2020. 2

[34] Lun Huang, Wenmin Wang, Jie Chen, and Xiao-Yong Wei. Attention on attention for image captioning, 2019. 7

[35] Yupan Huang, Tengchao Lv, Lei Cui, Yutong Lu, and Furu Wei. Layoutlmv3: Pre-training for document ai with unified text and image masking, 2022. 7

[36] Theo Jaunet, Corentin Kervadec, Romain Vuillemot, Grigory Antipov, Moez Baccouche, and Christian Wolf. Visqa: X-raying vision and language reasoning in transformers, 2021. 7

[37] Ziyue Jiang, Su Zhe, Zhou Zhao, Qian Yang, Yi Ren, Jinglin Liu, and Zhenhui Ye. Dict-tts: Learning to pronounce with prior dictionary knowledge for text-to-speech, 2022. 7

[38] Jaehyeon Kim, Sungwon Kim, Jungil Kong, and Sungroh Yoon. Glow-tts: A generative flow for text-to-speech via monotonic alignment search, 2020. 7

[39] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision, 2021. 6, 8, 9

[40] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. Stacked cross attention for image-text matching, 2018. 6

[41] Buyu Li, Yongchi Zhao, and Lu Sheng. Dancenet3d: Music based dance generation with parametric motion transformer. *ArXiv*, abs/2103.10206, 2021. 7

[42] Buyu Li, Yongchi Zhao, Zhelun Shi, and Lu Sheng. Danceformer: Music conditioned 3d dance generation with parametric motion transformer, 2021. 7

[43] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+language omni-representation pre-training, 2020. 2

[44] Ruilong Li, Shan Yang, David A. Ross, and Angjoo Kanazawa. Ai choreographer: Music conditioned 3d dance generation with aist++, 2021. 4

[45] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm, 2021. 7, 8, 9

[46] Zekang Li, Zongjia Li, Jinchao Zhang, Yang Feng, and Jie Zhou. Bridging text and video: A universal multimodal transformer for audio-visual scene-aware dialog. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:2476–2483, 2021. 7

[47] Junyang Lin, Rui Men, An Yang, Chang Zhou, Ming Ding, Yichang Zhang, Peng Wang, Ang Wang, Le Jiang, Xianyan Jia, Jie Zhang, Jianwei Zhang, Xu Zou, Zhikang Li, Xiaodong Deng, Jie Liu, Jinbao Xue, Huiling Zhou, Jianxin Ma, Jin Yu, Yong Li, Wei Lin, Jingren Zhou, Jie Tang, and Hongxia Yang. M6: A chinese multimodal pretrainer, 2021. 6

[48] Junyang Lin, An Yang, Yichang Zhang, Jie Liu, Jingren Zhou, and Hongxia Yang. Interbert: Vision-and-language interaction for multi-modal pretraining, 2020. 4, 6

[49] Kevin Lin, Linjie Li, Chung-Ching Lin, Faisal Ahmed, Zhe Gan, Zicheng Liu, Yumao Lu, and Lijuan Wang. Swinbert: End-to-end transformers with sparse attention for video captioning, 2021. 7

[50] Kevin Lin, Lijuan Wang, and Zicheng Liu. End-to-end human pose and mesh reconstruction with transformers, 2020. 7

[51] Xudong Lin, Gedas Bertasius, Jue Wang, Shih-Fu Chang, Devi Parikh, and Lorenzo Torresani. Vx2text: End-to-end

learning of video-based text generation from multimodal inputs, 2021. 7

[52] Haochen Liu, Zhiyu Huang, and Chen Lv. Strajnet: Multimodal hierarchical transformer for occupancy flow field prediction in autonomous driving, 2022. 2

[53] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks, 2019. 4, 6

[54] Yichao Lu, Himanshu Rai, Jason Chang, Boris Knyazev, Guangwei Yu, Shashank Shekhar, Graham W. Taylor, and Maksims Volkovs. Context-aware scene graph generation with seq2seq transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 15911–15921, 2021. 7

[55] Fengmao Lv, Xiang Chen, Yanyong Huang, Lixin Duan, and Guosheng Lin. Progressive modality reinforcement for human multimodal emotion recognition from unaligned multimodal sequences. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2554–2562, 2021. 2

[56] Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. Are multimodal transformers robust to missing modality? In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18156–18165, 2022. 8

[57] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. Attention bottlenecks for multimodal fusion, 2021. 9

[58] Duy-Kien Nguyen, Vedanuj Goswami, and Xinlei Chen. Movie: Revisiting modulated convolutions for visual counting and beyond, 2020. 6

[59] Andrew Owens and Alexei A. Efros. Audio-visual scene analysis with self-supervised multisensory features, 2018. 8

[60] Gao Peng, Zhengkai Jiang, Haoxuan You, Pan Lu, Steven Hoi, Xiaogang Wang, and Hongsheng Li. Dynamic fusion with intra- and inter- modality attention flow for visual question answering, 2018. 7

[61] Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer, 2017. 6

[62] Vadim Popov, Ivan Vovk, Vladimir Gogoryan, Tasnima Sadekova, and Mikhail Kudinov. Grad-tts: A diffusion probabilistic model for text-to-speech, 2021. 7

[63] Shraman Pramanick, Ewa M. Nowara, Joshua Gleason, Carlos D. Castillo, and Rama Chellappa. Where in the world is this image? transformer-based geo-localization in the wild, 2022. 7

[64] Yuankai Qi, Zizheng Pan, Yicong Hong, Ming-Hsuan Yang, Anton van den Hengel, and Qi Wu. The road to know-where: An object-and-room informed sequential bert for indoor vision-language navigation, 2021. 2

[65] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 2, 8

[66] Karthik Ramesh, Chao Xing, Wupeng Wang, Dong Wang, and Xiao Chen. Vset: A multimodal transformer for visual speech enhancement. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6658–6662, 2021. 7

[67] Ludan Ruan and Qin Jin. Survey: Transformer based video-language pre-training. *CoRR*, abs/2109.09920, 2021. 2

[68] Sachith Seneviratne, Damith Senanayake, Sanka Rasnayaka, Rajith Vidanaarachchi, and Jason Thompson. Dalle-urban: Capturing the urban design expertise of large text to image transformers, 2022. 7

[69] Dmitriy Serdyuk, Otavio Braga, and Olivier Siohan. Transformer-based video front-ends for audio-visual speech recognition for single and multi-person video, 2022. 7

[70] Bowen Shi, Wei-Ning Hsu, Kushal Lakhotia, and Abdelrahman Mohamed. Learning audio-visual speech representation by masked multimodal cluster prediction, 2022. 4

[71] Les Atlas Amil Khanzada Shuyun Tang, Xinying Hu and Mert Pilanci. Hierarchical multi-modal transformer for automatic detection of covid-19, 2022. 2

[72] Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. Vl-bert: Pre-training of generic visual-linguistic representations, 2019. 2, 6

[73] Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. *CoRR*, abs/1904.01766, 2019. 1, 2, 4, 7, 8

[74] Bastian Wittmann Johannes Paetzold Ivan Ezhov Hongwei Li Jiazhen Pan Sahand Sharifzadeh Georgios Kaissis Volker Tresp Bjoern Menze Suprosanna Shit, Rajat Koner. Relationformer: A unified framework for image-to-graph generation, 2022. 7

[75] Amir Tavanaei, Karim Bouyarmane, Iman Keivanloo, and Ismail Tutar. Mmt4: Multi modality to text transfer transformer. In *KDD 2022*, 2022. 2

[76] Tu-Khiem LeVan-Tu NinhMai-Khiem TranGraham Healy-Cathal GurrinMinh-Triet Tran. Avseeker: an active video retrieval engine at vbs2022, 2022. 7

[77] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J. Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences, 2019. 4, 7

[78] Guillermo Valle-Pé rez, Gustav Eje Henter, Jonas Beskow, Andre Holzapfel, Pierre-Yves Oudeyer, and Simon Alexanderson. Transflower. *ACM Transactions on Graphics*, 40(6):1–14, dec 2021. 7

[79] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017. 3

[80] W. Wang, R. Wang, and X. Chen. Topic scene graph generation by attention distillation from caption, 2021. 7

[81] Zhecan Wang, Noel Codella, Yen-Chun Chen, Luowei Zhou, Jianwei Yang, Xiyang Dai, Bin Xiao, Haoxuan You, Shih-Fu Chang, and Lu Yuan. Clip-td: Clip targeted distillation for vision-language tasks, 2022. 2

[82] Mengde Xu, Zheng Zhang, Fangyun Wei, Yutong Lin, Yue Cao, Han Hu, and Xiang Bai. A simple baseline for zero-shot semantic segmentation with pre-trained vision-language model, 2021. 2, 6

[83] Peng Xu and Xiatian Zhu. Deepchange: A large long-term person re-identification benchmark with clothes change, 2021. 2, 4, 5

[84] Xiangyu Xu and Chen Change Loy. 3d human texture estimation from a single image with transformers. 2021. 7

[85] Yoshihiro Yamazaki, Shota Orihashi, Ryo Masumura, Mihiro Uchida, and Akihiko Takashima. Audio visual scene-aware dialog generation with transformer-based video representations, 2022. 7

[86] Shen Yan, Xuehan Xiong, Anurag Arnab, Zhichao Lu, Mi Zhang, Chen Sun, and Cordelia Schmid. Multiview transformers for video recognition, 2022. 9

[87] Xu Yang, Hanwang Zhang, Guojun Qi, and Jianfei Cai. Causal attention for vision-language tasks, 2021. 7

[88] Kayo Yin and Jesse Read. Better sign language translation with STMC-transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5975–5989, Barcelona, Spain (Online), Dec. 2020. International Committee on Computational Linguistics. 2

[89] Zhou Yu, Zitian Jin, Jun Yu, Mingliang Xu, and Jianping Fan. Towards efficient and elastic visual question answering with doubly slimmable transformer, 2022. 7

[90] Heeseung Yun, Youngjae Yu, Wonsuk Yang, Kangil Lee, and Gunhee Kim. Pano-avqa: Grounded audio-visual question answering on 360° videos, 2021. 4

[91] Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. Merlot: Multimodal neural script knowledge models, 2021. 2

[92] Xunlin Zhan, Yangxin Wu, Xiao Dong, Yunchao Wei, Minlong Lu, Yichi Zhang, Hang Xu, and Xiaodan Liang. Product1m: Towards weakly supervised instance-level product retrieval via cross-modal pretraining, 2021. 4, 8

[93] Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493, 2020. 2

[94] Qian Zhang, Han Lu, Hasim Sak, Anshuman Tripathi, Erik McDermott, Stephen Koo, and Shankar Kumar. Transformer transducer: A streamable speech recognition model with transformer encoders and rnn-t loss, 2020. 1

[95] Yao Zhang, Nanjun He, Jiawei Yang, Yuexiang Li, Dong Wei, Yawen Huang, Yang Zhang, Zhiqiang He, and Yefeng Zheng. mmformer: Multimodal medical transformer for incomplete multimodal learning of brain tumor segmentation, 2022. 8

[96] Lichen Zhao, Daigang Cai, Lu Sheng, and Dong Xu. 3dvg-transformer: Relation modeling for visual grounding on point clouds. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 2908–2917, 2021. 7

[97] Ce Zheng, Matias Mendieta, Pu Wang, Aidong Lu, and Chen Chen. A lightweight graph transformer network for human mesh reconstruction from 2d human pose, 2021. 7

[98] Renjie Zheng, Junkun Chen, Mingbo Ma, and Liang Huang. Fused acoustic and text encoding for multimodal bilingual pretraining and speech translation, 2021. 7, 8

[99] Yiwu Zhong, Jing Shi, Jianwei Yang, Chenliang Xu, and Yin Li. Learning to generate scene graph from natural language supervision, 2021. 7

[100] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa, 2019. 2

[101] Chen Zhu, Wei Ping, Chaowei Xiao, Mohammad Shoeybi, Tom Goldstein, Anima Anandkumar, and Bryan Catanzaro. Long-short transformer: Efficient transformers for language and vision, 2021. 9

[102] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations, 2020. 2, 6