

Multimodal Road Closure Mapping via Cross-Modal Graph Attention

Izeddin Teetii
izzteeti@amazon.com
Amazon
US

Nidhi Sridhar
nidhsrid@amazon.com
Amazon
US

Vaibhav Kango
kangov@amazon.com
Amazon
US

Mohamed Moustafa
mmoustm@amazon.com
Amazon
US

Abstract

Accurate and timely identification of road closures is critical for large-scale navigation and logistics systems, yet remains challenging due to the need to jointly reason over visual evidence, spatial context, and road network topology. Existing approaches either rely on manual map updates or focus on detecting closure-related objects without resolving which road segment is affected. In this work, we introduce a unified multimodal and multitask framework that jointly performs *road closure classification* (no, partial, complete) and *road segment association* within a single learning architecture. Our model integrates graph-structured map data, satellite imagery augmented with projected street-level detections, and street-view images using modality-specific encoders and cross-modal transformer attention, enabling explicit reasoning over visual cues and network topology. Evaluated on nearly 20K real-world scenarios collected from a large-scale commercial vehicle fleet, our approach achieves **72.7%** closure classification accuracy and **60%** road segment association accuracy, outperforming frontier vision-language models by over **28 percentage points** on closure classification while being over **4,000× smaller** and **100× faster** at inference. Extensive ablations demonstrate the importance of cross-modal attention, edge-aware pooling, and graph message passing for accurate geospatial reasoning. This work establishes the first benchmark for joint road closure classification and segment-level attribution, and shows that task-specific architectural inductive biases substantially outperform general-purpose multimodal models for structured geospatial understanding.

CCS Concepts

• **Computing methodologies** → **Scene understanding**; • **Information systems** → **Geographic information systems**.

Keywords

Road closure detection · multimodal learning · graph attention networks · cross-modal fusion · road segment association · multi-task learning · geospatial reasoning

ACM Reference Format:

Izeddin Teetii, Vaibhav Kango, Nidhi Sridhar, and Mohamed Moustafa. 2026. Multimodal Road Closure Mapping via Cross-Modal Graph Attention. In *Proceedings of 34th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems (SIGSPATIAL '26)*. ACM, New York, NY, USA, 11 pages. <https://doi.org/10.1145/nnnnnnn.nnnnnnn>

1 Introduction

Accurate and timely road network information is fundamental to high-quality digital maps to enable safe and efficient delivery planning and navigation at scale. With billions of packages delivered annually by large-scale logistics operators, temporary road closures caused by construction, accidents, or events are a major source of routing failures. Yet most map systems rely on manual reporting, leading to delayed or incomplete updates that degrade navigation quality and operational efficiency.

Updating maps with road closures is more challenging than handling static or well-defined attributes such as traffic lights or stop signs. While detecting a traffic light unambiguously implies its presence, observing closure-related objects (e.g., barricades, drums, cones) does not necessarily indicate that a road is closed, nor does it specify which road segment is affected. Translating raw sensor detections into actionable map updates therefore requires solving two coupled problems: determining the closure type and identifying the impacted road segment. This is inherently a multimodal reasoning task.

Geospatial context information is often complementary but incomplete across individual modalities. Street-level images capture fine-grained object appearance but lack global spatial context; satellite imagery provides aerial view layout while missing small and especially transient objects; and map data encodes road topology without visual evidence. Effective road closure mapping requires jointly reasoning over visual, spatial, and graph-structured information. Our approach encodes each modality into a shared token space and fuses them through cross-modal attention, enabling road graph tokens to selectively attend to the most relevant visual evidence.

Recent vision-language models (VLMs) exhibit strong general-purpose multimodal capabilities, yet our experiments show that

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

SIGSPATIAL '26, Riverside, CA, USA

© 2026 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-x-xxxx-xxxx-x/YYYY/MM
<https://doi.org/10.1145/nnnnnnn.nnnnnnn>

they struggle with precise geospatial reasoning over structured maps. For example, Claude Opus achieves less than 45% accuracy on closure type classification and cannot meaningfully address road segment association due to its inability to reason over graph-structured spatial representations. These results suggest that general multimodal understanding does not readily transfer to tasks requiring fine-grained spatial reasoning over road networks, motivating architectures with explicit topological inductive biases.

In this work, we formally define the *road closure classification and road segment association problem* and propose a unified multimodal and multitask architecture as solution. To the best of our knowledge, this is the first work to jointly address closure classification and road segment association within a single learning framework. As discussed in Section 2, prior approaches primarily focus on detecting closure-related objects, without resolving closure type or performing automated map updates.

We argue that effective geospatial reasoning requires explicit architectural inductive biases. Road networks are inherently graph-structured, where connectivity and topology are as critical as visual appearance. Accordingly, we design a model that explicitly encodes road graph structure, spatial relationships, and cross-modal interactions. Our contributions are summarized as follows:

- (1) We formally define the road closure classification and road segment association tasks, providing a clear problem formulation and baseline for future research.
- (2) We propose a modular multimodal, multitask architecture that jointly performs closure classification and segment association.
- (3) We extend GATv2 [1] with an edge-aware decoder that conditions predictions on fused multimodal context and supports variable-length edge sets. On a large-scale real-world dataset, our model achieves 72.7% closure classification accuracy and 60% segment association accuracy, outperforming frontier VLMs by over 28% on closure classification while being over $4,000\times$ smaller and $100\times$ faster.

2 Related Work

2.1 Map Association of Detected Road Assets

Several works study the association of detected traffic-related objects with map elements, but not in the context of road closures or edge-level attribution. Zhang et al. [18] and Campbell et al. [2] address the association of detected objects, such as stop signs and traffic lights, to digital maps. While related in spirit, these approaches differ fundamentally from our setting. Zhang et al. combine learning-based detection with model- and rule-driven association, but do not report quantitative benchmarks or release code, limiting reproducibility and scalability. Campbell et al. localize traffic signs using street-view imagery alone, employing map data only during the association stage. This reliance on a single sensing modality limits robustness and fails to exploit complementary cues from aerial imagery and map topology, as demonstrated in our ablation study in Section 5.2. Importantly, neither work addresses road closures, temporary changes, or edge-level reasoning over road graphs, which are central to our problem setting, nor do they provide publicly available implementations.

2.2 Road Closure Detection and Roadwork Understanding

Only a limited number of studies explicitly address road closures, primarily from a detection perspective. Hasan et al. [8] propose a system for detecting road closure signs using YOLO-based object detection [13]. However, their work focuses solely on sign detection and does not describe how detected closures are integrated into a digital map, nor how ambiguities are resolved in complex scenarios such as intersections or multi-lane roads. Moreover, no quantitative evaluation is provided for closure-to-map integration accuracy. More recent datasets such as ROADWork [7] and methods for construction-site detection [6] advance visual understanding of roadwork scenes. While these efforts improve object-level perception, they do not address closure type classification, road segment association, or reasoning over structured road networks.

2.3 Graph-Based Reasoning on Road Networks

Graph neural networks (GNNs) have been widely applied to road networks for tasks such as traffic prediction and road attribute inference [9, 14, 17]. These works demonstrate that modeling road networks as graphs provides a strong inductive bias for reasoning over structured, topologically constrained data and enables effective propagation of contextual information across connected road segments. This motivates our use of GNNs for edge-level road closure reasoning. However, existing approaches do not consider closure attribution, multimodal fusion with visual cues, or the estimation of temporary road states, which are central to our problem.

In summary, prior work either focuses on detecting roadwork or closure-related objects without associating them to a digital map, or performs map association using rule-based heuristics. To the best of our knowledge, no existing method jointly integrates multimodal visual and map cues with explicit graph-based reasoning to perform edge-level road closure attribution together with closure type classification on an OSM graph.

3 Proposed Solution

3.1 Problem Definition

Given a local road network graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, a satellite image $\mathbf{I}_{\text{sat}}$ augmented with projected street-level detections, and a street-view image $\mathbf{I}_{\text{street}}$, we address two interconnected problems:

- (1) **Closure Classification:** Determine whether a road closure exists and, if present, classify its type as either *partial* (vehicle can proceed with restrictions) or *complete* (vehicle cannot proceed). This is formulated as a multi-class classification over $C=3$ classes: $\hat{y}_{\text{closure}} \in \{\text{no}, \text{partial}, \text{complete}\}$.
- (2) **Road Segment Association:** Identify which specific road segment is impacted by the detected closure. This is formulated as a classification problem over the set of M candidate edges in the local road network graph: $\hat{e} \in \mathcal{E}$.

Both tasks are solved jointly within a single forward pass, enabling shared representation learning across closure type and spatial attribution.

We propose a modular multimodal architecture that encodes each input modality independently, fuses them through cross-modal attention, and produces predictions via task-specific decoder heads.

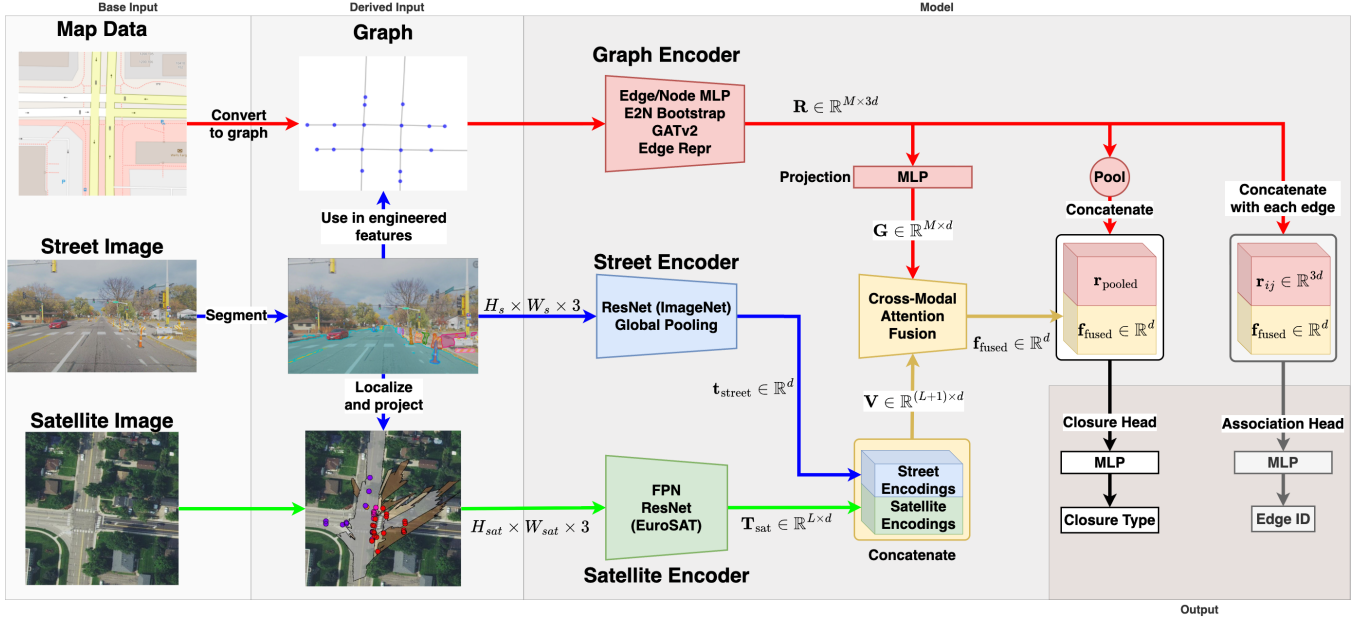


Figure 1: Overview of the proposed multimodal architecture.

As illustrated in Figure 1, the model takes three inputs, street-view images, satellite imagery augmented with projected detections, and map data, and jointly outputs closure classification and impacted road segment association.

3.2 Modality-Specific Encoders

Street Image Encoder. Street-view images $I_{\text{street}} \in \mathbb{R}^{H_{st} \times W_{st} \times 3}$ resemble natural images and primarily convey the presence and appearance of closure-related objects. We therefore employ a ResNet backbone pretrained on ImageNet [3] to extract a global feature vector $t_{\text{street}} \in \mathbb{R}^d$, where d denotes the hidden dimension.

Satellite Image Encoder. Satellite imagery provides a bird’s-eye view of the road network and is augmented with projected street-level detections to encode spatial context. Detected closure objects and road surfaces from street images are geometrically projected onto the satellite view, and the corresponding pixels are replaced with color-coded representations (e.g., barricades in red, signs in blue, road surfaces in gray), as shown in Figure 2b. This results in a three-channel image $I_{\text{satellite}} \in \mathbb{R}^{H_s \times W_s \times 3}$ that preserves the RGB format while explicitly highlighting detected elements.

This projection introduces a spatial prior that guides the model toward relevant spatial relationships between closure cues and the road layout, without requiring the satellite encoder to relearn object detection. To capture spatial structure at multiple scales, we employ a Feature Pyramid Network (FPN) [11] with a ResNet backbone pretrained on satellite imagery [10] to better align the feature extractor with the target domain. The FPN produces a set of multi-scale tokens $T_{\text{satellite}} \in \mathbb{R}^{L \times d}$, where L denotes the total number of spatial locations across pyramid levels.

Design Rationale. We intentionally adopt asymmetric encoders to reflect the differing information content of street and satellite

imagery. Satellite images encode rich spatial structure and road topology, making multi-scale feature extraction essential. In contrast, street images are perspective views centered on objects ahead of the vehicle, where global appearance cues are sufficient. This asymmetric design balances expressiveness and efficiency, enabling effective multimodal fusion while maintaining practical computational cost.

Graph Encoder. The graph encoder transforms raw map data into expressive *edge-level* representations suitable for road segment association. We adopt a bidirectional edge–node–edge design in which edge features initialize node representations, nodes exchange information through message passing to capture topology, and enriched node embeddings are used to refine edge representations. This ensures that each edge captures both its local attributes and its topological context within the road network. Standard formulations for node bootstrapping, attention computation, and residual updates are deferred to Appendix A.

Edge and Node Initialization. Raw edge attributes (e.g., length, angle, road class) are first encoded using a two-layer MLP with batch normalization to obtain edge embeddings $e_{ij} \in \mathbb{R}^d$. Node features (e.g., intersection properties and spatial cues) are linearly projected to the same hidden dimension. To bootstrap node representations with local road context, incoming edge embeddings are aggregated via mean pooling and combined with node features through a learned projection.

Topology-Aware Message Passing. Node embeddings are refined through multiple graph attention layers to propagate information across the road network. We employ GATv2 [1], which improves expressiveness over GAT [16], and extend it to incorporate edge features into the attention computation. Specifically, the unnormalized attention score between nodes i and j is computed

as:

$$\alpha_{ij} = \mathbf{a}^\top \text{LeakyReLU}(\mathbf{W}[\mathbf{h}_i \parallel \mathbf{h}_j \parallel \mathbf{e}_{ij}]), \quad (1)$$

where $\mathbf{h}_i, \mathbf{h}_j$ are node embeddings, $\mathbf{e}_{ij}$ is the edge embedding, and $\mathbf{a}, \mathbf{W}$ are learnable parameters. By applying the nonlinearity after concatenation, GATv2 enables dynamic attention rankings that depend on both query and key, unlike standard GAT. This allows message passing to account for node properties and the characteristics of the connecting road segments. Residual connections and layer normalization are applied to stabilize training.

Edge Representation Construction. After message passing, the final representation of each road segment is constructed by concatenating the embeddings of its source node, target node, and edge:

$$\mathbf{r}_{ij} = [\mathbf{h}_i \parallel \mathbf{h}_j \parallel \mathbf{e}_{ij}] \in \mathbb{R}^{3d}. \quad (2)$$

Edge representations are padded or truncated to a fixed maximum size M and accompanied by binary masks, yielding a batched tensor $\mathbf{R} \in \mathbb{R}^{B \times M \times 3d}$ for downstream fusion.

3.3 Cross-Modal Fusion

We fuse graph and visual information using cross-modal attention, where graph edge tokens query visual tokens from the satellite and street encoders. This design allows each road segment to attend selectively to relevant visual evidence.

Edge representations are first projected to the hidden dimension, while visual tokens from satellite imagery and the street-view encoder are concatenated into a shared set of visual tokens (batched as $\mathbf{T}_{\text{satellite}} \in \mathbb{R}^{B \times L \times d}$ and $\mathbf{t}_{\text{street}} \in \mathbb{R}^{B \times 1 \times d}$). To obtain a unified scene representation, we prepend a learnable CLS token to the graph tokens.

This design follows prior uses of CLS tokens in BERT [4] for sentence classification, Vision Transformers [5] for image classification, and CLIP [12] for vision-language alignment, and is adapted here for graph-vision multimodal fusion. The CLS token acts as an aggregation mechanism that learns to summarize the most relevant multimodal information required for downstream prediction.

Concretely, the projected edge representations $\mathbf{G} \in \mathbb{R}^{M \times d}$ (with the prepended CLS token) serve as queries, while the concatenated visual tokens $\mathbf{V} \in \mathbb{R}^{(L+1) \times d}$ form keys and values. Each cross-attention layer computes:

$$\mathbf{Q}^{(\ell+1)} = \text{LayerNorm}(\mathbf{Q}^{(\ell)} + \text{MHA}(\mathbf{Q}^{(\ell)}, \mathbf{V}, \mathbf{V})), \quad (3)$$

followed by a position-wise feed-forward network. This is stacked over multiple transformer blocks, enabling graph tokens to attend selectively to relevant visual evidence while the CLS token aggregates global multimodal context. Full formulations are provided in Appendix A.

The final fused scene representation is obtained from the CLS token after the last transformer layer and is denoted by $\mathbf{f}_{\text{fused}} \in \mathbb{R}^d$. This representation aggregates information across graph and visual modalities and conditions both downstream prediction heads.

3.4 Task-Specific Decoder Heads

Association Head. The association head predicts which road segment (graph edge) is affected by the closure. Each edge representation is conditioned on the fused scene context by concatenation:

$$\mathbf{z}_{ij} = [\mathbf{r}_{ij} \parallel \mathbf{f}_{\text{fused}}] \in \mathbb{R}^{4d}. \quad (4)$$

A two-layer MLP produces a scalar logit per edge:

$$s_{ij} = \text{MLP}_{\text{assoc}}(\mathbf{z}_{ij}) \in \mathbb{R}. \quad (5)$$

During training, logits corresponding to padded edges are masked by setting them to $-\infty$ prior to softmax. At inference, the predicted road segment is selected as $\hat{e} = \arg \max_j \text{softmax}(\mathbf{s})_j$.

Closure Classification Head. Closure type classification requires a global representation that respects the variable number of valid edges per graph. We therefore apply edge-aware mean pooling over valid edges only:

$$\mathbf{r}_{\text{pooled}} = \frac{1}{|\{(i, j) : m_{ij} = 1\}|} \sum_{(i, j) : m_{ij} = 1} \mathbf{r}_{ij}, \quad (6)$$

where $m_{ij} = 1$ denotes valid (non-padded) edges. This avoids representation dilution caused by naive pooling over padded edges, as confirmed in our ablation study (Section 5.2). The pooled representation is concatenated with the fused context and passed to a classification MLP:

$$\mathbf{y}_{\text{closure}} = \text{MLP}_{\text{closure}}([\mathbf{r}_{\text{pooled}} \parallel \mathbf{f}_{\text{fused}}]) \in \mathbb{R}^C, \quad (7)$$

where C denotes the number of closure classes (no, partial, full).

3.5 Learning Objective

We train the model using a multi-task objective with two cross-entropy losses: one for edge association over M candidates and one for closure classification over C classes. The total loss is given by

$$\mathcal{L} = \lambda_{\text{assoc}} \left(- \sum_{j=1}^M y_j \log \hat{y}_j \right) + \lambda_{\text{closure}} \left(- \sum_{c=1}^C y_c \log \hat{y}_c \right), \quad (8)$$

where y_j and y_c are one-hot ground-truth labels, $\hat{y}_j = \text{softmax}(\mathbf{s})_j$ and $\hat{y}_c = \text{softmax}(\mathbf{y}_{\text{closure}})_c$ are the predicted probabilities for association and closure classification, respectively, and λ_{assoc} and λ_{closure} balance the two objectives.

Association is optimized for all samples, including those without closures, as closure-related objects may be present without inducing an actual road closure.

4 Experiments

4.1 Dataset

We collected a real-world dataset using a large-scale commercial vehicle fleet, comprising 19,380 samples with detected closure objects such as barricades, drums, and cones. The dataset spans diverse geographic contexts across the US, including dense urban areas, suburban neighborhoods, rural highways, and residential streets. Each sample jointly provides street-level imagery, satellite imagery, structured map data, detected closure objects, and precise vehicle pose information (location and heading), enabling the study of multimodal road closure classification and road segment association.

Each sample consists of three modalities: (i) a forward-facing street-level image captured by the delivery vehicle (Figure 2c), (ii) a $100\text{ m} \times 100\text{ m}$ satellite image centered at the vehicle location with projected closure objects and surfaces (Figure 2b), and (iii) a local road network graph extracted from map data in the vicinity of the detected object. Details of data preprocessing, localization, and feature construction are provided in Appendix D.

Ground-truth annotations are provided by map and road-network domain experts. Each sample is labeled with the affected road segment and a closure type (*No Closure*, *Partial Closure*, or *Complete Closure*), with a naturally imbalanced distribution of 62%, 29%, and 9% respectively. We split the dataset into training (70%, 13,566 samples), validation (15%, 2,907 samples), and test (15%, 2,907 samples) sets, stratified by closure type.

Our graph encoder operates on a fixed maximum of $M=10$ edges per graph. In our dataset, the ground-truth road segment falls within the 10 nearest edges to the vehicle position in 94% of cases. Samples where the target edge exceeds M are excluded during preprocessing (approximately 6% of the data). Sequence bucketing or dynamic M selection are noted as potential future optimizations.

4.2 Implementation Details

All models are implemented in PyTorch and trained on a single NVIDIA Tesla V100 (16GB). Full architectural specifications and hyperparameters are provided in Appendix E.

4.3 Baselines

To our knowledge, this work is the first to jointly address road closure classification and affected road segment association using multimodal learning. With no public established benchmarks for this problem, we construct the following baselines:

- (1) **Claude Sonnet 3.5 v2.** A frontier vision-language model (VLM) evaluated on the closure classification task using street-level imagery. The model struggles with explicit spatial and topological reasoning over road networks.
- (2) **Claude Opus 4.5 v1.** A more advanced VLM with improved multi-step reasoning and vision capabilities. Similar to Sonnet, it is evaluated only on closure classification, as VLMs fundamentally lack the ability to reason over graph-structured spatial representations required for road segment association.
- (3) **Nearest-Edge Heuristic.** A rule-based baseline that assigns the closure to the road segment nearest to the median of the detected objects. This establishes a lower bound on task difficulty and isolates the contribution of learned representations over simple proximity-based reasoning.
- (4) **Task-Specific Variants.** To assess the benefit of joint learning, we train two single-task variants of our architecture: one optimized solely for closure classification and another solely for road segment association. These models serve as internal baselines in our ablations.

4.4 Evaluation Metrics

For baseline comparisons (Table 1), we report overall accuracy for each task, along with inference latency. For ablation studies (Tables 2–3), we report Precision, Recall, and F_1 score, macro-averaged

across classes and tasks. These complementary metrics provide deeper insight into model behavior under class imbalance and reveal performance trade-offs not captured by accuracy alone.

5 Results

5.1 Quantitative Results

Table 1 compares our proposed model against the baselines. General-purpose vision-language models (Claude Sonnet and Opus) are reported only for closure classification, as they cannot meaningfully address the association task due to their inability to reason over graph-structured road network representations. Our joint model achieves **72.7%** accuracy on closure classification and **59.9%** accuracy on segment association, substantially outperforming all baselines while operating with orders-of-magnitude lower latency and model size.

Table 1: Performance comparison with baseline methods

Model (params)	Closure/Assoc. Acc. (%)	Latency (s)
Nearest-Edge Heuristic (0)	— / 45.4	—
Claude Sonnet 3.5 v2 (>100B)	34.7 / —	7.0
Claude Opus 4.5 v1 (>100B)	44.3 / —	5.4
Ours (Separate) (12M each)	71.6 / 58.3	0.06
Ours (Joint) (23M)	72.7 / 59.9	0.05

To quantify uncertainty, we compute 95% bootstrap confidence intervals (10,000 resamples) on the test set. Our joint model achieves 72.7% closure classification accuracy (95% CI: [71.1%, 74.4%]) and 59.9% segment association accuracy (95% CI: [58.1%, 61.7%]). The narrow confidence intervals (≤ 3.6 percentage points) indicate stable performance estimates given the test set size of 2,907 samples.

Several key observations emerge from these results. First, large public VLMs struggle with structured geospatial reasoning despite prompt tuning and strong general reasoning capabilities, indicating poor transfer to map-centric tasks. Even the most capable model (Claude Opus) achieves only 44.3% on closure classification and cannot meaningfully address road segment association, confirming that general-purpose multimodal understanding does not substitute for explicit architectural inductive biases for graph reasoning and spatial alignment. Second, the nearest-edge heuristic achieves 45.4% on association through proximity alone, establishing a non-trivial lower bound; our model’s +14.5 point improvement over this heuristic confirms that learned multimodal representations contribute substantially beyond simple spatial proximity.

Joint training yields consistent performance gains (+1.1% for closure classification and +1.6% for association) over separate models, validating the effectiveness of multi-task learning. Closure type and affected road segment are geometrically coupled: complete closures typically block the full road width, whereas partial closures affect only a subset of the roadway.

Practical Considerations. In a potential application scenario, false positives (incorrectly predicting a closure) would cause unnecessary rerouting, while false negatives (missing a real closure) would result in vehicles encountering blocked roads. The asymmetric cost of these errors motivates operating the model as a

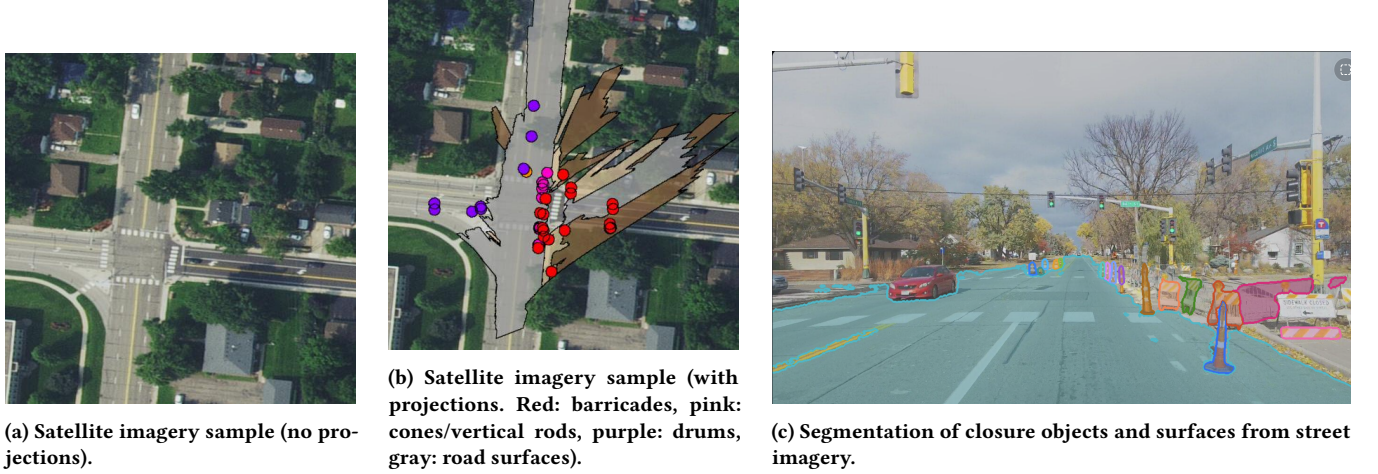


Figure 2: Input modalities for one data sample.

confidence-thresholded filter: only predictions exceeding a confidence threshold would be acted upon automatically, while lower-confidence cases can be deferred to human review or additional verification passes. This design allows practitioners to tune the precision–coverage tradeoff based on application requirements without retraining the model.

Finally, our approach offers substantial efficiency advantages. The joint model is over $100\times$ faster than VLM-based approaches at inference time while being over $4,000\times$ smaller (23M vs. $>100B$ parameters). This efficiency makes the model suitable for latency-sensitive applications where road closures must be detected and propagated rapidly. Qualitative results and illustrative anecdotes are provided in Appendix B.

5.2 Ablation Studies

Table 2 summarizes a systematic ablation over the input modalities used by our model. Table 3 reports ablations on fusion strategy, graph architecture, reasoning depth, and message passing.

Input Modalities. The graph-only baseline achieves a non-trivial performance ($F_1 = 35.9\%$, averaged across both tasks), indicating that engineered, vehicle-aware spatial features encode meaningful information for road closure reasoning. However, this modality alone is insufficient.

Table 2: Ablation study on input modalities.

Sat.	Sat.+Proj.	Graph	Street	P / R / F1 (%)	# Params
✗	✗	✓	✗	39.4 / 32.9 / 35.9	218K
✓	✗	✗	✗	22.6 / 26.9 / 24.6	12.32M
✓	✓	✗	✗	34.0 / 30.6 / 32.2	12.32M
✓	✓	✓	✗	52.9 / 38.0 / 44.2	12.37M
✓	✓	✓	✓	58.0 / 37.5 / 45.5	23.42M

Incorporating projection of detected closure objects onto satellite imagery substantially improves performance ($24.6\% \rightarrow 32.2\%$), highlighting the importance of explicitly encoding semantic spatial

context. The projection operation introduces a strong spatial prior by localizing closure-related objects within the aerial view, allowing the model to focus on spatial reasoning rather than implicit object localization.

Combining projected satellite imagery with the road network graph yields a further improvement ($F_1 = 44.2\%$), demonstrating the complementarity between visual spatial cues and topological structure. Finally, adding street-level imagery provides a modest but consistent gain ($F_1 = 45.5\%$), capturing fine-grained appearance and perspective information unavailable in overhead views. The street encoder contributes a single CLS token ($B \times 1 \times d$) by design, as the satellite imagery with projected detections already captures spatial detail at multiple scales. Multi-token extraction from street images was considered but deemed redundant given this overlap, and would further increase the parameter count without commensurate gains. Exploring richer street-view representations (e.g., spatial token grids) remains a direction for future work. Overall, these results validate the multimodal design, where each modality contributes distinct and complementary semantic information that cannot be fully recovered from any single source.

Fusion Strategy. Table 3 compares alternative strategies for fusing graph and visual modalities. Early fusion, which concatenates features prior to processing, performs poorly ($F_1 = 39.5\%$), as it forces the model to learn complex cross-modal interactions without explicit structural guidance. Late fusion improves performance ($F_1 = 42.8\%$) by enabling modality-specific feature extraction, but still lacks fine-grained interaction during representation learning. In contrast, cross-modal attention achieves the best performance ($F_1 = 45.5\%$), providing a $+2.7$ improvement over late fusion. This confirms our design choice (Section 3.3) of allowing graph edges to query visual tokens, enabling selective alignment between candidate road segments and relevant visual evidence rather than relying on a fixed global fusion.

Graph Pooling Strategy. We compare global pooling, which averages over all M edges including padding, with edge-aware pooling

Table 3: Ablation study on fusion, graph architecture, and message passing.

Component	Variant	P / R / F1 (%)
Fusion Strategy	Early Fusion	41.2 / 38.1 / 39.5
	Late Fusion	47.8 / 38.8 / 42.8
	Cross-Attention	58.0 / 37.5 / 45.5
Graph Pooling	Global	43.4 / 37.9 / 40.4
	Edge-Aware	58.0 / 37.5 / 45.5
Graph Encoder	GAT	41.0 / 36.9 / 38.8
	GATv2	58.0 / 37.5 / 45.5
GNN Depth	1 Layer	41.4 / 37.4 / 39.3
	2 Layers	53.6 / 38.4 / 44.8
	3 Layers	58.0 / 37.5 / 45.5
	4 Layers	53.4 / 37.4 / 43.9
Message Passing	Without	43.7 / 38.0 / 40.7
	With	58.0 / 37.5 / 45.5

that aggregates only valid edges. Edge-aware pooling yields a substantial improvement of +5.1 F_1 points over global pooling (45.5% vs. 40.4%). This gain highlights the importance of explicitly handling variable-length graphs: global pooling dilutes informative representations by averaging over padded edges, particularly in sparse graphs, whereas edge-aware pooling ensures that only meaningful edge information contributes to the closure classification.

Graph Encoder Architecture. Table 3 evaluates standard GAT against GATv2 for graph encoding. GATv2 significantly outperforms GAT (45.5% vs. 38.8%), demonstrating the benefit of more expressive, dynamic attention. By applying the nonlinearity after feature concatenation (Equation 1), GATv2 enables flexible attention rankings over neighboring nodes. This expressiveness is critical for road closure reasoning, where the relevance of neighboring road segments varies with closure type, spatial context, and network topology.

GNN Depth. Table 3 examines the effect of the number of message passing layers. A single GNN layer ($F_1 = 39.3\%$) captures only immediate neighborhood information, insufficient for understanding broader road network context. Two layers (44.8%) enable 2-hop reasoning, substantially improving performance. Three layers achieve optimal results (45.5%), allowing nodes to aggregate information from a sufficiently large neighborhood to understand local topology. Four layers (43.9%) show slight degradation, likely due to over-smoothing where node representations become too similar. Further analysis of over-smoothing and message passing mechanisms is provided in Appendix C.

Message Passing. Disabling message passing—using only initial node features without neighborhood aggregation—reduces F_1 by 4.8 points (40.7% vs. 45.5%). This demonstrates that topological context is crucial: message passing enables the model to reason

about how roads connect and form network structure, and to distinguish between similar-looking roads based on their position in the network.

Class Imbalance. The natural class distribution in our dataset is imbalanced (62% no closure, 29% partial, 9% complete), which disproportionately affects Complete Closure recall. With unweighted cross-entropy, per-class accuracy is 90.0% for No Closure, 56.8% for Partial Closure, and only 0.8% for Complete Closure. To mitigate this, we train a variant using inverse-frequency weighted cross-entropy for the closure classification loss. Class weighting recovers Complete Closure accuracy from 0.8% to 49.2%, while No Closure decreases from 90.0% to 73.9% and Partial Closure from 56.8% to 46.6%. Overall, macro F_1 improves from 47.9% to 53.0% (+5.1) at the cost of reducing overall classification accuracy from 72.7% to 63.8% (−8.9 points). Segment association accuracy is largely unaffected (59.9% → 58.7%). These results confirm that the architecture can learn all closure types when given appropriate training signal, and that the tradeoff between overall accuracy and minority class recall is tunable via loss weighting without architectural changes. Further improvements through focal loss or targeted data augmentation for complete closures are noted as future work.

Error Analysis. We analyze model performance across road types, network complexity, and detection density to identify systematic failure patterns. For road segment association, performance degrades substantially as road hierarchy decreases: highways achieve 70.1% accuracy compared to 54.1% on arterials, 47.6% on residential streets, and only 36.8% on service roads. This reflects the increasing topological ambiguity in denser road networks where more candidate edges compete for selection. Network complexity, measured by the number of nodes in the local road graph, exhibits a similar effect: scenes with simple graphs (≤ 4 nodes) achieve 72.2% association accuracy versus 54.3% for moderate-complexity graphs (5–7 nodes).

For closure classification, scenes with many detection events (≥ 16) are harder to classify (69.4%) compared to scenes with few detections (74.6%, ≤ 5 events). Multiple detected objects introduce competing visual signals that increase classification ambiguity, consistent with the failure mode discussed in Appendix B. Conversely, association accuracy *improves* with more detections (66.7% vs. 54.7%), as additional visual evidence provides stronger spatial cues for localizing the affected road segment. These opposing trends suggest that detection multiplicity benefits spatial reasoning while complicating closure type inference, motivating future work on multi-object disambiguation.

6 Conclusion

We presented a multimodal, multitask framework for mapping dynamic road closures. Our architecture encodes street-level imagery, satellite imagery with projected detections, and graph-structured road networks through cross-modal attention, enabling joint reasoning over visual cues and spatial topology. The proposed model substantially outperforms general-purpose vision–language model baselines while being significantly smaller and faster. This work establishes a benchmark for joint closure classification and road segment association, demonstrating that task-specific architectural

inductive biases are essential for structured geospatial reasoning. The framework generalizes to other road attributes such as lane restrictions and turn prohibitions.

Future Work. Future directions include inferring closure end times from temporal visual cues, handling multiple simultaneous closures within a single scene, and evaluating the framework in operational settings at scale.

References

- [1] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? *arXiv preprint arXiv:2105.14491*, 2021.
- [2] Andrew Campbell, Alan Both, and Qian (Chayn) Sun. Detecting and mapping traffic signs from google street view images using deep learning and gis. *Computers, Environment and Urban Systems*, 77:101350, 2019.
- [3] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, 2009.
- [4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2019.
- [5] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiuhua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2021.
- [6] Junchi Feng, Giles Hamilton-Fletcher, Nikhil Ballem, Michael Batavia, Yifei Wang, Juling Zhong, Maurizio Porfiri, and John-Ross Rizzo. Robust computer-vision based construction site detection for assistive-technology applications. *arXiv preprint arXiv:2503.04139*, 2025.
- [7] Anurag Ghosh, Robert Tamburo, Shen Zheng, Juan R Alvarez-Padilla, Hailiang Zhu, Michael Cardei, Nicholas Dunn, Christoph Mertz, and Srinivasa G Narasimhan. Roadwork dataset: learning to recognize, observe, analyze and drive through work zones. *arXiv preprint arXiv:2406.07661*, 2024.
- [8] Raiful Hasan and Ragib Hasan. Roadnote: Automated road closure detection using urban sensing. In *2022 IEEE 8th World Forum on Internet of Things (WF-IoT)*, pages 1–6, 2022.
- [9] Songtao He, Favyen Bastani, Satvat Jagwani, Edward Park, Sofiane Abbar, Mohammad Alizadeh, Hari Balakrishnan, Sanjay Chawla, Samuel Madden, and Mohammad Amin Sadeghi. Roadtagger: Robust road attribute inference with graph neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10965–10972, 2020.
- [10] Patrick Helber, Benjamin Bischke, Andreas Dengel, and Damian Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.
- [11] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2117–2125, 2017.
- [12] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.
- [13] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016.
- [14] Brahim Remmouche, Doukifli Boukraa, Anastasia Zakharova, Thierry Bouwmans, and Mokhtar Taffar. Long-term spatio-temporal graph attention network for traffic forecasting. *Expert Systems with Applications*, 288:128244, 2025.
- [15] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.
- [16] Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.
- [17] Can Zhang and Minglong Lei. A survey on spatio-temporal graph neural networks for traffic forecasting. In *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, pages 1417–1423, 2023.
- [18] Chaoquan Zhang, Hongchao Fan, and Wanzhi Li. Automated detecting and placing road objects from street-level images. *Computational urban science*, 1(1):18, 2021.

A Attention and Transformer Details

A.1 Edge-to-Node Bootstrapping

To initialize node representations with local road context, we aggregate incoming edge features and combine them with node attributes. Given raw node features $\mathbf{x}_i^{\text{node}}$ and encoded edge features $\mathbf{e}_{ji}$ from incoming edges $j \in \mathcal{N}_{\text{in}}(i)$, the initial node embedding is computed as:

$$\mathbf{h}_i^{(0)} = \mathbf{W}_{\text{node}} \mathbf{x}_i^{\text{node}} + \mathbf{W}_{\text{e2n}} \left(\frac{1}{|\mathcal{N}_{\text{in}}(i)|} \sum_{j \in \mathcal{N}_{\text{in}}(i)} \mathbf{e}_{ji} \right), \quad (9)$$

where $\mathbf{W}_{\text{node}}$ and $\mathbf{W}_{\text{e2n}}$ are learnable linear projections. This initialization allows node embeddings to encode both intersection-level attributes and the characteristics of connected road segments prior to message passing.

A.2 Edge-Aware Graph Attention

We extend the standard GATv2 formulation [1] to incorporate edge features into the attention computation. As introduced in Section 3, given node embeddings $\mathbf{h}_i$ and $\mathbf{h}_j$ and the corresponding edge embedding $\mathbf{e}_{ij}$, the unnormalized attention score is computed as in Equation 1. The normalized attention coefficients are obtained via softmax over the neighborhood of node i :

$$\tilde{\alpha}_{ij} = \text{softmax}_j(\alpha_{ij}). \quad (10)$$

Node representations are updated by aggregating messages from neighboring nodes:

$$\mathbf{h}_i^{(\ell+1)} = \sum_{j \in \mathcal{N}(i)} \tilde{\alpha}_{ij} \mathbf{W}_v \mathbf{h}_j^{(\ell)}. \quad (11)$$

Residual connections and layer normalization are applied to stabilize training:

$$\mathbf{h}_i^{(\ell+1)} = \text{LayerNorm} \left(\mathbf{h}_i^{(\ell)} + \text{ELU}(\mathbf{h}_i^{(\ell+1)}) \right). \quad (12)$$

A.3 Cross-Modal Transformer Fusion

Cross-modal fusion is implemented using a standard transformer architecture with multi-head attention [15]. Given graph tokens as queries and visual tokens as keys and values, fusion is performed through stacked cross-attention layers.

Let $\mathbf{G} \in \mathbb{R}^{M \times d}$ denote the projected edge representations and $\mathbf{V} \in \mathbb{R}^{(L+1) \times d}$ the visual tokens from the satellite and street encoders. We prepend a learnable CLS token to the graph tokens to enable global multimodal aggregation:

$$\mathbf{Q}^{(0)} = [\text{cls} \parallel \mathbf{G}] \in \mathbb{R}^{(1+M) \times d}. \quad (13)$$

For each transformer layer ℓ , cross-attention is applied as:

$$\mathbf{Q}^{(\ell+1)} = \text{LayerNorm} \left(\mathbf{Q}^{(\ell)} + \text{MHA}(\mathbf{Q}^{(\ell)}, \mathbf{V}, \mathbf{V}) \right), \quad (14)$$

followed by a position-wise feed-forward network:

$$\mathbf{Q}^{(\ell+1)} = \text{LayerNorm} \left(\mathbf{Q}^{(\ell+1)} + \text{FFN}(\mathbf{Q}^{(\ell+1)}) \right). \quad (15)$$

Here, MHA denotes multi-head attention with H heads, and FFN is a two-layer feed-forward network with GELU activation. Through cross-attention, graph tokens attend selectively to relevant visual evidence, while the CLS token aggregates global multimodal context.

The final fused representation is obtained from the CLS token after the last transformer layer.

B Qualitative Analysis and Anecdotes

Figure 3 presents qualitative examples of successful predictions. Our model correctly identifies complete closures on major highways (row 1) and partial closures on residential streets (row 2), demonstrating its ability to reason across diverse road types and closure scenarios.

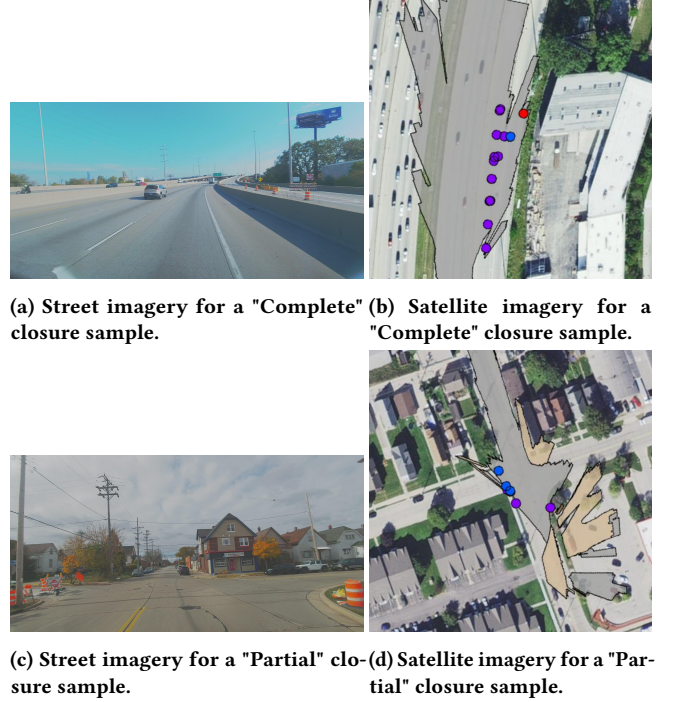
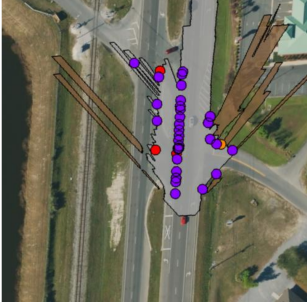


Figure 3: Two successful examples for partial and complete closures.

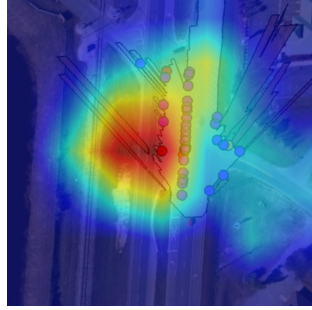
An illustrative failure case highlighting the task's complexity is shown in Figure 4. The scene contains a visually prominent line of drums, which the model has learned to associate with partial closures. Consequently, it attends to the adjacent road segment and predicts a partial closure. However, the decisive cue is a barely visible barricade located between the drums, indicating a complete closure on a different road segment. The model fails to capture this subtle signal and is instead dominated by the more salient drum pattern. This example exposes two challenges: strong learned correlations between object types and closure categories (drums $\rightarrow$ partial closure), and the difficulty of reasoning when multiple closure objects with different implications are present, where the model may focus on the more prominent ones while missing subtle but decisive indicators. These observations motivate future work on detecting and associating multiple closures within a single scene, beyond the current single-closure-per-sample assumption, as well as more expressive attention mechanisms to disambiguate competing visual evidence.



(a) Street imagery.



(b) Projected closure objects.



(c) Model activation visualization.

Figure 4: Anecdote showing an ambiguous sample.

C Additional Ablation Details

The GNN depth and message passing ablations reported in Table 3 (Section 5.2) are summarized here with additional context.

Over-smoothing at depth 4. The performance drop from 3 to 4 GNN layers ($45.5\% \rightarrow 43.9\%$) is consistent with over-smoothing, a known phenomenon in deep GNNs where repeated message passing causes node representations to converge, reducing discriminability. Given the small graph sizes in our setting ($M=10$ edges), 3 hops already cover the full local neighborhood for most samples, making deeper propagation unnecessary.

What message passing captures. Without message passing, the model relies solely on local edge and node attributes (distances, angles, road class). Enabling message passing allows each road segment to incorporate context from adjacent segments—for example, whether neighboring roads are also partially blocked, or whether the target segment connects to a dead end. This contextual reasoning accounts for the $+4.8 F_1$ improvement.

D Data Modalities and Preparation

The system integrates three complementary data modalities. We note that our contribution lies in the multimodal fusion architecture and learning framework for road closure classification and association, rather than in the data acquisition and preprocessing pipeline. Therefore, we treat segmentation, localization, and projection as preprocessing steps performed by existing methods, and focus our technical description on how these modalities are leveraged within our model. The three input modalities are:

D.1 Street-Level Imagery

Street images are captured via cameras mounted on fleet vehicles. These images undergo semantic segmentation to identify closure objects (arrow boards, barricades, cones, drums, TTC signs, vertical bars, and work vehicles) and surfaces of interest (road, sidewalk, and earth), as illustrated in Figure 2c. The road closure object categories are adapted from the ROADWork dataset [7]. Detected objects are localized in 3D using a pinhole camera model with structure-from-motion depth estimates, then projected to absolute geographic coordinates (latitude, longitude) via the vehicle’s GPS position and heading. GPS uncertainty (typically $\pm 3\text{--}5$ m in urban environments) introduces localization noise; however, the $100\text{ m} \times 100\text{ m}$ satellite crop used for aerial context provides sufficient spatial margin to accommodate this error without displacing projected objects outside the relevant image region.

D.2 Aerial Imagery

Based on the GPS coordinates where the vehicle detected a closure object, we retrieve a satellite image $I_{\text{sat}} \in \mathbb{R}^{H_s \times W_s \times 3}$ covering a $100\text{ m} \times 100\text{ m}$ area centered on the vehicle location, as shown in Figure 2a. While not real-time, these images provide crucial spatial context.

D.3 Graph-Structured Map Data

Map data (streets and intersections with attributes such as speed limits and road types, similar to OpenStreetMap format) is converted into a bidirectional graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where roads are edges and intersections are nodes. We engineer features for both:

Edge Features:

- Road class: Discrete values (10-102) indicating road type (service road, highway, urban street, etc.)
- Euclidean distance from edge’s closest point to detected closure object
- Euclidean distance from edge’s closest point to delivery vehicle
- Edge orientation angle relative to true north
- Angle between edge direction and vehicle heading

Node Features:

- Node’s (x, y) coordinates relative to delivery vehicle position
- Node degree (number of connected edges)
- Euclidean distance to detected closure object
- Euclidean distance to delivery vehicle
- Angle of node-to-object line relative to true north
- Angle of node-to-vehicle line relative to true north

These engineered features collectively encode three key aspects of the scene: the *road network structure* (road class, node degree, edge orientation), the *vehicle state* (distances and angles relative to vehicle position and heading), and the *closure object location* (distances and angles relative to detected objects). This representation enables the model to reason about spatial relationships between the road topology, the vehicle’s position, and the detected closure. All distance features use Euclidean distance, which closely approximates the geodesic (haversine) distance for the short ranges encountered in our application (typically within 100m). Feature normalization is applied as follows: distance features are min-max

normalized to $[0, 1]$, angular features are encoded as $(\sin \theta, \cos \theta)$ pairs to handle periodicity, and road class undergoes ordinal normalization where the lowest value (10, representing highways) maps to 0 and the highest value (102, representing parking aisles) maps to 1. This ordinal encoding preserves the inherent hierarchical relationship between road types, reflecting their relative importance in the road network. To manage computational complexity while maintaining task-relevant information, we limit the graph to a maximum of M edges nearest to the vehicle location. This design choice is motivated by the observation that road closures typically affect segments in the immediate vicinity of the vehicle; thus, the impacted edge is almost always among the closest edges, making distant edges unnecessary for accurate prediction.

E Hyperparameters and Model Specifications

This appendix provides detailed hyperparameters and architectural specifications for the proposed multimodal model.

E.1 Model Architecture Hyperparameters

Table 4: Model architecture hyperparameters and specifications.

Component	Parameter	Value
<i>Global Settings</i>		
	Hidden dimension (d)	512
	Dropout rate	0.2
<i>Graph Encoder (GATv2)</i>		
	Number of GAT layers	3
	Number of attention heads	8
	Node feature dimension	9
	Edge feature dimension	9
	Maximum edges per graph (M)	10
	Use edge bootstrap	True
	Use message passing	True
<i>Street Image Encoder</i>		
	Backbone architecture	ResNet-18
	Pretrained weights	ImageNet-1K
	Input resolution ($H_{st} \times W_{st}$)	224×224
	Output dimension	d (512)
<i>Aerial Image Encoder (FPN)</i>		
	Backbone architecture	ResNet-18
	Pretrained weights	EuroSAT
	Input resolution ($H_s \times W_s$)	224×224
	FPN output channels	d (512)
<i>Cross-Modal Fusion</i>		
	Fusion strategy	Cross-attention
	Number of transformer layers (N_{layers})	2
	Number of attention heads (H)	8
	Dropout	0.2

E.2 Training Hyperparameters

Table 5: Training hyperparameters and optimization settings.

Parameter	Value
Optimizer	AdamW
Learning rate	1×10^{-4}
Weight decay	1×10^{-4}
Batch size	32
Number of epochs	60
Learning rate scheduler	ReduceLROnPlateau
Scheduler patience	5 epochs
Scheduler factor	0.5
Loss weights ($\lambda_{assoc}, \lambda_{closure}$)	1.0, 1.0

E.3 Data Preprocessing

Table 6: Data preprocessing and augmentation specifications.

Parameter	Value/Method
<i>Image Normalization</i>	
Mean (RGB)	[0.485, 0.456, 0.406]
Std (RGB)	[0.229, 0.224, 0.225]
<i>Graph Features</i>	
Road class normalization	Ordinal (10-102 $\rightarrow$ 0-1)
Distance normalization	Min-max
Angle normalization	Sin and Cos
<i>Aerial Image Processing</i>	
Satellite image coverage	$100m \times 100m$

E.4 Computational Resources

Table 7: Computational resources and training time.

Resource	Specification
GPU	Tesla V100 (16GB)
Number of GPUs	1
Training time per epoch	~4 minutes
Total training time	~4 hours