Adaptive Memory-Augmented Video-Language Models for Fine-Grained Temporal Event Reasoning

Zhanhaoran Mao
School of Electrical Engineering and Computer Science, Oregon State University, Corvallis, OR, USA.
hellozhanhaoran@oregonstate.edu

Shane Miles
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
shane1990@ku.edu

Tejas J. Gandhi
Department of Computer Science and Engineering, University of Nevada, Reno, Reno, NV, USA.
gandhitejas@unr.edu

Abstract

Long-form video understanding is increasingly critical in domains ranging from autonomous surveillance to media analytics, yet contemporary video-language models remain limited in their capacity for fine-grained temporal event reasoning across extended sequences. This paper presents a systems-oriented examination of adaptive memory-augmented video-language architectures designed to capture and leverage long-range temporal dependencies for precise event localization, causal reasoning, and rare occurrence detection. We discuss the integration of dynamic external memory modules with transformer-based backbones, focusing on architectural trade-offs between read-write granularity, memory compression, and adaptive gating strategies that reconcile retention of salient information with computational feasibility. The analysis extends beyond algorithmic design to encompass infrastructure-level considerations, including distributed deployment, edge-cloud partitioning, latency constraints, and resilience to distributional shift. We further examine how memory-enabled video reasoning systems introduce novel governance challenges around fairness, bias propagation through accumulated knowledge, privacy of stored visual content, and the sustainability implications of persistent memory states. By synthesizing insights from computer vision, systems engineering, and AI policy, this work argues that adaptive memory augmentation for video-language models constitutes a socio-technical design problem requiring holistic frameworks that integrate robustness, equity, and resource efficiency. The discussion draws upon recent advances in video transformers, long-form video understanding, memory-augmented networks, and ethical AI deployment to articulate a forward-looking research agenda for temporally sophisticated and socially responsible video reasoning systems.

Keywords

Video-Language Models, Temporal Event Reasoning, Memory-Augmented Networks, Adaptive Architectures, Fairness in AI, Sustainable AI Systems.

1. Introduction

The rapid expansion of video data across security, healthcare, entertainment, and autonomous systems has elevated the demand for models capable of reasoning about events with fine temporal granularity. Modern video-language architectures, rooted in joint embedding spaces learned from massive image-text corpora and extended to video through spatiotemporal transformers and contrastive pre-training, have demonstrated impressive performance on tasks such as video captioning, question answering, and retrieval. Early successes were driven by adapting image-level transformer designs to sequences of frames through space-time attention factorization [1, 2, 3] and by jointly modeling visual and linguistic streams with cross-modal transformers [4]. These models, however, were primarily designed for short clips under one minute and implicitly assume that relevant visual evidence is densely present in a limited temporal window. In contrast, many real-world scenarios involve events unfolding over tens of minutes or hours, where causal connections span distant moments, action boundaries are subtle, and sparse infrequent events carry disproportionate semantic weight. The inability of fixed-context architectures to maintain a coherent internal representation across such extended durations constitutes a fundamental barrier to fine-grained temporal event reasoning.

Efforts to address long-form video understanding have progressed along multiple fronts. Dataset-driven approaches have expanded the temporal horizon of video benchmarks, most notably through large-scale egocentric video collections that capture daily activities over hours [5, 6]. Algorithmically, hierarchical and multiscale architectures have been proposed to capture temporal structure at varying resolutions, while temporal attention mechanisms with learned windowing schemes have reduced the quadratic cost of dense self-attention. Despite these advances, vanilla transformer-based architectures still discard information as soon as it exits the effective receptive field, forcing the model to compress all historical context into a fixed-length state vector, which often proves insufficient for recalling distant but causally relevant scenes. This limitation has motivated a growing interest in memory-augmented neural networks that maintain an external, dynamically addressable memory store, allowing explicit read and write operations over longer timescales than those permitted by internal hidden states. Foundational work on memory networks [7] and lifelong memory for rare events [8] established the potential of external memory for language and image tasks, while subsequent video captioning models incorporated attended memory banks that revisit both immediate and distant visual features during sentence generation [9]. Dynamic memory networks further allowed question-driven iterative memory updates, sharpening the retrieval of pertinent information [10].

The intersection of memory augmentation and hierarchical temporal representation is especially promising for fine-grained event reasoning. Hierarchical video understanding frameworks that construct pyramid-like representations over time enable models to reason about coarse story segments while simultaneously attending to fine-grained motion details. Multiscale vision transformers that operate on several temporal resolution streams have shown that such a strategy can preserve nuanced action dynamics without sacrificing the efficiency of global context aggregation [11]. Building on this direction, a recent hierarchical interleaved multi-stream motion encoding technique [12] demonstrates how to capture highly detailed temporal motion patterns across long videos by interleaving multiple encoding streams at different frame rates, effectively creating a rich motion-centric feature hierarchy that can be injected into downstream reasoning modules. This line of work points toward the possibility of integrating external memory not merely as a passive history buffer but as an active, structured repository that organizes incoming video streams into temporally indexed, motion-aware representations.

While algorithmic innovations are indispensable, the deployment of adaptive memory-augmented video-language systems in practical infrastructures raises a constellation of system-level and societal concerns that remain underexplored in the literature. Memory modules introduce persistent state that must be managed across inference sessions, with implications for latency, storage capacity, communication between edge and cloud nodes, and energy consumption. Moreover, because memory systems accumulate knowledge from sequences of past observations, they inherit the risks of bias amplification, fairness degradation, and privacy erosion that have become central to AI ethics discourse. This paper adopts a holistic perspective, treating adaptive memory-augmented video-language models not as isolated algorithmic artifacts but as socio-technical systems whose design must simultaneously satisfy computational, operational, ethical, and ecological requirements. The remainder of the paper is organized as follows. Section 2 examines architectural design choices and memory management schemes. Section 3 discusses temporal reasoning mechanisms and the interplay between memory state and event localization. Section 4 addresses infrastructure, deployment, and robustness considerations. Section 5 explores fairness, governance, and sustainability. Section 6 concludes with a synthesis of challenges and directions for future work.

2. Architectural Design and Adaptive Memory Management

Adaptive memory-augmented video-language architectures rest on the integration of an external memory module with a backbone capable of processing visual and textual streams. The memory module typically consists of a store of key-value pairs, where keys encode temporal and semantic context and values hold compressed visual features or cross-modal relational embeddings. Read operations retrieve memory entries relevant to a current temporal query, while write operations decide which new information to commit and which existing entries to update or evict. This read-write interface can be realized through soft attention mechanisms over the entire memory array, as in end-to-end trainable memory networks, or through a learnable controller that makes discrete addressing decisions, inspired by dynamic memory paradigms. The choice between continuous attention and discrete operations carries significant implications for scalability and differentiability; continuous soft memory is amenable to gradient-based optimization but incurs a computational cost that scales with memory size, whereas discrete addressing reduces the active working set but requires reinforcement learning or straight-through estimators that complicate training stability.

A central architectural trade-off concerns the granularity of memory entries. Fine-grained memory that stores per-frame or per-snippet features offers high temporal resolution but generates large memory banks that strain computational budgets during retrieval and writing. Coarser segment-level memory, on the other hand, groups frames into semantically coherent chunks and stores summary embeddings, reducing storage and retrieval costs at the risk of losing subtle temporal cues necessary for distinguishing visually similar but causally distinct events. Hierarchical memory structures that maintain multiple levels of granularity, from fine motion tokens to abstract event schemas, present a promising compromise. Such schemes align naturally with hierarchical video encoding backbones, where features at lower layers capture detailed motion and appearance while higher layers represent scene-level semantics. The interleaved multi-stream motion encoding approach [12] is particularly well-suited to feed such hierarchical memory banks, as it produces motion representations at distinct temporal resolutions that can be written into corresponding memory tiers, enabling the model

to answer questions about both instantaneous actions and long causal chains without overwhelming any single memory slice.

Adaptive gating mechanisms govern the dynamics of memory interaction, deciding when to read, which entries to retrieve, and how strongly to weight them relative to the current visual and textual inputs. A well-designed gating controller can condition these decisions on both the content of the query and the state of the memory, effectively implementing an attention-over-attention mechanism that filters out irrelevant historical clutter while amplifying context-dependent signals. The gating logic can be entirely parametric, relying on learned projections of the current multimodal embedding, or can incorporate explicit uncertainty estimates derived from the model's confidence in its ongoing event hypothesis. In either case, the controller must balance two competing objectives: maintaining long-term coherence by regularly refreshing memory with recent observations, and preventing catastrophic forgetting of rare but important earlier events. This necessitates a form of importance weighting that biases write operations toward temporally sparse or semantically surprising inputs, a principle reminiscent of lifelong memory strategies for rare event retention [8].

The interplay between memory management and cross-modal attention further shapes system architecture. Many video-language models rely on cross-attention layers where textual queries attend over video token sequences. When a memory module is introduced, textual queries can also attend over memory entries, effectively expanding the visual context window without increasing the number of active video frames processed by the encoder. However, double attention over both current frames and stored memories introduces a new computational burden. Strategies such as memory compression via hash-based addressing or product-key retrieval can reduce the number of memory entries searched, but they risk missing critical long-range correspondences. Alternatively, recurrent memory architectures that incrementally update a compact state vector as inspired by Transformer-XL's segment-level recurrence [13] offer a fixed memory footprint regardless of video length, yet they rely on the adequacy of the summary representation to capture all fine-grained details needed for later event reasoning. The design space is thus shaped by the tension between memory capacity, retrieval fidelity, computational cost, and the temporal granularity demanded by the downstream task.

Video-language pre-training has evolved to exploit contrastive objectives that align video segments with language captions. Models such as VideoCLIP [14] demonstrate that strong cross-modal alignments can be achieved through large-scale contrastive training on paired video-text data, producing representations that are highly effective for zero-shot video understanding. When integrated into a memory-augmented framework, pre-aligned representations reduce the burden on the memory controller to learn cross-modal correspondences from scratch, allowing it to specialize in temporal addressing and importance weighting. The memory module can store contrastively aligned embeddings, making it easier for the gating mechanism to retrieve entries on the basis of linguistic relevance. This modular decomposition, where pre-training handles cross-modal grounding and memory handles temporal persistence, encourages cleaner separation of concerns and enables reuse of state-of-the-art vision-language backbones without extensive architectural modification.

3. Temporal Reasoning Mechanisms and Event Localization

Fine-grained temporal event reasoning requires the model to localize events in time, determine their order, and infer causal relations that may span disconnected video segments. Memory-augmented architectures offer a structural advantage for such tasks because they decouple the temporal extent of reasoning from the model's immediate context window.

Rather than compressing all history into a single hidden vector, a query-driven read mechanism can selectively revisit memory entries corresponding to suspected event boundaries, reconstructing a timeline on demand. This aligns with the cognitive insight that human episodic memory supports mental time travel, allowing us to jump to specific moments in the past when constructing event narratives. In an architectural context, this capacity is realized through iterative memory access cycles where each read operation produces an updated belief about event structure, which in turn informs the next memory query.

A key challenge lies in maintaining temporal coherence across multiple memory retrievals. When a model retrieves a sequence of memory entries, those entries must preserve the relative temporal order of the original video while remaining accessible via semantic keys. This can be achieved by augmenting memory keys with positional encodings that encode temporal distance and ordering, and by structuring the memory as a linked list or differentiable priority queue that respects chronological constraints. Furthermore, the write policy must ensure that newly encoded observations do not overwrite temporally critical older entries unless an explicit consolidation step merges redundant or outdated information. Without such safeguards, the model can suffer from temporal confusion, where recalled events are misordered, undermining causal reasoning. Hierarchical motion encoding, which produces time-stamped features at several granularities, reduces this risk by naturally associating each memory entry with a well-defined temporal interval.

Event localization tasks, such as temporal action grounding in untrimmed videos, require the model to output precise start and end timestamps. Memory augmentation can enhance localization by allowing the model to compare candidate moments against a library of stored event prototypes and to propagate evidence from distant contextual snippets. Consider a video in which a person picks up a key early on and later unlocks a door; a purely feed-forward model that processes only a local sliding window might miss the causal link between the two actions, whereas a memory-augmented model that retrieves the earlier key-grasp event upon encountering the door scene can strengthen its confidence in the unlocking action and refine its boundaries. Adaptive memory gating is particularly useful in such scenarios: the controller can increase the read weighting for entries that exhibit high semantic affinity with the current query, effectively zooming out temporally to incorporate relevant earlier context while ignoring unrelated filler segments. This capability reduces the reliance on dense annotation of action boundaries during training, as the model learns to leverage its memory to disambiguate event extents.

Causal reasoning across long time intervals further stresses the memory system’s ability to store and link relational information. Beyond simple temporal ordering, many events involve precondition relations such that one event enables or prevents another. Representing these relations in memory requires not only storing visual features but also encoding the functional roles of objects and actions. One possibility is to structure memory entries as triples of subject, predicate, and object, akin to symbolic knowledge graph representations, while grounding each element in visual features. The memory controller can then perform graph traversal during inference, following causal links that span large temporal gaps. Although such hybrid neural-symbolic memory structures are still nascent, they point toward a future in which adaptive memory modules for video reasoning evolve from unstructured feature banks to organized knowledge repositories that support explicit reasoning primitives.

4. Infrastructure, Deployment, and Robustness

Deploying adaptive memory-augmented video-language models in real-world infrastructures demands architectural choices that extend far beyond academic prototyping. Video analytics pipelines in smart cities, retail, and industrial monitoring are typically organized around a hierarchy of cameras, edge nodes, and cloud data centers. The decision of where to place memory state and compute resources along this hierarchy has profound implications for latency, bandwidth, privacy, and energy consumption. Lightweight memory modules that store compressed video representations on edge devices enable low-latency event reasoning without continuous transmission of high-resolution video, aligning with the edge intelligence paradigm that pushes model inference and stateful processing to the network periphery [15, 16]. However, edge hardware is severely constrained in memory capacity and power, which restricts the amount of historical information that can be retained locally and forces aggressive compression and adaptivity in write policies. In contrast, centralizing the memory bank in the cloud allows unbounded storage and enables cross-camera aggregation, but introduces round-trip delays that may be unacceptable for time-critical applications such as anomaly detection in autonomous driving.

The robustness of memory-augmented reasoning systems to distributional shift is another pressing concern. Videos captured in dynamic environments undergo changes in lighting, viewpoint, and scene composition that can degrade the quality of visual features stored in memory. If features extracted under one condition are later queried under a different condition, the resulting memory retrievals can become noisy or misaligned. Systematic benchmarking of neural network robustness to common corruptions and perturbations [17] indicates that model performance often drops sharply under covariate shift, and the persistent nature of memory means that errors and degradation can accumulate over time, contaminating the knowledge base. Addressing this requires memory refresh strategies that periodically re-encode stored entries using the most recent encoder parameters or adaptively blend old features with newly encoded versions to maintain consistency. The memory controller can also be equipped with uncertainty estimates that down-weight entries recorded under degraded conditions, effectively implementing a form of self-assessment that promotes resilience.

Scalability challenges emerge from the compounding growth of memory as video streams accumulate. Even with aggressive compression, a continuously updating memory that serves multiple concurrent inference queries demands careful engineering of indexing structures, parallelization, and load balancing. Systems inspired by distributed key-value stores can distribute memory shards across multiple accelerators, but the need for random access with low latency favors in-memory storage rather than disk-based retrieval, which raises hardware cost and energy footprint. Furthermore, the interplay between memory size and model accuracy is not monotonic; above a certain threshold, adding more entries yields diminishing returns as the retrieval mechanism becomes overwhelmed by irrelevant distractors. This motivates the design of memory forgetting and consolidation policies that are themselves adaptive, pruning entries that have low relevance scores or have been superseded by more informative recent observations. Such policies must be carefully calibrated to avoid catastrophic forgetting of rare but safety-critical events, demonstrating once again the delicate equilibrium between retention and computational pragmatism.

From a systems engineering perspective, the lifecycle of memory-augmented models encompasses training, fine-tuning, and continuous online adaptation. Incremental learning scenarios in which the memory system is updated with new events over months or years challenge the assumption of static training distributions. The model must be capable of

incorporating novel event types without overwriting past knowledge, a requirement that intersects with ongoing research in continual learning and neural memory. While adaptive memory offers a natural substrate for such incremental updates, ensuring that fairness and performance metrics do not drift unacceptably over deployment time requires monitoring infrastructure and rollback mechanisms that are often absent in current AI operations frameworks.

5. Fairness, Governance, and Sustainability

Memory-augmented video-language systems that persist knowledge across time introduce distinctive fairness and governance risks. Because memory accumulates observations drawn from a potentially biased distribution of video data, the stored knowledge can encode and amplify societal biases present in the visual world. For instance, if a surveillance memory system disproportionately records interactions in certain neighborhoods due to uneven camera placement, its retrieved event representations may overrepresent correlations between demographic appearance and suspicious activities, leading to downstream reasoning that systematically disadvantages particular communities. Transparency and auditability become essential, as it is difficult for stakeholders to interrogate the opaque knowledge embedded in large memory banks. Model cards and similar documentation frameworks [18] provide a starting point by encouraging developers to disclose the provenance and characteristics of training and deployment data, but they do not yet adequately capture the dynamic, evolving nature of memory state. A more thorough governance approach would mandate periodic audits of memory content distributions, bias testing using counterfactual event pairs, and clear attribution of the sources that contributed to a reasoning outcome.

Privacy concerns are heightened when memory modules persistently store visual information that may include personally identifiable details, even in compressed feature form. While storing latent embeddings rather than raw pixels reduces re-identification risk, inversion attacks have demonstrated that significant visual detail can be reconstructed from deep features, raising the specter of unauthorized disclosure. Federated learning and on-device memory configurations can mitigate some risks by keeping data localized, but they complicate coordinated event reasoning across cameras. Differential privacy guarantees can be integrated into the write operations to bound the amount of individual-specific information retained in memory, yet such guarantees typically come at the cost of degraded utility for rare event detection. Policy frameworks must thus navigate the tension between privacy preservation and the societal benefits of robust event reasoning, potentially through graduated access controls that restrict memory queries to authorized use cases and enforce retention limits.

Sustainability constitutes an increasingly urgent dimension of large-scale AI systems. Training large video-language models with memory augmentation is computationally intensive, but the inference stage exerts a more continuous environmental impact, especially in always-on video analytics deployments. Persistent memory states that avoid re-encoding entire video histories can reduce redundant computation, thereby lowering operational carbon emissions relative to systems that reprocess sliding windows from scratch [19]. Conversely, the additional memory hardware required to store large embedding banks increases the embodied carbon of the infrastructure. A holistic carbon accounting must therefore weigh the reduced per-query energy against the increased resource provisioning. Adaptive memory that dynamically adjusts its capacity based on content complexity and query load offers a mechanism to align resource consumption with actual need, embodying the green AI principle

that computational budgets should be allocated in proportion to task difficulty rather than fixed a priori. Designing such adaptive resource allocators demands tight integration between machine learning orchestrators and energy-aware hardware scheduling, a frontier that remains largely unexplored in video reasoning contexts.

Finally, the governance of adaptive memory-augmented systems intersects with broader ethical principles for AI, including accountability, explainability, and the right to contest decisions. If a video-based event reasoning system makes a consequential determination, perhaps flagging an individual for security screening based on a sequence of stored visual events, the affected person should be able to understand which specific memories contributed to that decision and to challenge their accuracy or relevance. Implementing such contestability calls for memory modules that log retrieval provenance and support human-interpretable justifications, moving beyond compression toward structured event graphs that can be surfaced in explanation interfaces. The multi-disciplinary synthesis of technical, legal, and societal considerations mapped in consensus documents on AI ethics [20] underscores that no single stakeholder group can resolve these challenges in isolation; instead, the design of memory-augmented video-language architectures must be embedded in multi-stakeholder governance processes from the outset.

6. Conclusion

Adaptive memory-augmented video-language models represent a frontier in fine-grained temporal event reasoning, offering the capacity to span long temporal horizons, retrieve causally relevant past observations, and dynamically adapt memory allocation to the demands of complex video streams. This paper has examined this class of systems through a wide-angle lens that integrates architectural innovation with infrastructure, deployment, robustness, fairness, and sustainability considerations. The design space is rich with trade-offs: between memory granularity and computational cost, between persistent state and privacy, between localized edge inference and global cloud aggregation, and between adaptive forgetting and the retention of rare but significant events. Addressing these trade-offs demands more than isolated algorithmic progress; it requires a concerted effort to embed memory-augmented reasoning within holistic systems that are monitored for bias drift, audited for fairness, secured against reconstruction attacks, and optimized for carbon efficiency across their operational lifetimes. Future research should advance adaptive controllers that operate under explicit multi-objective regimes balancing accuracy, latency, energy, and equity metrics, while policy frameworks must evolve to provide legal clarity around persistent visual memory. Only through such integrated inquiry can the full potential of temporally aware video-language reasoning be realized in a manner that is technically robust, socially accountable, and environmentally responsible.

References

1. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In International Conference on Learning Representations.

2. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lucic, M., & Schmid, C. (2021). ViViT: A video vision transformer. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6836-6846).

3. Bertasius, G., Wang, H., & Torresani, L. (2021). Is space-time attention all you need for video understanding? In Proceedings of the International Conference on Machine Learning (pp. 813-824).

4. Sun, C., Myers, A., Vondrick, C., Murphy, K., & Schmid, C. (2019). VideoBERT: A joint model for video and language representation learning. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 7464-7473).

5. Wu, C.-Y., & Krahenbuhl, P. (2021). Towards long-form video understanding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1884-1894).

6. Grauman, K., Westbury, A., Byrne, E., Chavis, Z., Furnari, A., Girdhar, R., ... & Malik, J. (2022). Ego4D: Around the world in 3,000 hours of egocentric video. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 18995-19012).

7. Weston, J., Chopra, S., & Bordes, A. (2015). Memory networks. In International Conference on Learning Representations.

8. Kaiser, L., Nachum, O., Roy, A., & Bengio, S. (2017). Learning to remember rare events. In International Conference on Learning Representations.

9. Pei, W., Zhang, J., Wang, X., Ke, Y., & Shen, C. (2019). Memory-attended recurrent network for video captioning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 42(7), 1765-1778.

10. Kumar, A., Irsoy, O., Ondruska, P., Iyyer, M., Bradbury, J., Gulcehre, C., ... & Socher, R. (2016). Ask me anything: Dynamic memory networks for natural language processing. In Proceedings of the International Conference on Machine Learning (pp. 1378-1387).

11. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., & Feichtenhofer, C. (2021). Multiscale vision transformers. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6824-6835).

12. Jin, Haopeng, et al. "HY-Himmel Technical Report: Hierarchical Interleaved Multi-stream Motion Encoding for Long Video Understanding." arXiv preprint arXiv:2605.08158 (2026).

13. Dai, Z., Yang, Z., Yang, Y., Carbonell, J., Le, Q. V., & Salakhutdinov, R. (2019). Transformer-XL: Attentive language models beyond a fixed-length context. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 2978-2988).

14. Xu, H., Ghosh, G., Huang, P.-Y., Arora, P., Aminzadeh, M., Feichtenhofer, C., ... & Zettlemoyer, L. (2021). VideoCLIP: Contrastive pre-training for zero-shot video-text understanding. In Proceedings of the Conference on Empirical Methods in Natural Language Processing (pp. 6787-6800).

15. Zhang, H., Ananthanarayanan, G., Bodik, P., Philipose, M., Bahl, P., & Freedman, M. J. (2017). Live video analytics at scale with approximation and delay-tolerance. In Proceedings of the 14th USENIX Symposium on Networked Systems Design and Implementation (pp. 377-392).

16. Zhou, Z., Chen, X., Li, E., Zeng, L., Luo, K., & Zhang, J. (2019). Edge intelligence: Paving the last mile of artificial intelligence with edge computing. Proceedings of the IEEE, 107(8), 1738-1762.

17. Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. In International Conference on Learning Representations.

18. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. In Proceedings of the Conference on Fairness, Accountability, and Transparency (pp. 220-229).

19. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. Communications of the ACM, 63(12), 54-63.

20. Fjeld, J., Achten, N., Hilligoss, H., Nagy, A., & Srikumar, M. (2020). Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for AI. Berkman Klein Center for Internet & Society Research Publication.