

Reinforcement Learning for Visual-Language Reasoning with Multimodal Large Language Models

Romolo Mastandrea

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

Sarah-Jane Rourke

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

Ale Ferrandiz

Department of Computer Science and Engineering, University at Buffalo, Buffalo, NY, USA.
aferrandiz@buffalo.edu

Rachelle Gogan

Department of Electrical Engineering and Computer Science, University of Missouri,
Columbia, MO, USA.
gogan1977@missouri.edu

Abstract

Reinforcement learning has become a central mechanism for aligning multimodal large language models with complex visual-language reasoning tasks. Unlike conventional supervised fine-tuning, reinforcement learning enables models to optimize sequential decisions, trade off exploration and exploitation, and receive feedback on reasoning trajectories rather than only final answers. This paper examines the system-level implications of using reinforcement learning for visual-language reasoning with multimodal large language models. It argues that the integration of visual encoders, language backbones, reward models, orchestration layers, and deployment infrastructure creates structural trade-offs that cannot be resolved by model architecture alone. The discussion covers reward modeling, alignment, scalability, robustness, fairness, governance, sustainability, and policy. Particular attention is given to how reinforcement learning pipelines affect data curation, human feedback, compute allocation, latency, energy consumption, and accountability in real-world settings. The paper also considers how multimodal reasoning introduces unique failure modes, including hallucinated visual grounding, spurious correlations, and unsafe generalization across domains. By treating reinforcement learning as a socio-technical infrastructure rather than a narrow training technique, the analysis highlights the need for layered evaluation, transparent reward governance, and lifecycle management. The conclusion emphasizes that future progress depends on reconciling capability gains with robustness, fairness, and operational sustainability. The paper contributes a conceptual framework for researchers, engineers, and policymakers who must design, deploy, and regulate multimodal reasoning systems under uncertainty.

Keywords

reinforcement learning; multimodal large language models; visual-language reasoning; alignment; system governance; deployment sustainability.

1. Introduction

Multimodal large language models have rapidly expanded the boundary of artificial intelligence by coupling visual perception with linguistic reasoning. Systems that can interpret images, diagrams, screenshots, medical scans, and video frames while generating coherent natural language responses are no longer purely experimental. They are being integrated into interfaces for education, accessibility, industrial inspection, customer service, robotics, and scientific analysis. This expansion has made visual-language reasoning a critical test case for whether large models can move beyond pattern recognition toward robust, context-sensitive inference. Reinforcement learning offers one of the most promising pathways for improving that inference because it can optimize behavior over sequences rather than isolated outputs, as established in the broader reinforcement learning literature [1]. At the same time, the transformer architecture has provided the representational substrate for scaling language and multimodal models [2], while large-scale pretraining has demonstrated that general-purpose models can acquire broad capabilities from diverse data [3]. Visual-language pretraining has further shown that aligning images and text in a shared embedding space can support transfer across tasks [4]. The convergence of these developments creates both opportunity and risk.

The central problem is not simply whether reinforcement learning can improve benchmark scores. It is whether reinforcement learning can be embedded within a multimodal system in a way that preserves reliability, interpretability, fairness, and operational viability. A model that reasons well on curated academic datasets may still fail when deployed in hospitals, factories, or public services where visual inputs are noisy, distribution shifts are common, and mistakes carry material consequences. Reinforcement learning introduces additional complexity because reward signals must encode desired behavior, and those signals are themselves artifacts of human judgment, data selection, and institutional priorities. When rewards are poorly specified, models may optimize superficial cues rather than genuine visual-language understanding. When rewards are narrowly defined, models may become brittle. When rewards are opaque, accountability becomes difficult. This paper therefore treats reinforcement learning for visual-language reasoning as a system-level challenge involving architecture, data infrastructure, governance, and sustainability.

The argument proceeds through several interconnected themes. First, it situates reinforcement learning within the foundations of multimodal large language models and explains why sequential optimization matters for reasoning. Second, it analyzes architectural integration between visual encoders, language backbones, reward models, and orchestration layers. Third, it examines reward modeling and alignment as governance problems as much as technical ones. Fourth, it discusses training pipelines, scalability, and deployment constraints. Fifth, it explores robustness, fairness, policy, and sustainability. Finally, it offers case illustrations, cross-domain comparisons, and forward-looking perspectives. The aim is not to provide a narrow survey of algorithms but to develop a critical framework for understanding how reinforcement learning reshapes the socio-technical infrastructure of multimodal reasoning.

2. Foundations of Reinforcement Learning and Multimodal Large Language Models

Reinforcement learning is concerned with agents that learn through interaction, receiving feedback in the form of rewards and adjusting behavior to maximize long-term outcomes [1]. In traditional settings, this paradigm has been applied to games, control systems, robotics, and resource management. Its migration into language modeling has changed the scale and nature of the problem. Instead of learning a policy for navigating a physical environment, a language

model learns a policy for generating sequences of tokens, tool calls, or intermediate reasoning steps. The reward may come from human preferences, automated verifiers, task success metrics, or a combination of signals. This shift makes reinforcement learning a natural fit for visual-language reasoning because many multimodal tasks require multi-step inference: identifying objects, relating them spatially, interpreting text in images, drawing inferences, and producing an answer that satisfies both visual evidence and linguistic coherence.

The transformer architecture provides the computational foundation for these models [2]. Its attention mechanisms allow long-range dependencies between visual and textual representations, enabling a model to condition language generation on image features and prior context. Large-scale pretraining has shown that transformers can acquire broad linguistic and factual capabilities when trained on massive corpora [3]. In the multimodal domain, contrastive pretraining has demonstrated that images and text can be aligned through natural language supervision [4]. Models such as Flamingo have extended this alignment to few-shot visual-language tasks by interleaving visual and textual inputs [5]. BLIP-2 has shown that frozen image encoders and frozen language models can be bridged through lightweight querying mechanisms [6]. These developments establish the representational basis for visual-language reasoning, but they do not by themselves guarantee robust sequential decision-making. Reinforcement learning enters as a way to shape behavior beyond imitation.

The distinction between supervised fine-tuning and reinforcement learning is important for system design. Supervised fine-tuning trains a model to reproduce desired outputs from labeled examples. It is effective for style, format, and basic task execution, but it tends to treat each output independently. Reinforcement learning, by contrast, can optimize for outcomes that unfold over time. In visual-language reasoning, this might mean rewarding a model for correctly identifying a subtle visual cue, asking for clarification when the image is ambiguous, invoking an external tool when numerical precision is required, or explaining uncertainty instead of hallucinating. The reward function becomes a mechanism for encoding procedural norms, not merely answer correctness. However, this also means that the quality of the system depends heavily on the quality of the reward model, the diversity of the training environments, and the governance of feedback data.

3. Architectural Integration of Visual Encoders, Language Models, and Reward Systems

A multimodal large language model used for visual-language reasoning typically consists of several interacting components. A visual encoder converts images or video frames into representations. A projection or querying module maps those representations into a space that the language model can consume. A language backbone generates text, reasons over context, and may call external tools. A reward model or reward ensemble evaluates outputs or trajectories. An orchestration layer manages inference, tool use, memory, and safety filters. Reinforcement learning can intervene at multiple points in this stack. It can fine-tune the language backbone, adapt the multimodal projector, train a policy for tool selection, or optimize a controller that decides when to retrieve external information. The choice of intervention point determines the computational cost, the risk of catastrophic forgetting, and the difficulty of debugging failures.

Models such as Flamingo and BLIP-2 illustrate different integration philosophies [5], [6]. Flamingo emphasizes interleaved visual and textual conditioning with few-shot generalization, while BLIP-2 uses a lightweight bridge between frozen unimodal components. Visual instruction tuning, as in LLaVA, demonstrates that multimodal models can be trained to follow complex instructions involving images [7]. MiniGPT-4 similarly explores the

alignment of visual encoders with advanced language models [8]. These architectures show that visual-language capability can be assembled through modular design. Reinforcement learning adds another layer of modularity because the reward model can be separated from the policy model. This separation is useful for governance, as it allows independent auditing of the reward function. Yet it also creates a coupling problem: if the reward model is biased, outdated, or misaligned with deployment context, the policy may optimize in harmful directions even when the language backbone remains technically capable.

Architectural trade-offs are particularly acute in latency-sensitive systems. A model that performs multiple reasoning steps, calls external tools, and consults a reward model may achieve higher accuracy but at the cost of slower response times and greater energy use. Edge deployment imposes additional constraints because visual encoders and language models may need to be quantized, distilled, or partitioned across devices. In such settings, reinforcement learning may be used to learn efficient policies that decide when to perform full reasoning and when to use a cached or heuristic response. This creates a tension between adaptive computation and predictable behavior. A policy that skips reasoning under time pressure may improve user experience in ordinary cases but fail catastrophically in rare high-stakes cases. System designers must therefore treat reinforcement learning as part of a broader architecture of safeguards, fallbacks, and human escalation paths.

4. Reward Modeling and Alignment for Visual-Language Reasoning

Reward modeling is the point at which technical optimization meets normative choice. In reinforcement learning from human feedback, human preferences are collected and used to train a reward model that guides policy improvement [10], [11]. Proximal policy optimization has become a widely used method for stable policy updates in this setting [12]. Constitutional approaches to alignment have explored the use of explicit principles and AI-generated feedback to reduce reliance on direct human labeling [13]. Recent work on trajectory filtering and alignment for smaller parameter models suggests that reinforcement learning pipelines can be made more scalable by selecting high-quality reasoning paths rather than treating all trajectories equally [9]. This development is significant because it points toward more efficient alignment strategies, but it also raises questions about who defines trajectory quality and how filtering criteria affect fairness and generalization. The required reference appears here as [9] and is integrated into the broader discussion of reward design without making it the sole center of the paper.

In visual-language reasoning, reward modeling is especially difficult because correctness is often multimodal and context-dependent. A model may produce a linguistically fluent answer that is visually unsupported. It may identify an object correctly but misstate its spatial relation. It may describe a chart accurately but draw an invalid conclusion. A reward model trained only on final answer correctness may miss these intermediate failures. A reward model trained on human preferences may inherit annotator biases, cultural assumptions, and inconsistencies. A reward model trained on automated metrics may be gamed by models that exploit metric weaknesses. Alignment therefore requires a portfolio of reward signals, including task-specific verifiers, human evaluations, adversarial tests, and uncertainty-aware scoring. The governance challenge is to make these signals transparent enough for audit while keeping them flexible enough to adapt across domains.

Alignment also has a temporal dimension. Models are trained at one point in time and deployed in environments that change. A reward model that reflects acceptable behavior in a research laboratory may be inappropriate in a clinical setting, a classroom, or a public safety

context. Reinforcement learning can amplify this mismatch because policies may continue to adapt through online interaction. Without careful monitoring, a deployed system may drift toward behaviors that maximize short-term reward but violate long-term norms. This is why alignment should be treated as a lifecycle process rather than a one-time training phase. It requires continuous evaluation, versioned reward models, incident reporting, rollback mechanisms, and clear lines of accountability. In multimodal systems, this lifecycle must also account for changes in visual data distribution, camera hardware, user demographics, and regulatory requirements.

5. Training Pipelines, Data Infrastructure, and Scalable Deployment

Reinforcement learning pipelines for multimodal reasoning depend on data infrastructure that is often underestimated in technical discussions. The pipeline must ingest images, videos, text, metadata, annotations, preference rankings, and interaction logs. It must manage licensing, privacy, consent, and provenance. Datasheets for datasets provide a useful model for documenting how data was collected, what it contains, and what limitations it has [20]. Without such documentation, reinforcement learning systems may be trained on data that is unrepresentative, legally questionable, or ethically problematic. The reward model may then encode biases that are difficult to detect because they are buried in the training pipeline. Scalable deployment therefore requires not only compute resources but also data governance, audit trails, and reproducible preprocessing.

The training process itself involves multiple stages. Pretraining aligns visual and linguistic representations. Supervised instruction tuning teaches the model to follow multimodal prompts. Reward modeling captures preferences or correctness signals. Policy optimization adjusts behavior using reinforcement learning. Evaluation and red-teaming test for failure modes. Deployment introduces monitoring, logging, and human oversight. Each stage has its own infrastructure requirements. Large-scale language models such as LLaMA have shown that open and efficient foundation models can be adapted for many downstream uses [14]. However, the computational cost of reinforcement learning can be substantial, especially when reward models are large, rollouts are long, and environments are complex. Organizations with limited resources may struggle to compete, leading to concentration of capability among a few actors. This concentration raises concerns about market power, research access, and democratic oversight.

Scalability is not only a matter of hardware. It is also a matter of organizational process. Reinforcement learning experiments require version control for policies, reward models, data snapshots, and evaluation suites. They require incident response when a model produces harmful or unsafe outputs. They require legal review when visual data includes faces, medical information, or proprietary material. They require cross-functional teams that include machine learning engineers, domain experts, ethicists, and compliance officers. The more capable the model, the more important these non-model components become. A system that performs well in a benchmark but lacks robust operational governance is not truly scalable. It is merely a prototype with hidden liabilities.

6. Robustness, Generalization, and Failure Modes

Robustness in visual-language reasoning is difficult because the input space is high-dimensional and noisy. Images can be blurred, occluded, rotated, compressed, or adversarially perturbed. Text in images can be stylized, handwritten, or multilingual. Video adds temporal dynamics and motion blur. A model trained on clean, well-lit, English-language data may fail

when confronted with real-world variation. Reinforcement learning can improve robustness if the training environment includes diverse perturbations and if rewards penalize overconfident errors. However, reinforcement learning can also produce brittle policies if the reward model is too narrow. A policy may learn to exploit specific visual shortcuts, such as relying on background color instead of object shape. It may learn to ignore small text because the reward model does not test fine-grained reading. It may learn to produce verbose explanations that appear thoughtful but lack visual grounding.

Embodied and multimodal models such as PaLM-E illustrate the complexity of grounding language in perception and action [15]. Multimodal chain-of-thought reasoning has been proposed as a way to improve inference by generating intermediate rationales that incorporate visual information [16]. ReAct-style approaches interleave reasoning and acting, allowing models to use tools and external knowledge [17]. Toolformer explores how language models can learn to call APIs and use external tools [18]. These approaches expand capability but also expand the attack surface. A model that can call tools may call the wrong tool, pass sensitive data to an external service, or misinterpret tool outputs. A model that generates chain-of-thought rationales may produce plausible but misleading explanations. Reinforcement learning can optimize tool use and reasoning, but it must also be constrained by safety envelopes, permission systems, and runtime monitors.

Failure modes in multimodal reinforcement learning are often socio-technical. A hospital may deploy a model that performs well on radiology benchmarks but fails on rare conditions because the reward model was trained on a skewed dataset. A factory may deploy a visual inspection system that misses defects because the reinforcement learning policy learned to minimize false positives at the expense of false negatives. A public agency may deploy a document-understanding model that systematically misreads handwritten forms from certain communities. These failures are not purely technical; they reflect choices about data collection, reward weighting, deployment context, and oversight. Robustness therefore requires adversarial testing, domain-specific validation, uncertainty quantification, and mechanisms for affected users to contest decisions. It also requires a culture of incident reporting that treats failures as opportunities for system improvement rather than reputational threats.

7. Fairness, Governance, and Policy Implications

Fairness in multimodal reinforcement learning is complicated by the interplay of visual and linguistic data. Images can encode race, gender, age, disability, and socioeconomic context. Text can encode stereotypes, exclusionary language, and cultural biases. A model that learns from such data may reproduce or amplify inequities. Reinforcement learning can make this worse if the reward model is trained on preferences from a narrow group of annotators. It can also make fairness interventions more difficult because the policy may learn to satisfy a fairness metric in one context while violating it in another. Governance requires disaggregated evaluation, participatory design, and continuous monitoring. It also requires clarity about trade-offs: optimizing for one notion of fairness may reduce another, and different communities may have different priorities.

Policy frameworks for multimodal artificial intelligence are still emerging. Regulators may focus on transparency, data protection, safety, and non-discrimination. However, reinforcement learning adds specific challenges. Reward models may be proprietary, making it difficult for external auditors to understand why a system behaves as it does. Training data may come from multiple jurisdictions with different privacy laws. Deployment may cross

borders, creating conflicts between local norms and global model behavior. Governance therefore needs to operate at multiple levels. At the organizational level, companies should maintain model cards, datasheets, and reward model documentation. At the industry level, standards bodies can define evaluation protocols and incident reporting formats. At the public level, regulators can require impact assessments, audit rights, and redress mechanisms for high-risk applications. These measures are not obstacles to innovation; they are conditions for trustworthy deployment.

8. Sustainability, Cost, and Operational Trade-offs

The sustainability of reinforcement learning for multimodal reasoning is a growing concern. Training large models requires enormous compute, energy, and water for cooling. Reinforcement learning multiplies this cost because it involves repeated rollouts, reward model inference, and policy updates. Multimodal inputs add further expense because images and videos are more data-intensive than text. A single training run may consume significant resources, and frequent retraining for alignment or safety may compound the impact. Organizations must therefore consider carbon accounting, energy-efficient hardware, data center location, and model reuse. Smaller models and parameter-efficient methods can reduce cost, but they may also reduce capability. The trade-off between performance and sustainability is not merely technical; it is a policy choice about what level of capability is worth what environmental cost.

Operational trade-offs extend to latency, reliability, and maintenance. A reinforcement learning system that requires multiple reward model calls per query may be too slow for real-time applications. A system that uses large ensembles for safety may be too expensive to run at scale. A system that relies on human feedback for continual improvement may face bottlenecks in annotation capacity and quality control. These constraints shape what kinds of multimodal reasoning can be deployed responsibly. Emergency response, autonomous driving, and surgical assistance require extremely low latency and high reliability. Educational tutoring, content moderation, and document analysis may tolerate more latency but require fairness and explainability. A mature engineering practice would match reinforcement learning architecture to operational context rather than applying a single recipe everywhere.

9. Case Illustrations and Cross-Domain Comparisons

Consider a healthcare scenario in which a multimodal model assists radiologists by summarizing imaging studies and flagging anomalies. Reinforcement learning could be used to reward accurate identification of clinically relevant features, appropriate expressions of uncertainty, and adherence to reporting standards. However, the reward model would need to be trained with radiologist input, validated across patient populations, and updated as clinical guidelines change. The system would need to log its reasoning, defer to human experts in ambiguous cases, and protect patient privacy. If reinforcement learning optimized only for diagnostic accuracy, it might neglect communication quality, equity, or workflow integration. A successful deployment would require a layered reward structure and strong governance.

In manufacturing, a visual inspection model might use reinforcement learning to decide when to zoom in, when to request a higher-resolution image, and when to escalate to a human inspector. The reward could balance defect detection, false alarm reduction, and throughput. Here, robustness to lighting changes and camera wear is critical. In education, a multimodal tutor might use reinforcement learning to adapt explanations based on a student's diagrams, handwriting, or facial expressions. The reward could include learning gains, engagement, and

fairness across student groups. Each domain has different risk profiles, data constraints, and regulatory environments. Cross-domain comparison shows that there is no universal reward function. Instead, reinforcement learning must be contextualized through domain expertise, stakeholder participation, and iterative evaluation.

10. Evaluation, Benchmarking, and Human-in-the-Loop Governance

Evaluation of multimodal reinforcement learning systems must go beyond static benchmarks. Static datasets cannot capture the dynamic, interactive nature of deployed systems. They may also be contaminated by data leakage or overfitting. A more robust evaluation framework includes adversarial testing, distribution shift analysis, counterfactual fairness checks, and long-horizon task simulations. Human-in-the-loop governance is essential because many failures are only visible in context. Domain experts can identify unsafe recommendations, affected communities can report discriminatory outcomes, and operators can detect drift over time. These human inputs should feed back into reward model updates, but they must be managed carefully to avoid bias, fatigue, and conflicts of interest.

Human oversight also raises questions about responsibility. If a reinforcement learning system makes a harmful decision, who is accountable? The model developers, the reward model annotators, the deploying organization, or the operator who failed to intervene? Clear governance requires documented roles, escalation paths, and audit trails. It also requires meaningful contestability, meaning that affected individuals can challenge decisions and receive explanations that are understandable to them. In multimodal settings, explanations may need to reference visual evidence, textual reasoning, and external sources. This is technically challenging but necessary for trust. Evaluation and governance should therefore be co-designed with deployment, not added as an afterthought.

11. Future Directions

Future research should focus on reward models that are robust, interpretable, and adaptable across domains. This includes methods for uncertainty-aware reward estimation, multi-objective alignment, and participatory reward design. It also includes advances in efficient reinforcement learning that reduce compute and energy costs without sacrificing safety. Another direction is the development of modular architectures that allow components to be updated independently, so that a visual encoder can be improved without retraining the entire policy. Federated and privacy-preserving learning may allow organizations to share alignment insights without sharing sensitive data. Finally, policy research should explore how to regulate reinforcement learning pipelines without stifling beneficial innovation. This may involve tiered risk frameworks, mandatory documentation, third-party audits, and public reporting for high-risk applications.

The broader lesson is that reinforcement learning for visual-language reasoning is not merely a model optimization problem. It is a design problem that spans architecture, data, institutions, and values. Progress will depend on interdisciplinary collaboration among computer scientists, engineers, social scientists, legal scholars, and domain experts. It will also depend on a willingness to acknowledge uncertainty and to build systems that can fail safely. The most capable system is not necessarily the most trustworthy. Trustworthiness emerges from transparency, accountability, robustness, and responsiveness to affected communities. As multimodal models become more integrated into daily life, these qualities will matter as much as raw performance.

12. Conclusion

Reinforcement learning offers a powerful pathway for improving visual-language reasoning in multimodal large language models. It enables systems to optimize sequential behavior, incorporate human feedback, and adapt to complex tasks that require visual grounding and linguistic coherence. However, the integration of reinforcement learning into multimodal systems introduces significant structural trade-offs. Reward models encode normative choices. Training pipelines depend on data governance. Deployment raises concerns about robustness, fairness, sustainability, and accountability. Architectural decisions affect latency, cost, and safety. Governance mechanisms must operate across organizational, industry, and public levels. No single technical fix can address these challenges.

This paper has argued for a system-level perspective that treats reinforcement learning as part of a socio-technical infrastructure. Such a perspective recognizes that model capability alone is insufficient. Trustworthy multimodal reasoning requires layered evaluation, transparent reward governance, human oversight, and lifecycle management. It requires attention to environmental cost and equitable access. It requires policy frameworks that protect rights without freezing progress. The field is moving quickly, and the temptation to prioritize benchmark gains is strong. Yet the long-term success of multimodal artificial intelligence will depend on whether systems can reason not only accurately but also responsibly. Reinforcement learning can help achieve that goal, but only if it is embedded within architectures and institutions that value robustness, fairness, and accountability as much as performance.

References

1. Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.
2. 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.
3. Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877–1901.
4. 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.
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. 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.

7. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. *Advances in Neural Information Processing Systems*, 36, 34892–34916.
8. Zhu, D., Chen, J., Shen, X., Li, X., & Elhoseiny, M. (2023). MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.
9. Li, Q. (2026, May). TrajLite PPO: Scalable Reasoning Path Filtering and Alignment for Small Parameter Large Language Models. In *2026 2nd International Conference on Artificial Intelligence and Digital Ethics (ICAIDE)* (pp. 29-32). IEEE.
10. Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., Schulman, J., Hilton, J., Kelton, F., Miller, L., Simens, M., Askell, A., Welinder, P., Christiano, P., Leike, J., & Lowe, R. (2022). Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35, 27730–27744.
11. Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., & Amodei, D. (2017). Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30, 4299–4307.
12. Schulman, J., Wolski, F., Dhariwal, P., Radford, A., & Klimov, O. (2017). Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.
13. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., Chen, C., Olsson, C., Olah, C., Hernandez, D., Drain, D., Ganguli, D., Li, D., Tran-Johnson, E., Perez, E., ... Kaplan, J. (2022). Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.
14. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.
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. Zhang, Z., Zhang, A., Li, M., Zhao, H., Karypis, G., & Smola, A. (2023). Multimodal chain-of-thought reasoning in language models. *arXiv preprint arXiv:2302.00923*.
17. Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K., & Cao, Y. (2023). ReAct: Synergizing reasoning and acting in language models. *International Conference on Learning Representations*.
18. Schick, T., Dwivedi-Yu, J., Dessì, R., Raileanu, R., Lomeli, M., Zettlemoyer, L., Cancedda, N., & Scialom, T. (2023). Toolformer: Language models can teach themselves to use tools. *Advances in Neural Information Processing Systems*, 36, 68539–68551.
19. Amodei, D., Olah, C., Steinhardt, J., Christiano, P., Schulman, J., & Mané, D. (2016). Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

20. 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.