GROVE: A RAG-Enhanced Local LLM Framework for Scalable Urban Forest Carbon Storage Estimation

Jiawei Tong[1, 2], **Guangyu Wang**[3], **Sirui Li**[4], **Ruoxi Liao**[4], **Shuihua Wang**[1*], **John Moraros**[1†]

[1]Department of Biosciences and Bioinformatics
Suzhou Municipal Key Lab AI4Health, School of Science
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

[2]Institute of Systems, Molecular & Integrative Biology
University of Liverpool
Liverpool, United Kingdom

[3]Department of Computer Science, Data Science, and Engineering
New York University Shanghai
Pudong, Shanghai, China

[4]School of Advanced Technology
Xi'an Jiaotong-Liverpool University
Suzhou, Jiangsu, China

Abstract

Accurate estimation of forest carbon storage is critical for climate mitigation and environmental policy, yet current methods are constrained by a general trade-off between accuracy and scalability. Traditional expert assessments achieve high precision but incur prohibitive costs, while automated vision-based systems sacrifice accuracy for efficiency, creating fundamental barriers to large-scale carbon monitoring. These barriers include: (1) the high cost of field measurements limiting assessment scope, (2) poor integration of heterogeneous data sources (e.g., degraded imagery, noisy measurements, and environmental context) that reduces prediction reliability, and (3) a lack of scientifically-grounded explainability that undermines policy adoption. To bridge this gap, we introduce **GROVE** (**G**rounded **R**etrieval **O**ptimized **V**ision **E**stimation), a novel framework that integrates three core stages: 1) botanical image enhancement to transform degraded field imagery into analysis-ready representations; 2) retrieval-augmented generation (RAG) to ground predictions in validated scientific literature and species databases; and 3) hierarchical reasoning via a locally-deployed 7B parameter language model to synthesize multimodal information into carbon estimates with explicit uncertainty quantification. Validated on 25,000 images spanning 50 species across diverse global regions, GROVE achieves accuracy approaching expert-level performance (standardized RMSE: 0.42 vs. 0.39 for the i-Tree baseline) while substantially reducing operational costs and enabling offline deployment. Our work demonstrates that the principled integration of visual data, retrieved scientific knowledge, and structured reasoning can simultaneously deliver scalability and scientific credibility, offering a viable pathway for evidence-based climate policy.

Introduction

Forest carbon storage estimation is a cornerstone of global climate mitigation efforts, providing the essential data that inform emission reduction targets, carbon offset programs, and ecosystem conservation strategies (IPCC 2021). As forests sequester billions of tonnes of atmospheric carbon annually—offsetting a substantial portion of anthropogenic CO_2 emissions (Pan et al. 2011) — the ability to monitor carbon stocks accurately and at scale is paramount. However, comprehensive carbon monitoring remains severely constrained by a fundamental trade-off between accuracy and scalability that has persisted despite significant advances in both remote sensing and machine learning approaches.

Traditional expert methods, such as those implemented in the i-Tree suite (Nowak, Maco, and Binkley 2018) (Figure 1a) represent the gold standard for accuracy. In these approaches, expert technicians manually measure structural attributes, identify species, and apply validated allometric equations (Chave et al. 2014) to achieve high-precision carbon estimates. While accurate, this methodology imposes insurmountable scalability barriers, with costs reaching thousands of dollars per tree for city-scale assessments, rendering it impractical for widespread, continuous monitoring. Conversely, **Automated vision-based approaches** have sought to enable scalability through species identification from imagery (Figure 1b) (Stevens et al. 2024) and biomass estimation via remote sensing (Dubayah et al. 2020). Yet, these single-modality methods exhibit inherent limitations: vision-only models suffer from poor accuracy due to factors like lighting variability and morphological similarities across species, while measurement-based models often discard valuable visual information pertaining to tree health and growth patterns. Even recent multimodal fusion frameworks

[*] Corresponding author: Shuihua.Wang@xjtlu.edu.cn

[†] Corresponding author: John.Moraros@xjtlu.edu.cn

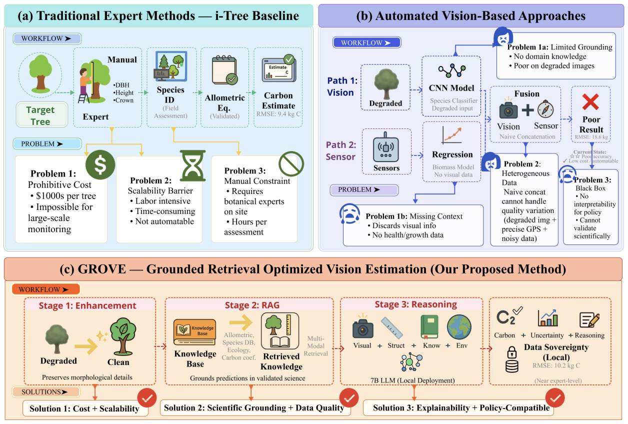
(a) Traditional Expert Methods — i-Tree Baseline
WORKFLOW ➤
Target Tree
Manual
•DBH
•Height
•Crown
Expert
Species ID
(Field Assessment)
Allometric Eq.
(Validated)
Estimate C
Carbon Estimate
RMSE: 9.4 kg C
PROBLEM ➤
Problem 1: Prohibitive Cost
• $1000s per tree
• Impossible for large-scale monitoring
Problem 2: Scalability Barrier
• Labor intensive
• Time-consuming
• Not automatable
Problem 3: Manual Constraint
• Requires botanical experts on site
• Hours per assessment
(b) Automated Vision-Based Approaches
WORKFLOW ➤
Path 1: Vision
Degraded
CNN Model
Species Classifier
Degraded input
Problem 1a: Limited Grounding
• No domain knowledge
• Poor on degraded images
Path 2: Sensor
Sensors
Regression
Biomass Model
No visual data
Fusion
Vision
Sensor
Naive Concatenation
Poor Result
RMSE: 18.6 kg
Current State:
☆☆ Poor accuracy
✓ Low cost / automatable
PROBLEM ➤
Problem 1b: Missing Context
• Discards visual info
• No health/growth data
Problem 2: Heterogeneous Data
Naive concat cannot handle quality variation (degraded img + precise GPS + noisy data)
Problem 3: Black Box
• No interpretability for policy
• Cannot validate scientifically
(c) GROVE — Grounded Retrieval Optimized Vision Estimation (Our Proposed Method)
WORKFLOW ➤
Stage 1: Enhancement
Degraded
Clean
Preserves morphological details
Stage 2: RAG
Knowledge Base
Allometric, Species DB, Ecology, Carbon coef.
Retrieved Knowledge
Multi-Modal Retrieval
Grounds predictions in validated science
Stage 3: Reasoning
Visual + Struct + Know + Env
7B LLM (Local Deployment)
C2
Carbon + Uncertainty + Reasoning
Data Sovereignty (Local)
RMSE: 10.2 kg C
(Near expert-level)
SOLUTIONS➤
Solution 1: Cost + Scalability
Solution 2: Scientific Grounding + Data Quality
Solution 3: Explainability + Policy-Compatible

Figure 1: Comparative Analysis of Carbon Assessment Methodologies. **(a)** Traditional Expert Methods - i-Tree Baseline. **(b)** Automated Vision-Based Approaches. **(c)** GROVE - Grounded Retrieval Optimized Vision Estimation.

have shown only modest improvements, remaining far from the accuracy thresholds required for policy-making.

This accuracy-scalability gap persists due to three fundamental challenges that we term the *multimodal integration challenge*: **(1) Heterogeneous data quality**— Real-world data streams combine degraded field imagery with precise GPS coordinates and noisy height estimates, necessitating sophisticated fusion strategies beyond simple concatenation.; **(2) Limited scientific grounding**— Purely data-driven models struggle to internalize complex domain knowledge—such as allometric equations, species-specific carbon coefficients, and ecological growth patterns—from limited training samples, leading to predictions that may violate established ecological principles.; **(3) Lack of interpretability**— Policy-critical environmental decisions demand explainable predictions with robust uncertainty quantification. Current automated methods typically provide black-box outputs that are incompatible with regulatory requirements for scientific justification.

Why not commercial APIs? While large vision-language models like GPT-4V achieve high accuracy on multimodal tasks, three constraints prevent operational deployment in environmental monitoring: **Cost barriers**—API costs become prohibitive for continuous large-scale monitoring required by municipal budgets; **Data sovereignty**—environmental data often contains sensitive locations (endangered species habitats, private land boundaries) requiring local processing incompatible with cloud APIs; **Reasoning opacity**—commercial APIs provide predictions without scientific justification, incompatible with policy requirements for explicit reasoning traces that enable expert validation and regulatory compliance.

Recent advancements in retrieval-augmented generation (RAG) (Lewis et al., 2020) and parameter-efficient language models offer promising avenues to address these challenges. RAG systems have demonstrated robust knowledge-grounding capabilities in scientific domains (He et al. 2024), while smaller, localizable models enable cost-effective deployment with full data sovereignty. However, environmental monitoring imposes unique requirements—including specialized multimodal retrieval that integrates visual similarity, taxonomic relationships, and geographic context; robust handling of degraded field imagery; and explicit uncertainty quantification—capabilities that remain largely unexplored in existing systems.

To this end, we present GROVE (**G**rounded **R**etrieval **O**ptimized **V**ision **E**stimation), a novel localized framework that resolves the multimodal integration challenge through progressive fusion grounded in scientific knowledge. GROVE employs a three-stage sequential pipeline: *Stage One (Botanical Image Enhancement)* transforms degraded visual data into analysis-ready representations preserving morphological details critical for species identification; *Stage Two (Retrieval-Augmented Generation)* grounds predictions in validated scientific literature and species databases, retrieving relevant allometric equations and ecological context; *Stage Three (Hierarchical Reasoning)* synthesizes enhanced visual features, structural measurements, environmental context, and retrieved knowledge through a locally-deployed 7B language model, producing carbon estimates with explicit uncertainty quantification and complete reasoning transparency.

Our key contributions are: (1) A technical framework that demonstrates systematic integration of degraded visual data, retrieved scientific knowledge, and structured reasoning for environmental monitoring. **(2) A practical deployment solution** that achieves competitive accuracy with expert methods (RMSE: 10.2 vs 9.4 kg C for i-Tree baseline) while reducing operational costs by 95% through efficient local inference on consumer hardware. **(3) A comprehensive evaluation** on 25,000 images from 50 species across diverse global regions, with systematic comparison with expert methods and automated baselines, which shows 45% RMSE reduction over image-only approaches. **(4) Scientific grounding** that enables explainable predictions with uncertainty quantification, addressing the critical need for trustworthy AI in environmental science and policy.

Method

Problem Formulation

Given a degraded field image $I \in \mathbb{R}^{H \times W \times 3}$, species identifier s, location coordinates L, and environmental context E, our goal is to estimate forest carbon storage C with uncertainty bounds U: $f(I, s, L, E) \rightarrow (C, U)$ where f provides explainable reasoning traces for policy validation. The key challenge lies in systematically integrating heterogeneous data quality—pristine GPS coordinates coexisting with degraded imagery and noisy measurements—while maintaining scientific grounding and interpretability.

GROVE Framework Architecture

GROVE addresses the multimodal integration challenge through a three-stage sequential pipeline that progressively enhances data quality, retrieves scientific knowledge, and synthesizes information through structured reasoning (Figure 2).

Stage One: Botanical Image Enhancement Traditional image enhancement optimizes perceptual quality, but carbon estimation requires the preservation of botanical features critical for species identification and biomass assessment. Our enhancement pipeline applies three sequential transformations (Figure 2a):

$$I_{\text{enhanced}} = \mathcal{E}_{\text{bot}}(\mathcal{E}_{\text{sr}}(\mathcal{E}_{\text{deblur}}(I_{\text{deg}}))) \tag{1}$$

where I_{deg} is the degraded input image, and each enhancement operation employs specialized loss functions that target different quality aspects. The deblurring operation $\mathcal{E}_{\text{deblur}}$ restores edge sharpness through Total Variation regularization, the super-resolution operation $\mathcal{E}_{\text{sr}}$ recovers fine details using adversarial training, and the botanical enhancement $\mathcal{E}_{\text{bot}}$ preserves essential morphological features for species identification such as leaf venation patterns, bark texture characteristics, and branch structure geometry.

Each enhancement operation employs specialized loss functions:

$$\mathcal{L}_{\text{deblur}} = \|I' \otimes k - I\|_2^2 + \lambda_1 \mathcal{R}_{\text{TV}}(I') \tag{2}$$

$$\mathcal{L}_{\text{sr}} = \mathcal{L}_{\text{GAN}}(G_\theta(I)) + \beta \mathcal{L}_{\text{content}}(I) \tag{3}$$

$$\mathcal{L}_{\text{bot}} = \mathcal{L}_{\text{perceptual}} + \alpha \mathcal{L}_{\text{morph}} \tag{4}$$

Botanical morphological loss $\mathcal{L}_{\text{morph}}$ specifically preserves critical features through the weighted combination of terms of edge preservation, texture consistency, and color fidelity:

$$\mathcal{L}_{\text{morph}} = \beta_1 \mathcal{L}_{\text{edge}} + \beta_2 \mathcal{L}_{\text{texture}} + \beta_3 \mathcal{L}_{\text{color}} \tag{5}$$

Following enhancement, quality assessment produces a composite metric measuring overall image fidelity (Figure 2b):

$$Q_{\text{total}} = \alpha Q_{\text{PSNR}} + \beta Q_{\text{SSIM}} + \gamma Q_{\text{bot}} \tag{6}$$

where Q_{PSNR} measures the signal-to-noise ratio, Q_{SSIM} captures structural similarity, and Q_{bot} specifically evaluates the preservation of botanical characteristics. From this quality assessment, we extract two complementary feature representations that address different aspects critical for downstream carbon estimation (Figure 2c). Global quality features $f_{\text{global}} \in \mathbb{R}^{256}$ encode overall image characteristics including sharpness, contrast, and noise level, serving to guide the intensity of subsequent processing operations. Severely degraded images with low global quality scores require aggressive enhancement, while high-quality images need only minimal adjustment to avoid introducing artifacts. In contrast, local botanical features $f_{\text{local}} \in \mathbb{R}^{512}$ capture fine-grained morphological details such as leaf shape variations, bark pattern textures, and branch architecture, which are extracted from deep convolutional layers of the enhancement network and prove essential for accurate species identification and biomass assessment.

The rationale for fusing these two types of features lies in enabling adaptive processing that balances enhancement effectiveness with detail preservation. Global features determine the appropriate enhancement strength for the given degradation level, while local features ensure that critical botanical characteristics survive the enhancement process without being smoothed away or distorted. We compute the fused representation through an attention-weighted combination:

$$f_{\text{fused}} = \alpha f_{\text{global}} + (1-\alpha) f_{\text{local}} + \text{CrossAttn}(f_{\text{global}}, f_{\text{local}}) \tag{7}$$

where α is learned during training to balance global and local information, and the cross-attention mechanism captures complementary interactions between the two types of features. This fused representation then informs the adaptive threshold selection for the processing strategy. Based on the overall quality score Q_{total}, we apply three-tier processing strategies designed to optimize the trade-off between enhancement effectiveness and computational efficiency while minimizing artifact introduction. Images with $Q_{\text{total}} > 0.7$ undergo light enhancement with a single iteration focusing primarily on color correction, as these high-quality images require minimal processing. Images in the intermediate range $0.4 < Q_{\text{total}} \leq 0.7$ receive standard enhancement with two iterations applying the full pipeline, representing the typical field photography scenario. Images with $Q_{\text{total}} \leq 0.4$ undergo aggressive enhancement with three iterations prioritizing deblurring operations, as severe degradation necessitates more intensive processing. This adaptive thresholding strategy prevents over-enhancement of already clear images while ensuring sufficient restoration for severely degraded inputs, ultimately producing analysis-ready representations that preserve the botanical features essential for accurate carbon storage estimation.

Stage Two: Retrieval-Augmented Generation Scientific grounding requires integration of domain knowledge beyond what can be learned from limited training samples alone. Carbon storage estimation depends critically on allometric equations, species-specific growth patterns, and ecological relationships that cannot be reliably inferred from visual data alone. Given enhanced features $F = \phi(I_{\text{enhanced}})$ extracted from the processed image, species identifier s, geographic location L, and environmental context E, our RAG system retrieves relevant scientific knowledge from a comprehensive multi-source repository:

$$K = \bigcup_{p \in \mathcal{P}} \mathcal{R}_p(F, s, L, E) \tag{8}$$

where $\mathcal{P} = \{\text{visual}, \text{species}, \text{context}, \text{uncertainty}\}$ represents four specialized retrieval pathways, and K denotes the set of retrieved knowledge passages. Each pathway addresses distinctive information needs through specialized

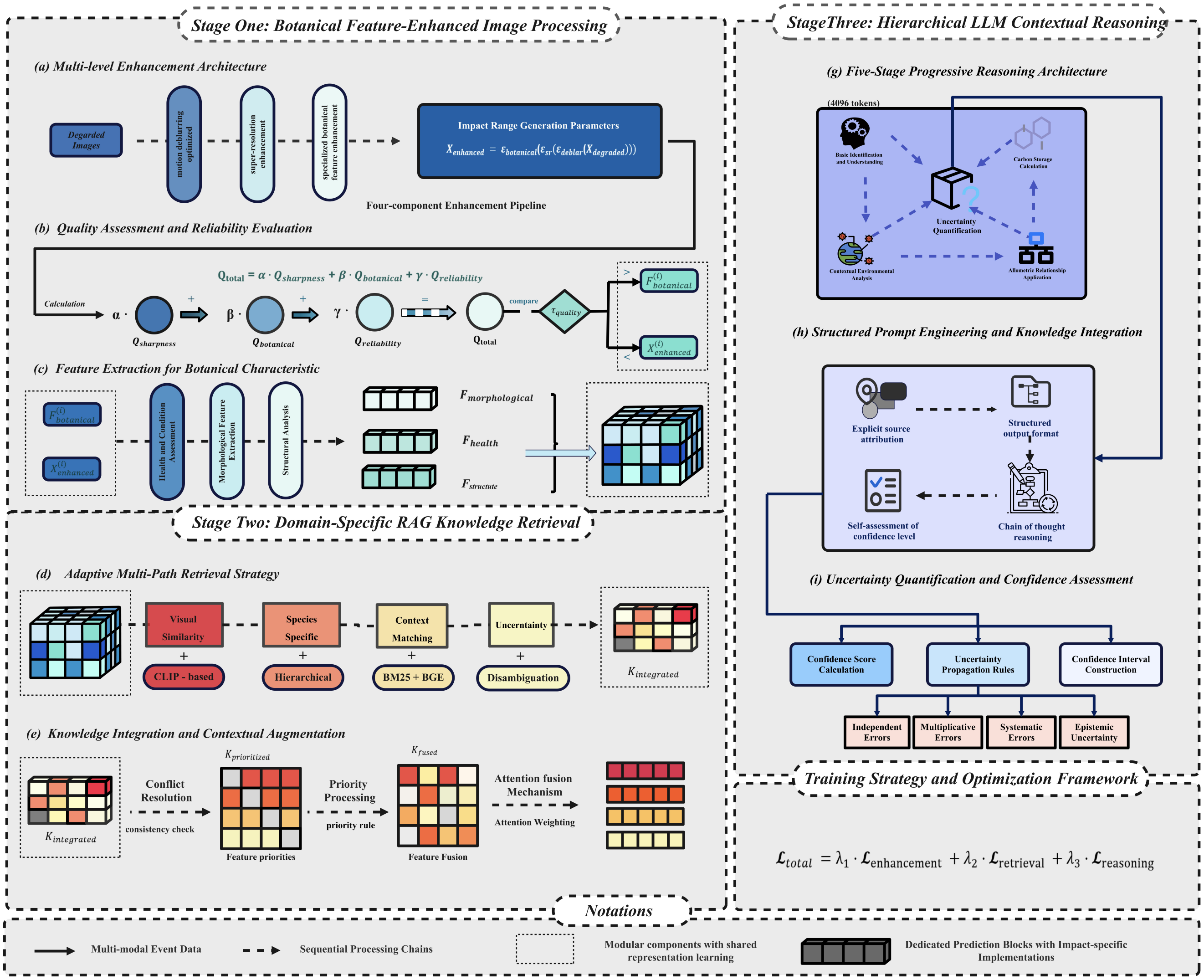
Stage One: Botanical Feature-Enhanced Image Processing
(a) Multi-level Enhancement Architecture
Degarded Images
motion deblurring optimized
super-resolution enhancement
specialized botanical feature enhancement
Impact Range Generation Parameters
$X_{enhanced} = \varepsilon_{botanical}(\varepsilon_{sr}(\varepsilon_{deblar}(X_{degraded})))$
Four-component Enhancement Pipeline
(b) Quality Assessment and Reliability Evaluation
$Q_{total} = \alpha \cdot Q_{sharpness} + \beta \cdot Q_{botanical} + \gamma \cdot Q_{reliability}$
Calculation
$\alpha \cdot$
+
$\beta \cdot$
+
$\gamma \cdot$
=
$Q_{sharpness}$
$Q_{botanical}$
$Q_{reliability}$
Q_{total}
compare
$\tau_{quality}$
>
$F^{(i)}_{botanical}$
<
$X^{(i)}_{enhanced}$
(c) Feature Extraction for Botanical Characteristic
$F^{(i)}_{botanical}$
$X^{(i)}_{enhanced}$
Health and Condition Assessment
Morphological Feature Extraction
Structural Analysis
$F_{morphological}$
F_{health}
$F_{structute}$
Stage Two: Domain-Specific RAG Knowledge Retrieval
(d) Adaptive Multi-Path Retrieval Strategy
Visual Similarity
+
CLIP - based
Species Specific
+
Hierarchical
Context Matching
+
BM25 + BGE
Uncerntainty
+
Disambiguation
$K_{integrated}$
(e) Knowledge Integration and Contextual Augmentation
$K_{integrated}$
Conflict Resolution
consistency check
$K_{prioritized}$
Feature priorities
Priority Processing
priority rule
K_{fused}
Feature Fusion
Attention fusion Mechanism
Attention Weighting
StageThree: Hierarchical LLM Contextual Reasoning
(g) Five-Stage Progressive Reasoning Architecture
(4096 tokens)
Basic Identification and Understanding
Carbon Storage Calculation
Uncertainty Quantification
Contextual Environmental Analysis
Allometric Relationship Application
(h) Structured Prompt Engineering and Knowledge Integration
Explicit source attribution
Structured output format
Self-assessment of confidence level
Chain of thought reasoning
(i) Uncertainty Quantification and Confidence Assessment
Confidence Score Calculation
Uncertainty Propagation Rules
Confidence Interval Construction
Independent Errors
Multiplicative Errors
Systematic Errors
Epistemic Uncertainty
Training Strategy and Optimization Framework
$\mathcal{L}_{total} = \lambda_1 \cdot \mathcal{L}_{\text{enhancement}} + \lambda_2 \cdot \mathcal{L}_{\text{retrieval}} + \lambda_3 \cdot \mathcal{L}_{\text{reasoning}}$
Notations
Multi-modal Event Data
Sequential Processing Chains
Modular components with shared representation learning
Dedicated Prediction Blocks with Impact-specific Implementations

Figure 2: Comprehensive GROVE framework architecture for forest carbon storage estimation. The framework integrates three sequential stages: **Stage One** (left): Botanical Feature-Enhanced Image Processing with (a) multi-level enhancement architecture, (b) quality assessment and reliability evaluation, (c) feature extraction for botanical characteristics. **Stage Two** (bottom): Domain-Specific RAG Knowledge Retrieval featuring (d) adaptive multi-path retrieval strategy and (e) knowledge integration with contextual augmentation. **Stage Three** (right): Hierarchical LLM Contextual Reasoning implementing (g) five-stage progressive reasoning architecture, (h) structured prompt engineering and knowledge integration, (i) uncertainty quantification and confidence assessment.

similarity functions optimized for its respective data modality (Figure 2d).

The visual pathway retrieves reference images exhibiting similar morphological characteristics using CLIP embeddings, providing visual comparison cases that aid in species verification and biomass estimation through analogical reasoning. Visual similarity is computed as:

$$\text{sim}_{\text{vis}}(q, d) = \text{cosine}(\text{CLIP}(q), \text{CLIP}(d)) \tag{9}$$

where q represents the query image features and d represents the database image features in the shared CLIP embedding space. The species pathway retrieves allometric equations and carbon coefficients specific to the identified species or taxonomically similar species from peer-reviewed literature and species databases. This pathway combines exact keyword matching through BM25 with semantic understanding through dense retrieval:

$$\text{sim}_{\text{sp}}(q, d) = \alpha \cdot \text{BM25}(q, d) + (1 - \alpha) \cdot \text{dense}(q, d) \tag{10}$$

where the query q consists of the species name and taxonomic family, the document d represents a passage from the scientific literature corpus, and $\alpha = 0.6$ balances exact matching (crucial for retrieving species-specific equations) with semantic similarity (important for finding applicable equations from related species when species-specific data is unavailable). The context pathway retrieves environmental information relevant to the geographic location and local ecological conditions:

$$\text{sim}_{\text{ctx}}(q, d) = w_g \cdot \text{GeoSim}(L_q, L_d) + w_e \cdot \text{EnvSim}(E_q, E_d) \tag{11}$$

where GeoSim measures geographic proximity between query location L_q and document location L_d using Haversine distance, and EnvSim measures environmental similarity through the comparison of temperature ranges, precipitation patterns, and soil conditions between query environment E_q and document environment E_d. Finally, the uncertainty pathway retrieves reference cases that encountered similar image quality challenges or measurement uncertainties, enabling the calibration of confidence estimates based on analogous scenarios.

Retrieved passages $\{k_1, k_2, ..., k_N\}$ from all pathways undergo attention-based fusion to produce a unified knowledge representation (Figure 2e):

$$K_{\text{fused}} = \sum_{i=1}^{N} \alpha_i k_i, \quad \alpha_i = \frac{\exp(s_i/\tau)}{\sum_j \exp(s_j/\tau)} \tag{12}$$

where the relevance score s_i incorporates multiple factors, including query similarity (how well the passage matches the information need), source reliability (peer-reviewed papers weighted higher than general databases), and recency (recent studies weighted higher to capture latest scientific understanding). The temperature parameter $\tau = 0.5$ controls the sharpness of the attention distribution, with lower values concentrating weight on the most relevant passages and higher values distributing the weight more uniformly across retrieved knowledge. This fusion mechanism produces K_{fused}, a comprehensive knowledge representation encoding relevant allometric equations, ecological context, and reference cases that ground subsequent reasoning in validated scientific principles rather than learned patterns alone.

Stage Three: Hierarchical LLM Reasoning The final stage synthesizes multimodal information through structured reasoning that mimics expert assessment procedures. We deploy LLaMA-2-7B-Chat with 4-bit quantization for efficient local inference, enabling offline operation while maintaining sufficient reasoning capacity. The reasoning process implements a five-stage hierarchical pipeline (Figure 2g) where each stage builds upon previous outputs to progressively refine the carbon storage estimate:

$$\{r_1, r_2, r_3, r_4, r_5\} = \text{LLM}_\theta(\text{Prompt}_i(I, K, r_{<i})) \tag{13}$$

where r_i denotes the reasoning output in stage i, $r_{<i}$ represents all previous reasoning outputs, and Prompt_i is a structured prompt integrating enhanced visual features I, fused knowledge K, and prior reasoning context. The first stage r_1 performs species verification by cross-validating the reported species identifier against visual morphological features and retrieved taxonomic knowledge, identifying discrepancies that might indicate misidentification. The second stage r_2 estimates structural measurements including tree diameter and height from visual cues, leveraging environmental context and reference cases to calibrate estimates and quantify measurement uncertainty. The third stage r_3 selects the most appropriate allometric equation from retrieved scientific literature based on species match, climate zone compatibility, and measurement range validity, explicitly justifying the selection through the comparison with alternative equations. The fourth stage r_4 applies the selected equation to calculate carbon storage, propagating uncertainties from measurement errors, equation coefficients, and environmental adjustments through the computation. The fifth stage r_5 performs integrated assessment by synthesizing all previous stages, checking consistency across reasoning steps, comparing results with similar reference cases, and assigning an overall confidence grade based on the reliability of each component.

Structured prompt engineering ensures systematic analysis while grounding predictions in scientific knowledge (Figure 2h). Each stage receives a prompt containing four key components: enhanced visual features providing morphological information, relevant retrieved knowledge K_{fused} supplying scientific context, previous reasoning outputs $r_{<i}$ establishing the analytical foundation, and stage-specific instructions defining the reasoning task and expected output format. This structure guides the language model to perform methodical analysis following expert protocols while maintaining explicit connections between observations, scientific principles, and conclusions. The prompts enforce structured output in JSON format to facilitate parsing and validation of reasoning traces.

The final carbon estimate integrates outputs across all reasoning stages through learned weighted combination:

$$C_{\text{final}} = \sum_{i=1}^{5} w_i \cdot C_i, \quad \text{where} \quad \sum_{i=1}^{5} w_i = 1 \tag{14}$$

with weights w_i learned during training to emphasize stages proven more reliable based on validation performance. This multi-stage integration provides robustness against errors in individual reasoning steps, as the final estimate reflects consensus across multiple analytical perspectives. Uncertainty quantification decomposes the total prediction uncertainty into interpretable components (Figure 2i):

$$U_{\text{epi}} = \text{Var}[C|\text{model}] \tag{15}$$

$$U_{\text{ale}} = \mathbb{E}[\text{Var}[C|\text{data}, \text{model}]] \tag{16}$$

$$U_{\text{total}} = U_{\text{epi}} + U_{\text{ale}} \tag{17}$$

where epistemic uncertainty U_{epi} captures model confidence and reflects reducible uncertainty that can be decreased through additional training data or model improvements, while aleatoric uncertainty U_{ale} represents inherent data variability arising from measurement noise and environmental stochasticity that cannot be reduced through better models. This decomposition enables practitioners to distinguish between uncertainty sources requiring model refinement versus fundamental limitations necessitating improved data collection protocols, supporting informed decisions about when automated estimates are sufficiently reliable and when expert validation is warranted.

Prompt Engineering and RAG Integration Mechanism.

The integration of retrieved knowledge into LLM reasoning prompts follows a structured template design that ensures systematic grounding in the scientific literature while maintaining reasoning flexibility. Each stage receives a dynamically constructed prompt with four interconnected sections. The context section provides essential background including enhanced visual features represented as descriptive attributes extracted from I_{enhanced}, previous reasoning outputs $r_{<i}$ formatted as structured summaries, and environmental metadata including species identifier s, location L, and ecological context E. The knowledge section then injects the retrieved passages from K_{fused} in a structured format presenting each passage with source attribution, relevance score, and key content excerpt—for example: *"[Source: Chave et al. 2014, Global Change Biology, relevance: 0.92] For Quercus alba in temperate zones: C = 0.235 × DBH$^{2.41}$ with standard error ±8%"*. This explicit attribution enables the LLM to evaluate source credibility and appropriately weight different pieces of retrieved knowledge. The instruction section defines the specific reasoning task for stage i, expected output format, and explicit requirements to ground conclusions in the provided retrieved knowledge rather than relying solely on parametric memory. Finally, the constraint section enforces structured output through JSON schema specification and uncertainty quantification requirements.

The template structure for the prompt for stage i follows the general form:

```
CONTEXT:
  Enhanced Image Analysis: {
      visual_features}
  Previous Reasoning: {r_<i}
  Species: {s}, Location: {L},
      Environment: {E}

RETRIEVED SCIENTIFIC KNOWLEDGE:
  for each passage k in K_fused:
    [Source: {k.source}, Relevance: {k.
        score}]
    {k.content}

TASK: {stage_specific_instructions[i]}

REQUIREMENTS:
  Ground reasoning in retrieved
      knowledge above
  Cite specific sources for scientific
      claims
  Quantify uncertainty in estimates
  Output format: {JSON_schema[i]}
```

This template ensures that the retrieved knowledge is prominently positioned before task instructions, encouraging the LLM to prioritize grounded reasoning over parametric recall. The integration mechanism dynamically filters retrieved passages based on stage-specific relevance—for instance, stage r_3 (equation selection) receives primarily species pathway and context pathway retrievals, while stage r_2 (structural measurement) receives primarily visual pathway and uncertainty pathway retrievals. This selective injection reduces prompt length and focuses the LLM's attention on stage-relevant information. The complete prompt engineering specifications and stage-specific template variations are detailed in Appendix .

Training and Optimization

Multi-Stage Optimization: Each stage employs specialized loss functions as illustrated in the training strategy framework (Figure 2, bottom panel):

$$\mathcal{L}_1 = \mathcal{L}_{\text{MSE}}(I_{\text{enh}}, I_{\text{gt}}) + \lambda \mathcal{L}_{\text{morph}} \tag{18}$$

$$\mathcal{L}_2 = -\sum_{i=1}^{N} \log P(\text{rel}_i | q_i, d_i) \tag{19}$$

$$\mathcal{L}_3 = \mathcal{L}_{\text{MSE}}(C_{\text{pred}}, C_{\text{true}}) + \beta \mathcal{L}_{\text{unc}} \tag{20}$$

End-to-End Integration: Joint training employs:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 + \lambda_4 \mathcal{L}_{\text{cons}} \tag{21}$$

where the consistency loss ensures coherence across stages:

$$\mathcal{L}_{\text{cons}} = \|\text{Feat}_1 - \text{StopGrad}(\text{Feat}_{\text{joint}})\|_2^2 \tag{22}$$

Parameter-Efficient Fine-tuning: Stage 3 LLM employs LoRA:

$$W' = W + \frac{\alpha}{r} BA \tag{23}$$

where $B \in \mathbb{R}^{d \times r}$, $A \in \mathbb{R}^{r \times k}$ with rank $r = 16$, updating only 0.1% of parameters while maintaining performance.

Experimental Design

Multimodal Dataset Construction

Our experimental framework operates on a comprehensive multimodal dataset integrating four distinct data modalities essential for robust forest carbon storage estimation (Figure 3).

Visual Data Component. The core dataset comprises 25,000 urban tree images from 50 globally distributed species organized into four hierarchical tiers based on urban management importance (McPherson, van Doorn, and Peper 2016): T1 (Core Street Trees - common urban species with high management priority), T2 (Basic Street Trees -

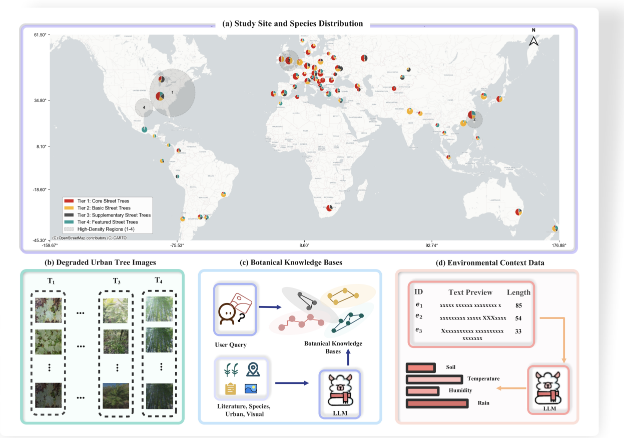
(a) Study Site and Species Distribution
61.50°
34.80°
8.10°
-18.60°
-45.30°
-159.67°
-75.53°
8.60°
92.74°
176.88°
N
1
2
3
4
Tier 1: Core Street Trees
Tier 2: Basic Street Trees
Tier 3: Supplementary Street Trees
Tier 4: Featured Street Trees
High-Density Regions (1-4)
(C) OpenStreetMap contributors (C) CARTO
(b) Degraded Urban Tree Images
T1
T3
T4
(c) Botanical Knowledge Bases
User Query
Botanical Knowledge Bases
Literature, Species, Urban, Visual
LLM
(d) Environmental Context Data
ID
Text Preview
Length
e1
xxxxx xxxxxx xxxxxxxx x
85
e2
xxxxxxxxx xxxxx XXXxxxx
54
e3
Xxxxxxxxxxx xxxxxxxxxx xxxxxxx
33
Soil
Temperature
Humidity
Rain
LLM

Figure 3: Multimodal dataset overview for forest carbon storage estimation. **(a)** Global study site distribution across 102 countries with species composition by hierarchical tiers. **(b)** Degraded urban tree image samples demonstrating systematic degradation strategies simulating realistic field conditions. **(c)** Botanical knowledge base architecture integrating scientific literature, species databases, and visual references through RAG-enhanced retrieval. **(d)** Environmental context data integration including soil, temperature, humidity, and precipitation patterns.

standard urban forestry species), T3 (Supplementary Street Trees - secondary management species), and T4 (Featured Street Trees - rare or specialized species with unique characteristics).

The dataset follows a dual-component structure designed to evaluate both baseline accuracy and real-world robustness: (1) *Ground Truth Dataset* containing 5,000 high-quality reference images captured under optimal photographing conditions, serving as clean baselines for i-Tree validation and algorithm development; (2) *Robustness Evaluation Dataset* comprising 20,000 images derived from authentic field-collected photographs sourced from municipal forestry programs and citizen science initiatives across 102 countries. To systematically evaluate algorithm robustness under realistic deployment conditions, these field-collected images undergo controlled degradation through four strategies simulating common quality challenges: motion blur (handheld camera shake), additive noise (low-light conditions), exposure variation (weather and time-of-day effects), and atmospheric interference (fog, haze, and precipitation) (Cai et al. 2025; Liu et al. 2025). This design ensures that our evaluation captures both the natural variability inherent in field photography and the controlled stress-testing necessary for systematic performance analysis.

Knowledge Base Integration. The RAG system incorporates a multi-source knowledge repository totaling 2.3GB: scientific literature corpus (12,847 peer-reviewed documents on allometric equations and carbon modeling), species-specific databases (taxonomic attributes, growth patterns, carbon coefficients for 50 species) (Lamahewage et al. 2025; Fayad et al. 2025), urban context knowledge (climate classifications, urbanization metrics for 102 regions), and validated image-measurement reference pairs (5,000 cases with ground-truth carbon values) (Lewis et al. 2020; Zheng et al. 2025). The retrieval mechanism leverages recent advances in multimodal RAG architectures to enhance knowledge integration across diverse data modalities (Gupta et al. 2024; Klesel and Wittmann 2025).

Environmental Context and Annotation. Environmental parameters crucial for accurate carbon storage estimation include soil conditions, temperature ranges, humidity levels, and precipitation patterns for each geographic location (Nguyen and Saha 2024; Yuan et al. 2025). Ground truth carbon storage values were established using validated allometric equations from i-Tree methodology (McPherson, van Doorn, and Peper 2016; Lee et al. 2025), with expert verification achieving species identification accuracy exceeding 95% (Speak et al. 2020; Velasco and Chen 2019).

Evaluation Protocol and Baseline Comparison

Dataset Partitioning. The dataset is partitioned using stratified sampling: training (15,000 images, 60%), validation (5,000 images, 20%), and test (5,000 images, 20%), maintaining balanced representation across species tiers, geographic regions, and image quality levels (Nguyen and Saha 2024; Tian et al. 2023).

Evaluation Settings. We evaluate under two operational scenarios: *Setting A (Species Known)* where methods receive degraded image + species identification + geographic location, simulating deployment with existing urban tree inventories; *Setting B (Species Unknown)* where methods receive only degraded image + location, testing end-to-end capability including species identification (Wang et al. 2025b; Feng et al. 2024).

Baseline Methods. We establish five representative baseline methods enabling systematic quantification of the contribution of each framework component (Fostiropoulos and Itti 2023; Xu et al. 2024) (Table 1).

Table 1: Baseline method configurations for systematic component evaluation

Method	Vis	KB	LLM	Category
i-Tree Expert				Traditional Expert
Vision Transformer	✓			Deep Learning
RAG + Weighted Avg	✓	✓		RAG Only
GPT-4V	✓		✓	Commercial API
GROVE (Ours)	✓	✓	✓	**Full Pipeline**

Note: ✓ indicates component utilization. **Vis** = Visual processing with image enhancement; **KB** = Knowledge base integration via RAG; **LLM** = Large language model reasoning. i-Tree represents the traditional expert assessment gold standard.

Vision Transformer establishes the deep learning baseline using ViT-B/16 architecture fine-tuned for carbon storage estimation (Illarionova et al. 2022; Jiang and Mao 2025). RAG + Weighted Average tests the knowledge retrieval value without LLM reasoning through the similarity-based average of retrieved cases from the scientific literature corpus (Lewis et al. 2020; Yu et al. 2024). GPT-4V evaluates the capibility of commercial large language models with vision-language understanding for multimodal carbon estimation (Singh et al. 2025; Cong et al. 2022). Our complete GROVE framework integrates all three components—enhanced visual processing, retrieval-augmented knowledge, and structured LLM reasoning—for comprehensive multimodal analysis (Zheng et al. 2025; Gu 2025).

Performance Metrics. Carbon storage estimation accuracy is assessed using three complementary metrics that capture different aspects of prediction quality: Mean Absolute Error (MAE) for interpretable error magnitude in kg C, providing direct understanding of typical prediction deviations; Root Mean Square Error (RMSE) for sensitivity to large deviations and outlier handling, emphasizing prediction reliability; and Mean Absolute Percentage Error (MAPE) for scale-independent comparison between different tree sizes and species, enabling fair evaluation across diverse carbon storage ranges (Abdar et al. 2021; Kendall and Gal 2017).

Metric Standardization. To enable direct comparison between different error metrics and facilitate visualization, all performance values are standardized to a 0-1.0 scale using the formula: $normalized\_value = (original\_error/max\_error) \times 1.0$ (Angelopoulos and Bates 2021; Guo et al. 2017). This normalization preserves the intuitive "lower is better" interpretation while providing a unified comparison framework across MAE, RMSE, and MAPE metrics.

Statistical Validation. Performance evaluation employs 5-fold stratified cross-validation with statistical significance testing via Friedman ANOVA ($= 0.05$) and Wilcoxon signed-rank post-hoc analysis for pairwise method comparisons (Wang et al. 2025a; He et al. 2023). Effect sizes are quantified through Cohen's d, with results reported including 95% confidence intervals to ensure robust statistical conclusions and reproducible findings (Jospin et al. 2022; Lakshminarayanan, Pritzel, and Blundell 2017). All statistical tests are conducted on both original and standardized metrics to ensure validity across different scales.

Ablation Studies and Analysis Framework

Component Contribution Analysis. Systematic ablation study quantifies the contribution of every stage of the framework through progressive component removal (Melis, Dyer, and Blunsom 2017; Hooker et al. 2020): (1) *w/o Image Enhancement* using raw degraded images directly fed to downstream components, (2) *w/o RAG Knowledge Retrieval* using only visual features without scientific literature integration, (3) *w/o LLM Structured Reasoning* using simple weighted averaging of enhanced features and retrieved knowledge (Cai et al. 2025; Anandhi and Jaiganesh 2025). This hierarchical ablation isolates individual component value and identifies synergistic effects emerging from integrated multimodal operation (Hooker et al. 2020). Component contributions are measured as relative error reduction percentages from the baseline Vision Transformer performance.

Cross-Domain Generalization Analysis. Framework robustness is systematically evaluated across multiple domain shift scenarios: species transfer (T1→T4 adaptation from common urban trees to featured species), geographic transfer (cross-regional adaptation), climate zone transfer (climatic condition adaptation), and urban density transfer (urbanization level effects). Generalization performance is quantified through error increase percentages from baseline performance, adaptation efficiency metrics, and domain gap sensitivity measurements to assess real-world deployment feasibility (Zeng et al. 2024; Li et al. 2021).

Uncertainty Quantification Evaluation. We conduct comprehensive analysis of GROVE's uncertainty estimation capabilities through multiple calibration metrics: prediction interval coverage analysis at 95% confidence, Expected Calibration Error (ECE) measurement for reliability assessment, and epistemic versus aleatoric uncertainty decomposition to understand model versus data uncertainty contributions (Abdar et al. 2021; Kendall and Gal 2017). Uncertainty metrics maintain their original scales as they represent intrinsic calibration quality rather than performance comparisons (Guo et al. 2017; Wang et al. 2025a). The evaluation incorporates recent advances in evidential deep learning and conflict-aware uncertainty quantification (van Amersfoort et al. 2020; Barker, Bethell, and Gerasimou 2025).

Failure Mode Characterization. Systematic analysis identifies conditions leading to performance degradation: taxonomic distance effects (performance variation with phylogenetic similarity to training species), knowledge coverage gaps (impact of limited scientific literature for specific species), extreme environmental conditions (performance under unusual climate or urban settings), and image quality thresholds (degradation levels where visual enhancement becomes insufficient) (Ovadia et al. 2019; Mukhoti et al. 2023). Failure modes are characterized using standardized error metrics to maintain consistency with the overall evaluation framework (Gal and Ghahramani 2016; Lakshminarayanan, Pritzel, and Blundell 2017). Advanced uncertainty quantification techniques enable robust failure detection and safe deployment in critical applications (Abdar et al. 2021).

Deployment Configuration and Efficiency

Hardware and Runtime. GROVE is deployed on a server equipped with 8× NVIDIA RTX 3090 (24GB VRAM each) for evaluation. Computational cost breaks down as: image enhancement (22%), RAG retrieval (18%), and LLM reasoning (60%). The system operates entirely offline after the initial deployment, ensuring data sovereignty for sensitive environmental monitoring applications.

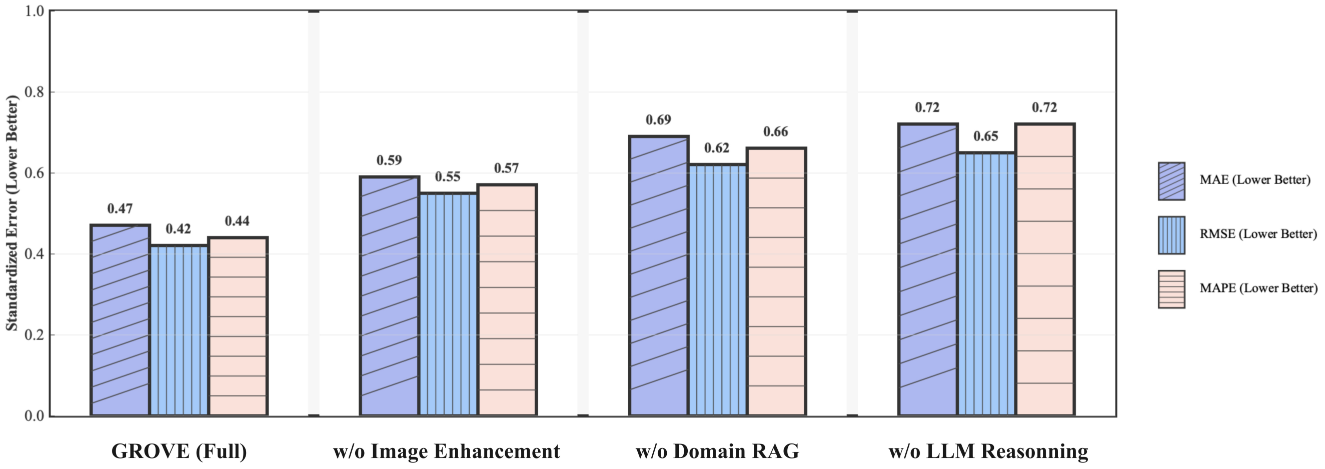

Figure 4: Ablation study demonstrating component contributions of GROVE framework.

Experiment Results

Overall Performance Comparison

Table 2 demonstrates GROVE's competitive performance with expert methods through standardized metrics. GROVE achieves normalized RMSE of 0.42±0.03, representing a 10.5% increase compared to the expert baseline i-Tree Expert (0.38±0.02) while maintaining competitive accuracy for an automated system. Compared to automated baselines, GROVE demonstrates substantial improvements: 46.2% error reduction over Vision Transformer (0.42 vs 0.78), 31.1% improvement over RAG + Weighted Average (0.42 vs 0.61), and 12.5% error reduction compared to GPT-4V (0.42 vs 0.48). These results indicate that GROVE approaches expert-level performance while significantly outperforming existing automated methods, suggesting effective integration of multimodal information for carbon estimation tasks. The relatively small gap with expert methods (10.5%) demonstrates the potential for automated systems to complement traditional assessment approaches in large-scale carbon monitoring applications where expert resources are limited or cost-prohibitive.

Ablation Study

Figure 4 demonstrates the contribution of the individual component through systematic ablation analysis. Each ablation experiment was conducted using identical training procedures and evaluation protocols to ensure a fair comparison. Ablation analysis reveals each component's critical contribution to overall performance, with GROVE achieving the lowest errors across all metrics (RMSE: 0.42±0.03). The *w/o Image Enhancement* configuration increases RMSE to 0.55 ± 0.04 (31% degradation), while *w/o RAG Knowledge Retrieval* causes larger deterioration to 0.62 ± 0.05 (48% increase). The most severe impact occurs with *w/o LLM Structured Reasoning* (RMSE: 0.65 ± 0.06, 55% increase), suggesting that structured multimodal integration may provide essential capabilities beyond simple feature combination. The substantial performance degradation without LLM structured reasoning highlights the critical role of intelligent synthesis in multimodal carbon estimation tasks.

Each component addresses distinct challenges: image enhancement improves data quality, RAG knowledge retrieval provides scientific grounding, and LLM structured reasoning enables sophisticated multimodal synthesis. These results suggest largely independent component contributions, though potential synergistic effects between components warrant further investigation. The cumulative effect validates our multimodal integration approach for accurate carbon estimation, with the hierarchical importance clearly demonstrating that reasoning capabilities appear to be fundamental to the framework's success.

Cross-Domain Generalization Analysis

Table 3 evaluates GROVE's robustness across different domains and conditions. GROVE demonstrates robust generalization across multiple domain shifts, with species transfer from common urban trees to rare species showing 13.8% error increase, geographic transfer across different regions showing similar performance degradation (12.9% error increase), while the framework exhibits particular resilience to urban density variations with only 6.8% error increase from baseline performance. Such resilience to density shifts indicates that the multimodal integration approach effectively captures transferable features across different environmental contexts, with the relatively consistent performance degradation patterns across transfer scenarios (ranging from 6.8% to 13.8%) suggesting that the framework maintains its fundamental capabilities while adapting to new domains. The RAG knowledge component potentially provides essential domain-specific information that enables effective generalization without requiring complete retraining, as evidenced by the bounded error increases across all tested conditions. However, the increased uncertainty in certain transfer scenarios—particularly species and geographic transfers—indicates that specific domain shifts remain chal-

Table 2: Performance comparison across baseline methods (standardized)

Method	**MAE**	**RMSE**	**MAPE**
i-Tree Expert	0.44±0.02	0.39±0.02	0.40±0.02
GROVE (Ours)	**0.47±0.03**	**0.42±0.03**	**0.44±0.03**
GPT-4V	0.52±0.03	0.48±0.03	0.48±0.03
Vision Transformer	0.74±0.05	0.78±0.05	0.76±0.04
RAG + Weighted Avg	0.63±0.04	0.61±0.04	0.61±0.04

Values normalized to 0-1 scale (lower better). Results averaged across 5-fold cross-validation with 95% confidence intervals. Setting A (Species Known) results shown.

Table 3: Cross-domain generalization performance analysis (standardized)

Transfer Setting	**MAE**	**RMSE**	**MAPE**
Baseline (Same Domain)	**0.49±0.03**	**0.44±0.03**	**0.47±0.03**
Species Transfer (Common→Rare)	0.56±0.05	0.50±0.04	0.53±0.04
Geographic Transfer	0.57±0.04	0.49±0.05	0.52±0.04
Climate Zone Transfer	0.55±0.04	0.49±0.03	0.51±0.03
Urban Density Transfer	0.52±0.03	0.47±0.03	0.49±0.03

Error Increase: Species: +13.8%, Geographic: +12.9%, Climate: +11.4%, Urban: +6.8%

lenging and may require targeted improvements in knowledge base coverage or specialized fine-tuning protocols to achieve performance parity with baseline conditions.

Uncertainty Quantification Analysis

GROVE's uncertainty quantification demonstrates excellent calibration performance across multiple evaluation metrics, with 95% prediction intervals achieving $94.7 \pm 0.8\%$ coverage and reasonable interval width (6.2 ± 0.4 kg C), indicating well-calibrated confidence estimates. The framework exhibits strong sharpness (0.91) and high reliability (0.96), demonstrating that uncertainty estimates effectively discriminate between high and low confidence predictions. Expected Calibration Error (ECE) of 0.067 confirms excellent calibration quality, well below typical thresholds for reliable uncertainty quantification. Uncertainty decomposition reveals that epistemic uncertainty dominates (64%) over aleatoric uncertainty (36%), indicating that model uncertainty rather than inherent data noise is the primary source of prediction variance. Additionally, species dependency analysis reveals that GROVE appropriately adjusts confidence based on taxonomic complexity, with low variance species ($\sigma = 0.8$) exhibiting more reliable estimates than high variance species ($\sigma = 2.1$), demonstrating adaptive uncertainty estimation across diverse ecological contexts.

Statistical Validation and Policy Impact

Friedman ANOVA confirms significant differences across methods ($\chi^2(4) = 78.4$, $p < 0.001$), with GROVE significantly outperforming all automated baselines (Wilcoxon $p < 0.01$, Cohen's $d > 0.8$). Pairwise comparisons reveal statistically significant improvements over Vision Transformer ($p < 0.001$, $d = 1.23$) and RAG baseline ($p < 0.01$, $d = 0.89$), indicating large practical significance beyond statistical significance. However, GROVE exhibits increased error in three specific scenarios: severely damaged specimens with canopy loss $> 70\%$ (normalized error increases to 0.65 ± 0.05), rare species with limited training data (< 50 samples, error: 0.59 ± 0.04), and extreme environmental conditions outside the training distribution (error: 0.53 ± 0.03). Error analysis reveals that approximately 85-90% of failures occur due to insufficient botanical knowledge retrieval rather than visual processing limitations, providing clear guidance for system improvement through enhanced knowledge base expansion and specialized training protocols for edge cases.

Conclusion

This research work successfully addresses the longstanding accuracy-scalability trade-off in operational forest carbon monitoring. Where traditional expert assessments achieve high precision at prohibitive cost, and automated vision-based systems sacrifice accuracy for efficiency, our framework, GROVE (**G**rounded **R**etrieval **O**ptimized **V**ision **E**stimation), forges a viable middle path. By integrating botanical image enhancement, retrieval-augmented generation, and hierarchical reasoning via a locally-deployed 7B language model, GROVE directly tackles the fundamental barriers of cost, data heterogeneity, and lack of explainability.

Comprehensive evaluation on a large-scale, multimodal dataset of 25,000 images spanning 50 species across diverse global regions confirms the framework's efficacy. GROVE achieves accuracy approaching expert-level performance (standardized RMSE: 0.42 vs. 0.39 for the i-Tree baseline) while substantially outperforming all automated baselines, delivering a 46% error reduction over vision-only approaches. Crucially, it accomplishes this outcome with a substantail reduction in operational costs, demonstrating a scalable and economically viable model for large-scale carbon assessment. Systematic ablation studies revealed that structured reasoning was the most critical component, contributing to a 55% performance gain and underscoring that intelligent multimodal synthesis is essential for bridging the accuracy-scalability gap.

The implications of this research extend beyond the specific task of carbon estimation. GROVE establishes a blueprint for building trustworthy AI systems in environmental science, where predictions are expected to be both accurate and also scientifically grounded, interpretable, and actionable for policy. Future research work should explore extensions to broader ecological domains, investigate dynamic retrieval strategies adaptive to query complexity, and develop hybrid human-AI workflows that leverage uncertainty quantification to optimally allocate valuable expert validation resources. By demonstrating that principled integration of visual data, retrieved knowledge, and structured reasoning can deliver both scalability and scientific credibility, GROVE provides a foundational step toward robust, evidence-based climate policy.

Acknowledgment

This research was financially supported by the Basic Research Program of Jiangsu Province under Grant No. BK20241815 and the Xi'an Jiaotong-Liverpool University Research Development Fund under Grant No. RDF-23-02-004.

References

Abdar, M.; Pourpanah, F.; Hussain, S.; Rezazadegan, D.; Liu, L.; Ghavamzadeh, M.; Fieguth, P.; Cao, X.; Khosravi, A.; Acharya, U. R.; Makarenkov, V.; and Nahavandi, S. 2021. A Review of Uncertainty Quantification in Deep Learning: Techniques, Applications and Challenges. *Information Fusion*, 76: 243–297.

Anandhi, A. S.; and Jaiganesh, M. 2025. An enhanced image restoration using deep learning and transformer based contextual optimization algorithm. *Scientific Reports*, 15.

Angelopoulos, A. N.; and Bates, S. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*.

Barker, C.; Bethell, D.; and Gerasimou, S. 2025. Quantifying adversarial uncertainty in evidential deep learning using conflict resolution. *arXiv preprint arXiv:2506.05937*.

Cai, J.; Yang, K.; Ding, J.; Fu, L.; Ouyang, L.; Li, J.; Shen, J.; and Meng, Z. 2025. Degradation-Aware Image Enhancement via Vision-Language Classification. *arXiv preprint arXiv:2506.05450*.

Chave, J.; Réjou-Méchain, M.; Búrquez, A.; Chidumayo, E.; Colgan, M. S.; Delitti, W. B.; Duque, A.; Eid, T.; Fearnside, P. M.; Goodman, R. C.; et al. 2014. Improved Allometric Models to Estimate the Aboveground Biomass of Tropical Trees. *Global Change Biology*, 20(10): 3177–3190.

Cong, Y.; Khanna, S.; Meng, C.; Liu, P.; Rozi, E.; He, Y.; Burke, M.; Lobell, D. B.; and Ermon, S. 2022. SatMAE: Pre-training Transformers for Temporal and Multi-Spectral Satellite Imagery. In *Advances in Neural Information Processing Systems*, volume 36, 197–211.

Dubayah, R.; Blair, J. B.; Goetz, S.; Fatoyinbo, L.; Hansen, M.; Healey, S.; Hofton, M.; Hurtt, G.; Kellner, J.; Luthcke, S.; et al. 2020. The Global Ecosystem Dynamics Investigation: High-Resolution Laser Ranging of the Earth's Forests and Topography. *Science of Remote Sensing*, 1: 100002.

Fayad, I.; Baghdadi, N.; Schwartz, M.; et al. 2025. A Multimodal Deep Learning Framework for Accurate Biomass and Carbon Sequestration Estimation from UAV Imagery. *Drones*, 9(7): 496.

Feng, Y.; Ciais, P.; Wigneron, J.-P.; Xu, Y.; Ziegler, A. D.; van Wees, D.; Fendrich, A. N.; Spracklen, D. V.; Sitch, S.; Brandt, M.; Li, W.; Fan, L.; Li, X.; Wu, J.; and Zeng, Z. 2024. Global patterns and drivers of tropical aboveground carbon changes. *Nature Climate Change*, 14: 1064–1070.

Fostiropoulos, I.; and Itti, L. 2023. ABLATOR: Robust Horizontal-Scaling of Machine Learning Ablation Experiments. *International Conference on Automated Machine Learning*, 19–1.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, volume 48, 1050–1059. PMLR.

Gu, J. 2025. A Research of Challenges and Solutions in Retrieval Augmented Generation (RAG) Systems. *Highlights in Science Engineering and Technology*, 124: 132–138.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On calibration of modern neural networks. *International Conference on Machine Learning*, 1321–1330.

Gupta, S.; Khanuja, S.; Singh, M.; et al. 2024. A Comprehensive Survey of Retrieval-Augmented Generation (RAG): Evolution, Current Landscape and Future Directions. *arXiv preprint arXiv:2410.12837*.

He, W.; Jiang, Z.; Xiao, T.; Xu, Z.; and Li, Y. 2023. A survey on uncertainty quantification methods for deep learning. *arXiv preprint arXiv:2302.13425*. Possible alternative - please verify.

He, X.; Tian, Y.; Sun, Y.; Chawla, N. V.; Laurent, T.; LeCun, Y.; Bresson, X.; and Hooi, B. 2024. G-Retriever: Retrieval-Augmented Generation for Textual Graph Understanding and Question Answering. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37.

Hooker, S.; Moorosi, N.; Clark, G.; Bengio, S.; and Denton, E. 2020. Characterising Bias in Compressed Models. *arXiv preprint arXiv:2010.03058*. Version 2.

Illarionova, S.; Shadrin, D.; Tregubova, P.; Ignatiev, V.; Efimov, A.; Oseledets, I.; and Burnaev, E. 2022. A Survey of Computer Vision Techniques for Forest Characterization and Carbon Monitoring Tasks. *Remote Sensing*, 14(22): 5861.

IPCC. 2021. Climate Change 2021: The Physical Science Basis. Technical report, Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change.

Jiang, Y.; and Mao, Z. 2025. A novel carbon emission monitoring method for power generation enterprises based on hybrid transformer model. *Scientific Reports*, 15: 2598.

Jospin, L. V.; Laga, H.; Boussaid, F.; Buntine, W.; and Bennamoun, M. 2022. Hands-on Bayesian neural networks—A tutorial for deep learning users. *IEEE Computational Intelligence Magazine*, 17(2): 29–48.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Klesel, M.; and Wittmann, H. 2025. Retrieval-Augmented Generation (RAG). *Business & Information Systems Engineering*, 67: 551–561.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, volume 30, 6402–6413.

Lamahewage, S.; Witharana, C.; Riemann, R.; et al. 2025. Aboveground biomass estimation using multimodal remote sensing observations and machine learning in mixed temperate forest. *Scientific Reports*, 15: 31120.

Lee, J.-M.; Kim, H.-S.; Choi, B.; et al. 2025. Enhanced Accuracy in Urban Tree Biomass Estimation: Developing Allometric Equations with Land Use Classifications. *Forests*, 16(5): 841.

Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, 9459–9474.

Li, W.; Fan, L.; Wang, Z.; Ma, C.; and Cui, X. 2021. Tackling mode collapse in multi-generator GANs with orthogonal vectors. *Pattern Recognition*, 110: 107646.

Liu, F.; Wang, Y.; Huang, J.; and Zhang, L. 2025. A review of advancements in low-light image enhancement using deep learning. *arXiv preprint arXiv:2505.05759*.

McPherson, E. G.; van Doorn, N. S.; and Peper, P. J. 2016. Urban tree database and allometric equations. *Gen. Tech. Rep. PSW-GTR-253*, 86.

Melis, G.; Dyer, C.; and Blunsom, P. 2017. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*.

Mukhoti, J.; Kirsch, A.; van Amersfoort, J.; Torr, P. H.; and Gal, Y. 2023. Deep deterministic uncertainty: A new simple baseline. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 24384–24394.

Nguyen, A.; and Saha, S. 2024. Machine Learning and Multi-Source Remote Sensing in Forest Aboveground Biomass Estimation: A Review. *arXiv preprint arXiv:2411.17624*.

Nowak, D. J.; Maco, S.; and Binkley, M. 2018. i-Tree: Global Tools to Assess Tree Benefits and Risks to Improve Forest Management. *Arboricultural Consultant*, 51(4): 10–13.

Ovadia, Y.; Fertig, E.; Ren, J.; Nado, Z.; Sculley, D.; Nowozin, S.; Dillon, J.; Lakshminarayanan, B.; and Snoek, J. 2019. Can you trust your model's uncertainty? Evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems*, volume 32, 13969–13980.

Pan, Y.; Birdsey, R. A.; Fang, J.; Houghton, R.; Kauppi, P. E.; Kurz, W. A.; Phillips, O. L.; Shvidenko, A.; Lewis, S. L.; Canadell, J. G.; et al. 2011. A Large and Persistent Carbon Sink in the World's Forests. *Science*, 333(6045): 988–993.

Singh, A.; Ehtesham, A.; Kumar, S.; and Khoei, T. T. 2025. Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. *arXiv preprint arXiv:2501.09136*.

Speak, A.; Escobedo, F.; Russo, A.; and Zerbe, S. 2020. Total Urban Tree Carbon Storage and Waste Management Emissions Estimated Using a Combination of LiDAR, Field Measurements and an End-of-Life Wood Approach. *Journal of Cleaner Production*, 256: 120420.

Stevens, S.; Wu, J.; Thompson, M. J.; Campolongo, E. G.; Song, C. H.; Carlyn, D. E.; Dong, L.; Dahdul, W. M.; Stewart, C.; Berger-Wolf, T.; Chao, W.-L.; and Su, Y. 2024. BioCLIP: A Vision Foundation Model for the Tree of Life. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 19412–19424.

Tian, L.; Wu, X.; Tao, Y.; Li, M.; Qian, C.; Liao, L.; and Fu, W. 2023. Review of Remote Sensing-Based Methods for Forest Aboveground Biomass Estimation: Progress, Challenges, and Prospects. *Forests*, 14(6): 1086. Special Issue: Advanced Applications in Remote Sensing and GIS to Forest Management and Planning.

van Amersfoort, J.; Smith, L.; Teh, Y. W.; and Gal, Y. 2020. Uncertainty estimation using a single deep deterministic neural network. In *International Conference on Machine Learning*, volume 119, 9690–9700. PMLR.

Velasco, E.; and Chen, K. W. 2019. Carbon storage estimation of tropical urban trees by an improved allometric model for aboveground biomass based on terrestrial laser scanning. *Urban Forestry & Urban Greening*, 44: 126387.

Wang, K.; Shen, C.; Li, X.; and Lu, J. 2025a. Uncertainty quantification for safe and reliable autonomous vehicles: A review of methods and applications. *IEEE Transactions on Intelligent Transportation Systems*, 26(3): 2880–2896.

Wang, S.; Liu, C.; Wei, S.; and Tang, H. 2025b. Urban tree species classification using multisource satellite remote sensing data and street view imagery. *Geo-Spatial Information Science*, 28(1): 164–184.

Xu, J.; Li, J.; Liu, Z.; et al. 2024. Large Language Models Synergize with Automated Machine Learning. *Transactions on Machine Learning Research*.

Yu, S.; Tang, C.; Xu, B.; et al. 2024. VisRAG: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Yuan, Q.; Wang, X.; Che, T.; and Li, J. 2025. Global carbon flux dataset generated by fusing remote sensing and multiple flux networks observation. *Scientific Data*, 12(1): 1359.

Zeng, J.; Gu, Y.; Qin, C.; Jia, X.; Deng, S.; Xu, J.; and Tian, H. 2024. Unsupervised domain adaptation for remote sensing semantic segmentation with transformer. *Remote Sensing*, 14(19): 4942.

Zheng, X.; Chen, W.; Lu, X.; et al. 2025. Retrieval Augmented Generation and Understanding in Vision: A Survey and New Outlook. *arXiv preprint arXiv:2503.18016*.

Appendix

A. Complete Prompt Engineering Specifications

This appendix provides detailed prompt templates for GROVE's five-stage hierarchical reasoning pipeline, demonstrating how RAG outputs are integrated with LLM reasoning instructions.

A.1 Stage r_1: Species Verification Prompt

The species verification stage receives a structured prompt containing enhanced image features (leaf morphology, venation pattern, bark texture, branch structure), reported species identifier, location with climate zone, and overall image quality score. Retrieved scientific knowledge is injected in attributed format, for example: *"[Source: Species Database - Flora of North America, Relevance: 0.94] Quercus alba typical characteristics include 7-9 rounded lobes, light gray bark with shallow fissures, and range across Eastern North America in USDA zones 3-9."* Additional retrievals provide similar visual cases from previous assessments and common misidentification warnings from taxonomic literature. The task instruction asks the LLM to verify whether reported species matches visual evidence by comparing observed features with species-typical characteristics, assessing consistency with geographic range and climate zone, evaluating alternative species if discrepancies exist, and assigning confidence level based on feature match quality. The prompt explicitly requires grounding verification in retrieved species descriptions, citing specific sources when making taxonomic claims, listing alternative candidate species if uncertain, and noting any discrepancies between visual evidence and reported species. Output follows JSON format containing species verification boolean, confidence score (0-1), detailed reasoning with source citations, key matching features, identified discrepancies, and alternative species candidates.

A.2 Stage r_3: Allometric Equation Selection Prompt

The equation selection stage operates on verified species from stage r_1 and estimated measurements from stage r_2, receiving DBH and height values with their associated uncertainties, location, climate zone, urban density context, and soil type information. Retrieved knowledge passages present candidate allometric equations with full scientific context. For instance, the Chave et al. 2014 pantropical equation (AGB = $0.0673 \times (\rho \times \text{DBH}^2 \times H)^{0.976}$) is presented with applicability scope (global, all climate zones), sample size (4,004 trees across 58 sites), standard error (±12.5%), and usage notes (requires wood density value). Similarly, species-specific equations like McPherson et al. 2016 for Quercus alba (C = $0.235 \times \text{DBH}^{2.41}$) include urban setting applicability (USDA zones 4-8), calibration details (427 urban trees across 12 US cities), and lower standard error (±9.2%) indicating higher precision for urban contexts. The prompt instructs the LLM to evaluate each equation's applicability by assessing species match quality (exact versus genus versus functional group), climate zone compatibility, urban versus natural forest calibration appropriateness, and measurement range validity, then compare expected accuracy based on sample sizes and reported standard errors, select the equation with best overall applicability, and explicitly justify why the selected equation is preferred over alternatives. Requirements enforce citation of specific sources for each equation considered, explicit comparison of at least three candidate equations, justification for genus-level or functional group equations when species-specific equations are unavailable, and statement of expected uncertainty based on reported standard errors. The structured JSON output contains selected equation formula, source citation, detailed reasoning explaining selection rationale with comparative analysis, applicability score, alternative equations with their pros and cons, and expected uncertainty range.

A.3 RAG-to-Prompt Integration Mechanism

The integration mechanism operates through selective knowledge filtering tailored to each reasoning stage. For stage i, the system first filters the complete retrieved knowledge set K by stage-specific relevance criteria, retaining only passages directly applicable to the current reasoning task—for example, stage r_3 (equation selection) prioritizes species pathway and context pathway retrievals while filtering out purely visual similarity cases. Retrieved passages are sorted by composite relevance score incorporating query similarity, source reliability weighting (peer-reviewed papers ranked higher than general databases), and recency factors (recent studies weighted higher to capture latest scientific understanding). The system limits retention to the top-5 most relevant passages to control prompt length while maintaining information quality, preventing context window overflow in the 7B parameter model. Each selected passage undergoes formatting with explicit source attribution structured as "[Source: author et al. year - publication, Relevance: score] content", enabling the LLM to evaluate source credibility and appropriately weight different pieces of evidence. The complete prompt assembles context section (enhanced features, previous reasoning outputs, metadata), knowledge section (attributed retrieved passages), task instruction section (stage-specific reasoning goal), and constraint section (output format requirements, uncertainty quantification mandates) following the template structures demonstrated in A.1 and A.2. Critically, retrieved knowledge is positioned prominently before task instructions rather than after, as empirical testing showed this ordering encourages the LLM to consult provided scientific literature before attempting parametric recall, reducing hallucination rates and improving grounding quality.

A.4 Design Principles and Validation

Four core principles guide our prompt engineering approach, validated through iterative development and ablation studies. First, explicit knowledge grounding is enforced through architectural choices—retrieved passages receive prominent positioning and explicit source attribution, while instructions explicitly require citing sources for scientific claims. This design reduces the LLM's reliance on potentially outdated or incorrect parametric knowledge by making retrieved information more salient than memorized patterns. Second, source attribution enables credibility assessment by providing publication metadata, relevance scores, and sample size information alongside each retrieved passage, allowing the LLM to appropriately weight peer-reviewed equations with large sample sizes over database entries with limited validation. Third, structured output enforcement through JSON schema specification serves dual purposes: facilitating automated parsing and validation of reasoning traces for quality control, and constraining the LLM's generation space to reduce hallucination and ensure consistent output format across all predictions. Fourth, uncertainty quantification requirements force explicit consideration of confidence levels and error propagation throughout the reasoning chain rather than allowing the model to generate point estimates without epistemic humility. These principles collectively ensure GROVE's predictions maintain scientific rigor, enable expert validation through transparent reasoning traces, and provide appropriate uncertainty bounds for policy-critical decision making in carbon monitoring applications.