

Knowledge-Enhanced Vision-Language Models for Accurate Image Understanding

Temistocle Mortati

Department of Computer Science, Binghamton University, Binghamton, NY, USA.
mortati580@binghamton.edu

Hans Jorg Kohler

Department of Computer Science, University of New Hampshire, Durham, NH, USA.
hansjorgk@unh.edu

Gaby Butte

Department of Electrical Engineering and Computer Science, University of Kansas, Lawrence, KS, USA.
butte1990@ku.edu

Abstract

Vision-language models have become central to image understanding because they align visual perception with natural language reasoning, yet their accuracy remains uneven in domains where factual precision, spatial relations, and context-sensitive meaning are essential. This paper examines knowledge-enhanced vision-language models as socio-technical systems rather than isolated model architectures. It argues that accurate image understanding depends on the combined design of multimodal encoders, structured and unstructured knowledge sources, retrieval infrastructure, adaptation mechanisms, governance instruments, and evaluation regimes. The analysis first reviews architectural foundations, including transformer-based fusion, contrastive pretraining, generative vision-language modeling, and parameter-efficient adaptation. It then considers how external knowledge, knowledge graphs, retrieval augmentation, and instruction tuning can improve grounding while introducing latency, maintenance, and provenance trade-offs. The paper also explores deployment sustainability, robustness, fairness, and policy implications, emphasizing that knowledge enhancement is not merely a technical upgrade but an institutional commitment. Through cross-domain comparisons in healthcare, industrial inspection, public administration, and embodied systems, the discussion shows that accurate image understanding requires traceable evidence, calibrated uncertainty, accessible documentation, and continuous monitoring. The conclusion contends that the most reliable knowledge-enhanced systems will be those that treat knowledge as a governed infrastructure, not as a static database attached to a model. Such systems must balance accuracy with accountability, performance with sustainability, and automation with human oversight. The paper contributes a system-level research agenda for designing vision-language models that are accurate, robust, fair, and governable across high-stakes visual environments.

Keywords

vision-language models, knowledge enhancement, image understanding, retrieval augmentation, multimodal systems, governance, robustness, fairness, deployment sustainability.

1. Introduction

Vision-language models have rapidly moved from laboratory demonstrations to operational infrastructures that classify images, answer visual questions, generate captions, inspect defects, and support embodied decision-making. Their appeal lies in the promise of a shared representational space in which visual content and natural language meaning can be aligned through large-scale pretraining [2]. Early contrastive systems demonstrated that image-text pairs could support zero-shot recognition and flexible prompting, while later generative and multimodal architectures extended these capabilities toward open-ended visual reasoning and dialogue [4], [5]. At the same time, the growing dependence on such models has exposed a persistent problem: fluent outputs are not necessarily accurate outputs. A model may describe a scene plausibly while missing a critical object, misidentifying a relation, or importing irrelevant linguistic priors into visual interpretation. This gap between fluency and factual precision motivates the turn toward knowledge-enhanced vision-language models.

Knowledge enhancement refers to the deliberate integration of external evidence, structured representations, domain ontologies, retrieval mechanisms, memory stores, and human-curated resources into the multimodal reasoning process. The rationale is straightforward. A model trained only on broad image-text corpora may learn general associations but lack the specificity required for clinical imaging, industrial safety, legal evidence, or scientific observation. Retrieval-augmented approaches have shown that language models can benefit from non-parametric memory and external documents [7], while earlier work on retrieval-augmented pretraining demonstrated the value of learned access to knowledge sources [8]. In parallel, knowledge graph embeddings and relational representations provide compact ways to encode entities, relations, and constraints that are difficult to infer from raw pixels alone [9]. When these resources are coupled with vision-language architectures, the system can potentially condition image interpretation on verified facts, domain rules, and context-specific taxonomies.

However, knowledge enhancement introduces structural trade-offs that cannot be reduced to model accuracy alone. External knowledge must be acquired, cleaned, versioned, indexed, secured, and updated. Retrieval adds latency and failure modes. Structured knowledge may be incomplete or biased. Domain ontologies may conflict across institutions. Governance becomes necessary because the system’s answer depends on sources that may be proprietary, outdated, or contested. Foundation model research has already emphasized that such systems concentrate technical, economic, and social risks, including opacity, dependency, and environmental cost [10]. Accurate image understanding therefore requires a system-level perspective that treats knowledge, model, infrastructure, and policy as co-designed elements.

This paper develops that perspective. It asks how knowledge-enhanced vision-language models can be designed for accurate image understanding without sacrificing robustness, fairness, sustainability, or accountability. The discussion is organized around architectural foundations, integration mechanisms, governance and deployment, robustness and fairness, case illustrations, and policy implications. A central claim is that knowledge enhancement is most effective when knowledge is treated as governed infrastructure rather than as an auxiliary database. This means that provenance, version control, access control, documentation, evaluation, and human oversight must be designed into the system from the beginning. The paper also situates hallucination mitigation within broader reliability engineering, noting that attention-level interventions and token-exclusion strategies can complement, but not replace, institutional verification [11].

2. Architectural Foundations of Knowledge-Enhanced Vision-Language Systems

The architectural basis of modern vision-language models rests on transformer networks, which allow attention mechanisms to model long-range dependencies across modalities [1]. In contrastive image-text models, separate encoders are trained to align matching pairs in a shared embedding space, enabling flexible zero-shot transfer and retrieval [2]. Other architectures reduce reliance on region supervision or convolutional pipelines by processing visual patches and text tokens within unified transformer stacks [3]. Generative approaches then combine visual encoding with language decoding, allowing the model to produce captions, answers, and descriptions rather than only similarity scores [4]. Few-shot multimodal systems further demonstrate that a single model can condition on interleaved images and text, supporting tasks that require in-context visual reasoning [5]. These developments provide the representational foundation upon which knowledge enhancement can be layered.

Knowledge enhancement can enter the architecture at several points. It may shape pretraining through weakly supervised image-text objectives [12], it may be injected during instruction tuning through domain-specific visual question-answer data [14], or it may be attached at inference time through retrieval modules that supply documents, tables, or structured facts [7]. Parameter-efficient adaptation methods such as low-rank adaptation allow domain knowledge to be incorporated without full retraining, which is important when institutions cannot afford repeated large-scale updates [6]. More recent multimodal systems combine frozen image encoders with large language models, creating modular pipelines in which visual representations are projected into a language model’s reasoning space [13]. This modularity is attractive for knowledge enhancement because retrieval, memory, and domain adapters can be changed independently of the core visual encoder.

Yet modularity also creates integration risk. If the visual encoder is frozen and the language model is adapted, errors may arise from representational mismatch rather than from insufficient knowledge. If retrieval is added late, the model may ignore retrieved evidence when linguistic priors are strong. If structured knowledge is converted into text before being passed to the language model, relational constraints may be flattened into prose, losing formal properties that a knowledge graph could preserve [9]. If knowledge is stored in an external vector index, the system must maintain embedding consistency across model versions. These issues are not merely engineering details; they determine whether knowledge enhancement improves accuracy or merely increases architectural complexity. The most robust designs therefore specify clear interfaces between perception, retrieval, reasoning, and verification, and they treat each interface as a site of potential failure.

A further architectural consideration is the distinction between parametric and non-parametric knowledge. Parametric knowledge is encoded in model weights and is difficult to audit or update. Non-parametric knowledge resides in external stores, making it easier to version, inspect, and correct. Retrieval-augmented generation illustrates this advantage by separating the language model from a document index [7], and retrieval-augmented pretraining shows that models can learn to query external evidence during training [8]. For image understanding, non-parametric knowledge may include product manuals, diagnostic guidelines, safety regulations, geospatial metadata, or prior inspection reports. The trade-off is that retrieval introduces dependence on index quality, query formulation, and ranking. A knowledge-enhanced vision-language system must therefore be evaluated not only by answer accuracy but also by retrieval precision, evidence attribution, latency, and failure recovery.

3. Knowledge Integration Mechanisms and Structural Trade-offs

The first major integration mechanism is structured knowledge representation. Knowledge graphs encode entities and relations in ways that support symbolic queries and constraint checking [9]. In image understanding, a knowledge graph might represent anatomical structures, machine components, legal categories, or supply-chain relationships. When a vision-language model identifies a visual entity, the graph can supply additional relations that are not visible in the image, such as known failure modes, compatible parts, or regulatory classifications. This can improve accuracy when the image alone is ambiguous. However, structured knowledge requires maintenance. Ontologies drift, entities multiply, and relation labels may be interpreted differently across communities. A system that depends on a knowledge graph inherits the graph’s coverage gaps and curation biases.

The second mechanism is retrieval augmentation. By storing documents, reports, or examples in an external index, the system can retrieve context relevant to the current image and question [7]. Retrieval is particularly useful for long-tail visual domains where pretraining data are sparse. A maintenance engineer inspecting a corroded pipe may benefit from retrieved inspection histories, material specifications, and safety codes. A physician interpreting a rare presentation may benefit from retrieved case notes and guidelines. Yet retrieval quality depends on the query generated from the image and question, the embedding model used for indexing, and the ranking strategy. If the retrieved evidence is irrelevant, the language model may still produce a confident answer. If the evidence is relevant but contradictory, the system needs a conflict-resolution policy. These challenges make retrieval a governance issue as much as an algorithmic one.

The third mechanism is instruction tuning and dialogue-based grounding. Visual instruction tuning trains models to follow natural language instructions about images, which improves controllability and task generalization [14]. In knowledge-enhanced settings, instruction tuning can teach the model to request evidence, cite sources, express uncertainty, or defer to a human expert. Embodied multimodal systems extend this logic to robotic and interactive settings, where visual understanding must support planning and action [15]. However, instruction tuning can also encode the norms of its training data. If the instructions reward fluency over verifiability, the model may learn to hide uncertainty. If they reward compliance over caution, the model may produce unsafe recommendations. The design of instruction data is therefore a policy instrument, not only a technical dataset.

The fourth mechanism is parameter-efficient adaptation. Low-rank adaptation and related methods allow a model to be specialized for a domain without modifying all weights [6]. This is valuable for sustainability because it reduces compute and storage requirements. It also supports institutional autonomy: a hospital, factory, or public agency can maintain a domain adapter without retraining a foundation model. Yet adapters can fragment accountability. If multiple adapters exist for different departments, which one produced a given answer? If an adapter is updated, how are prior outputs audited? If adapters are traded or reused, how are their training data documented? These questions connect directly to datasheets and model cards, which provide structured documentation for datasets and models [16], [17]. Knowledge enhancement through adapters therefore requires metadata standards that travel with the adapter, not only with the base model.

Across these mechanisms, the central trade-off is between accuracy and governability. More knowledge sources can improve factual precision, but they also increase system complexity, latency, and attack surface. More structured knowledge can support constraint checking, but it may be brittle when the world is ambiguous. More retrieval can reduce hallucination, but it

may introduce irrelevant or malicious content. More adaptation can improve domain fit, but it can complicate versioning and oversight. A system-level approach must therefore optimize not for a single accuracy metric but for a portfolio of properties: factual correctness, provenance, calibration, auditability, fairness, energy cost, and recoverability. This portfolio view aligns with broader critiques of large models, which warn that scale alone does not resolve social and epistemic risks [10], [18].

4. Governance, Infrastructure, and Deployment Sustainability

Governance begins with documentation. Datasheets for datasets encourage creators to record motivation, composition, collection processes, and recommended uses [16]. Model cards provide analogous reporting for models, including intended use, performance characteristics, and limitations [17]. In knowledge-enhanced vision-language systems, documentation must extend further to knowledge sources. A retrieval index should have a source register that records ownership, update frequency, licensing, coverage, and known gaps. A knowledge graph should have an ontology charter that explains entity definitions, relation semantics, and conflict resolution. An adapter should have a training data statement that describes domain data, annotation procedures, and evaluation results. Without such documentation, accurate image understanding may be reproducible only within the original team, which undermines institutional trust.

Infrastructure is equally important. Knowledge-enhanced models depend on compute, storage, network access, and data pipelines. If retrieval is centralized, latency and availability become critical. If knowledge is distributed across institutions, interoperability standards become critical. If models are deployed at the edge, energy and memory constraints may limit the size of the knowledge store. Technical debt in machine learning systems arises from entangled dependencies, undeclared consumers, configuration drift, and hidden feedback loops [20]. Knowledge-enhanced vision-language systems are especially vulnerable to such debt because they combine multiple fast-moving components: visual encoders, language models, vector databases, ontologies, adapters, and evaluation scripts. Sustainable deployment requires lifecycle management, not only initial integration.

Deployment sustainability also involves environmental and economic cost. Large multimodal models require substantial training and inference resources. Knowledge enhancement can reduce the need for retraining by allowing external updates, but retrieval and larger context windows increase inference cost. Organizations must therefore decide where knowledge should reside: in model weights, in adapters, in retrieval indexes, or in human workflows. A system that retrieves every possible document may be accurate but unsustainable. A system that compresses all knowledge into weights may be efficient but unauditably. The practical path is tiered knowledge architecture, where frequently used, stable knowledge is cached or parameterized, while volatile or high-stakes knowledge is retrieved and verified in real time. This tiering should be evaluated against service-level objectives that include accuracy, latency, cost, and carbon intensity.

Policy implications follow from these governance and infrastructure choices. Public-sector deployment may require procurement rules that demand provenance, audit logs, and human review. Healthcare deployment may require compliance with privacy and clinical validation standards. Industrial deployment may require safety certification and traceability. In each case, the knowledge-enhanced system must be able to explain which sources informed an output. This is not merely a technical feature; it is a condition for accountability. The taxonomy of risks posed by language models emphasizes harms such as misinformation, discrimination,

privacy violation, and malicious use [19]. Vision-language systems add risks related to surveillance, misidentification, and automated interpretation of visual evidence. Governance must therefore include access control, purpose limitation, retention limits, and redress mechanisms.

5. Robustness, Fairness, and Evaluation in Socio-Technical Settings

Robustness in knowledge-enhanced vision-language models means more than resistance to adversarial perturbations. It includes robustness to distribution shift, missing knowledge, contradictory evidence, corrupted retrieval, and ambiguous queries. A model may perform well on benchmark images but fail when lighting, camera angle, or cultural context changes. A retrieval index may return outdated guidance. A knowledge graph may lack entities for a minority population. An attention-head intervention may reduce certain hallucination patterns, but it cannot guarantee factual accuracy across domains [11]. Robustness engineering therefore requires red-teaming, stress testing, ablation studies, and continuous monitoring in deployment. It also requires graceful degradation: when evidence is insufficient, the system should express uncertainty or defer to a human rather than generate a confident answer.

Fairness is inseparable from robustness because visual and textual data are socially structured. Image datasets may underrepresent certain skin tones, body types, architectural styles, or geographic regions. Language data may encode stereotypes that influence captioning and visual question answering. Retrieval indexes may privilege dominant languages and institutional perspectives. Knowledge graphs may formalize categories in ways that exclude or misclassify marginalized groups. These biases can compound across modalities. A knowledge-enhanced system that retrieves historical documents may reproduce historical prejudice. A model card can document known limitations, but documentation alone does not correct harm [17]. Fairness requires participatory design, disaggregated evaluation, and mechanisms for affected communities to contest outputs. It also requires attention to the political economy of knowledge: whose documents are indexed, whose ontologies are authoritative, and whose interpretations are treated as ground truth.

Evaluation must correspondingly move beyond single accuracy scores. Useful evaluation regimes include factual correctness, evidence attribution, calibration, abstention behavior, subgroup performance, latency, cost, and user trust. In high-stakes settings, evaluation should include human-in-the-loop workflows and incident reporting. The stochastic parrots critique warns that large language models can amplify bias and environmental harm while appearing authoritative [18]. Vision-language models intensify this risk because images are often treated as objective evidence. A model that mislabels a person, object, or event can have legal, medical, or social consequences. Evaluation should therefore include adversarial scenarios, ambiguous cases, and cross-domain transfer tests. It should also examine whether the system improves human decision-making or merely shifts responsibility to an opaque model.

A key evaluation challenge is the provenance of knowledge. If a model cites a retrieved document, is the citation faithful? If a knowledge graph supports an inference, is the relation valid? If an adapter specializes a model, does it preserve general capabilities? These questions require traceable pipelines and audit logs. They also require standardized benchmarks that include source metadata. Without such standards, accuracy claims are difficult to compare across systems. The field needs evaluation suites that test not only what the model knows but also how it knows, how it handles uncertainty, and how it behaves when knowledge is contested. This is especially important for policy, where decisions may affect rights, safety, and resource allocation.

6. Case Illustrations and Cross-Domain Comparisons

Healthcare imaging illustrates both the promise and the difficulty of knowledge enhancement. A vision-language model might interpret a radiology image, retrieve clinical guidelines, and generate a preliminary report. Knowledge graphs could encode anatomical relations, diagnostic criteria, and contraindications. Instruction tuning could teach the model to cite evidence and flag uncertainty [14]. However, patient data are sensitive, guidelines change, and errors can harm. The system must integrate privacy-preserving infrastructure, clinical validation, and human oversight. A retrieved document may be authoritative but not applicable to the patient. A knowledge graph may be incomplete for rare conditions. The value of knowledge enhancement lies in supporting clinicians, not replacing clinical judgment. Evaluation must therefore measure sensitivity, specificity, calibration, and workflow impact.

Industrial inspection offers a different configuration. A vision-language system may examine images of equipment, retrieve maintenance histories, and recommend actions. Knowledge graphs can represent component hierarchies, failure modes, and safety procedures. Edge deployment may require low latency and offline operation. Here, the trade-off between retrieval and local caching becomes prominent. Frequently used safety rules can be stored locally, while rare incidents can be retrieved from a central repository when connectivity allows. Model cards and datasheets can document known limitations [16], [17]. Technical debt can accumulate if sensors, models, and knowledge bases are updated independently [20]. Sustainable deployment requires maintenance schedules for data, models, and ontologies, not only for physical equipment.

Public administration and civic image understanding raise governance questions. A system might analyze satellite imagery, traffic footage, or infrastructure photos to support planning, disaster response, or code enforcement. Knowledge enhancement can incorporate zoning laws, demographic data, and historical records. Yet such systems can enable surveillance, discrimination, and mission creep. The taxonomy of risks for language models identifies misinformation, privacy, and malicious use as central concerns [19]. Vision-language systems add the risk of automated visual classification being treated as factual. Governance must therefore include purpose limitation, transparency, appeal mechanisms, and restrictions on secondary use. Cross-domain comparison shows that the same architectural choice, such as retrieval augmentation, has different ethical weight depending on context. In industrial inspection, a false negative may cause equipment failure. In civic surveillance, a false positive may violate civil liberties.

Embodied and robotic systems provide another comparative lens. Multimodal models can support navigation, manipulation, and human-robot interaction by combining visual perception with language instructions [15]. Knowledge enhancement can supply maps, object affordances, and safety constraints. However, embodied systems operate in dynamic environments where errors have physical consequences. Robustness requires real-time uncertainty estimation and fail-safe behaviors. Fairness concerns may arise when robots are deployed in public spaces or care settings. Sustainability concerns include energy consumption and hardware lifecycle. These cases show that knowledge-enhanced vision-language models are not a single technology but a family of socio-technical arrangements. Their accuracy depends on how knowledge is curated, how inference is bounded, and how accountability is allocated.

7. Policy Implications and Forward-Looking Perspectives

Policy for knowledge-enhanced vision-language models should address procurement, transparency, liability, and public participation. Procurement can require documentation of knowledge sources, evaluation results, and update procedures. Transparency can require user-facing explanations of evidence and uncertainty. Liability can clarify responsibility when a model, knowledge source, or human operator contributes to harm. Public participation can ensure that affected communities have a voice in deployment decisions, especially in surveillance, healthcare, and policing. These measures are not obstacles to innovation; they are conditions for trustworthy adoption. Foundation model research has shown that risks are distributed across the lifecycle, from data collection to deployment [10]. Governance must therefore be lifecycle-oriented rather than focused only on model release.

A forward-looking research agenda should prioritize interoperable knowledge infrastructures, provenance-aware multimodal pretraining, uncertainty calibration, and human-AI teaming. Interoperability requires standard metadata for ontologies, retrieval indexes, and adapters. Provenance-aware pretraining could teach models to associate visual features with evidence types rather than only with labels. Uncertainty calibration should distinguish between perceptual ambiguity, knowledge gaps, and reasoning failures. Human-AI teaming should design interfaces that reveal evidence, invite contestation, and support escalation. These directions align with calls for datasheets and model cards, but they extend them to knowledge artifacts [16], [17]. They also connect to broader concerns about bias, environmental cost, and concentrated power in artificial intelligence [18], [19].

Another priority is robust evaluation across languages and cultures. Vision-language models are often trained on English-dominant data, and knowledge sources may reflect specific legal or medical traditions. A model deployed globally must handle translation, localization, and conflicting norms. Retrieval systems must avoid privileging dominant sources. Knowledge graphs must represent multiple perspectives where appropriate. Evaluation should include cross-lingual visual question answering, culturally specific image interpretation, and community-defined accuracy criteria. These are difficult tasks, but they are necessary if knowledge-enhanced systems are to support pluralistic societies. The alternative is a narrow form of accuracy that performs well on benchmarks while failing in situated practice.

Finally, sustainability and governance must be linked. A system that is accurate but environmentally costly, or fair in one region but extractive in another, is not genuinely sustainable. Knowledge enhancement can reduce retraining needs by enabling external updates, but it can also increase inference complexity. Policy should encourage efficiency, reuse, and shared public knowledge infrastructures. It should also support maintenance: knowledge bases, ontologies, and evaluation suites require ongoing funding and stewardship. Technical debt in machine learning systems often arises when maintenance is deferred [20]. The same is true for knowledge infrastructures. Without sustained investment, even well-designed systems decay into outdated or inconsistent resources.

8. Conclusion

Knowledge-enhanced vision-language models offer a promising path toward accurate image understanding, but accuracy is not achieved by adding external data to a model. It emerges from the careful alignment of architecture, knowledge representation, retrieval infrastructure, adaptation strategy, evaluation, and governance. Transformer-based multimodal systems provide the representational foundation [1], while contrastive, generative, and instruction-tuned models enable flexible visual reasoning [2], [4], [14]. Retrieval augmentation and knowledge graphs can supply factual and relational context that raw image-text pretraining

may lack [7], [9]. Parameter-efficient adaptation and modular design can make domain specialization more sustainable [6], [13]. Yet each enhancement introduces new dependencies, failure modes, and accountability questions.

The central argument of this paper is that knowledge must be treated as governed infrastructure. This means documenting sources, versioning knowledge, monitoring retrieval quality, calibrating uncertainty, evaluating subgroup performance, and providing mechanisms for contestation. It also means recognizing that hallucination mitigation, including attention-level interventions, is one part of a broader reliability program [11]. Robustness, fairness, sustainability, and policy are not peripheral concerns; they shape whether a system can be trusted in healthcare, industry, public administration, or embodied settings. The most accurate systems will be those that can explain their evidence, acknowledge their limits, and adapt responsibly when knowledge changes.

Future work should develop interoperable standards for knowledge provenance, multimodal evaluation suites that measure calibration and fairness, and governance frameworks that support human oversight without stifling innovation. It should also investigate how knowledge-enhanced models behave under distribution shift, adversarial retrieval, and cross-cultural deployment. The research agenda is therefore both technical and institutional. Accurate image understanding is not only a model property; it is a property of the entire socio-technical system that produces, validates, and uses visual knowledge.

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. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
2. 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. *Proceedings of the 38th International Conference on Machine Learning*, 139, 8748-8763.
3. Kim, W., Son, B., & Kim, I. (2021). ViLT: Vision-and-language transformer without convolution or region supervision. *Proceedings of the 38th International Conference on Machine Learning*, 139, 5583-5594.
4. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *Proceedings of the 39th International Conference on Machine Learning*, 162, 12888-12900.
5. 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. *Advances in Neural Information Processing Systems*, 35, 23716-23736.
6. Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). LoRA: Low-rank adaptation of large language models. *International Conference on Learning Representations*.
7. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S., & Kiela, D. (2020). Retrieval-augmented

generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459-9474.

8. Guu, K., Lee, K., Tung, Z., Pasupat, P., & Chang, M.-W. (2020). REALM: Retrieval-augmented language model pre-training. *Proceedings of the 37th International Conference on Machine Learning*, 119, 3929-3938.
9. Bordes, A., Usunier, N., Garcia-Durán, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. *Advances in Neural Information Processing Systems*, 26, 2787-2795.
10. 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., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv:2108.07258*.
11. 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).
12. Wang, Z., Yu, J., Yu, A. W., Dai, Z., Tsvetkov, Y., & Cao, Y. (2021). SimVLM: Simple visual language model pretraining with weak supervision. *International Conference on Learning Representations*.
13. Li, J., Li, D., Savarese, S., & Hoi, S. (2023). BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Proceedings of the 40th International Conference on Machine Learning*, 202, 19730-19742.
14. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36.
15. Driess, D., Xia, F., Sajjadi, M. S. M., Lynch, C., Chowdhery, A., Ichter, B., Wahid, A., Tompson, J., Vuong, Q., Yu, T., Huang, W., Chebotar, Y., Sermanet, P., Duckworth, D., Levine, S., Vanhoucke, V., Hausman, K., Toussaint, M., Greff, K., ... Florence, P. (2023). PaLM-E: An embodied multimodal language model. *Proceedings of the 40th International Conference on Machine Learning*, 202, 8469-8488.
16. 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.
17. 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. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220-229.
18. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, 610-623.
19. Weidinger, L., Uesato, J., Rauh, M., Griffin, C., Huang, P.-S., Mellor, J., Glaese, A., Cheng, M., Balle, B., Kasirzadeh, A., Biles, C., Brown, S., Kenton, Z., Hawkins, W., Stepleton, T., Birhane, A., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 214-229.

20. 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. *Advances in Neural Information Processing Systems*, 28, 2503-2511.