Transferring the Long Tail: Composing Geometric and LLM-Based Stitching for Cross-Provider Map Migration

Shaishav M. Maisuria
Amazon
Seattle, WA, USA
smaisur@amazon.com

Vishal Kumar
Amazon
Seattle, WA, USA
vishku@amazon.com

Ravi Garg
Amazon
Seattle, WA, USA
ravigarg@amazon.com

***Abstract*—Base map migration across different map providers necessitates bringing the migrated map to parity with the pre-migration base map. This becomes critical when map application owners make their own geometric edits to the base map to improve quality for downstream applications. For example, last-mile transportation applications may add the final leg of roads leading to a shipping door, segments often missing in off-the-shelf mapping products. Such edits can be transferred to a new provider's base map using map conflation models and algorithms. While attribute edit migration has been studied in the literature, geometric edit migration remains largely unexplored. We present an LLM and machine-learning based system that uses a road correspondence ML model to identify equivalent roads between two maps, followed by a context-matching stitching model based on geometric rules. We demonstrate that high-precision stitching requirements lead to a loss of recall in the ML model, and we recover part of that recall-loss tail using multimodal LLM-based agent orchestration. The pipeline has four stages, three deterministic and one agentic. They are (1) a deduplication stage that removes roads already present in the new map, (2) the geometric stitcher, (3) a reachability pass that commits topologically easy connections without any model call, and (4) an LLM selection agent. The agent is given a stitching proposal tool that generates geometry-rule-based candidate connections, so that in the common case the LLM chooses among precomputed candidates rather than predicting coordinates, together with tools whose inputs include satellite imagery in addition to map data, and a validation tool that checks for gross geometric issues. We evaluate on two datasets. On a reproducible single-region open-data benchmark, an OpenStreetMap-to-Overture migration of $12{,}403$ source roads over $2{,}065$ km^2, the correspondence model detects 84.2% of the genuinely missing roads, the geometric stitcher connects 53.3% of them, and the reachability pass and LLM agent raise connected recall to 70.5%. On four regions of manually audited custom edits at a fixed 99% precision operating point, the stitcher alone reaches $\sim 74\%$ recall; deduplication lifts this to 80.5%, the deterministic reachability and re-admission passes to 88.3%, and the LLM agent, recovering the remaining tail, to 93%. Applied over independently-processable clusters of edits, this multi-step design enables scaled and incremental migration of bulk custom edits between base map providers while keeping overall compute and latency low by offloading the vast majority of edits to a lightweight ML model, with the LLM rescuing only the tail cases.**

***Index Terms*—map conflation, cross-provider migration, road stitching, large language model agents, geospatial data integration, last-mile logistics**

I. Introduction

Modern navigation systems are rarely powered by a single base-map provider. A navigation stack may combine open-source data such as OpenStreetMap [1] or the Overture Maps Foundation transportation layer [2] with commercial providers whose coverage, licensing terms, or refresh cadences better serve particular regions. Because providers digitize the world from different imagery at different times, no two road networks align perfectly: segment boundaries fall at different points, junction topologies diverge, and geometric fidelity tracks the vintage of the satellite capture. Such variation is uneven enough within a single provider to be an object of study in itself [3].

Over the lifetime of a deployment, operators accumulate a substantial layer of custom edits on top of their licensed base map: new subdivision roads, corrected driveways, restored service-road connectivity, and geometry that reflects ground truth the provider has not yet captured. In last-mile transportation this layer is operationally critical, because routing quality depends on the final leg to a delivery door, and routing at that scale is now studied on public industrial data [4]. When business, licensing, or coverage considerations motivate a switch to a different base map, these years of edits do not transfer automatically. The source map (the operator's existing map) and the target map (the new base map) differ significantly, so migration requires selectively replaying the operator's own additions onto the target network: transferring what the source has that the target lacks, without disturbing what the target already represents correctly.

Prior work on map conflation has focused predominantly on the attribute dimension of this problem, matching road segments across providers and migrating properties such as speed limits, turn restrictions, and functional class [5]–[7]. The geometric dimension, transferring road shapes that exist only in the source into the target network so that they are topologically connected and routable, has received far less attention despite being more consequential. A routing engine simply ignores any geometry that is not joined at valid junctions, and downstream consumers such as map matching

depend on a connected network [8], so an imprecise geometric stitcher produces floating fragments rather than usable roads.

Geometric edit migration is difficult for several compounding reasons. First, the candidate junction to which a new road must attach may lie beyond a conservative spatial search window, especially in areas of high geometric misalignment between providers. Second, some new roads must connect through other new roads that are themselves being migrated in the same batch, creating ordering dependencies. Third, local imagery can be ambiguous, making it unclear whether a candidate connection is valid without broader spatial context. A stitcher tuned for high precision of geometry stitching therefore trades recall for accuracy, leaving a long tail of valid roads behind: roads that pass the "is this genuinely new?" test yet fail the "where exactly do I attach it?" test.

This paper presents a hybrid system that closes that long tail by composing a fast, auditable geometric stitcher with a multimodal large-language-model (LLM) agent. A road-correspondence ML model first identifies source roads genuinely absent from the target; a context-matching stitcher attaches each to a compatible target junction within a fixed distance window; and the roads it cannot confidently connect pass to an LLM agent that reasons over satellite imagery and network context to select among code-precomputed candidate connections. Grounded vision-language models make reasoning over overhead imagery feasible enough to build on [9], yet general-purpose foundation models still trail task-specific ones on multimodal geospatial benchmarks [10], which is part of why we constrain what the model may decide rather than asking it to produce geometry. Its only unconstrained output is a vision-derived connection waypoint that a validation rule engine checks before acceptance, and failures are routed to human review rather than dropped, so a custom edit is either migrated or queued for an operator. On audited multi-region edits the automated stages reach 93% recall at 99% precision. Prior work has applied language models to the *matching* half of conflation [11]; to our knowledge this is the first reported system to place an LLM in the *geometry-transfer* half, recovering connectivity for roads a high-precision stitcher declined to commit.

We make the following contributions in this paper:

1) We propose a **two-tier pipeline** for cross-provider geometry migration of a base map that couples an ML-powered high-throughput road context-based stitcher with a multimodal LLM agent that operates exclusively on the stitcher's recall-loss data. The cheap tier is realized as three deterministic stages and the expensive tier as the agent. The tiers are connected by a well-defined handoff interface, and every migrated road segment is labeled with its origin stage (ML stitcher or LLM agent) to enable independent quality auditing and differentiated downstream review.
2) We propose a **layered candidate-selection contract** for embedding LLMs in precision-critical geospatial workflows. The LLM reasons over imagery, but in the common case it only chooses among precomputed, code-validated connection candidates and cannot influence geometry; a narrow third level lets it trace waypoints where no candidate fits. The invariant is that no geometry reaches the target map without passing deterministic validation, so the worst case of a bad sample is a rejected proposal, not a committed spurious road (Section IV-C). We describe the pattern for one setting and do not claim to have tested its transfer to others.
3) We define a reproducible open-data benchmark for geometric edit migration, an OpenStreetMap-to-Overture transfer over $2{,}065$ km^2 of Seattle whose target is built by rolling back the Overture segments their own citations attribute to the migrated edits. On it the correspondence model detects 84.2% of the missing roads, the geometric stitcher connects 53.3% and the composed pipeline 70.5%. On audited multi-region edits the stitcher achieves $\sim$$74\%$ recall at 99% precision and the composed pipeline reaches 93%, of which 4.7 points are the agent's. We report recall after each stage rather than one number, because a deterministic deduplication pre-filter accounts for part of the lift and only one of the four stages is agentic.

Section II reviews related work, Section III states the migration problem and the transfer diff, Section IV describes the system, Section V presents the benchmark, results and discussion, and Section VI concludes.

II. Related Work

Map conflation. Conflation, the matching and merging of features across maps, was formalized by Saalfeld [12] and surveyed by Ruiz et al. [5], who classify approaches by the geometric and topological cues they exploit. Score-based multi-stage matchers extend this to heterogeneous road patterns [6]. These methods merge whole maps rather than diffing two providers or recovering the roads that matching leaves unconnected.

Learning-based conflation. Recent systems make matching efficient or more accurate with learning and graph structure: MAYUR prunes candidates with early pruning and rank joins [13], Map Stitcher reconciles differences via graph sampling [14], KRAFT solves a merge as constrained optimization over a knowledge graph [15], and FlexMap Fusion conflates HD maps with OpenStreetMap [16]. Route-based conflation aligns linear-referencing-system maps to OpenStreetMap [7], a migration setting close to ours that transfers attributes where we transfer geometry, and high-precision conflation of sensor-observed traffic signs is a related industrial line [17]. All of these decide *whether* and *where* features correspond; none pairs a geometric stitcher with a downstream agent that recovers the geometry the stitcher discards.

Road extraction and repair from imagery. A separate line infers road graphs directly from overhead imagery: DeepRoadMapper repairs missing connections [18], RoadTracer grows a graph by iterative decoding [19], and Sat2Graph encodes it as a tensor [20]. This is the closest mechanism to our multi-round chaining, which likewise extends connectivity

hop by hop from imagery, but the task differs: these systems build a network from imagery alone, whereas we attach specific, already-authored source geometry into an existing target network at a fixed high-precision operating point and commit nothing the target's deterministic validators reject.

LLMs for conflation and spatial reasoning. Recent LLM work assesses a language model's ability to *match* embedded road lines [11], i.e. entity resolution rather than geometry recovery, and without a high-precision stitching stage in the loop. Our agent recovers connectivity for roads the geometric stitcher could not attach, selecting among code-validated candidates in the common case (Section IV-C sets out the three levels of action space), and Section V reports connected recall after each stage so that the agent's share is not conflated with the deterministic stages'.

LLM agents. Our agent follows the tool-augmented reasoning pattern [21], and its vote aggregation is a reject-only form of self-consistency across samples [22]. Prior conflation work has concentrated on attribute transfer or on centerline alignment; the geometric transfer of edits that have no target counterpart, and its recovery at a fixed high-precision operating point, is the gap this paper fills.

III. Problem: Cross-Provider Migration and the Transfer Diff

Setting. Let S be a source map carrying a provider's custom road edits and T a target base map to which the provider is migrating. The two were edited independently, so they disagree, and migration is the selective transfer of source-only content into T driven by the diff between them rather than a full merge.

The transfer diff. The *semantic* diff $D^*(S,T)$ is the set of source roads with no real counterpart in T. A matching step approximates it by an *acceptable counterpart* test, aligning a source road to a target road when their geometries agree within a road-class-aware buffer and their endpoints and heading correspond. The computed diff $\hat{D}(S,T)$ is fallible: it contains source roads flagged missing that are present under a different digitization. Migration must add D^* but observes only $\hat{D}$; roads that match feed the attribute-transfer half and are out of scope here.

Geometry transfer (stitching). Write $M = \hat{D}(S,T) \setminus P$ for the *stitchable set*, the computed diff after deduplication (Section IV-B) has removed the subset P it determines are already present. For each $r \in M$, a *correct integration* joins r into T's network at the correct junctions so that r is routable; an integration that cannot be made confidently is withheld.

The residual tail. The ML stitcher partitions M into $A \cup R$, where A is the set it integrates at the fixed precision operating point, and R is the *recall-loss* residual, holding back roads whose junction lies beyond its connection window, whose connection depends on another road in R, or whose local geometry is ambiguous. On the audited-edit regions at $p^* \approx 0.99$ precision the stitcher recovers 74.1% of the whole audited diff, P included; on the open benchmark of Section V, which divides by M as defined above, R is far larger. Either way R is the long tail this paper targets.

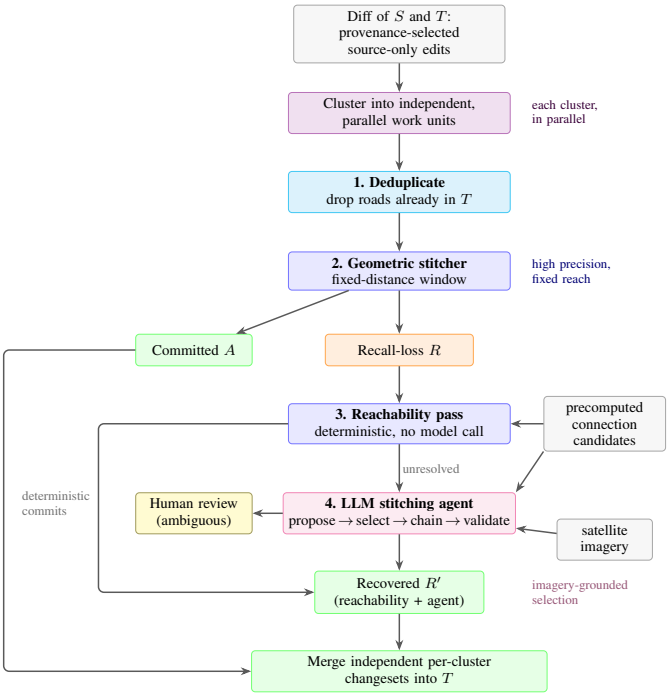


Fig. 1. Composed geometry-transfer pipeline. Provenance-selected source-only edits are grouped into independent, parallel clusters. Within each cluster the roads are deduplicated, then the geometric stitcher commits what it can connect (A) and withholds a recall-loss tail (R); a deterministic reachability pass resolves the topologically easy part of R without a model call, and the LLM agent recovers roads from what remains. Clusters produce independent changesets that merge into T; ambiguous cases go to human review.

Objective. Given R and the local target network, recover a subset $R' \subseteq R$ of correct integrations maximizing recall $(|A|+|R'|)/|M|$ while holding precision at p^*. Because $|M|$ is the denominator, improving deduplication and improving stitching are not interchangeable: the first changes which roads are in scope, the second how many in-scope roads are connected, and Section V reports them separately.

IV. Composed Geometry-Transfer System

Given a source map S and a target base map T, a diff identifies source-only roads, and geometry transfer must connect each genuine one into T's network. Our system realizes this as a *four*-stage composition (Fig. 1), each stage more expensive and less constrained than the one before it and invoked only on what the previous stage could not resolve. First, *deduplication* removes candidates that are in fact already present in the target, so only genuinely missing roads flow downstream. Second, a machine-learned *geometric stitcher* attaches the majority of those roads at high precision and withholds the ones it cannot confidently connect to a recall-loss tail. Third, a cheap *deterministic reachability pass* resolves the tail roads that are already topologically close to the network, committing them without invoking any model. Fourth, an *LLM agent*

recovers roads from what remains. A small residual that none of them resolves confidently is routed to human review.

A. Candidate identification and clustering

Selecting the operator's own edits. We first keep only the roads an operator actually added or changed. Every element carries provenance markers recording which editor authored its attributes and its geometry, and operator edits fall in a reserved identifier band, so we select exactly those edits and leave the rest of the base map alone: we replay the operator's own work, not the whole map.

Clustering into independent work units. We then build a graph whose nodes are map elements and whose edges join elements sharing a vertex or a membership; its connected components become the clusters, oversized components are split by road class and size, and each cluster gets a bounding box. Because the graph covers only what changed between two consecutive source versions and components share no state, clusters run in parallel and a refresh revisits only the roads that moved.

B. Deduplication and geometric stitching

Road correspondence and removing already-present roads. A diff also flags roads that are already in the target but drawn differently, and committing a second copy corrupts the network. The road-correspondence model removes these before any stitching, using three deterministic tests. The body-coverage test asks whether at least half the source length falls within a road-class-aware buffer of existing target roads, excluding non-road features such as building footprints. The routable-duplicate test asks whether the target network can already route between the road's endpoints along a path that tracks the road rather than detouring. A geometric flag may be confirmed by an advisory vision check asking whether the two lines are the same real-world road; an unconfirmed flag falls back to the geometric decision, so the stage's verdict is deterministic either way. A third test catches roads running a few meters to the side of an existing one and roughly parallel to it, usually the same road digitized twice, and routes them to human review rather than dropping them. Roads flagged by any test are recorded as already-present and excluded.

Context-matching geometric stitcher. The genuinely missing roads enter a context-matching geometric stitcher, which attaches each to a compatible target junction within a fixed distance window, or withholds it to a recall-loss set if none is found (Fig. 2(a)): because the true junction is farther than the window, because the connection depends on another not-yet-stitched road, or because the local geometry is ambiguous. On the audited-edit regions, tuned to 99% precision, this stage recovers $\sim 74\%$ of the audited diff (already-present roads included); on the open benchmark of Section V a substantial share is withheld. In both cases the recall-loss set is the long tail the LLM agent targets. We describe this stage by its inputs, outputs, and operating point; its model internals are orthogonal to the agentic recovery that is this paper's focus.

Algorithm 1: Geometry-free candidate selection for one recall-loss road

Input: road $r \in R$; local target network G_T; optional overlay I
Output: decision $d(r) \in \{\text{ACCEPT}(c), \text{REJECT}, \text{HOLD}\}$

```
C ← BUILDCANDIDATES(r, G_T)      // junction snaps + imagery-inferred connections
foreach c ∈ C do
    cache c by id; expose signals σ(c) = {snap dist, angle, class-compat, name-sim, is-junction}
    withhold raw geometry/coordinates of c from the model
if C = ∅ then
    W ← LLM-WAYPOINTS(I)      // Level 3: only path where the model shapes geometry
    return VALIDATE(W)
        // building/length/snap/topology checks, else reject
ĉ ← LLM({σ(c)}_{c∈C}, I)      // Levels 1--2: model returns a candidate id
if ĉ ∉ cache then
    return REJECT          // fabricated id rejected
else if VOTE(ĉ) downgrades then
    return REJECT or HOLD      // reject-only over model samples
else
    return ACCEPT(ĉ) with cached geometry
```

A deterministic reachability pass. Not every road in the tail needs the model. Before the agent runs, a cheaper pass groups the withheld roads by their graph distance from a road that is already connected and resolves the reachable ones itself. It selects from the same precomputed, code-validated candidates the agent later chooses among (Section IV-C) and commits one only when it is anchored on already-connected geometry, short, and class-compatible: the forced picks, where a single trivially valid option exists. It then walks shortest paths outward to pull in chains of missing roads that hang off an already-connected end. This is the deterministic answer to the ordering dependency of Section I: these roads failed the stitcher not because their connections are hard but because, judged one road at a time, the road they connect through was not yet part of the network. The pass applies the same rules in dependency order, adding sequencing rather than tolerance, and it differs from the agent's multi-round chaining in what a hop requires: here every hop is an existing trivial candidate, whereas the agent chains through hops whose geometry must first be inferred from imagery. Only the roads it cannot settle go on to the LLM.

Algorithm 2: Multi-round chain stitching over the residual R

Input: residual roads R; target network G_T; max rounds K
Output: recovered set $R' \subseteq R$
$R' \leftarrow \emptyset$; $F \leftarrow$ CONNECTEDENDPOINTS(G_T) // reachable frontier
for $k \leftarrow 1$ **to** K **do**
  progress $\leftarrow$ FALSE
  foreach $r \in R \setminus R'$ *with a candidate anchored on* F **do**
    $d \leftarrow$ SELECTCANDIDATE(r, $G_T \cup R'$) // Alg. 1
    if $d =$ ACCEPT(c) **then**
      $R' \leftarrow R' \cup \{r\}$; add r's endpoints to F;
      progress $\leftarrow$ TRUE
  if $\neg$ *progress* **then break**
return R'

C. The LLM stitching agent

The agent recovers roads from the recall-loss tail. Because fabricated geometry is hard to recover from, deterministic code precomputes a small set of candidate connections for each withheld road, and the LLM's job is to *select* among them rather than to draw or copy geometry.

What the agent receives, and what it may return. For one withheld road the agent receives an annotated aerial overlay, the nearby target roads and obstacles within a fixed radius, and a table of precomputed candidates with non-geometric match signals only: connection distance, incidence angle, road-class compatibility, name similarity, and whether the target point is a junction. It never sees raw coordinates. It returns one candidate identifier, and the tool rejects any identifier absent from the cache, so a fabricated connection cannot enter the output (Algorithm 1).

Candidate-selection contract, in three levels. The contract is not uniform, so we state the gradation rather than claim the model never originates geometry. The proposal tool generates candidates from geometric rules together with the local imagery, each cached under a stable identifier. *Level 1, select-only:* the candidate snaps the road's endpoint to a compatible junction and the agent returns an identifier, so it cannot influence geometry. *Level 2, constrained-propose:* where the connection is not a direct snap the candidate carries a connecting segment whose shape the tool inferred from imagery under geometric rules, so the shape originates in the tool rather than in free generation. *Level 3, free-propose:* where no candidate fits, the agent traces sparse waypoints along a drivable path in the imagery.

Level 3 is the only path on which the model's own output determines shape, and multi-round chaining also runs on it. The invariant that holds across all three levels is therefore narrower than "the LLM never emits geometry". It is that *no geometry reaches the target map without passing deterministic validation*: building-crossing, length, snap, and topology checks run on every Level-3 output before it is materialized. The worst case of a bad sample is a rejected proposal, not a committed spurious road. How much of the recovered recall depends on this least constrained level is not instrumented separately in these runs, which we note as a limitation for anyone adopting the pattern.

Multi-round chaining. Some missing roads connect to the network only through an intermediate stretch that is itself absent from the target. For these the agent works over several rounds (Algorithm 2), inferring the connecting geometry hop by hop from the imagery and stitching inward from an already-connected endpoint until the road reaches the network. Each hop is validated before the next, and the road's source location is used only to fetch the relevant imagery, never as geometry to be copied.

Our claims are about the composition and the geometry-free contract rather than a specific model (Table V), and they improve as the underlying model improves.

D. Layered validation and output

Because a stitched connection is speculative, no single check gates it. Each proposed acceptance must survive a stack of independent validators, ordered cheap-to-expensive so most bad candidates are eliminated before costly steps run.

Pre-scoring and multi-signal agreement. Before the LLM is invoked, a deterministic scorer discards implausible candidates and flags borderline ones, so model reasoning is spent only where it can change the outcome. A candidate then becomes eligible only when several independent signals agree: connection distance, incidence angle, road-class compatibility, name similarity, junction presence, and visual confirmation in the imagery. No single signal can carry an acceptance.

Vote aggregation. The accept-or-reject decision is sampled several times independently and aggregated by majority, ties resolved conservatively. Aggregation over model samples is reject-only: it can downgrade an acceptance but never upgrade one, so sampling variance can only cost recall, never inject a wrong connection.

Two re-admission paths, and why they differ. The stack is not monotone, and the two places a held connection can be re-admitted are not equivalent. The geometric guardrail is deterministic: it blocks any connection whose materialized geometry crosses a building, divider, or other obstacle, and conversely re-admits a held connection when code proves the geometry clears every obstacle by a margin, countering a measured over-rejection mode without reopening the variance channel the vote closes. Visual reconsideration is different: for borderline rejects it re-renders the imagery at closer zoom and asks the model again, a second judgment rather than a proof. What holds uniformly is the guarantee of Section IV-C: no geometry reaches the target map without passing the deterministic validators, so a re-admission can restore a connection code accepts but never manufacture one code rejects. The obstacle tests read the target provider's own features, so grade separation is caught only when the target map records it.

Post-materialization audit and topology check. After a connection becomes committed geometry, a final audit rechecks it against the same obstacle tests and trims or drops any segment that fails, catching errors introduced when candidates are reconciled into one changeset. Only geometry that also connects into the routable network, leaving no floating fragment and no newly disconnected component, is committed.

Human review and provenance. Connections the validation stack holds as too ambiguous go to a human review queue rather than being committed or discarded. Accepted connections form a changeset that creates the new geometry and makes topology-only edits at the junctions where it splices in; where an approved road partly overlaps an existing one, the new parts are created, the coincident parts become topology-only edits, and the gap between them is bridged, so a half-existing road is merged rather than duplicated. Agent-authored output carries a provenance tag distinct from the geometric stage's, so downstream review can apply different policies to each.

V. Results and Discussion

We report two evaluations. The first is a *reproducible open-data benchmark* on an OpenStreetMap-to-Overture migration of the Seattle area, built entirely from published data so the numbers can be independently re-derived. The second reported conflation run replays manually audited custom edits onto four further OpenStreetMap regions, which is where we observe the effect of network density; it scores roads over the full audited diff, so the two evaluations are reported side by side as complementary views rather than one series. The open benchmark establishes the pipeline's baseline and shows the residual concentrating in a narrow confidence band, which is what the selective agent stage targets.

A. A reproducible open-data benchmark

Substrate. The source S is the 12,403 OpenStreetMap roads *created* during the edit window inside a 2,065 km^2 Seattle bounding box, in eight local road classes (residential, service, driveway, parking aisle, alley, unclassified, track, living street). The box runs from $(-122.45496,\ 47.48889)$ at lower left to $(-121.85863,\ 47.90477)$ at upper right, in WGS 84 lon/lat. The target T is the Overture transportation theme [2] (GeoParquet roads snapshot, filtered to the same box), 327,152 road segments and 510,119 connectors, converted to ways and nodes with topology preserved through connector-seeded, coordinate-keyed node identity. Table I pins the versions; both sides are public, and the edit window and class filters are the only parameters.

Two target variants. Overture's road subtype includes 119,135 footway segments here, far more pedestrian linework than a vehicular base map carries, and a footway along a driveway is a plausible false correspondence. The *vehicular* target keeps vehicular classes only and is our headline; the *full-Overture* target adds footways and is the harder bound. This is the biggest sensitivity we found, so we report both.

TABLE I
Source and target map versions used in the open benchmark of this section, with the window over which source edits were accumulated. Every version is a published artifact, so the benchmark is independently re-derivable.

Map	Role	Version
OpenStreetMap history	Source edits	2024-08-01 to 2026-04-10
Overture transportation	Target base map	2024-07-22.0

Metrics. Work is partitioned into 1,583 spatial clusters, each source way in exactly one, with no shared state between them. Every open-benchmark recall divides by the roads actually *missing*: the eligible set less the roads deduplication recognizes as already present. Deduplication therefore earns no recall on this benchmark; it only decides which roads are in scope (its false fires, reported below, are the exception). We keep two recalls apart. *Detection recall* is the share the correspondence model declares new; it is diagnostic only, because a detected road that is never connected is of no use to a router. *Connected recall* is the share actually joined to the target network. Table II reports both by road class after each stage, and Table III the disposition of every missing road. All counts are on the full 12,403-way source set.

Correctness at the operating point. The stitcher runs at its fixed auto-commit gate, and our correctness evidence here is geometric and duplicate-based, so these figures are not comparable to the 99%-precision figures of Section V-D. The applied population audits clean: all 5,235 connection points created for the 5,692 stitcher-applied ways lie at 0.000 m on the target network, with no wrong-junction snaps or degenerate geometry. The pipeline emits zero duplicate creates out of 12,403 (95% Wilson upper bound 0.031%), and deduplication false-fires on 35 of the 10,624 ways verified absent from the target (0.33%).

B. Model recall on the open benchmark

Of the 12,403 eligible source ways, deduplication suppresses 1,723 as already present, leaving 10,680 genuinely missing. On the vehicular target the correspondence model detects 8,990 of those (84.2%) as new (Table II). The geometric stitcher connects 5,692 of them, a connected recall of 53.3% and a detect-to-connect conversion of 63.3%; the other 3,298 detected roads form the recall-loss tail. On that tail the reachability pass connects a further 564 roads without a model call and the LLM agent 1,271, for a connected recall of 70.5%; 1,197 roads go to human review and 266 are stopped by the road-class or minimum-length gate (Table III). Every eligible way lands in exactly one bucket (1,690 model-missed + 7,527 connected + 1,463 queued = 10,680).

Footways cost 17.9 points of detection recall on the full-Overture target because they absorb correspondences, a driveway being matched to the pedestrian way alongside it. That is a substrate property, not a modelling failure, and it is why a vehicular-provider migration is easier than the full-Overture bound suggests.

Where recall is lost. The residual splits cleanly into two stages, and the split is the open benchmark's main finding.

Model-stage loss (1,690 roads, 15.8%). Roads the correspondence model declares already present. All score at or below its decision threshold, 68.8% below 0.4 and 31.2% in 0.4–0.5, so the bucket mixes confident errors with near-misses. On a 250-way audited sample from the full-Overture run none is a hidden duplicate; the characteristic failure is a driveway matched to the street it serves, 10 to 15 m away and roughly parallel.

Stitch-stage loss (3,298 roads, 30.9%). Roads correctly detected as new that the stitcher declined to commit. 2,308 of them (70.0%) score between 0.5 and 0.8. The auto-commit confidence gate withholds these, not any geometric obstacle. Of the remaining 990 that clear the gate, 266 are stopped by the road-class or minimum-length gates and 724 find no admissible connection in the window.

The tail is therefore concentrated, not diffuse. Two thirds of the loss sits at the stitch stage, and 70% of that in one confidence band of roads the model already believes are new and for which candidates exist. That is the population the selection agent of Section IV-C is built for, where the decision is which precomputed candidate to accept rather than whether the road is new; it recovers 1,271 of the 3,298.

C. Results by road class

The model is strongest exactly where the motivating application needs it (Table II). Driveways, the closest public analogue to a last-mile final leg, reach 93.0% model recall. Residential streets are the weakest of the volume classes at 66.0%; they are long and are the class most often matched to an adjacent parallel road, typically the street a driveway serves. Tracks reach 95.8% detection, yet most are then dropped by a road-class gate before stitching, so recovering them is a configuration change, not research. Connected recall follows the same order, 80.0% on driveways against 52.8% on residential streets; parking aisles convert worst at the stitcher (41.2%) because the aisle's far end often has no node inside the window, a geometry failure the later stages cannot repair.

Figure 3 shows ordinary committed output; most driveway additions are short stubs off an existing street.

D. Behavior across road-network density

The open benchmark covers one metropolitan area, so on its own it cannot show how the pipeline behaves as intersection density falls. Table IV comes from a different experiment: manually audited custom edits replayed onto four further OpenStreetMap regions, from dense urban to rural, at a fixed 99% precision operating point. These cover every mode of transport, so the footway effect of Section V-B applies throughout and accounts for part of the lower recall.

Two things carry over. First, network density dominates. The stitcher holds at 92.1% in the dense urban region and falls to 47.3% in the rural one, the expected hard case for any fixed-window method, matching the open benchmark's finding that long residential roads are harder than short driveways. Second, the lift decomposes rather than being a single number. Dedu-

TABLE II

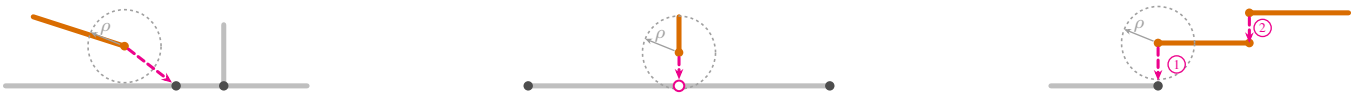
OPEN OPENSTREETMAP-TO-OVERTURE BENCHMARK (SECTION V-A), VEHICULAR TARGET. RECALL BY ROAD CLASS: COUNTS, WITH SHARES OF THE CLASS'S *Missing* ROADS BENEATH. *Missing* LEAVES OUT ROADS DEDUPLICATION FOUND ALREADY IN THE TARGET, SO EVERY SHARE IS OVER ROADS THAT GENUINELY HAD TO BE ADDED. *Detected* IS DETECTION RECALL; THE NEXT THREE COLUMNS ARE CUMULATIVE *connected* RECALL AFTER THE STITCHER, THE REACHABILITY PASS AND THE LLM AGENT. *Queued* ROADS WERE DETECTED BUT CONNECTED BY NO STAGE. "OTHER" POOLS TRACK, UNCLASSIFIED, ALLEY AND LIVING STREET.

			Connected (cumulative)			
Class	Missing	Detected	Stitcher	+ Reach.	+ Agent	Queued
driveway	5,358	4,985 (93.0%)	3,332 (62.2%)	3,546 (66.2%)	4,288 (80.0%)	697 (13.0%)
service (other)	2,554	2,044 (80.0%)	1,303 (51.0%)	1,571 (61.5%)	1,767 (69.2%)	277 (10.8%)
residential	1,848	1,220 (66.0%)	714 (38.6%)	761 (41.2%)	975 (52.8%)	245 (13.3%)
parking aisle	566	438 (77.4%)	233 (41.2%)	255 (45.1%)	343 (60.6%)	95 (16.8%)
other (4 classes)	354	303 (85.6%)	110 (31.1%)	123 (34.7%)	154 (43.5%)	149 (42.1%)
vehicular target	**10,680**	**8,990 (84.2%)**	**5,692 (53.3%)**	**6,256 (58.6%)**	**7,527 (70.5%)**	**1,463 (13.7%)**

TABLE III

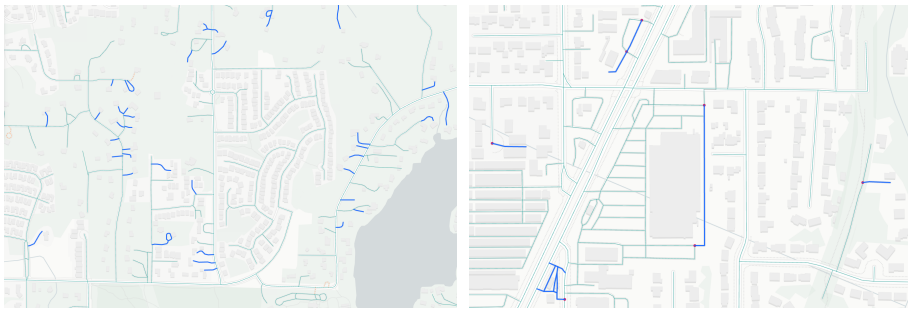
OPEN OPENSTREETMAP-TO-OVERTURE BENCHMARK: WHERE THE 10,680 MISSING ROADS ENDED UP AFTER THE COMPLETE PIPELINE. INDENTED ROWS PARTITION THE DETECTED ROADS; SHARES ARE OF ALL MISSING ROADS, SO THE THREE CONNECTED ROWS SUM TO THE 70.5% CONNECTED RECALL.

Bucket	Roads	Share
Detected as new by correspondence model	8,990	84.2%
connected by the geometric stitcher	5,692	53.3%
connected by the reachability pass	564	5.3%
connected by the LLM agent	1,271	11.9%
queued for human review	1,197	11.2%
dropped by road-class or min.-length gate	266	2.5%
Declared already present (model-missed)	1,690	15.8%
Missing	**10,680**	**100%**



(a) stub beyond window (b) mid-segment T (c) multi-hop chain

Fig. 2. Recovery scenarios the fixed-window stitcher drops and the agent is designed for. Gray: target network. Orange: a source road withheld to the recall-loss set R. Dotted circle: the stitcher's window, radius ρ. Magenta dashed: the proposed connection; the hollow node in (b) is made by splitting a segment, and (c)'s numbers are chaining rounds. Chains whose every hop is a trivial candidate are resolved by the deterministic pass.



(a) driveway additions (b) a service road joining two cross streets

Fig. 3. Committed output on the open benchmark. **Blue**: source roads successfully stitched and committed. **Green**: the existing target network. **Orange dashed**: roads that could not be stitched, withheld for low confidence or because no node overlap fell inside the proximity threshold. **Crimson** dots in (b): connection points onto the target network. Map data © OpenStreetMap contributors.

TABLE IV
A DIFFERENT EXPERIMENT FROM TABLES II–III: MANUALLY AUDITED CUSTOM EDITS REPLAYED ONTO FOUR FURTHER OPENSTREETMAP REGIONS, AT 99% PRECISION, ALL MODES OF TRANSPORT; COLUMNS CUMULATIVE, LAST ROW THE UNWEIGHTED MEAN. RECALL IS OVER THE WHOLE AUDITED DIFF, SO A ROAD CORRECTLY RECOGNIZED AS ALREADY PRESENT COUNTS AS RESOLVED AND DEDUPLICATION RAISES RECALL HERE, WHEREAS TABLE II DROPS THOSE ROADS BEFORE COUNTING. THE TWO ANSWER DIFFERENT QUESTIONS AND DO NOT COMBINE.

Region	Setting	Stitcher	+dedup.	+reach.	+re-adm.	+agent
A	suburban	71.2%	75.6%	81.7%	84.9%	90.6%
B	mixed	85.7%	96.2%	96.9%	97.0%	97.3%
C	dense urban	92.1%	93.5%	95.2%	95.6%	97.1%
D	rural	47.3%	56.7%	69.1%	75.8%	87.0%
Mean	–	74.1%	80.5%	85.7%	88.3%	**93.0%**

plication, by resolving the roads that were already present, moves the mean from 74.1% to 80.5%, the reachability pass to 85.7%, geometric re-admission to 88.3% and the LLM agent to 93.0%. The two deterministic stages after deduplication therefore account for 7.8 of the 12.5 points and the agent for 4.7, concentrated in the rural region (11.2 points), where the fixed-window stitcher is weakest, and negligible in the mixed region, which is already at 97.0% before the agent runs.

E. Cost, latency, and agent configuration

Deterministic stages. The three deterministic stages cost 91 ms of CPU per source way on the vehicular target, 1,126 s of cluster CPU over 2,065 km^2, rising to 144 ms against the full Overture road subtype. The slowest cluster takes 21 s, and all 1,583 clusters completed without error on both targets. Clusters share no state, so throughput scales with worker count and an incremental refresh revisits only the roads that moved.

Agent configuration. Table V fixes what Section IV-C left open. The prompt has three parts: (i) the task statement and the contract, which instructs the model to return exactly one candidate identifier from the table or the token HOLD, and never coordinates; (ii) the candidate table of non-geometric signals of Algorithm 1; and (iii) the annotated overlay, two aerial tiles at zoom 19 centered on the withheld road's endpoint, with a third tile at zoom 20 added only on visual reconsideration.

TABLE V
LLM AGENT CONFIGURATION AND MEASURED COST. TOKEN, COST AND LATENCY FIGURES ARE PER CONNECTED COMPONENT THE AGENT EXAMINES, A WITHHELD ROAD TOGETHER WITH THE CHAIN IT CONNECTS THROUGH, SUMMED OVER EVERY VOTE AND CHAINING ROUND; P90 IS THE HIGHEST MEASURED PERCENTILE.

Configuration	
Model	Claude Opus 5 (Amazon Bedrock)
Sampling	adaptive thinking; no temperature
Votes per decision k; ties	3; HOLD
Max chaining rounds K	5
Imagery per call	2 tiles at zoom 19
Measured per connected component, p50 (p90)	
Input tokens	184,814 (246,727)
served from prompt cache	92%
Output tokens, with reasoning	9,038 (15,637)
Wall-clock latency	145 s (249 s)
Cost at public list price	$0.80

Level 3 uses a separate prompt that asks for sparse waypoints along a drivable path and is issued only when the candidate set is empty. The prompts are not released.

Agent cost and latency. We report cost per connected component the agent examines, a withheld road together with the chain it connects through. At the median a component consumes 184,814 input tokens, 92% of them served from the prompt cache since the contract and tool schemas are shared across turns, and 9,038 output tokens including reasoning, over about twelve model turns and fifteen tool calls. Latency is 145 s at the median and 249 s at p90, dominated by reasoning tokens rather than tool execution; the deterministic validators run in milliseconds, so the layered validation of Section IV-D is not where the time goes. At public on-demand list price the median component costs $0.80, so the 3,298-road tail costs at most about $2,600 if every road were its own component, against about $8,500 for an agent run on every missing road. These figures are an upper bound, measured on country-scale workloads of the agent stack; the stitching agent's candidate table is short and its chaining runs only on the hard cases. The reduction from running the agent on the tail alone is the constant factor of the tail's share, $1/0.31 \approx 3.2\times$, whatever the per-component price. Acceptances per level, vote downgrades

and imagery–geometry conflicts were not instrumented in these runs.

When imagery and geometry disagree. The two signals are not peers, and the ordering answers what happens when they conflict. Geometric validation is a gate the model cannot open: a candidate that crosses a building or divider, exceeds the length bound, or fails the snap or topology check is never offered to the model, and a Level-3 trace that fails those checks is rejected whatever the imagery shows. Imagery decides only among candidates that have already passed. Within that set a visual verdict can still contradict the geometric signals, typically a short, class-compatible snap whose path the tile shows blocked by a fence, wall or tree line the target map does not record. Nor is such a candidate accepted on geometry alone, since acceptance requires several independent signals to agree (Section IV-D); the decision is HOLD, after one more look at closer zoom for borderline cases, and the road goes to human review with the disagreeing signals attached. The asymmetry is deliberate: imagery can veto a geometrically plausible candidate but never admit one that geometry rejects, so a stale or mis-registered tile costs recall, never precision.

The benchmark relocated the bottleneck. The most useful outcome of moving to open data is the diagnosis rather than the headline. The bottleneck is neither the correspondence model, which finds 84.2% of missing roads, nor geometric impossibility. It is a confidence gate. $2{,}308$ roads, over a fifth of the missing set, are detected as new and then withheld for want of confidence in which connection to choose. A selective agent is the right shape of intervention there (Fig. 2), and it recovers $1{,}271$ of those roads; a threshold sweep on the stitcher is the cheapest complement to try and has not been measured here. Replacing the stitcher with the agent would instead pay agent cost on every road.

Limitations. (i) The 93% rests on the audited-edit regions, where the reference was set by trained map editors who also maintain operational map data and the custom edits themselves are not released. Unlike the open benchmark, these regions also credit roads correctly recognized as already present, so the two sets must not be combined. (ii) Overture's transportation theme is substantially derived from OpenStreetMap, so this is closer to a map and its own derivative than to two independently edited providers, since some correspondences are recoverable from shared identifiers rather than geometry, and measured recall is optimistic relative to a genuinely independent pair. (iii) Deduplication coverage is a single-best-edge metric, so very short ways can show inflated coverage from crossing roads, and one-ended stitches remain possible in the evaluated code, since the both-ends gate is not enabled. (iv) The LLM stage samples, so identical input can yield different proposals across runs; reject-only voting mitigates rather than eliminates this, and chaining adds a second-order dependence. (v) One-to-many correspondences are not modeled, so the near-parallel deduplication test can read two carriageways as one road digitized twice. (vi) Splicing splits a segment, so linearly referenced attributes must be re-referenced onto the parts, which makes the attribute and geometry halves less separable than Section I presents.

VI. Conclusions and Future Work

We have presented a hybrid pipeline for geometric edit migration that composes a fast ML-powered context-based stitcher with a multimodal LLM agent, so that operators can swap base-map providers while carrying over years of accumulated road-geometry edits, committing only geometry that joins the target network at valid junctions. On a reproducible open-data OpenStreetMap-to-Overture benchmark the correspondence model detects 84.2% of the missing roads, the geometric stitcher connects 53.3% of them and the composed pipeline 70.5%. On audited multi-region edits the stitcher reaches $\sim$$74\%$ recall at 99% precision and the composed pipeline 93%; deduplication accounts for 6.4 of those points, the deterministic reachability and re-admission stages for 7.8, and the LLM agent for 4.7. The two tiers meet at a clean handoff that labels ML-stitched and agent-stitched roads separately, and a human-review queue takes the connections too ambiguous for any stage. Because the agent runs only on the residual tail, its spend is a constant factor below an agent-on-every-road design, which bulk migration needs to stay affordable as area grows.

Several directions would strengthen this work. A controlled comparison between the LLM selection agent and a deterministic nearest-compatible selector over the same candidates would attribute the recall lift more sharply, and a threshold sweep on the stitcher would establish how much of the residual is untuned reach rather than missing evidence. Evaluating against a genuinely independent commercial target would test the harder tail that an OpenStreetMap-derived target understates; restricting the present benchmark to Overture segments whose provenance is not OpenStreetMap is a cheaper first step. Finally, we plan to learn the geometric proposal thresholds per region rather than fixing them globally.

References

[1] M. Haklay and P. Weber, "OpenStreetMap: User-generated street maps," *IEEE Pervasive Computing*, vol. 7, no. 4, pp. 12–18, 2008. [Online]. Available: https://doi.org/10.1109/MPRV.2008.80

[2] Overture Maps Foundation, "Overture Maps Foundation: Transportation theme," 2026, open map data. Accessed 2026-08-21. [Online]. Available: https://overturemaps.org/

[3] C. Barron, P. Neis, and A. Zipf, "A comprehensive framework for intrinsic OpenStreetMap quality analysis," *Transactions in GIS*, vol. 18, no. 6, pp. 877–895, 2014. [Online]. Available: https://doi.org/10.1111/tgis.12073

[4] D. Merchán, J. Arora, J. Pachon, K. Konduri, M. Winkenbach, S. Parks, and J. Noszek, "2021 Amazon last mile routing research challenge: Data set," *Transportation Science*, vol. 58, no. 1, pp. 8–11, 2024. [Online]. Available: https://doi.org/10.1287/trsc.2022.1173

[5] J. J. Ruiz, F. J. Ariza, M. A. Ureña, and E. B. Blázquez, "Digital map conflation: A review of the process and a proposal for classification," *International Journal of Geographical Information Science*, vol. 25, no. 9, pp. 1439–1466, 2011. [Online]. Available: https://doi.org/10.1080/13658816.2010.519707

[6] M. Hacar and T. Gökgöz, "A new, score-based multi-stage matching approach for road network conflation in different road patterns," *ISPRS International Journal of Geo-Information*, vol. 8, no. 2, p. 81, 2019. [Online]. Available: https://doi.org/10.3390/ijgi8020081

[7] G. Ali, N. Feierabend, P. Doshi, W. Chung, S. Babiceanu, and M. Fontaine, "Automated route-based conflation between linear referencing system maps and OpenStreetMap using open-source tools," in *Proc. IEEE Int. Conf. Intelligent Transportation Systems (ITSC)*, 2025, pp. 4443–4450. [Online]. Available: https://doi.org/10.1109/ITSC60802.2025.11423782

[8] P. Newson and J. Krumm, "Hidden Markov map matching through noise and sparseness," in *Proc. ACM SIGSPATIAL*, 2009, pp. 336–343. [Online]. Available: https://doi.org/10.1145/1653771.1653818

[9] K. Kuckreja, M. S. Danish, M. Naseer, A. Das, S. Khan, and F. S. Khan, "GeoChat: Grounded large vision-language model for remote sensing," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2024, pp. 27 831–27 840. [Online]. Available: https://doi.org/10.1109/CVPR52733.2024.02629

[10] G. Mai, W. Huang, J. Sun, S. Song, D. Mishra, N. Liu, S. Gao, T. Liu, G. Cong, Y. Hu, C. Cundy, Z. Li, R. Zhu, and N. Lao, "On the opportunities and challenges of foundation models for GeoAI (vision paper)," *ACM Transactions on Spatial Algorithms and Systems*, vol. 10, no. 2, pp. 1–46, 2024. [Online]. Available: https://doi.org/10.1145/3653070

[11] M. Hacar and Ö. Öztürk Hacar, "LLM-based map conflation: Performance assessment on matching embedded road lines," *ISPRS International Journal of Geo-Information*, vol. 15, no. 4, p. 144, 2026. [Online]. Available: https://doi.org/10.3390/ijgi15040144

[12] A. Saalfeld, "Conflation: Automated map compilation," *International Journal of Geographical Information Systems*, vol. 2, no. 3, pp. 217–228, 1988. [Online]. Available: https://doi.org/10.1080/02693798808927897

[13] G. Agarwal, L. V. S. Lakshmanan, X. Gao, K. Ventullo, S. Mohapatra, and S. Basu, "MAYUR: Map conflAtion using earlY prUning and Rank join," in *Proc. ACM SIGSPATIAL*, 2021, pp. 550–553. [Online]. Available: https://doi.org/10.1145/3474717.3484258

[14] E. H. Sereshgi and C. Wenk, "Map Stitcher: Graph sampling-based map conflation," in *Proc. ACM SIGSPATIAL Workshop on Spatial Big Data and AI for Industrial Applications*, 2024, pp. 5–15. [Online]. Available: https://doi.org/10.1145/3681766.3699604

[15] F. Hashemi and L. V. S. Lakshmanan, "KRAFT: A knowledge graph-based framework for automated map conflation," in *Proc. ACM Int. Conf. Information and Knowledge Management (CIKM)*, 2025, pp. 802–812. [Online]. Available: https://doi.org/10.1145/3746252.3761291

[16] M. Leitenstern, F. Sauerbeck, D. Kulmer, and J. Betz, "FlexMap fusion: Georeferencing and automated conflation of HD maps with OpenStreetMap," in *Proc. IEEE Intelligent Vehicles Symp. (IV)*, 2024, pp. 272–278. [Online]. Available: https://doi.org/10.1109/IV55156.2024.10588414

[17] V. Kango, H. M. Eraqi, and M. Moustafa, "High precision map conflation of fleet sourced traffic signs," in *Proc. IEEE Int. Geoscience and Remote Sensing Symp. (IGARSS)*, 2024. [Online]. Available: https://doi.org/10.1109/IGARSS53475.2024.10642165

[18] G. Mattyus, W. Luo, and R. Urtasun, "DeepRoadMapper: Extracting road topology from aerial images," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017, pp. 3458–3466. [Online]. Available: https://doi.org/10.1109/ICCV.2017.372

[19] F. Bastani, S. He, S. Abbar, M. Alizadeh, H. Balakrishnan, S. Chawla, S. Madden, and D. DeWitt, "RoadTracer: Automatic extraction of road networks from aerial images," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 4720–4728. [Online]. Available: https://doi.org/10.1109/CVPR.2018.00496

[20] S. He, F. Bastani, S. Jagwani, M. Alizadeh, H. Balakrishnan, S. Chawla, M. M. Elshrif, S. Madden, and M. A. Sadeghi, "Sat2Graph: Road graph extraction through graph-tensor encoding," in *Proc. European Conf. Computer Vision (ECCV)*, 2020, pp. 51–67. [Online]. Available: https://doi.org/10.1007/978-3-030-58586-0_4

[21] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, and Y. Cao, "ReAct: Synergizing reasoning and acting in language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023. [Online]. Available: https://arxiv.org/abs/2210.03629

[22] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou, "Self-consistency improves chain of thought reasoning in language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023. [Online]. Available: https://arxiv.org/abs/2203.11171