Video-and-Language (VidL) models and their cognitive relevance

Anne Zonneveld[1] Albert Gatt[2] Iacer Calixto[3]
[1]Amsterdam Brain and Cognition Center, University of Amsterdam
[2]Department of Information and Computing Sciences, Utrecht University
[3]Department of Medical Informatics, Amsterdam UMC, University of Amsterdam

Abstract

In this paper we give a narrative review of multi-modal video-language (VidL) models. We introduce the current landscape of VidL models and benchmarks, and draw inspiration from neuroscience and cognitive science to propose avenues for future research in VidL models in particular and artificial intelligence (AI) in general. We argue that iterative feedback loops between AI, neuroscience, and cognitive science are essential to spur progress across these disciplines. We motivate why we focus specifically on VidL models and their benchmarks as a promising type of model to bring improvements in AI and categorise current VidL efforts across multiple 'cognitive relevance axioms'. Finally, we provide suggestions on how to effectively incorporate this interdisciplinary viewpoint into research on VidL models in particular and AI in general. In doing so, we hope to create awareness of the potential of VidL models to narrow the gap between neuroscience, cognitive science, and AI.

1. Introduction

Human intelligence seamlessly combines *sub-symbolic* perceptual signals—which are multi-modal and span across, for instance, vision and audition—with *symbolic* human language. The quest for artificial intelligence (AI) can be said to begin with a focus on mimicking human intelligence or behaviour, e.g., with Turing stating that human behaviour is a must-use guide for developing AI [119].This cognitive (or behavioural) approach is central to AI and has informed the field since its beginning. However, neuroscience also had an important influence on early developments in AI, especially as inspiration in the creation of artificial neural networks (ANNs), which were originally informed by the architecture of the brain and by the properties of real neurons [79, 94]. More recently, we have seen a resurgence of interest in the intersection between AI models and neuroscience [144] and also by research explicitly comparing ANNs to predicting brain activity, for example in the context of object recognition [138, 99] and language processing [98]. Another pillar of AI research that is becoming increasingly more relevant has to do with AI systems and their applicability and performance; or, in other words, concerns with engineering efforts and progress in solving concrete problems in vision [69, 95] and language [96, 96].

AI models that jointly reason upon visual and linguistic inputs have attracted particular attention in recent years. Reasons include the availability of data to train these models and the reduced costs of high-performance computing infrastructure to train and deploy such models [109]. Models include image-language (IL) models, which learn representations that combine static images and text (e.g., CLIP [88]), and video-language (VidL) models, which receive videos and text as inputs (e.g., VideoBERT [1]). Although research on IL models has boomed over the last years and great progress has been made,[1] *many of the most interesting capabilities in an AI model we would associate with human intelligence require an explicit time dimension*, which is missing in IL models. Models that learn from videos are an interesting alternative since time is built-in in VidL models. We thus argue that *VidL models offer an ideal test-bed for training and evaluating AI models that mimick some of the most interesting facets of human intelligence*.

Our main contributions are: *(i)* We provide a narrative review and discuss the current landscape of existing VidL models and benchmarks. *(ii)* We take inspiration from neuroscience and cognitive science to review existing VidL benchmarks in terms of their *cognitive relevance*. *(iii)* Finally, we propose avenues for future research and provide suggestions on how to effectively incorporate this interdisciplinary viewpoint into research in VidL models and in AI.

In Sections 1.1 and 1.2 we further elaborate on why we think this is a fruitful and necessary approach. In Section 2 we introduce the current landscape of VidL models and benchmarks, and categorise both according to our proposed 'cognitive relevance axioms'. Finally, in Section 3 we discuss shortcomings in current VidL models and benchmarks and recommend promising research avenues for VidL and AI research.

[1]See [146] for a recent overview on IL models.

1.1. The importance of interdisciplinary AI research

The fields of neuroscience, cognitive science and AI have a rich and intertwined history [49, 83]. The synergy between these three different fields (often referred to as cognitive computational neuroscience) has yielded a large body of work comparing the behaviour and activity of biological and artificial systems (as referred to in [97]). Although this procedure of comparison between biological and artificial systems could effectively lead to progress, it should be done with care, as these different disciplines and their designed models can have very different goals [83]. The field of AI is mainly occupied with benchmarking and engineering goals, and adapting models for real-world applications. Neuroscience, on the other hand, focuses on models with neurally plausible algorithms and representations, viewing such models as hypotheses to be tested against neural data. Finally, models in cognitive science are as good as they can predict and explain human cognitive behaviour [58].

[83] proposes a three-dimensional rubric—including engineering, neural plausibility and human-like behaviour—to evaluate models in cognitive computational neuroscience research. Iterative feedback loops between these dimensions can lead to *generative*, *corroborative* or *corrective* contributions to model development, which are needed for the advancement of AI [83, 144]. For example, one could use successful behavioural models to predict brain activity, corroborating these models' neural plausibility. This could lead to the generation of new behavioural tasks to enable differentiation between similar neurally plausible models. A mismatch between human behaviour and that of such neurally plausible model could then lead to the correction of the original model.

1.2. Focusing on multimodal video-language models

There has been some work relating image-language (IL) models to human behaviour and neuroscience (e.g., [78, 123]) and video models (without language) to human behaviour and neuroscience (e.g., [61, 128]). However, to the best of our knowledge there is a gap in research relating VidL models to human behaviour and neuroscience (with the notable exception of [113]). However, we argue that VidL models are highly cognitively relevant, since they are multimodal and include an explicit temporal dimension. We elaborate further on these two points below.

VidL models are multimodal In line with embodied theories [8] which state that our human understanding of the conceptual world around us is formed by language and internal bodily states from multiple modalities (e.g., vision, somato-sensory system, olfaction), multi-modal models leverage multi-modal information, combining symbolic and sensorily grounded representations [90]. An important discussion regarding the limitation of AI systems has to do with the symbol grounding problem [48], whereby an AI system that has access only to symbolic representations (e.g., text) is said to be unable to learn word meanings; in order to do that, such AI system would need access to external knowledge, e.g., knowledge about how to *ground* these symbols in the real world [9]. Multi-modal models address this problem to some extent, as they are trained on different types of data and integrate information from different channels (e.g., text, images, videos, audio). These models *could* be grounded—at least according to [48]'s definition of 'grounding'—, implying that their representations *could* in principle encode "real word meanings" and be suitable for applications that need a joint understanding of inputs to solve complex problems.

Moreover, behavioural and eye movement research also show that there are interactions between linguistic and non-linguistic information, i.e., showing that language affects speakers' colour perception and discrimination [129] and that linguistic and visual input are jointly used in disambiguation and reference resolution [112]. Even for abstract words without any visual referents, including additional multi-modal information (e.g., sensorimotor or social information) can ground meaning [14, 126, 29].

VidL models include a temporal dimension VidL models are theoretically able to answer more cognitively relevant questions than IL models, since videos include the temporal dimension and allow for *understanding of spatial-temporal dynamics* and the *grounding of language in actions*. This enables the comprehension of events and their compositional structure and further supports the understanding of notions of agency, cause and effect, and object permanence, which are essential for human intelligence [87, 105]. The use of video is arguably a better approximation to the perceptual flow humans are exposed to during development: dynamically changing environments perceived from different viewpoints, under varying lighting conditions and occlusions, and subject to self- or externally induced motion and saccadic eye movement [77].

2. Landscape

Video-language models seek a nonlinear mapping between the video and language modalities, i.e effective alignment of vision and language. Textual data is represented as a sequence of words (thus including a temporal dimension due to its sequential nature) and video data is represented as a sequence of static frames (thus including temporal and spatial dimensions). Within the VidL model, word embeddings and video representations are fused using an early or late fusion strategy to obtain semantically rich and

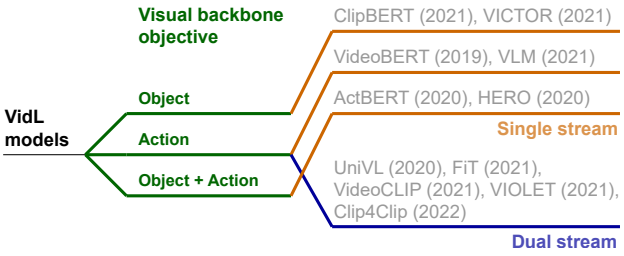

Figure 1. Schematic overview of selected VidL models organized along different dimensions (best viewed in colour).

grounded multimodal representations. Based on the success story of the pre-trained BERT model [32], VidL models often apply the *pretrain-then-finetune paradigm*, i.e., models are first pre-trained on basic fundamental tasks to learn general-purpose representations that can then subsequently be fine-tuned for specific downstream tasks. By using this paradigm, models do not have to be retrained from scratch every time for different tasks [93].

2.1. Model architecture and pretraining procedure

In Table 1, we categorise the growing body of VidL models along a few different dimensions. Many of the architectural insights are carried over from IL to VidL models, which is why we also discuss relevant aspects of IL models.

Single- vs. dual-stream architectures There are many different methods for fusing video representations and word embeddings, but overall one can make a distinction between single-stream and dual-stream architectures. In the case of single-stream architectures, video and language representations are jointly processed by one single cross-modal encoder, while in the case of dual-stream architectures video and language representations are encoded separately and then fused. In this way, dual-stream architectures can apply different processing regimes on different streams or modalities [124], which yields different multimodal representations compared to single-stream architectures [19]. In the context of IL models, some argue for superiority of one framework over the other [74, 22], while others state both perform equivalently when controlled for differences in training data and hyper parameters [18]. However, in the context of VidL models, not much research has been done yet on this topic.

Model family VidL models can be divided into convolutional neural network (CNN) based architectures and Transformer-based architectures. Early work on IL models focused mainly on CNN-based architectures, aggregating visual features extracted from object-trained 3D CNNs with static word embeddings [56, 17, 116] based on language models like Word2Vec [82] or GloVe [86]. A later wave of IL models relied on region proposal networks (e.g. LXMERT [111]), especially faster recurrent convolutional networks (RCNNs) [89]. More recent work focuses on Transformer-based architectures, as they do not have a locality bias like CNNs and can learn interactions between non-local contexts [118]. IL models like [57] use Transformers for the visual backbone and concatenate image patches and language tokens before processing them through a single Transformer stack [57]. In the context of VidL models, Transformer-based architectures can be used for temporal and/or spatial attention mechanisms for both video and language. In the case of single-stream VidL Transformer architectures, videos are first converted to 'visual tokens' which can be used as direct input for a single cross-modal transformer module that processes both language and video. This conversion is however based on video features extracted with CNN architectures, e.g. ResNet [50], ResNet3D [117] or S3D [133].

Visual backbone objective The visual backbone of VidL models can be based on either purely object-trained architectures or purely action-trained architectures or on a combination of both. For object-trained architectures, the most used datasets are ImageNet [31] and VisualGenome [59]. For action-trained architectures, the most used datasets are Kinetics [55] and Howto100M [80]. The idea behind using a combination of both object and action-trained architectures, is that doing so produces global-local multi-modal representations. The object-trained architecture provides local regional object features, while the action-trained architecture yields a clip-level action representation [150].

Pretraining tasks Different models are trained using different pre-training tasks, depending on their architecture and objectives. Overall, we note that some models focus on *reconstructing objectives* while others focus on *matching objectives*. Commonly used pre-training tasks focusing on reconstructing objectives include Masked Language Modelling (MLM)[32], Masked Frame Modelling (MFM)[67], Masked Visual-token Modelling (MVM)[1, 42], Masked Object Classification (MOC)[150], Masked Action Classification (MAC)[150], Frame Order Modelling (FOM) [67], Sentence Order Modelling (SOM)[62], Masked Modal Matching (MMM)[75, 135] and Language Reconstruction (LR)[75]. Pre-training tasks focusing on matching objectives include Cross-Modal Matching (CMM)[1, 150] and Video Subtitle Matching (VSM)[67]. We provide a detailed description of each pre-training task in Appendix A.

2.2. Downstream tasks and Benchmarks

After pre-training, VL models can be fine-tuned on one or more specific downstream benchmark tasks. In the this

Model	Stream	*Visual backbone (object)*	*Visual backbone (action)*	Pre-training tasks
VideoBERT[1]	single	N.A.	S3D[133]; Kinetics[55]	MLM, MVM, CMM
ActBERT[150]	single	Faster R-CNN[89]; Visual Genome[59]	ResNet-3D[117]; Kinetics[55]	MLM, MAC, MOC, CMM
HERO[67]	single	2D Resnet-50 [50]; ImageNet[31]	SlowFast[40]; Kinetics[55]	MLM, MFM, VSM, FOM
UniVL[75]	dual	N.A.	S3D[133]; Howto100M[81]	MLM, MFM, MMM, CMM, LR
VLM[135]	single	N.A.	S3D[133]; Howto100M[81]	MLM, MFM, MMM
VICTOR[62]	single	Inception-V4[110]; ImageNet[31]	N.A.	MLM, FOM, SOM, LR
ClipBERT[63]	single	2D Resnet-50 [50]; ImageNet[31]	N.A.	MLM, MOC, CMM
VIOLET [42]	dual	N.A.	Video Swin Transformer[73]; Kinetics[55]	MLM, MVM
FiT[7]	dual	N.A.	Space Time Transformer[10]; WebVid2M[7]	CMM
VideoCLIP[136]	dual	N.A.	S3D[133]; Howto100M[81]	CMM
CLIP4CLIP[76]	dual	N.A.	CLIP-ViT [88]; Howto100M[81]	CMM

Table 1. Overview of state-of-the-art Transformer-based VidL models. In the visual backbone object and action objectives, we first show the architecture used followed by the training dataset separated by a semi-colon ';'.

section we will give a general overview of the different types of benchmark tasks that are currently in use.

Text-based video retrieval tasks can generally be divided into two types: video retrieval (VR) [127] and video corpus moment retrieval (VCMR)[37]. VR requires the model to find the most relevant video clip from the available data set based on a natural language query and rank all candidate videos. This tests global alignment of video and language information. VCMR goes one step further, requiring the model to find the most relevant video clip and the most relevant moment within this video clip based on the query, testing for both global and local alignment. The goal of explicitly testing on moment retrieval is to force models to actively leverage temporal information to solve the task. Video-subtitle corpus moment retrieval (VSCMR) [65] extends this task to a multi-channel setup where both video and textual information need to be considered.

Video captioning requires the model to generate descriptions that portray the essence of the most interesting content of the visual scene. To do so, the model has to capture specific image regions [54] and understand temporal relationships [141]. In different variations of the task the model has to produce either a single sentence describing the whole video or a short paragraph [141]. This is motivated by the idea that most videos, especially longer videos, contain more than one event and are semantically rich. A single sentence yields a highly compressed summary. The risk here is that the sentence is uninformative. On the other hand, allowing multiple sentences may yield irrelevant and redundant descriptions [51]. Thus, ideally the model is able to comprehend what content is most relevant to perform the right amount of abstraction.

Video question answering (VidQA) tests the model's ability to answer open-ended or multiple choice questions about video content. VidQA provides fine-grained visual understanding [142] and in a typical automatic evaluation scenario one compares the model's generated output against one or more ground truth answers [4]. Answers are grounded in both space and time and thus require models to localize relevant moments and detect referred entities, i.e. spatio-temporal reasoning [64]. We distinguish three subtypes of VidQA: factoid, compositional, and social VidQA.

Factoid VidQA. This is the most basic version of VidQA and involves questions that target the existence of objects, object relationships or action recognition, i.e. 'what', 'who', and 'where' questions. Models perform relatively well for these types of questions, as they do not require exhaustive reasoning or inference about physical events [47]. More focus on 'why' and 'how' questions could shift research towards the core of human intelligence [147], as this type of reasoning goes beyond purely recognising the available content.

Compositional VidQA. This tests the compositional spatio-temporal reasoning abilities of a model. Compositionality is a core semantic property of natural language, usually characterised as "the meaning of the whole is a function of the meaning of its parts" [25]. A visual event can be decomposed into separate elements, e.g. objects, actions and their relationships, that together give meaning to the scene [11, 121]. Human infants implement decomposition from a young age and use ideas of object permanence, solidity and continuity [87] to perform inference about past and future, i.e. compositional reasoning allows for for generalisation to new domains and logical rules [60]. This type of reasoning is at the core of human development [105]. Compositional VidQA tasks can include questions about relationships between actions and objects and about the order in which actions take place [47].

Social VidQA. This task targets social intelligence. Social interaction is a fundamental 'arena of language use' [24] and has been argued, in the context of AI models, to be the highest possible level of grounding for language models [12]. This type of tasks address this goal by posing questions about social situations and mental states, e.g., Q: "Are they getting along?" A: "No, nobody seems to be smiling".

Audiovisual scene-aware dialogue (AVSAD) focuses on dialogue instead of one-shot question answering as done

with VidQA, since human communications entails sequential dialogue, often using referents [2]. AVSAD tasks take inspiration from the visual dialogue task [27]—which requires a model to hold a dialogue about a single static image and its description—but instead focuses on video and audio, and answers need to be grounded in video and in the preceding dialogue containing co-references.

Video language inference (VLI) is introduced by [71] and can be seen as a extension of VidQA, inspired by natural language inference benchmarks [15]. VLI takes a video clip with aligned subtitles as premise and a textual description as hypothesis. The model is then required to judge the relationship between the premise and hypothesis as an entailment, contradiction or neutral. Performing these inferences requires various reasoning skills, from more surface-level grounding (i.e. identifying objects), to more in-depth commonsense reasoning (e.g. about human intentions or causal relationships).

Future event prediction (FEP) is introduced by [66] and is an alternative inference benchmark that tests a model's ability to describe logical events in the future (hypothesis) based on a video clip with aligned subtitles (premise). This requires thorough understanding of video and language dynamics, but also implies multimodal commonsense knowledge. FEP focuses on predicting high-level future events, contrary to other existing video prediction tasks that mostly focus on low-level vision or semantics, i.e. predicting future frames [122, 68] or action labels [36, 101].

Action segmentation (AS) requires a model to parse a video into different actions at the frame level, i.e. the model has to assign each frame to a specified action label [114]. For example, the activity of 'cooking an egg' could consist of the following action steps: 'take out a pan', 'fill the pan with water', 'boil water', 'put the egg in the water' etc. The model is required to leverage sequential information to be able to determine such action boundaries. Action segmentation helps understanding of what actions are being performed, how far they have progressed and how actions will evolve in the future [34].

Action localisation (AL) requires the model to predict a predefined action label for a video, and localise the start and end point of the action [114]. Recognition of human actions shows understanding of dynamics and complex human intention reasoning [151].

Multimodal sentiment analysis (MSA) requires the model to integrate verbal and non-verbal information for the detection of sentiment. Including non-verbal information is essential, as this can change the perception of expressed sentiment. An example is how the statement "This movie is sick" can either be perceived positively when combined with a smile or negatively when combined with a frown [143]. Thus, an important element to identify the true meaning of the communicated language is the recognition of human emotion and intentions [104], i.e., theory of mind.

2.3. Cognitive relevance axioms

Multiple datasets are available for finetuning VidL models on different tasks. As many of these tasks are designed with very practical engineering applications in mind, we ask to what extent they are also cognitively relevant. In Table 2 we highlight these cognitive relevance dimensions for each VidL task and fine-tuning dataset.

Level of grounding This is the most important axiom to assess cognitive relevance. [12] defines levels of grounding or 'World Scopes' (WSs) that encompass a spectrum ranging from least to most grounded. Ideally, models should go beyond just text to consider the contextual and social foundations of language to learn word meanings. Usage-based theories of language additionally state that functionality is the source of meaning [130], i.e. that language acquires its meaning through its relation to the physical world and the social interactions it enables. Thus, although models must be able to recognise objects, people and activities to comprehend the language describing them, to better mimic humans models must go further and understand notions of causality, commonsense physics and social interactions.

[12] proposes the following hierarchy: *Corpus (WS1)*: Models learn only from text, typically consisting of small(er) curated datasets. *Internet (WS2)*: Models still learn only from text but now at Internet-scale, which can lead to unforeseen emerging capabilities [16]. *Perception (WS3)*: Models learn from multiple modalities, e.g., visual, auditory and somatosensory data. Multi-modality yields physical heuristics that we use in methaphors and abstract concepts. For example, we are able to understand the meaning of the word 'soft' as we ourselves have physically experienced this tactile sensation. *Embodiment and action (WS4)*: Multimodal sensory experiences are fundamental for comprehension of action-oriented categories [115]. Language is additionally grounded in these action-oriented categories to facilitate communication [103]. Understanding of actions and their relationships to our environment, again, translates to understanding of metaphors like 'a distant concern' (grounded in the idea that far-away things have little effect on local space). *Social interactions (WS5)*: These are the foundational use case of language [12, 24]. If the physical world grounds language, human intentions give language a purpose. Thus, to properly comprehend

Task	Sub-task	Dataset	MC?	Video Content	Text types	Query source	ML?	Grounding
Retrieval	VR	YC2R[148]	✓	Instructional cooking	Subs + Description	Crowd		WS4
		VATEX-EN-R[125]		Human activities	Description	Crowd	✓	WS3
		MSRVTT-R[137]		Various	Description	Crowd		WS3
		ActivityNet Captions[59]		Human activities	Description	Crowd		WS4
		MPII-MD[92]		Movies	Description	Professional		WS3
	VCMR	DiDeMo[3]		Flickr	Description	Crowd		WS4
		Charades-STA[44]		Scripted	Description	Semi-automatic		WS4
	VSCMR	TVR[65]	✓	TV shows	Subs+ Description	Crowd		WS4
		How2R[67]	✓	Instructional	Subs+ Description	Crowd	✓	WS4
Captioning	Sentence	TVC[65]	✓	TV shows	Subs + description	Crowd		WS4
		YC2C[148]	✓	Instructional cooking	Subs + Description	Crowd		WS4
		VATEX-EN-C[125]		Human activities	Description	Crowd	✓	WS3
	Paragraph	Youtube-Clips[20]		Various	Description	Crowd		WS3
		TACoS-MultiLevel[91]		Instructional cooking	Description	Crowd		WS3
Q&A	Factoid	Video-QA[145]		Various	QA	Semi-automatic		WS3
		How2-QA[67]	✓	Instructional	Subs + QA	Crowd	✓	WS3
		MSRVTT-QA[134]		Various	QA	Crowd		WS3
		ActivityNet-QA[142]		Various	QA	Crowd		WS4
	Compositional	TVQA[64]	✓	TV shows	Subs + QA	Crowd		WS4
		AGQA[47]		Charades (scripted)	QA	Automatic		WS4
		CLEVRER[139]		Synthetic	QA	Automatic		WS4
	Social	Social-IQ[143]	✓	Youtube	Subs + QA	Crowd		WS5
Dialogue	—	AVSAD[2]	✓	Charades (scripted)	Dialogue	Crowd		WS5
Inference	VLI	VIOLIN[71]	✓	TV shows + Youtube	Subs+ Description	Crowd		WS5
	FEP	VLEP[66]	✓	TV shows + Youtube	Subs	Crowd		WS4

Table 2. Different types and subtypes of VidL benchmarks and their datasets. We categorise each entry according to a number of dimensions indicating their cognitive relevance. **MC**: Multi-channel. **ML**: Multilingual. **Subs**: Subtitles. **QA**: Question and Answer.

language, one must have some form of social intelligence or rather *theory of mind*, i.e., the ability to consider mental states of others, including emotions, beliefs, and intentions [5]. According to usage-based theories of language, understanding of others' mental states is necessary to infer the communicative content others try to convey [107].

Video content Different datasets contain different types of video, which may affect temporal dependencies and generalisability. We can define four axes with regards to video content: (i) diversity, (ii) instructions, (iii) naturalness, and (iv) socialness. *(i)* Some datasets contain large amounts of clips and many different topics (e.g., entertainment, sports, people, science), while other datasets are more restricted and focus on specific types of video content (e.g., only cooking videos). The more diverse a dataset is, the more different viewpoints, activities and people it includes and thus the more general it is. *(ii)* Narrated instructional videos have a very strong temporal dependency between visual content and subtitles, i.e., the natural language actually refers to the visual elements of the scene at each moment. This is not necessarily true in non-instructional videos, e.g., dialogue subtitles may refer to non-physical events or concepts that are not depicted in the current scene. *(iii)* Some datasets contain natural videos, while a few contain synthetically composed videos. The latter allow for a fully controlled environment, though it is questionable whether their results generalise to real world situations. *(iv)* The extent to which datasets include socialness varies. For example, instructional videos do not typically include social cues in interactive settings. Films or television series, which contain many characters, provide richer social interactions.

Multi- or single-channel data Most often, VidL datasets have videos and dialogues in natural language (subtitles). Some datasets only include the visual information of videos, i.e. single-channel data, while others also include audio, subtitles, or both. We consider multichannel data as more cognitively relevant, as humans also process many different forms of sensory input (e.g., vision, somatosensation, audition) during perception.

Text type Generally, three different text types can be distinguished: subtitles, descriptions and question-answer (Q&A) pairs. Multichannel data is accompanied with subtitles (and optionally other text types), while single-channel data is accompanied by either descriptions or Q&A pairs. Q&A pairs are not used as input data, but mostly as queries.

Query source The vast majority of available datasets use queries collected via crowd-sourcing and often include human validation. However, as human annotation is costly and labour-intensive, some datasets have automatically or semi-automatically generated queries, e.g., leveraging logic

templates in combination with video descriptions [139, 21] or scene-graphs [47]. This allows for more granular control of the composition of the questions and the distribution of included concepts. However, overall human annotations are more natural and ensure higher quality.

Multilingual textual data In cognitive science there is an overreliance on English, which has yielded biased results about human cognition and language [13]. Similarly, in machine learning there has also been an overreliance on 'Western' visual and linguistic input data, yielding geographical biases in, e.g., object recognition [30, 100] and language models [38, 39]. These biases also translate to VidL models: performance decreases for 'non-Western' images [140] and multilingual data sets [70]. Thus, to ensure VidL model generalisation, we need multicultural and multilingual data.

3. Discussion

3.1. Current shortcomings

When evaluating the results of Table 2 using the proposed cognitive relevance axioms, it is clear that most downstream benchmarks are grounded in WS3 or WS4, i.e. perceptions or actions. Thus, benchmarks regarding WS5, sociality, are missing and not widely available. As stressed before, understanding of social situations is fundamental for understanding natural language. Therefore, we encourage the development of more datasets that are grounded in sociality, i.e., datasets that contain visual content and queries regarding social situations. Furthermore, datasets that consist of instructional videos have high temporal dependencies but their topics are often restricted (e.g., only cooking videos), which limits their generalisability. Lastly, most datasets are leveraged for single-utterance scenarios, while human communication entails sequential dialogues [2]. Thus, we believe future benchmarks should include multiple question-answering or turn-taking scenarios, e.g., similarly to the Visual Dialogue task [27] but using videos instead of static images.

3.2. Cognitively relevant vs. cognitively inspired

Overall, Table 2 shows that some benchmarks are more cognitively relevant than others. However, instead of using benchmarks which are retrospectively argued to be cognitively relevant, one should design benchmarks that explicitly target specific cognitive properties, i.e., *cognitively inspired benchmarks*. For example, the AGENT benchmark [102] was designed to test core intuitive psychological reasoning based on developmental psychology. Intuitive psychology is the ability to reason about other agents' mental states, solely based on their observed actions. This ability is already available to infants from a young age. Based on cognitive tests to probe the understanding of intuitive psychology in infants, the AGENT benchmark focuses on four categories: attribution of goal preferences [132], action efficiency [46], cost-reward trade-off [72], and unobserved constraints [26], and it leverages the Violation of Expectation (VoE) paradigm from developmental psychology. Additionally, ADEPT [103] and OFPR-Net [28] are benchmarks that use the same paradigm to test the principles of understanding of intuitive physics, i.e. object permanence, solidity and continuity, as highlighted by classic developmental studies on physical understanding [106].

Currently, these cognitively inspired benchmarks do not include any form of textual information, but this could potentially be added in the form of textual frame descriptions or textual queries. For example, one could prompt the model to answer a multiple-choice question on the expected trajectory of an occluded object. These cognitively inspired tasks are fundamentally very different from the cognitively relevant VidL benchmarks reviewed above, as they explicitly test for specific cognitive properties of interest, instead of testing on a more general application level. This is a necessary step to bridge the gap between cognitive science, AI and neuroscience, as has already been suggested, for example, in the context of linguistic phenomena [85].

3.3. Fair human-machine comparisons

Using cognitively relevant or inspired benchmarks, one also needs proper evaluation methods not only to assess model performance but also whether a model is able to reach 'human-level' performance. Historically, most machine learning research has been focused on measuring successes, i.e., what makes systems perform equally well as humans. However, it might be more informative to focus on a system's failures, or how differently it behaves compared to humans—e.g., what kinds of errors systems and humans make, and what kinds of corner cases they fail in—as accuracy scores alone cannot distinguish between different behavioural strategies [45]. One suggestion is to use the method of error consistency [45], i.e., a quantitative measurement of whether two black-box systems (e.g., DNNs and the brain) systematically make similar errors on the same input.

It is also possible that two systems may use similar behaviour strategies but show different behavioural output [41]. Not all behavioural output differences reflect systematic differences (consider, for example, two persons who respond differently to the same stimulus). How, then, should one properly evaluate human-machine differences? To clarify this problem, one should make the distinction between *competence* and *performance*, a distinction originally drawn by Chomsky [23] to characterise the difference between a native speaker's knowledge of language and their actual use of language in specific situations. In terms of the current argument, competence is a characterisation of a sys-

tem's internal states. Performance characterises how a system behaves under specific conditions. Intelligent systems might often actually know more than their behaviour shows due to performance constraints. For example, biological and artificial systems retrieve sensory input differently, use different hardware and have different modes of behavioural response. Thus, we should evaluate human-machine differences in light of these performance constraints, using the following principles [41] (as explained with examples from the field of computer vision):

(i) Placing human-like constraints on machines To investigate whether human constraints can account for human-machine differences, one could actively implement these constraints in an artificial system and see whether differences in behaviour decrease or disappear. An example comes from the study of adversarial images, i.e., images with very small perturbations that are invisible to the human eye but that are misclassified by artificial systems, which are often used as an argument for different information processing between humans and machines [84, 6]. When implementing human-like constraints, such as a human-like fovea, general robustness against perturbations improves [33] and adversarial examples generated based on this constrained architecture actually also seem to mislead humans [35].

(ii) Placing machine-like constraints on humans Machines also have constraints that humans do not have, e.g., they have a very limited and specific vocabulary often based on labelled datasets. This may have as result that even if machines hypothetically could 'see' images similarly to humans (competence), they would not be able to produce human-like classifications (performance). This constraint is clear in, e.g., a study on noisy images, for which humans find no objects, but which DNNs classify as an object [84] since DNNs must always predict a label. However, when humans are also forced to choose a label from the same set of predefined labels, the likelihood that the human agrees with the DNN is above chance [149]. On a similar note, algorithms often produce a rank-order list of their predictions[2] yielding richer responses and performance measurements than a single label, e.g., top-5 vs. top-1 accuracy. Thus, an alternative could be to ask humans to provide a top-k ranking over possible labels to see how this matches with the machine predicted top-k labels.

(iii) Species-specific task alignment It is not always possible or desirable to equalise constraints among humans and machines; sometimes, it may be desirable to accommodate unique constraints, even when this yields different tasks for different systems. For example, in comparative reward-learning studies where different agents perform the same task the type of reward is adjusted to the agent, e.g., human adults could receive money while rodents could receive sugar water. This principle of 'species-specific task alignment' can also be applied between humans and machines. For instance, in a study on 'atomic vision' [108, 43] both humans and machines had to classify an image based on a minimally cropped patch that was selected using human psychological experiments, yielding different results for humans and machines [120]. The original conclusion was that humans use 'atomic vision' to classify images while machines do not. However, when using minimally cropped patches based on 'machine psychophysics', machine performance showed a similar discrete flooring as human performance on minimally cropped patches selected using human psychophysics. Thus, in this task-aligned setting, both humans and machines actually showed 'atomic recognition'.

Although it is not necessary or even possible to apply all three principles at the same time, these principles do provide a useful framework to help researchers think in terms of testing constraints. Consideration of these principles facilitates making fair human-machine comparisons, which is essential to enable iterative feedback loops between neuroscience, cognitive science and AI.

As formulated above, these three principles have a *behavioural* emphasis, in that they focus on the design of tasks and analysis of performance. A more ambitious principle also suggests itself:

(iv) Neurologically-inspired design Can models which ground linguistic meaning in temporal and visual data benefit from studies of the neural substrates of these abilities in humans? This would place VidL models on a similar footing as, say, models of visual attention [53, 52, 131], but would also come closer to addressing recent calls to extend the classic Turing test in AI to fully embodied systems [144]. Part of this enterprise would involve revisiting current VidL architectures (see Figure 1) in terms of their anatomical plausibility, and leveraging ablation methods to study the role of different model components in a new, neurologically-inspired light. While neurological plausibility may not be a goal of all VidL models, such neuro-inspired design would also provide the cognitive neurosciences with mathematically solid tools to test hypothesis about the neural substrates of grounded cognition.

Using these frameworks for fair human-machine comparisons in combination with cognitively relevant VidL benchmark tasks, we hope to narrow the gap between AI and cognitive science research—and potentially also with neuroscience—and to make meaningful progress across these interdisciplinary fields.

[2]This includes any neural network trained to minimise categorical cross-entropy and that uses softmax as the activation function.

References

[1] Chen , Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, October 2019. 1, 3, 4, 15

[2] Huda Alamri, Vincent Cartillier, Abhishek Das, Jue Wang, Anoop Cherian, Irfan Essa, Dhruv Batra, Tim K Marks, Chiori Hori, Peter Anderson, et al. Audio visual scene-aware dialog. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7558–7567, 2019. 5, 6, 7

[3] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan Russell. Localizing moments in video with natural language. In *Proceedings of the IEEE international conference on computer vision*, pages 5803–5812, 2017. 6

[4] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015. 4

[5] Ian A Apperly and Stephen A Butterfill. Do humans have two systems to track beliefs and belief-like states? *Psychological review*, 116(4):953, 2009. 6

[6] Anish Athalye, Logan Engstrom, Andrew Ilyas, and Kevin Kwok. Synthesizing robust adversarial examples. In *International conference on machine learning*, pages 284–293. PMLR, 2018. 8

[7] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 1728–1738, 2021. 4

[8] Lawrence W Barsalou. Grounded cognition. *Annu. Rev. Psychol.*, 59:617–645, 2008. 2

[9] Emily M Bender and Alexander Koller. Climbing towards NLU : On Meaning , Form , and Understanding in the Age of Data. In *Proceedings ofthe 58th Annual Meeting ofthe Association for Computational Linguistics (ACL'20)*, pages 5185–5198, 2020. 2

[10] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, volume 2, page 4, 2021. 4

[11] Irving Biederman. Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2):115, 1987. 4

[12] Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online, Nov. 2020. Association for Computational Linguistics. 4, 5

[13] Damián E Blasi, Joseph Henrich, Evangelia Adamou, David Kemmerer, and Asifa Majid. Over-reliance on english hinders cognitive science. *Trends in cognitive sciences*, 26(12):1153–1170, 2022. 7

[14] Anna M Borghi, Ferdinand Binkofski, Cristiano Castelfranchi, Felice Cimatti, Claudia Scorolli, and Luca Tummolini. The challenge of abstract concepts. *Psychological Bulletin*, 143(3):263, 2017. 2

[15] Samuel R Bowman, Gabor Angeli, Christopher Potts, and Christopher D Manning. A large annotated corpus for learning natural language inference. *arXiv preprint arXiv:1508.05326*, 2015. 5

[16] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 5

[17] Elia Bruni, Nam-Khanh Tran, and Marco Baroni. Multimodal distributional semantics. *Journal of artificial intelligence research*, 49:1–47, 2014. 3

[18] Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language berts. *Transactions of the Association for Computational Linguistics*, 9:978–994, 2021. 3

[19] Jize Cao, Zhe Gan, Yu Cheng, Licheng Yu, Yen-Chun Chen, and Jingjing Liu. Behind the scene: Revealing the secrets of pre-trained vision-and-language models. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pages 565–580. Springer, 2020. 3

[20] David Chen and William B Dolan. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200, 2011. 6

[21] Hongshen Chen, Xiaorui Liu, Dawei Yin, and Jiliang Tang. A survey on dialogue systems: Recent advances and new frontiers. *Acm Sigkdd Explorations Newsletter*, 19(2):25–35, 2017. 7

[22] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer, 2020. 3

[23] Noam Chomsky. *Aspects of the theory of syntax*. MIT Press, Cambridge, MA, 1965. 7

[24] Herbert H Clark. *Arenas of language use*. University of Chicago Press, 1992. 4, 5

[25] Maxwell John Cresswell. *Logics and languages*. Routledge, 2016. 4

[26] Gergely Csibra, Szilvia Bíró, Orsolya Koós, and György Gergely. One-year-old infants use teleological representations of actions productively. *Cognitive science*, 27(1):111–133, 2003. 7

[27] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José MF Moura, Devi Parikh, and Dhruv

Batra. Visual dialog. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 326–335, 2017. 5, 7

[28] Arijit Dasgupta, Jiafei Duan, Marcelo H Ang Jr, Yi Lin, Su-hua Wang, Renée Baillargeon, and Cheston Tan. A benchmark for modeling violation-of-expectation in physical reasoning across event categories. *arXiv preprint arXiv:2111.08826*, 2021. 7

[29] Simon De Deyne, Danielle J Navarro, Guillem Collell, and Andrew Perfors. Visual and affective multimodal models of word meaning in language and mind. *Cognitive Science*, 45(1):e12922, 2021. 2

[30] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59, 2019. 7

[31] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 4

[32] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018. 3, 15

[33] Arturo Deza and Talia Konkle. Emergent properties of foveated perceptual systems. *arXiv preprint arXiv:2006.07991*, 2020. 8

[34] Guodong Ding, Fadime Sener, and Angela Yao. Temporal action segmentation: An analysis of modern technique. *arXiv preprint arXiv:2210.10352*, 2022. 5

[35] Gamaleldin Elsayed, Shreya Shankar, Brian Cheung, Nicolas Papernot, Alexey Kurakin, Ian Goodfellow, and Jascha Sohl-Dickstein. Adversarial examples that fool both computer vision and time-limited humans. *Advances in neural information processing systems*, 31, 2018. 8

[36] Dave Epstein, Boyuan Chen, and Carl Vondrick. Oops! predicting unintentional action in video. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 919–929, 2020. 5

[37] Victor Escorcia, Mattia Soldan, Josef Sivic, Bernard Ghanem, and Bryan Russell. Finding moments in video collections using natural language. *arXiv preprint arXiv:1907.12763*, 2019. 4

[38] Fahim Faisal and Antonios Anastasopoulos. Geographic and geopolitical biases of language models. *arXiv preprint arXiv:2212.10408*, 2022. 7

[39] Fahim Faisal, Yinkai Wang, and Antonios Anastasopoulos. Dataset geography: Mapping language data to language users. *arXiv preprint arXiv:2112.03497*, 2021. 7

[40] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019. 4

[41] Chaz Firestone. Performance vs. competence in human–machine comparisons. *Proceedings of the National Academy of Sciences*, 117(43):26562–26571, 2020. 7, 8

[42] Tsu-Jui Fu, Linjie Li, Zhe Gan, Kevin Lin, William Yang Wang, Lijuan Wang, and Zicheng Liu. Violet: End-to-end video-language transformers with masked visual-token modeling. *arXiv preprint arXiv:2111.12681*, 2021. 3, 4, 15

[43] Christina M Funke, Judy Borowski, Karolina Stosio, Wieland Brendel, Thomas SA Wallis, and Matthias Bethge. Five points to check when comparing visual perception in humans and machines. *Journal of Vision*, 21(3):16–16, 2021. 8

[44] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 6

[45] Robert Geirhos, Kristof Meding, and Felix A Wichmann. Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33:13890–13902, 2020. 7

[46] György Gergely and Gergely Csibra. Teleological reasoning in infancy: The naıve theory of rational action. *Trends in cognitive sciences*, 7(7):287–292, 2003. 7

[47] Madeleine Grunde-McLaughlin, Ranjay Krishna, and Maneesh Agrawala. Agqa: A benchmark for compositional spatio-temporal reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11287–11297, 2021. 4, 6, 7

[48] Stevan Harnad. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335–346, 1990. 2

[49] Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258, 2017. 2

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 3, 4

[51] Saiful Islam, Aurpan Dash, Ashek Seum, Amir Hossain Raj, Tonmoy Hossain, and Faisal Muhammad Shah. Exploring video captioning techniques: A comprehensive survey on deep learning methods. *SN Computer Science*, 2(2):1–28, 2021. 4

[52] L Itti. Models of bottom-up attention and saliency. In L Itti, G Rees, and J K Tsotsos, editors, *Neurobiology of Attention*, pages 576–582. Elsevier, San Diego, Ca., 2005. 8

[53] L Itti and C Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, Mar. 2001. 8

[54] Justin Johnson, Andrej Karpathy, and Li Fei-Fei. Densecap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4565–4574, 2016. 4

[55] Will Kay, Joao Carreira, Karen Simonyan, Brian Zhang, Chloe Hillier, Sudheendra Vijayanarasimhan, Fabio Viola, Tim Green, Trevor Back, Paul Natsev, et al. The kinetics human action video dataset. *arXiv preprint arXiv:1705.06950*, 2017. 3, 4

[56] Douwe Kiela and Léon Bottou. Learning image embeddings using convolutional neural networks for improved

multi-modal semantics. In *Proceedings of the 2014 Conference on empirical methods in natural language processing (EMNLP)*, pages 36–45, 2014. 3

[57] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 3

[58] John W Krakauer, Asif A Ghazanfar, Alex Gomez-Marin, Malcolm A MacIver, and David Poeppel. Neuroscience needs behavior: correcting a reductionist bias. *Neuron*, 93(3):480–490, 2017. 2

[59] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 3, 4, 6

[60] Brenden Lake and Marco Baroni. Generalization without systematicity: On the compositional skills of sequence-to-sequence recurrent networks. In *International conference on machine learning*, pages 2873–2882. PMLR, 2018. 4

[61] Lynn Le, Luca Ambrogioni, Katja Seeliger, Yağmur Güçlütürk, Marcel van Gerven, and Umut Güçlü. Brain2pix: Fully convolutional naturalistic video frame reconstruction from brain activity. *Frontiers in Neuroscience*, 16:940972, 2022. 2

[62] Chenyi Lei, Shixian Luo, Yong Liu, Wanggui He, Jiamang Wang, Guoxin Wang, Haihong Tang, Chunyan Miao, and Houqiang Li. Understanding chinese video and language via contrastive multimodal pre-training. In *Proceedings of the 29th ACM International Conference on Multimedia*, pages 2567–2576, 2021. 3, 4, 15

[63] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7331–7341, 2021. 4

[64] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*, 2019. 4, 6

[65] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer, 2020. 4, 6

[66] Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. What is more likely to happen next? video-and-language future event prediction. *arXiv preprint arXiv:2010.07999*, 2020. 5, 6

[67] Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. Hero: Hierarchical encoder for video+ language omni-representation pre-training. *arXiv preprint arXiv:2005.00200*, 2020. 3, 4, 6, 15

[68] Xiaodan Liang, Lisa Lee, Wei Dai, and Eric P Xing. Dual motion gan for future-flow embedded video prediction. In *proceedings of the IEEE international conference on computer vision*, pages 1744–1752, 2017. 5

[69] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1

[70] Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. *arXiv preprint arXiv:2109.13238*, 2021. 7

[71] Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910, 2020. 5, 6

[72] Shari Liu, Tomer D Ullman, Joshua B Tenenbaum, and Elizabeth S Spelke. Ten-month-old infants infer the value of goals from the costs of actions. *Science*, 358(6366):1038–1041, 2017. 7

[73] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022. 4

[74] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *Advances in neural information processing systems*, 32, 2019. 3

[75] Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*, 2020. 3, 4, 15

[76] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508:293–304, 2022. 4, 15

[77] Wei Ji Ma and Benjamin Peters. A neural network walks into a lab: towards using deep nets as models for human behavior. *arXiv preprint arXiv:2005.02181*, 2020. 2

[78] Raja Marjieh, Pol van Rijn, Ilia Sucholutsky, Theodore R Sumers, Harin Lee, Thomas L Griffiths, and Nori Jacoby. Words are all you need? capturing human sensory similarity with textual descriptors. *arXiv preprint arXiv:2206.04105*, 2022. 2

[79] Warren S. McCulloch and Walter Pitts. A logical calculus of the ideas immanent in nervous activity. *The bulletin of mathematical biophysics*, 5(4):115–133, Dec 1943. 1

[80] Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889, 2020. 3

[81] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings*

of the IEEE/CVF international conference on computer vision, pages 2630–2640, 2019. 4

[82] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013. 3

[83] Ida Momennejad. A rubric for human-like agents and neuroai. *Philosophical Transactions of the Royal Society B*, 378(1869):20210446, 2023. 2

[84] Anh Nguyen, Jason Yosinski, and Jeff Clune. Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 427–436, 2015. 8

[85] Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. *arXiv preprint arXiv:2112.07566*, 2021. 7

[86] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014. 3

[87] Jean Piaget. *The origins of intelligence in children.* W W Norton & Co, 1952. 2, 4

[88] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 1, 4

[89] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, 28, 2015. 3, 4

[90] Brian Riordan and Michael N Jones. Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, 3(2):303–345, 2011. 2

[91] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition: 36th German Conference, GCPR 2014, Münster, Germany, September 2-5, 2014, Proceedings 36*, pages 184–195. Springer, 2014. 6

[92] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3202–3212, 2015. 6

[93] Ludan Ruan and Qin Jin. Survey: Transformer based video-language pre-training. *AI Open*, 2022. 3

[94] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, Oct 1986. 1

[95] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115:211–252, 2015. 1

[96] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. Interpretation of natural language rules in conversational machine reading. *arXiv preprint arXiv:1809.01494*, 2018. 1

[97] Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, 22(1):55–67, 2021. 2

[98] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021. 1

[99] Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib J Majaj, Rishi Rajalingham, Elias B Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Franziska Geiger, et al. Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, page 407007, 2018. 1

[100] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*, 2017. 7

[101] Yuge Shi, Basura Fernando, and Richard Hartley. Action anticipation with rbf kernelized feature mapping rnn. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 301–317, 2018. 5

[102] Tianmin Shu, Abhishek Bhandwaldar, Chuang Gan, Kevin Smith, Shari Liu, Dan Gutfreund, Elizabeth Spelke, Joshua Tenenbaum, and Tomer Ullman. Agent: A benchmark for core psychological reasoning. In *International Conference on Machine Learning*, pages 9614–9625. PMLR, 2021. 7

[103] Linda Smith and Michael Gasser. The development of embodied cognition: Six lessons from babies. *Artificial life*, 11(1-2):13–29, 2005. 5, 7

[104] Mohammad Soleymani, David Garcia, Brendan Jou, Björn Schuller, Shih-Fu Chang, and Maja Pantic. A survey of multimodal sentiment analysis. *Image and Vision Computing*, 65:3–14, 2017. 5

[105] Elizabeth S Spelke. Core knowledge. *American psychologist*, 55(11):1233, 2000. 2, 4

[106] Elizabeth S Spelke and Katherine D Kinzler. Core knowledge. *Developmental science*, 10(1):89–96, 2007. 7

[107] Dan Sperber and Deirdre Wilson. *Relevance: Communication and cognition*, volume 142. Citeseer, 1986. 6

[108] Sanjana Srivastava, Guy Ben-Yosef, and Xavier Boix. Minimal images in deep neural networks: Fragile object recognition in natural images. *arXiv preprint arXiv:1902.03227*, 2019. 8

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

[110] Christian Szegedy, Sergey Ioffe, Vincent Vanhoucke, and Alexander Alemi. Inception-v4, inception-resnet and the impact of residual connections on learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017. 4

[111] Hao Tan and Mohit Bansal. Lxmert: Learning cross-modality encoder representations from transformers. *arXiv preprint arXiv:1908.07490*, 2019. 3

[112] Michael K Tanenhaus, Michael J Spivey-Knowlton, Kathleen M Eberhard, and Julie C Sedivy. Integration of visual and linguistic information in spoken language comprehension. *Science*, 268(5217):1632–1634, 1995. 2

[113] Jerry Tang, Meng Du, Vy A Vo, Vasudev Lal, and Alexander G Huth. Brain encoding models based on multimodal transformers can transfer across language and vision. *arXiv preprint arXiv:2305.12248*, 2023. 2

[114] Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. Coin: A large-scale dataset for comprehensive instructional video analysis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1207–1216, 2019. 5

[115] Esther Thelen and Linda B Smith. *A dynamic systems approach to the development of cognition and action*. MIT press, 1994. 5

[116] Atousa Torabi, Niket Tandon, and Leonid Sigal. Learning language-visual embedding for movie understanding with natural-language. *arXiv preprint arXiv:1609.08124*, 2016. 3

[117] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 3, 4

[118] Shikhar Tuli, Ishita Dasgupta, Erin Grant, and Thomas L Griffiths. Are convolutional neural networks or transformers more like human vision? *arXiv preprint arXiv:2105.07197*, 2021. 3

[119] A. M. TURING. I.—COMPUTING MACHINERY AND INTELLIGENCE. *Mind*, LIX(236):433–460, 10 1950. 1

[120] Shimon Ullman, Liav Assif, Ethan Fetaya, and Daniel Harari. Atoms of recognition in human and computer vision. *Proceedings of the National Academy of Sciences*, 113(10):2744–2749, 2016. 8

[121] Melissa Le-Hoa Vo. The meaning and structure of scenes. *Vision Research*, 181:10–20, 2021. 4

[122] Carl Vondrick, Hamed Pirsiavash, and Antonio Torralba. Generating videos with scene dynamics. *Advances in neural information processing systems*, 29, 2016. 5

[123] Aria Yuan Wang, Kendrick Kay, Thomas Naselaris, Michael J Tarr, and Leila Wehbe. Natural language supervision with a large and diverse dataset builds better models of human high-level visual cortex. *bioRxiv*, 2023. 2

[124] Huansha Wang, Ruiyang Huang, and Jianpeng Zhang. Research progress on vision–language multimodal pretraining model technology. *Electronics*, 11(21):3556, 2022. 3

[125] Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4581–4591, 2019. 6

[126] Xiaosha Wang, Wei Wu, Zhenhua Ling, Yangwen Xu, Yuxing Fang, Xiaoying Wang, Jeffrey R Binder, Weiwei Men, Jia-Hong Gao, and Yanchao Bi. Organizational principles of abstract words in the human brain. *Cerebral Cortex*, 28(12):4305–4318, 2018. 2

[127] Xiaohan Wang, Linchao Zhu, and Yi Yang. T2vlad: global-local sequence alignment for text-video retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5079–5088, 2021. 4

[128] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu. Neural encoding and decoding with deep learning for dynamic natural vision. *Cerebral cortex*, 28(12):4136–4160, 2018. 2

[129] Jonathan Winawer, Nathan Witthoft, Michael C Frank, Lisa Wu, Alex R Wade, and Lera Boroditsky. Russian blues reveal effects of language on color discrimination. *Proceedings of the national academy of sciences*, 104(19):7780–7785, 2007. 2

[130] Ludwig Wittgenstein. *Philosophical investigations*. John Wiley & Sons, 2010. 5

[131] Jeremy M Wolfe. Guided Search 4.0 Current Progress With a Model of Visual Search. In Wayne D Gray, editor, *Integrated Models of Cognitive Systems*, volume 1, pages 99–119. Oxford University Press, New York, 2007. ISSN: 15347362. 8

[132] Amanda L Woodward. Infants selectively encode the goal object of an actor's reach. *Cognition*, 69(1):1–34, 1998. 7

[133] Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu, and Kevin Murphy. Rethinking spatiotemporal feature learning: Speed-accuracy trade-offs in video classification. In *Proceedings of the European conference on computer vision (ECCV)*, pages 305–321, 2018. 3, 4

[134] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653, 2017. 6

[135] Hu Xu, Gargi Ghosh, Po-Yao Huang, Prahal Arora, Masoumeh Aminzadeh, Christoph Feichtenhofer, Florian Metze, and Luke Zettlemoyer. Vlm: Task-agnostic video-language model pre-training for video understanding. *arXiv preprint arXiv:2105.09996*, 2021. 3, 4, 15

[136] Hu Xu, Gargi Ghosh, Po-Yao Huang, Dmytro Okhonko, Armen Aghajanyan, Florian Metze, Luke Zettlemoyer, and Christoph Feichtenhofer. Videoclip: Contrastive pre-training for zero-shot video-text understanding. *arXiv preprint arXiv:2109.14084*, 2021. 4

[137] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296, 2016. 6

[138] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the national academy of sciences*, 111(23):8619–8624, 2014. 1

[139] Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. Clevrer: Collision events for video representation and reasoning. *arXiv preprint arXiv:1910.01442*, 2019. 6, 7

[140] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. Broaden the vision: Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*, 2021. 7

[141] Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. Video paragraph captioning using hierarchical recurrent neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4584–4593, 2016. 4

[142] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134, 2019. 4, 6

[143] Amir Zadeh, Rowan Zellers, Eli Pincus, and Louis-Philippe Morency. Multimodal sentiment intensity analysis in videos: Facial gestures and verbal messages. *IEEE Intelligent Systems*, 31(6):82–88, 2016. 5, 6

[144] Anthony Zador, Sean Escola, Blake Richards, Bence Ölveczky, Yoshua Bengio, Kwabena Boahen, Matthew Botvinick, Dmitri Chklovskii, Anne Churchland, Claudia Clopath, et al. Catalyzing next-generation artificial intelligence through neuroai. *Nature communications*, 14(1):1597, 2023. 1, 2, 8

[145] Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. Leveraging video descriptions to learn video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31, 2017. 6

[146] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *arXiv preprint arXiv:2304.00685*, 2023. 1

[147] Yaoyao Zhong, Junbin Xiao, Wei Ji, Yicong Li, Weihong Deng, and Tat-Seng Chua. Video question answering: Datasets, algorithms and challenges. *arXiv preprint arXiv:2203.01225*, 2022. 4

[148] Luowei Zhou, Chenliang Xu, and Jason Corso. Towards automatic learning of procedures from web instructional videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. 6

[149] Zhenglong Zhou and Chaz Firestone. Humans can decipher adversarial images. *Nature communications*, 10(1):1334, 2019. 8

[150] Linchao Zhu and Yi Yang. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755, 2020. 3, 4, 15

[151] Yi Zhu, Xinyu Li, Chunhui Liu, Mohammadreza Zolfaghari, Yuanjun Xiong, Chongruo Wu, Zhi Zhang, Joseph Tighe, R Manmatha, and Mu Li. A comprehensive study of deep video action recognition. *arXiv preprint arXiv:2012.06567*, 2020. 5

A. Pretraining tasks

Below a short list of common pretraining tasks.

Masked Language Modelling (MLM)[32] requires the model to predict masked words based on their surrounding words and their visually aligned video frames.

Masked Frame Modelling (MFM)[67] requires the model to predict masked out video frame features (as extracted with CNNs), given the text and remaining video frames.

Masked Visual-token Modelling (MVM) is similar to MFM, except that it uses 'tokenized' video frames instead of video frame features. Video frames are translated into discrete visual tokens, which can be used to reconstruct masked (regions of) video frames. The method is first used by [1] over the temporal dimension and later by [42] in both the temporal and spatial dimension.

Masked Object Classification (MOC) [150] is also similar to MFM but requires the model to predict masked out regional object features, instead of frame video features.

Masked Action Classification (MAC) [150] requires the model to predict masked out action features based on the remaining linguistic features and object features.

Video Subtitle Matching (VSM) [67] requires the model to predict whether a subtitle matches the input video, as well as to retrieve the relevant moment of localization, ensuring global and local temporal alignment.

Masked Modal Modelling (MMM) [75, 135] requires the model to predict all tokens from a completely masked out modality, based on the tokens from a other modality.

Frame Order Modelling (FOM)[67] requires the model to reconstruct the original timestamps of a set of randomly shuffled video frames, explicitly ensuring temporal alignment.

Sentence Order Modelling (SOM) [62] requires the model to reconstruct the original sentence order in a set of randomly selected and shuffled sentences.

Cross-Modal Matching (CMM) was introduced as 'the linguistic-visual alignment classification objective' [1], while [150] later called it cross-modal matching. By adding a linear layer followed by a sigmoid activation function on top of the output of the first token ([CLS]), a cross-modality score is achieved that indicates the relevance of the linguistic information and visual features. Alternatively, a similarity calculation module can be added to the network which calculates and optimzes the representational similarity between visual and textual information [76].

Language Reconstruction (LR) [75] requires the model to reconstruct words based on masked ground-truth text and video. LR is different from MLM in that LR focuses on next word prediction, i.e. the model only attends to previous word and video tokens when predicting the next word.