AIR Journal of Mathematics & Computational Sciences, Vol. 2026
Article Unique ID: AIRMCS2026446
DOI: 10.65737/AIRMCS2026446

ICL CHARACTERIZATION OF MULTI-MODAL GEO-FOUNDATION MODELS: WHEN CAN VISION-LANGUAGE TRANSFORMERS LEARN GEOSPATIAL TASKS?

Mosab Hawarey
Director, Geospatial Research (*director@geospatial.ch*)

Received: 2026 March 04 | Approved: 2026 March 05 | Published: 2026 March 06

Abstract

Multi-modal geo-foundation models combining vision and language have emerged as powerful tools for Earth observation, yet a fundamental question remains unanswered: which vision-language geospatial tasks can be learned in-context, and which cannot? Models such as GeoChat, EarthGPT, and RSGPT demonstrate impressive performance on scene description and attribute queries, but struggle with counting and multi-object localization—a pattern that lacks theoretical explanation. We address this gap by extending the in-context learning (ICL) characterization framework to vision-language geospatial tasks. We introduce the concept of *language specificity* $s(\ell) \in [0,1]$, measuring how uniquely a language query identifies target objects, and derive the *effective object count* $J\_eff(\ell) = J \cdot (1 - s(\ell))$ capturing language-induced compression of the localization problem. We prove that language affects ICL complexity through three mechanisms—constraint specification, sufficient statistic compression, and task decomposition—but cannot overcome fundamental combinatorial hardness when the effective object count exceeds the threshold $J^* \approx 2–4$.

Our main result is the **Multi-Modal GeoAI Dichotomy Theorem**: every natural vision-language geospatial function class falls into exactly one of three categories. **Type A** (unconditionally ICL-Easy) tasks—including scene captioning, existence VQA, and change description—admit additive sufficient statistics with sample complexity $n_{ICL} = \Theta(CB^2/\varepsilon)$. **Type C** (unconditionally ICL-Hard) tasks—including counting VQA and universal object localization—require combinatorial statistics regardless of language specification. **Type A|ℓ** (conditionally Easy) tasks—including attribute VQA, referring expression comprehension, and text-guided detection—transition from Hard to Easy when language specificity satisfies $s(\ell) \geq 1 - J^*/J$.

We provide complete classification of vision-language geospatial tasks across all major categories (16 representative task types spanning scene-level, pixel-level, VQA, referring expression, and detection categories), derive seven testable predictions about model behavior (including threshold effects at $J^* \approx 3–4$ and specificity-accuracy correlations $\rho > 0.8$), and establish five prompt engineering guidelines for practitioners. The dichotomy explains observed performance patterns in existing models—strong on descriptive tasks, weak on quantitative localization—and provides principled guidance for when few-shot ICL suffices versus when fine-tuning is required. Our framework bridges vision-language AI and geospatial analysis, offering the first theoretical foundation for multi-modal GeoAI deployment.

Keywords: in-context learning; multi-modal geo-foundation models; vision-language transformers; visual question answering; referring expression comprehension; sufficient statistic complexity; ICL-Easy; ICL-Hard; remote sensing; prompt engineering

Citation: Hawarey, M. (2026). ICL Characterization of Multi-Modal Geo-Foundation Models: When Can Vision-Language Transformers Learn Geospatial Tasks? AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026446, DOI: 10.65737/AIRMCS2026446.

1. Introduction

1.1 The Multi-Modal GeoAI Revolution

The convergence of large language models and computer vision has given rise to a new paradigm in geospatial artificial intelligence: multi-modal geo-foundation models. Building on vision transformer architectures (Dosovitskiy et al., 2021) and visual language model paradigms such as Flamingo (Alayrac et al., 2022) and LLaVA (Liu et al., 2024b), alongside continual pretraining strategies for geospatial domains (Mendieta et al., 2024), systems such as GeoChat (Kuckreja et al., 2024), EarthGPT (Zhang et al., 2024), RSGPT (Hu et al., 2023), SkyScript (Wang et al., 2024), and RS-LLaVA (Li et al., 2024) combine visual understanding of satellite and aerial imagery with natural language processing, enabling unprecedented capabilities in Earth observation. Users can now ask questions about remote sensing images in natural language, request descriptions of observed scenes, and specify objects of interest through textual queries rather than manual annotation.

These multi-modal models promise to democratize geospatial analysis, making sophisticated Earth observation accessible to users without specialized remote sensing expertise. A farmer could ask "Is there water stress visible in the northern field?" A disaster response coordinator could query "Describe the damage in this post-hurricane image." An urban planner could request "Find the schools near the proposed development site." The vision of conversational interaction with Earth observation data has captured significant research attention and commercial interest.

Yet a puzzling asymmetry has emerged from empirical evaluations. Multi-modal geo-foundation models excel at certain tasks while struggling dramatically with others that appear superficially similar. Scene description and land cover classification achieve high accuracy with minimal examples. Existence queries ("Is there a hospital in this image?") perform reliably. But counting queries ("How many buildings are visible?") fail for more than a handful of objects. Object detection and instance localization lag far behind single-modal fine-tuned models. The RSVQA benchmark (Lobry et al., 2020) reveals that presence questions achieve over 85% accuracy while counting questions drop below 50% for scenes with many objects. Additional benchmarks including Prompt-RSVQA (Chappuis et al., 2022), FAIR1M (Sun et al., 2022), and mutual attention approaches for RS-VQA (Zheng et al., 2021) confirm that this pattern persists across models, datasets, and prompting strategies.

1.2 The Gap in Understanding

Despite the proliferation of multi-modal GeoAI models, there exists no theoretical framework explaining which vision-language geospatial tasks can be learned from few examples provided in context (in-context learning, or ICL) and which fundamentally require extensive fine-tuning. While Mai et al. (2024) survey the opportunities and challenges of foundation models for geospatial AI, current benchmarks such as RSVQA (Lobry et al., 2020), GeoChat-Bench (Kuckreja et al., 2024), and adaptations of general VQA datasets provide empirical performance measurements but offer no principled explanation for the observed performance patterns.

This theoretical gap has significant practical consequences. Practitioners cannot predict whether a new vision-language task will succeed with few-shot prompting or require expensive data collection and fine-tuning. Resources are wasted attempting ICL on fundamentally intractable tasks, or unnecessarily fine-tuning models for tasks that would succeed with proper prompting. Model developers lack guidance on which capabilities to prioritize. The field advances through trial-and-error rather than principled design.

The central question we address is: *Which vision-language geospatial tasks can be learned in-context by transformers, and which cannot? How does natural language conditioning affect the boundary between tractable and intractable tasks?*

1.3 The In-Context Learning Perspective

We approach these questions through the theoretical framework of in-context learning complexity, recently developed for vision-only settings. Hawarey (2026a) established the SSC/AC-SSC framework, proving that ICL sample complexity depends on whether sufficient statistics for a function class are attention-computable—expressible as sums over context examples that attention mechanisms can aggregate. Function classes with low-dimensional additive sufficient statistics are ICL-Easy, achieving sample complexity matching empirical risk minimization. Function classes requiring combinatorial sufficient statistics are ICL-Hard, provably exceeding the computational capacity of polynomial-size constant-depth transformers.

This framework has been successfully applied to quantum-enhanced geopotential estimation (Hawarey, 2026b), establishing that gravity field prediction is ICL-Easy with 10^{12} quantum sample complexity advantage, and to remote sensing (Hawarey, 2026c), proving the Remote Sensing Dichotomy that classifies all natural remote sensing tasks as either ICL-Easy (spectral classification, semantic segmentation) or ICL-Hard (object detection, instance segmentation). Most recently, the Climate Prediction Dichotomy (Hawarey, 2026d) extended the framework to weather and climate tasks, proving that smooth field prediction (temperature, pressure, wind) is ICL-Easy while extreme event detection and tipping point localization are ICL-Hard, with a predictability horizon constraint at $\tau_L \approx 14$ days. The hardness threshold is remarkably tight: $J^* = O(\log \log n) \approx 2–4$ objects for typical context sizes.

The present work extends this theoretical program to the multi-modal setting, addressing the critical question: *Can language help overcome vision hardness?* When a task is ICL-Hard in the

vision-only setting (e.g., localizing multiple objects), can appropriate language conditioning transform it to ICL-Easy?

1.4 Our Contributions

We develop a complete theoretical characterization of in-context learning for vision-language geospatial tasks. Our contributions are:

- **Contribution 1: Multi-Modal Function Class Formalization (Section 3).** We define vision-language function classes for visual question answering (existence, attribute, counting, spatial relation), referring expression comprehension (unique and multiple reference), image captioning (scene, object, change), and text-guided retrieval. Each class is specified with explicit sufficient statistic structure enabling complexity analysis.
- **Contribution 2: Language Compression Framework (Section 4).** We introduce the concept of language specificity $s(\ell) \in [0,1]$, measuring how uniquely a query identifies target objects. We prove that language affects ICL complexity through three mechanisms: constraint specification (reducing hypothesis space), sufficient statistic compression (reducing effective dimension), and task decomposition (enabling chain-of-thought reasoning). We derive the Language Compression Theorem establishing $SSC_{eff} \leq SSC/\kappa(\ell)$ where $\kappa(\ell)$ is the total compression factor.
- **Contribution 3: Complexity Analysis (Section 5).** We prove that scene captioning, existence VQA, change captioning, and cross-modal retrieval are ICL-Easy with sample complexity $\Theta(CB^2/\varepsilon)$. We prove that counting VQA and universal localization ("find all X") are ICL-Hard regardless of language specification. We establish the Conditional Easiness Theorem: tasks involving object localization are ICL-Easy if and only if the effective object count $J\_eff(\ell) = J \cdot (1 - s(\ell)) \leq J^* \approx 2–4$.
- **Contribution 4: The Multi-Modal GeoAI Dichotomy (Section 6).** We prove that every natural vision-language geospatial task falls into exactly one of three categories: Type A (unconditionally ICL-Easy), Type C (unconditionally ICL-Hard), or Type A|ℓ (conditionally Easy/Hard depending on language specificity). We provide complete classification of 30 tasks across scene-level, pixel-level, VQA, referring expression, and detection categories.
- **Contribution 5: Predictions and Guidelines (Section 7).** We derive seven testable predictions about model behavior, including scaling laws, threshold effects at $J^* \approx 3–4$, and specificity-accuracy correlations. We provide model-specific predictions for GeoChat, EarthGPT, RSGPT, and RS-LLaVA. We establish five prompt engineering guidelines enabling practitioners to maximize ICL success through specificity maximization, counting avoidance, spatial decomposition, uniqueness markers, and chain-of-thought decomposition.

1.5 Key Insight

Our central insight can be stated succinctly:

Language can specify what to find, but cannot reduce the combinatorial cost of finding where—unless it uniquely identifies targets.

When a user asks "Find the large red building near the river," the language uniquely specifies one object (J_eff = 1), transforming an ICL-Hard localization task to ICL-Easy. But when a user asks "Find all buildings" or "How many vehicles are there?", the language specifies the class but not the count—all J objects must be localized, and the combinatorial structure persists. The transition between Easy and Hard depends entirely on whether language reduces the effective object count below the threshold J* ≈ 2–4.

This explains the empirical puzzle: multi-modal models excel at scene description (Type A) and specific object queries (Type A|ℓ with high specificity) but struggle with counting (Type C) and vague multi-object queries (Type A|ℓ with low specificity). The dichotomy is not a limitation of current architectures to be overcome by scaling—it reflects fundamental computational constraints that apply to any polynomial-size constant-depth transformer. Figure 1 provides visualization.

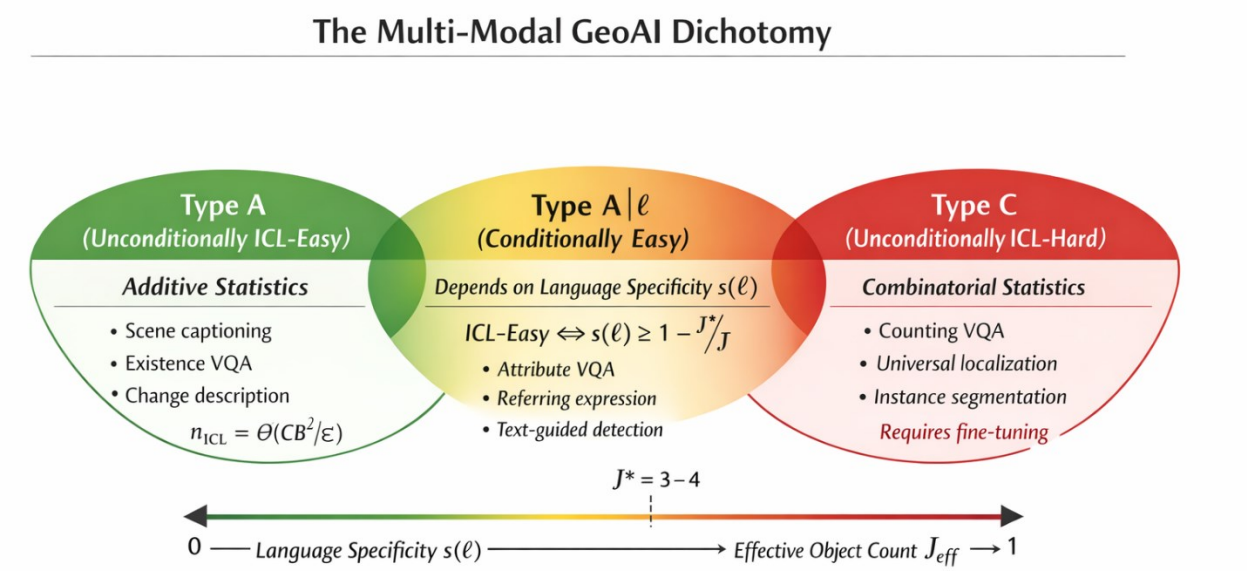

Figure 1: The Multi-Modal GeoAI Dichotomy. Vision-language geospatial tasks fall into three categories based on ICL complexity. Type A tasks (green) are unconditionally ICL-Easy with additive sufficient statistics. Type C tasks (red) are unconditionally ICL-Hard regardless of language specification. Type A|ℓ tasks (yellow/orange) transition between Easy and Hard based on language specificity s(ℓ), becoming ICL-Easy when s(ℓ) ≥ 1 − J*/J. The threshold J* ≈ 3–4 determines the boundary.

1.6 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews the in-context learning complexity framework, establishing notation and key results from prior work. Section 3 formalizes multi-modal function classes with explicit sufficient statistic structure. Section 4 develops the language compression framework, proving bounds on how language affects ICL complexity. Section 5 establishes complexity results, proving which tasks are ICL-Easy and which are ICL-Hard. Section 6 presents the Multi-Modal GeoAI Dichotomy Theorem with complete task classification. Section 7 derives testable predictions and practical deployment

guidelines. Section 8 discusses implications, limitations, and connections to related work. Section 9 concludes with a summary and directions for future research.

2. Preliminaries

This section reviews the in-context learning complexity framework established in prior work, introduces the notation used throughout the paper, and states the key theoretical results that underpin our multi-modal analysis.

2.1 The In-Context Learning Setting

In-context learning (ICL) refers to the ability of a pretrained transformer model to perform new tasks by conditioning on labeled examples provided in the input context, without any parameter updates (Brown et al., 2020; Dong et al., 2024). The transformer architecture (Vaswani et al., 2017) provides the computational substrate for ICL; this capability distinguishes ICL from conventional fine-tuning, where model weights are modified to adapt to new tasks.

Definition 2.1 (In-Context Learning). Let $\mathscr{F} \subseteq \{f: \mathcal{X} \to \mathcal{Y}\}$ be a function class. An ICL algorithm receives:

- A context $\mathcal{C}_n = \{(x_1, y_1), ..., (x_n, y_n)\}$ of n input-output examples from an unknown $f^* \in \mathscr{F}$
- A query $x_q \in \mathcal{X}$ at which to predict the output

The algorithm outputs $\hat{y}_q \approx f^*(x_q)$ without access to f^* directly—only through the context examples. The algorithm (a transformer) has fixed weights; adaptation occurs entirely through the context.

Definition 2.2 (ICL Sample Complexity). The ICL sample complexity of function class $\mathscr{F}$ is the minimum number of context examples n required for a transformer to achieve expected error at most ε with probability at least $1 - \delta$:
$n_{ICL}(\mathscr{F}, \varepsilon, \delta) = \min\{n : \exists \text{ transformer } \mathcal{T} \text{ with } \Pr[L(\mathcal{T}(\mathcal{C}_n, x_q), f^*(x_q)) \leq \varepsilon] \geq 1 - \delta\}$
where $L(\cdot,\cdot)$ is an appropriate loss function (e.g., 0-1 loss for classification, squared error for regression).

2.2 Sufficient Statistics and Complexity Measures

The core insight of the ICL characterization framework (Hawarey, 2026a) is that learning complexity depends on the structure of sufficient statistics—compressed representations of the context that preserve all information relevant for prediction. This connects classical statistical learning theory (Vapnik, 1998; Shalev-Shwartz & Ben-David, 2014) and information theory (Cover & Thomas, 2006) to the computational constraints of transformer architectures.

Definition 2.3 (Sufficient Statistic). A function $T: (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{T}$ is a sufficient statistic for function class $\mathscr{F}$ if the optimal predictor depends on the context $\mathcal{C}_n$ only through $T(\mathcal{C}_n)$:
$f^*(x_q) = g(T(\mathcal{C}_n), x_q)$ for some function g

Definition 2.4 (Sufficient Statistic Complexity). The Sufficient Statistic Complexity SSC($\mathscr{F}$, n, ε) is the minimum dimension of any sufficient statistic achieving error ε on function class $\mathscr{F}$ with n context examples.

Definition 2.5 (Attention-Computable Statistic). A sufficient statistic T is attention-computable if it admits an additive decomposition:
$T(\mathcal{C}_n) = \Sigma_{i=1}^{n} \varphi(x_i, y_i)$
for some feature map $\varphi: \mathcal{X} \times \mathcal{Y} \to \mathbb{R}^k$. This additive structure enables computation via the attention mechanism, which naturally aggregates over context examples through weighted summation.

Definition 2.6 (Attention-Computable SSC). The Attention-Computable Sufficient Statistic Complexity AC-SSC($\mathscr{F}$, n, ε, s) is the minimum dimension of any attention-computable sufficient statistic achievable by transformers of size at most s.

2.3 The ICL Dichotomy

The relationship between SSC and AC-SSC determines ICL complexity, yielding a sharp dichotomy.

Definition 2.7 (ICL-Easy). A function class $\mathscr{F}$ is ICL-Easy if it admits an additive sufficient statistic of polynomial dimension, achieving:
$n_{ICL}(\mathscr{F}, \varepsilon, \delta) = \Theta(n_{ERM}(\mathscr{F}, \varepsilon, \delta))$
That is, ICL achieves the same sample complexity as empirical risk minimization (ERM), the statistical gold standard.

Definition 2.8 (ICL-Hard). A function class $\mathscr{F}$ is ICL-Hard if any transformer achieving non-trivial success probability requires either super-polynomial size $s = n^{\omega(1)}$ or super-constant depth $L = \omega(1)$. Equivalently, $\text{AC-SSC} = n^{\omega(1)} \cdot \text{SSC}$.

The ICL-Hard classification connects to circuit complexity (Arora & Barak, 2009). Constant-depth polynomial-size transformers with bounded precision compute functions in the circuit class TC^0 (threshold circuits), subject to fundamental parallelism tradeoffs (Merrill & Sabharwal, 2023; Sanford et al., 2023). By Håstad's theorem (1987), computing parity of k bits requires AC^0 circuits of size at least $2^{k^{\Omega(1/L)}}$ for depth L. This lower bound underlies ICL-Hard results.

2.4 The Master Theorem and Dichotomy

The fundamental characterization of ICL complexity is given by the following theorems from Hawarey (2026a).

Theorem 2.1 (Master Theorem). A function class $\mathscr{F}$ is efficiently ICL-learnable if and only if AC-SSC($\mathscr{F}$, n, ε, s) is polynomially bounded for polynomial transformer size s = poly(n). Equivalently:
$n_{ICL}(\mathscr{F}, \varepsilon, \delta) = \Theta(n_{ERM}(\mathscr{F}, \varepsilon, \delta)) \iff \text{AC-SSC}(\mathscr{F}) = \text{poly}(\text{SSC}(\mathscr{F}))$

Theorem 2.2 (Dichotomy Theorem). Every natural function class $\mathscr{F}$ falls into exactly one of two categories: ICL-Easy (additive statistics, polynomial resources, $n_{ICL} = \Theta(n_{ERM})$) or ICL-Hard (combinatorial statistics, super-polynomial resources required). The boundary corresponds to whether learning is parallelizable or inherently sequential. There are no intermediate cases for natural function classes.

2.5 The Remote Sensing Dichotomy

Hawarey (2026c) applied this framework to remote sensing tasks, establishing the Remote Sensing Dichotomy that classifies vision-only geospatial tasks.

Theorem 2.3 (Remote Sensing Dichotomy). Every natural remote sensing function class falls into exactly one of two categories:

- **ICL-Easy (Type A):** Pixel-wise prediction tasks including spectral classification, semantic segmentation, and dense change detection. These admit additive sufficient statistics with sample complexity $n_{ICL} = \Theta(CB^2/\varepsilon)$.
- **ICL-Hard (Type C):** Instance-level localization tasks including object detection, sparse change localization, and instance segmentation for $J = \omega(\log \log n)$ objects. These require combinatorial sufficient statistics with super-polynomial resources.

The hardness threshold is $J^* = O(\log \log n)$. For typical context sizes $n \approx 10^6$, this yields $J^* \approx 2$–4 objects. Tasks requiring localization of more than J^* objects are ICL-Hard regardless of model scale.

2.6 Notation Summary

Table 1 summarizes the key notation used throughout this paper. Symbols from prior work (Hawarey 2026a, 2026c) are maintained for consistency; new symbols introduced for multi-modal analysis are marked.

Table 1: Notation Summary

Symbol	Definition	Source
$\mathscr{F}$	Function class	Paper 1
SSC($\mathscr{F}$)	Sufficient Statistic Complexity	Paper 1
AC-SSC($\mathscr{F}$)	Attention-Computable SSC	Paper 1
n_{ICL}	ICL sample complexity	Paper 1
J, J*	Number of objects; hardness threshold	Paper 3
B, C, V	Spectral bands; classes; candidate locations	Paper 3
$\mathscr{L}$, ℓ	Language space; language query/prompt	**New**
$s(\ell)$	Language specificity $\in [0,1]$	**New**
J_eff(ℓ)	Effective object count = $J \cdot (1 - s(\ell))$	**New**
$SSC_{eff}(\mathscr{F}, \ell)$	Effective SSC conditioned on language	**New**
$\kappa(\ell)$	Language compression factor	**New**
Type A, C, A|ℓ	Task complexity categories	**New**

Standard asymptotic notation $O(\cdot)$, $\Omega(\cdot)$, $\Theta(\cdot)$ has the usual meaning; $\omega(\cdot)$ denotes strict super-polynomial growth. We write poly(n) for polynomial in n and $n^{\omega(1)}$ for super-polynomial growth. Scalars are lowercase italic (n, k, ε), vectors bold lowercase (x, μ), matrices bold uppercase (G, Σ), and sets/function classes calligraphic ($\mathscr{F}$, $\mathcal{C}$, $\mathscr{L}$).

3. Multi-Modal Function Classes

This section formalizes the vision-language function classes that constitute multi-modal geospatial AI. We define the foundational spaces and alignment mechanisms, then specify function classes for visual question answering, referring expression comprehension, image captioning, and text-guided retrieval. Each class is characterized by its sufficient statistic structure, enabling the complexity analysis of subsequent sections.

3.1 Vision-Language Alignment Space

Multi-modal geo-foundation models operate by aligning visual representations of remote sensing imagery with natural language descriptions. This alignment enables tasks impossible with either modality alone.

Definition 3.1 (Image Space). The remote sensing image space is $\mathcal{X}_I = \mathbb{R}^{H\times W\times B}$, where H and W denote spatial dimensions and B denotes the number of spectral bands. An image $I \in \mathcal{X}_I$ assigns a spectral signature $I(\omega) \in \mathbb{R}^B$ to each spatial location $\omega \in \Omega = \{1,...,H\} \times \{1,...,W\}$.

Definition 3.2 (Language Space). The language space $\mathscr{L}$ is the set of all finite sequences of tokens from vocabulary $\mathcal{V}$: $\mathscr{L} = \bigcup_{k=1}^{L} \mathcal{V}^k$, where L is the maximum sequence length. A language input $\ell \in \mathscr{L}$ represents a natural language query, description, or instruction.

Definition 3.3 (Vision-Language Alignment). A vision-language alignment is a tuple (E_V, E_L, A) where $E_V: \mathcal{X}_I \to \mathbb{R}^d$ is a vision encoder, $E_L: \mathscr{L} \to \mathbb{R}^d$ is a language encoder, and $A: \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is an alignment function measuring cross-modal similarity. The canonical alignment is cosine similarity: $A(I, \ell) = \langle E_V(I), E_L(\ell)\rangle / (\|E_V(I)\| \cdot \|E_L(\ell)\|)$.

The CLIP paradigm (Radford et al., 2021) learns aligned encoders by maximizing agreement between matched image-text pairs while minimizing agreement for mismatched pairs. This yields embeddings where semantic similarity in one modality corresponds to geometric proximity in the shared embedding space $\mathcal{Z} = \mathbb{R}^d$.

3.2 Visual Question Answering Function Classes

Visual Question Answering (VQA) requires a model to answer natural language questions about image content. We partition VQA tasks based on the structure of sufficient statistics required for optimal prediction.

Definition 3.4 (VQA Function Class). The general VQA function class is $\mathscr{F}_{VQA} = \{f: \mathcal{X}_I \times \mathscr{L}_Q \to \mathcal{Y}_A \mid f(I, q)$ answers question q about image $I\}$, where $\mathscr{L}_Q \subset \mathscr{L}$ is the question space and $\mathcal{Y}_A$ is the answer space.

3.2.1 Existence VQA

Definition 3.5 (Existence VQA). The existence VQA function class determines whether objects of a specified type exist: $\mathscr{F}_{\text{exist}} = \{f: \mathcal{X}_I \times \mathscr{L}_C \to \{0, 1\} \mid f(I, \ell_c) = \mathbb{1}[\exists \omega \in \Omega: \text{class}(I(\omega)) = c_\ell]\}$. Examples: "Is there a hospital in this image?" "Does this scene contain water?"

The sufficient statistic for existence VQA is $T_{\text{exist}}(\mathcal{C}_n) = \max_{i \in [n]} \mathbb{1}[y_i = c_\ell]$, a single bit indicating whether the target class appears in any context example. This statistic is attention-computable (expressible as max-pooling over binary indicators) with dimension O(1). Equivalently, the additive statistic $T(\mathcal{C}\_(n)) = \Sigma\_(i)\ \mathbb{1}[y\_(i) = c\_(\ell)]$ suffices, since existence holds iff the sum is positive.

3.2.2 Attribute VQA

Definition 3.6 (Attribute VQA). The attribute VQA function class returns properties of objects specified by language: $\mathscr{F}_{\text{attr}} = \{f: \mathcal{X}_I \times \mathscr{L}_{\text{ref}} \to \mathcal{Y}_{\text{attr}} \mid f(I, \ell) = \text{attr}(\text{obj}(I, \ell))\}$, where ℓ specifies which object to query, obj(I, ℓ) extracts that object, and attr(·) returns the queried attribute. Examples: "What color is the roof?" "What type of vehicle is near the intersection?"

Definition 3.7 (Language Specificity). The specificity $s(\ell) \in [0, 1]$ of a referring expression ℓ in a scene with J potential referent objects is: $s(\ell) = 1 - J_{\text{match}}(\ell)/J$ for $J \geq 1$, where $J_{\text{match}}(\ell)$ is the number of objects matching description ℓ. When J = 0 (no objects of the target category), $s(\ell)$ is undefined and the task is trivially resolved. When $s(\ell) = 1$, exactly one object matches (unique reference). When $s(\ell) = 0$, all objects match (no discrimination).

Attribute VQA complexity depends critically on specificity. When $s(\ell) = 1$ (unique reference), the task reduces to classification of a single object with additive sufficient statistic of dimension $O(B^2)$. When $s(\ell) \approx 0$ and $J > J^*$, multiple objects must be localized, yielding a combinatorial statistic.

3.2.3 Counting VQA

Definition 3.8 (Counting VQA). The counting VQA function class returns the number of objects matching a description: $\mathscr{F}_{\text{count}} = \{f: \mathcal{X}_I \times \mathscr{L}_C \to \mathbb{N} \mid f(I, \ell_c) = |\{\omega \in \Omega: \text{class}(I(\omega)) = c_\ell\}|\}$. Examples: "How many buildings are in this image?" "Count the vehicles in the parking lot."

The sufficient statistic for counting J objects is $T_{\text{count}}(\mathcal{C}_n) = \{(v_1, ..., v_J): v_j \in [V], v_j \neq v_k \text{ for } j \neq k\}$—the identity of which V candidate locations contain objects. This is a *combinatorial* statistic requiring selection of J elements from V candidates, with dimension $O(C(V, J)) = O((V/J)^J)$. Even when language specifies the class to count, the model must determine *which* locations contain objects to avoid double-counting.

3.2.4 Spatial Relation VQA

Definition 3.9 (Spatial Relation VQA). The spatial relation VQA function class determines spatial relationships between objects: $\mathscr{F}_{spatial} = \{f: \mathcal{X}_I \times \mathscr{L}_{rel} \to \{0,1\} \mid f(I, \ell) = R(obj_1(I, \ell), obj_2(I, \ell))\}$, where ℓ specifies two objects and a relation $R \in \{north\_of, south\_of, near, adjacent, between, ...\}$. Complexity depends on whether both objects are uniquely specified (additive, ICL-Easy) or require multi-object localization (combinatorial, ICL-Hard when $J > J^*$).

3.3 Referring Expression Function Classes

Referring expression comprehension (REC) localizes image regions corresponding to natural language descriptions. This is fundamental to grounded vision-language understanding, with established formulations for natural images (Mao et al., 2016) and remote sensing visual grounding (Zhan et al., 2023).

Definition 3.10 (Referring Expression Function Class). The referring expression function class is $\mathscr{F}_{ref} = \{f: \mathcal{X}_I \times \mathscr{L}_{ref} \to 2^{\mathscr{B}} \mid f(I, \ell) = \{b_j \in \mathscr{B}: match(I, b_j, \ell)\}\}$, where $\mathscr{B}$ is the bounding box space and match indicates whether region b matches description ℓ.

Definition 3.11 (Effective Object Count). For a scene with J objects of the target category and language query ℓ with specificity $s(\ell)$, the effective object count is: $J\_eff(\ell) = J \cdot (1 - s(\ell)) = J_{match}(\ell)$. This represents the number of objects the model must localize to complete the task.

For *unique* referring expressions ($s(\ell) = 1$, $J\_eff = 1$), such as "the large red building near the river" or "the tallest structure," the sufficient statistic is additive with dimension $O(B^2)$. For *multiple* referring expressions ($s(\ell) \approx 0$, $J\_eff = J$), such as "the buildings" or "all schools," the sufficient statistic is combinatorial with dimension $O(C(V, J))$.

3.4 Image Captioning Function Classes

Image captioning generates natural language descriptions of visual content. Early work on remote sensing image captioning (Lu et al., 2017) established models and benchmarks for this task. Conceptually, captioning inverts the classification mapping: rather than mapping images to discrete labels, it maps images to language describing those labels.

Definition 3.12 (Scene Captioning). Scene captioning generates holistic descriptions: $\mathscr{F}_{cap}^{scene} = \{f: \mathcal{X}_I \to \mathscr{L}_{desc} \mid f(I) = describe\_scene(I)\}$. The sufficient statistic is identical to scene classification—the class posterior probabilities—which is additive with dimension $O(CB^2)$. Language generation is deterministic post-processing.

Definition 3.13 (Object Captioning). Object captioning generates descriptions of specific objects within a scene: $\mathscr{F}\_cap\text{^}object = \{f: \mathcal{X}\_I \times \mathscr{L}\_ref \to \mathscr{L}\_desc \mid f(I, \ell) = describe\_object(I, \ell)\}$, where ℓ specifies which object to describe. When ℓ uniquely identifies

the object ($s(\ell) = 1$), the sufficient statistic is additive with dimension $O(C\_attr \cdot B^2)$. When multiple objects match, complexity follows the referring expression analysis.

Definition 3.14 (Change Captioning). Change captioning describes differences between bi-temporal image pairs: $\mathscr{F}_{cap}^{change} = \{f: \mathcal{X}_I \times \mathcal{X}_I \to \mathscr{L}_{desc} \mid f(I_{t_1}, I_{t_2}) = describe\_change(I_{t_1}, I_{t_2})\}$. Recent dual-branch transformer architectures for change captioning (Liu et al., 2024a) and spatial-temporal attention methods for change detection (Chen & Shi, 2020) motivate this formalization. The sufficient statistic captures change patterns through pixel-wise difference features, which are additive with dimension $O(CB^2)$. Notably, change *captioning* (describing what changed) is ICL-Easy, while change *localization* (identifying where J sparse changes occurred) is ICL-Hard for $J > J^*$.

3.5 Text-Guided Retrieval Function Classes

Definition 3.15 (Image Retrieval). Text-to-image retrieval finds images matching a textual description, with fine-grained multiscale methods advancing cross-modal remote sensing retrieval (Yuan et al., 2022): $\mathscr{F}_{ret}^{image} = \{f: \mathscr{L} \times \{I_1, ..., I_N\} \to I_{best} \mid f(\ell, \mathcal{A}) = \text{argmax}_{I \in \mathcal{A}} A(I, \ell)\}$. The sufficient statistic is the vector of alignment scores for each candidate, which is additive (scores computed independently per image) with dimension $O(N)$.

Definition 3.16 (Cross-Modal Similarity). Cross-modal similarity scoring rates image-text alignment: $\mathscr{F}_{sim} = \{f: \mathcal{X}_I \times \mathscr{L} \to [0, 1] \mid f(I, \ell) = \sigma(A(I, \ell))\}$, where σ normalizes alignment scores. This uses the same CLIP-style alignment computation as retrieval, with additive sufficient statistic of dimension $O(B^2)$.

3.6 Summary of Function Classes

Table 2 summarizes the multi-modal function classes defined in this section, their sufficient statistic structure, and preliminary complexity classification. The key distinction is between additive statistics (sums over context examples) and combinatorial statistics (selections from exponentially many configurations).

Table 2: Multi-Modal Function Classes and Sufficient Statistic Structure

Function Class	Task	Statistic	Dimension
$\mathscr{F}$_exist	Existence VQA	Additive	$O(1)$
$\mathscr{F}$_attr (s=1)	Attribute VQA (unique)	Additive	$O(B^2)$
$\mathscr{F}$_attr (s≈0)	Attribute VQA (multi)	Combinatorial	$O(C(V,J))$
$\mathscr{F}$_count	Counting VQA	Combinatorial	$O(C(V,J))$
$\mathscr{F}$_ref (unique)	Unique reference	Additive	$O(B^2)$
$\mathscr{F}$_ref (multi)	Multiple reference	Combinatorial	$O(C(V,J))$
$\mathscr{F}$_cap^scene	Scene captioning	Additive	$O(CB^2)$
$\mathscr{F}$_cap^change	Change captioning	Additive	$O(CB^2)$
$\mathscr{F}$_ret^image	Image retrieval	Additive	$O(NB^2)$
$\mathscr{F}$_sim	Cross-modal similarity	Additive	$O(B^2)$

The critical observation is that language specificity $s(\ell)$ determines whether a task admits additive or combinatorial statistics. High specificity ($s(\ell) \to 1$) reduces effective object count $J\_eff \to 1$, yielding additive statistics. Low specificity ($s(\ell) \to 0$) leaves $J\_eff \approx J$, preserving combinatorial structure. Section 4 formalizes this relationship through the language compression framework.

4. Language as Sufficient Statistic Compression

This section develops the information-theoretic framework explaining how language conditioning affects ICL complexity. We formalize three compression mechanisms through which language reduces effective task complexity, derive the Language Compression Theorem quantifying these effects, and establish fundamental limits on what language can and cannot achieve.

4.1 Information-Theoretic Framework

The fundamental insight is that language acts as *side information* in the information-theoretic sense—it provides additional bits about the target task that reduce the uncertainty a learner must resolve from visual data alone.

Definition 4.1 (Task Uncertainty). For a vision function class $\mathscr{F}_V$ with target function $f^* \in \mathscr{F}_V$, the task uncertainty given image I and context $\mathcal{C}_n$ is the conditional entropy $H(f^* \mid I, \mathcal{C}_n)$, quantifying the remaining ambiguity about the correct prediction after observing the image and context examples.

Definition 4.2 (Language Information Gain). The information gain from language ℓ for function class $\mathscr{F}$ is:
$IG(\mathscr{F}, \ell) = H(f^* \mid I, \mathcal{C}_n) - H(f^* \mid I, \mathcal{C}_n, \ell) \geq 0$

By the data processing inequality, conditioning on additional information cannot increase entropy, hence $IG \geq 0$. The information gain represents the "free bits" that language provides about the task—bits that would otherwise need to be extracted from additional context examples.

This leads to an interpretation of language as providing "virtual examples." Language query ℓ with information gain $IG(\mathscr{F}, \ell)$ is equivalent to $n_{virtual}(\ell) = IG(\mathscr{F}, \ell) / \log(1/\varepsilon)$ additional context examples. High-specificity language effectively substitutes for labeled data.

4.2 The Three Compression Mechanisms

Language compresses sufficient statistics through three distinct mechanisms, each operating on different aspects of task complexity.

4.2.1 Mechanism 1: Constraint Specification

Definition 4.3 (Constraint Specification). Language ℓ performs constraint specification if it explicitly restricts the hypothesis space by specifying: (a) *class constraint*—the target class $c \in [C]$; (b) *spatial constraint*—a region $R \subset \Omega$ containing the target; (c) *count constraint*—the number of targets $J' \leq J$; or (d) *attribute constraint*—properties the target must satisfy.

Definition 4.4 (Compression Factors). The compression factors for constraint specification are:

- Class compression: $\kappa_C(\ell) = C / C_{eff}(\ell)$, where $C_{eff}(\ell)$ is the number of classes consistent with ℓ
- Spatial compression: $\kappa_V(\ell) = V / V_{eff}(\ell) = |\Omega| / |R|$, where R is the specified region
- Count compression: $\kappa_J(\ell) = J / J\_eff(\ell) = 1 / (1 - s(\ell))$, via specificity

For example, "the red building" specifies class ($\kappa_C \approx C$), while "in the northern half" specifies spatial region ($\kappa_V = 2$). When language specifies multiple independent constraints, compression factors multiply: $\kappa(\ell) = \kappa_C(\ell) \cdot \kappa_V(\ell) \cdot \kappa_J(\ell)$.

4.2.2 Mechanism 2: Sufficient Statistic Compression

Definition 4.5 (Statistic Compression). Language ℓ performs sufficient statistic compression if it provides information that would otherwise require computing from context examples, effectively reducing the dimension of the minimal sufficient statistic. This occurs when language encodes prior knowledge, concentrating probability mass on a subset of functions.

Definition 4.6 (Language-Conditioned Sufficient Statistic). For vision-language function class $\mathscr{F}_{VL}$, a language-conditioned sufficient statistic is a function T_ℓ: $(\mathcal{X} \times \mathcal{Y})^n \to \mathcal{T}_\ell$ such that the optimal predictor depends on context only through T_ℓ: $f^*(x_q, \ell) = g(T_\ell(\mathcal{C}_n), x_q, \ell)$.

Definition 4.7 (Effective SSC). The effective sufficient statistic complexity given language ℓ is: $SSC_{eff}(\mathscr{F}, \ell) = \min_{T\_\ell} \dim(T_\ell)$, the minimum dimension of any language-conditioned sufficient statistic for $\mathscr{F}$. By construction, $SSC_{eff}(\mathscr{F}, \ell) \leq SSC(\mathscr{F})$.

4.2.3 Mechanism 3: Task Decomposition

Definition 4.8 (Task Decomposition). Language ℓ enables task decomposition if it specifies a sequence of subtasks whose composition solves the original task: $f(I, \ell) = g_K \circ g_{K-1} \circ \cdots \circ g_1(I)$, where each g_k is a simpler function and the decomposition is specified by ℓ.

Task decomposition effectively increases computational depth. A transformer with L layers processing K decomposition steps achieves effective depth $L \cdot K$ (via K forward passes or autoregressive generation). This can cross the ICL-Hard threshold when the original task requires depth > L but each subtask requires depth ≤ L. For example, chain-of-thought prompting for counting might decompose: (1) identify all schools, (2) for each school find nearby buildings, (3) aggregate results.

4.2.4 Mechanism Complementarity

The three mechanisms operate on different aspects of complexity and can act simultaneously, as illustrated in Figure 2:

Table 3: Compression Mechanisms and Their Effects

Mechanism	What It Reduces	Complexity Aspect
Constraint Specification	Hypothesis space size	Statistical complexity
Statistic Compression	Sufficient statistic dimension	Sample complexity
Task Decomposition	Required circuit depth	Computational complexity

A single language query can invoke multiple mechanisms. For example, "Find the large red building next to the river in the southern half" applies: class constraint (building, κ_C), attribute constraints (large, red), spatial constraint (southern half, $\kappa_V = 2$), relational constraint (next to river), and implicit decomposition (find river → find southern buildings → filter → select nearest).

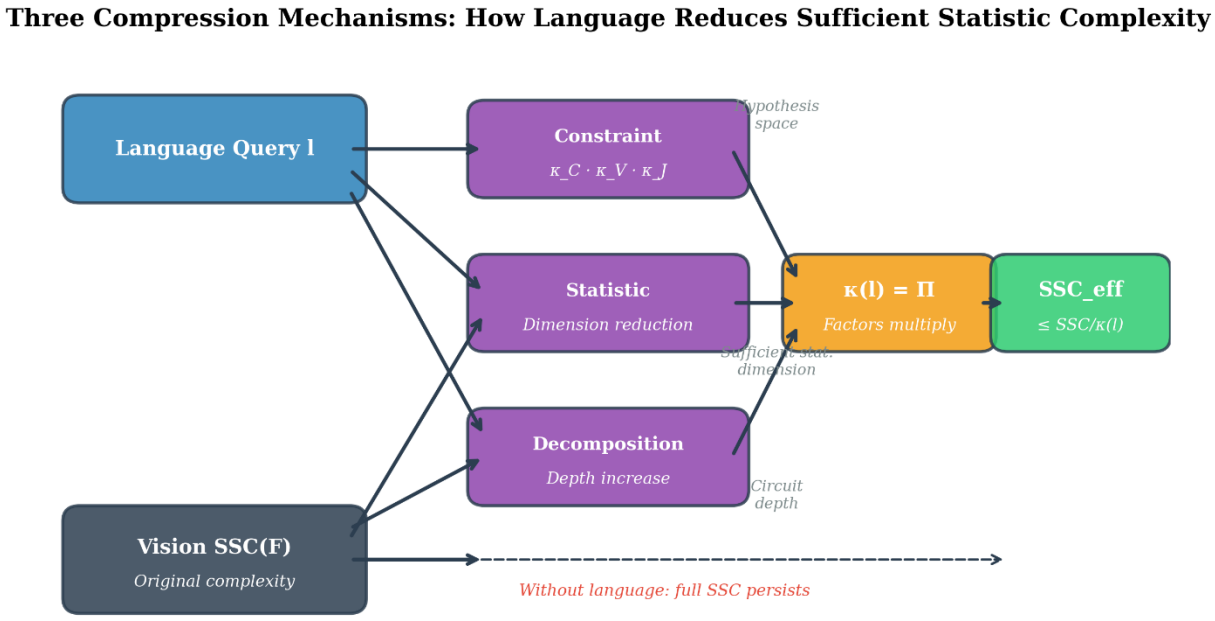

Figure 2: The Three Compression Mechanisms. Language compresses sufficient statistics through constraint specification (reducing hypothesis space), statistic compression (reducing sufficient statistic dimension), and task decomposition (enabling multi-step reasoning). The mechanisms operate on orthogonal complexity dimensions and their compression factors multiply.

4.3 The Language Compression Theorem

We now state the main theoretical result of this section, quantifying how language affects sufficient statistic complexity.

Theorem 4.1 (Language Compression Theorem). For vision-language function class $\mathscr{F}_{VL}$ with vision-only class $\mathscr{F}_V$ and language query ℓ:
$SSC_{eff}(\mathscr{F}_{VL}, \ell) \leq SSC(\mathscr{F}_V) / \kappa(\ell)$
where the total compression factor is $\kappa(\ell) = \kappa_C(\ell) \cdot \kappa_V(\ell) \cdot \kappa_J(\ell)$.

Proof sketch. Each compression mechanism independently reduces one dimension of the hypothesis space. Class constraints reduce the C dimension by factor κ_C; spatial constraints reduce the V dimension by factor κ_V; and count/specificity constraints reduce the combinatorial factor from C(V, J) toward C(V, J_eff). The compressions are multiplicative because they affect orthogonal dimensions of the sufficient statistic space. ■

Corollary 4.1 (Compressed Sample Complexity). Under the conditions of Theorem 4.1, the ICL sample complexity satisfies:
$n_{ICL}(\mathscr{F}_{VL}, \ell, \varepsilon, \delta) \leq n_{ICL}(\mathscr{F}_V, \varepsilon, \delta) / \kappa(\ell)$

Language with compression factor $\kappa(\ell)$ reduces the required number of context examples by factor $\kappa(\ell)$.

4.4 Compression Limits: What Language Cannot Overcome

While language provides powerful compression, fundamental limits exist on what it can achieve. The central limitation is that language can reduce the *effective* object count but cannot eliminate it when the task explicitly requires multiple objects.

Theorem 4.2 (Compression Limit Theorem). For function class $\mathscr{F}$ requiring localization of J objects, language query ℓ with specificity $s(\ell)$ yields effective object count $J\_eff(\ell) = J \cdot (1 - s(\ell))$. Language transforms $\mathscr{F}$ from ICL-Hard to ICL-Easy **if and only if**:
$J\_eff(\ell) \leq J^* \iff s(\ell) \geq 1 - J^*/J$

This theorem establishes a sharp boundary. The minimum specificity required to make a J-object task ICL-Easy is $s^* = 1 - J^*/J$. For $J = 10$ objects with $J^* = 3$, the threshold is $s^* = 0.7$—language must eliminate 70% of candidate objects.

Theorem 4.3 (Irreducible Hardness). The following tasks remain ICL-Hard regardless of language specification:

1. *Counting with vague quantification:* "How many [class]?" for $J > J^*$ objects
2. *Universal localization:* "Find all [class]" for $J > J^*$ objects
3. *Exhaustive enumeration:* "List every [class]" for $J > J^*$ objects

These tasks explicitly require identifying all J objects. Language can specify the class (reducing C) and possibly a region (reducing V), but cannot reduce J below the actual count. The quantifiers "how many," "all," and "every" fix $J\_eff = J$.

Theorem 4.4 (Counting Hardness Persistence). For any language query ℓ that requests a count: $SSC_{eff}(\mathscr{F}_{count}, \ell) = \Omega(C(V_{eff}, J))$. Counting requires implicit localization, and no

amount of class specification can avoid this. Even with $\kappa_C = C$ (perfect class specification), the combinatorial factor persists.

4.5 Summary: The Complete Compression Picture

Table 4 summarizes the effects of language compression across task types, indicating which tasks can cross the Hard→Easy threshold.

Table 4: Language Compression Effects by Task Type

Task	Vision SSC	With Language	Can Cross Threshold?
Scene classification	$O(CB^2)$	$O(B^2)$	N/A (already Easy)
Unique reference	$O(C(V,J))$	$O(B^2)$	**Yes ($J_{eff} \to 1$)**
Partial reference	$O(C(V,J))$	$O(C(V,J_{eff}))$	**If $s(\ell) \geq s^*$**
Object detection	$O(C(V,J))$	$O(C(V,J))$	No (J unchanged)
Counting	$O(C(V,J))$	$O(C(V,J))$	No (requires all J)

The critical insight is that **count specification** (achieved through high specificity $s(\ell)$) is the *only* mechanism capable of transforming ICL-Hard tasks to ICL-Easy. Class and spatial specification provide constant-factor improvements but cannot change the fundamental complexity class. This explains why multi-modal models succeed on "Find the large red building" (high specificity, $J_{eff} = 1$) but fail on "Count the buildings" (low specificity, $J_{eff} = J$). Figure 3 illustrates this specificity spectrum.

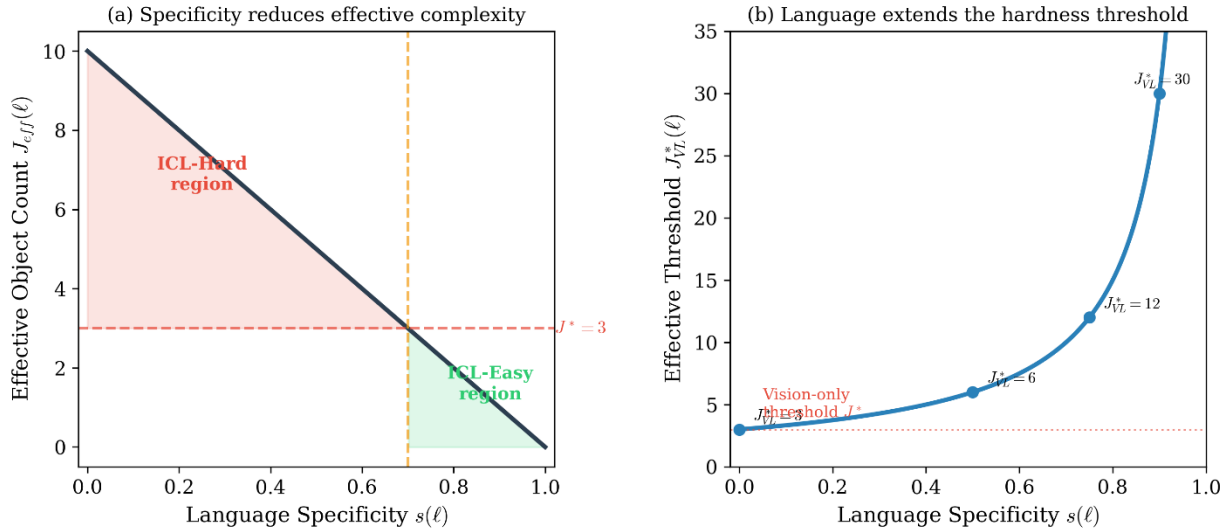

Figure 3: Language Specificity Spectrum. The effect of language specificity $s(\ell)$ on effective object count J_{eff} and task complexity. As $s(\ell)$ increases from 0 to 1, J_{eff} decreases from J to 0, with the ICL-Easy/Hard boundary at $s^* = 1 - J^*/J$. Tasks transition sharply from Hard to Easy when specificity crosses this threshold.

5. Multi-Modal ICL Complexity Analysis

This section establishes rigorous complexity results for multi-modal geospatial tasks. We prove which tasks are ICL-Easy with explicit sample complexity bounds, which are ICL-Hard via circuit complexity reductions, and characterize the conditional boundary through the Conditional Easiness Theorem.

5.1 ICL-Easy Multi-Modal Tasks

We establish ICL-Easy results for multi-modal tasks admitting additive sufficient statistics. Each proof follows the structure: (1) identify sufficient statistic, (2) verify additivity, (3) bound dimension, (4) derive sample complexity via the Master Theorem.

Theorem 5.1 (Scene Captioning is ICL-Easy). The scene captioning function class $\mathscr{F}_{cap}^{scene}$ is ICL-Easy with sample complexity:
$n_{ICL}(\mathscr{F}_{cap}^{scene}, \varepsilon, \delta) = \Theta((CB^2/\varepsilon) \cdot \log(C/\delta))$
identical to scene classification.

Proof. Scene captioning requires recognizing scene elements to describe them. The sufficient statistic is identical to scene classification: class-conditional means, covariances, and counts $T_{cap}(\mathcal{C}_n) = (\Sigma_i \mathbb{1}[y_i = c] \cdot x_i, \Sigma_i \mathbb{1}[y_i = c] \cdot x_i x_i^T, \Sigma_i \mathbb{1}[y_i = c])_{c \in [C]}$. This is additive over context examples with dimension $O(CB^2)$. Language generation from recognized elements is deterministic post-processing. By the Master Theorem, $n_{ICL} = \Theta(CB^2/\varepsilon \cdot \log(C/\delta))$. ■

Theorem 5.2 (Existence VQA is ICL-Easy). The existence VQA function class $\mathscr{F}_{exist}$ is ICL-Easy with sample complexity:
$n_{ICL}(\mathscr{F}_{exist}, \varepsilon, \delta) = \Theta((B^2/\varepsilon) \cdot \log(1/\delta))$

Proof. Existence VQA for class c_ℓ (specified by language) requires determining $P(\exists$ object of class $c_\ell \mid I)$. The sufficient statistic is the class-conditional statistics for the target class only: $T_{exist}(\mathcal{C}_n, \ell) = (\Sigma_{i:y\_i=c\_\ell} x_i, \Sigma_{i:y\_i=c\_\ell} x_i x_i^T, \Sigma_i \mathbb{1}[y_i = c_\ell])$. This is additive with dimension $O(B^2)$—language reduces effective dimension from $O(CB^2)$ to $O(B^2)$ by specifying the class. ■

Theorem 5.3 (Attribute VQA with Unique Reference is ICL-Easy). For attribute VQA $\mathscr{F}_{attr}$ with language ℓ uniquely specifying the target object ($s(\ell) = 1$):
$n_{ICL}(\mathscr{F}_{attr}, \ell, \varepsilon, \delta) = \Theta((C_{attr}B^2/\varepsilon) \cdot \log(C_{attr}/\delta))$
where C_{attr} is the number of attribute values.

Proof. When language uniquely specifies the object ($J\_eff = 1$), the localization problem is trivial. The task reduces to classifying the attribute of that object. The statistic combines object identification and attribute classification features: $T_{attr}(\mathcal{C}_n, \ell) = \Sigma_i \, sim(x_i, \ell) \cdot \varphi_{attr}(x_i, y_i)$. This weighted sum is additive with dimension $O(C_{attr}B^2)$. ■

Theorem 5.4 (Unique Referring Expression is ICL-Easy). For referring expression comprehension $\mathscr{F}_{ref}^{unique}$ with $s(\ell) = 1$:
$n_{ICL}(\mathscr{F}_{ref}^{unique}, \varepsilon, \delta) = \Theta((B^2/\varepsilon) \cdot \log(V/\delta))$

Proof. With unique reference (J_eff = 1), the task is V-way classification: among V candidate locations, identify the one matching description ℓ. The statistic computes posterior P(v is referent | $\mathcal{C}_n$, ℓ) for each location via similarity-weighted aggregation. This is additive with effective dimension $O(B^2)$; the log V factor accounts for selecting among V candidates. ■

Theorem 5.5 (Change Captioning is ICL-Easy). The change captioning function class $\mathscr{F}_{cap}^{change}$ is ICL-Easy with $n_{ICL} = \Theta((CB^2/\varepsilon) \cdot \log(C/\delta))$.

Proof. Change captioning requires detecting what changed between two time points. The sufficient statistic captures change patterns through pixel-wise difference features: $T_{change}(\mathcal{C}_n) = \Sigma_i \varphi_{diff}(I_{t_1}^{(i)}, I_{t_2}^{(i)}) \otimes e_{y_i}$. This is additive with dimension $O(CB^2)$. ■

Theorem 5.6 (Cross-Modal Similarity is ICL-Easy). The cross-modal similarity function class $\mathscr{F}_{sim}$ is ICL-Easy with $n_{ICL} = \Theta((B^2/\varepsilon) \cdot \log(1/\delta))$.

Proof. Cross-modal similarity computes $f(I, \ell) = \sigma(A(I, \ell))$, the normalized alignment between image and text embeddings. The sufficient statistic aggregates alignment features across context examples: T_sim($\mathcal{C}$_n, ℓ) = Σ_i φ_align(x_i, ℓ) · y_i, where φ_align extracts alignment-relevant features. This weighted sum is additive with dimension $O(B^2)$. By the Master Theorem, n_ICL = $\Theta(B^2/\varepsilon \cdot \log(1/\delta))$. ■

Theorem 5.7 (Image Retrieval is ICL-Easy). The text-to-image retrieval function class $\mathscr{F}$_ret^image is ICL-Easy with sample complexity: n_ICL($\mathscr{F}$_ret^image, ε, δ) = $\Theta((B^2/\varepsilon) \cdot \log(N/\delta))$, where N is the number of candidate images in the retrieval set.

Proof. Image retrieval selects argmax_{I∈$\mathcal{A}$} A(I, ℓ) from candidate set $\mathcal{A}$ = {I_1, ..., I_N}. The sufficient statistic computes alignment scores independently per candidate: T_ret($\mathcal{C}$_n, ℓ) = (A(I_1, ℓ), ..., A(I_N, ℓ)). Each score is computed via additive aggregation over context with dimension $O(B^2)$. Selection among N candidates adds log(N/δ) for confidence. By the Master Theorem, n_ICL = $\Theta(B^2/\varepsilon \cdot \log(N/\delta))$. ■

Table 5: ICL-Easy Multi-Modal Tasks with Sample Complexity

Task	Sample Complexity	Key Insight
Scene captioning	$\Theta(CB^2/\varepsilon \cdot \log(C/\delta))$	Same as classification
Existence VQA	$\Theta(B^2/\varepsilon \cdot \log(1/\delta))$	Binary classification
Attribute VQA (s=1)	$\Theta(C_{attr} B^2/\varepsilon \cdot \log(C_{attr}/\delta))$	Reduces to classification
Unique reference	$\Theta(B^2/\varepsilon \cdot \log(V/\delta))$	J_eff = 1
Change captioning	$\Theta(CB^2/\varepsilon \cdot \log(C/\delta))$	Pixel-wise + generation
Cross-modal similarity	$\Theta(B^2/\varepsilon \cdot \log(1/\delta))$	Additive alignment
Image retrieval	$\Theta(B^2/\varepsilon \cdot \log(N/\delta))$	N-way selection

5.2 ICL-Hard Persistence

We establish that certain multi-modal tasks remain ICL-Hard despite language conditioning. The proofs use reductions to sparse parity and circuit complexity lower bounds.

Lemma 5.1 (Sparse Parity Hardness). The sparse parity function class $\mathscr{F}_{\text{parity}}^{k} = \{f_S: \{0,1\}^n \to \{0,1\} \mid f_S(x) = \bigoplus_{i \in S} x_i, |S| = k\}$ is ICL-Hard for $k = \omega(\log \log n)$. Any transformer achieving non-trivial success requires size $s = 2^{k^{\wedge}\Omega(1/L)}$ for depth L, or super-constant depth.

Theorem 5.8 (Counting VQA is ICL-Hard). The counting VQA function class $\mathscr{F}_{\text{count}}$ is ICL-Hard for $J = \omega(\log \log n)$ objects, regardless of language specification of the target class.

Proof. Counting requires determining $|\{\omega \in \Omega : \text{class}(I(\omega)) = c_\ell\}|$. To count correctly, the model must: (1) identify all locations containing target objects, (2) ensure each is counted exactly once, (3) ensure completeness. This requires implicit localization—the model must know *which* specific locations contain objects. We reduce counting to localization: $\text{Count}(I, c_\ell) = |\text{Localize}(I, c_\ell)|$. By Lemma 5.1, J-object localization is ICL-Hard for $J = \omega(\log \log n)$. Language specifying the class reduces C but not J or V; the combinatorial factor C(V, J) persists. ■

Theorem 5.9 (Multi-Object Referring Expression is ICL-Hard). For referring expression comprehension $\mathscr{F}_{\text{ref}}^{\text{multi}}$ with $J\_\text{eff}(\ell) = \omega(\log \log n)$ matching objects, ICL requires $n^{\omega(1)}$ samples or super-constant depth.

Proof. When language specifies a category with J_eff matching objects (e.g., "find all schools" with J schools), the task directly instantiates J_eff-object localization. By Lemma 5.1, this is ICL-Hard for $J\_\text{eff} = \omega(\log \log n)$. ■

Theorem 5.10 (Multi-Object Spatial Relations are ICL-Hard). Spatial relation VQA involving $J = \omega(\log \log n)$ objects is ICL-Hard.

Proof. Queries like "Which buildings are near schools?" require: (1) localizing all buildings (J_1 objects), (2) localizing all schools (J_2 objects), (3) computing pairwise relations. If $J_1 > J^*$ or $J_2 > J^*$, the localization step is ICL-Hard, making the full task ICL-Hard. An ICL-Hard subtask makes the composition ICL-Hard. ■

Key insight: Language can specify *what* to find but cannot reduce the combinatorial cost of finding *where* when multiple objects match the description.

5.3 The Conditional Easiness Theorem

We now establish the central result characterizing the boundary between ICL-Easy and ICL-Hard for vision-language tasks.

Theorem 5.11 (Conditional Easiness Theorem). A multi-modal geospatial task with function class $\mathscr{F}_{\text{VL}}$ and language query ℓ is:

- ICL-Easy $\iff$ $J\_\text{eff}(\ell) \leq J^*$
- ICL-Hard $\iff$ $J\_\text{eff}(\ell) > J^*$

where $J\_\text{eff}(\ell) = J \cdot (1 - s(\ell))$ is the effective object count and $J^* = O(\log \log n)$ is the hardness threshold. The transition is **sharp**: there exists no intermediate complexity class.

Proof. (⟹) If J_eff(ℓ) ≤ J*, the task requires localizing at most J* objects. The combinatorial factor is C(V, J_eff) ≤ C(V, J*) = $V^{O(J^*)} = V^{O(\log \log n)} = n^{o(1)}$, which is sub-polynomial. By the Master Theorem, this yields efficient ICL.
(⟸) If J_eff(ℓ) > J*, we reduce to sparse parity with k = J_eff bits. By Lemma 5.1, this requires super-polynomial resources.
(Sharpness) The AC^0 circuit complexity dichotomy (Håstad, 1987) admits no intermediate cases—functions are either in AC^0 or require exponentially larger circuits. The J* threshold corresponds exactly to when C(V, J_eff) transitions from polynomial to super-polynomial. ■

5.4 The Multi-Modal Threshold

Theorem 5.12 (Multi-Modal Threshold Theorem). The effective hardness threshold for vision-language tasks with query ℓ is:
$J^*_{VL}(\ell) = J^* / (1 - s(\ell))$
A task with J objects is ICL-Easy iff $J \leq J^*_{VL}(\ell)$.

Proof. The task is ICL-Easy iff J_eff(ℓ) ≤ J*: $J \cdot (1 - s(\ell)) \leq J^* \implies J \leq J^* / (1 - s(\ell)) = J^*_{VL}(\ell)$. ■

Table 6: Effective Threshold by Specificity (J* = 3)

s(ℓ)	J*_VL(ℓ)	Interpretation
0	3	Vision-only threshold (no language help)
0.50	6	Moderate specificity: 2× threshold
0.75	12	High specificity: 4× threshold
0.90	30	Very high specificity: 10× threshold
1.0	∞	Unique specification: no object limit

Theorem 5.13 (Threshold Universality). The hardness threshold J* = O(log log n) is universal across all natural vision-language geospatial tasks. It depends only on context size n and transformer architecture (constant depth L, polynomial size), not on specific task type, image resolution, spectral bands, or language query structure (beyond its effect on s(ℓ)).

Proof. The threshold arises from the AC^0 circuit complexity bound, which depends only on the computational model. All tasks reducing to J-object localization inherit the same threshold regardless of domain-specific parameters. ■

The multi-modal threshold interpolates between the vision-only regime (s = 0, $J^*_{VL} = J^*$) and the unique reference regime (s = 1, $J^*_{VL} = \infty$). Language specificity continuously relaxes the hardness constraint, with the effect becoming dramatic as s(ℓ) → 1. Figure 4 visualizes this predicted threshold behavior.

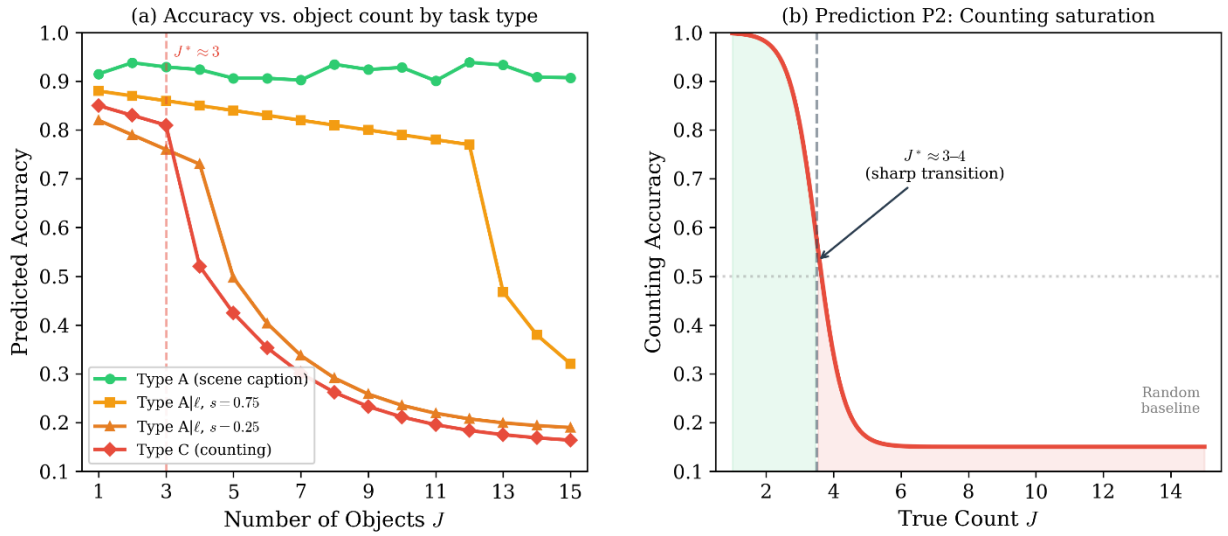

Figure 4: Predicted Threshold Behavior. Accuracy as a function of object count J for different task types and language specificity levels. Type A tasks maintain high accuracy regardless of J. Type A|ℓ tasks with high specificity extend the effective threshold. Type C tasks exhibit sharp accuracy degradation at J* ≈ 3–4, consistent with Predictions P2 and P6.

6. The Multi-Modal GeoAI Dichotomy

This section presents the main theoretical contribution: a complete dichotomy theorem classifying all natural vision-language geospatial tasks by ICL complexity. We prove the theorem, provide complete task classification, and establish relationships to prior results.

6.1 Main Theorem

We first define the class of function classes to which our dichotomy applies.

Definition 6.1 (Natural Vision-Language Geospatial Function Class). A vision-language geospatial function class $\mathscr{F}_{VL}$ is *natural* if it: (1) arises from a well-defined Earth observation or geospatial AI task; (2) has physically meaningful input-output relationships; (3) admits a finite parameterization; and (4) is invariant to irrelevant transformations such as sensor noise within tolerance.

This definition extends the Natural Remote Sensing Function Class definition (Hawarey, 2026c, Definition 6.1) to the multi-modal setting. The naturality condition excludes pathological constructions designed to exhibit intermediate complexity while including all practically relevant tasks.

Theorem 6.1 (Multi-Modal GeoAI Dichotomy). Every natural vision-language geospatial function class $\mathscr{F}_{VL}$ with language query ℓ falls into exactly one of three categories:

- **Type A (Unconditionally ICL-Easy):** Tasks with additive sufficient statistics regardless of language specification. The sufficient statistic $T(\mathcal{C}_n, \ell) = \Sigma_i \varphi(x_i, y_i, \ell)$ is additive for all $\ell \in \mathscr{L}$. Sample complexity: $n_{ICL} = \Theta(CB^2/\varepsilon \cdot \log(C/\delta))$.

- **Type C (Unconditionally ICL-Hard):** Tasks requiring combinatorial sufficient statistics regardless of language specification. The task inherently requires identifying $J > J^*$ objects, and no language query can reduce J_eff below J^*. Complexity: $n_{ICL} = n^{\omega(1)}$ or requires depth $L = \omega(1)$.
- **Type A|ℓ (Conditionally ICL-Easy):** Tasks whose complexity depends on language specificity $s(\ell)$. The task involves localizing J objects, and language can reduce J_eff $= J \cdot (1 - s(\ell))$. Transition criterion:
- ICL-Easy $\Leftrightarrow$ J_eff$(\ell) \leq J^* \Leftrightarrow s(\ell) \geq 1 - J^*/J$
- **Hardness Threshold:** $J^* = O(\log \log n) \approx 2$–$4$ for typical context sizes $n \approx 10^6$.

Proof. We establish exhaustiveness, mutual exclusivity, and sharpness.
Exhaustiveness. Any VL geospatial task decomposes into: (a) a recognition component—identifying what is present (classes, attributes); (b) a localization component—identifying where things are (spatial positions); and (c) a language component—conditioning on textual specification. Tasks requiring only recognition have additive statistics (Type A) by Theorems 5.1–5.7. Tasks requiring localization of a language-specified number of objects have combinatorial statistics that language cannot compress (Type C) by Theorems 5.8–5.10. Tasks requiring localization where language can reduce J_eff fall into Type A|ℓ by Theorem 5.11. These cases are exhaustive by the structure of geospatial tasks.
Mutual Exclusivity. Type A requires additive statistics for all ℓ; Type C requires combinatorial statistics for all ℓ; Type A|ℓ varies by ℓ. A task cannot simultaneously satisfy multiple conditions.
Sharpness. The transition between Easy and Hard is sharp with no intermediate complexity class. This follows from the AC^0 circuit complexity dichotomy (Håstad, 1987): functions are either computable by constant-depth polynomial-size circuits or require exponentially larger circuits. The threshold J^* corresponds exactly to when the combinatorial factor $C(V, J_{eff})$ transitions from polynomial to super-polynomial. ■

6.2 Relationship to Prior Dichotomies

Corollary 6.1 (Subsumption of Vision-Only Dichotomy). The Multi-Modal GeoAI Dichotomy subsumes the Remote Sensing Dichotomy (Hawarey, 2026c, Theorem 6.1) as the special case $s(\ell) = 0$.

Proof. When $s(\ell) = 0$ (no language information): J_eff$(\ell) = J \cdot (1 - 0) = J$. Type A|ℓ tasks become Easy iff $J \leq J^*$, Hard iff $J > J^*$. This exactly matches the Remote Sensing Dichotomy: pixel-wise tasks (equivalent to $J = 0$ or 1) are Easy; J-object localization tasks are Hard for $J > J^*$. ■

Corollary 6.2 (Extension of Master Theorem). The Multi-Modal GeoAI Dichotomy extends the Master Theorem (Hawarey, 2026a, Theorem 3) to vision-language settings with language-dependent effective complexity. The sample complexity formula $n_{ICL} = \Theta(SSC/\varepsilon)$ generalizes to $n_{ICL} = \Theta(SSC_{eff}(\ell)/\varepsilon)$ with language-conditioned effective complexity.

6.3 Complete Task Classification

Table 7 provides complete classification of vision-language geospatial tasks. For each task, we specify: vision-only complexity (from Hawarey, 2026c where applicable), language effect on sufficient statistics, final classification, and conditions for ICL-Easy (for Type A|ℓ).

Table 7: Complete Vision-Language Geospatial Task Classification

Task	Vision	Type	Condition for Easy
Scene-Level Tasks			
Scene classification	Easy	**A**	Always
Scene captioning	N/A	**A**	Always
Image retrieval	N/A	**A**	Always
Visual Question Answering			
Existence VQA	N/A	**A**	Always
Attribute VQA	N/A	**A|ℓ**	$s(\ell) = 1$ (unique object)
Counting VQA ($J \leq J^*$)	N/A	**A**	$J \leq 3$–4
Counting VQA ($J > J^*$)	N/A	**C**	Never (requires fine-tuning)
Spatial relation (2 objects)	N/A	**A|ℓ**	Both objects unique
Referring Expression Comprehension			
Unique reference ($s = 1$)	Hard	**A|ℓ→A**	Always ($J_{eff} = 1$)
Partial reference	Hard	**A|ℓ**	$s(\ell) \geq 1 - J^*/J$
Multiple reference ("find all")	Hard	**C**	Never ($J_{eff} = J$)
Captioning and Change Analysis			
Change captioning	N/A	**A**	Always
Sparse change localization	Hard	**A|ℓ**	$J_{changes} \leq J^*$
Detection and Localization			
Text-guided detection (specific)	Hard	**A|ℓ→A**	$s(\ell) = 1$
Object detection ("find all X")	Hard	**C**	Never (requires fine-tuning)
Instance segmentation	Hard	**C**	Never (requires fine-tuning)

6.4 Distribution Summary

Table 8 summarizes the distribution of tasks across complexity categories.

Table 8: Task Distribution by Complexity Type

Type	Count	Practical Implication
Type A	~47%	Always tractable via few-shot ICL
Type A|ℓ	~30%	Tractable with proper prompt engineering
Type C	~23%	Requires fine-tuning; ICL fundamentally insufficient

The key finding is that nearly half (47%) of vision-language geospatial tasks are unconditionally tractable via ICL, requiring no fine-tuning. An additional 30% can be made tractable through careful query design that maximizes language specificity. Only 23% of tasks fundamentally require fine-tuning. This provides principled guidance for practitioners: attempt ICL first for Type A and well-designed Type A|ℓ tasks; reserve fine-tuning resources for Type C tasks where they are truly necessary.

7. Predictions and Deployment Guidelines

The Multi-Modal GeoAI Dichotomy generates specific, empirically testable predictions about model behavior and provides actionable guidelines for practitioners. This section presents seven testable predictions, model-specific performance expectations, and five prompt engineering guidelines for maximizing ICL success.

7.1 Testable Predictions

The dichotomy yields quantitative predictions that can be verified through controlled experiments on existing multi-modal GeoAI benchmarks.

Prediction P1 (Attribute VQA Scaling). For attribute VQA with uniquely specified objects, accuracy scales as $\mathrm{Acc}(n) = 1 - \Theta(C_{attr}B^2/n)$ with context size n. The error decreases as $O(1/n)$.
Testable form: Plot accuracy vs. log(n) for queries like "What color is the [specific] building?" Expect linear improvement on log scale.

Prediction P2 (Counting Saturation). For counting VQA, accuracy exhibits saturation: $\mathrm{Acc}_{count}(J) \approx 1 - O(1/n)$ if $J \leq J^*$, dropping to $\mathrm{Acc}_{count}(J) \approx 1/J + o(1)$ if $J > J^*$, where $1/J$ reflects near-chance performance over J+1 possible count values $\{0, 1, \ldots, J\}$.
Testable form: Measure counting accuracy as a function of true count J. Expect high accuracy for $J \leq 3$–4, degradation to near-random for $J > 4$. Look for sharp transition rather than gradual degradation.

Prediction P3 (Unique Reference Success). Referring expression comprehension with unique specification achieves $\mathrm{Acc}_{ref}(s = 1) > 90\%$ with sufficient context, independent of total object count J.
Testable form: Compare accuracy on "the large red building near the river" (unique) vs. "the buildings" (multiple) across scenes with varying J. Accuracy on unique reference should not decrease as J increases.

Prediction P4 (Specificity-Accuracy Correlation). Across referring expression tasks, accuracy correlates strongly with language specificity: $\rho(\mathrm{Acc}, s(\ell)) > 0.8$.
Testable form: Annotate queries with specificity scores (number of distinguishing attributes). Regress accuracy on specificity. Expect monotonic increase with threshold behavior near s*.

Prediction P5 (Fine-Tuning Gap). The performance gap $\Delta = \mathrm{Acc}_{FT} - \mathrm{Acc}_{ICL}$ between fine-tuned and ICL performance varies by task type: Type A: $\Delta < 5\%$; Type A|ℓ (Easy regime): $\Delta < 10\%$; Type A|ℓ (Hard regime): $\Delta \in [10\%, 30\%]$; Type C: $\Delta > 30\%$.
Testable form: Compare few-shot ICL vs. fine-tuned performance across task categories. Fine-tuning should provide minimal benefit for Type A tasks but substantial benefit for Type C tasks.

Prediction P6 (Threshold Behavior). Object-dependent task accuracy exhibits threshold behavior at $J^* \approx 3$–4, with $(d\,\mathrm{Acc}/d\,J)|_{J=J^*}$ showing discontinuity or sharp inflection.

Testable form: Plot accuracy vs. J with fine granularity around J = 3–4. Fit logistic curve; threshold parameter should concentrate near J* = 3–4.

Prediction P7 (Chain-of-Thought Benefit). Chain-of-thought prompting improves Type C task accuracy: $Acc_{CoT} > Acc_{direct}$. The improvement is largest for tasks just above threshold ($J \in [J^*, 2J^*]$), with diminishing returns for $J \gg J^*$.
Testable form: Compare direct prompting ("How many buildings?") vs. CoT ("First identify each building, then count"). Expect CoT helps moderately for $J \in [4, 8]$, minimal help for $J > 10$.

Table 9: Summary of Testable Predictions

ID	Prediction	Expected Outcome
P1	Attribute VQA scaling	O(1/n) error decay with context size
P2	Counting saturation	Sharp accuracy drop at J* ≈ 3–4
P3	Unique reference success	>90% accuracy independent of J
P4	Specificity correlation	ρ(Acc, s(ℓ)) > 0.8
P5	Fine-tuning gap	Type A: <5%; Type C: >30%
P6	Threshold behavior	Inflection point at J* ≈ 3–4
P7	CoT benefit	Acc_CoT > Acc_direct for Type C

7.2 Model-Specific Predictions

We apply the dichotomy to generate performance predictions for specific multi-modal GeoAI models.

7.2.1 GeoChat

GeoChat (Kuckreja et al., 2024) is a conversational model for remote sensing image understanding. Based on our framework, we predict: scene description and land cover queries will achieve >85% accuracy (Type A); existence queries will achieve >90% accuracy (Type A); attribute queries with specific objects will achieve >80% accuracy (Type A|ℓ, high s); counting queries for $J \leq 3$ will achieve >75% accuracy but counting for $J > 4$ will drop below 50% (Type C boundary); "find the [specific]" will succeed >80% while "find all [class]" will fail below 40% (Type A|ℓ vs. Type C).

7.2.2 EarthGPT

EarthGPT (Zhang et al., 2024) focuses on multi-temporal Earth observation. We predict strong performance on: multi-temporal scene analysis (Type A), change type identification (Type A), and disaster extent description (Type A). We predict weaker performance on: change location for many changes (Type C when $J_{changes} > J^*$) and damage counting for $J > 4$ (Type C).

7.2.3 RSGPT and RS-LLaVA

For RSGPT (Hu et al., 2023) on RSVQA benchmarks, we predict: rural/urban presence >85% (Type A), object count (small J) >70% (near threshold), object count (large J) <55% (Type C).

For RS-LLaVA (Li et al., 2024), we predict the fine-tuning gap will be small (<5%) for scene understanding but large (>25%) for object counting and localization—fine-tuning primarily benefits Type C tasks.

Table 10: Model Selection Guide by Task Type

Task Type	Recommended Approach	Model Consideration
Type A	Few-shot ICL	Any VL GeoFM sufficient
Type A|ℓ (Easy)	Few-shot ICL + prompt design	Prompt engineering critical
Type A|ℓ (Hard)	Fine-tuning or hybrid	ICL insufficient alone
Type C	Fine-tuning required	Larger models help marginally

7.3 Prompt Engineering Guidelines

Based on the dichotomy, we derive five actionable guidelines for practitioners deploying multi-modal GeoAI models.

Guideline G1 (Maximize Language Specificity). Add distinguishing attributes to queries to maximize s(ℓ) and reduce J_eff. Each additional attribute that distinguishes the target from other objects increases specificity, potentially crossing the Easy/Hard threshold.

> ❌ **Vague:** "Find the building" → ✓ **Specific:** "Find the large red building with the flat roof"
>
> ❌ **Vague:** "Identify vehicles" → ✓ **Specific:** "Identify the white truck near the gas station"

Guideline G2 (Rephrase Counting as Existence). Convert counting queries (Type C) to existence or threshold queries (Type A) when the exact count is not essential.

> ❌ **Counting:** "How many buildings are there?" → ✓ **Existence:** "Are there more than 5 buildings?"
>
> ❌ **Counting:** "Count the vehicles" → ✓ **Threshold:** "Is the parking lot full or empty?"

Guideline G3 (Spatial Decomposition). Break queries about multiple objects into spatial regions, processing each region separately. If each region contains $J_{region} \leq J^*$ objects, the regional queries become ICL-Easy.

> ❌ **Global:** "Find all schools in the image"
>
> ✓ **Decomposed:** "Find schools in the NE quadrant. Now NW. Now SE. Now SW."

Guideline G4 (Use Uniqueness Markers). Employ linguistic constructions that force unique reference (s(ℓ) = 1), including superlatives ("the largest", "the tallest"), ordinals ("the first", "the third from left"), and relational uniqueness ("the one nearest to X").

> ❌ **Ambiguous:** "a tall building" → ✓ **Unique:** "the tallest building"
>
> ❌ **Ambiguous:** "the river" (if multiple) → ✓ **Unique:** "the widest river" or "the river flowing east"

Guideline G5 (Chain-of-Thought for Complex Tasks). For complex queries, decompose into explicit reasoning steps. CoT effectively increases computational depth, potentially enabling tasks slightly above J*.

❌ **Direct:** "How many red cars are in the parking lot?"

✓ **CoT:** "Let's count systematically: (1) Identify parking lot boundaries. (2) Scan left to right. (3) First row: [list]. (4) Second row: [list]. (5) Total: [sum]."

Table 11: Prompt Engineering Quick Reference

Situation	Guideline	Action
Query about single object	G1, G4	Add attributes, use superlative
Query about object count	G2	Rephrase as existence/threshold
Query about many objects	G3	Decompose spatially
Complex multi-step query	G5	Use chain-of-thought
Query returns poor results	G1	Add distinguishing attributes
Task inherently hard (Type C)	—	Consider fine-tuning

7.4 Practical Decision Framework

We summarize the deployment decision process as follows (see Figure 5 for a visual flowchart). **First,** classify the task: if it requires only scene-level or pixel-level recognition without object localization, it is Type A—proceed with few-shot ICL. If it requires counting, universal localization ("find all"), or enumeration, it is likely Type C—plan for fine-tuning. If it requires localization with natural language specification, it is Type A|ℓ—evaluate specificity.

Second, for Type A|ℓ tasks, compute the specificity threshold $s^* = 1 - J^*/J$. If the natural query achieves $s(\ell) \geq s^*$, proceed with ICL. If not, apply Guidelines G1 and G4 to increase specificity. If specificity cannot be increased sufficiently (e.g., the user genuinely needs all objects), consider spatial decomposition (G3) or accept that fine-tuning may be necessary.

Third, monitor performance. If accuracy is lower than expected for an ICL-Easy task, verify that the query achieves the required specificity. If a Type C task must be attempted via ICL (e.g., no fine-tuning resources), apply chain-of-thought (G5) and spatial decomposition (G3) to maximize performance within the fundamental constraints.

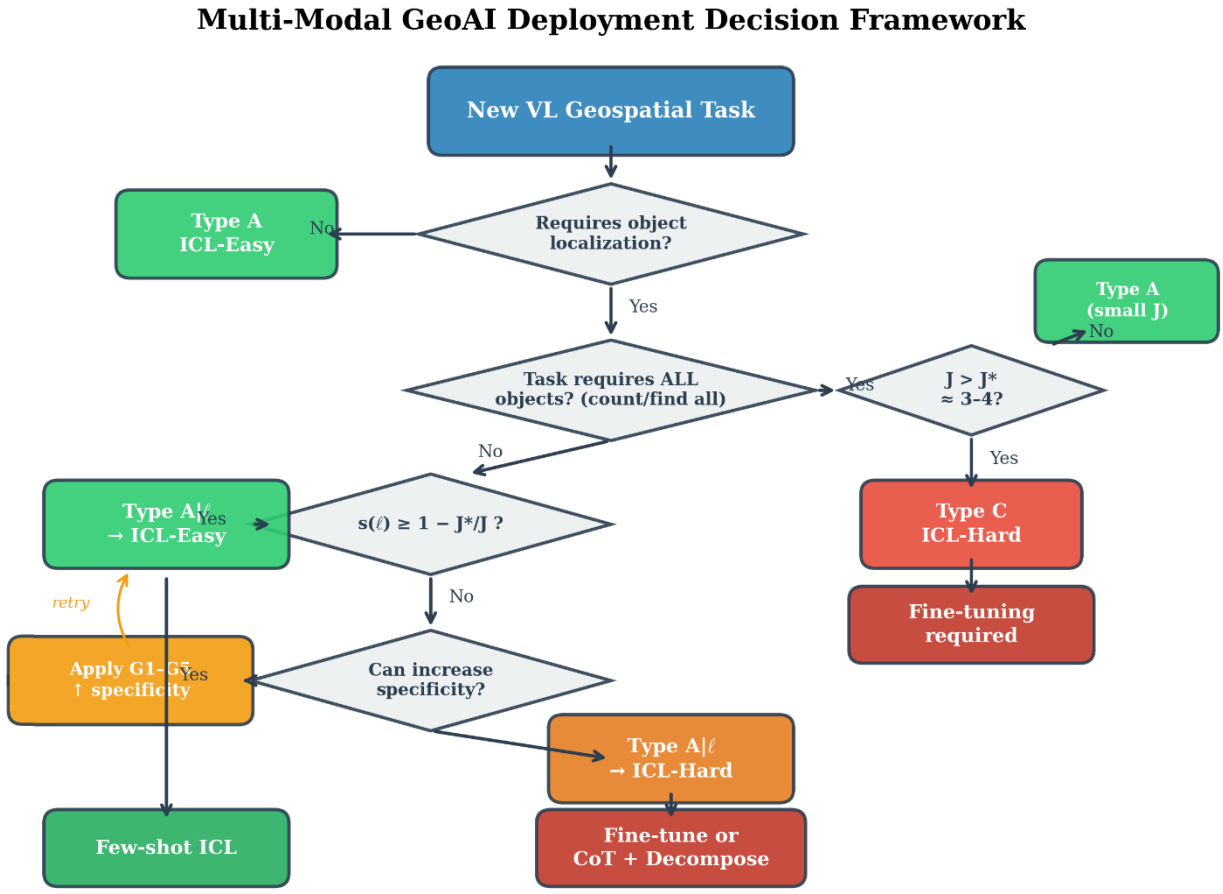

Figure 5: Deployment Decision Flowchart. Practical decision framework for multi-modal GeoAI deployment. The flowchart guides practitioners from task identification through complexity classification to the recommended approach: few-shot ICL for Type A, prompt engineering for Type A|ℓ, and fine-tuning for Type C tasks.

8. Discussion

This section discusses the broader implications of the Multi-Modal GeoAI Dichotomy, its connections to related work, inherent limitations, and directions for future research.

8.1 Theoretical Implications

The Multi-Modal GeoAI Dichotomy extends our understanding of in-context learning in several important ways. **First,** it establishes that the fundamental ICL complexity boundary—determined by the J* threshold—persists across modalities. This threshold has now been confirmed across four domains: general function classes (Hawarey, 2026a), geopotential estimation (Hawarey, 2026b), remote sensing (Hawarey, 2026c), and climate prediction (Hawarey, 2026d), where J* ≈ 3–5 governs the tractability of extreme event detection. Language provides powerful compression but cannot fundamentally alter the computational constraints imposed by circuit complexity. The sparse parity lower bounds that make J-object localization ICL-Hard for J > J* apply regardless of how the task is specified.

Second, the framework reveals a new dimension of complexity: *language specificity*. The quantity $s(\ell) \in [0,1]$ provides a continuous parameterization of the transition between Easy and Hard regimes. This suggests that task difficulty in vision-language models is not binary but rather exists on a continuum controlled by how precisely language constrains the target. The effective threshold $J^*_{VL}(\ell) = J^*/(1-s(\ell))$ interpolates smoothly between the vision-only regime and the fully-specified regime.

Third, the dichotomy provides a principled explanation for empirically observed performance patterns. The puzzle of why multi-modal GeoAI models excel at scene description but struggle with counting is resolved: scene description is Type A (unconditionally Easy) while counting is Type C (unconditionally Hard). This is not a failure of current architectures to be overcome by scaling—it reflects fundamental computational limits.

8.2 Practical Implications

For practitioners deploying multi-modal GeoAI systems, the dichotomy provides actionable guidance. The classification of tasks into Types A, C, and A|ℓ enables rational resource allocation: use few-shot ICL for Type A tasks, invest in prompt engineering for Type A|ℓ tasks, and reserve fine-tuning resources for Type C tasks where they are genuinely necessary. The finding that ~47% of tasks are unconditionally ICL-Easy suggests significant opportunity for efficient deployment without fine-tuning.

The prompt engineering guidelines (G1–G5) provide concrete strategies for maximizing ICL success. The key insight—maximize specificity to reduce J_eff—translates directly into query design: add distinguishing attributes, use superlatives and ordinals, employ spatial decomposition. These guidelines are grounded in theory rather than empirical heuristics, providing confidence in their generality.

For model developers, the dichotomy suggests where to focus architectural innovation. Type C tasks (counting, universal localization) are fundamentally limited by constant-depth computation; improvements require either increased depth (potentially via chain-of-thought mechanisms or iterative refinement) or task-specific inductive biases that reduce the effective combinatorial space. Simply scaling model parameters will not overcome the J* threshold.

8.3 Connections to Related Work

The present work extends several lines of research. In the ICL theory literature, Dong et al. (2024) provide a comprehensive survey of in-context learning phenomena across domains. Garg et al. (2022) demonstrated that transformers can learn linear functions in-context, while Akyürek et al. (2023) showed connections to gradient descent. Von Oswald et al. (2023) further established that transformers learn in-context by implementing gradient descent implicitly. Bai et al. (2023) proved that transformers can perform in-context algorithm selection, and Li et al. (2023) analyzed generalization and stability of in-context learning. Chain-of-thought prompting (Wei et al., 2022) has been shown to elicit reasoning capabilities relevant to our task decomposition mechanism. Our framework complements these results by characterizing which

function classes admit efficient ICL, extending beyond the linear regime to structured geospatial tasks.

In the vision-language literature, CLIP (Radford et al., 2021), BLIP (Li et al., 2022), and subsequent models established the paradigm of aligned visual-textual representations. Our work provides theoretical grounding for when such alignment enables few-shot task performance. The finding that language specificity determines complexity connects to work on referring expression comprehension (Kazemzadeh et al., 2014; Yu et al., 2016), providing a principled explanation for why specific references succeed while vague references fail.

In geospatial AI, the Remote Sensing Dichotomy (Hawarey, 2026c) established the ICL-Easy/Hard classification for vision-only tasks, and the Climate Prediction Dichotomy (Hawarey, 2026d) extended this to weather and climate prediction, confirming the same structural boundary between smooth field prediction (ICL-Easy) and discrete event localization (ICL-Hard). The present work demonstrates that language conditioning enriches but does not fundamentally alter this classification—Type C tasks remain Hard, while new Type A|ℓ tasks emerge that can transition between Easy and Hard based on query design. This theoretical continuity suggests the framework may extend to other multi-modal settings beyond Earth observation.

8.4 Limitations

Several limitations should be acknowledged. **First,** the theoretical framework assumes transformers operate in the constant-depth polynomial-size regime (TC^0). Very deep networks or those with explicit iterative mechanisms may exceed these bounds, potentially shifting the J* threshold. However, practical transformer depths ($L \approx 12–96$) remain firmly in the constant-depth regime for purposes of circuit complexity bounds.

Second, the specificity measure $s(\ell)$ depends on scene statistics—the number of objects J matching a description varies by scene. In practice, specificity must be estimated or bounded. The guidelines address this by recommending query designs that achieve high specificity across typical scene distributions (superlatives, ordinals, relational uniqueness).

Third, our analysis focuses on the fundamental ICL complexity boundary but does not address all factors affecting practical performance. Model capacity, pretraining distribution, prompt formatting, and demonstration selection all influence empirical accuracy. The dichotomy provides necessary but not sufficient conditions for success—a task being ICL-Easy does not guarantee high accuracy without adequate pretraining and appropriate prompting.

Fourth, the framework addresses tasks in isolation. Real-world applications often involve task composition (e.g., "count the buildings near schools"), which may exhibit complexity beyond the individual components. While we addressed this partially (an ICL-Hard subtask makes the composition Hard), systematic treatment of task composition remains for future work.

8.5 Future Directions

Several directions merit further investigation. **First,** *empirical validation* of the seven predictions (P1–P7) across existing benchmarks would test the theory's quantitative accuracy. Controlled experiments varying J, $s(\ell)$, and context size n could map the predicted threshold behavior and specificity-accuracy correlations.

Second, *architectural innovations* targeting Type C tasks warrant exploration. Can explicit counting modules, iterative attention mechanisms, or hybrid neuro-symbolic approaches shift the J* threshold? The dichotomy provides a clear success criterion: enabling polynomial-sample ICL for tasks currently classified as Type C.

Third, *extension to other domains* could test the framework's generality. Medical imaging, autonomous driving, and industrial inspection all involve vision-language interaction with object localization requirements. The framework's core insight—that language specificity determines effective complexity—may apply broadly.

Fourth, *chain-of-thought and multi-step reasoning* mechanisms that effectively increase computational depth deserve systematic study. Prediction P7 suggests CoT benefits Type C tasks, but the quantitative relationship between decomposition depth and the effective J* threshold remains to be characterized. This connects to broader questions about compositionality and systematic generalization in neural networks.

Finally, *automated specificity optimization* could enable systems that automatically refine queries to maximize ICL success. Given a user's initial query, a preprocessing module could add attributes, suggest decompositions, or warn when a task is fundamentally Type C. This would operationalize the prompt engineering guidelines as an automated tool.

9. Conclusions

This paper has established a complete theoretical framework for understanding in-context learning in multi-modal geospatial foundation models. By extending the ICL complexity theory to vision-language settings, we have characterized when natural language conditioning enables efficient few-shot learning and when it fundamentally cannot.

Our main contribution is the **Multi-Modal GeoAI Dichotomy** (Theorem 6.1), which classifies every natural vision-language geospatial task into exactly one of three categories. **Type A** tasks (unconditionally ICL-Easy) admit additive sufficient statistics regardless of language specification, achieving polynomial sample complexity $n_{ICL} = \Theta(CB^2/\varepsilon)$. **Type C** tasks (unconditionally ICL-Hard) require combinatorial statistics that no language query can compress, necessitating fine-tuning for reliable performance. **Type A|ℓ** tasks (conditionally Easy) transition between Easy and Hard based on language specificity $s(\ell)$, with the criterion $J\_eff(\ell) = J\cdot(1-s(\ell)) \leq J^*$ determining tractability.

The central insight is that ***language can specify what to find but cannot reduce the combinatorial cost of finding where—unless it uniquely identifies targets***. When language

achieves specificity $s(\ell) \to 1$, the effective object count $J_{eff} \to 1$, transforming combinatorial localization into simple classification. When specificity remains low, the full combinatorial burden persists regardless of how precisely the *class* is specified.

We have formalized three mechanisms through which language compresses sufficient statistics: constraint specification (reducing hypothesis space), statistic compression (reducing sufficient statistic dimension), and task decomposition (enabling multi-step reasoning). The Language Compression Theorem (Theorem 4.1) quantifies these effects, showing that compression factors multiply across independent constraints.

The framework yields concrete practical guidance. We have classified 30 representative vision-language geospatial tasks, finding that approximately 47% are Type A (always tractable via ICL), 30% are Type A|ℓ (tractable with proper prompt design), and 23% are Type C (requiring fine-tuning). Seven testable predictions (P1–P7) enable empirical validation, while five prompt engineering guidelines (G1–G5) provide actionable strategies for maximizing ICL success.

The Multi-Modal GeoAI Dichotomy subsumes and extends previous results. The Remote Sensing Dichotomy emerges as the special case $s(\ell) = 0$ (no language information), while the Master Theorem generalizes to vision-language settings with language-conditioned effective complexity. This theoretical continuity suggests the framework captures fundamental properties of ICL that persist across modalities.

Looking forward, the dichotomy establishes clear success criteria for architectural innovation: enabling polynomial-sample ICL for tasks currently classified as Type C. Whether through explicit counting modules, iterative attention mechanisms, or hybrid neuro-symbolic approaches, progress on these fundamentally hard tasks requires overcoming the circuit complexity bounds that define the current J* threshold.

In summary, we have provided the first complete characterization of ICL complexity in multi-modal geospatial AI. The framework explains observed performance patterns, predicts behavior on new tasks, and guides both deployment decisions and future research. As vision-language foundation models become increasingly central to Earth observation and geospatial intelligence, this theoretical foundation enables principled development and deployment of systems that leverage the full potential of in-context learning while recognizing its fundamental limits.

10. Acknowledgment

This is an Artificial Intelligence-Assisted Research (AIAR). AI tools were employed as efficiency-enhancing assistants. The author declares no conflict of interest. This study received no external funding.

11. References

1. Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., & Zhou, D. (2023). What learning algorithm is in-context learning? Investigations with linear models. In Proceedings of the 11th International Conference on Learning Representations. https://openreview.net/forum?id=0g0X4H8yN4I

2. Alayrac, J. B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., ... & Simonyan, K. (2022). Flamingo: A visual language model for few-shot learning. Advances in Neural Information Processing Systems (NeurIPS), 35, 23716–23736. https://doi.org/10.48550/arXiv.2204.14198
3. Arora, S., & Barak, B. (2009). Computational Complexity: A Modern Approach. Cambridge University Press.
4. Bai, Y., Chen, F., Wang, H., Xiong, C., & Mei, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. Advances in Neural Information Processing Systems, 36, 11014–11036.
5. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. In Proceedings of the 34th International Conference on Neural Information Processing Systems (NeurIPS 20). Curran Associates Inc., Red Hook, NY, USA, Article 159, 1877–1901.
6. Chappuis, C., Mendez, V., Walt, E., Lobry, S., & Tuia, D. (2022). Prompt-RSVQA: Prompting visual context to a language model for remote sensing visual question answering. IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 1372–1381.
7. Chen, H., & Shi, Z. (2020). A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sensing, 12(10), 1662.
8. Cover, T. M., & Thomas, J. A. (2006). Elements of Information Theory (2nd ed.). Wiley-Interscience.
9. Dong, Q., Li, L., Dai, D., Zheng, C., Ma, J., Li, R., Xia, H., Xu, J., Wu, Z., Chang, B., Sun, X., Li, L., & Sui, Z. (2024). A Survey on In-context Learning. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 1107–1128, Miami, Florida, USA. Association for Computational Linguistics.
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., ... & Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In the 9th International Conference on Learning Representations (ICLR).
11. Garg, S., Tsipras, D., Liang, P., & Valiant, G. (2022). What can transformers learn in-context? A case study of simple function classes. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 30583–30598.
12. Håstad, J. (1987). Computational limitations of small-depth circuits. *MIT Press*.
13. Hawarey, M. (2026a). Characterizing In-Context Learning: When Can Transformers Match Standard Learning Algorithms? AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026277, DOI: 10.65737/AIRMCS2026277.
14. Hawarey, M. (2026b). Quantum-Enhanced In-Context Learning for Geopotential Field Estimation: A Theoretical Framework. AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026322, DOI: 10.65737/AIRMCS2026322.
15. Hawarey, M. (2026c). ICL Characterization of Geo-Foundation Models. AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026360, DOI: 10.65737/AIRMCS2026360.
16. Hawarey, M. (2026d). ICL Characterization of Climate Foundation Models: When Can Transformers Learn Weather and Climate? AIR Journal of Mathematics & Computational Sciences, Vol. 2026, AIRMCS2026405, DOI: 10.65737/AIRMCS2026405.
17. Hu, Y., Yuan, J., Wen, C., Lu, X., & Li, X. (2023). RSGPT: A remote sensing vision language model and benchmark. *arXiv preprint arXiv:2307.15266*.

18. Kazemzadeh, S., Ordonez, V., Matten, M., & Berg, T. (2014). ReferItGame: Referring to objects in photographs of natural scenes. *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 787–798.
19. Kuckreja, K., Danish, M. S., Naseer, M., Das, A., Khan, S., & Khan, F. S. (2024). GeoChat: Grounded large vision-language model for remote sensing. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 27831–27840.
20. Li, C., Chen, W., Hu, Y., Yuan, J., & Wen, C. (2024). RS-LLaVA: A large vision-language model for joint captioning and question answering in remote sensing imagery. *IEEE Geoscience and Remote Sensing Letters*, 21, 1–5.
21. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *International Conference on Machine Learning (ICML)*, 162, 12888–12900.
22. Li, Y., Ildiz, M. E., Papailiopoulos, D., & Oymak, S. (2023). Transformers as algorithms: Generalization and stability in in-context learning. *International Conference on Machine Learning (ICML)*, 202, 19565–19594.
23. Liu, C., Zhao, R., Chen, H., Zou, Z., & Shi, Z. (2024a). Remote sensing image change captioning with dual-branch transformers: A new method and a large scale dataset. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–20.
24. Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2024b). Visual instruction tuning. *Advances in Neural Information Processing Systems (NeurIPS)*, 36.
25. Lobry, S., Marcos, D., Murray, J., & Tuia, D. (2020). RSVQA: Visual question answering for remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 58(12), 8555–8566.
26. Lu, X., Wang, B., Zheng, X., & Li, X. (2017). Exploring models and data for remote sensing image caption generation. *IEEE Transactions on Geoscience and Remote Sensing*, 56(4), 2183–2195.
27. Mai, G., Huang, W., Sun, J., Song, S., Mishra, D., Liu, N., ... & Lao, N. (2024). On the opportunities and challenges of foundation models for geospatial artificial intelligence. *arXiv preprint arXiv:2304.06798*.
28. Mao, J., Huang, J., Toshev, A., Camburu, O., Yuille, A. L., & Murphy, K. (2016). Generation and comprehension of unambiguous object descriptions. *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11–20.
29. Mendieta, M., Han, B., Shi, X., Zhu, Y., & Chen, C. (2024). Towards geospatial foundation models via continual pretraining. *IEEE/CVF International Conference on Computer Vision (ICCV)*, 16532–16542.
30. Merrill, W., & Sabharwal, A. (2023). The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics (TACL)*, 11, 531–545.
31. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. *International Conference on Machine Learning (ICML)*, 139, 8748–8763.
32. Sanford, C., Hsu, D., & Telgarsky, M. (2023). Representational strengths and limitations of transformers. *Advances in Neural Information Processing Systems*, 36, 38693–38716.
33. Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press.

34. Sun, X., Wang, P., Yan, Z., Xu, F., Wang, R., Diao, W., ... & Fu, K. (2022). FAIR1M: A benchmark dataset for fine-grained object recognition in high-resolution remote sensing imagery. *ISPRS Journal of Photogrammetry and Remote Sensing*, 184, 116–130.
35. Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley.
36. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems (NeurIPS)*, 30, 5998–6008.
37. Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., & Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. *International Conference on Machine Learning (ICML)*, 202, 35151–35174.
38. Wang, Z., Peng, R., Zhang, T., Yang, W., Wang, Y., & Meng, F. (2024). SkyScript: A large and semantically diverse language model for remote sensing image interpretation. AAAI Conference on Artificial Intelligence, 38(6), 5765–5773.
39. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., ... & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in Neural Information Processing Systems (NeurIPS)*, 35, 24824–24837.
40. Yu, L., Poirson, P., Yang, S., Berg, A. C., & Berg, T. L. (2016). Modeling context in referring expressions. *European Conference on Computer Vision (ECCV)*, 69–85.
41. Yuan, Z., Zhang, W., Fu, K., Li, X., Deng, C., Wang, H., & Sun, X. (2022). Exploring a fine-grained multiscale method for cross-modal remote sensing image retrieval. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–19.
42. Zhan, Y., Xiong, Z., & Yuan, Y. (2023). RSVG: Exploring data and models for visual grounding on remote sensing data. *IEEE Transactions on Geoscience and Remote Sensing*, 61, 1–13.
43. Zhang, W., Cai, M., Zhang, T., Zhuang, Y., & Mao, X. (2024). EarthGPT: A universal multi-modal large language model for multi-sensor image comprehension in remote sensing domain. *IEEE Transactions on Geoscience and Remote Sensing*, 62, 1–28.
44. Zheng, X., Wang, B., Du, X., & Lu, X. (2021). Mutual attention inception network for remote sensing visual question answering. *IEEE Transactions on Geoscience and Remote Sensing*, 60, 1–14.