Efficient Spatiotemporal Token Compression for Large-Scale Multimodal Video Retrieval

Sagar Kamar
Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
sagark@ku.edu

Otis Cook
Department of Computer Science, University of North Texas, Denton, TX, USA.
hellootis@unt.edu

Abstract

The rapid growth of multimodal video archives demands retrieval systems that can efficiently index and search across vast collections of video, audio, and textual metadata. Transformer-based architectures have become the backbone of such systems, converting raw media streams into sequences of high-dimensional embeddings or tokens. However, the quadratic complexity of attention mechanisms and the sheer number of tokens generated from spatiotemporal video features create formidable computational and storage bottlenecks. This paper presents a comprehensive system-level analysis of spatiotemporal token compression as a strategic intervention for large-scale multimodal video retrieval. We examine the landscape of token reduction techniques—including pruning, merging, and learned selection—and their integration into end-to-end retrieval pipelines. The discussion extends beyond algorithmic innovation to address infrastructure design choices, trade-offs between compression ratio and retrieval relevance, systems-wide robustness, and the governance challenges that emerge when retrieval results shape public information access. We further investigate energy consumption, lifecycle carbon costs, fairness in multimodal representations, and the policy implications of deploying compressed retrieval systems in ethically sensitive domains. By synthesizing insights from computer vision, natural language processing, systems engineering, and technology policy, the paper identifies critical structural tensions and offers a forward-looking perspective on building sustainable, equitable, and high-performance video retrieval systems. The analysis underscores that efficient token compression is not merely a computational optimization but a sociotechnical design choice with far-reaching consequences for data governance, accessibility, and public trust in AI-driven information retrieval.

Keywords

multimodal video retrieval, spatiotemporal token compression, vision transformers, large-scale systems, retrieval infrastructure, fairness, sustainability, AI governance.

1. Introduction

The exponential proliferation of video data across domains such as news broadcasting, social media, surveillance, and cultural heritage has catalyzed the development of large-scale multimodal retrieval systems capable of connecting queries expressed in natural language, images, or other modalities to relevant video segments. Contemporary architectures achieve this cross-modal alignment by projecting diverse media into a shared embedding space,

commonly using transformer-based encoders pretrained on massive image-text or video-text pairs. Foundational vision transformers demonstrated that a pure self-attention mechanism can yield state-of-the-art performance on image recognition by treating an image as a sequence of patch embeddings, and subsequent works extended this paradigm to video by factorizing attention across spatial and temporal dimensions, producing spatiotemporal tokens that encode motion and appearance cues [1, 2]. The introduction of contrastive language-image pretraining further popularized dual-encoder designs that align visual tokens with textual representations, enabling efficient similarity search in a latent joint space [3].

As video modalities become richer and frame rates increase, the volume of spatiotemporal tokens generated from even a single short clip rapidly inflates. Factorized video transformers such as ViViT process clips by extracting tubelet embeddings along spatial and temporal axes, compounding the token count [4]. Masked autoencoding approaches like VideoMAE demonstrate that a high proportion of tokens can be masked during pretraining without degrading downstream performance, hinting at substantial redundancy in dense token sequences [5]. Meanwhile, large-scale video-language models such as BLIP that unify understanding and generation across multimodal data further intensify the computational demands on retrieval backbones, necessitating careful management of token throughput at both training and inference stages [6]. The gap between the ballooning token budgets required for state-of-the-art representations and the latency, memory, and energy constraints of production retrieval systems has emerged as a central research challenge.

This paper frames token compression not as an isolated model-level optimization but as a systems-wide design variable that cuts across the architecture, deployment infrastructure, sustainability, and governance of multimodal video retrieval engines. We argue that efficient spatiotemporal token compression must be evaluated using a multifaceted lens that accounts for structural trade-offs, fairness implications, and the lifecycle impacts of large-scale AI operations. The following sections systematically unpack the compression design space, examine the infrastructural scaffolding required to serve compressed retrieval indices at scale, and explore the sociotechnical risks that compressed representations may inadvertently amplify.

2. Tokenization and Retrieval Architectures for Multimodal Video

Modern multimodal retrieval pipelines for video generally adopt a two-stage paradigm: an offline indexing phase encodes video content into a compact vector representation, and an online query phase maps user queries into the same latent space for nearest neighbor search. The encoder architecture profoundly influences the token geometry and the subsequent compression strategy. Early video retrieval systems relied on pretrained 3D convolutional networks that produce clip-level feature vectors, but the transformer revolution replaced these with patch-based tokenizers that decompose each frame into a grid of non-overlapping patches and then process the resulting sequence with self-attention. A common variant factorizes attention along spatial and temporal dimensions to control computational complexity while preserving motion sensitivity, but even factorized operations must still attend over tens of thousands of tokens for long videos, making token reduction a prerequisite for scalable indexing.

The dual-encoder design popularized by CLIP and its video adaptations employs separate transformers for visual and text modalities, projecting them into a shared embedding space via contrastive learning [3, 7]. In video settings, the visual encoder must condense a variable-length sequence of spatiotemporal tokens into a fixed-size representation suitable for fast

inner-product search. Commonly, a simple temporal pooling or the classification token embedding serves as the aggregated video vector, but such naive aggregation discards fine-grained temporal information that could be vital for moment-level retrieval queries. To mitigate this loss, recent systems adopt hierarchical indexing strategies that retain a set of per-clip embeddings alongside a global video fingerprint, effectively multiplying the number of tokens stored per video. Hybrid encoders that fuse vision and language within a single backbone further complicate token management, because cross-attention between modalities can require maintaining large intermediate token buffers during inference, pushing retrieval latency beyond acceptable thresholds for interactive applications.

The retrieval backend relies heavily on approximate nearest neighbor libraries such as FAISS, which are optimized for high-dimensional dense vectors but scale linearly with the dimensionality and logarithmically with the dataset size under appropriate index structures [8]. The product of token count per video and embedding dimension thus directly determines the memory footprint and query time of the index. For a dataset containing hundreds of millions of video clips, even a modest token budget of a few thousand per clip can overflow the memory capacity of a single machine, forcing sharding across distributed nodes. Every additional token preserved in the compressed representation increases the storage cost linearly and the search cost multiplicatively with the number of probes in the index, creating a sharp economic incentive for aggressive yet quality-preserving compression.

3. Spatiotemporal Token Compression Paradigms

The compulsion to reduce token counts without sacrificing retrieval quality has given rise to a rich family of compression techniques that can be broadly categorized into pruning, merging, and learned token selection. Each paradigm offers distinct trade-offs between compression ratio, representational fidelity, and computational overhead, and their suitability varies with the retrieval task, video genre, and downstream fairness requirements.

Token pruning methods drop less informative tokens based on an importance score predicted by a lightweight gating module, which is often trained jointly with the main task. The DynamicViT framework first demonstrated that conditional computation could eliminate a large fraction of image tokens by ranking them according to a learned significance metric, a principle later extended to video by adding a temporal gating dimension [9]. Pruning can achieve dramatic compression ratios—exceeding 80% token reduction in many video benchmarks—while preserving classification accuracy, but it introduces a non-differentiable sampling step that has historically complicated end-to-end training. Recent advances replace hard selection with differentiable top-k operators or straight-through estimators, enabling gradient flow through the compression module and making token pruning more amenable to retrieval-oriented fine-tuning. A major structural drawback of pruning, however, is that it permanently discards visual evidence that might be irrelevant for the training objective but critical for open-ended retrieval queries where a user may seek semantically peripheral visual details.

Merging-based approaches take a less destructive path by combining redundant tokens into fewer aggregated representatives. Token merging as popularized by ToMe for images and later adapted for video operates by iteratively grouping tokens with high cosine similarity in the key space of transformer layers, averaging their feature vectors, and passing the merged representation forward [10]. For spatiotemporal video tokens, merging can be applied independently to spatial patches within each frame or across temporal neighbors, yielding spatially consistent segmentation-like groupings that preserve object boundaries and motion

coherence. Merging is fully differentiable and does not introduce additional learnable parameters beyond the similarity thresholding logic, making it appealing for latency-sensitive retrieval backends that cannot afford auxiliary networks. However, the quality of merging degrades when faced with fast camera motion or frequent shot changes, because naive similarity metrics may erroneously fuse tokens from distinct semantic entities, blurring discriminative video features and reducing retrieval precision for fine-grained queries.

Learned token selection with a dedicated tokenization bottleneck, exemplified by the TokenLearner family, inserts a small convolutional or attention-based module that adaptively generates a compact summary of the visual input by learning to attend to informative spatial locations [11]. This approach can be seen as a middle ground: it does not hard-drop tokens but instead recomputes a fixed small set of output tokens that attend over the full input token sequence, compressing information into a learnable codebook that is jointly optimized with the retrieval objective. The fixed output size guarantees predictable memory usage, which is invaluable for deployment on resource-constrained edge nodes serving retrieval requests. The modular design also permits hierarchical application, where a first-stage TokenLearner reduces per-frame tokens and a second-stage temporal aggregator compresses the frame-level representation, aligning naturally with the multi-granular retrieval architectures discussed earlier. The primary cost is the additional forward pass through the selection module, which can offset some of the efficiency gains if not carefully implemented.

Hybrid strategies have emerged that interleave the above paradigms across transformer layers. For instance, early layers may perform aggressive pruning to quickly shed background tokens, while deeper layers apply cautious merging to consolidate semantic regions before the final representation is pooled for retrieval. Such schedules can be dynamically adjusted based on video complexity estimated through low-cost motion heuristics, enabling content-adaptive compression that allocates a larger token budget to visually dynamic segments and a sparser budget to static scenes. The design space thus expands beyond static compression ratios to runtime-contingent policies, which in turn demand new orchestration logic within the retrieval pipeline to handle variable-length compressed representations.

Recent work on hierarchical interleaved multi-stream motion encoding further enriches the token compression landscape by decomposing video into parallel streams that encode frame-level appearance descriptors, short-term motion residuals, and long-range temporal context tokens, and then interleaving these streams within a unified transformer backbone that selectively attends across streams based on task-specific gating signals [15]. This design allows the system to allocate token capacity asymmetrically across information streams, preserving fine-grained motion detail in the high-resolution motion stream while aggressively compressing static background regions. The interleaved architecture also facilitates asynchronous decoding of compressed tokens during retrieval, where a query that requires only coarse temporal localization can bypass the motion stream entirely, further reducing computational expenditure per query. Such architectural innovations highlight a broader trend in which token compression becomes inseparable from the overall design of the representation space, rather than being a post-hoc performance patch.

The selection among these paradigms must be governed by careful systems analysis that accounts for the statistical properties of the video corpus, the query distribution, the latency budget, and the marginal utility of additional tokens for retrieval accuracy. Human-centered evaluations that measure the effect of compression on search result relevance and user

satisfaction remain scarce, and this gap represents a critical opportunity for future interdisciplinary research.

4. Infrastructure for Large-Scale Compressed Retrieval

Deploying token-compressed retrieval pipelines at web scale requires a carefully engineered infrastructure that spans data ingestion, encoding, index construction, query serving, and continuous model update cycles. The architectural design of such infrastructure must reconcile several competing demands: the need for high throughput to ingest live video streams, low-latency query response even under peak load, cost-efficient storage of billions of compressed embeddings, and resilience against hardware failures and data drift.

A common pattern distributes the encoding workload across a fleet of GPU-accelerated instances that run the pretrained video transformer with integrated compression modules. Because token compression reduces the size of intermediate activations, it also lowers the memory bandwidth pressure and allows higher batch sizes during offline indexing, effectively amortizing the cost of the compression-specific computation. However, this benefit is partially offset by the need to maintain multiple encoding pipelines when different compression granularities are deployed for different retrieval tiers—for example, a coarse index for initial filtering and a fine-grained index for re-ranking. A microservice architecture that encapsulates each encoder variant as a scalable, stateless service with a consistent gRPC or REST interface enables dynamic resource allocation and simplifies A/B testing of new compression algorithms in production traffic.

Index construction for compressed representations must handle variable-length token sequences if dynamic compression policies vary across videos. Sharding strategies that partition the corpus by video identifier or by semantic category can balance index size and query parallelism, but they introduce cross-shard communication overhead when a query needs to probe multiple shards. Hierarchical indexing schemes, such as inverted file with product quantization, can accommodate compressed tokens by treating each token as a separate vector or by aggregating multiple tokens into a single query-passage through learned weighted pooling. The optimal shard granularity and replication factor are sensitive to the compression ratio: more aggressive compression yields smaller per-video footprints that allow a larger fraction of the index to reside in memory, reducing disk I/O during query time and improving tail latency.

Sustainability considerations increasingly influence infrastructure decisions. The electricity consumption and embedded carbon footprint of large-scale retrieval systems have drawn scrutiny from both researchers and regulators. Token compression directly reduces the floating-point operations required per video encoding, which translates into lower energy consumption per video indexed. However, the lifecycle costs of the specialized hardware accelerators needed to run custom compression modules and the network energy consumed in distributed index lookups must also be accounted for [16]. Data center operators are beginning to incorporate energy proportionality into their orchestration logic, scheduling batch indexing jobs during periods of high renewable energy availability and powering down idle GPU nodes. Compression algorithms that offer a tunable efficiency-accuracy knob enable operators to shift dynamically toward more compression when the carbon intensity of the grid is high, embodying a form of carbon-aware retrieval that aligns systems performance with environmental responsibility.

Data governance mechanisms are integral to the infrastructure layer. The pipelines that transcode and tokenize video must respect data residency constraints, consent preferences, and intellectual property boundaries, especially when dealing with user-generated content or archival footage with mixed rights. Compressed embeddings, while abstracted from raw pixels, can still leak membership information or be susceptible to model inversion attacks that reconstruct recognizable visual features, raising data protection obligations under regulations such as the General Data Protection Regulation [20]. Access control and audit logging must extend to the token compression stage, ensuring that any party that queries the retrieval system cannot infer whether a specific video was part of the indexing corpus through crafted query sequences. Privacy-preserving token compression techniques based on differential privacy or adversarial learning are still in their infancy but will become essential as retrieval systems are deployed in domains such as healthcare video analysis and legal evidence search.

5. Deployment, Robustness, and Fairness

Moving a spatiotemporal token compression system from a controlled offline benchmark to a live deployment serving heterogeneous users and queries reveals a host of robustness and fairness challenges that are often invisible during academic evaluation. Compression artifacts can differentially affect retrieval performance across demographic groups, content genres, and query types, leading to systematic disparities in information access that demand careful fairness auditing.

Robustness concerns arise from distribution shifts between the data used to train the compression modules and the videos encountered in production. A pruning network that learns to predict token importance based on a pretraining corpus dominated by curated studio footage may perform poorly on noisy user-generated videos with unstable camera work, occlusion, or unusual framing, dropping tokens that are critical for recognizing rare objects or actions. Similarly, merging algorithms that rely on global similarity thresholds may over-merge tokens in low-light scenes where the signal-to-noise ratio is already degraded, causing a non-linear loss of retrieval recall for queries referencing subtle visual attributes. Content-adaptive compression policies that adjust thresholds based on low-level video statistics can mitigate these failures, but they require online estimation of compression safety, which adds computational control logic that must itself be robust against adversarial inputs crafted to trigger pathological token discard.

Fairness in multimodal retrieval is a multidimensional construct that intersects with representation bias in the underlying pretrained vision-language models, unequal performance across skin tones, genders, or geographic regions, and disparate impacts of compression-induced information loss on marginalized query topics. Vision transformers pretrained on Internet-scale image-text pairs have been shown to encode harmful stereotypes and to exhibit performance disparities in person-related recognition tasks, and these biases propagate into the video domain when temporal aggregation amplifies spurious correlations [17, 18]. Token compression can exacerbate these inequities if the gating or merging criteria inadvertently penalize visual features correlated with protected attributes. For example, a learned importance score trained with a cross-modal contrastive objective might systematically assign lower weight to tokens covering head coverings or culturally specific attire, because such visual patterns are underrepresented in the retrieval training queries, leading to higher retrieval failure rates for queries targeted at those populations. Mitigating these risks requires integrating fairness constraints into the compression training objective, conducting disaggregated evaluation of retrieval quality across demographic slices, and establishing

transparent documentation practices such as model cards that report compression-induced fairness metrics [21].

Deployment monitoring infrastructure must continuously track token compression behavior alongside retrieval quality indicators. Canary queries that probe known edge cases, demographic slices, and rare visual concepts should be injected into live traffic to alert operators when a model update or data drift causes a significant degradation in fairness or recall. The token budget itself can become a governance lever, with platform policies specifying minimum representation thresholds for certain protected visual categories that must be preserved during compression, backed by automated enforcement in the token selection logic. The design of such fairness-preserving compression constraints remains an open research area that sits at the intersection of systems engineering and AI ethics.

6. Sustainability and Policy Implications

The environmental footprint of large-scale AI systems has become a pressing policy concern, and multimodal video retrieval represents an especially energy-intensive application due to its reliance on high-throughput tensor operations and dense nearest neighbor search index scans. Token compression contributes directly to sustainability by reducing the computational load per video, but its net effect depends on how operators leverage the freed capacity. If efficiency gains are reinvested into indexing more videos or serving higher query volumes—a phenomenon known as the rebound effect—the absolute energy consumption may remain unchanged or even increase. Sustainability-aware system governance thus requires coupling token compression with aggregate energy budgets and demand-side policies that incentivize reductions in overall compute usage rather than merely per-unit efficiency.

Policy frameworks such as the European Union AI Act are beginning to impose documentation obligations on high-risk AI systems, and large-scale content retrieval engines that influence public access to information could plausibly fall under such categorization [19]. Compressed video embeddings that serve as the backbone of retrieval indices may need to be registered as part of the technical documentation, accompanied by evidence that compression does not degrade the system's robustness and fairness below acceptable thresholds. These regulatory dynamics create a new class of compliance requirements for token compression research: algorithms must be auditable, and the rationale for retaining or discarding certain spatiotemporal tokens must be interpretable by third-party assessors. The opacity of learned compression gating mechanisms clashes with the regulatory demand for explainability, prompting a need for inherently interpretable compression modules or post-hoc explanation tools that can associate compressed token sets with their contribution to retrieval outcomes.

International data governance further shapes the topology of retrieval infrastructure. Regulations that restrict cross-border transfer of personal data may mandate that raw video footage be processed and tokenized within national or regional boundaries, but the resulting compressed embeddings—if deemed sufficiently anonymized—might be transferred to central index servers for global search. The legal characterization of a compressed embedding as personal data or anonymous derived data is unsettled, creating uncertainty that affects infrastructure investment decisions. Multinational platforms are thus forced to architect their token compression and retrieval pipelines with geo-fencing capabilities, duplicate inference nodes in multiple jurisdictions, and cryptographic protocols that provide verifiable assurances about the irreversibility of compression.

At a societal level, the adoption of aggressive token compression in public-facing video search services raises concerns about content transparency and archival fidelity. When a retrieval engine systematically drops visual tokens from its index to meet throughput targets, it performs a form of representational filtering that can shape collective memory and historical narratives. This structural effect becomes especially salient for archival institutions and journalism platforms that prioritize fidelity over speed. Governance models that separate the roles of content ingestion with full token preservation and retrieval serving with configurable compression could offer a balanced path, ensuring that high-fidelity representations remain accessible for evidentiary and cultural heritage purposes while everyday search traffic benefits from efficiency gains.

7. Conclusion

Efficient spatiotemporal token compression stands as both a computational necessity and a sociotechnical inflection point for the next generation of multimodal video retrieval systems. The landscape of compression techniques—encompassing pruning, merging, learned selection, and hybrid interleaved architectures—offers a spectrum of trade-offs that intersect with latency, storage, retrieval precision, and fairness. Infrastructure decisions regarding sharding, index construction, and carbon-aware scheduling amplify the downstream effects of token budget choices. As these systems grow to index billions of videos, the implications for environmental sustainability, data protection, and equitable access to information become inseparable from the core compression algorithm design.

We have argued that the evaluation of token compression must move beyond accuracy-throughput charts to incorporate fairness audits, robustness under distribution shift, lifecycle energy analysis, and compliance with emerging AI governance frameworks. The design of compressed representations that are simultaneously efficient, interpretable, and fairness-aware remains an open frontier. Progress will demand collaboration across computer vision, systems engineering, law, and public policy to construct retrieval infrastructures that uphold societal values while managing the relentless scale of video data. Only through such interdisciplinary integration can the research community ensure that token compression serves as an enabler of trustworthy and sustainable video retrieval rather than a hidden source of brittleness and inequity.

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. 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). PMLR.

3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In International Conference on Machine Learning (pp. 8748–8763). PMLR.

4. 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).

5. Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked autoencoders are data-efficient learners for self-supervised video pre-training. In Advances in Neural Information Processing Systems (Vol. 35, pp. 10078–10093).

6. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In Proceedings of the International Conference on Machine Learning (pp. 12888–12900). PMLR.

7. Xu, H., Ghosh, G., Huang, P.-Y., Okhonko, D., Aghajanyan, A., Metze, F., Zettlemoyer, L., & Feichtenhofer, C. (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). Association for Computational Linguistics.

8. Johnson, J., Douze, M., & Jégou, H. (2019). Billion-scale similarity search with GPUs. IEEE Transactions on Big Data, 7(3), 535–547.

9. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., & Hsieh, C.-J. (2021). DynamicViT: Efficient vision transformers with dynamic token sparsification. In Advances in Neural Information Processing Systems (Vol. 34, pp. 13937–13949).

10. Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., & Hoffman, J. (2023). Token merging: Your ViT but faster. In International Conference on Learning Representations.

11. Ryoo, M. S., Piergiovanni, A. J., Arnab, A., Dehghani, M., & Angelova, A. (2021). TokenLearner: Adaptive space-time tokenization for videos. In Advances in Neural Information Processing Systems (Vol. 34, pp. 12786–12797).

12. Feichtenhofer, C., Fan, H., Yang, F., & He, K. (2023). VideoMAE V2: Scaling video masked autoencoders with dual masking. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 14549–14560).

13. Wang, Y., Li, K., Li, Y., He, Y., Huang, B., Zhao, Z., Zhang, H., Xu, J., Liu, Y., Wang, Z., Xing, S., Chen, G., Pan, J., Feng, Y., Yu, J., Lu, F., Wang, S., Chen, K., & Sun, J. (2022). InternVideo: General video foundation models via generative and discriminative learning. arXiv preprint arXiv:2212.03191.

14. Feichtenhofer, C., Fan, H., Malik, J., & He, K. (2019). SlowFast networks for video recognition. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 6202–6211).

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

16. Patterson, D., Gonzalez, J., Le, Q., Liang, C., Munguia, L.-M., Rothchild, D., So, D., Texier, M., & Dean, J. (2021). Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350.

17. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (pp. 77–91). PMLR.

18. Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. Communications of the ACM, 64(12), 86–92.

19. European Commission. (2021). Proposal for a regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). COM(2021) 206 final.

20. European Parliament and Council of the European Union. (2016). Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data (General Data Protection Regulation). Official Journal of the European Union, L119, 1–88.

21. Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I. D., & Gebru, T. (2019). Model cards for model reporting. In Conference on Fairness, Accountability, and Transparency (pp. 220–229). PMLR.