GeoGround: Uncertainty-Weighted Multi-Task Learning for Geo-Alignment and Address Defect Detection

Srinivas Virinchi, Aman Gulati, Anoop Saladi
International Machine Learning, Amazon, Bengaluru, India
{virins,amangula,saladias}@amazon.com

Abstract

Address intelligence in e-commerce demands accurate geocoding and proactive defect detection under strict sub-50 ms latency constraints. These tasks are inherently coupled: precise spatial grounding provides a strong prior for defect propensity, yet prior approaches optimize them independently. While generative LLMs offer rich semantic representations, they lack spatial inductive bias and fail to meet real-time serving requirements. We introduce GeoGround, a multi-task learning framework that jointly models coordinate grounding and address defect detection. The model combines a hierarchical spatial grounding objective with Focal Loss for defect classification, using uncertainty-based task weighting to balance optimization under severe class imbalance. To strengthen supervision, we curate a large-scale noisy address dataset using LLM-assisted data construction, augmenting the training corpus with signals that are costly to obtain manually. GeoGround achieves 5.86× gains in address defect detection precision and up to 4.86× improvements in location prediction accuracy over strong encoder baselines, while remaining 75× more efficient than decoder LLMs such as Qwen2-1.5B. A two-week online A/B test in a large-scale delivery pipeline confirms real-world impact, yielding a 50 bps uplift in defect detection, a 40 bps gain in location prediction, and an estimated operational savings of $3.09M annually.

1 Introduction

Building robust address intelligence for e-commerce requires solving two interlinked tasks using free-form customer addresses. Although these inputs are frequently incomplete, misspelled, or strategically obfuscated, they serve a dual role: enabling precise last-mile delivery and signaling localized address defects (e.g., loss, damage, or fraud). The first task, Location[1] Prediction, requires geospatial awareness to cluster semantically diverse but geographically proximate inputs (e.g., `XXX E 52nd St, NY` and `XXX E. 52, b/w Park/Mad., NYC`)[2] while maintaining cold-start robustness for fragmentary data. The second task, Defect Detection, identifies geographically correlated operational risks. As shown in Fig. 1, these defects often manifest as localized hotspots while adjacent areas remain unaffected; this spatial correlation confirms that geographic location is a critical prior for predicting defect propensity.

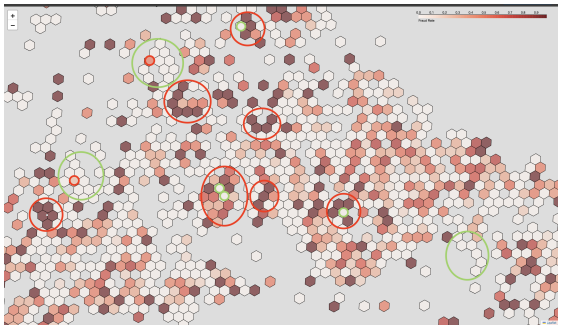

Figure 1: Spatial distribution of defect rates across an H3 (Brodsky et al., 2018) grid, ranging from white (low, ≈0) to dark red (high, ≈1). Red circles indicate defect clusters; green denote low-defect regions.

Problem. We aim to learn a unified representation that maps raw address text to geocode while simultaneously predicting operational defect propensity. A key challenge is the disparity in data availability: while the model is supervised using paired address–geocode data during training, it must infer both precise location and defect risk using only raw address text at inference. This dual-task objective must overcome extreme label imbalance ($< 5\%$ positives) while satisfying strict sub-50 ms production latency constraints.

Motivation. Existing approaches (Govind and Sohoney, 2022; Singh et al., 2025; Govind et al.,

[1]The terms location and geocode are used interchangeably to refer to the latitude and longitude of an address.

[2]Hereafter, fine-grained address details (e.g., unit or house numbers) are masked with `X` to preserve privacy.

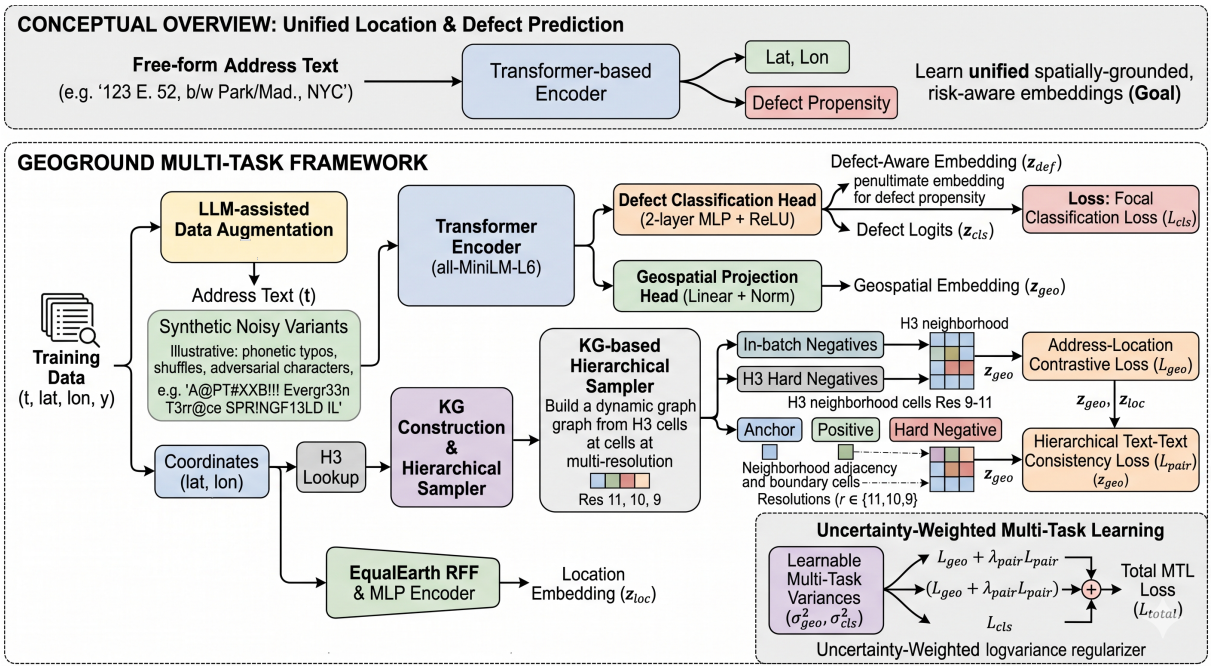

Figure 2: Learning from paired address–geocode data to predict both geolocation and defect propensity from text alone. The proposed GeoGround framework jointly optimizes geospatial alignment and defect prediction via Multi-Task Learning (MTL), enhanced by hierarchical spatial consistency and uncertainty-weighted loss balancing.

2025) remain decoupled: they either discretize space, align text to coordinates, or model task-specific risks without capturing the intrinsic synergy between address structure and operational defects. Such decoupling underutilizes shared geospatial priors and leads to unstable defect detection. Central to bridging this gap is the H3 (Brodsky, 2018b) index—a hexagonal hierarchical grid that discretizes the globe into multi-resolution identifiers (see Appendix A for details). By leveraging H3's uniform hexagonal symmetry, we enable stable neighborhood modeling and enforce spatial consistency across hierarchical resolutions (house → street → locality). Embedding this hierarchy into a unified representation allows the model to leverage shared spatial priors, which is essential for robust, real-time address intelligence.

1.1 Contributions.

To solve these challenges, we introduce GeoGround, a unified multi-task framework that jointly optimizes for geographic grounding and defect awareness. Our key contributions include:

- **Unified MTL Architecture:** We integrate geospatial alignment and defect prediction into a single end-to-end model, jointly optimizing two tasks: *(i) Address Defect Detection* and *(ii) Location Prediction*. Using Uncertainty-Weighted MTL (Kendall et al., 2018), we dynamically balance heterogeneous losses to prevent gradient interference and ensure stable convergence. GeoGround is a modular and extensible framework; its components are designed to be task-agnostic, allowing them to be deployed in any domain where a synergy exists between spatial grounding and operational risk prediction.

- **Multi-Resolution Hierarchical Supervision:** We propose a dual-loss strategy coupling address–location alignment with a hierarchical text–text contrastive objective. By sampling hard negatives from multi-resolution H3 neighborhoods (Res 9–11), GeoGround injects fine-grained spatial priors to improve cold-start robustness.

- **LLM-Based Data Augmentation:** We leverage LLMs to synthetically generate a robust dataset of incomplete, misspelled, and strategically obfuscated addresses, ensuring model precision against real-world adversarial and fragmentary inputs.

- **Empirical and Production Impact:** GeoGround outperforms strong encoder baselines with $5.86\times$ higher defect precision and $4.86\times$ better location accuracy, while being

75× more efficient than decoder-LLMs. A two-week A/B test yields production uplifts of 50 bps (defects) and 40 bps (location), totaling $3.09M in annual savings.

2 Related Work

Spatial Indexing and Geospatial Representation. Hierarchical spatial indices like H3 (Brodsky, 2018a), S2 Geometry (Veach, 2017), and Geohash (Morton, 1966) map coordinates into discrete, hierarchical tokens. While Geohash and S2 utilize rectangular or quadtree-based tiling, H3's hexagonal uniformity minimizes shape distortion, benefiting neighborhood-based learning tasks (Kothari and Sohoney, 2022). Unlike prior work that treats these grids as static labels, we leverage multi-resolution H3 cells (res. 9–11) to form anchor–positive pairs. This imposes a hierarchical consistency that internalizes spatial proximity directly into the embedding space, removing the need for geocode lookups during inference.

Contrastive, Graph, and LLM-Based Learning. Geospatial modeling has transitioned from task-specific interpolation (Govind and Sohoney, 2022) to unified contrastive frameworks. GeoCLIP (Wu and et al., 2022) and AddressBind (Govind et al., 2025) align text with coordinates, while graph-driven approaches (Penumadu et al., 2022; Yuan and et al., 2021; Klemmer et al., 2023) refine grounding through local connectivity. Recently, Large Language Models (LLMs) like Llama-3.2-1B (Dubey et al., 2024) and Qwen2-1.5B (Yang et al., 2024) have enabled generative address parsing. Despite their semantic depth, LLMs lack native spatial inductive bias and struggle with geographic continuity. Our approach bridges this gap by combining the linguistic capacity of compact encoders with multi-scale spatial priors sampled from a knowledge graph.

Multi-Task Modeling and Loss Balancing. Address intelligence frequently separates geocoding and defect detection into disjoint tasks (Qian and et al., 2021), limiting generalization. Multi-task learning (MTL) provides a unified alternative (Ruder, 2017; Liu and et al., 2019), though heterogeneous objectives can suffer from gradient interference. We utilize uncertainty-weighted MTL (Kendall et al., 2018) to balance geocoding and classification objectives, alongside Focal Loss (Lin et al., 2017) to manage defect class imbalance. This architecture yields a single, low-latency embedding that is simultaneously geospatially grounded and defect-sensitive.

3 Multi-Task Geospatial Grounding Architecture

GeoGround addresses these requirements through a unified Multi-Task Learning (MTL) framework (Fig. 2), guided by the philosophy that address embeddings should not just map to geocodes, but internalize the hierarchical nature of space itself. This design follows a three-stage narrative: First, we inject *scalable spatial priors* by using the H3 hierarchy as a supervision scaffold. This ensures that the representation of an address is anchored to its neighborhood, maintaining consistency across multiple levels of granularity. Second, we move beyond simple alignment; the model is forced to *internalize spatial hierarchy* through a text–text consistency signal. This allows the encoder to remain robust against common address noise while preserving the underlying geographic structure. Finally, to ensure *stable optimization* across these heterogeneous goals, we employ uncertainty-based weighting. This balances the contrastive grounding and imbalanced classification tasks dynamically, preventing gradient interference and allowing the model to learn a truly unified representation of both location and address defect propensity.

3.1 Encoders

Address Encoder. Given an input address text t, a pre-trained Transformer encoder f_θ (e.g., *all-MiniLM-L6-v2* (Reimers and Gurevych, 2019)) produces a base text embedding

$$\mathbf{z}_{\text{text}} = f_\theta(t) \in \mathbb{R}^{d_h}$$

This embedding serves as a shared latent representation for both spatial and defect-aware supervision. A linear projection transforms it into a geospatial embedding:

$$\mathbf{z}_{\text{geo}} = \text{Norm}\big(\text{Linear}_{\text{geo}}(\mathbf{z}_{\text{text}})\big),$$

where $\text{Norm}(\cdot)$ denotes ℓ_2 normalization, ensuring $\mathbf{z}_{\text{geo}} \in \mathbb{R}^d$ lies on the unit hypersphere for cosine-based contrastive learning.

Location Encoder. Each address also has associated geocodes (lat, lon). To enable metric learning in a Euclidean space, these are first projected to the Equal Earth coordinate system, yielding planar coordinates (x, y) through a deterministic mapping

$$(x, y) = \Pi_{\text{EqualEarth}}(\text{lat}, \text{lon}),$$

where $\Pi_{\text{EqualEarth}}(\cdot)$ denotes the Equal Earth projection function (Šavrič et al., 2019) that preserves area while minimizing global distortion. The projected coordinates (x, y) are then encoded as Random Fourier Features (RFF):

$$\phi_{\text{RFF}}(x, y) = [\cos(\Omega[x, y]^\top), \sin(\Omega[x, y]^\top)],$$

where Ω is a is a fixed buffer sampled from a Gaussian distribution to approximate a Gaussian RBF kernel that controls the resolution of spatial encoding (Rahimi and Recht, 2007). The resulting feature vector is processed by a small MLP (Rumelhart et al., 1986) g_ϕ to produce

$$\mathbf{z}_{\text{loc}} = \text{Norm}(g_\phi(\phi_{\text{RFF}}(x, y))),$$

yielding a normalized coordinate embedding $\mathbf{z}_{\text{loc}} \in \mathbb{R}^d$ that lies in the same cosine metric space as $\mathbf{z}_{\text{geo}}$.

Defect-Aware Encoder. To model defect-related semantics, the text embedding $\mathbf{z}_{\text{text}}$ is passed through a two-layer MLP classifier head that yields both the penultimate embedding and output logits:

$$\mathbf{z}_{\text{def}} = \text{ReLU}(\text{Linear}_1(\mathbf{z}_{\text{text}})), \quad \mathbf{z}_{\text{cls}} = \text{Linear}_2(\mathbf{z}_{\text{def}})$$

where $\mathbf{z}_{\text{def}} \in \mathbb{R}^{d'}$ represents a *defect-aware embedding* capturing operational risk semantics. The defect propensity probability is then obtained via a softmax layer $\mathbf{p} = \text{softmax}(\mathbf{z}_{\text{cls}})$, with p_1 denoting the predicted defect score.

3.2 Training Objectives

Knowledge Graph Construction. To enable adaptive hierarchical supervision, GeoGround builds a dynamic multi-resolution H3 graph $\mathcal{G} = (V, E)$ at every mini-batch. Nodes correspond to address embeddings, while edges encode neighborhood adjacency derived from H3 indices at resolutions $\{11, 10, 9\}$. GeoGround retrieves on-the-fly hard positives and hard negatives by querying multi-resolution H3 neighborhoods and cross-resolution boundary cells. This graph provides a unified sampling mechanism used across all contrastive objectives, while maintaining efficient per-batch complexity $\mathcal{O}(|V| \log |V|)$ through fast H3 lookup.

Geospatial Alignment. Positive and negative pairs for this objective are drawn from the dynamic H3 graph $\mathcal{G}$. We align address embeddings $\mathbf{z}_{\text{geo}}$ with geocode embeddings $\mathbf{z}_{\text{loc}}$ using an InfoNCE (Van den Oord et al., 2018) loss. For each sample i, the negative pool $\mathcal{N}_i$ consists of in-batch negatives, $4N$ H3 hard negatives, and N random global negatives, all obtained via the KG-based sampler. The objective is:

$$\mathcal{L}_{\text{geo},i} = -\log \frac{\exp\big(\text{sim}(\mathbf{z}_{\text{geo},i}, \mathbf{z}_{\text{loc},i})/\tau\big)}{D_i},$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity and τ is a temperature scalar. The denominator D_i aggregates the positive pair and all $6N - 1$ negatives:

$$D_i = \exp\left(\frac{\text{sim}(\mathbf{z}_{\text{geo},i}, \mathbf{z}_{\text{loc},i})}{\tau}\right) + \sum_{\mathbf{z}_k \in \mathcal{N}_i} \exp\left(\frac{\text{sim}(\mathbf{z}_{\text{geo},i}, \mathbf{z}_{\text{loc},k})}{\tau}\right).$$

This large heterogeneous negative set ($6N$ terms) promotes fine-grained spatial discrimination. The final loss $\mathcal{L}_{\text{geo}}$ is averaged over all i.

Hierarchical Text–Text Consistency. All triplets (a, p, n) used here are sampled via the same KG-based hierarchical sampler operating over $\mathcal{G}$. We draw positives and negatives at resolutions $r \in \{11, 10, 9\}$ and enforce that the anchor a is closer to its positive p than to its negative n:

$$\mathcal{L}^{(r)}_{\text{pair}} = -\log \frac{e^{s_{ap}/\tau}}{e^{s_{ap}/\tau} + e^{s_{an}/\tau}}$$
$$s_{ap} = \text{sim}(\mathbf{z}_a, \mathbf{z}_p), \quad s_{an} = \text{sim}(\mathbf{z}_a, \mathbf{z}_n)$$

where $\text{sim}(\cdot, \cdot)$ is cosine similarity and τ is a temperature scalar. The overall text–text consistency loss aggregates across resolutions:

$$\mathcal{L}_{\text{pair}} = \sum_r w_r \, \mathbb{E}\left[\mathcal{L}^{(r)}_{\text{pair}}\right],$$

with $w_{11} > w_{10} > w_9$ assigning greater weight to finer spatial structure.

Address Defect Detection. The classification head is trained using the focal loss (Lin et al., 2017):

$$\mathcal{L}_{\text{cls}} = -\alpha_y (1 - p_y)^\gamma \log(p_y),$$

where α_y is a class-balancing coefficient and γ controls the focusing parameter (we use $\gamma = 2$).

Multi-Task Optimization. The overall objective optimizes three losses arising from two tasks: (i) a geospatial task combining $\mathcal{L}_{\text{geo}}$ and the hierarchical text consistency term $\mathcal{L}_{\text{pair}}$, and (ii) a classification task with loss $\mathcal{L}_{\text{cls}}$. We use an uncertainty-weighted framework (Kendall et al., 2018) to replace manual tuning with learnable task variances. The total loss is:

$$\mathcal{L}_{\text{total}} = \tfrac{1}{2} e^{-\log \sigma^2_{\text{geo}}} (\mathcal{L}_{\text{geo}} + \lambda_{\text{pair}} \mathcal{L}_{\text{pair}}) + \tfrac{1}{2} e^{-\log \sigma^2_{\text{cls}}} \mathcal{L}_{\text{cls}} + \tfrac{1}{2} (\log \sigma^2_{\text{geo}} + \log \sigma^2_{\text{cls}})$$

where $\log \sigma^2_{\text{geo}}$ and $\log \sigma^2_{\text{cls}}$ are learnable log-variances for the geospatial and classification tasks. The hyperparameter λ_{pair} acts as a fixed structural prior that calibrates the influence of hierarchical H3 consistency against point-level coordinate precision, ensuring spatial alignment across resolutions within the geospatial objective. Tasks with higher uncertainty (larger σ^2) are down-weighted, allowing more reliable objectives to guide the shared encoder. The log-variance regularizer prevents degenerate solutions, yielding stable multi-task training and an adaptive balance across spatial, textual, and classification objectives.

Data Augmentation. Addresses associated with operational defects often contain high entropy artifacts such as vague locality references, incorrect street names or ZIP codes, token reordering, phonetic distortions, and artificial character injection (e.g., *"APT#XXB!!! 12OO M@IN ST.,,, Spr1ngfie1d IL 627O4 U.S.A."*). While human interpretable, these inputs frequently destabilize traditional encoders and degrade spatial grounding. Although we manually curate high quality defect cases via auditor review, such adversarial patterns remain sparse in production logs, limiting generalization from clean data alone. To scale coverage, we augment the training set with 2% synthetic samples generated via In Context Learning (Brown et al., 2020) (Listing 1), where structured corruption including phonetic noise, structural shuffling, field level perturbations, omissions, and adversarial character injection is introduced while preserving geographic intent. Each synthetic variant inherits its original defect label, explicitly supervising the link between textual ambiguity and operational risk, thereby mitigating *obfuscation bias* and improving robustness under strict real time constraints.

3.3 Training and Inference

Each mini-batch includes paired samples $(t, \text{lat}, \text{lon}, y)$ for geospatial and classification supervision, and triplets (a, p, n, r) generated on-the-fly from the H3 graph for hierarchical consistency. Both encoders are trained jointly to minimize $\mathcal{L}$total, with all embeddings ℓ_2-normalized for metric stability. At inference, only the address encoder is used: $z$geo supports spatial retrieval or clustering, z_{def} encodes defect-aware semantics, and p_1 provides the defect propensity score. This enables fully text-only inference, making GeoGround operationally efficient at scale.

4 Experimental Setup

Datasets. Our dataset comprises tens of millions of address–geocode–label triplets with a severely imbalanced binary defect label (positives $< 5\%$); evaluation is conducted on a temporally held-out split to simulate deployment. These labels are derived from a 14-day window of verified carrier loss reports and fraud investigations, specifically targeting three defect categories: packages reported lost or stolen, damage claims on non-returnable goods, and fraudulent returns where genuine items are replaced with fake items. As detailed earlier, we augment the training distribution with 2% synthetic noisy samples. The 2% augmentation ratio was a Pareto-optimal choice identified through sensitivity analysis, where it served as the "Goldilocks" point for balancing adversarial defect robustness with geocoding accuracy while remaining within the budgetary and latency constraints of production-scale LLM generation. Critically, while training utilizes multimodal signals, inference relies strictly on raw address text to maintain low-latency production requirements.

Baselines. We benchmark `GeoGround` against three model families selected to satisfy a production-mandated < 50 ms latency constraint. Encoders: `RoBERTa-General` (125M) (Liu et al., 2019), the domain-adapted `RoBERTa-Address` (Kothari and Sohoney, 2022), and a `GraphSAGE`-augmented variant (Hamilton et al., 2017) representing graph-contextualized neighborhood information. Spatial-aware Models: Task-specific benchmarks including `H3-aware triplet models` (Singh et al., 2025), `Multilingual-E5-large` (560M) (Wang et al., 2024), and the multimodal `AddressBind` framework (Govind et al., 2025). Lightweight Decoders: `Llama-3.2-1B` and `Qwen2-1.5B`, both fine-tuned for sequence classification to evaluate generative backbones within the target latency budget. We omit larger architectures ($>$ 7B) as they exceed real-time deployment requirements without prohibitive compression.

GeoGround. We utilize a pre-trained Transformer encoder f_θ (e.g., *all-MiniLM-L6-v2* (Reimers and Gurevych, 2019)) to evaluate three architectural configurations[3]. `GeoGround-Frozen` trains a lightweight MLP head on fixed embeddings to isolate the encoder's intrinsic representation

[3]Due to corporate privacy compliance, the proprietary codebase and datasets cannot be shared.

quality. `GeoGround-E2E` jointly fine-tunes f_θ and the classifier to enable tighter task coupling. `GeoGround-MTL` further extends this with multi-task objectives for geospatial alignment and defect propensity. Across all variants, each mini-batch ($\mathcal{N}$) incorporates $4\mathcal{N}$ random and $\mathcal{N}$ hard-negative geocodes to ensure robust spatial alignment during training.

Evaluation. Models generate 384-dimensional embeddings and are trained for up to 10 epochs. We evaluate across three production settings: (1) Defect Detection: reported via top-$k\%$ precision ($k \in \{x_1, x_2, \dots\}$); (2) Geocoding: measured by mean Haversine distance (Inman, 1835) following a nearest-neighbor retrieval-aggregation procedure; and (3) Spatial Accuracy: calculated as Location Prediction Accuracy (precision@y_n) and Defect rate (defect@z_n) for various radii. To isolate *Embedding Quality*, we report HitRate and MRR for geospatial retrieval. To assess *operational performance*, we measure inference throughput and latency on our target hardware. Due to proprietary constraints, we report improvements relative to RoBERTa-General. Training is conducted on an instance equipped with four NVIDIA L4 GPUs (24 GB VRAM each) and 192 GB of system RAM.

5 Results

Performance Analysis. We evaluate geocoding and defect detection in a strict two-stage setup to isolate representational quality. As shown in Table 1, `GeoGround-Frozen` outperforms all baselines, including fine-tuned generative decoders `Llama-3.2-1B` and `Qwen2-1.5B`. While these LLMs leverage vast pre-training for strong semantic defect detection, they offer marginal gains on short, structured address text and trail specialized models like `RoBERTa-Triplet-H3` in location accuracy due to a lack of explicit geospatial inductive bias. These results establish `GeoGround` as a superior representational backbone, with the multi-task formulation (`GeoGround-MTL`) achieving the most significant gains. By jointly optimizing for geolocation and defect cues, `GeoGround-MTL` amplifies defect precision by $6.45\times$ and location accuracy by $5.36\times$ over the baseline. Further, the model achieves the highest Silhouette Score ($7.97\times$ relative to the baseline), indicating that the joint objective forces the embedding manifold into significantly more compact and geographically well-separated regions. GeoGround outperforms all baselines with up to $4.8\times$ HitRate improvement; for detailed results on the geospatial retrieval task, refer to Appendix C due to space constraints.

Ablation Analysis. Ablations confirm that GeoGround-MTL's performance is driven by the synergy of three core pillars (Table 1). First, the *geolocation objective* serves as the primary spatial anchor; its removal precipitates the sharpest decline in Defect Precision ($5.86\times \rightarrow 2.08\times$). Second, *focal loss* and *data augmentation* are essential for imbalanced learning, providing the supervision necessary to identify rare ($< 5\%$) defects. Third, *KG-based contrastive learning* and *uncertainty weighting* provide structural robustness; the former disambiguates semantically similar but distant addresses via hard-negative sampling, while the latter stabilizes convergence against high-variance geospatial gradients.

Error Characterization. A post-hoc analysis of model errors reveals three primary failure modes: Cold-Start Addresses, where new constructions are not yet reflected in the Knowledge Graph or H3 spatial priors; Granular Ambiguity, occurring in high-density complexes where specific "Unit/-Suite" identifiers are absent from the input text; and Extreme Transliteration, where rare phonetic noise exceeds the coverage of our 2% noisy data augmentation. Quantitatively, these modes account for approximately 35%, 45%, and 20% of remaining errors, respectively. By identifying these gaps, we demonstrate that GeoGround's failures are largely driven by external data freshness and input quality rather than architectural limitations, providing a clear roadmap for next steps.

Operational Efficiency. Benchmarking on a single NVIDIA L4 GPU (batch size 64) reveals that GeoGround maintains a sub-10ms latency, split across text tokenization (1.2ms), transformer encoding (6.8ms), and task-specific projection heads (1.4ms), enabling a throughput of ∼6,762 samples/s as detailed in Table 2. In contrast, generative decoders (`Llama`, `Qwen`) impose a prohibitive computational tax; `Qwen2-1.5B`, for example, requires over 700ms per batch—a $75\times$ latency increase that violates strict real-time production Service Level Agreements (SLA). While generative backbones can offer marginal representational depth, their deployment is currently infeasible for high-volume logistics. GeoGround thus occupies the optimal Pareto front, delivering state-of-the-art accuracy within a fraction of the inference budget.

Production Impact. A 2-week online A/B test

Table 1: Comparative performance for location prediction, defect detection, and clustering quality. Ratios (×) are relative to the Roberta-General baseline; ↑ and ↓ denote whether higher or lower values represent improvement.

Category	Model	Address Defect Detection↑		Location Prediction Accuracy↑			Location Prediction Defects↓		Address Clustering↑
		Prec@x1%	Prec@x2%	Prec@y1%	Prec@y2%	Prec@y3%	Def@z1%	Def@z2%	Silhouette Score
Baselines	Roberta-General	1.000x	1.000x	1.000x	1.000x	1.000x	1.000x	1.000x	1.000x
	RoBERTa-Address	1.303x	1.423x	1.080x	1.417x	1.536x	0.994x	0.901x	1.842x
	RoBERTa-GraphSage	1.321x	1.453x	0.811x	0.894x	1.101x	1.366x	1.181x	1.157x
	Multilingual-E5-Large	1.425x	1.596x	1.078x	1.416x	1.534x	0.994x	0.902x	2.105x
	RoBERTa-Triplet-H3	1.511x	1.518x	1.739x	2.283x	2.475x	0.617x	0.561x	5.868x
	RoBERTa-Triplet-H3-VM	1.532x	1.541x	1.753x	2.186x	2.373x	0.634x	0.613x	6.026x
	AddressBind	1.683x	1.738x	1.816x	2.381x	2.581x	0.592x	0.538x	6.578x
	Llama-3.2-1B	1.842x	1.915x	1.652x	1.915x	2.145x	0.720x	0.650x	2.421x
	Qwen2-1.5B	1.921x	2.044x	1.690x	2.044x	2.256x	0.685x	0.612x	2.763x
GeoGround Variants	GeoGround-Frozen	2.081x	2.429x	1.962x	2.536x	2.735x	0.566x	0.493x	7.105x
	GeoGround-E2E	2.837x	3.112x	1.502x	1.823x	1.811x	0.601x	0.581x	6.315x
	GeoGround-MTL	**5.863x**	**6.453x**	**3.777x**	**4.863x**	**5.362x**	**0.482x**	**0.310x**	**7.973x**
Ablation	GeoGround - Uncertainty	5.555x	6.101x	3.475x	4.434x	4.928x	0.485x	0.318x	7.815x
	GeoGround - KG Contrastive	5.083x	5.804x	3.128x	3.935x	4.340x	0.501x	0.339x	7.526x
	GeoGround - Focal Loss	3.901x	4.316x	2.321x	2.866x	3.159x	0.534x	0.373x	7.210x
	GeoGround-Data Augmentation	4.120x	4.580x	3.650x	4.710x	5.210x	0.512x	0.345x	7.657x
	GeoGround - Geolocation	2.081x	2.452x	1.962x	2.536x	2.735x	0.566x	0.492x	7.078x

Table 2: Inference throughput and latency

Model	Architecture Class	Throughput (samples/ms)	Latency (ms/batch)
RoBERTa-General	RoBERTa-base (Standard)	0.678	94.33
RoBERTa-Address	RoBERTa-base (Standard)	0.666	96.10
RoBERTa-Triplet-H3	RoBERTa-base (Standard)	0.661	96.79
RoBERTa-Triplet-H3-Varying-Margin	RoBERTa-base (Standard)	0.661	96.79
RoBERTa-Address-GraphSage	GNN (Text-to-Node)	0.650	98.45
Multilingual-E5-large-Instruct	XLM-R-large (Massive)	0.183	349.09
Llama-3.2-1B	Decoder (Generative)	0.112	571.42
Qwen2-1.5B	Decoder (Generative)	0.089	719.10
AddressBind	MiniLM-L6 (Compact)	6.662	9.61
GeoGround	MiniLM-L6 (Compact)	**6.762**	**9.47**

in a high-scale delivery pipeline confirmed GeoGround's utility, yielding a 50 bps uplift in defect detection and a 40 bps gain in location prediction accuracy. These technical improvements translate to an estimated $3.09M in annual operational savings.

6 Conclusion

We presented GeoGround, a multi task framework for geospatially grounded, defect aware address embeddings. By integrating hierarchical spatial consistency with uncertainty weighted optimization, GeoGround outperforms state of the art baselines while meeting strict production SLAs.

References

Isaac Brodsky. 2018a. H3: Uber's hexagonal hierarchical spatial index. https://www.uber.com/en-SE/blog/h3/.

Isaac Brodsky. 2018b. H3: Uber's hexagonal hierarchical spatial index. https://www.uber.com/blog/h3/. Accessed: 2026-02-14.

Isaac Brodsky, David Crain, Gergely Fekete, Hao Li, Nick Mattingly, and Ben Worsham. 2018. H3: Uber's hexagonal hierarchical spatial index. *Uber Engineering Blog*. https://eng.uber.com/h3/.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Govind, Sayan Putatunda, and Saurabh Sohoney. 2025. Addressbind: Cross-modal alignment of addresses and geocodes for last-mile transportation systems. In *Proceedings of the 33rd ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*. ACM.

Govind and Saurabh Sohoney. 2022. Learning geolocations for cold-start and hard-to-resolve addresses via deep metric learning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: Industry Track (EMNLP)*.

William L. Hamilton, Zhitao Ying, and Jure Leskovec. 2017. Inductive representation learning on large graphs. In *Advances in Neural Information Processing Systems*.

James Inman. 1835. Navigation and nautical astronomy: For the use of british seamen.

Alex Kendall, Yarin Gal, and Roberto Cipolla. 2018. Multi-Task Learning Using Uncertainty to Weigh Losses for Scene Geometry and Semantics. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7482–7491.

Konstantin Klemmer, Nathan S. Safir, and Daniel B. Neill. 2023. Positional encoder graph neural net-

works for geographic data. In *Proceedings of AISTATS 2023*, volume 206, pages 1379–1389.

Govind Kothari and Saurabh Sohoney. 2022. Geospatially informed models for geocoding unstructured addresses. *arXiv preprint arXiv:2203.01234*.

Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. 2017. Focal loss for dense object detection. *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Saining Liu and et al. 2019. End-to-end multi-task learning with attention. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Li, Sanqiang Ko, Justin Mahoney, Yangfu Chen, Emily Glass, Xiaodong Le, Veselin Stoyanov, Jacob Eisenstein, Anca Rus, Mohit Vijayanarasimhan, Fei Deng, Ali Ghazvininejad, Mike Guler, Scott Johnson, Luke Lewis, and Luke Zettlemoyer. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Guy Macdonald Morton. 1966. A computer oriented geodetic data base and a new technique in file sequencing. Technical report, IBM Ltd. Ottawa, Canada.

Vamsi Krishna Penumadu, Nitesh Methani, and Saurabh Sohoney. 2022. Learning geospatially aware place embeddings via weak supervision. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management (CIKM)*, pages 4071–4080.

Hui Qian and et al. 2021. Seq2seq geocoding for chinese addresses with hierarchical attention. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*.

Ali Rahimi and Benjamin Recht. 2007. Random features for large-scale kernel machines. *Proceedings of the Neural Information Processing Systems (NeurIPS)*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Sebastian Ruder. 2017. An overview of multi-task learning in deep neural networks. *arXiv preprint arXiv:1706.05098*.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. 1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Uddeshya Singh, Gowtham Bellala, Ravi Shankar Devanapalli, and Vikas Goel. 2025. Geo-spatially informed models for geocoding unstructured addresses. In *Proceedings of the 31st International Conference on Computational Linguistics: Industry Track*, pages 236–242.

Aaron Van den Oord, Yazhe Li, and Igor Babakhin. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Eric Veach. 2017. S2 geometry library. https://s2geometry.io/.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2024. Multilingual e5 text embeddings: A technical report. *arXiv preprint arXiv:2402.05672*.

Jiachen Wu and et al. 2022. Exploring vision-language models for geographic location prediction. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, and … et al. 2024. Qwen2 technical report. *arXiv preprint*.

Jianing Yuan and et al. 2021. Place representation learning with graph neural networks for warm-start geocoding. In *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*.

Bojan Šavrič, Tom Patterson, and Bernhard Jenny. 2019. The equal earth map projection. *International Journal of Geographical Information Science*, 33(3):454–465.

A Background

We utilize the H3 hexagonal hierarchical spatial index to discretize the coordinate space into a structured grid for multi-resolution analysis. Unlike traditional square-based subdivisions, H3's hexagonal geometry ensures uniform distance between a cell and all six immediate neighbors, facilitating more stable neighborhood modeling and reducing edge-effect artifacts during spatial aggregation. The hierarchical nature of H3 allows GeoGround to internalize multi-scale spatial priors—from street-level detail to neighborhood-level context—by jointly optimizing consistency across various resolutions (e.g., resolutions 9, 10, and 11). Furthermore, the integer-based representation of H3 cells enables high-performance spatial joins and lookups, which is essential for meeting the sub 50 ms real-time latency requirements of production logistics pipelines.

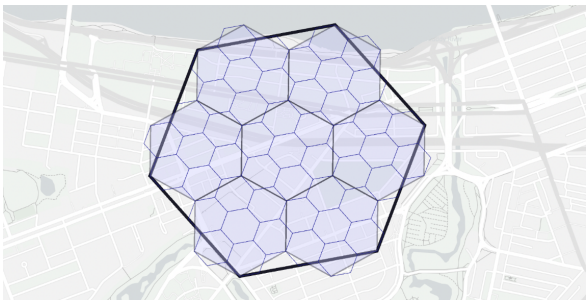

Figure 3: Demonstration of h3 grid and dependency between different resolutions

B Noise Data Augmentation

```
System: You are an expert in geospatial
    data augmentation for adversarial
    robustness. Generate realistic, high
    -entropy address variations that
    preserve latent geographic intent
    but introduce structured noise
    likely to challenge traditional
    encoders.

Important Constraints:
- All clean addresses include a ZIP/
    postal code.
- Preserve geographic plausibility and
    human interpretability.
- Do not change the underlying city or
    state.
- If corrupting the ZIP code, keep it
    syntactically plausible.
- Generate exactly ONE noisy variation
    per corruption type.
- Output ONLY valid JSON.

Corruption Types:
1. phonetic_typo
2. structural_shuffling
3. token_merging
4. abbreviation_drift
5. omission_vague
6. adversarial_injection  (artificial
    characters, excessive punctuation,
    case distortion, letter-number swaps
    , symbol insertion)

### Examples:

Example 1
Clean Address: 1200000742 Evergreen
    Terrace, Springfield, IL 62704
Output:
{
  "phonetic_typo": "1200000742 Evergren
      Terace, Springfeild, IL 62704",
  "structural_shuffling": "Springfield,
      IL 62704, 1200000742 Evergreen
      Terrace",
  "token_merging": "1200000742
      EvergreenTerrace, Springfield,
      IL62704",
  "abbreviation_drift": "1200000742
      Evergreen Ter., Springfield, IL
      6270",
  "omission_vague": "Evergreen Terrace,
      near downtown Springfield, IL",
  "adversarial_injection":
      "1200000742!!! Evergr33n T3rrace
      ,,, SPR!NGF13LD IL 627O4 U.S.A."
}

Example 2
Clean Address: 161232323200 Pennsylvania
    Avenue NW, Washington, DC 20500
Output:
{
  "phonetic_typo": "161232323200
      Pensylvania Ave NW, Washngton, DC
      20500",
  "structural_shuffling": "Washington,
      DC 20500, Pennsylvania Avenue NW
      1600",
  "token_merging": "161232323200
      PennsylvaniaAveNW,
      WashingtonDC20500",
  "abbreviation_drift": "161232323200
      Pennsylvania Ave., Washington, DC
      2050",
  "omission_vague": "Pennsylvania Avenue
      NW, Washington, DC",
  "adversarial_injection":
      "161232323200!!! P3NNSYLV@N!A AVE
      NW,,, W@SH1NGT0N DC 2050O U.S.A."
}

### Target Task:
Clean Address: {input_address}

Generate noisy variations for all six
    corruption types above.
Output ONLY valid JSON in the same
    format.
```

Listing 1: ICL Prompt for Multi-type Adversarial Address Noise Generation

C Embedding Quality

Geospatial Retrieval Task. We further assess fine-grained spatial fidelity through a *Geospatial Retrieval Task*, designed to measure how effectively embeddings encode neighborhood-level location semantics. Given a query address, the objective is to retrieve its true geospatial neighbor from a 1000-candidate pool sampled from diverse regions. This setting captures a realistic use case in logistics systems, where nearest-neighbor retrieval is used for warm-start candidate generation and regional assignment.

We report *hitrate@z1*—the probability that the top retrieved address matches the true neighbor—and *MRR*, which reflects the retrieval rank of the nearest ground-truth match. The three difficulty tiers (x1, y1, z1) correspond to increasingly challenging candidate distributions, with z1 representing the most geographically diverse pool.

Table 3: Geospatial Retrieval Task. Higher indicates better geospatial alignment.

Model	hitrate@x1 ↑	hitrate@y1 ↑	hitrate@z1 ↑	MRR@x1 ↑	MRR@y1 ↑	MRR@z1 ↑
RoBERTa-General	1.000x	1.000x	1.000x	1.000x	1.000x	1.000x
RoBERTa-Address	1.098x	1.210x	1.320x	1.032x	1.059x	1.114x
Llama-3.2-1B	1.842x	2.115x	2.450x	1.520x	1.785x	2.012x
Qwen2-1.5B	2.015x	2.340x	2.680x	1.685x	1.920x	2.245x
RoBERTa-Address-GraphSage	1.050x	1.180x	1.220x	1.010x	1.021x	1.081x
RoBERTa-Triplet-H3	3.000x	3.247x	3.454x	2.581x	2.676x	2.800x
Multilingual-E5-large-Instruct	2.984x	2.012x	3.014x	1.935x	2.343x	2.852x
AddressBind	3.686x	4.205x	4.536x	3.097x	3.265x	3.429x
GeoGround	**4.000x**	**4.500x**	**4.800x**	**3.300x**	**3.500x**	**3.700x**

GeoGround delivers the strongest retrieval performance across all tiers, achieving up to a $4.8\times$ improvement in hitrate and $3.7\times$ improvement in MRR over RoBERTa-General. Notably, it surpasses even AddressBind—despite using the same underlying MiniLM backbone—highlighting the value of combining hierarchical H3 alignment with uncertainty-weighted multi-task training.

These results underscore two important findings. First, spatial contrastive learning alone (e.g., in AddressBind or Triplet-H3) improves retrieval substantially, but the addition of defect supervision and uncertainty weighting produces embeddings that are not only geospatially aligned but also more robust to noisy and ambiguous address text. Second, the strong retrieval accuracy complements the earlier classification results (Table 1), demonstrating that a single unified encoder can maintain spatial fidelity while simultaneously improving defect prediction—without requiring model duplication or multiple fine-tuning stages.