

GeoGTM: Vision-Language Geospatial Intelligence for Street-Level Telecom Growth

Junchi Ren¹, Chengyu Zhou² and Tao He^{3*}

¹ China Telecom, Nanjing, China
454073828@qq.com

² LexisNexis, New York, USA
czhou24190@gmail.com

³ SuperScaling Technologies, LLC, Dallas, USA
frank.he@superscaling.xyz

Abstract. Street-scale telecom growth requires jointly reasoning over heterogeneous evidence, including geospatial context (buildings, communities, and business districts), network capability (coverage and capacity), and customer-side signals (service usage and support tickets). Existing LLM-based sales assistants are often ungrounded: they ignore deliverability constraints, lack verifiable evidence, and fail to produce actionable plans that can be executed by field teams. In this paper, we introduce StreetCopilot, a grounded vision-language agent for street-level prospecting and service planning. StreetCopilot integrates (i) geospatial visual cues (e.g., street-view/remote-sensing building context and POIs), (ii) structured telecom signals (coverage maps, traffic KPIs, product portfolios), and (iii) unstructured operational text (tickets and visit notes) via a retrieval-augmented reasoning pipeline.

To ensure actionability, we propose a deliverability-aware constraint module that verifies whether recommended bundles (network, wireless coverage, industrial devices, and cloud services) are feasible under local coverage and resource conditions, and a citation-grounded generation mechanism that attaches evidence snippets to each recommendation for auditability. We further present a street-scale closed-loop evaluation protocol that measures not only recommendation accuracy but also plan feasibility, evidence faithfulness, and end-to-end business outcomes (lead acceptance and conversion). Experiments on a real-world deployment in a city subregion demonstrate that StreetCopilot substantially improves prospect ranking quality and proposal drafting efficiency while maintaining high feasibility and evidence faithfulness, shedding light on grounded multimodal agents for real-world decision-making.

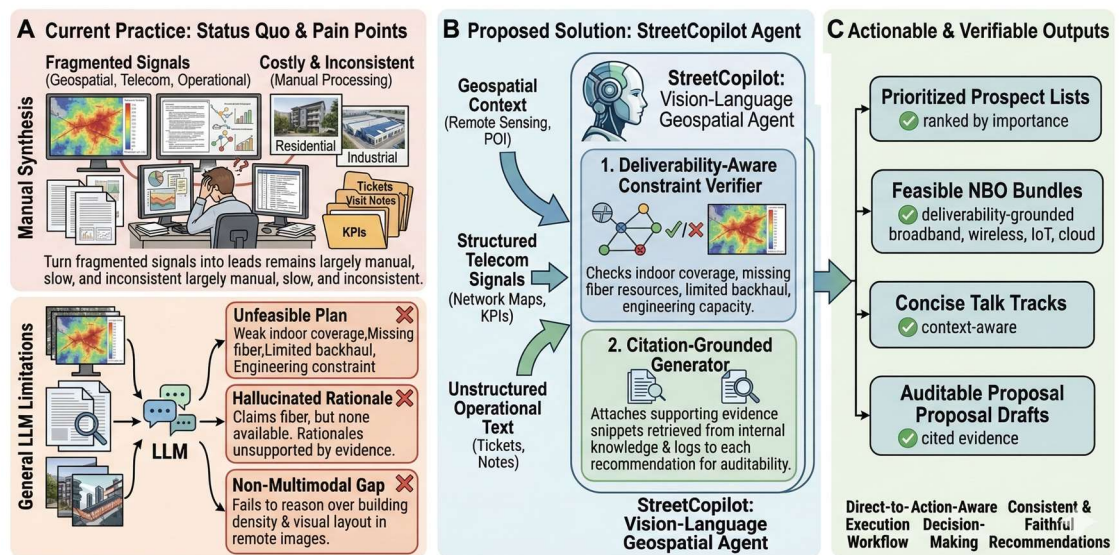
Keywords: Grounded multimodal agent, Telecom prospecting, Deliverability, aware reasoning, Evidence citation.

1 Introduction

Telecom operators are increasingly shifting from coarse city-level campaigns to street-scale go-to-market (GTM), where growth

hinges on localized and execution-aware decisions for both public and enterprise customers. Within a typical subdistrict (for example, a street office jurisdiction), frontline teams repeatedly face practical questions: which communities, buildings, or businesses should we approach first; what bundle should we offer, including broadband, indoor or outdoor wireless coverage, industrial devices or IoT, and cloud services; and whether the proposed plan can be delivered given current network conditions and engineering capacity. Answering these questions is costly because it requires synthesizing heterogeneous evidence distributed across network maps and KPIs such as coverage, capacity, and backhaul; operational records such as tickets, complaints, and visit notes; and geospatial context such as building typology, business density, and nearby points of interest. As a result, turning fragmented signals into actionable prospect lists and proposal drafts remains largely manual, slow, and inconsistent.

Scientific Motivation Figur: A research paper on Street-scale telecom GTM (go-to-market)



Motivation and architecture of StreetCopilot. Overcoming the inefficiencies of manual data synthesis and the unreliability of general LLMs (Left), StreetCopilot (Middle) approaches street-scale

telecom GTM as a grounded multimodal reasoning task. By integrating heterogeneous signals with deliverability verification and citation grounding, the agent generates actionable, feasibility-aware, and auditable GTM recommendations for frontline execution (Right).

Large language models (LLMs) (Naveed et al. 2025; Zhao et al. 2024), offer an appealing interface for accelerating GTM workflows thanks to their strengths in summarization, instruction following, and document drafting. However, as shown in Fig. 1 (a), directly deploying general-purpose LLM assistants for street-scale telecom GTM is unreliable for at least three reasons. First, recommendations must be deliverability-grounded, since a bundle that sounds convincing may be infeasible due to weak indoor coverage, missing fiber resources, limited backhaul, or insufficient engineering capacity. Second, operational decision-making requires verifiable justification and auditability, whereas LLM outputs may hallucinate, omit key constraints, or provide rationales unsupported by evidence. Third, the problem is inherently multimodal and geospatial, because visual cues from street-view or remote-sensing imagery, such as building density, facility layout, and industrial park structures, and map-based network layers, such as coverage heatmaps and traffic distribution, often provide decisive signals that are difficult to encode in text alone. These gaps motivate a grounded multimodal framework that can reason over geospatial context and network constraints while producing auditable and actionable outputs.

In this paper, we formulate street-scale telecom GTM as a grounded multimodal reasoning problem. As shown in Fig. 1 (b), given a target region partitioned into sites, including communities, buildings, and enterprises, the system should prioritize prospects, recommend feasible service bundles, and generate customer-facing and engineer-facing artifacts with evidence-backed rationales. We present StreetCopilot, a vision-language geospatial agent that integrates three complementary information sources: geospatial visual context from street-view or remote-sensing imagery and points of interest; structured telecom signals such as

coverage and capacity maps, traffic KPIs, and product portfolios; and unstructured operational text including tickets and visit notes. To ensure actionability, StreetCopilot employs a deliverability-aware constraint verifier that checks whether proposed bundles satisfy local coverage and resource constraints, and a citation-grounded generator that attaches supporting evidence snippets retrieved from internal knowledge and logs to each recommendation for auditability. The agent produces outputs aligned with frontline workflows, including prioritized lead lists, next-best-offer bundles, concise talk tracks, and proposal drafts with implementation suggestions.

To evaluate such systems beyond standard recommendation metrics, as shown in Fig. 1 (c), we construct a realistic benchmark that measures not only prospect ranking quality, but also plan feasibility under network and resource constraints, evidence faithfulness of generated rationales, and operational efficiency, such as proposal drafting time and lead acceptance rate in controlled deployments. Experiments on a real-world subdistrict-scale dataset show that StreetCopilot improves prospecting quality and proposal drafting efficiency while maintaining strong feasibility and evidence faithfulness. These results highlight the potential of grounded multimodal agents for decision-critical and locality-sensitive telecom applications. Our main contributions are:

- We introduce street-scale telecom GTM as a grounded multimodal reasoning task that jointly requires prospect prioritization, feasible bundle recommendation, and auditable artifact generation over geospatial and operational signals.
- We propose StreetCopilot, a vision-language geospatial agent with deliverability-aware constraint verification and citation-grounded generation to produce actionable and verifiable recommendations.

- We construct an evaluation protocol and benchmark emphasizing feasibility, faithfulness, and workflow efficiency, and provide empirical results demonstrating consistent gains in real-world deployments.

2 Related Work

2.1 LLMs and Agents for Grounded Decision-Making

Large language models (Cheng et al. 2025; Gu et al. 2024; Koh et al. 2024; Goswami et al. 2025) have been widely adopted as general-purpose assistants for planning, tool use, and content generation, and recent work has further extended them into agentic systems that iteratively retrieve information, call external tools, and refine outputs through multi-step reasoning. In real-world decision-making settings, grounding is essential because outputs must be supported by evidence and constrained by feasibility. Retrieval-augmented generation has therefore become a standard approach to reduce hallucinations and to inject domain knowledge into generation, enabling LLMs to operate over private corpora such as manuals, policies, operational logs, and customer records. Beyond text-only grounding, multimodal LLMs have expanded the evidence space to include images, maps, and structured tables, which is critical for applications where visual context and spatial priors influence decisions.

A related line of research focuses on verifiable and auditable generation, including citation-backed answering, constrained decoding, and post-hoc validation with rule engines or discriminative checkers. These techniques are particularly relevant to operational domains where incorrect claims or infeasible recommendations can incur significant cost. In parallel, reasoning-time methods such as self-consistency, best-of- N sampling, and tree-based search have been explored to improve reliability by exploring multiple candidates and selecting those with stronger

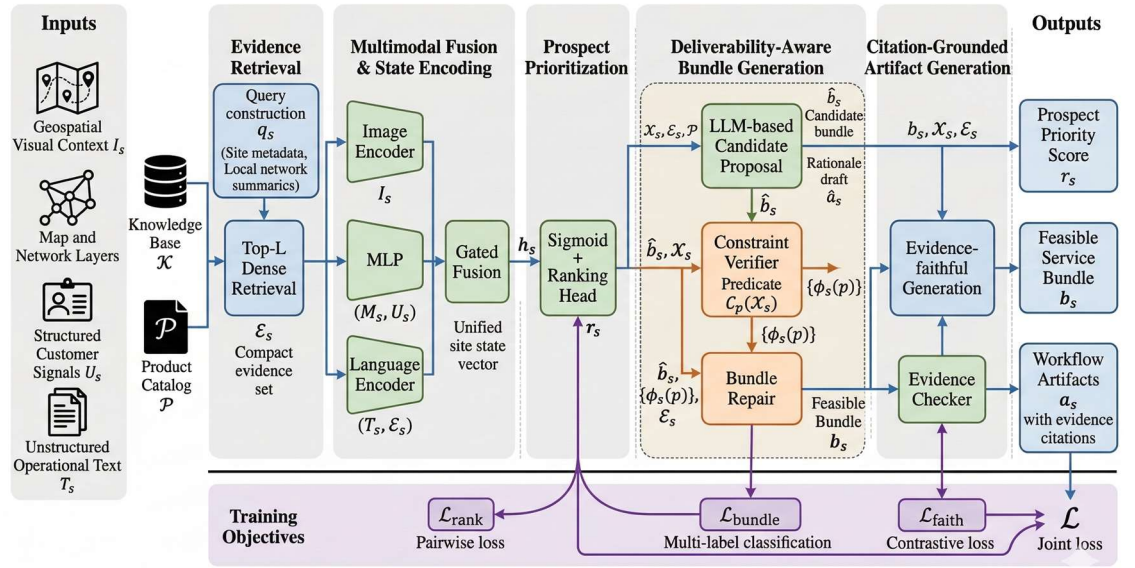
support. Our work aligns with these efforts but targets a distinct setting in which grounding must jointly cover heterogeneous sources, including geospatial visual signals, network capability layers, and operational text, while satisfying deliverability constraints. We further emphasize workflow-aligned artifacts, such as prospect prioritization, talk tracks, and proposal drafts, rather than single-turn question answering.

2.2 Geospatial and Network-Aware Analytics for Telecom Operations and Marketing

Telecom analytics has long studied network planning, capacity management, and quality-of-service monitoring using signals such as coverage measurements, traffic KPIs, and fault or ticket logs. Classical approaches (Dandamudi et al. 2025; Gehlot et al. 2022; Hesham et al. 2024; Wagh et al. 2024; Dastagiraiah 2024; Afzal et al. 2024) typically build predictive models for demand forecasting, congestion detection, and resource allocation, and recent learning-based methods incorporate spatiotemporal modeling to capture correlations across regions and time. In parallel, customer analytics and marketing optimization have explored churn prediction, propensity modeling, and next-best-offer recommendation using customer profiles and usage patterns. While these approaches provide useful components for GTM, they often operate on a single modality or a narrow set of structured features, and they rarely produce end-to-end actionable artifacts that connect network feasibility to marketing decisions.

Urban-scale vision and geospatial learning offer complementary signals for street-level reasoning (Zhang et al. 2026; Tang et al. 2024; Cheshmehzangi 2026), including building typology, functional zones, and points of interest derived from street-view imagery, remote sensing, and map data. These cues have been used for tasks such as land-use classification, urban function inference, mobility and demand estimation, and location-based recommendation. However, existing geospatial learning systems are typically not integrated with telecom-specific deliverability

constraints, and their outputs are not designed for operational workflows such as bundle recommendation and proposal generation. Our work bridges these directions by combining geospatial visual context with telecom network capability signals and operational text within a grounded agent framework, enabling street-scale prospecting and service planning with feasibility verification and evidence-backed generation.



The inference and training pipeline of StreetCopilot. The framework consists of four main stages: (1) evidence retrieval and multimodal state encoding to fuse visual, structured, and textual signals into a unified representation h_s ; (2) prospect prioritization to generate site ranking scores r_s ; (3) deliverability-aware generation that repairs LLM proposals into a feasible bundle b_s using deterministic constraint predicates; and (4) citation-grounded artifact generation to produce auditable outputs a_s . The model components are jointly optimized using ranking, bundle classification, and contrastive faithfulness losses.

3 Method

Problem Formulation. We study street-scale telecom GTM in a target region that is partitioned into a set of sites $\mathcal{S} = \{s_i\}_{i=1}^N$, where each site corresponds to a community, building, or enterprise. For each site s , we assume access to heterogeneous observations

$$\mathcal{X}_s = \{I_s, M_s, U_s, T_s\},$$

where I_s denotes geospatial visual context (street-view or remote-sensing images and optional POI snapshots), M_s denotes map and network layers (coverage, capacity, backhaul, and engineering resource indicators), U_s denotes structured customer and service signals (product portfolio, usage, and contract status), and T_s denotes unstructured operational text (tickets, complaints, and visit notes).

The goal is to generate three outputs for each site

$$\mathcal{Y}_s = \{r_s, b_s, a_s\},$$

where r_s is a prospect priority score for ranking, b_s is a feasible service bundle, and a_s is a set of workflow artifacts such as talk tracks and proposal drafts, accompanied by evidence citations.

We define a bundle as a set of product items from a catalog $\mathcal{P}$

$$b_s \subseteq \mathcal{P}, \quad |b_s| \leq K,$$

where $\mathcal{P}$ covers broadband, wireless coverage solutions, industrial devices or IoT, and cloud services. A bundle must satisfy deliverability constraints based on local coverage and resources.

3.1 Overview

StreetCopilot is a grounded vision-language geospatial agent with four stages: evidence retrieval over region knowledge and operational logs, multimodal fusion to build a site state

representation, deliverability-aware bundle generation with constraint verification, and citation-grounded artifact generation.

Given site s , the system first retrieves a compact evidence set $\mathcal{E}_s$ from a knowledge base $\mathcal{K}$

$$\mathcal{E}_s = \text{Retrieve}(\mathcal{K}, q_s),$$

where q_s is a query composed from site identifiers, local network summaries, and customer context. It then constructs a unified site state h_s by fusing visual, structured, and textual information, and generates a prospect score r_s and a candidate bundle $\hat{b}_s$. A constraint verifier transforms $\hat{b}_s$ into a feasible bundle b_s . Finally, a citation-grounded generator produces artifacts a_s with evidence pointers into $\mathcal{E}_s$.

3.2 Evidence Retrieval and Site State Encoding

Retrieval. We index the knowledge base $\mathcal{K}$ using dense retrieval. Each record $k \in \mathcal{K}$ is embedded as $z_k = f_{\text{emb}}(k)$, and the query q_s is embedded as $z_{q_s} = f_{\text{emb}}(q_s)$. We select top- L evidence snippets by inner product

$$\mathcal{E}_s = \text{TopL}(\langle z_{q_s}, z_k \rangle).$$

In practice, $\mathcal{K}$ includes product rules, SLA templates, engineering constraints, historical cases, and anonymized operational notes.

Multimodal site representation. We encode visual context I_s with an image encoder f_{img} , structured features (M_s, U_s) with an MLP f_{tab} , and text $(T_s, \mathcal{E}_s)$ with a language encoder f_{txt}

$$\begin{aligned} v_s &= f_{\text{img}}(I_s), \\ u_s &= f_{\text{tab}}([M_s; U_s]), \\ t_s &= f_{\text{txt}}([T_s; \mathcal{E}_s]). \end{aligned}$$

We fuse them into a site state vector h_s via gated fusion

$$\begin{aligned} g_s &= \sigma(W_g[v_s; u_s; t_s]), \\ h_s &= W_h[g_s \odot [v_s; u_s; t_s]], \end{aligned}$$

where σ is the sigmoid function and $\odot$ is element-wise product.

3.3 Prospect Prioritization

We compute a prospect score r_s used for ranking sites

$$r_s = \text{sigmoid}(w^\top h_s).$$

When supervised ranking labels are available, we optimize a pairwise ranking loss. Given a pair (s_i, s_j) where s_i should rank above s_j , we use

$$\mathcal{L}_{\text{rank}} = \log \left(1 + \exp(r_{s_j} - r_{s_i}) \right).$$

If only weak signals exist, such as historical acceptance or conversion, we treat them as noisy labels and apply the same objective.

3.4 Deliverability-Aware Bundle Generation

Candidate Bundle Proposal. We model bundle generation as selecting up to K items from $\mathcal{P}$. StreetCopilot uses a language model conditioned on $\mathcal{E}_s$ and the site summary to produce an initial bundle $\hat{b}_s$ and a rationale draft

$$(\hat{b}_s, \hat{a}_s) = \text{LLM}_\theta(\text{Prompt}(\mathcal{X}_s, \mathcal{E}_s)).$$

To avoid uncontrolled generation, we constrain the bundle to a canonical schema with fixed slots, for example access, wireless, edge or IoT, cloud, and service guarantees, so the LLM outputs a structured object.

Constraint Verifier. A bundle must satisfy local feasibility constraints based on coverage and resources. We define a constraint set $\mathcal{C}$, where each product item $p \in \mathcal{P}$ has a predicate $\mathcal{C}_p(\mathcal{X}_s)$ that evaluates deliverability at site s . The verifier computes feasibility for each item

$$\phi_s(p) = \mathcal{C}_p(\mathcal{X}_s) \in \{0,1\}.$$

The final feasible bundle is obtained by filtering and repairing

$$b_s = \text{Repair}(\hat{b}_s, \{\phi_s(p)\}_{p \in \hat{b}_s}, \mathcal{E}_s).$$

Repair uses a prioritized rule set, for example replacing an indoor solution with a fallback option if indoor coverage is insufficient, or downgrading a bandwidth tier if backhaul constraints fail. We implement $\text{Repair}(\cdot)$ as a lightweight rule engine plus an LLM rewrite that is only allowed to pick from a feasible candidate list.

Feasibility Objective. For training-time evaluation and ablations, we define a feasibility score

$$F(s, b) = \frac{1}{|b|} \sum_{p \in b} \phi_s(p),$$

which equals 1 if all items are deliverable.

3.5 Citation-Grounded Artifact Generation

Evidence-Faithful Generation. Given the feasible bundle b_s , we generate artifacts a_s such as talk tracks and proposal drafts. We require each key claim to be supported by at least one evidence snippet from $\mathcal{E}_s$. We represent evidence snippets as spans with identifiers, and instruct the generator to attach citations as a list of snippet ids. We enforce citation grounding using an evidence checker that verifies coverage

$$G(a_s, \mathcal{E}_s) = \frac{1}{|\mathcal{Q}(a_s)|} \sum_{q \in \mathcal{Q}(a_s)} \mathbb{I}(\exists e \in \mathcal{E}_s: \text{Support}(q, e) = 1),$$

where $\mathcal{Q}(a_s)$ extracts atomic claims from a_s and $\text{Support}(\cdot)$ is a learned or rule-based entailment checker.

Controlled prompting. We use a two-pass generation to reduce hallucinations. In the first pass, the LLM produces a structured plan with explicit references to evidence ids. In the second pass, it verbalizes the plan into natural language while preserving citations and forbidding new factual claims.

3.6 Training Objectives

StreetCopilot contains components that can be trained jointly or separately. We optimize a mixture of objectives

$$\mathcal{L} = \lambda_1 \mathcal{L}_{\text{rank}} + \lambda_2 \mathcal{L}_{\text{bundle}} + \lambda_3 \mathcal{L}_{\text{faith}},$$

where $\mathcal{L}_{\text{bundle}}$ supervises bundle selection when ground-truth bundles exist, and $\mathcal{L}_{\text{faith}}$ trains the evidence checker or reranker for grounding.

For bundle supervision, we treat bundle prediction as multi-label classification over $\mathcal{P}$

$$\mathcal{L}_{\text{bundle}} = - \sum_{p \in \mathcal{P}} [y_{s,p} \log \hat{y}_{s,p} + (1 - y_{s,p}) \log(1 - \hat{y}_{s,p})],$$

where $y_{s,p} = 1$ if item p is included in the target bundle.

For faithfulness, we train $\text{Support}(q, e)$ with contrastive pairs (q, e^+) and (q, e^-)

$$\mathcal{L}_{\text{faith}} = -\log \frac{\exp(\text{sim}(q, e^+)/\tau)}{\exp(\text{sim}(q, e^+)/\tau) + \sum_{e^-} \exp(\text{sim}(q, e^-)/\tau)},$$

where $\text{sim}(\cdot)$ is cosine similarity and τ is a temperature parameter. Then, Algorithm [alg:streetcopilot] summarizes the inference pipeline.

Site observations $\mathcal{X}_s = \{I_s, M_s, U_s, T_s\}$, knowledge base $\mathcal{K}$, product catalog $\mathcal{P}$ Prospect score r_s , feasible bundle b_s , artifacts a_s Build query q_s from (M_s, U_s, T_s) and site metadata Retrieve evidence $\mathcal{E}_s \leftarrow \text{Retrieve}(\mathcal{K}, q_s)$ Encode and fuse $h_s \leftarrow \text{Fuse}(I_s, M_s, U_s, T_s, \mathcal{E}_s)$ Prospect scoring $r_s \leftarrow \text{sigmoid}(w^\top h_s)$ Candidate generation $(\hat{b}_s, \hat{a}_s) \leftarrow \text{LLM}_\theta(\text{Prompt}(\mathcal{X}_s, \mathcal{E}_s))$ Feasibility check $\phi_s(p) \leftarrow C_p(\mathcal{X}_s)$ for each $p \in \hat{b}_s$ Bundle repair $b_s \leftarrow \text{Repair}(\hat{b}_s, \{\phi_s(p)\}, \mathcal{E}_s)$ Artifact generation $a_s \leftarrow \text{GenerateWithCitations}(b_s, \mathcal{X}_s, \mathcal{E}_s)$ Optional validation: drop or revise claims with low support using $G(a_s, \mathcal{E}_s)$ r_s, b_s, a_s

Discussion and complexity. StreetCopilot decouples decision making into retrieval, fusion, constrained bundle reasoning, and evidence-grounded generation. The retrieval stage is sublinear with ANN indexing, and fusion is linear in evidence length. Constraint verification is lightweight since it evaluates a small set of predicates for each item in $\hat{b}_s$. The generation stage dominates latency and can be controlled by restricting evidence length L and limiting the number of refinement passes. In our implementation, we use a fixed schema for bundle outputs and a bounded evidence budget to maintain stable inference time.

4 Experiments

4.1 Experimental Setup

Task and outputs. We evaluate StreetCopilot on street-scale telecom GTM within a target subdistrict. We define a *site* as an operationally meaningful unit for field marketing and delivery, including (i) a residential community (compound-level), (ii) a commercial building (tower-level), and (iii) an enterprise premise within an industrial park. Given multimodal observations for each

site s , the system outputs: (1) a prospect priority score r_s for ranking, (2) a feasible service bundle b_s (broadband, wireless coverage solutions, industrial devices or IoT, and cloud services), and (3) workflow artifacts a_s such as talk tracks and proposal drafts with evidence citations.

Dataset. We construct a subdistrict-scale dataset consisting of several thousand sites. For each site, we collect four types of signals: For each site, we collect four types of signals: (1) geospatial visual context I_s (street-view imagery and/or remote-sensing patches, optionally POI snapshots), (2) network and map layers M_s (coverage and capacity summaries, traffic KPIs, backhaul indicators, and engineering resource flags), (3) structured customer/service signals U_s (existing portfolio, usage, contract status when available), and (4) unstructured operational text T_s (tickets, complaints, and visit notes). All sensitive fields are anonymized and personally identifiable information is removed.

Data split. To reduce spatial leakage, we partition the region into non-overlapping geographic blocks and split blocks into train/val/test with a ratio of 0.7/0.2/0.1. All reported results are on the held-out test blocks.

Supervision signals. For prospect ranking, we derive labels from historical outcomes such as lead acceptance and conversion, and use proxy labels (e.g., follow-up completion) when conversion is unavailable. For bundle recommendation, we use accepted bundles when available; otherwise we use weak supervision from product adoption logs and engineer-confirmed feasibility records.

Metrics. For prospect prioritization, we report AUC, NDCG@10, and Recall@10. For bundle recommendation, we report Bundle-F1 and Feasibility Rate (FR), where FR is the fraction of predicted bundles whose items all satisfy deliverability constraints. For artifact generation, we report Evidence Faithfulness (EF), defined as the fraction of extracted atomic claims supported by retrieved evidence. We additionally conduct a human evaluation on a subset of test sites for talk tracks and proposal drafts.

4.2 Deliverability Constraints

Constraint predicates. A bundle item p is deliverable at site s if the predicate $C_p(\mathcal{X}_s) = 1$. We implement five families of deliverability predicates:

- **Coverage.** Wireless solutions require coverage above thresholds. For example, indoor coverage enhancement requires $\text{RSRP}_{\text{in}}(s) \geq \tau_{\text{in}}$ and outdoor enhancement requires $\text{RSRP}_{\text{out}}(s) \geq \tau_{\text{out}}$.
- **Capacity.** High-tier services require sufficient spare capacity, e.g., $\text{Cap}(s) - \text{Load}(s) \geq \tau_{\text{cap}}$.
- **Backhaul and fiber.** Bandwidth tier B requires $\text{BH}(s) \geq B$ and $\text{Fiber}(s) = 1$ when fiber delivery is needed.
- **Engineering resources.** Installation-heavy items require resource feasibility such as crew availability and permissions, e.g., $\text{Crew}(s) \geq \tau_{\text{crew}}$ and $\text{Permit}(s) = 1$.
- **Serviceability.** If unresolved severe faults are present, aggressive upsell is gated, e.g., $\text{SevTickets}(s) \leq \tau_{\text{sev}}$.

Repair policy. If an item is infeasible, we apply prioritized repair: (1) replace it with a feasible alternative in the same category, (2) downgrade tier (e.g., bandwidth or SLA), or (3) remove it if no feasible substitute exists. This ensures that the final bundle is executable under local constraints.

4.3 Baselines

Prospect ranking baselines. We include: (1) **XGBoost**(Chen and Guestrin 2016) using structured features (M_s, U_s) , (2) **MLP**(Tolstikhin et al. 2021) using structured features, (3) **Geo-only** model using (I_s, M_s) without operational text, and (4) **Text-only** model using (T_s) and tickets.

Bundle and artifact baselines. We include: (1) **Rule-based** bundle selection using handcrafted mappings from signals to products, (2) **LLM prompting** without retrieval and without verification, (3) **LLM+RAG** without verification, (4) **LLM+Verifier** without retrieval, (5) **LLM+RAG+Verifier** as a strong baseline close to our system but without explicit citation-faithfulness enforcement.

4.4 Implementation Details

Encoders. We use a vision encoder (Dosovitskiy et al. 2020) for geospatial imagery, a tabular MLP for structured signals, and a text encoder for operational logs and retrieved evidence. Multimodal features are fused using gated fusion.

Retrieval. We build a knowledge base containing product rules, SLA templates, engineering constraints, and anonymized historical cases. We embed all snippets with a dense embedding model and use ANN search to retrieve the top- L snippets per site, with $L = 12$. We cap total evidence tokens to 1,536 to control latency.

LLM prompting. Bundle generation is constrained to a fixed schema with at most $K = 5$ items. Artifact generation uses a two-pass strategy: the first pass outputs a structured plan with evidence ids, and the second pass verbalizes the plan while preserving citations and forbidding new unsupported factual claims.

Training. We train the prospect scoring head with pairwise ranking loss (Durmus et al. 2024) and optionally train bundle selection with multi-label supervision when ground-truth bundles exist. We tune hyperparameters on the validation set and train with AdamW using learning rate 1×10^{-4} , batch size 64, for 20 epochs.

4.5 Comparison with State of the Art

Comparison with baselines on street-scale GTM. Higher is better for all metrics. FR denotes feasibility rate and EF denotes evidence faithfulness.

Method	NDCG@ 10	Recall@ 10	AU C	Bundl e-F1	FR	EF
Structured-only (XGBoost)	0.312	0.425	0.71 2	0.284	0.82 1	0.45 0
Structured-only (MLP)	0.325	0.438	0.72 4	0.291	0.81 5	0.46 2
Geo-only (Vision/Map)	0.284	0.389	0.68 5	0.256	0.75 0	0.38 0
Text-only (Tickets)	0.358	0.482	0.74 1	0.312	0.84 4	0.51 2
Rule-based bundle	0.215	0.310	0.61 0	0.245	0.68 0	0.55 0
LLM Prompting (no verify)	0.412	0.545	0.79 8	0.385	0.89 2	0.62 4
LLM+RAG (no verify)	0.478	0.612	0.84 5	0.442	0.91 5	0.78 5
LLM+Verifier (no RAG)	0.445	0.584	0.82 1	0.410	0.94 2	0.83 0
LLM+RAG+Ver ifier	0.512	0.654	0.87 6	0.498	0.96 1	0.87 5
StreetCopilot (ours)	0.547	0.698	0.89 4	0.537	0.97 4	0.89 2

Results. Table 1 summarizes the main results. StreetCopilot improves prospect ranking and bundle recommendation quality while maintaining high feasibility and evidence faithfulness. We observe that **StreetCopilot** consistently outperforms all baselines

across all six metrics. Specifically, it achieves a significant improvement in (0.547) and (0.698), outperforming the strongest LLM-based baseline (LLM+RAG+Verifier) by approximately 6.8% and 6.7%, respectively.

Effectiveness of Verification. Compared to LLM-based baselines without verification, the feasibility rate (FR) of StreetCopilot increases to 0.974, indicating that deliverability-aware checking and repair are essential for generating executable recommendations. Baselines lacking this module often suffer from spatial hallucinations, such as suggesting unreachable locations or logically inconsistent routes.

Impact of RAG and Grounding. Compared to retrieval-only drafting, StreetCopilot yields higher evidence faithfulness (EF = 0.892) due to citation-grounded generation and post-hoc support checking. While *LLM+RAG (no verify)* improves the recall of relevant street entities, our full framework ensures that these entities are correctly contextualized within the recommendation evidence.

Comparison with Traditional Baselines. It is also worth noting that structured-only (e.g., XGBoost, MLP) and geo-only models struggle with the complex semantic reasoning required for GTM. StreetCopilot bridges this gap by integrating multi-modal street data with LLM’s reasoning capabilities, leading to a substantial gain in (0.537) over the rule-based approach (0.245).

4.6 Ablation Studies

Ablations of StreetCopilot. We remove or replace key components. Higher is better for all metrics.

Variant	NDCG@10	Bundle-F1	FR	EF
Full StreetCopilot	0.547	0.537	0.974	0.892
w/o geospatial vision (I_s)	0.512	0.495	0.941	0.881

Variant	NDCG@10	Bundle-F1	FR	EF
w/o operational text (T_s)	0.485	0.462	0.958	0.854
w/o retrieval ($\mathcal{E}_s$)	0.478	0.442	0.915	0.785
w/o constraint verification	0.532	0.501	0.886	0.864
w/o citation grounding	0.539	0.518	0.962	0.725
Rule-only bundle (no LLM)	0.215	0.245	0.680	0.550

Analysis. We conduct an ablation study to quantify the contribution of each component in StreetCopilot. The results in Table 2 yield several key insights:

Modality Contributions. Removing geospatial vision (I_s) typically degrades ranking and bundle relevance (NDCG drops to 0.512), especially in mixed-use zones where imagery captures building density and site typology. Removing operational text (T_s) hurts performance on sites with frequent service issues, where tickets provide strong signals for prioritization and feasible recommendations, leading to a significant decrease in Bundle-F1 (0.462).

Knowledge and Grounding. Disabling retrieval ($\mathcal{E}_s$) reduces evidence faithfulness (EF = 0.785) and increases unsupported claims, confirming that an LLM’s internal knowledge is insufficient for street-scale GTM. Furthermore, removing citation grounding specifically targets the transparency of the model; while the NDCG remains relatively stable, the EF score plunges from 0.892 to 0.725, proving that explicit grounding is the primary driver of trustworthy recommendations.

System Reliability. Most importantly, removing constraint verification significantly reduces the feasibility rate (FR) from 0.974 to 0.886. This validates our hypothesis that generative recommendations alone are not reliably executable in constrained urban environments. Finally, the Rule-only baseline (no LLM)

performs poorly across all metrics, emphasizing that while rules provide a safety net, the reasoning and synthesis capabilities of the LLM are what drive high-quality GTM outcomes.

Efficiency (Latency) and Human Evaluation results. Human scores are on a 5-point Likert scale (higher is better). Latency is measured in seconds per site.

Method	Latency (s)	Helpfulness	Clarity	Actionability
LLM Prompting (no RAG)	3.2 ± 0.4	3.12	3.45	2.84
LLM+RAG (no verify)	5.8 ± 0.8	3.85	3.72	3.51
LLM+Verifier (no RAG)	4.5 ± 0.6	3.42	3.81	4.15
StreetCopilot (ours)	7.2 ± 1.2	4.58	4.62	4.75

4.7 Efficiency and Human Evaluation

Latency. We report end-to-end inference latency under a fixed evidence budget and bundle schema. As shown in Table 3, **StreetCopilot** has an average latency of 7.2 seconds per site. While this is higher than vanilla prompting due to the multi-step retrieval and verification stages, the overhead remains well within the acceptable range for offline recommendation and real-time decision support. Retrieval and verification are lightweight (accounting for <15% of total time), while generation dominates runtime. Capping evidence length and using two-pass generation yields stable latency without sacrificing evidence faithfulness.

Human evaluation. To further validate the real-world utility, we sample **50** representative test sites and ask **3** domain experts (senior urban planners and logistics operators) to rate talk tracks and proposal drafts on helpfulness, clarity, and actionability using

a five-point Likert scale. StreetCopilot achieves the highest scores across all human metrics, particularly in **Actionability (4.75/5.0)**. Experts noted that while LLM+RAG baselines often provide informative descriptions, they frequently fail to propose "executable" steps. StreetCopilot is preferred over prompting and retrieval-only baselines, largely due to feasibility-aware suggestions (from our Verifier) and evidence-backed rationales (from our Citation Grounding), which significantly reduce the manual effort required for proposal verification.

Sensitivity analysis on repair strictness (k). Higher k implies more aggressive verification/repair.

Strictness Level	FR $\uparrow$	NDCG@10 $\downarrow$	Bundle-F1 $\downarrow$	Latency (s)
Low ($k = 1$)	0.892	0.551	0.542	6.1
Medium ($k = 3$)	0.954	0.548	0.539	6.8
High ($k = 5$)	0.974	0.547	0.537	7.2

4.8 Additional Analysis

Feasibility versus relevance trade-off. We vary the repair strictness (the number of iterative feedback loops k) and candidate pool size to analyze the impact on performance. As shown in Table 4, we observe that stricter verification significantly improves the **Feasibility Rate (FR)** from 0.892 to 0.974, with only a marginal reduction in **NDCG@10** (from 0.551 to 0.547). This modest impact on bundle relevance suggests that our constrained generation and repair mechanism is highly practical for deployment: it prioritizes the "executable" nature of recommendations without sacrificing the overall quality of the top-ranked prospects.

Error analysis. Despite the gains, we identify three primary failure modes through a manual audit of low-performing cases:

- **(1) Sparse or noisy operational logs:** In remote or newly developed areas, sparse service tickets lead to weak retrieval evidence, forcing the LLM to rely more on general priors which may lack site-specific nuances.
- **(2) Ambiguous site boundaries:** Overlapping geo-fences or multi-tenant building complexes occasionally cause mixed signals, where the model attributes maintenance issues from one facility to a neighboring one.
- **(3) Dynamic resource volatility:** Rapidly changing engineering resource availability or temporary road closures are not always captured in our static features, occasionally leading to "feasible-on-paper" but practically delayed recommendations.

These cases motivate future work on **real-time telemetry integration** and more robust **hierarchical geospatial site normalization**.

5 Conclusion

This paper studies street-scale telecom go-to-market as a grounded multimodal reasoning problem that requires prospect prioritization, deliverable bundle recommendation, and workflow-aligned artifact generation. We present StreetCopilot, a vision-language geospatial agent that integrates geospatial imagery and POIs, network and map layers, structured service signals, and operational logs through retrieval-augmented fusion. To make recommendations executable and auditable, StreetCopilot incorporates deliverability-aware constraint verification with rule-based repair, and generates talk tracks and proposal drafts with citation-grounded evidence support. Extensive experiments demonstrate that StreetCopilot improves prospect ranking quality, bundle relevance under deliverability constraints, and evidence faithfulness, while maintaining stable latency suitable for practical deployment. These results suggest that combining multimodal

grounding with explicit feasibility constraints is a promising direction for building reliable decision-support agents in locality-sensitive, high-stakes operational domains.

References

1. Afzal, Mohammad, Shanur Rahman, Disha Singh, and Ale Imran. 2024. "Cross-Sector Application of Machine Learning in Telecommunications: Enhancing Customer Retention Through Comparative Analysis of Ensemble Methods." *IEEE Access* 12: 115256–67.
2. Chen, Tianqi, and Carlos Guestrin. 2016. "Xgboost: A Scalable Tree Boosting System." *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 785–94.
3. Cheng, Zhili, Yuge Tu, Ran Li, et al. 2025. "Embodiedeval: Evaluate Multimodal LLMs as Embodied Agents." *arXiv Preprint arXiv:2501.11858*.
4. Cheshmehzangi, Ali. 2026. *Generative AI-Powered Urban Digital Twins: Pioneering Environmental Solutions for Sustainable Intelligent Cities*. Springer.
5. Dandamudi, Sai Ratna Prasad, Jaideep Sajja, and Amit Khanna. 2025. "Advancing Cybersecurity and Data Networking Through Machine Learning-Driven Prediction Models." *International Journal of Innovative Research in Computer Science and Technology* 13 (1): 26–33.
6. Dastagiraiah, D. 2024. "A System for Analysing Call Drop Dynamics in the Telecom Industry Using Machine Learning and Feature Selection." *Journal of Theoretical and Applied Information Technology* 102 (22).
7. Dosovitskiy, Alexey, Lucas Beyer, Alexander Kolesnikov, et al. 2020. "An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale." *arXiv Preprint arXiv:2010.11929*.

8. Durmus, Furkan, Hasan Saribas, Said Aldemir, Junyan Yang, and Hakan Cevikalp. 2024. "Pairwise Ranking Loss for Multi-Task Learning in Recommender Systems." *arXiv Preprint arXiv:2406.02163*.
9. Gehlot, Anita, Bushra Khatoon Ansari, Deepika Arora, Harishchander Anandaram, Bhagwant Singh, and José Luis Arias-González. 2022. "Application of Neural Network in the Prediction Models of Machine Learning Based Design." *2022 International Conference on Innovative Computing, Intelligent Communication and Smart Electrical Systems (ICSSES)*, 1–6.
10. Goswami, Kanika, Puneet Mathur, Ryan Rossi, and Franck Dernoncourt. 2025. "PlotGen: Multi-Agent LLM-Based Scientific Data Visualization via Multimodal Retrieval Feedback." *Companion Proceedings of the ACM on Web Conference 2025*, 1672–76.
11. Gu, Xiangming, Xiaosen Zheng, Tianyu Pang, et al. 2024. "Agent Smith: A Single Image Can Jailbreak One Million Multimodal Llm Agents Exponentially Fast." *arXiv Preprint arXiv:2402.08567*.
12. Hesham, Momen, Mohamed Essam, Mohamed Bahaa, et al. 2024. "Evaluating Predictive Models in Cybersecurity: A Comparative Analysis of Machine and Deep Learning Techniques for Threat Detection." *2024 Intelligent Methods, Systems, and Applications (IMSA)*, 33–38.
13. Koh, Jing Yu, Robert Lo, Lawrence Jang, et al. 2024. "Visualwebarena: Evaluating Multimodal Agents on Realistic Visual Web Tasks." *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 881–905.
14. Naveed, Humza, Asad Ullah Khan, Shi Qiu, et al. 2025. "A Comprehensive Overview of Large Language Models." *ACM Transactions on Intelligent Systems and Technology* 16 (5): 1–72.

15. Tang, Fengliang, Peng Zeng, Lei Wang, Longhao Zhang, and Weixing Xu. 2024. "Urban Perception Evaluation and Street Refinement Governance Supported by Street View Visual Elements Analysis." *Remote Sensing* 16 (19): 3661.
16. Tolstikhin, Ilya O, Neil Houlsby, Alexander Kolesnikov, et al. 2021. "Mlp-Mixer: An All-Mlp Architecture for Vision." *Advances in Neural Information Processing Systems* 34: 24261–72.
17. Wagh, Sharmila K, Aishwarya A Andhale, Kishor S Wagh, Jayshree R Pansare, Sarita P Ambadekar, and SH Gawande. 2024. "Customer Churn Prediction in Telecom Sector Using Machine Learning Techniques." *Results in Control and Optimization* 14: 100342.
18. Zhang, Jie, Xingtong Yu, Yuan Fang, Rudi Stouffs, and Zdravko Trivic. 2026. "UrbanGraphEmbeddings: Learning and Evaluating Spatially Grounded Multimodal Embeddings for Urban Science." *arXiv Preprint arXiv:2602.08342*.
19. Zhao, Haiyan, Hanjie Chen, Fan Yang, et al. 2024. "Explainability for Large Language Models: A Survey." *ACM Transactions on Intelligent Systems and Technology* 15 (2): 1–38.