

Multimodal Scientific Claim Verification Using Large Language Models and Retrieval-Augmented Generation

Milan J. Raiyani¹, Mayur M. Jani²

¹ PG Scholar, School of Engineering and Technology, Dr. Subhash University, Junagadh, Gujarat, India
(ORCID: 0009-0002-1894-7553, Email: milan.raiyani@dsuni.ac.in)

² Assistant Professor, IT Department, School of Engineering and Technology, Dr. Subhash University, Junagadh, Gujarat, India (ORCID: 0009-0001-0272-3793, Email: mayur.jani@dsuni.ac.in)

Abstract - Scientific claim verification is essential for evidence-based scientific communication, yet decisive evidence in research papers is often distributed across prose, tables, figures, captions, equations, and page-level layout. Text-only verification can therefore retrieve relevant sentences while overlooking the visual or structural evidence needed to interpret a claim correctly. This paper proposes a layout-aware multimodal Retrieval-Augmented Generation (RAG) framework for verifying complex scientific claims from complete research PDFs. The proposed pipeline preserves document structure, extracts and links textual, tabular, and visual evidence, performs hybrid lexical-semantic retrieval with metadata and visual signals, and applies staged cross-modal reasoning using Large Language Models (LLMs) and Vision-Language Models (VLMs). Each verification output is designed to contain a decision—Supported, Refuted, or Not Enough Evidence—together with traceable source-page references and uncertainty information. A comparative literature analysis covering 30 relevant studies shows that existing work provides strong foundations for text retrieval, table reasoning, document layout understanding, multimodal verification, and hallucination control, but these capabilities remain fragmented across tasks and benchmarks. The proposed Phase-I work therefore focuses on an integrated architecture and a reproducible evaluation plan rather than reporting unperformed experiments. The planned evaluation will compare text-only, text-plus-table, layout-unaware multimodal, and complete layout-aware multimodal configurations using classification, evidence retrieval, citation correctness, hallucination, abstention, latency, and explainability measures.

Keywords - Claim Verification, Evidence Grounding, Large Language Models, Multimodal Retrieval-Augmented Generation, Scientific PDFs, Table Understanding, Vision-Language Models.

1. INTRODUCTION

The volume, heterogeneity, and structural complexity of scientific literature make reliable verification of factual and empirical claims increasingly difficult. A claim may depend on a sentence in the discussion, a numerical value in a table, a trend in a plot, a definition in a caption, or a relationship between information located on different pages. SciFact formalized scientific claim verification as a retrieval-and-reasoning problem in which systems retrieve evidence and identify rationales supporting or refuting a claim [1]. VerT5erini subsequently integrated retrieval, sentence

selection, and label prediction in a sequence-to-sequence framework [2], while MultiVerS explored full-document context and weak supervision for scientific verification [3]. These studies established a strong text-centered foundation, but they do not by themselves represent the complete visual and structural evidence available in scientific papers.

The challenge becomes more explicit when claims require table or figure interpretation. SCITAB introduced expert-verified claims grounded in scientific tables and demonstrated that compositional reasoning over tabular evidence remains difficult for modern models [5]. ChartCheck addressed fact-checking over chart images, while newer multimodal benchmarks such as SciVer and MuSciClaims evaluate claim verification with visual evidence [6]-[8]. MuSciClaims reports that even strong vision-language models can struggle with evidence localization and cross-modal aggregation [8]. SciClaimEval further broadens the setting by providing authentic claims with figures and tables from 180 papers across three domains and 1,664 annotated samples [9]. These findings reinforce the need to preserve the relationships among textual, tabular, visual, and layout information.

Retrieval-Augmented Generation provides a natural mechanism for grounding model decisions in external evidence rather than relying only on parametric knowledge [12]. However, retrieval quality depends on the representation of the source document. Dense retrieval methods such as Dense Passage Retrieval and Contriever improve semantic matching, while ColBERT supports fine-grained late-interaction retrieval [13]-[15]. In parallel, document-understanding research has demonstrated the importance of two-dimensional layout and visual context through LayoutLM, LayoutLMv2, LayoutLMv3, DocLayNet, PubLayNet, DocFormer, Donut, and Pix2Struct [19]-[25]. Table-specific and chart-specific methods, including TAPAS, Table Transformer, DePlot, and ChartQA, further demonstrate that structured numerical evidence requires specialized representations [26]-[29].

Recent reliability-oriented RAG research also shows that retrieval alone does not guarantee faithful generation. Self-RAG introduces adaptive retrieval and self-reflection, while RAGTruth provides a corpus for analyzing hallucinations in retrieval-augmented generation [16], [17]. These developments motivate an evidence-first verification pipeline in which the system can abstain when evidence is insufficient and can expose the provenance of the evidence used for its decision. The present work therefore targets an integrated scientific-PDF verification framework rather than an unrestricted question-answering system.

2. LITERATURE REVIEW AND COMPARATIVE ANALYSIS

The literature was organized into five capability groups: (i) scientific claim verification, (ii) multimodal claim and chart/table verification, (iii) document-layout understanding, (iv) retrieval and RAG, and (v) reliability and hallucination control. Table 1 summarizes representative quantitative facts reported by selected benchmark studies. These values are reported from the cited works and are not results produced by the proposed system.

Study / benchmark	Year	Evidence / scale	Reported finding relevant to this work
SciFact [1]	2020	1.4K expert-written claims	Established scientific claim verification with evidence and rationales
VerT5erini [2]	2021	SciFact + COVID-19 sets	Joint retrieval, sentence selection and label prediction
MultiVerS [3]	2022	Full-document context	Uses document-level context and weak supervision
SciFact-Open [4]	2022	500K research abstracts	Smaller-corpus systems showed ≥ 15 F1 drop on open-domain evaluation
SCITAB [5]	2023	1.2K expert-verified claims	Table grounding and compositional reasoning remain difficult
MuSciClaims [8]	2025	Multimodal benchmark	Best reported model reached about 0.72 F1 in the cited ACL version
SciClaimEval [9]	2026	1,664 samples / 180 papers / 3 domains	Figure-based verification remains challenging; 11 multimodal models evaluated

Table 1. Selected quantitative characteristics of established scientific claim-verification benchmarks.

A second comparison is useful for identifying the integration gap. Existing studies often optimize one evidence modality or one stage of the pipeline. Scientific claim-verification datasets focus on retrieval and reasoning; document-understanding models preserve layout; table and chart systems interpret structured visual data; and RAG reliability methods control unsupported generation. The proposed research combines these capabilities at the scientific-PDF level, with explicit provenance and uncertainty as first-class outputs. [10]

Capability	Representative studies	Primary strength	Remaining limitation for this research
Text claim verification	SciFact, VerT5erini, MultiVerS	Evidence retrieval and rationale selection	Primarily text/document-level evidence
Open-domain retrieval	SciFact-Open, DPR, ColBERT, Contriever	Large-scale semantic/lexical retrieval	Does not preserve complete PDF multimodality

Table reasoning	SCITAB, TAPAS, Table Transformer	Structured numerical reasoning	Limited integration with figures and page layout
Chart/visual reasoning	ChartCheck, ChartQA, DePlot	Visual and numerical chart interpretation	Often benchmark-specific rather than full-paper verification
Document layout	LayoutLM family, DocLayNet, PubLayNet, DocFormer	Spatial and structural representation	Not designed specifically for claim-verification decisions
Multimodal claim verification	SciVer, MuSciClaims, SciClaimEval, ToolSciVer	Cross-modal evidence and reasoning	Evidence integration, localization and reliability remain difficult
RAG reliability	Self-RAG, RAGTruth	Retrieval-aware generation and hallucination analysis	General RAG rather than scientific-PDF evidence graphs

Table 2. Capability-level comparison used to derive the research gap.

3. RESEARCH OBJECTIVE AND SPECIFIC OBJECTIVES

Main Objective: To design and evaluate a layout-aware multimodal Retrieval-Augmented Generation framework that verifies complex scientific claims by jointly retrieving and reasoning over textual passages, structured tables, figures, captions, and related mathematical or visual evidence from complete scientific PDF documents.

Specific Objectives:

- (1) To develop a scientific-PDF processing pipeline that preserves page layout, section hierarchy, captions, tables, figures, and mathematical content where technically feasible.
- (2) To construct a unified evidence representation and multimodal index containing lexical and dense text representations, structured table representations, visual representations, and links to original page locations.
- (3) To implement hybrid lexical, semantic, visual, and metadata-aware evidence retrieval with a reranking stage.
- (4) To design a staged multi-hop verification procedure that connects evidence across pages and modalities before producing a final decision.
- (5) To generate evidence-grounded outputs containing the verification label, supporting or contradicting evidence, page references, confidence indicators, and uncertainty information.
- (6) To evaluate the proposed framework against text-only and single-modality baselines using classification, evidence retrieval, reliability, latency, and explainability measures.

4. PROBLEM STATEMENT AND RESEARCH GAP

Existing scientific claim-verification systems frequently assume that decisive evidence can be represented as a text passage or a single structured object. This assumption is inadequate for claims whose meaning depends on multiple components of a complete paper. For example, a statement that one method outperforms a baseline may require a results paragraph, a metric definition, a table containing several columns, and a figure or caption that clarifies the experimental setting. SciFact-Open also demonstrates that

systems trained on smaller corpora can lose performance when moved to a substantially larger open-domain evidence collection [4]. Table-grounded and multimodal benchmarks further reveal difficulties in evidence localization and compositional reasoning [5], [8], [9].

Research Gap: The literature does not indicate a single dissertation-scale scientific-PDF pipeline that explicitly unifies five requirements: (1) preservation of scientific-document layout; (2) linked multimodal evidence units; (3) hybrid lexical-semantic-visual retrieval; (4) explicit multi-hop cross-modal verification; and (5) evidence-grounded, uncertainty-aware reporting. Importantly, the proposed contribution does not claim that each individual component is novel in isolation. The research contribution is the design, integration, and later empirical evaluation of these components as one traceable verification workflow for complete scientific papers.

5. PROPOSED METHODOLOGY

5.1 Scientific PDF Ingestion and Layout Preservation

The framework accepts a complete scientific PDF and produces a structured document representation. The processing layer combines text extraction, page segmentation, figure and table detection, table-structure recovery, caption association, equation preservation where feasible, and page-level metadata. Each evidence unit retains source page, section title, modality, and bounding-region information when available. Layout-aware representation is motivated by document-understanding research showing that spatial and structural information is useful for visually rich documents [19]–[23], [20]–[25]

5.2 Claim Extraction and Normalization

The system supports two claim-entry modes. A researcher may submit a specific claim, or a claim-extraction stage may identify atomic scientific claims from selected document content. Each claim is normalized into entities, relations, quantities, comparison operators, and candidate evidence types. Decomposing complex claims into atomic units is intended to reduce ambiguity and make retrieval and verification traceable.

5.3 Multimodal Evidence Representation and Indexing

Text units are indexed using lexical terms and dense embeddings. Tables are represented both as structured rows and columns and as rendered visual regions, allowing numerical relationships to be processed through structured and visual representations. Figures and charts are stored with captions and page locations and represented through visual embeddings. A lightweight metadata graph links document components, such as a figure with its caption, a table with its discussion paragraph, and a claim with candidate evidence units. [26], [27] [30]

5.4 Hybrid Multimodal Retrieval

Evidence retrieval combines lexical matching, dense semantic retrieval, metadata filtering, and visual similarity. The first stage retrieves a broad candidate set. A reranking stage considers claim entities, numerical terms, section context, modality compatibility, and cross-reference relationships. The design is consistent with evidence that lexical and semantic retrieval provide complementary retrieval signals and that dense retrieval is useful for open-domain evidence discovery [4], [13]–[15]. [13], [14], [15]

5.5 Cross-Modal Multi-Hop Verification

Verification proceeds in stages. First, the system identifies the claim type and evidence requirements. Second, it inspects relevant text, table, and figure candidates. Third, it reconciles numerical values, units, labels, trend directions, and comparison operators. Fourth, it determines whether the combined evidence supports, contradicts, or fails to resolve the claim. The final classifier returns Supported, Refuted, or Not Enough Evidence (NEI) and attaches the evidence snippets or visual regions used for the decision.

5.6 Hallucination and Reliability Control

Reliability is treated as a design requirement. The proposed system will use evidence-only prompting, source-page citations, numerical consistency checks, contradiction checks, and explicit abstention when decisive evidence is missing. Self-reflective retrieval ideas and hallucination-focused evaluation such as Self-RAG and RAGTruth provide relevant methodological guidance [16], [17]. The verifier will not be forced to choose Supported or Refuted when the available document evidence is insufficient. [16]–[18]

6. PROPOSED ARCHITECTURE

The proposed six-stage architecture consists of: (1) scientific research PDF input; (2) PDF parsing and layout preservation; (3) claim extraction and normalization; (4) multimodal RAG evidence retrieval and reranking; (5) LLM/VLM cross-modal reasoning; and (6) final scientific claim verification with evidence and uncertainty. The architecture treats paragraphs, tables, figures, captions, and metadata as connected evidence rather than independent inputs. The expected output is a traceable verdict linked to the source evidence.

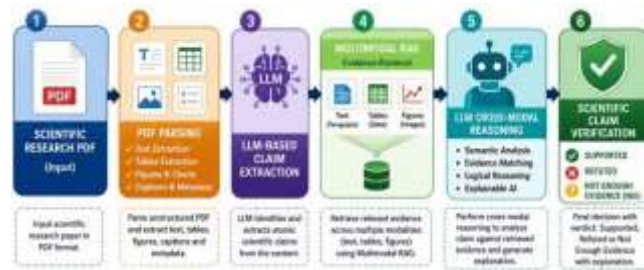


Fig. 1. Proposed Layout-Aware Multimodal Framework For Scientific Claim Verification.

Stage	Input / operation	Output
1	Scientific PDF ingestion	Page images, text and document metadata
2	Layout and structure parsing	Text blocks, tables, figures, captions, page regions
3	Claim extraction and normalization	Atomic claim representation + evidence requirements
4	Hybrid multimodal retrieval	Ranked text/table/figure evidence candidates
5	Cross-modal LLM/VLM reasoning	Evidence reconciliation and verification reasoning
6	Decision and provenance layer	Supported / Refuted / NEI + page references + uncertainty

Table 3. Proposed six-stage architecture and expected intermediate outputs.

7. EVALUATION PLAN

7.1 Datasets and Evaluation Setting

The evaluation will use publicly available datasets where task definitions and licensing permit. SciFact provides text-based scientific claim verification [1]; SciFact-Open evaluates open-domain verification over 500K research abstracts [4]; SCITAB targets table-grounded claims [5]; SciVer and MuSciClaims address multimodal scientific claim verification [7], [8]; ChartCheck and ChartQA provide chart-centric evidence settings [6], [29]; and SciClaimEval provides a newer cross-modal setting with authentic claims, figures, and tables [9]. A supplementary manually annotated collection of complete scientific PDFs may be created for claims that require cross-page and cross-modal evidence.

7.2 Metrics and Baselines

Evaluation dimension	Planned metric	Purpose
Classification	Accuracy, Precision, Recall, Macro-F1	Measure final claim-verification quality
Evidence retrieval	Evidence Precision, Evidence Recall, Recall@K	Measure whether decisive evidence is retrieved
Citation/provenance	Citation correctness, page-reference accuracy	Measure traceability of decisions
Reliability	Hallucination rate, abstention quality	Measure unsupported decisions and safe refusal
Efficiency	Average response latency	Measure practical system cost
Explainability	Evidence sufficiency and qualitative expert assessment	Assess whether decisions are understandable and auditable

Table 4. Planned evaluation dimensions and metrics. No numerical result in this table is claimed as an experiment of the present study.

7.3 Baselines and Ablation Studies

The main comparisons will include: (1) text-only retrieval and verification; (2) text-plus-table verification; (3) multimodal retrieval without layout metadata; and (4) the complete proposed framework. Ablation experiments will remove one design component at a time, including page/layout metadata, table structure, figure representations, hybrid retrieval, reranking, and multi-hop reasoning. Error analysis will categorize failures into retrieval misses, table-structure errors, visual-localization errors, numerical inconsistencies, cross-modal reasoning errors, and unsupported generation.

8. EXPECTED CONTRIBUTIONS

- (1) A unified layout-aware representation for scientific-PDF evidence.
- (2) A multimodal retrieval strategy linking textual, tabular, and visual evidence.
- (3) A staged multi-hop verification procedure for claims requiring evidence across modalities and pages.
- (4) An evidence-grounded output format with page-level traceability and uncertainty handling.
- (5) An experimental comparison with text-only and single-modality baselines.
- (6) A reproducible evaluation protocol suitable for later dissertation implementation and experimentation.

9. LIMITATIONS AND ETHICAL CONSIDERATIONS

The proposed system is intended to assist researchers rather than replace peer review or expert scientific judgment. PDF parsing may introduce OCR, table-extraction, or figure-association errors. Dense charts, small labels, mathematical notation, incomplete documents, and ambiguous claims may reduce reliability. The system must therefore communicate uncertainty and preserve source references. Dataset licenses, copyright and author rights, privacy considerations, and responsible handling of scientific documents will be respected. Claims related to medical, legal, safety-critical, or other high-consequence decisions should receive appropriate expert review before action is taken.

10. CONCLUSION

This paper presents a dissertation-stage research framework for multimodal scientific claim verification using Large Language Models and Retrieval-Augmented Generation. The proposed approach extends text-centered scientific claim verification toward complete scientific PDFs by preserving layout, linking multimodal evidence units, applying hybrid retrieval, and performing staged cross-modal reasoning. The literature indicates that scientific claim verification, table reasoning, visual evidence localization, document-layout understanding, and RAG reliability are active but fragmented areas [1]-[30]. The proposed research addresses the integration gap by treating evidence provenance and uncertainty as explicit components of the verification workflow. Because this work is at Dissertation Phase-I, the manuscript deliberately reports the proposed methodology and evaluation plan rather than claiming experimental results that have not yet been obtained. Subsequent phases will implement the system, conduct controlled experiments and ablations, and report measured outcomes against the defined baselines and metrics. [11]

REFERENCES

1. Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., and Hajishirzi, H., —Fact or Fiction: Verifying Scientific Claims,| Proceedings of EMNLP, pp. 7534–7550, 2020.
2. Pradeep, R., Ma, X., Nogueira, R., and Lin, J., —Scientific Claim Verification with VerT5erini,| Proceedings of the 12th Workshop on Health Text Mining and Information Analysis, pp. 94–103, 2021.
3. Wadden, D., Lo, K., Wang, L. L., Cohan, A., Beltagy, I., and Hajishirzi, H., —MultiVerS: Improving Scientific Claim Verification with Weak Supervision and Full-Document Context,| Findings of NAACL, pp. 61–76, 2022.
4. Wadden, D., Lo, K., Kuehl, B., Cohan, A., Beltagy, I., Wang, L. L., and Hajishirzi, H., —SciFact-Open: Towards Open-Domain Scientific Claim Verification,| Findings of EMNLP, pp. 4719–4734, 2022.
5. Lu, X., Pan, L., Liu, Q., Nakov, P., and Kan, M.-Y., —SCITAB: A Challenging Benchmark for Compositional Reasoning and Claim Verification on Scientific Tables,| Proceedings of EMNLP, pp. 7787–7813, 2023.
6. Akhtar, M., Subedi, N., Gupta, V., Tahmasebi, S., Cocarascu, O., and Simperl, E., —ChartCheck: Explainable Fact-Checking over Real-World Chart Images,| Findings of ACL, pp. 13921–13937, 2024.
7. Wang, C., Shen, Y., Kuang, Z., Cohan, A., and Zhao, Y., —SciVer: Evaluating Foundation Models for Multimodal Scientific Claim Verification,| Proceedings of ACL, pp. 8562–8579, 2025.
8. Lal, Y. K., Bandham, M., Hasan, M. S., Kashi, A., Koupaei, M., and Balasubramanian, N., —MuSciClaims: Multimodal Scientific Claim Verification,| Proceedings of IJCNLP-AAACL, pp. 3285–3307, 2025.

9. Ho, X., Wu, Y.-A., Kumar, S., Xia, T. C., Boudin, F., Greiner-Petter, A., and Aizawa, A., —SciClaimEval: Cross-modal Claim Verification in Scientific Papers,| Proceedings of LREC, 2026.
10. Ho, X., Kumar, S., Wu, Y.-A., Boudin, F., Takasu, A., and Aizawa, A., —Table-Text Alignment: Explaining Claim Verification Against Tables in Scientific Papers,| Findings of EMNLP, pp. 2509–2517, 2025.
11. Zhou, B., Shi, P., Kamoi, R., Zhang, N., and Zhang, R., —ToolSciVer: Multimodal Scientific Claim Verification with Visual Tool Augmented Reinforcement Learning,| arXiv:2607.16131, 2026.
12. Lewis, P., Perez, E., Piktus, A., et al., —Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks,| Advances in Neural Information Processing Systems, vol. 33, 2020.
13. Karpukhin, V., Oguz, B., Min, S., et al., —Dense Passage Retrieval for Open-Domain Question Answering,| Proceedings of EMNLP, pp. 6769–6781, 2020.
14. Khattab, O., and Zaharia, M., —ColBERT: Efficient and Effective Passage Search via Contextualized Late Interaction over BERT,| Proceedings of SIGIR, pp. 39–48, 2020.
15. Izacard, G., Caron, M., Hosseini, L., et al., —Unsupervised Dense Information Retrieval with Contrastive Learning,| Transactions on Machine Learning Research, 2022.
16. Asai, A., Wu, Z., Wang, Y., Sil, A., and Hajishirzi, H., —Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection,| Proceedings of ICLR, 2024.
17. Niu, C., Wu, Y., Zhu, J., Xu, S., Shum, K., Zhong, R., Song, J., and Zhang, T., —RAGTruth: A Hallucination Corpus for Developing Trustworthy Retrieval-Augmented Language Models,| Proceedings of ACL, pp. 10862–10878, 2024.
18. Yan, S.-Q., Gu, J.-C., Zhu, Y., and Ling, Z.-H., —Corrective Retrieval Augmented Generation,| arXiv:2401.15884, 2024.
19. Xu, Y., Li, M., Cui, L., Huang, S., Wei, F., and Zhou, M., —LayoutLM: Pre-training of Text and Layout for Document Image Understanding,| Proceedings of KDD, pp. 1192–1200, 2020.
20. Xu, Y., Xu, Y., Lv, T., et al., —LayoutLMv2: Multi-modal Pre-training for Visually-rich Document Understanding,| Proceedings of ACL-IJCNLP, pp. 2579–2591, 2021.
21. Huang, Y., Lv, T., Cui, L., Lu, Y., and Wei, F., —LayoutLMv3: Pre-training for Document AI with Unified Text and Image Masking,| Proceedings of ACM Multimedia, pp. 4083–4091, 2022.
22. Pfizmann, B., Auer, C., Dolfi, M., Nassar, A. S., and Staar, P., —DocLayNet: A Large Human-Annotated Dataset for Document-Layout Segmentation,| Proceedings of KDD, pp. 3743–3751, 2022.
23. Zhong, X., Tang, J., and Jimeno-Yepes, A., —PubLayNet: Largest Dataset Ever for Document Layout Analysis,| Proceedings of ICDAR, pp. 1015–1022, 2019.
24. Appalaraju, S., Jasani, B., Udpikar, A., and Manmatha, R., —DocFormer: End-to-End Transformer for Document Understanding,| Proceedings of ICCV, pp. 993–1002, 2021.
25. Kim, G., Hong, T., Yim, M., et al., —OCR-Free Document Understanding Transformer,| Proceedings of ECCV, pp. 498–517, 2022.
26. Herzig, J., Nowak, P. K., Müller, T., Piccinno, F., and Eisenschlos, J. M., —TaPas: Weakly Supervised Table Parsing via Pre-training,| Proceedings of ACL, pp. 4320–4333, 2020.
27. Smock, B., Pesala, R., and Abraham, R., —PubTables-1M: Towards Comprehensive Table Extraction from Unstructured Documents,| Proceedings of CVPR, pp. 4634–4642, 2022.
28. Liu, F., Eisenschlos, J. M., Piccinno, F., et al., —DePlot: One-shot Visual Language Reasoning by Plot-to-Table Translation,| Findings of ACL, pp. 13–22, 2023.
29. Masry, A., Long, D. X., Tan, J. Q., Joty, S., and Hoque, E., —ChartQA: A Benchmark for Question Answering about Charts with Visual and Logical Reasoning,| Findings of ACL, pp. 2263–2279, 2022.
30. Radford, A., Kim, J. W., Hallacy, C., et al., —Learning Transferable Visual Models From Natural Language Supervision,| Proceedings of ICML, vol. 139, pp. 8748–8763, 2021.