Context-Aware Multimodal RAG Frameworks for Regional Real Estate Markets Using Dialect-Specific LLMs

Rohit Pise
Department of Computer Applications (MCA)Shri Vaishnav Vidyapeeth Vishwavidyalaya, Indore, India

***Abstract*—Regional real estate markets are characterized by fragmented listing sources, heterogeneous documentation (floor plans, brochures, registry excerpts, photographs), and buyer-seller communication that occurs predominantly in local dialects rather than standard national languages. Conventional retrieval-augmented generation (RAG) pipelines, built around monolingual text retrieval and generic embedding models, degrade sharply under these conditions because they cannot align code-mixed or dialectal queries with formally written property metadata, nor can they reason jointly over text and non-text artifacts such as site images, floor plans, and geospatial layers. This paper proposes Context-Aware Multimodal RAG (CAM-RAG), a framework that combines (i) a multimodal retrieval layer that jointly indexes textual listings, structured metadata, images, and geospatial context; (ii) a dialect-adaptation layer that normalizes regional and code-mixed queries prior to retrieval while preserving semantic intent; and (iii) a context-aware fusion and generation module that grounds large language model (LLM) outputs in retrieved evidence to reduce hallucination in price, location, and legal-status claims. We motivate the design against prior work in retrieval-augmented generation, multimodal retrieval, and low-resource/dialectal language modelling, and we present a concrete system architecture, a staged fine-tuning methodology for dialect adaptation, and an evaluation protocol comprising retrieval-quality, generation-faithfulness, and dialect-robustness metrics. We further outline a data governance and bias-auditing procedure appropriate for a regulated domain such as property transactions. The contribution of this paper is architectural and methodological: it defines a reproducible framework and evaluation protocol intended to be instantiated and empirically validated on region-specific real estate corpora in follow-up work, rather than reporting closed empirical results.**

***Index Terms*—Retrieval-Augmented Generation, Multimodal Retrieval, Dialect-Specific Language Models, Real Estate Information Systems, Context-Aware AI, Low-Resource NLP, PropTech**

I. INTRODUCTION

The digitization of real estate markets has produced large volumes of heterogeneous data: structured listing databases, unstructured brochure text, floor-plan images, geotagged photographs, and buyer queries submitted through chat and voice interfaces. In many regional markets, particularly across linguistically diverse countries such as India, buyer and agent communication is conducted in local dialects, transliterated scripts, or code-mixed language (for example, Hindi-English or Marathi-English mixing), while the underlying listing data is authored in formal, standardized language. This mismatch between the register of user queries and the register of indexed content is a primary source of retrieval failure in deployed property-search assistants.

Retrieval-augmented generation has emerged as the dominant architecture for grounding large language model outputs in verifiable, up-to-date external evidence, mitigating the hallucination risk that arises when LLMs rely solely on parametric memory. However, most production RAG systems assume a single modality (text) and a single, relatively standardized language register. Real estate discovery, by contrast, is inherently multimodal: a prospective buyer's intent is only fully specified by combining textual constraints (budget, locality, legal status), visual constraints (layout, natural light, finishing quality visible in photographs or floor plans), and geospatial constraints (proximity to transit, schools, or flood-prone zones).

This paper proposes Context-Aware Multimodal RAG (CAM-RAG), a framework designed specifically for regional real estate platforms. The framework's central contribution is a dialect-adaptation layer that sits upstream of retrieval and re-expresses dialectal or code-mixed queries into a normalized intermediate representation without discarding locally significant terms (for example, colloquial property-type or locality names that have no standard-language equivalent). This normalized representation drives a multimodal retriever that jointly scores textual and visual evidence, and the retrieved evidence is passed to a generation module that is explicitly constrained to cite retrieved fields, reducing the risk of fabricated price or legal-status statements in a domain where such errors carry financial and legal consequences.

1.1 Motivation

The motivation for this work is threefold. First, existing property-search assistants built on generic LLM APIs frequently misinterpret dialectal queries, either failing to retrieve relevant listings or silently substituting a semantically adjacent but incorrect locality. Second, textual RAG alone cannot answer queries that are fundamentally visual or spatial in nature ("a flat that gets afternoon sun" or "walking distance to the metro"), which are common in real estate discourse. Third, the regulatory and financial stakes of property transactions make ungrounded generation particularly costly, motivating an architecture in which every generated claim is traceable to a retrieved source document.

1.2 Research Questions

- RQ1: How can a retrieval pipeline be adapted to jointly index and rank textual, visual, and geospatial real estate evidence under a single relevance model?
- RQ2: What adaptation strategy allows an LLM-based query encoder to robustly interpret dialectal and code-mixed real estate queries without a full re-training of the base language model?
- RQ3: How can generation be constrained so that price, locality, and legal-status claims are verifiably grounded in retrieved evidence, and how should faithfulness be measured in this domain?

1.3 Contributions

- A modular system architecture (CAM-RAG) integrating multimodal retrieval, dialect adaptation, and citation-constrained generation for regional real estate applications.
- A staged methodology for dialect adaptation that combines lightweight instruction-tuning with a retrieval-side normalization dictionary, avoiding full model retraining.
- An evaluation protocol with concrete metrics for retrieval quality, dialect robustness, and generation faithfulness, designed for reproducibility on region-specific corpora.
- A discussion of governance, fairness, and regulatory considerations specific to deploying generative AI in property transactions.

II. LITERATURE REVIEW

2.1 Retrieval-Augmented Generation

Retrieval-augmented generation was introduced as a way to combine a parametric language model with a non-parametric memory accessed through dense retrieval, allowing the model to condition generation on documents retrieved at inference time rather than relying solely on knowledge encoded during training (Lewis et al., 2020). Subsequent work extended this idea with in-context retrieval augmentation applied iteratively during generation (Guu et al., 2020; Izacard & Grave, 2021), and with retrieval strategies that interleave generation and retrieval steps to handle multi-hop reasoning (Trivedi et al., 2023). These systems are predominantly evaluated on open-domain question answering over English text corpora and do not address multimodal evidence or dialectal query variation.

2.2 Multimodal Retrieval and Generation

Multimodal retrieval systems such as CLIP established joint text-image embedding spaces that enable cross-modal similarity search (Radford et al., 2021), and later work extended retrieval augmentation to multimodal settings, retrieving images or image-text pairs to condition downstream generation (Chen et al., 2022; Yasunaga et al., 2023). These approaches demonstrate that jointly indexing heterogeneous modalities improves grounding for tasks where a single modality is insufficient to resolve a query, which is directly applicable to property search but has not, to our knowledge, been instantiated for structured real estate metadata combined with floor plans and geospatial layers.

2.3 Dialect-Specific and Low-Resource Language Modelling

Work on low-resource and dialectal NLP has shown that standard multilingual language models trained predominantly on high-resource, standardized text underperform on dialectal and code-mixed input, and that targeted adaptation—through continued pretraining, instruction-tuning on dialectal data, or script-normalization pipelines—recovers much of the performance gap without requiring a model trained from scratch on the dialect (Aji et al., 2022; Khanuja et al., 2020, on GLUECoS for code-mixed Hindi-English). Parameter-efficient fine-tuning methods such as LoRA (Hu et al., 2022) make such adaptation tractable for teams without large-scale compute, which is a practical constraint for regional PropTech platforms.

2.4 AI Applications in Real Estate (PropTech)

Applied work on AI in real estate has largely focused on price prediction using structured tabular features and, more recently, on computer-vision-based property valuation from images (Poursaeed et al., 2018; Bency et al., 2017). Conversational and generative interfaces for property search remain relatively underexplored in the academic literature compared to industry deployments, and existing academic treatments rarely address language variation or dialectal query handling, despite this being a first-order usability concern in linguistically diverse markets.

2.5 Research Gap

The reviewed literature establishes three separate threads—text RAG, multimodal retrieval, and dialect adaptation—but no existing framework, to our knowledge, unifies them for the specific constraints of regional real estate search: heterogeneous structured and unstructured evidence, dialectal and code-mixed queries, and a domain where ungrounded generation carries outsized financial risk. CAM-RAG is proposed to close this gap by treating dialect normalization as a first-class stage of the retrieval pipeline rather than a downstream generation concern, and by constraining generation to be traceable to retrieved multimodal evidence.

Table 1 situates CAM-RAG relative to the lines of prior work reviewed above, highlighting that no existing system combines full multimodal coverage, dedicated dialect adaptation, and citation-constrained generation within a single pipeline.

Table 1. Comparative positioning of CAM-RAG against related lines of work.

System / Line of Work	Modality Coverage	Dialect / Code-Mix Handling	Grounded Generation	Real Estate Domain Fit
Classic RAG (Lewis et al., 2020)	Text only	None	Partial	Low
Multimodal retrieval (CLIP-based)	Text + Image	None	Not addressed	Medium
Code-mixed NLP benchmarks (GLUECoS)	Text only	Strong (evaluation only)	N/A	Low
Vision-based property valuation	Image + tabular	None	N/A	Medium
CAM-RAG (proposed)	Text + Image + Geospatial + Structured	Dedicated adaptation layer	Citation-constrained	High (purpose-built)

III. PROBLEM STATEMENT

Let a user query q be expressed in a regional dialect or code-mixed form, and let the target corpus D consist of heterogeneous documents d = (t, v, g, m), where t is free text, v is one or more images (photographs or floor plans), g is geospatial metadata, and m is structured metadata (price, area, legal status, amenities). A conventional text-only RAG pipeline retrieves documents by scoring a dense embedding of q against embeddings of t alone, which fails in two respects: (a) the embedding of a dialectal q is poorly aligned with the embedding of standard-register t, producing low recall on relevant listings; and (b) query intent that is only resolvable through v or g (visual or spatial constraints) is invisible to a text-only retriever. The problem this paper addresses is the design of a retrieval and generation architecture that resolves both failure modes jointly, while keeping generated answers traceable to specific fields of specific retrieved documents.

IV. PROPOSED FRAMEWORK: CAM-RAG

CAM-RAG is organized into four cooperating modules: a Dialect Adaptation Layer, a Multimodal Indexing and Retrieval Layer, a Context Fusion module, and a Citation-Constrained Generation module. Figure 1 (described narratively below, as this document does not embed a rendered diagram) depicts the pipeline as a left-to-right flow from raw user query to grounded response.

4.1 System Overview

A raw query enters the Dialect Adaptation Layer, which produces (i) a normalized-language rewrite of the query for embedding purposes and (ii) a preserved set of locality-specific or colloquial terms that are retained verbatim as retrieval filters, since these terms often correspond precisely to named localities or vernacular property-type labels that a normalized rewrite would otherwise erase. Both outputs are passed to the Multimodal Indexing and Retrieval Layer, which independently scores candidate documents along text, image, and geospatial channels and combines these scores through a learned weighting function. The top-k fused candidates are passed to the Context Fusion module, which assembles a compact, field-labelled evidence package (price, locality, area, legal status, amenities, thumbnail description) per candidate. Finally, the Citation-Constrained Generation module produces a natural-language response in which every factual assertion is required to map to a specific field of a specific evidence package, enabling a post-hoc verifier to check claim-to-source alignment.

4.2 Dialect Adaptation Layer

This layer combines two complementary mechanisms. The first is a lightweight, parameter-efficient instruction-tuned adapter (following the LoRA paradigm of Hu et al., 2022) applied to a base multilingual encoder, trained on a small parallel corpus of dialectal and code-mixed real estate queries paired with standardized rewrites. The second is a maintained locality and colloquialism dictionary that maps vernacular terms to canonical locality identifiers, which is used to inject retrieval filters even when the adapter's rewrite is imperfect. This dual mechanism is designed so that dialect coverage can be extended incrementally by updating the dictionary, without requiring adapter retraining for every new term.

4.3 Multimodal Indexing and Retrieval Layer

Textual fields are embedded with a sentence-level dense encoder; property images and floor plans are embedded with a vision-language encoder in a shared or bridged embedding space (in the tradition of CLIP-style contrastive pretraining); geospatial metadata is represented as a feature vector capturing proximity to points of interest and is scored through a separate lightweight ranker. A late-fusion re-ranker combines the three channel scores using weights that can be tuned per query type (for example, up-weighting the geospatial channel when the normalized query contains proximity language).

4.4 Context Fusion Module

Rather than concatenating raw retrieved text into the generation prompt, the Context Fusion module extracts a fixed schema of fields from each candidate document and serializes them into a structured, labelled block. This design choice is intended to make the citation-constraint step in the generation module tractable: because every fact available to the generator is already field-labelled, a generated sentence referencing "price" can be checked against the literal price field of the cited document.

4.5 Citation-Constrained Generation Module

The generation module is prompted to produce a response together with inline field-level citations, and a lightweight post-generation verifier cross-checks each numeric or categorical claim against the source field it cites, flagging unsupported claims for either regeneration or suppression. This mirrors the general strategy of constrained, verifiable generation used in high-stakes RAG applications, applied here specifically to price, area, locality, and legal-status fields, which are the categories of error most likely to cause financial or legal harm to a user.

V. METHODOLOGY

5.1 Corpus Design

A region-specific evaluation corpus should combine (i) a structured listing dataset scraped or licensed from regional property portals, including price, area, amenities, and legal-status fields; (ii) an associated image set of property photographs and, where available, floor plans; (iii) a geospatial layer of points of interest; and (iv) a query set collected from real user interactions or elicited through structured user studies, deliberately sampling dialectal and code-mixed phrasing alongside standard-register queries to allow paired comparison.

5.2 Adapter Training Protocol

The dialect-adaptation adapter is proposed to be trained in three stages: (1) continued masked-language-model pretraining on unlabeled dialectal real estate text to adapt subword statistics to the target register; (2) parallel-corpus instruction-tuning mapping dialectal queries to normalized rewrites; and (3) retrieval-in-the-loop fine-tuning, in which the adapter is further updated using retrieval outcomes as a weak supervision signal, penalizing rewrites that lead to low-relevance retrieval results.

5.3 Evaluation Metrics

The evaluation protocol is organized around three axes:

- Retrieval quality: Recall@k and nDCG@k computed separately for standard-register and dialectal query subsets, allowing direct measurement of the dialect performance gap before and after adaptation.
- Dialect robustness: a Dialect Robustness Ratio, defined as the retrieval metric on dialectal queries divided by the same metric on paired standard-register queries, with a ratio approaching 1.0 indicating successful adaptation.
- Generation faithfulness: a claim-level precision metric computed by the post-generation verifier, reporting the proportion of numeric or categorical claims in a generated response that are supported by an exact field match in a cited source document.
- User-perceived relevance: a human evaluation protocol using Likert-scale relevance and trust ratings collected from regional users across both dialectal and standard-register query sessions.

This protocol is proposed rather than executed within the scope of this paper; Section 7 discusses anticipated outcome patterns with reference to comparable results reported in the adjacent literature.

Table 2 summarizes the proposed evaluation metrics, their definitions, and the target behavior each metric is designed to detect.

Table 2. Proposed evaluation metrics for CAM-RAG.

Metric	Definition	Target Behavior
Recall@k	Proportion of relevant listings present in top-k retrieved results	Parity between dialectal and standard-register query subsets
nDCG@k	Rank-sensitive relevance score over top-k results	Higher-ranked results more relevant, independent of query register
Dialect Robustness Ratio	Dialectal metric divided by paired standard-register metric	Approach 1.0 after adaptation
Claim-level Precision	Share of generated numeric/categorical claims matching a cited field exactly	High precision; unsupported claims suppressed, not fabricated
User Trust Rating	Likert-scale post-interaction survey score	Comparable trust across dialect groups

VI. ANTICIPATED OUTCOMES AND DISCUSSION

Because this paper presents a framework and evaluation protocol rather than a completed empirical study, outcomes are discussed here as anticipated patterns grounded in comparable published results, to be confirmed or revised through the corpus-specific evaluation described in Section 5. Prior work on dialect and code-mixed adaptation reports that targeted, parameter-efficient adaptation recovers a substantial share of the performance gap relative to standard-register input, without requiring full model retraining (Hu et al., 2022; Khanuja et al., 2020); we expect a similar pattern to hold for the Dialect Robustness Ratio defined in Section 5.3, with the ratio improving materially after the staged adapter training but not necessarily reaching full parity, particularly for queries containing rare or highly local colloquialisms not represented in the adaptation corpus.

For the multimodal retrieval channel, we anticipate the largest relative improvement on queries that contain visual or spatial constraints not resolvable from text alone, consistent with findings in the broader multimodal retrieval literature that cross-modal grounding disproportionately benefits queries whose intent spans modalities (Chen et al., 2022). For claim-level generation precision, the citation-constrained design is expected to trade a small amount of response fluency and completeness for a substantial reduction in unsupported numeric claims, mirroring the general precision-recall trade-off observed when constrained decoding or verification steps are added to RAG pipelines.

A remaining open question is how the framework performs when the adaptation corpus itself contains regional bias—for example, over-representing certain localities or property categories—since this could propagate into systematically unequal retrieval quality across neighborhoods. We treat this as a governance concern addressed in Section 7 rather than a purely technical one.

VII. GOVERNANCE, FAIRNESS, AND LIMITATIONS

7.1 Data Governance

Property data frequently includes personally identifiable information about sellers and occupants, and in many jurisdictions is subject to data-protection and consumer-protection regulation. A deployment of CAM-RAG should include an access-control layer restricting which fields are surfaced to which user roles, an audit log of retrieved evidence per generated response to support dispute resolution, and a data-retention policy aligned with local regulatory requirements.

7.2 Fairness Across Localities and Dialect Groups

Because the adaptation corpus is necessarily finite, there is a risk that users speaking less-represented dialects, or searching in less-documented localities, receive systematically lower retrieval quality. We recommend that the Dialect Robustness Ratio (Section 5.3) be monitored per dialect subgroup rather than only in aggregate, so that disparities are visible rather than averaged away.

7.3 Limitations

- This paper does not report closed empirical results; the evaluation protocol in Section 5 has not yet been executed on a live corpus.
- The framework assumes availability of a parallel dialectal-to-standard query corpus for adapter training, which may require dedicated data collection in markets where no such resource currently exists.
- Citation-constrained generation reduces but does not eliminate hallucination risk; the post-generation verifier is only as reliable as the field-matching logic underlying it, and edge cases (approximate or range-valued fields) require additional handling not detailed here.
- Geospatial scoring assumes reasonably complete points-of-interest data, which may be sparse in some regional markets.

VIII. CONCLUSION AND FUTURE WORK

This paper proposed CAM-RAG, a context-aware, multimodal retrieval-augmented generation framework designed for regional real estate markets characterized by dialectal query variation and heterogeneous, multimodal listing evidence. The framework's central design contribution is treating dialect adaptation as a first-class retrieval-side concern rather than leaving it to the generation model, combined with a citation-constrained generation stage intended to keep price, locality, and legal-status claims traceable to retrieved evidence. We presented a concrete architecture, a staged adapter-training methodology, and an evaluation protocol spanning retrieval quality, dialect robustness, and generation faithfulness.

Future work should prioritize instantiating this framework on a real, region-specific corpus—for instance, a mid-sized Indian metropolitan market with documented dialectal query variation—to obtain the empirical results this paper deliberately withholds from claiming in advance. Further extensions include incorporating voice-based query input, which would introduce an additional automatic speech recognition error source layered on top of dialectal variation, and extending the geospatial channel to incorporate temporal signals such as under-construction status and expected possession timelines.

REFERENCES

[1] Aji A.F., Winata G.I., Koto F., Cahyawijaya S., Romadhony A., Mahendra R., Kurniawan K., Moeljadi D., Prasojo R.E., Baldwin T., Lau J.H. and Ruder S.; 2022. One country, 700+ languages: NLP challenges for underrepresented languages and dialects in Indonesia. Proc. 60th Annu. Meet. Assoc. Comput. Linguistics (Vol. 1: Long Papers). Association for Computational Linguistics, Dublin: 7226-7249.

[2] Bency A.J., Rallapalli S., Ganti R.K., Srivatsa M. and Manjunath B.S.; 2017. Beyond spatial auto-regressive models: Predicting housing prices with satellite imagery. IEEE Winter Conf. Appl. Comput. Vision (WACV). IEEE, Santa Rosa, CA: 320-329.

[3] Chen W., Hu H., Chen X., Verga P. and Cohen W.W.; 2022. MuRAG: Multimodal retrieval-augmented generator for open question answering over images and text. Proc. 2022 Conf. Empirical Methods in Natural Language Processing. Association for Computational Linguistics, Abu Dhabi: 5558-5570.

[4] Guu K., Lee K., Tung Z., Pasupat P. and Chang M.W.; 2020. REALM: Retrieval-augmented language model pre-training. Proc. 37th Int. Conf. Machine Learning. PMLR: 3929-3938.

[5] Hu E.J., Shen Y., Wallis P., Allen-Zhu Z., Li Y., Wang S., Wang L. and Chen W.; 2022. LoRA: Low-rank adaptation of large language models. International Conference on Learning Representations (ICLR).

[6] Izacard G. and Grave E.; 2021. Leveraging passage retrieval with generative models for open domain question answering. Proc. 16th Conf. Eur. Chapter Assoc. Comput. Linguistics: Main Volume. Association for Computational Linguistics: 874-880.

[7] Khanuja S., Dandapat S., Srinivasan A., Sitaram S. and Choudhury M.; 2020. GLUECoS: An evaluation benchmark for code-switched NLP. Proc. 58th Annu. Meet. Assoc. Comput. Linguistics. Association for Computational Linguistics, online: 3575-3585.

[8] Lewis P., Perez E., Piktus A., Petroni F., Karpukhin V., Goyal N., Kuttler H., Lewis M., Yih W., Rocktaschel T., Riedel S. and Kiela D.; 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. Adv. Neural Inf. Process. Syst., 33: 9459-9474.

[9] Poursaeed O., Matera T. and Belongie S.; 2018. Vision-based real estate price estimation. Mach. Vision Appl., 29(4): 667-676.

[10] Radford A., Kim J.W., Hallacy C., Ramesh A., Goh G., Agarwal S., Sastry G., Askell A., Mishkin P., Clark J., Krueger G. and Sutskever I.; 2021. Learning transferable visual models from natural language supervision. Proc. 38th Int. Conf. Machine Learning. PMLR, 139: 8748-8763.

[11] Trivedi H., Balasubramanian N., Khot T. and Sabharwal A.; 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. Proc. 61st Annu. Meet. Assoc. Comput. Linguistics (Vol. 1: Long Papers). Association for Computational Linguistics, Toronto: 10014-10037.

[12] Yasunaga M., Aghajanyan A., Shi W., James R., Leskovec J., Liang P., Lewis M., Zettlemoyer L. and Yih W.; 2023. Retrieval-augmented multimodal language modeling. Proc. 40th Int. Conf. Machine Learning. PMLR: 39755-39769.