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

Anne Zonneveld\textsuperscript{1}  Albert Gatt\textsuperscript{2}  Iacer Calixto\textsuperscript{3}

\textsuperscript{1}Amsterdam Brain and Cognition Center, University of Amsterdam
\textsuperscript{2}Department of Information and Computing Sciences, Utrecht University
\textsuperscript{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 \cite{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 \cite{79, 94}. More recently, we have seen a resurgence of interest in the intersection between AI models and neuroscience \cite{144} and also by research explicitly comparing ANNs to predicting brain activity, for example in the context of object recognition \cite{138, 99} and language processing \cite{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 \cite{69, 95} and language \cite{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 \cite{109}. Models include image-language (IL) models, which learn representations that combine static images and text (e.g., CLIP \cite{88}), and video-language (VidL) models, which receive videos and text as inputs (e.g., VideoBERT \cite{1}). Although research on IL models has boomed over the last years and great progress has been made, 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 mimic 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.

\textsuperscript{1}See \cite{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 Visual Genome [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 (VVM)[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 ??.
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 ‘;’.

<table>
<thead>
<tr>
<th>Model</th>
<th>Stream</th>
<th>Visual backbone (object)</th>
<th>Visual backbone (action)</th>
<th>Pre-training tasks</th>
</tr>
</thead>
<tbody>
<tr>
<td>VideoBERT [14]</td>
<td>single</td>
<td>N.A.</td>
<td>S3D [133]; Kinetics [55]</td>
<td>MLM, MVM, CMM</td>
</tr>
<tr>
<td>UniVL [75]</td>
<td>dual</td>
<td>N.A.</td>
<td>S3D [133]; Howto100M [81]</td>
<td>MLM, MFM, MMM, CMM, LR</td>
</tr>
<tr>
<td>VLM [35]</td>
<td>single</td>
<td>N.A.</td>
<td>S3D [133]; Howto100M [81]</td>
<td>MLM, MFM, MMM</td>
</tr>
<tr>
<td>ClipBERT [63]</td>
<td>single</td>
<td>2D ResNet-50 [50]; ImageNet [31]</td>
<td>N.A.</td>
<td>MLM, MOC, CMM</td>
</tr>
<tr>
<td>VIOLET [42]</td>
<td>dual</td>
<td>N.A.</td>
<td>Video Swin Transformer [73]; Kinetics [55]</td>
<td>MLM, MVM</td>
</tr>
<tr>
<td>VideoCLIP [136]</td>
<td>dual</td>
<td>N.A.</td>
<td>S3D [133]; Howto100M [81]</td>
<td>CMM</td>
</tr>
<tr>
<td>CLIP4CLIP [76]</td>
<td>dual</td>
<td>N.A.</td>
<td>CLIP-ViT [88]; Howto100M [81]</td>
<td>CMM</td>
</tr>
</tbody>
</table>

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 sub-types 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. Compositional-ity 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 an 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 [145]. 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 metaphors 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
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

<table>
<thead>
<tr>
<th>Task</th>
<th>Sub-task</th>
<th>Dataset</th>
<th>MC?</th>
<th>Video Content</th>
<th>Text types</th>
<th>Query source</th>
<th>ML?</th>
<th>Grounding</th>
</tr>
</thead>
<tbody>
<tr>
<td>Retrieval</td>
<td>VR</td>
<td>YC2R[145]</td>
<td>✓</td>
<td>Instructional cooking</td>
<td>Subs + Description</td>
<td>Crowd</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>VATEX-EN-R[125]</td>
<td></td>
<td>Human activities</td>
<td>Description</td>
<td>Crowd</td>
<td>✓</td>
<td>WS4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSRVTT-R[73]</td>
<td></td>
<td>Various</td>
<td>Description</td>
<td>Crowd</td>
<td>WS3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ActivityNet Captions[59]</td>
<td></td>
<td>Human activities</td>
<td>Description</td>
<td>Crowd</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>MPII-MD[92]</td>
<td></td>
<td>Movies</td>
<td>Description</td>
<td>Professional</td>
<td>WS3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VCMR</td>
<td>DiDeMo[3]</td>
<td>✓</td>
<td>Flickr</td>
<td>Description</td>
<td>Crowd</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Charades-STA[44]</td>
<td></td>
<td>Scripted</td>
<td>Description</td>
<td>Semi-automatic</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VSCMR</td>
<td>TVR[65]</td>
<td>✓</td>
<td>TV shows</td>
<td>Subs+ Description</td>
<td>Crowd</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>How2R[67]</td>
<td>✓</td>
<td>Instructional</td>
<td>Subs+ Description</td>
<td>Crowd</td>
<td>✓</td>
<td>WS4</td>
</tr>
<tr>
<td>Captioning</td>
<td>Sentence</td>
<td>TVC[65]</td>
<td>✓</td>
<td>TV shows</td>
<td>Subs + Description</td>
<td>Crowd</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>YC2C[145]</td>
<td>✓</td>
<td>Instructional cooking</td>
<td>Subs + Description</td>
<td>Crowd</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>VATEX-EN-C[125]</td>
<td></td>
<td>Human activities</td>
<td>Description</td>
<td>Crowd</td>
<td>✓</td>
<td>WS3</td>
</tr>
<tr>
<td></td>
<td>Paragraph</td>
<td>Youtube-Clips[29]</td>
<td></td>
<td>Various</td>
<td>Description</td>
<td>Crowd</td>
<td>WS3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>TACoS-MultiLevel[91]</td>
<td></td>
<td>Instructional cooking</td>
<td>Description</td>
<td>Crowd</td>
<td>WS3</td>
<td></td>
</tr>
<tr>
<td>Q&amp;A</td>
<td>Factoid</td>
<td>Video-QA[145]</td>
<td>✓</td>
<td>Various</td>
<td>QA</td>
<td>Semi-automatic</td>
<td>WS3</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>How2-QA[67]</td>
<td>✓</td>
<td>Instructional</td>
<td>Subs + QA</td>
<td>Crowd</td>
<td>✓</td>
<td>WS3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>MSRVTT-QA[134]</td>
<td></td>
<td>Various</td>
<td>QA</td>
<td>Crowd</td>
<td>WS3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Compositional</td>
<td>TVQA[64]</td>
<td>✓</td>
<td>TV shows</td>
<td>Subs + QA</td>
<td>Crowd</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>AQQA[47]</td>
<td></td>
<td>Charades (scripted)</td>
<td>QA</td>
<td>Automatic</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>CLEVRER[139]</td>
<td></td>
<td>Synthetic</td>
<td>QA</td>
<td>Automatic</td>
<td>WS4</td>
<td></td>
</tr>
<tr>
<td>Social</td>
<td>Dialogue</td>
<td>Social-IQ[143]</td>
<td>✓</td>
<td>Youtube</td>
<td>Subs + QA</td>
<td>Crowd</td>
<td>WS5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>VLI</td>
<td>VIOLIN[71]</td>
<td>✓</td>
<td>TV shows + Youtube</td>
<td>Subs+ Description</td>
<td>Crowd</td>
<td>WS5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>FEP</td>
<td>VLEP[66]</td>
<td>✓</td>
<td>TV shows + Youtube</td>
<td>Subs</td>
<td>Crowd</td>
<td>WS4</td>
<td></td>
</tr>
</tbody>
</table>

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. **QA**: Question and Answer.
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: attribute 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, 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.
References


[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


[76] Huaiashao 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


[126] Xiaoshang 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


