

Parameter-Efficient Fine-Tuning for Hallucination-Aware Multimodal Models

Ekkehard Kabus

Department of Computer Science, University of Central Florida, Orlando, FL, USA.

ekkehardkabus@ucf.edu

Abstract

Multimodal models increasingly combine visual, textual, auditory, and structured data to support decision-making in domains such as healthcare, mobility, industrial inspection, education, and public administration. Their capacity to generate fluent cross-modal descriptions and recommendations, however, is accompanied by hallucination risks that can undermine trust, safety, and regulatory compliance. This paper examines parameter-efficient fine-tuning as a system-level strategy for hallucination-aware multimodal models. Rather than treating adaptation as a narrow optimization problem, the analysis frames it as an architectural, infrastructural, and governance concern. It reviews adapter-based, low-rank, prompt-based, and quantization-aware approaches, then connects them to multimodal hallucination sources such as weak grounding, attention misallocation, dataset bias, and evaluation gaps. The discussion emphasizes structural trade-offs involving memory, latency, modularity, observability, fairness, sustainability, and policy alignment. It argues that parameter-efficient fine-tuning can reduce computational barriers and enable domain-specific correction, but it can also concentrate risk when hallucination-aware objectives are absent from training data, evaluation pipelines, and deployment controls. The paper concludes that robust multimodal systems require layered governance, continuous auditing, modality-aware benchmarks, and infrastructure that makes adaptation reversible, traceable, and accountable. Such an approach positions parameter-efficient fine-tuning not as a shortcut for model customization, but as a component of socio-technical assurance for generative multimodal infrastructures.

Keywords

parameter-efficient fine-tuning; multimodal models; hallucination; vision-language systems; model governance; robustness; sustainability; responsible artificial intelligence.

1. Introduction

Parameter-efficient fine-tuning has emerged as a central strategy for adapting large models without retraining all parameters [1]. In conventional full fine-tuning, every weight in a neural network is updated, which creates substantial memory, compute, and storage burdens. Parameter-efficient methods instead modify or inject a small number of trainable components, allowing organizations to specialize foundation models while preserving most of the original architecture [2]. This shift is especially consequential for multimodal systems, where image encoders, language decoders, cross-modal projectors, and retrieval components may each require different adaptation policies. The appeal is straightforward: lower cost, faster iteration, and easier deployment across heterogeneous environments. Yet the deeper significance lies in how parameter-efficient fine-tuning changes the governance and failure modes of generative multimodal systems.

Hallucination in multimodal models is not merely a linguistic error. It often arises when a model produces content that is unsupported by visual evidence, misaligned with retrieved

documents, or inconsistent with prior conversational context. A system may describe objects that are absent, infer relationships that were never depicted, or generate confident clinical, legal, or safety recommendations from ambiguous inputs. Such failures are socio-technical because they emerge from the interaction among model architecture, training data, user interface, organizational incentives, and evaluation practices. Parameter-efficient fine-tuning can mitigate some of these failures by enabling targeted corrections, domain-specific grounding, and lighter-weight safety tuning [3]. At the same time, it can obscure accountability when adapters are distributed, reused, or combined without comprehensive testing [4]. The central question is therefore not whether parameter-efficient fine-tuning is computationally attractive, but whether it can be designed as part of a credible hallucination-aware infrastructure.

This paper develops that question through a system-level analysis. It first reviews parameter-efficient fine-tuning approaches and their implications for multimodal adaptation [5]. It then examines hallucination sources in vision-language architectures and the limits of current evaluation practices [6]. Subsequent sections analyze architectural trade-offs, governance mechanisms, deployment infrastructure, robustness, fairness, sustainability, and policy implications. The argument is that hallucination-aware multimodal systems require more than efficient training methods. They require an assurance architecture in which adaptation, monitoring, escalation, and redress are treated as first-class design concerns.

2. Parameter-Efficient Fine-Tuning as a Systems Strategy

Parameter-efficient fine-tuning encompasses several technical families, each with distinct operational consequences. Adapter-based methods insert small trainable modules into transformer layers while freezing the original model weights [1]. Low-rank adaptation approximates weight updates through low-rank matrices, reducing trainable parameters while retaining expressive capacity [2]. Prompt-based methods prepend or optimize continuous prompts, shifting the adaptation burden from model weights to input-space or prefix-space representations [3]. Prompt tuning further demonstrated that learned soft prompts can approach full fine-tuning performance under sufficient scale [4]. More recent work has shown that few-shot parameter-efficient fine-tuning can outperform in-context learning in cost and accuracy for several tasks [5]. Quantization-aware methods such as QLoRA combine low-rank adaptation with quantized base models, enabling fine-tuning of very large models on limited hardware [6].

From a systems perspective, these methods alter the location of change. In full fine-tuning, the model artifact itself becomes the site of adaptation, which simplifies some deployment pipelines but makes versioning expensive. In parameter-efficient fine-tuning, adaptation can be separated from the base model, stored as small deltas, and composed at inference time. This modularity supports multi-tenant platforms, edge deployment, and rapid rollback. However, it also introduces configuration complexity. An organization may maintain dozens of adapters for different departments, languages, modalities, or risk thresholds. The resulting system resembles a software supply chain more than a single trained model. Without disciplined artifact registries, dependency tracking, and permission models, parameter-efficient fine-tuning can create hidden technical debt that is difficult to observe until failures occur [18].

The infrastructure implications are substantial. Serving systems must load a base model and dynamically apply one or more adapters, which affects latency, memory bandwidth, and batching strategy. In multimodal settings, adaptation may need to occur at different points:

vision encoder, cross-modal projector, language decoder, or retrieval ranker. Each location carries different trade-offs. Adapting the vision encoder can improve visual grounding but may damage generalization across imaging conditions. Adapting the language decoder can improve stylistic and factual conformity but may increase hallucination if visual evidence is weakly represented. Adapting the cross-modal projector can strengthen alignment but may require paired data that is scarce or biased. Parameter-efficient fine-tuning thus becomes an architectural decision about where to intervene, not merely how to reduce trainable parameters.

3. Hallucination in Multimodal Architectures

Multimodal hallucination is shaped by the architecture that binds modalities together. Contrastive image-text models such as CLIP provide strong zero-shot alignment between visual and textual representations [7]. Bootstrapping methods such as BLIP-2 connect frozen image encoders with large language models through lightweight querying transformers, enabling flexible vision-language generation [8]. Systems such as MiniGPT-4 demonstrate how a small set of trainable projection layers can align a visual encoder with an instruction-following language model [9]. These architectures create powerful capabilities but also introduce failure points. The language model may overrule visual evidence, the projector may lose fine-grained spatial information, and the image encoder may encode dataset biases that are invisible to the user. Hallucination can therefore emerge from the interface between components rather than from a single defective module.

Recent work on vocabulary hijacking in large vision-language models has examined how critical attention heads and inert tokens affect hallucination mitigation [10]. The significance of this line of inquiry is system-level: it suggests that hallucination is not only a data or decoding problem, but also a problem of how attention distributes probability mass across vocabulary and modality tokens. If certain tokens absorb attention without contributing grounded information, the model may generate fluent but unsupported content. Parameter-efficient fine-tuning can interact with this phenomenon by modifying attention behavior, prompt representations, or projector outputs. A carefully designed adapter may reduce the influence of inert tokens, while a poorly designed adapter may amplify them. The difference depends on evaluation targets, training examples, and the degree to which visual evidence is preserved during adaptation.

Hallucination research in natural language generation has produced extensive taxonomies and detection methods [11]. Work on faithfulness and factuality in summarization shows that generated text can be fluent while contradicting source documents [12]. In image captioning, object hallucination has been documented as a persistent problem in which models mention objects not present in the image [13]. Conditional generation research has also developed methods for detecting hallucinated content by comparing generated claims with source evidence [14]. These findings apply directly to multimodal systems, where evidence may be visual, textual, retrieved, or conversational. A model may hallucinate by misreading an image, by ignoring a retrieved passage, by overgeneralizing from a caption, or by following a conversational prior that conflicts with the current input. Parameter-efficient fine-tuning must therefore be evaluated against multiple evidence channels, not only against language fluency or task accuracy.

4. Architectural Integration and Structural Trade-Offs

Integrating parameter-efficient fine-tuning into hallucination-aware multimodal systems requires choices about modularity, composition, and observability. Foundation model research has emphasized that large pretrained systems concentrate capability and risk in ways that complicate downstream accountability [15]. A hospital deploying a multimodal triage assistant, for example, may fine-tune a vision-language model to recognize wound conditions and generate preliminary notes. If the adaptation is stored as a low-rank adapter, the hospital can update it without retraining the base model. Yet the final output depends on the base model, the adapter, the retrieval index, the prompt template, and the user interface. If a hallucination occurs, determining whether the cause lies in the adapter, the base model, or the data pipeline may be difficult. Architectural integration must therefore include provenance metadata, adapter versioning, and evidence traceability.

There are also trade-offs between specialization and generalization. A parameter-efficient adapter trained on a narrow domain may improve accuracy within that domain while degrading performance on out-of-distribution inputs. In multimodal settings, this risk is amplified because visual conditions vary widely across cameras, lighting, resolution, and cultural context. An adapter that performs well on curated benchmark images may fail on mobile phone photographs from a different population. Similarly, prompt-based adaptation may be sensitive to tokenization and language, especially for low-resource languages or mixed-script inputs. Low-rank adaptation may be more stable but still inherits biases from the base model and fine-tuning data. The design challenge is to create adaptation mechanisms that improve grounding without creating brittle domain silos.

Composition is another structural concern. Multiple adapters may be combined for multilingual, multi-domain, or multi-task use. Composition can be sequential, parallel, or gated, each with different implications for latency and interpretability. Parallel composition may preserve modularity but increase memory bandwidth. Sequential composition may enable cumulative refinement but make credit assignment harder. Gated composition may allow dynamic routing but introduce new failure modes if the gate is miscalibrated. In hallucination-aware systems, composition must be accompanied by conflict detection. If two adapters encode contradictory factual norms, the system should be able to detect inconsistency rather than silently blending outputs. This suggests that parameter-efficient fine-tuning should be embedded in a broader orchestration layer that manages adapters as governed artifacts, not merely as performance patches.

5. Governance, Evaluation, and Deployment Infrastructure

Governance of multimodal hallucination cannot be reduced to model-level metrics. Risk taxonomies for language models identify harms related to misinformation, bias, privacy, security, and autonomy [16]. These risks become more complex when models process images, documents, and sensor data. A governance framework must define who can create adapters, what data they may use, how they are evaluated, where they are deployed, and how incidents are reported. It must also address the political economy of large models, including the concentration of computational resources and the tendency of large-scale systems to encode unexamined assumptions [17]. Parameter-efficient fine-tuning can democratize adaptation, but it can also distribute unaccountable modifications across many actors. Governance must therefore balance openness with traceability.

Evaluation infrastructure is a critical bottleneck. Static benchmarks quickly become saturated or gamed, and they rarely capture the long-tail conditions that produce hallucinations in practice. A robust evaluation pipeline for hallucination-aware multimodal models should

combine automated factuality checks, human review, red-teaming, and field monitoring. It should test grounding across modalities, including image regions, OCR text, retrieved documents, and temporal context. It should also measure abstention behavior. A model that says it cannot determine whether a lesion is malignant is safer than one that hallucinates a diagnosis. Parameter-efficient fine-tuning can be used to train abstention and escalation behavior, but only if the training data and reward signals include negative examples and uncertainty cases. Evaluation must therefore be designed alongside adaptation, not after deployment.

Deployment infrastructure must support observability and reversibility. Adapters should be versioned, signed, and linked to evaluation reports. Inference logs should capture enough context to reconstruct why a particular output occurred, while respecting privacy and data minimization. Rollback should be possible without retraining the base model. Canary deployment should be standard for high-risk domains. Because multimodal systems often operate across cloud, edge, and on-premise environments, the infrastructure must handle heterogeneous hardware and intermittent connectivity. In edge settings, adapters may be cached locally, which raises questions about synchronization, staleness, and security. A compromised or outdated adapter could produce systematic hallucinations across many devices. The supply chain risk is analogous to software dependency risk, and it requires similar controls: provenance, integrity checks, and rapid revocation.

6. Robustness, Fairness, and Sustainability

Robustness in multimodal systems is not only about adversarial attacks. It also concerns distribution shift, sensor degradation, occlusion, language variation, and cultural context. Energy and policy considerations in deep learning have shown that computational scale carries environmental and economic costs [19]. Green AI arguments further emphasize that efficiency should be measured alongside accuracy, not treated as a secondary concern [20]. Parameter-efficient fine-tuning is attractive because it reduces the energy required for adaptation. However, the total environmental footprint depends on deployment patterns. If many small adapters are served simultaneously on inefficient infrastructure, or if base models are repeatedly reloaded, the savings may be reduced. Sustainability must be evaluated at the system level, including inference, storage, networking, and hardware lifecycle.

Fairness is equally systemic. A multimodal model may perform well on average while failing disproportionately for users with darker skin tones, non-standard dialects, disabilities, or low-bandwidth devices. Parameter-efficient fine-tuning can address some disparities by enabling targeted adaptation with community-specific data. But it can also fragment accountability if fairness is assessed only within separate adapters. A system may have a high-performing adapter for one population and a poorly evaluated adapter for another, creating a two-tier service. Governance should require intersectional evaluation across adapters and base models. It should also examine whether adaptation data was collected with consent and whether benefits return to the communities that provided data. Fairness is not achieved by efficiency alone; it requires deliberate inclusion in problem framing, data collection, evaluation, and redress.

Robustness and fairness intersect with hallucination through uncertainty. When a model is uncertain, it should be able to express uncertainty, request additional evidence, or defer to a human. Parameter-efficient fine-tuning can shape these behaviors by training adapters on examples of ambiguity and escalation. However, if the base model is overconfident, an adapter may struggle to override that tendency. Architectural support for uncertainty may

include ensemble methods, retrieval verification, and cross-modal consistency checks. For example, a system could compare a generated caption against object detection outputs, or a clinical note against structured lab results. These checks require additional infrastructure but can reduce the risk of fluent hallucination. The design principle is that efficiency should not come at the cost of verifiability.

7. Policy and Forward-Looking Perspectives

Policy frameworks for artificial intelligence increasingly emphasize risk management, transparency, and accountability. Multimodal hallucination is difficult to regulate because harm is context-dependent. A hallucinated caption in a museum guide may be inconvenient; a hallucinated diagnosis in a clinical decision support system may be dangerous. Parameter-efficient fine-tuning complicates regulation because the adapted model may not be a single artifact. Regulators may need to examine the base model, adapter, data provenance, evaluation reports, and deployment context together. This suggests a shift from model-centric approval to lifecycle assurance. Organizations should maintain documentation for each adapter, including intended use, limitations, evaluation results, and monitoring plans. Third-party audits should be able to reconstruct the adaptation chain without accessing sensitive raw data.

Looking forward, several technical and institutional developments are likely to shape the field. First, hallucination-aware fine-tuning objectives may become standard, combining task loss with grounding, consistency, and abstention terms. Second, multimodal evaluation may move toward evidence-linked benchmarks that require models to cite visual regions or retrieved passages. Third, adapter marketplaces and model hubs will need security and governance features comparable to software package repositories. Fourth, edge deployment will require lightweight verification mechanisms that can run alongside inference. Fifth, policy will likely demand incident reporting and redress mechanisms for high-risk applications. Parameter-efficient fine-tuning can support these developments if it is designed with auditability and reversibility in mind, rather than as a purely computational optimization.

Cross-domain comparisons are instructive. In software engineering, dependency management matured through package registries, vulnerability databases, and reproducible builds. In aviation, safety-critical systems use redundancy, certification, and incident investigation. In medicine, interventions are evaluated through trials, post-market surveillance, and clinical governance. Multimodal artificial intelligence lacks equivalent maturity. Parameter-efficient fine-tuning offers an opportunity to build modularity that resembles these mature systems, but only if adapters are treated as governed components with clear interfaces and failure semantics. The alternative is a proliferation of opaque patches that improve benchmark scores while increasing systemic fragility.

8. Conclusion

Parameter-efficient fine-tuning is a powerful enabler for adapting multimodal models under resource constraints. It reduces the cost of specialization, supports modular deployment, and creates opportunities for domain-specific hallucination mitigation. However, efficiency does not automatically produce safety, fairness, or accountability. Hallucination in multimodal systems emerges from the interaction of architecture, data, evaluation, governance, and user context. Parameter-efficient methods can either reduce or amplify these risks depending on how they are designed and deployed. The paper has argued for a system-level approach that treats adaptation as part of a broader assurance infrastructure. This includes versioned

adapters, evidence traceability, modality-aware evaluation, abstention behavior, fairness auditing, environmental accounting, and policy-aligned governance. The next generation of multimodal systems should not ask only how to fine-tune more efficiently. It should ask how to fine-tune in ways that make hallucination visible, contestable, and correctable across the full lifecycle of deployment.

References

1. Houlisby, N., Giurciu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. In *Proceedings of the 36th International Conference on Machine Learning* (Vol. 97, pp. 2790–2799). PMLR.
2. 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. In *International Conference on Learning Representations*. OpenReview.
3. Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing* (Volume 1: Long Papers) (pp. 4582–4597). Association for Computational Linguistics.
4. Lester, B., Al-Rfou, R., & Constant, N. (2021). The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 3045–3059). Association for Computational Linguistics.
5. Liu, H., Tam, D., Muqeeth, M., Mohta, J., Huang, T., Bansal, M., & Raffel, C. (2022). Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning. In *Advances in Neural Information Processing Systems* (Vol. 35, pp. 1950–1965). Curran Associates.
6. Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems* (Vol. 36, pp. 10088–10115). Curran Associates.
7. 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.
8. 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.
9. Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv*. <https://arxiv.org/abs/2304.10592>
10. Chen, Y., Li, J., Yao, W., Ma, X., Du, G., Wang, W., & Li, J. (2026, July). Vocabulary hijacking in vlms: 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).

11. Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y. J., Madotto, A., & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), Article 248, 1–38. <https://doi.org/10.1145/3571730>
12. Maynez, J., Narayan, S., Bohnet, B., & McDonald, R. (2020). On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 1906–1919)*. Association for Computational Linguistics.
13. Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 4035–4045)*. Association for Computational Linguistics.
14. Zhou, Y., Lei, T., Liu, H., Du, N., Huang, Y., Zhao, V., Dai, A. M., Le, Q. V., Laudon, J., & Gao, J. (2021). Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021 (pp. 1393–1404)*. Association for Computational Linguistics.
15. 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. S., Chen, A., Creel, K., Davis, J. Q., Demszky, D., ... Liang, P. (2021). On the opportunities and risks of foundation models. *arXiv*. <https://arxiv.org/abs/2108.07258>
16. 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., He, L., Hendricks, L. A., Rimell, L., Isaac, W., ... Gabriel, I. (2022). Taxonomy of risks posed by language models. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (pp. 214–229)*. Association for Computing Machinery. <https://doi.org/10.1145/3531146.3533088>
17. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (pp. 610–623)*. Association for Computing Machinery. <https://doi.org/10.1145/3442188.3445922>
18. 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.
19. 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. <https://doi.org/10.18653/v1/P19-1355>
20. Schwartz, R., Dodge, J., Smith, N. A., & Etzioni, O. (2020). Green AI. *Communications of the ACM*, 63(12), 54–63. <https://doi.org/10.1145/3381831>