

Adaptive Context Filtering for Efficient Vision-Language Model Inference

Manu Anton

School of Information Technology, University of Cincinnati, Cincinnati, OH, USA.
manuanton703@uc.edu

Nadja Kallert

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

Tijmen Gerritse

Department of Computer Science, University of North Texas, Denton, TX, USA.
tijmen.gerritse@unt.edu

Justyna Szylar

School of Computing, Clemson University, Clemson, SC, USA.
justynaszylar94@clemson.edu

Abstract

Vision-language models have become a central infrastructure for multimodal artificial intelligence, supporting applications that range from visual question answering and document understanding to robotics, accessibility tools, and scientific discovery. Despite their capabilities, these models remain expensive to run because they process large numbers of visual tokens alongside linguistic context, often invoking complex cross-modal attention and repeated memory transfers during inference. This paper examines adaptive context filtering as a system-level strategy for improving the efficiency of vision-language model inference without treating efficiency as a purely local optimization problem. Adaptive context filtering dynamically selects, compresses, reorders, or defers visual and textual context according to the query, the deployment environment, and the operational risk of the task. The paper argues that such filtering must be understood through architecture, infrastructure, governance, sustainability, fairness, and robustness rather than only through benchmark accuracy. It analyzes structural trade-offs between latency, memory bandwidth, energy consumption, and evidentiary completeness, and it discusses how filtering can reduce hallucination when guided by attention behavior, yet also introduce new failure modes when relevant evidence is discarded. The paper further considers policy implications for transparency, auditability, procurement, and carbon-aware deployment. Cross-domain case illustrations show that the same filtering mechanism can be beneficial in low-stakes search but dangerous in medical or autonomous settings unless uncertainty, fallback, and human oversight are designed into the system. The conclusion proposes that adaptive context filtering should be governed as a socio-technical capability, with evaluation that includes fairness, robustness, environmental cost, and accountability.

Keywords

adaptive context filtering; vision-language models; multimodal inference; efficient machine learning; system architecture; robustness; fairness; sustainability; AI governance.

1. Introduction

Vision-language models have rapidly moved from research prototypes into operational systems that interpret images, diagrams, user interfaces, sensor streams, and natural-language instructions in a single computational pipeline [1], [2], [3]. Their promise lies in the ability to connect visual evidence with linguistic reasoning, enabling tasks such as visual question answering, image captioning, document analysis, robotic manipulation, and assistive scene description [4], [5], [6], [7]. However, this multimodal capacity produces a distinctive form of computational burden. A vision-language model does not process only a short prompt; it often consumes hundreds or thousands of visual tokens, each of which participates in attention operations, memory movement, and cross-modal fusion. During inference, the cost is not limited to floating-point arithmetic. It includes memory bandwidth, cache pressure, kernel launch overhead, storage access, network transmission, and the energy required to sustain data center or edge hardware. As a result, efficient vision-language inference is a systems problem rather than a single-model problem.

Adaptive context filtering responds to this problem by treating context as a negotiable resource. Instead of passing every visual region, every text span, and every historical interaction through the full model, an adaptive system estimates which context is likely to matter for the current query and filters the remainder according to explicit or learned policies. Filtering may involve removing redundant visual tokens, merging similar patches, prioritizing regions associated with linguistic references, caching stable scene representations, or routing difficult queries to a more expensive execution path [8], [9], [10]. The central claim of this paper is that adaptive context filtering is not merely a compression technique. It is an architectural and governance choice that determines what evidence the model can see and what evidence it can ignore. That choice shapes accuracy, robustness, fairness, latency, cost, and public accountability.

This paper develops a long-form system-level analysis of adaptive context filtering for vision-language model inference. It begins with the technical foundations of multimodal transformers and the sources of inference inefficiency. It then presents a conceptual architecture for adaptive filtering, followed by an examination of efficiency trade-offs, robustness and hallucination, fairness and governance, deployment and sustainability, evaluation practices, and cross-domain case illustrations. The discussion is intentionally interdisciplinary because the deployment of vision-language models affects not only engineering metrics but also institutional trust, environmental policy, and social equity. The paper concludes that adaptive context filtering can be a valuable mechanism for efficient inference only when it is accompanied by uncertainty-aware fallback, transparent documentation, and evaluation that reflects the real conditions of use.

2. Foundations and Motivation

The technical foundation of modern vision-language models rests on transformer attention, which allows a model to weigh relationships among tokens across long sequences [1]. Vision transformers extended this logic to images by dividing them into patches and treating patch embeddings as tokens [2]. Contrastive image-text pretraining then demonstrated that visual and linguistic representations could be aligned in a shared embedding space, enabling zero-shot transfer and flexible downstream use [3]. Later multimodal architectures combined

frozen visual encoders with large language models, added cross-attention layers, or projected visual features into a language model’s token space [4], [5]. Instruction tuning further improved the ability of these systems to follow user requests and generalize across tasks [6], [7].

The motivating inefficiency arises because visual tokens are numerous and often highly redundant. A single high-resolution image may generate hundreds or thousands of patch tokens, and a video or multi-page document may generate far more. Many of those tokens contain background texture, repeated patterns, or information irrelevant to the query. Nevertheless, standard inference passes them through attention layers whose cost grows with sequence length. Sparse attention and adaptive representation methods have shown that not all tokens or positions require equal computation [11], [12]. In vision-language settings, however, the problem is more complex because relevance is determined jointly by visual content and linguistic intent. A region that seems unimportant in isolation may become critical when the query asks about a small object, a spatial relation, or a subtle textual label.

This complexity motivates adaptive context filtering. The goal is not to remove context arbitrarily but to allocate computation where the marginal value of additional context is highest. For example, a query about the color of a car may require only a small region around the vehicle, while a query about traffic safety may require pedestrians, signs, road boundaries, and temporal motion cues. A document question may require optical character recognition features in one column, while a chart question may require axis labels and legends. The filtering policy therefore has to estimate relevance, uncertainty, and risk. It also has to decide whether filtering should happen before the language model, inside cross-modal attention, or during decoding. These decisions determine whether the system saves computation or merely shifts it elsewhere.

3. Architecture of Adaptive Context Filtering

A system-level architecture for adaptive context filtering can be understood as four cooperating layers. The first layer is context estimation, which extracts candidate visual regions, text spans, metadata, and temporal segments. The second layer is relevance scoring, which uses query features, task priors, attention statistics, or lightweight classifiers to estimate which candidates should be retained. The third layer is execution policy, which determines whether to prune, merge, cache, defer, or route the filtered context to a smaller or larger model. The fourth layer is feedback and observability, which records what was filtered, why it was filtered, and how the final output performed. This architecture is not a single algorithm; it is a runtime discipline that connects model design, serving infrastructure, and monitoring.

Visual token reduction is one of the most direct forms of adaptive filtering. Prior work has shown that many vision-language models can discard a substantial fraction of visual tokens after early layers with limited accuracy loss, because later layers focus on a smaller set of semantically meaningful representations [8]. Other methods adaptively merge tokens or select key visual features based on attention or similarity [9]. Survey work on efficient multimodal models has catalogued pruning, quantization, distillation, and token merging as complementary strategies [10]. In an adaptive context filtering architecture, these methods become policy options rather than fixed model properties. A high-risk medical query might retain more tokens and use a conservative fallback, while a low-risk image search query might aggressively filter and accept a faster approximate answer.

The architecture also has to manage cross-modal alignment. If visual tokens are removed before the language model receives them, the remaining tokens must preserve the spatial and semantic relations needed for reasoning. Compositional tasks that depend on relations among objects are particularly sensitive to aggressive filtering [13]. Therefore, an effective system may maintain a compact but structured scene representation, such as object-level summaries, region graphs, or cached embeddings, rather than a simple reduced set of patches. It may also use a router that decides whether a query requires detailed visual grounding, broad scene understanding, or mostly linguistic reasoning. Foundation model research has emphasized that such systems are general-purpose infrastructures whose behavior is shaped by scale, data, and deployment context [14]. Adaptive filtering is therefore a way to specialize inference without retraining the entire model for every task.

4. Efficiency, Latency, and Resource Trade-offs

The primary appeal of adaptive context filtering is efficiency, but efficiency is multidimensional. Computational cost includes arithmetic operations, attention calculations, and decoding steps. Latency includes queueing delay, preprocessing, network transfer, and postprocessing. Memory cost includes model weights, key-value caches, activation buffers, and temporary visual embeddings. Energy cost depends on hardware utilization, data movement, cooling, and the carbon intensity of the electricity supply. A filtering policy that reduces arithmetic operations but increases data movement may not improve real-world efficiency. Similarly, a policy that lowers average latency but produces unpredictable worst-case latency may be unsuitable for interactive or safety-critical systems. Sparse attention and adaptive input representations show that computational savings depend on matching the method to the hardware and sequence structure [11], [12].

Batching and serving introduce additional trade-offs. In cloud inference, large batches improve accelerator utilization but can increase tail latency for individual users. Adaptive filtering can make batches heterogeneous because different queries retain different amounts of context. If not managed carefully, this heterogeneity can reduce hardware efficiency and increase scheduling complexity. At the edge, memory bandwidth and thermal limits may dominate, making aggressive filtering more valuable than raw arithmetic reduction. A deployment may therefore use different filtering policies in different tiers: a strict policy on mobile devices, a balanced policy on edge servers, and a permissive policy in data centers. The architecture must expose these policies as configurable parameters rather than hidden model behavior.

There are also organizational trade-offs. Machine learning systems accumulate hidden technical debt when data dependencies, configuration, and feedback loops are poorly documented [23]. Adaptive filtering adds another layer of configuration that can drift over time. If a filtering threshold is tuned for one dataset, it may fail silently when the input distribution changes. If logs do not record which context was removed, debugging becomes difficult. Efficiency gains can therefore be undermined by operational opacity. A mature system needs versioned filtering policies, reproducible evaluation, and clear ownership of decisions about what context the model is allowed to discard.

5. Robustness, Hallucination, and Reliability

Filtering directly affects the evidence available to a vision-language model, and therefore affects hallucination. When relevant visual evidence is removed, the language model may fill gaps with plausible but incorrect text. Recent work on vocabulary hijacking in large vision-

language models has examined critical attention heads and the role of inert tokens in mitigating hallucination, suggesting that not all tokens contribute equally to factual grounding and that attention behavior can reveal which representations matter [15]. This insight supports adaptive filtering that preserves evidence-bearing tokens while excluding inert or distracting ones. However, the same insight also warns that filtering policies based on superficial salience may remove tokens that are statistically quiet but semantically decisive. A small sign, a subtle lesion, or a faint annotation may not dominate attention yet may determine the correct answer.

Robustness must therefore be evaluated under distribution shift, adversarial perturbation, and ambiguous queries. A system that filters aggressively in ordinary scenes may fail when lighting, occlusion, or domain-specific appearance changes. Safety engineering for artificial intelligence emphasizes the need to anticipate specification failures, reward hacking, and negative side effects in deployed systems [24]. In vision-language inference, a negative side effect can be a confident answer produced without the visual evidence needed to justify it. Foundation model research similarly warns that general-purpose systems can exhibit emergent behaviors and unanticipated failure modes when deployed outside their training distribution [14]. Adaptive filtering should therefore include uncertainty estimation, abstention, and fallback to unfiltered or higher-cost inference when confidence is low.

Reliability also depends on temporal consistency. In video or interactive settings, a model may filter different regions across frames, causing flickering, inconsistency, or unstable reasoning. A robust system may maintain persistent object tracks, scene memory, or key-frame summaries so that filtering is coherent over time. It may also separate filtering for perception from filtering for reasoning. Perception may require broad spatial coverage, while reasoning may require only a few grounded entities. By distinguishing these stages, the system can preserve safety-relevant context while reducing unnecessary language-model computation.

6. Fairness, Governance, and Policy

Adaptive context filtering can reproduce or amplify social bias. If a filtering policy is trained on data that underrepresents certain skin tones, clothing styles, disabilities, dialects, or cultural contexts, it may systematically discard visual evidence associated with those groups. Foundational work on gender and skin-type classification demonstrated that commercial systems can exhibit intersectional accuracy disparities [16]. Broader surveys of bias in machine learning show that such disparities often arise from data, objective functions, and deployment contexts rather than from a single faulty component [17]. In vision-language systems, filtering introduces another site where bias can enter: the decision about which visual regions and textual cues are considered relevant. A query about a person in a wheelchair, a religious garment, or a non-Western street scene may be poorly served if the relevance model has not learned to treat those features as important.

Governance frameworks for artificial intelligence emphasize transparency, accountability, fairness, privacy, and human oversight [20], [25]. Adaptive filtering complicates these principles because the model’s output may depend on context that was never processed. If a system cannot explain which regions were filtered and why, affected users cannot contest a decision. If a public agency procures such a system, it needs contractual guarantees about logging, auditing, and redress. Principles alone are insufficient without institutional mechanisms that assign responsibility and enable enforcement [21]. A governance-aware filtering architecture should therefore produce auditable records of retained and discarded

context, preserve privacy by avoiding unnecessary storage of sensitive regions, and provide human-review pathways for high-stakes decisions.

Human-centered design offers a complementary perspective. Systems should be designed around human goals, controllability, and responsibility rather than only around model-centric benchmarks [22]. In practice, this means allowing users or operators to adjust filtering aggressiveness, inspect evidence, or request a more thorough analysis. It also means documenting filtering behavior in model cards, system cards, or deployment reports. Policy implications extend to procurement standards, safety certification, and sector-specific regulation. A medical imaging tool, for example, may need to preserve all clinically relevant regions and provide uncertainty estimates, while an image search tool may legitimately prioritize speed. The same technical mechanism can be acceptable in one context and unacceptable in another.

7. Deployment, Infrastructure, and Sustainability

Deployment environments shape the value of adaptive context filtering. In cloud data centers, the dominant concerns may be throughput, accelerator utilization, and cost per query. At the edge, concerns include battery life, thermal throttling, intermittent connectivity, and privacy. In hybrid architectures, filtering can decide whether a query is processed locally, sent to a regional server, or escalated to a central model. This tiered approach can reduce network transmission and improve response time, but it also creates consistency challenges. If different tiers use different filtering policies, the same query may produce different answers depending on where it is executed. Governance must therefore address service-level expectations and user awareness.

Sustainability is a central but often neglected dimension. Training large models has significant energy and carbon costs [18], and inference at scale can become a dominant operational expense. Green AI research argues for reporting efficiency, energy use, and carbon emissions alongside accuracy [19]. Adaptive context filtering can contribute to sustainability by reducing unnecessary computation, but it can also encourage more frequent use or more complex models if efficiency gains are reinvested in additional capability. This rebound effect means that efficiency improvements do not automatically reduce total environmental impact. Carbon-aware scheduling, hardware-aware filtering, and transparent energy reporting are needed to connect filtering policy to sustainability outcomes.

Operational infrastructure also includes monitoring, logging, and incident response. Hidden technical debt arises when model behavior depends on fragile data pipelines and undocumented configurations [23]. Filtering policies should be versioned, tested, and rolled back when failures occur. Trustworthy artificial intelligence frameworks emphasize robustness, privacy, explainability, and accountability across the lifecycle [25]. For adaptive filtering, this means treating the filtering policy as a governed artifact. It should have an owner, a change log, an evaluation suite, and a process for handling drift. Without such infrastructure, efficiency gains may be achieved at the cost of reliability and institutional trust.

8. Evaluation, Benchmarks, and Operational Metrics

Evaluating adaptive context filtering requires more than top-line accuracy. Standard vision-language benchmarks measure tasks such as image captioning, visual question answering, and retrieval, but they may not reveal when a model succeeds for the wrong reason or fails on a rare but important subpopulation [3], [13]. A filtering system should be evaluated on accuracy, calibration, robustness, fairness, latency, memory use, and energy consumption.

Compositional benchmarks are especially important because they test whether the model preserves relations among objects, attributes, and actions after filtering [13]. If a model can answer simple object-presence questions but fails on spatial or relational queries, the filtering policy may be removing context that supports compositional reasoning.

Operational metrics should reflect deployment realities. Average latency is less useful than tail latency for interactive systems. Throughput is less useful than cost per successful task when errors require human correction. Energy per query and energy per correct answer reveal whether efficiency gains are meaningful. Token retention ratio, region recall, and evidence coverage can help diagnose whether filtering is too aggressive. Fairness metrics should be disaggregated by relevant groups and contexts, following the lesson that aggregate accuracy can hide intersectional disparities [16]. Qualitative evaluation by domain experts is also necessary in medical, legal, industrial, and accessibility settings.

Human-centered evaluation should include whether users understand the system’s limitations and can contest its outputs [22]. If a user is not told that the system filtered most of an image, they may over-trust a confident answer. If an operator cannot inspect which regions were discarded, they cannot diagnose a failure. Evaluation should therefore include transparency, controllability, and redress. Benchmarks may need to be paired with deployment simulations that vary network conditions, hardware tiers, privacy constraints, and risk thresholds. The goal is not to find a single optimal filtering policy but to characterize the operating envelope within which a policy is safe and useful.

9. Case Illustrations and Cross-Domain Comparisons

In medical imaging, adaptive context filtering could reduce the cost of analyzing large radiology studies, but it must be conservative. A model that filters out a small nodule or subtle fracture may produce a dangerous false negative. Here, filtering should be used to prioritize regions for detailed analysis rather than to remove them permanently from the reasoning process. Uncertainty estimates and human review are essential. In industrial inspection, filtering may be more acceptable when defects are well-defined and camera positions are controlled, but rare defect classes still require broad coverage. In autonomous driving, filtering must preserve pedestrians, signs, road boundaries, and temporal motion cues; latency savings cannot come at the cost of safety. In education and accessibility, filtering may help describe scenes quickly, but it must avoid omitting information that users need to navigate or understand social context.

Cross-domain comparison shows that the value of filtering depends on the cost of error and the predictability of the task. Low-stakes image search can tolerate aggressive filtering because a missed detail can be corrected by a follow-up query. High-stakes decision support cannot rely on the same policy. Vision-language architectures differ as well. Some models use frozen encoders and lightweight projection layers, making token reduction relatively easy [5]. Others use cross-attention or instruction-tuned fusion, where filtering interacts with language generation in more complex ways [6], [7]. Efficient multimodal surveys suggest that no single method dominates across all settings [10]. A system that works well for short queries may fail for long documents or video. A system optimized for data centers may be unsuitable for mobile devices.

The lesson from these cases is that adaptive context filtering should be designed as a risk-aware service. It should expose different operating modes, preserve evidence for high-stakes tasks, and record decisions for audit. It should also be evaluated in the domain where it will

be used, not only on general benchmarks. A filtering policy that improves average efficiency but increases rare catastrophic errors is not a successful system. The architecture must therefore connect technical performance to institutional tolerance for risk. This connection is not automatic; it requires deliberation among engineers, domain experts, users, and policymakers.

10. Future Directions

Future research should develop filtering policies that are learned but interpretable. Learned routers and attention-based selectors can adapt to query complexity [9], [11], but they must provide explanations for why context was discarded. Causal evaluation could help determine whether retained context is sufficient for correct reasoning rather than merely correlated with it. Uncertainty-aware filtering could decide when to abstain, request more context, or escalate to a larger model. Temporal filtering for video and embodied agents could maintain coherent scene memory while reducing per-frame computation. These directions require new benchmarks that measure evidence coverage, calibration, and robustness under distribution shift.

Governance research should explore standards for documenting filtering behavior. Foundation model transparency initiatives have shown that capability claims are difficult to verify without standardized reporting [14]. Professional and policy frameworks for artificial intelligence ethics emphasize accountability, privacy, and human oversight [20]. Adaptive filtering should be included in model cards, system cards, and procurement requirements. Auditors may need access to filtered-context logs, though privacy constraints may limit retention. Federated evaluation and third-party auditing could help verify fairness and robustness without centralizing sensitive data. Regulatory approaches may differ by sector, but common principles such as traceability, contestability, and safety margins are broadly applicable.

Safety research should examine failure modes specific to filtering. If a model is trained to rely on filtered context, it may become overconfident when the filter removes critical evidence [24]. Adversarial attacks could target the filtering policy itself, causing a system to ignore a stop sign, a warning label, or a medical indicator. Defenses may include redundancy, multi-view filtering, and fallback to unfiltered inference under uncertainty. The long-term goal should be not maximum compression but adaptive allocation of computation according to task value and risk. Efficient inference is valuable only when it preserves the reliability and legitimacy of the overall system.

11. Conclusion

Adaptive context filtering offers a promising path toward efficient vision-language model inference by treating visual and textual context as a resource to be allocated rather than a burden to be processed uniformly. It can reduce latency, memory pressure, energy consumption, and cost, especially when visual tokens are redundant or irrelevant to the query. Yet the same mechanism can remove evidence, amplify bias, obscure accountability, and create new safety risks. This paper has argued that adaptive context filtering must therefore be understood as a socio-technical capability. Its architecture must include relevance estimation, execution policy, fallback, observability, and governance. Its evaluation must include robustness, fairness, calibration, sustainability, and human oversight alongside accuracy. Its deployment must respect domain-specific risk tolerances and provide transparent documentation and redress. When designed with these commitments, adaptive context filtering can support efficient multimodal inference without sacrificing the evidentiary and

ethical standards that responsible artificial intelligence requires. When designed only for speed, it risks making vision-language systems faster but less trustworthy.

References

1. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems* (Vol. 30, pp. 5998-6008). Curran Associates.
2. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.
3. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning* (Vol. 139, pp. 8748-8763). PMLR.
4. Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J., Borgeaud, S., ... Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 23716-23736). Curran Associates.
5. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning* (Vol. 202, pp. 19730-19742). PMLR.
6. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 34892-34916). Curran Associates.
7. Dai, W., Li, J., Li, D., Tiong, A. M. H., Zhao, J., Wang, W., Li, B., Fung, P., & Hoi, S. (2023). InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 49250-49267). Curran Associates.
8. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., & Chang, B. (2024). An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision* (pp. 19-35). Springer.
9. Shang, Y., Cai, B., Xu, G., Lee, Y. J., & Yan, Y. (2024). LLaVA-PruMerge: Adaptive token reduction for efficient large multimodal models. *arXiv preprint arXiv:2403.15388*.
10. Jin, Y., Li, J., Liu, Y., Qin, T., Wang, Y., & Liu, T. (2024). Efficient multimodal large language models: A survey. *arXiv preprint arXiv:2405.10739*.
11. Child, R., Gray, S., Radford, A., & Sutskever, I. (2019). Generating long sequences with sparse transformers. *arXiv preprint arXiv:1904.10509*.
12. Baevski, A., & Auli, M. (2019). Adaptive input representations for neural language modeling. In *International Conference on Learning Representations*.
13. Thrush, T., Jiang, R., Bartolo, M., Singh, A., Williams, A., Kiela, D., & Ross, C. (2022). Winoground: Probing vision and language models for visio-linguistic compositionality.

In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 5238-5248). IEEE.

14. Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. arXiv preprint arXiv:2108.07258.
15. Chen, Y., Li, J., Yao, W., Ma, X., Du, G., Wang, W., & Li, J. (2026, July). Vocabulary hijacking in lvm: Unveiling critical attention heads by excluding inert tokens to mitigate hallucination. In Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 38458-38485).
16. Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In Conference on Fairness, Accountability and Transparency (Vol. 81, pp. 77-91). PMLR.
17. Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2021). A survey on bias and fairness in machine learning. *ACM Computing Surveys*, 54(6), 1-35.
18. Strubell, E., Ganesh, A., & McCallum, A. (2019). Energy and policy considerations for deep learning in NLP. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3645-3650). Association for Computational Linguistics.
19. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54-63.
20. Floridi, L., & Cowls, J. (2019). A unified framework of five principles for AI in society. *Harvard Data Science Review*, 1(1).
21. Mittelstadt, B. (2019). Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence*, 1(11), 501-507.
22. Shneiderman, B. (2020). *Human-centered AI*. Oxford University Press.
23. Sculley, D., Holt, G., Golovin, D., Davydov, E., Phillips, T., Ebner, D., Chaudhary, V., Young, M., Crespo, J.-F., & Dennison, D. (2015). Hidden technical debt in machine learning systems. In *Advances in Neural Information Processing Systems* (Vol. 28, pp. 2503-2511). Curran Associates.
24. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. arXiv preprint arXiv:1606.06565.
25. Kaur, D., Uslu, S., Rittichier, K. J., & Durresi, A. (2022). Trustworthy artificial intelligence: A review. *ACM Computing Surveys*, 55(2), 1-38.