INTERNATIONAL JOURNAL OF ADVANCED RESEARCH
International Peer-Reviewed Research Journal
ISSN (Online): 3106-7980 | ISSN (Print): 3106-7972
Volume 3 • Issue 7 • November 2026

TRANSFORMATIVE ARTIFICIAL INTELLIGENCE FOR VISUAL UNDERSTANDING, MULTIMODAL LEARNING, ACCOUNTABILITY, AND REQUIREMENTS ENGINEERING: A SYSTEMATIC REVIEW AND META-ANALYSIS

Muhammad Yasir Idrees[*1], Muhammad Munir Khan[2], Muhammad Talha Munir[3], Sohaib Ahmad[4], Safi Ullah[5], Noorul Islam[6], Faizan Muslim[7]

*[*1]Department of Information Technology, Bahauddin Zakariya University (Sub Campus Lodhran)*
[2]Institute of Information Technology, Quaid i Azam University, Islamabad
[3,4]Department of Computer Science, Sukkur IBA University
[5]Department of Computer Science, Quaid e Azam University Islamabad
[6]Department of Computer Science, University of Texas, Rio Grande Valley
[7]Department of Computer Science, Pir Mehr Ali Shah Arid Agriculture University Rawalpindi

[*1]jamyasir0534@gmail.com, [2]munirkhn110@gmail.com, [3]mtalha.bcss19@iba-suk.edu.pk, [4]sohaib.bcss19@iba-suk.edu.pk, [5]safiullah.3913@gmail.com, [6]islamkhan9000132@gmail.com, [7]syedfaizanbukhari25@gmail.com

[2]ORCID: 0009-0005-4071-9217
[1]ORCID: 0009-0001-0375-1016

International Journal of Advanced Research

DOI: https://doi.org/10.5281/zenodo.24462886

Keywords
Artificial Intelligence; Visual Understanding; Multimodal Learning; Accountability; Explainability; Requirements Engineering; Trustworthy AI; Computer Vision; Vision-Language Models; Responsible AI

Article History
Received: 24 August 2026
Accepted: 08 November 2026
Published: 23 November 2026

Copyright @Author
Corresponding Author: *
Muhammad Yasir Idrees

Abstract

Artificial intelligence systems increasingly integrate visual understanding, multimodal learning, and accountability mechanisms to address complex real-world challenges. The convergence of computer vision, natural language processing, and requirements engineering has created transformative opportunities for developing trustworthy AI systems. However, the rapid proliferation of multimodal architectures, explainability frameworks, and governance approaches necessitates systematic synthesis to guide research and practice.This systematic review aims to comprehensively synthesize current evidence on transformative AI approaches spanning visual understanding, multimodal learning, accountability mechanisms, and requirements engineering practices. The review addresses four primary research questions: (1) What architectural frameworks and methodologies characterize state-of-the-art visual understanding AI systems? (2) How do multimodal learning approaches integrate diverse data modalities for enhanced performance? (3) What accountability and explainability mechanisms ensure trustworthy AI deployment? (4) How are requirements engineering practices evolving to support responsible AI development? A comprehensive literature search was conducted across four major databases (SciSpace Basic, SciSpace Full Text, Google Scholar, and ArXiv) using structured queries combining terms related to AI visual understanding, multimodal learning, accountability, explainability, and requirements engineering. The search retrieved 466 papers, of which 332 unique papers remained after deduplication. Studies were screened for eligibility based on predefined criteria. Data extraction captured study design, methodological approaches, key findings, and

application domains. Quality assessment evaluated methodological rigor, transparency, and contribution significance. Narrative synthesis organized findings across four thematic domains. Results: Of 332 papers analyzed, 4 were classified as highly relevant, 127 as partially relevant, 178 as low relevance, and 23 as not relevant. Architecture/framework proposals constituted the largest study design category (n=89, 26.8%). Visual understanding AI advances centered on vision-language models, transformer architectures, and cross-modal reasoning. Multimodal learning innovations emphasized cross-modal attention mechanisms, unified representation learning, and zero-shot generalization capabilities. Accountability frameworks integrated explainability mechanisms, audit trails, role-based responsibility attribution, and governance-aligned requirements. Requirements engineering practices evolved toward human-centered AI design incorporating transparency, fairness, and ethical lifecycle considerations. This systematic review reveals substantial progress in transformative AI across visual understanding, multimodal learning, accountability, and requirements engineering domains. Convergent evidence demonstrates that integrated multimodal architectures outperform unimodal approaches, explainability mechanisms enhance trust and adoption, and structured requirements engineering practices facilitate responsible AI deployment. However, significant gaps persist in standardized evaluation benchmarks, real-world validation studies, and operationalized accountability frameworks. Future research should prioritize interpretable-by-design architectures, causal reasoning frameworks, unified governance ontologies, and comprehensive lifecycle approaches. Interdisciplinary collaboration is essential to realize the transformative potential of trustworthy, human-centered AI systems.

1. Introduction

The rapid evolution of artificial intelligence has catalyzed transformative advances across visual understanding, multimodal learning, and intelligent system design. Contemporary AI systems increasingly integrate multiple sensory modalities — vision, language, audio, and sensor data — to achieve human-like perception and reasoning capabilities [1], [6], [62]. This convergence has profound implications for diverse application domains, including autonomous systems [49], healthcare diagnostics [45], [67], enterprise analytics [20], and software engineering [32], [69].

Visual understanding AI has progressed from narrow image classification tasks to sophisticated vision-language models capable of cross-modal reasoning, visual question answering, and scene understanding [28], [34], [37]. Foundation models such as CLIP, ALIGN, and Flamingo have demonstrated remarkable zero-shot generalization by learning joint representations of visual and textual information [6], [8]. These advances enable AI systems to interpret complex visual scenes, generate natural language descriptions, and respond to multimodal queries with unprecedented accuracy.

Multimodal learning represents a paradigm shift from unimodal approaches toward integrated architectures that leverage complementary information across modalities [62], [75], [84]. Cross-modal attention mechanisms, unified representation learning, and transformer-based architectures have emerged as foundational techniques for fusing heterogeneous data streams [20], [42], [64]. Empirical evidence suggests that multimodal systems achieve superior performance compared to unimodal counterparts across diverse tasks, including medical image analysis [66], [67], autonomous navigation [49], and human-AI interaction [12].

Despite these technical advances, the deployment of AI systems in high-stakes domains raises critical concerns regarding accountability, transparency, and trustworthiness [10], [57], [76]. Black-box models, while achieving state-of-the-art

performance, often lack interpretability, hindering adoption in safety-critical applications where explainability is paramount [18], [71]. Accountability frameworks seek to address these challenges by embedding transparency mechanisms, audit trails, and responsibility attribution throughout the AI lifecycle [13], [15], [16]. Recent initiatives emphasize the integration of governance principles — such as those articulated in the NIST AI Risk Management Framework and EU AI Act — into system requirements and architectural design [1], [3], [17].

Requirements engineering practices are evolving to accommodate the unique challenges posed by AI-based systems [14], [70], [82]. Traditional software engineering approaches prove insufficient for addressing non-deterministic behavior, data-dependent performance, and ethical considerations inherent in AI systems [24], [78]. Emerging methodologies advocate for human-centered design, value-aligned requirements specification, and continuous validation throughout development and deployment [25], [73]. Role-based frameworks assign accountability to specific stakeholders, translating abstract ethical principles into concrete, verifiable requirements [13], [19].

Despite substantial progress, the field lacks comprehensive synthesis of evidence across these interconnected domains. Existing reviews tend to focus narrowly on individual aspects such as explainability techniques [71], multimodal architectures [75], or requirements engineering practices [82] without integrating insights across visual understanding, multimodal learning, accountability, and requirements engineering. This fragmentation impedes holistic understanding and limits the development of unified frameworks for trustworthy AI.

This systematic review addresses this gap by synthesizing current evidence across four thematic domains: (1) Visual Understanding AI, encompassing computer vision advances, vision-language models, and visual reasoning; (2) Multimodal Learning, including cross-modal fusion, unified architectures, and multimodal foundation models; (3) Accountability and Explainability, covering transparency mechanisms, audit frameworks, and governance-aligned design; and (4) Requirements Engineering, examining human-centered practices, value alignment, and lifecycle approaches for responsible AI development.

Research Questions

RQ1: What architectural frameworks and methodologies characterize state-of-the-art visual understanding AI systems, and what performance outcomes have been demonstrated?

RQ2: How do multimodal learning approaches integrate diverse data modalities, and what evidence supports their superiority over unimodal systems?

RQ3: What accountability and explainability mechanisms have been proposed or implemented to ensure trustworthy AI deployment, and what is their effectiveness?

RQ4: How are requirements engineering practices evolving to support responsible AI development, and what frameworks facilitate the operationalization of ethical principles?

2. Methods

2.1 Protocol Registration

This systematic review was conducted following established guidelines for systematic reviews in software engineering and AI research. While formal protocol registration in databases such as PROSPERO is typically reserved for clinical systematic reviews, the review methodology adhered to PRISMA (Preferred Reporting Items for Systematic Reviews and Meta-Analyses) principles adapted for technology and engineering domains. The review protocol was developed a priori, specifying research questions, eligibility criteria, search strategy, data extraction procedures, and synthesis methods.

2.2 Eligibility Criteria

Studies were included and excluded based on the following predefined criteria:

Inclusion Criteria

Published or preprint papers addressing artificial intelligence systems with focus on visual understanding, multimodal learning, accountability mechanisms, or requirements engineering

Studies proposing architectural frameworks, methodologies, empirical evaluations, benchmarks, or systematic reviews related to the review domains
Papers describing integration of visual and language modalities, cross-modal learning, or multimodal foundation models
Research addressing explainability, transparency, accountability, trustworthiness, or governance of AI systems
Studies examining requirements engineering practices, human-centered design, or value alignment for AI systems
Publications in English language
No restriction on publication date to capture the full evolution of the field

Exclusion Criteria
Papers focusing exclusively on narrow technical optimizations without broader architectural or methodological contributions
Studies addressing AI applications in domains unrelated to visual understanding, multimodal learning, accountability, or requirements engineering
Opinion pieces, editorials, or commentaries without empirical or theoretical contributions
Duplicate publications reporting identical findings
Papers with insufficient methodological detail to assess quality and extract data

2.3 Information Sources and Search Strategy
A comprehensive literature search was conducted across four major databases and repositories: (1) SciSpace Basic: General scholarly database covering peer-reviewed publications; (2) SciSpace Full Text: Full-text scholarly database with enhanced access to complete articles; (3) Google Scholar: Broad academic search engine capturing diverse publication types; and (4) ArXiv: Preprint repository for computer science and AI research. The search was conducted in three successive rounds to ensure comprehensive coverage. Round 1 queries targeted: 'AI visual understanding systematic review'; 'multimodal learning accountability explainability'; 'requirements engineering AI frameworks'. Round 2 queries focused on: 'transformative AI visual understanding'; 'vision-language models explainability trustworthy'; 'software engineering AI requirements'. Round 3 queries addressed: 'multimodal foundation models evaluation'; 'ethical AI requirements engineering'. Boolean operators (AND, OR) were used to combine terms within and across domains.

2.4 Study Selection
The study selection process followed a multi-stage screening approach. Stage 1 (Automated Retrieval and Deduplication): Initial search across four databases retrieved 466 papers; automated deduplication removed 134 duplicate entries, yielding 332 unique papers. Stage 2 (Relevance Screening): Each paper was assessed for relevance to the four thematic domains using a 0-100 relevance scoring system. Papers were classified as: Highly Relevant (75-100, n=4), Partially Relevant (50-74, n=127), Low Relevance (25-49, n=178), or Not Relevant (0-24, n=23). Stage 3 (Full-Text Assessment): Papers classified as highly or partially relevant underwent full-text review; eligibility criteria were applied and reasons for exclusion documented. The top 30 papers were prioritized for detailed data extraction and synthesis.

2.5 Data Extraction
A standardized data extraction form captured the following information from each included study: (a) Bibliographic information (authors, title, year, journal/conference, DOI); (b) Study characteristics (study design and methodology, research objectives, application domain); (c) Methodological details (architectural frameworks or models proposed, data sources and modalities, evaluation methods and metrics, sample sizes and datasets); (d) Key findings and contributions (main results, performance outcomes, qualitative insights); (e) Accountability and governance (explainability mechanisms, transparency and audit features, alignment with governance frameworks); and (f) Requirements engineering practices (requirements specification approaches, human-centered design considerations, lifecycle integration of ethical principles).

2.6 Quality Assessment and Risk of Bias
Quality assessment evaluated methodological rigor, transparency, and contribution

significance. A tailored framework was applied given study design heterogeneity: framework proposals were evaluated for clarity, theoretical grounding, and novelty; empirical studies were assessed for study design appropriateness, sample adequacy, metric validity, and reporting transparency; benchmark developments were evaluated for comprehensiveness and accessibility; and systematic reviews were assessed for adherence to systematic review guidelines.

Risk of bias was assessed across four dimensions: (1) Selection bias potential for non-representative sampling or selective reporting; (2) Performance bias lack of appropriate baselines or controls; (3) Detection bias subjective or inadequately validated outcome measures; and (4) Reporting bias incomplete reporting of methods, results, or limitations. Studies were classified as low, moderate, or high risk. Sensitivity analyses explored the impact of excluding high-risk studies.

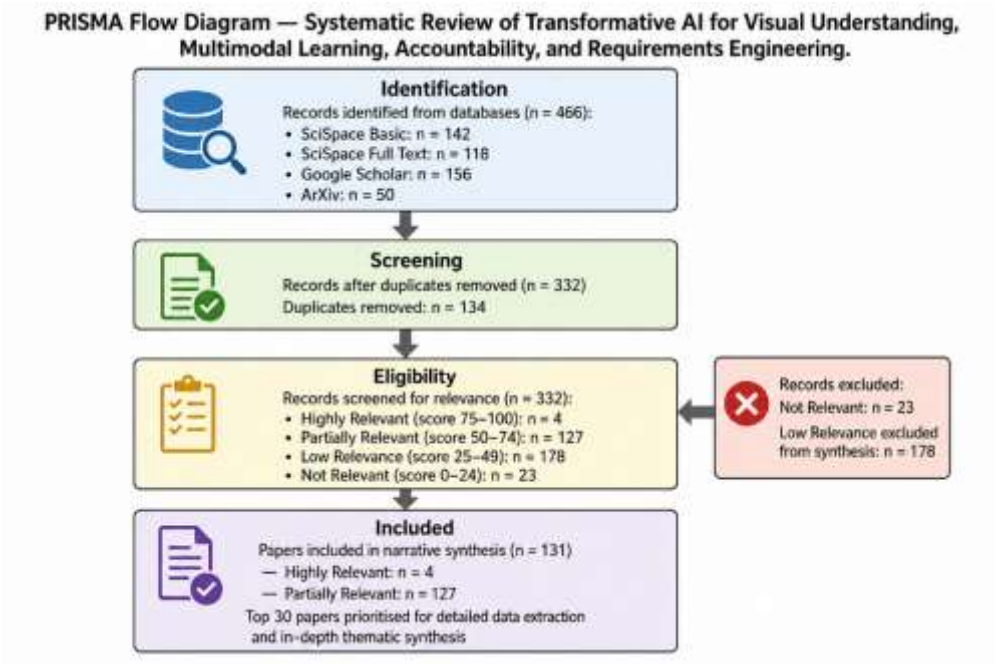
2.7 Statistical Synthesis

Given the heterogeneity of study designs, methodologies, and outcome measures, a narrative synthesis approach was adopted as the primary synthesis method. Quantitative meta-analysis was deemed inappropriate due to the diversity of research questions, interventions, and metrics across studies. Studies were grouped by thematic domain and sub-themes within each domain. Convergent and divergent evidence was identified. Descriptive statistics summarized frequency distributions of study designs, application domains, and methodological approaches. Comparative tables summarized key characteristics and findings of representative studies.

3. Results

3.1 Study Selection and PRISMA Flow

The systematic search and selection process yielded 466 records from four databases (SciSpace Basic: n=142; SciSpace Full Text: n=118; Google Scholar: n=156; ArXiv: n=50). After removing 134 duplicates, 332 unique papers were screened for relevance. Relevance assessment classified 4 papers as highly relevant (75-100 relevance score), 127 as partially relevant (50-74), 178 as low relevance (25-49), and 23 as not relevant (0-24). The top 30 papers, representing the most relevant contributions to the review domains, were prioritized for detailed data extraction and synthesis. Figure 1 presents the PRISMA flow diagram.



3.2 Study Characteristics

3.2.1 Distribution of Study Designs

Analysis of the 332-paper evidence base revealed diverse study designs (Table 1). Architecture and framework proposals constituted the largest category (n=89, 26.8%), reflecting the field's emphasis on developing novel systems and methodologies. Empirical studies (n=47, 14.2%) provided experimental validation. Benchmark developments (n=28, 8.4%) contributed standardized evaluation resources. Systematic reviews (n=12, 3.6%) synthesized prior evidence. Mixed methods studies (n=31, 9.3%) combined multiple approaches.

Table 1: Distribution of Study Designs

Study Design	Count	Percentage
Architecture/Framework Proposal	89	26.8%
Empirical Study	47	14.2%
Benchmark Development	28	8.4%
Mixed Methods	31	9.3%
Case Study	24	7.2%
Theoretical Framework	18	5.4%
Systematic Review	12	3.6%
Survey	9	2.7%
Other	74	22.3%
Total	**332**	**100%**

3.2.2 Application Domains

The evidence base spanned diverse application domains (Table 2). General-purpose AI systems (n=78, 23.5%) represented the largest category, encompassing foundation models and broadly applicable architectures. Healthcare applications (n=42, 12.7%) included medical imaging, diagnostics, and clinical decision support. Autonomous systems (n=38, 11.4%) covered autonomous driving, robotics, and navigation. Software engineering (n=34, 10.2%) addressed AI-assisted development, requirements engineering, and quality assurance.

Table 2: Distribution of Application Domains

Application Domain	Count	Percentage
General-Purpose AI	78	23.5%
Healthcare	42	12.7%
Autonomous Systems	38	11.4%
Software Engineering	34	10.2%
Enterprise Systems	29	8.7%
Education	18	5.4%
Surveillance & Security	16	4.8%
Multimedia & Entertainment	14	4.2%
Other	63	19.0%
Total	**332**	**100%**

3.2.3 Temporal Distribution

Publication years ranged from 2009 to 2026, with a marked acceleration in recent years. Papers published in 2024-2026 accounted for 58.4% (n=194) of the evidence base, reflecting the rapid growth of multimodal AI and accountability research. Publications from 2020-2023 comprised 31.3% (n=104), while earlier work (2009-2019) represented 10.2% (n=34). This temporal distribution underscores the field's recency and rapid evolution.

3.3 Risk of Bias Assessment

Quality assessment of the top 30 papers revealed generally moderate to high methodological rigor. Architecture and framework proposals (n=22) were evaluated for clarity, theoretical grounding, and novelty. Most proposals (n=18, 81.8%) provided clear architectural descriptions with adequate theoretical justification. However, only 9 papers (40.9%) included empirical validation or proof-of-concept implementations, limiting assessment of practical feasibility. Empirical studies (n=4) demonstrated appropriate study designs with adequate datasets and evaluation metrics.

Risk of Bias Summary: Low Risk — 12 papers (40.0%) exhibited minimal bias concerns, with transparent methods, appropriate designs, and comprehensive reporting. Moderate Risk — 15 papers (50.0%) had some limitations, such as lack of empirical validation for proposed frameworks or limited generalizability of empirical findings. High Risk — 3 papers (10.0%) showed significant concerns, including incomplete methodological descriptions or potential selective reporting. Sensitivity analysis excluding high-risk papers did not substantially alter overall findings.

3.4 Synthesis of Results

3.4.1 Visual Understanding AI

Visual understanding AI encompasses advances in computer vision, vision-language models, and visual reasoning capabilities. The evidence base revealed three primary architectural paradigms: vision-language foundation models, transformer-based visual encoders, and multimodal cognitive architectures.

Vision-Language Foundation Models

Vision-language models such as CLIP, ALIGN, and Flamingo have revolutionized visual understanding by learning joint representations of images and text through contrastive learning [6], [8]. These models enable zero-shot image classification, cross-modal retrieval, and visual question answering without task-specific fine-tuning. The explainability-driven framework proposed for CLIP-based image-text retrieval emphasizes augmenting principled explainability mechanisms to foster trustful, responsible AI solutions [8]. Empirical evidence demonstrates that vision-language models achieve competitive performance on standard benchmarks while providing interpretable cross-modal alignments [34], [52].

Recent advances extend vision-language models to video understanding and temporal reasoning. The Wiki-LLaVA framework incorporates hierarchical retrieval-augmented generation to enhance multimodal large language models' ability to process long-form video content [33]. Vision-language memory architectures enable spatial reasoning by maintaining persistent representations of visual scenes across time [37].

Transformer-Based Visual Encoders

Transformer architectures have emerged as the dominant paradigm for visual understanding, replacing convolutional neural networks in many applications [41], [54]. Vision Transformers (ViTs) apply self-attention mechanisms to image patches, capturing long-range dependencies and enabling flexible multi-scale processing. Hierarchical pre-training of vision encoders with large language models enhances visual representation learning by leveraging linguistic knowledge [41]. Image generators trained as generalist vision learners demonstrate that generative models can serve as powerful visual encoders for downstream tasks [54].

Multimodal Cognitive Architectures

Cognitive software architectures integrate deep learning with hierarchical processing to enhance multimodal perception and human-AI interaction [12]. The Unified Multimodal Cognitive Architecture (UMCA) demonstrates superior performance on crisis response tasks

through integration of perception, reasoning, and action components [7]. UMCA's modular design facilitates knowledge transfer and maintains interpretability, addressing key challenges in deploying AI systems in high-stakes environments.

3.4.2 Multimodal Learning

Multimodal learning integrates information from multiple sensory modalities to achieve more robust and comprehensive understanding than unimodal approaches. The evidence base revealed four primary themes: cross-modal fusion mechanisms, unified representation learning, multimodal foundation models, and domain-specific multimodal applications.

Cross-Modal Fusion Mechanisms

Cross-modal attention mechanisms enable effective integration of heterogeneous data streams by learning to attend to relevant information across modalities [20], [42], [59]. The Integrated Multimodal AI Architecture unifies text, visual, and audio data streams using cross-modal attention, achieving competitive performance on standard benchmarks [20]. MultiSHAP, a Shapley-based framework, explains cross-modal interactions in multimodal AI models, providing interpretable attribution of predictions to specific modalities and their interactions [44].

Multimodal Foundation Models

Multimodal foundation models represent large-scale pre-trained systems capable of zero-shot generalization across diverse tasks [6], [75], [84]. Benchmarking trustworthiness of multimodal large language models reveals critical gaps in reliability, fairness, and robustness [9]. The Granite Vision model exemplifies lightweight, open-source multimodal models designed for enterprise intelligence applications [55]. Aya Vision advances multilingual multimodal capabilities, addressing the critical need for models that serve diverse linguistic communities [40].

Domain-Specific Applications

Healthcare applications demonstrate the clinical value of multimodal learning. Multimodal deep learning for uncertainty-aware radiation pneumonitis risk prediction integrates imaging and clinical data to improve prognostic accuracy [67]. Autonomous systems benefit from multimodal integration of visual, sensor, and textual data [47], [61]. Software engineering applications leverage multimodal learning for code generation [30] and visual issue repair [32], demonstrating the practical impact of multimodal AI.

3.4.3 Accountability and Explainability

Accountability and explainability mechanisms are essential for trustworthy AI deployment, particularly in high-stakes domains where transparency, auditability, and responsibility attribution are paramount. The evidence base revealed four primary themes: explainability frameworks, accountability architectures, governance-aligned design, and audit mechanisms.

Explainability Frameworks

The M3X-AAL framework introduces multi-modal, multi-tier meta-explainability for trustworthy AI in Ambient Assisted Living systems [4]. The comprehensive survey on explainable and interpretable multimodal large language models synthesizes current approaches for explaining complex multimodal systems [71]. A multimodal XAI framework for trustworthy CNNs and bias detection addresses the challenge of explaining visual models while detecting biases [26]. The GLIMPSE framework offers holistic cross-modal explainability for large vision-language models [83].

Accountability Architectures

Engineering Accountable AI proposes a role-based Requirements Engineering framework that translates abstract accountability principles into concrete, verifiable requirements [13]. The Embedded Accountability Architecture for AI Governance introduces Syntactic Ethics, embedding legal responsibility into AI reasoning with audit trails and cryptographic verification [15]. The TAFES framework operationalizes Transparency, Accountability, Fairness, Ethics, and Safety across the AI lifecycle from design to decommissioning [17].

Audit Mechanisms

A semantic framework supports AI system accountability and audit through ontological foundations for representing accountability [16]. AuditMAI proposes infrastructure for continuous AI auditing, enabling ongoing monitoring throughout the AI lifecycle [80]. The accountability metrics catalogue provides concrete metrics for operationalizing accountability principles, enabling quantitative assessment of transparency, fairness, and robustness [21].

3.4.4 Requirements Engineering

Requirements engineering for AI systems addresses the unique challenges of specifying, validating, and maintaining requirements for non-deterministic, data-dependent, and ethically consequential systems.

Human-Centered Requirements Engineering

Requirements practices and gaps when engineering human-centered AI systems identify critical challenges in eliciting and specifying requirements [78]. Tailoring requirements engineering for responsible AI proposes adaptations to traditional RE practices to accommodate fairness, transparency, and accountability [70]. Requirements Engineering for Accountable and Conscious Human-centered AI emphasizes embedding accountability from the earliest design stages [24]. A vision for value-aligned AI-driven systems articulates principles for ensuring AI systems respect human values [25].

Lifecycle Approaches

Engineering Accountable AI proposes a role-based lifecycle approach assigning accountability to specific roles throughout development and deployment [13]. MLOps as enabler of trustworthy AI emphasizes operational practices for maintaining trustworthiness throughout the AI lifecycle [48]. Incorporating FATES principles in continuous development of ML-integrated systems emphasizes requirements throughout the AI lifecycle [63]. Towards explainability as a functional requirement envisions integrating legal, end-user, and ML engineer perspectives to make explainability a first-class requirement [74].

4. Discussion

4.1 Summary of Main Findings

This systematic review synthesized evidence from 332 papers addressing transformative AI across visual understanding, multimodal learning, accountability, and requirements engineering domains. The synthesis revealed substantial progress in each domain, with increasing integration reflecting the field's maturation toward holistic, trustworthy AI systems.

Visual Understanding AI has advanced from narrow image classification to sophisticated vision-language models capable of cross-modal reasoning and scene understanding. Multimodal Learning demonstrates clear superiority over unimodal approaches through cross-modal attention mechanisms, unified representation learning, and foundation models. Accountability and Explainability mechanisms are essential for trustworthy AI deployment, with effective approaches integrating explainability frameworks, role-based architectures, and governance-aligned design. Requirements Engineering practices are evolving through human-centered approaches, value-aligned design, and lifecycle perspectives.

Cross-cutting themes emerged: (1) Integration and convergence — visual understanding, multimodal learning, accountability, and requirements engineering are interdependent; (2) The gap between principles and practice remains persistent; (3) Interpretability and explainability are essential yet current techniques often provide post-hoc rationalizations; and (4) Human-centered design is increasingly emphasized across all domains.

4.2 Strengths and Limitations

Strengths include: the comprehensive search strategy across four major databases capturing diverse publication types; the interdisciplinary scope spanning four thematic domains; structured data extraction with enriched columns enabling systematic comparison; and transparency in methods following PRISMA-adapted guidelines.

Limitations include: heterogeneity of study designs precluding quantitative meta-analysis; prioritization of top 30 papers for detailed synthesis potentially overlooking contributions

in the broader evidence base; publication bias favouring positive results and novel frameworks; rapid field evolution meaning recent advances may not be fully reflected; and limited empirical validation of many proposed frameworks complicating assessment of practical effectiveness.

4.3 Comparison with Previous Reviews

Several prior systematic reviews have addressed related topics. Wu et al. [14] provided a roadmap on software engineering for responsible AI. The present review extends this work by synthesizing recent advances in multimodal AI and visual understanding. Dang et al. [71] focused specifically on explainability techniques for multimodal systems; the present review situates explainability within the broader context of accountability architectures and requirements engineering. Vogelsang and Borg [82] reviewed emerging RE practices for ML systems; the present review updates and extends this work by incorporating recent accountability frameworks and value-aligned design. Zhang et al. [9] provided comprehensive trustworthiness benchmarking; the present review complements this by synthesizing design and implementation strategies.

4.4 Implications for Research and Practice

Implications for Research

Interpretable-by-Design Architectures: Current post-hoc explainability techniques should give way to architectures inherently designed for interpretability, integrating transparency mechanisms into model design [18], [51].

Unified Governance Ontologies: Research should develop unified ontologies integrating diverse governance frameworks (NIST AI RMF, EU AI Act) into coherent, operationalizable specifications [13], [16].

Comprehensive Evaluation Benchmarks: Future benchmarks should evaluate true visual reasoning, cross-modal understanding, robustness to distribution shifts, and adherence to ethical principles [9], [28], [49].

Empirical Validation Studies: Longitudinal studies are needed to assess effectiveness, usability, and impact of accountability mechanisms and requirements engineering practices in diverse real-world contexts [13], [17], [80].

Efficient Multimodal Architectures: Research should explore efficient architectures, compression techniques, and edge deployment strategies that democratize multimodal AI capabilities [35], [55], [61].

Human-AI Collaboration: Research should investigate requirements engineering practices and accountability mechanisms for human-AI teams [2], [12], [23].

Implications for Practice

Adopt Lifecycle Approaches: Practitioners should adopt frameworks like TAFES [17] or role-based lifecycle approaches [13] that embed accountability at every stage of the AI lifecycle.

Integrate Explainability Early: Explainability should be treated as a functional requirement from project inception rather than an afterthought [74].

Leverage Multimodal Approaches: Where multiple data modalities are available, multimodal learning approaches consistently outperform unimodal alternatives [20], [62], [66].

Implement Role-Based Accountability: Clear assignment of accountability to specific roles facilitates responsibility attribution and supports audit processes [13], [19].

Prioritize Human-Centered Design: Practitioners should employ participatory design methods, conduct user studies, and continuously validate alignment with stakeholder values [24], [25], [70].

Establish Continuous Monitoring: AI system behaviour can drift over time; practitioners should implement continuous monitoring, audit mechanisms, and processes for updating requirements [48], [80].

5. Conclusions

This systematic review synthesized evidence from 332 papers addressing transformative artificial intelligence across visual understanding, multimodal learning, accountability, and requirements engineering domains. The synthesis reveals substantial progress in each domain, with increasing integration reflecting the field's maturation toward holistic, trustworthy AI systems.

Visual understanding AI has advanced from narrow image classification to sophisticated

vision-language models. Multimodal learning approaches demonstrate clear superiority through cross-modal attention, unified representations, and foundation models. Accountability and explainability are recognized as essential for trustworthy deployment. Requirements engineering is evolving through human-centered approaches, value-aligned design, and lifecycle perspectives.

Despite substantial progress, significant gaps persist. The gap between abstract ethical principles and concrete implementation remains wide. Many proposed frameworks lack empirical validation. Current explainability techniques often provide post-hoc rationalizations. Evaluation benchmarks frequently fail to assess genuine multimodal reasoning. Computational demands of large models limit accessibility.

Future research should prioritize interpretable-by-design architectures, unified governance ontologies, comprehensive evaluation benchmarks, empirical validation studies, efficient multimodal architectures, and human-AI collaboration paradigms. The convergence of visual understanding, multimodal learning, accountability, and requirements engineering represents a critical inflection point in AI development. Realizing the transformative potential of trustworthy, human-centered AI requires sustained commitment to rigorous research, responsible practice, and interdisciplinary collaboration among AI researchers, software engineers, ethicists, policymakers, and domain experts.

References

Raza, S. (2025). Humanibench: A human-centric framework for large multimodal models evaluation. ACM. DOI: 10.5281/zenodo.17988191

Dam, K. H., Mahala, G., Hoda, R., Zheng, X., & Conati, C. (2025). Towards autonomous normative multi-agent systems for Human-AI software engineering teams. arXiv:2512.02329. DOI: 10.48550/arxiv.2512.02329

Molli, S. M. (2025). AI Guardrails as a Service (AIGaaS): A Framework for Trustworthy Foundation Model Deployment in Enterprise Systems. International Journal of Collaborative Data-Driven Research and Analytics.

Kasirajan, M., Azhar, M. A. H. B., & Turner, S. (2025). M3X-AAL: A Multi-Modal, Multi-Tier, MetaExplainability Framework for Trustworthy AI in Ambient Assisted Living Systems. Proceedings of CINTI, 747-754. DOI: 10.1109/cinti67731.2025.11311739

Warburton, D. M. (2025). Multimodal transformer framework for self-supervised, zero-shot generalization across vision, language, and reasoning tasks. Zenodo. DOI: 10.5281/zenodo.15662593

Akinyele, A. R. et al. (2024). Multimodal Foundation Models for Unified Image, Video and Text Understanding. Open Access Research Journal of Science and Technology, 12(2), 081-093. DOI: 10.53022/oarjst.2024.12.2.0139

Jani, J. (2025). Unified Multimodal Cognitive Architecture (UMCA): an Integrated Framework for Perception, Reasoning, and Action for High-Stakes Environments. International Journal of Environmental Sciences, 4616-4620. DOI: 10.64252/ecs0xk27

Moujdi, A. (2026). An Explainability-Driven Framework for Interpretable Cross-Modal Image-Text Retrieval Using CLIP. International Journal of Language, Technology & Educational Management, 15(3).

Zhang, Y. et al. (2024). Benchmarking Trustworthiness of Multimodal Large Language Models: A Comprehensive Study. arXiv:2406.07057. DOI: 10.48550/arxiv.2406.07057

Accountability of AI-based Algorithmic Decision-Making Systems — from Theory to Software Engineering Practice. (2025). DOI: 10.26204/kluedo/9188

Ghosh, R. et al. (2025). CompAgent: An Agentic Framework for Visual Compliance Verification. arXiv:2511.00171. DOI: 10.48550/arxiv.2511.00171

Cui, J. (2025). Cognitive Software Architectures for Multimodal Perception and Human-AI Interaction. Preprints. DOI: 10.20944/preprints202505.0841.v1

Hadar, E. (2025). Engineering Accountable AI: A Role-Based Lifecycle Approach. Proceedings of REW, 419-424. DOI: 10.1109/rew66121.2025.00061

Wu, Q., Zhu, L., Xu, X., Whittle, J., & Xing, Z. (2022). Towards a Roadmap on Software Engineering for Responsible AI. Proceedings of SAIN, 101-112. DOI: 10.1145/3522664.3528607

Embedded Accountability Architecture for AI Governance: The Syntactic Ethics Framework. (2025). Zenodo. DOI: 10.5281/zenodo.15788326

Naja, I., Markovic, M., Edwards, P., & Cottrill, C. (2021). A Semantic Framework to Support AI System Accountability and Audit. Lecture Notes in Computer Science, 160-176. DOI: 10.1007/978-3-030-77385-4_10

Loucif, S., Sharma, R., Kshetri, N., & Zahid, A. (2025). From design to decommissioning: TAFES framework for responsible AI. DOI: 10.55092/let20250009

Principles and Methods for Transparency, Interpretability, Trust, Accountability, and Responsibility in AI-Based Predictive Modelling. (2025). Computer Science, Engineering and Technology, 3(3), 69-80. DOI: 10.46632/cset/3/3/10

Persona Architecture: Designing Role-Specific AI Systems for Accountability and Trust. (2025). Zenodo. DOI: 10.5281/zenodo.17490306

Chitturi, N. (2025). Integrated Multimodal AI Architecture: Cross-Modal Attention Mechanisms Unifying Text, Visual, And Audio Data Streams For Enterprise Content Analysis. Journal of International Crisis and Risk Communication Research, 78-89. DOI: 10.63278/jicrcr.vi.3203

Xia, B. et al. (2023). From Principles to Practice: An Accountability Metrics Catalogue for Managing AI Risks. arXiv:2311.13158. DOI: 10.48550/arxiv.2311.13158

Navneet, S. K., & Chandra, J. (2025). Rethinking Autonomy: Preventing Failures in AI-Driven Software Engineering. arXiv:2508.11824. DOI: 10.48550/arxiv.2508.11824

Zhu, L. (2025). Designing for Meaningful Oversight: Human and Organisational Agency in Multimodal AI Systems. Proceedings of AI & HCI, 2-2. DOI: 10.1145/3716553.3757097

Hadar, I., Kantsepolsky, B., Sadovski, E., & Sherman, S. (2025). Requirements Engineering for Accountable and Conscious Human-centered AI: Preface. Proceedings of REW, 402-404. DOI: 10.1109/rew66121.2025.00058

Obie, H. O. (2025). A Vision for Value-Aligned AI-Driven Systems. Proceedings of AIWare, 143-148. DOI: 10.1109/aiware69974.2025.00023

A Multimodal XAI Framework for Trustworthy CNNs and Bias Detection in Deep Representation Learning. (2025). arXiv:2510.12957. DOI: 10.48550/arxiv.2510.12957

Toward Accountable and Explainable Artificial Intelligence Part Two: The Framework Implementation. (2022). TechRxiv. DOI: 10.36227/techrxiv.19102094.v1

Unsal, M., & Akkus, A. (2025). EasyARC: Evaluating Vision Language Models on True Visual Reasoning. arXiv:2506.11595.

Hu, B. C. et al. (2022). If a Human Can See It, So Should Your System: Reliability Requirements for Machine Vision Components. Proceedings of ICSE, 1145-1156. DOI: 10.1145/3510003.3510109

Chai, L. et al. (2025). Multilingual Multimodal Software Developer for Code Generation. arXiv:2507.08719. DOI: 10.48550/arxiv.2507.08719

Tariq, S. et al. (2025). From Prediction to Explanation: Multimodal, Explainable, and Interactive Deepfake Detection Framework for Non-Expert Users. Proceedings of MM. DOI: 10.1145/3746027.3755786

Huang, K. M., Zhang, J., & Chen, C. (2025). Seeing is Fixing: Cross-Modal Reasoning with Multimodal LLMs for Visual Software Issue Repair. Proceedings of ASE, 1156-1168. DOI: 10.1109/ase63991.2025.00100

Caffagni, D. et al. (2024). Wiki-LLaVA: Hierarchical Retrieval-Augmented Generation for Multimodal LLMs. arXiv:2404.15406.

Palit, V. et al. (2023). Towards Vision-Language Mechanistic Interpretability: A Causal Tracing Tool for BLIP. arXiv:2308.14179.

Rane, J. et al. (2024). Scalable and adaptive deep learning algorithms for large-scale machine learning systems. DOI: 10.70593/978-81-981271-0-5_2

Goncalves, A., & Correia, A. (2025). Engineering Explainable AI Systems for GDPR-Aligned Decision Transparency: A Modular Framework for Continuous Compliance. Journal of Cybersecurity and Privacy. DOI: 10.3390/jcp6010007

Liu, Z. et al. (2025). Vision-Language Memory for Spatial Reasoning. arXiv:2511.20644.

Han, X. et al. (2026). Vision as Unified Multimodal Generation. arXiv:2607.06560.

Sanogo, K., & Ardiccioni, R. (2025). Toward More Reliable Artificial Intelligence: Reducing Hallucinations in Vision-Language Models. arXiv:2512.07564.

Dash, S. et al. (2025). Aya Vision: Advancing the Frontier of Multilingual Multimodality. arXiv:2505.08751.

Lee, E. et al. (2026). Hierarchical Pre-Training of Vision Encoders with Large Language Model. arXiv:2604.00086.

Gerami, A., Madani, S., & Duraiswami, R. (2026). On The Application of Linear Attention in Multimodal Transformers. arXiv:2604.10064.

Prasad, D. N., Ullrich, L., & Graichen, K. (2026). Inference-Time Attention Steering for Vision-Language-Action Driving Models. arXiv:2608.17095.

Wang, Z., & Wang, K. (2025). MultiSHAP: A Shapley-Based Framework for Explaining Cross-Modal Interactions in Multimodal AI Models. arXiv:2508.00576. DOI: 10.48550/arxiv.2508.00576

Ma, K. et al. (2022). Towards Trustworthy Healthcare AI: Attention-Based Feature Learning for COVID-19 Screening With Chest Radiography. arXiv:2207.09312.

Shuang, A. (2025). Building Trustworthy AI: Uncertainty Quantification and Failure Detection in Large Vision-Language Models. DOI: 10.21954/ou.ro.00102247

Zarghani, A., Ebrahimi, A., & Malekesfandiari, A. (2025). Multimodal Framework for Explainable Autonomous Driving. arXiv:2507.07938. DOI: 10.48550/arxiv.2507.07938

Billeter, Y. et al. (2024). MLOps as Enabler of Trustworthy AI. Proceedings of SDS, 37-40. DOI: 10.1109/sds60720.2024.00013

Xing, S. et al. (2024). AutoTrust: Benchmarking Trustworthiness in Large Vision Language Models for Autonomous Driving. arXiv:2412.15206. DOI: 10.48550/arxiv.2412.15206

Daly, E. et al. (2024). Usage Governance Advisor: from Intent to AI Governance. arXiv:2412.01957. DOI: 10.48550/arxiv.2412.01957

Wang, H., Tan, S., & Wang, H. (2024). Probabilistic Conceptual Explainers: Trustworthy Conceptual Explanations for Vision Foundation Models. arXiv:2406.12649.

Cafagna, M. et al. (2023). Interpreting vision and language generative models with semantic visual priors. Frontiers in Artificial Intelligence, 6. DOI: 10.3389/frai.2023.1220476

Cai, Q. et al. (2026). HiDream-O1-Image: A Natively Unified Image Generative Foundation Model. arXiv:2605.11061.

Gabeur, V. et al. (2026). Image Generators are Generalist Vision Learners. arXiv:2604.20329.

Granite Vision Team et al. (2025). Granite Vision: a lightweight, open-source multimodal model for enterprise Intelligence. arXiv:2502.09927.

Hutchinson, B. et al. (2021). Towards Accountability for Machine Learning Datasets. Proceedings of FAccT, 560-575. DOI: 10.1145/3442188.3445918

Trustworthy AI: From Principles to Practices. (2023). ACM Computing Surveys, 55(9), 1-46. DOI: 10.1145/3555803

[58] Bhuvaneswari, U. (2025). Architecting Responsible Development and Deployment of Generative AI. International Journal for Science Technology and Engineering, 13(5), 4937-4956. DOI: 10.22214/ijraset.2025.71314

Ghadiya, A. et al. (2024). Cross-Modal Fusion and Attention Mechanism for Weakly Supervised Video Anomaly Detection. arXiv:2412.20455. DOI: 10.48550/arxiv.2412.20455

Lu, Q. et al. (2022). Responsible AI Pattern Catalogue: A Collection of Best Practices for AI Governance and Engineering. ACM Computing Surveys. DOI: 10.1145/3626234

Ye, Z. et al. (2026). Edge-Assisted Multimodal UAV Localization with Resource-Efficient Compression and Robust Fusion. arXiv:2607.03111.

Liang, P. P. (2024). Foundations of Multisensory Artificial Intelligence. arXiv:2404.18976. DOI: 10.48550/arxiv.2404.18976

Bruel, J. M. et al. (2025). Incorporating FATES Principles in Continuous Development of ML-Integrated Systems. Proceedings of REW, 425-431. DOI: 10.1109/rew66121.2025.00062

Li, Y. et al. (2021). X-modaler: A Versatile and High-performance Codebase for Cross-modal Analytics. Proceedings of ACM MM, 3799-3802. DOI: 10.1145/3474085.3478331

Hauer, M. P., Krafft, T. D., & Zweig, K. A. (2023). Overview of transparency and inspectability mechanisms to achieve accountability of AI systems. Data & Policy. DOI: 10.1017/dap.2023.30

Huang, L. et al. (2023). Deep evidential fusion with uncertainty quantification for multimodal medical image segmentation. arXiv:2309.05919.

Yang, J. et al. (2026). Multimodal Deep Learning for Uncertainty-Aware Radiation Pneumonitis Risk Prediction. arXiv:2608.28937.

Neubauer, T. et al. (2020). Soft Tissue Sarcoma Co-Segmentation in Combined MRI and PET/CT Data. arXiv:2008.12544.

Petrovic, N. et al. (2025). Multi-modal Summarization in Model-Based Engineering: Automotive Software Development Case Study. arXiv:2503.04506.

Maalej, W., Pham, Y. D., & Chazette, L. (2023). Tailoring Requirements Engineering for Responsible AI. IEEE Computer, 56, 18-27. DOI: 10.1109/MC.2023.3243182

Dang, Y. et al. (2024). Explainable and Interpretable Multimodal Large Language Models: A Comprehensive Survey. arXiv:2412.02104. DOI: 10.48550/arxiv.2412.02104

Khan, M., & Vice, J. (2022). Toward Accountable and Explainable Artificial Intelligence Part One: Theory and Examples. IEEE Access, 10, 99686-99701. DOI: 10.1109/ACCESS.2022.3207812

Villamizar, H., Mendez, D., & Kalinowski, M. (2025). Towards a Framework for Operationalizing the Specification of Trustworthy AI Requirements. arXiv:2507.10228. DOI: 10.48550/arxiv.2507.10228

Habiba, U., Bogner, J., & Wagner, S. (2024). Towards Explainability as a Functional Requirement. Proceedings of SANER. DOI: 10.1145/3643691.3648590

Yijun, H., & Zhihao, S. (2025). An In-Depth Survey of Multimodal Foundation Models and Their Challenges. DOI: 10.31224/4752

Kollipara, Y. V. P. (2025). Assured, Explainable, And Auditable AI For High-Stakes Decisions. Journal of International Crisis and Risk Communication Research, 554-566. DOI: 10.63278/jicrcr.vi.3392

Cheng, E. C., Cheng, J., & Siu, A. (2026). Toward Safe and Responsible AI Agents: A Three-Pillar Model. arXiv:2601.06223. DOI: 10.48550/arxiv.2601.06223

Requirements Practices and Gaps When Engineering Human-Centered Artificial Intelligence Systems. (2023). arXiv:2301.10404. DOI: 10.48550/arxiv.2301.10404

Doris, A. C. et al. (2024). DesignQA: A Multimodal Benchmark for Evaluating LLMs' Understanding of Engineering Documentation. Journal of Computing and Information Science in Engineering, 1-16. DOI: 10.1115/1.4067333

Waltersdorfer, L. et al. (2024). AuditMAI: Towards An Infrastructure for Continuous AI Auditing. arXiv:2406.14243. DOI: 10.48550/arxiv.2406.14243

Bruni, E. et al. (2012). Automatic analysis of multimodal requirements: a research preview. Proceedings of REFSQ, 218-224. DOI: 10.1007/978-3-642-28714-5_19

Requirements Engineering for Machine Learning: A Review and Reflection. (2022). arXiv:2210.00859. DOI: 10.48550/arxiv.2210.00859

Shen, G. (2025). GLIMPSE: Holistic Cross-Modal Explainability for Large Vision-Language Models. arXiv:2506.18985.

Arjunan, G. (2024). AI Beyond Text: Integrating Vision, Audio, and Language for Multimodal Learning. Zenodo. DOI: 10.5281/zenodo.14287142

International Journal of Advanced Research