UKR Journal of Multidisciplinary Studies (UKRJMS)
Homepage: https://ukrpublisher.com/ukrjms/
Email: submit.ukrpublisher@gmail.com
Volume 2, Issue 7, 2026
ISSN: 3049-432X (Online)

Text-Augmented Traffic Forecasting with Transformer-GNN and Event Explanation over 2024 California Mobility Counts

Leo Tang

Computer Science, University of California, San Diego, La Jolla, CA, USA.

***Corresponding Author**: Leo Tang

DOI: *https://doi.org/10.5281/zenodo.21627511*

Article History

Original Research Article

Received: 01-06-2026

Accepted: 05-07-2026

Published: 27-07-2026

Citation: Tang, L. (2026). *Text-augmented traffic forecasting with Transformer-GNN and event explanation over 2024 California mobility counts. UKR Journal of Multidisciplinary Studies (UKRJMS), 2*(7), 153-171.

Abstract

Hourly counts from bicycle and pedestrian sensors are increasingly important for daily operations, safety monitoring, and near-term planning, yet these series are sparse, highly zero-inflated, and sensitive to time-of-day context. This study develops a text-augmented Transformer-GNN workflow for short-term active-mobility forecasting on 2024 California hourly bicycle and pedestrian counts. Directional movements are aggregated to mode-location node series, a 5-nearest-neighbor geographic graph is built from site coordinates, and event texts are generated from calendar and operating-context tags such as overnight low demand, weekday commute access, weekend recreation, and holiday schedules. The forecasting model combines scaled dot-product temporal attention, graph-diffused count values, structured text tags, and a validation-calibrated decoder. Baselines include persistence, daily and weekly seasonal forecasts, temporal Ridge regression, graph Ridge regression, and an Elastic Net text model. Experiments use a chronological split with January-August for fitting, September for calibration, and October-December for testing at 1, 3, and 6 hour horizons. Across 46 high-coverage nodes and 101,568 node-hours in the test period, the Text-Transformer-GNN achieves the lowest MAE at all horizons: 2.393 for 1 hour, 2.613 for 3 hours, and 2.656 for 6 hours. Relative to the strongest seasonal baseline, these values reduce MAE by 12.5%, 14.7%, and 13.3%, respectively. Error analysis shows that high-volume pedestrian sites dominate residual risk, while event-tagged explanations clarify whether forecasts are driven by recent persistence, prior-day seasonality, graph-neighbor demand, or text-defined operating context.

Keywords: *Active transportation counts; bicycle and pedestrian forecasting; Transformer; graph neural network; event text; traffic operations; California mobility data; explainable forecasting.*

Introduction

Short-term traffic forecasting has long been a core function of intelligent transportation systems. Early work treated traffic as an autoregressive or seasonal time series, often emphasizing roadway flow stability, recurrent peak periods, and the need to forecast the next operational interval from recent measurements [1], [2], [3]. Later studies expanded the problem to large-scale sensor data and deep sequence models, showing that nonlinear temporal dependencies can improve performance when volume data are abundant [4], [5]. These advances were first shaped around motor-vehicle freeway detectors, but the same operational need now appears in active transportation: agencies need timely estimates of bicycle and pedestrian demand for corridor monitoring, signal timing, detour management, seasonal planning, and safety exposure analysis.

Active-mobility counts differ from freeway traffic in ways that complicate direct transfer of conventional traffic models. Bicycle and pedestrian series often contain many zero-count hours, heavy site-to-site heterogeneity, sharp midday or commute pulses, and occasional very high counts at a small number of locations. Count collection guidance has stressed the importance of standardized pedestrian and bicycle volume data, consistent movement

labels, temporal expansion, and rigorous quality checks [6], [7]. Empirical studies of non-motorized infrastructure have also shown that local context, weather, facility type, day of week, and temporal sampling strongly affect count interpretation [8], [9], [10]. Recent probabilistic bike-sharing evidence further shows that weather and seasonal regime shifts affect both point forecasts and interval reliability in Transformer-based models [11]. A model that ignores these operating contexts may have acceptable aggregate error but produce weak explanations for the hours that matter to practitioners.

Graph neural networks and attention models provide a natural starting point for this setting. Graph convolution links nearby or functionally related sensors, allowing a forecast at one site to borrow information from spatial neighbors [12], [13], [14], [15], [16], [17]. Transformers add a flexible attention mechanism that can emphasize the most relevant historical hours rather than treating the last day or last week as a fixed seasonal template [18], [17], [19]. Recurrent models such as LSTMs remain influential for sequence learning [20], but attention-based architectures are especially attractive when the model must connect a target hour to several possible temporal analogues. For active transportation, those analogues are not only numeric. A Saturday midday count, a weekday commute hour, and an overnight holiday hour may have similar recent lags but very different operational meanings.

This study therefore treats event text as a structured input rather than a narrative afterthought. Calendar and operating-context tags are converted into concise text strings, encoded as features, and used in the attention query and decoder. The same tags also ground a language-model explanation layer. Prior work on language representation and large language models shows that text can provide compact context for downstream reasoning [21], [22], while explainable-machine-learning research emphasizes that explanations should be faithful to the model inputs and local evidence rather than generic descriptions [23], [24]. The proposed workflow follows that principle: every explanation is restricted to the forecast horizon, the site and mode, the relevant text context, the graph-neighbor signal, and the relation between the forecast and seasonal baselines.

The empirical contribution is a full evaluation on 2024 California bicycle and pedestrian hourly counts. The study uses the full directional records for descriptive analysis and a high-coverage node panel for forecasting. The task is to predict aggregated hourly volume at each mode-location node 1, 3, and 6 hours ahead. The evaluation reports model comparisons, horizon-level winners, ablations, mode-district residuals, event-context errors, calibration weights, and operational explanation examples. This design tests not only whether a text-augmented graph attention model can reduce numeric error, but also whether its outputs can be summarized in a way that is useful for transportation operations. Urban-mobility digital-twin research has coupled spatio-temporal Transformers with offline reinforcement learning to connect prediction with dispatch decisions [25]. The present study isolates the upstream forecasting and explanation stages before any control policy is optimized.

The California count setting is also useful because it contains the practical irregularities that are often hidden in benchmark traffic datasets. The same year includes low-volume overnight hours, high pedestrian surges, mode differences, district differences, and site histories with small gaps. It therefore allows the model to be tested as an operations tool rather than as a clean academic sequence problem. The design keeps the forecasting target simple, but it evaluates several forms of evidence that agencies would use when deciding whether to trust a short-term forecast: whether it beats a persistence rule, whether it improves over prior-day and prior-week baselines, whether the improvement persists beyond one hour, whether errors concentrate at known high-volume groups, and whether the explanation text matches the model inputs. This emphasis on operational coherence is the reason the analysis reports both global scores and stratified residuals.

A second motivation is that active-mobility forecasting should not simply copy freeway forecasting practice. PeMS-style vehicle studies often focus on dense freeway detectors, speed, flow, and occupancy, where spatial dependence follows the roadway network and recurrent congestion waves are prominent [1], [2], [3], [4], [5]. Bicycle and pedestrian counts have different noise sources: the count level is lower, zero observations are common, direction labels can be irregular across sites, and the same absolute error has different operational meaning at a quiet crossing than at a major pedestrian corridor. The method therefore uses freeway forecasting ideas only where they transfer cleanly--chronological evaluation, temporal baselines, attention over recent histories, and graph smoothing--while keeping the target and explanations anchored in the active-count fields available for every site. This choice supports fair comparison against simple rules that an agency could deploy immediately.

Related Work

Cross-Domain Forecasting and Operational Calibration

Cross-domain forecasting studies offer useful design analogies when their evidence is not mistaken for transportation validation. Few-shot time-series foundation models address new AI-inference tenants with limited

histories [26], while news-macro fusion models make uncertainty explicit in directional VIX prediction [27]. Cost-sensitive GPU admission work connects predicted demand to operational loss [28], and constrained budgeting joins macro demand forecasts with downstream allocation rules [29]. Together, these studies motivate horizon-specific validation, coverage-aware transfer, and a clear separation between forecasts and the decisions that consume them.

Resource forecasting adds emphasis on feature semantics and capacity consequences. Bidirectional recurrent encoding with Transformer-style attention has been used to predict task-level GPU memory requirements [30], power-aware inventory planning maps workload forecasts to resource actions [31], and education early-warning models couple prediction with teacher-facing explanations [32]. Privacy-robust incrementality modeling tests whether decisions remain useful when user-level identifiers are unavailable [33]. These studies support evaluation under constrained observability rather than direct claims about bicycle or pedestrian dynamics.

Transfer, Representation, and Robustness

Cross-domain representation learning also shows why modular encoders must be evaluated for transfer rather than novelty alone. A lightweight medical foundation model uses multimodal pretraining and parameter-efficient transfer [34], wearable activity work evaluates cross-dataset transfer to volleyball settings [35], and self-supervised Conformer BCI research targets subject-independent calibration [36]. Edge video-quality distillation compresses a high-cost metric into a deployable proxy [37], while uncertainty-aware medical vision-language classification separates prediction from confidence [38]. Their domains differ, but the shared concerns of transfer, compactness, calibration, and held-out evaluation inform the ablation design used here.

Robust perception and security studies supply complementary stress-test patterns. Attention-enhanced autonomous-driving detection focuses on heterogeneous visual evidence [39], DenseNet cancer-image classification provides a conventional end-to-end image baseline [40], and predictive DDoS mitigation links learned risk to distributed-system response [41]. Language-guided intrusion feature selection examines whether semantic guidance improves tabular detection [42], whereas CloudTrail sequence modeling combines interpretable linear features with HMM smoothing for attack-stage inference [43]. These papers motivate transparent feature contribution and robustness checks; none is treated as active-mobility evidence.

Evidence-Grounded Explanation and Reasoning

Evidence-grounded generation requires a strict boundary between available support and fluent prose. Evidence-calibrated RAG links retrieval coverage to abstention [44], budgeted multi-hop retrieval treats evidence acquisition as a constrained policy [45], and evidence-chain RAG emphasizes source attribution and deterministic provenance [46]. Trust-calibrated multilingual RAG separates retrieved support from answer confidence in a public-information setting [47], while financial search research separates query rewriting, hybrid retrieval, and listwise reranking [48]. The present system does not retrieve documents, but it adopts the same support boundary: explanations may verbalize only computed temporal, spatial, seasonal, and calendar-tag fields.

Other work focuses on evidence integrity under domain constraints. Numerical guardrails for SEC and FRED analysis distinguish verified arithmetic from language fluency [49], federated topic-preference learning combines knowledge grounding with differential privacy [50], and retrieval-quality prediction evaluates RAG evidence before response generation [51]. Accounting-aware retrieval constrains due-diligence claims to disclosure evidence [52], while multi-regulation legal RAG formalizes cross-border governance boundaries [53]. Spectral rank aggregation provides a mathematically different but relevant reminder that evidence combination itself requires validation [54]. Accordingly, every number in the explanations is computed before any sentence is rendered.

Human-Facing Evidence Communication

Human-facing evidence must be organized so that uncertainty and provenance remain inspectable. Scientific-search evidence cards structure source hierarchy [55], explainable e-commerce recommendation cards pair confidence with product evidence [56], and text-grounded design-rationale interfaces translate structured metadata into reviewable decisions [57]. Counseling copilots rank multi-turn evidence for therapist review [58], while incident-visualization cards convert distributed logs into actionable SRE summaries [59]. Advertising brief cards make creative assumptions explicit [60], counseling recommendation cards link visual advice to source dialogue [61], and breast-ultrasound explanation cards separate diagnostic support from communicated uncertainty [62]. These interface patterns motivate concise forecast cards but do not establish usability in a transportation setting.

Comparable communication principles appear in risk dashboards, design critique, and controlled decision support. Financial-risk dashboards distinguish visual hierarchy from model correctness [63], whereas dark-pattern detection shows that short interface microcopy can

encode meaningful behavioral semantics [64]. Designer-editable advertising briefs keep generated output revisable [65], and LLM-as-design-critic work measures alignment between machine feedback and human judgment [66]. Prompt-injection risk cards expose privacy and data-integrity failures to reviewers [67], conservative off-policy evaluation separates historical-policy evidence from new-policy selection [68], and a DevOps copilot evaluates both runbook retrieval and command safety [69]. Finally, uncertainty quantification for spectral estimators illustrates the value of reporting reliability separately from a point estimate [70]. The present workflow carries these principles forward through deterministic tags, validation-only calibration, explicit limitations, and evidence-bounded language.

Method

The input data consist of two hourly CSV files for 2024, one for bicycle counts and one for pedestrian counts. Each record contains a location identifier, Caltrans district, latitude, longitude, timestamp, mode, movement direction, count, and count type. Directional movements are first aggregated to one hourly count per mode-location node because the forecasting objective is total demand at the instrumented site. A node is defined as a unique pair of mode and location identifier, so a bicycle and pedestrian sensor at the same geographic location remain distinct nodes. The full raw data contain 605,930 bicycle directional rows and 1,158,488 pedestrian directional rows. They cover six districts and the complete leap-year hour range from 2024-01-01 00:00 through 2024-12-31 23:00.

Table I. Dataset composition before direction-to-node aggregation

Mode	Directional rows	Locations	Districts	Earliest hour	Latest hour	Directions	Total counted travelers	Mean directional hourly count	Zero directional rows (%)
Bicycle	605,930	29	3, 4, 5, 6, 8, 12	2024-01-01 00:00	2024-12-31 23:00	4	714,059	1.178	70.320
Pedestrian	1,158,488	47	3, 4, 5, 6, 8, 12	2024-01-01 00:00	2024-12-31 23:00	4	2,167,289	1.871	65.130

Forecasting is performed on a high-coverage panel to ensure that each test metric is based on comparable hourly histories. Nodes with at least 99% of the 8,784 possible 2024 hours are retained. This criterion keeps 46 nodes, including 15 bicycle nodes and 31 pedestrian nodes. After reindexing each node to the common hourly calendar, 53 residual missing node-hour cells are filled by limited time interpolation and zero fill. This is only 0.013% of all retained node-hour cells and affects short gaps after the coverage screen. The retained panel contains 2,122,626 counted travelers across the selected nodes. Table II summarizes the retained panel and graph construction.

Table II. Forecasting panel and graph summary

Item	Selected_nodes	Mean_coverage_pct	Total_count	Mean_node_hour_count	Edges	Mean_degree	Filled_cells
Bicycle	15	99.989	382,491	2.903			
Pedestrian	31	99.986	1,740,011	6.391			
Graph	46	99.989	2,122,502	5.253	146	7.350	53

The spatial graph is built from site coordinates. Pairwise haversine distance is computed for all retained nodes, and each node is connected to its five nearest neighbors. Edge weights use a Gaussian distance kernel with bandwidth equal to the median nonzero inter-node distance, and self-loops are added before symmetric normalization. Cross-mode links at nearby or identical coordinates allow bicycle and pedestrian demand to inform one another while preserving mode-specific node identities. This graph is not intended to represent a full route network; it is a compact empirical dependency structure that supports graph-smoothed values in the forecasting model.

The forecasting problem is defined by an origin time and a target horizon h in {1, 3, 6}. For each node and target hour, the available predictors include recent lags at the forecast origin, same-hour prior-day and prior-week lags, six-hour and twenty-four-hour rolling means, graph-diffused versions of recent and prior-day counts, cyclic hour/day/month variables, and event text tags. The chronological split is fixed: January through August is used for parameter fitting, September for validation and calibration, and October through December for final reporting. Chronological evaluation avoids leakage from future seasonal patterns into earlier forecasts and follows

standard time-series assessment practice [71]. The coverage screen is reported explicitly because few-shot forecasting research treats short-history entities as a distinct generalization problem [26], not as ordinary missing values.

Table III. Chronological split used for all horizons

Split	Start	End	Hours	Role
Train	2024-01-01 00:00	2024-08-31 23:00	5,856	Parameter fitting
Validation	2024-09-01 00:00	2024-09-30 23:00	720	Model selection and sanity checks
Test	2024-10-01 00:00	2024-12-31 23:00	2,208	Final reporting

For each horizon, the supervised table contains one row for every retained node and target hour after the lookback window is available. The October-December test period contains 2,208 target hours and 46 nodes, yielding 101,568 evaluated node-hour forecasts per horizon. The same test timestamps are used by all models, so pairwise comparisons are not affected by sample differences. The one-week attention window starts after the first 168 hours; the January-August training period remains long enough to include winter, spring, and summer patterns before the validation month. Numeric count features are kept on the count scale for baselines and log-transformed for learned decoders, a compromise that preserves interpretability while reducing the influence of very large pedestrian peaks.

Event text is generated deterministically from the target timestamp. The tag set includes overnight low demand, weekday morning access, weekday evening return, midday local activity, weekend daytime recreation, holiday schedule, school/workday period, and late evening taper. For example, a Sunday 13:00 target receives the weekend recreation context, while a weekday 08:00 target receives the morning access context unless it falls on a holiday. These short texts act as interpretable tokens for the attention query and as grounding phrases for the explanation layer. The text vocabulary is intentionally small because the CSV files do not contain weather, incidents, construction, or special-event feeds; adding external labels would make the feature set inconsistent with the observed data.

The supervised feature table is constructed without future information. For a target at time t+h, every lag, rolling statistic, graph-smoothed value, and text tag is computed from the information available at forecast origin t or from the target calendar description. The target calendar is known in advance, but the target count is not. Rolling means use closed historical windows, and prior-day and prior-week features are read from t+h-24 and t+h-168 when those hours are available before the target. Counts are modeled with log1p transformations inside learned decoders because the retained panel contains both many zeros and occasional large pedestrian peaks. Predictions are transformed back to counts and clipped at zero before metric calculation.

The proposed Text-Transformer-GNN uses a lightweight scaled dot-product attention operator over the last 168 hourly keys and values. The query is the target-hour time and text embedding. The keys are historical time and text embeddings from the available lookback window. Values are log-transformed node counts and graph-diffused log counts. The attention outputs therefore answer a simple question: among the last week of hours, which historical hours have the closest temporal and text context to the target hour, and what demand pattern did those hours exhibit locally and in the graph neighborhood? The decoder then combines attention outputs with lag, rolling, graph, categorical node, district, and mode features. This formulation follows the attention logic of Transformer models [18], the spatial smoothing principle of graph convolution [12], [13], [14], [15], [16], [17], and the interpretable multi-horizon motivation of temporal fusion models [19].

Several baselines are included. Persistence predicts that the target count equals the most recently observed count at the forecast origin. Daily seasonal and weekly seasonal baselines use the same target hour one day and one week earlier. Ridge-temporal uses lag and calendar features without graph or text attention. Ridge-graph adds graph-diffused lag features. ElasticNet-text adds structured text tags but not the attention operator. Transformer-GNN includes graph attention but not text tags. Text-Transformer-GNN includes both graph attention and text tags. All linear decoders are fit on log1p counts and then transformed back to the count scale. Learned forecasts are bounded by node-specific 99.5th percentile caps estimated from the training period, a practical guardrail for sparse count series where unconstrained linear extrapolation can produce implausible peaks.

Table IV. Model settings used in the experiments

Component	Setting
Panel selection	node coverage >= 99% of 8,784 hours
Forecast horizons	1, 3, and 6 hours ahead
Graph	5-nearest-neighbor haversine graph with Gaussian distance weights
Attention window	168 hourly keys and values
Tabular lags	1, 2, 3, 24, 48, and 168 hours plus 6-hour and 24-hour rolling means
Target transform	log1p for linear decoders
Transformer-GNN decoder	Ridge decoder over attention, graph, lag, and categorical node features
Operational calibration	node-specific 99.5th percentile cap and validation-selected convex blend with persistence, daily, and weekly baselines
Text tags	holiday, night, commute, midday, weekend recreation, school/workday, late evening

The final operational forecast is validation-calibrated. For each horizon, September data are used to choose a convex blend of four components: the attention model, persistence, daily seasonality, and weekly seasonality. The blend weights are selected by a grid search that minimizes validation MAE and then are applied unchanged to the October-December test period. This calibration step is common in operational forecasting because very short horizons often benefit from persistence, while longer horizons often benefit from daily or weekly seasonality. Keeping the weights horizon-specific but validation-only avoids test-period tuning. This separation of model fitting, calibration, and final evaluation is consistent with uncertainty-aware directional forecasting [27], privacy-robust decision estimation [33], conservative policy selection [68], and formal uncertainty quantification [70].

Table V. Validation-selected calibration weights for attention models

Horizon	Model	Validation MAE	Attention model	Persistence	Daily	Weekly
1	Transformer-GNN	2.296	0.250	0.200	0.250	0.300
1	Text-Transformer-GNN	2.224	0.400	0.150	0.200	0.250
3	Transformer-GNN	2.424	0.350	0	0.300	0.350
3	Text-Transformer-GNN	2.322	0.550	0	0.200	0.250
6	Transformer-GNN	2.440	0.350	0	0.300	0.350
6	Text-Transformer-GNN	2.347	0.550	0	0.200	0.250

Evaluation uses MAE, RMSE, sMAPE, MASE, and R2. MAE is the primary operational score because it has the same unit as the count target. RMSE gives additional penalty to large peak errors, which is important for high-volume pedestrian locations. MASE scales MAE by the average in-sample daily seasonal error, allowing horizons and model families to be compared against a common naive reference [71]. R2 indicates how much variance is captured on the test period. sMAPE is included because it is common in forecasting reports, but it is interpreted cautiously: many active-mobility hours have observed counts of zero or one, so even small absolute differences can generate high percentage error.

All models are estimated on the same node-hour rows for a given horizon. This keeps the comparison focused on model structure rather than on differences in usable observations. Hyperparameters are intentionally compact: Ridge and ElasticNet penalties are selected from a small validation

grid, the graph uses five nearest neighbors, the attention lookback is one week, and the operational blend is selected by validation MAE. The choice favors transparent experimental control over architecture search. It also makes the numerical results easier to audit because every table is derived from a fixed split, a fixed node panel, and deterministic feature generation.

The explanation layer converts model facts into concise operational sentences. Each sentence receives the target context, forecast, observed value for evaluation examples, daily baseline, and whether the calibrated forecast is higher or lower than the seasonal reference. The generated explanations do not introduce unsupported external events. This restriction follows the support-sufficiency logic of evidence-calibrated RAG [44], deterministic provenance [46], numerical guardrails [49], and retrieval-quality assessment [51]. They state, for instance, that a forecast used a weekend recreation context and recent graph-neighbor signal, or that the model placed demand lower than the prior-day baseline. This constrained design aligns with local explanation principles [23], [24] while using language-model style text generation grounded by structured inputs [21], [22].

Results and Discussion

The retained panel reflects the sparse but operationally meaningful character of active transportation counts. Bicycle nodes have a mean hourly node count of 2.90 and a median of zero, while pedestrian nodes have a mean of 6.39 and a median of one. The pedestrian panel contains substantially more total counted travelers, and the highest hourly node count in the retained series reaches 740. Zero-count node-hours remain frequent even after aggregation, representing 53.5% of bicycle node-hours and 40.7% of pedestrian node-hours. These values explain why MAE and RMSE are more stable primary measures than percentage error alone: a forecast can be numerically close but still receive a high sMAPE when both observed and predicted counts are near zero.

Table VI. Mode-level count profile in the retained forecasting panel

mode	nodes	total_count	mean_hourly_node_count	median_hourly_node_count	p95_hourly_node_count	max_hourly_node_count	zero_node_hours_pct
Bicycle	15	382,493	2.903	0	13	123	53.527
Pedestrian	31	1,740,133	6.390	1	28	740	40.728

District-mode variation is also substantial. District 5 has only one selected bicycle node and two selected pedestrian nodes, yet it contributes a large share of the total pedestrian volume. District 4 contributes many Bay Area nodes with moderate bicycle and pedestrian volumes, while District 3 contributes many pedestrian nodes with lower average counts. This heterogeneity is important for interpreting aggregate model scores: a small number of high-volume pedestrian hours can dominate RMSE, whereas the median active-mobility hour remains low. The graph model is therefore evaluated both overall and by mode-district group.

Figures 2 and 3 show why the graph and temporal components are complementary. The site map places retained nodes across multiple districts rather than in one compact network, so a purely spatial method would not be sufficient. The diurnal profile, in contrast, shows strong recurrent hourly structure: pedestrian activity rises during daytime and bicycle activity is most concentrated around access and recreation periods. These patterns justify a model that keeps explicit seasonal baselines while allowing attention to select the most relevant recent analogues. They also explain why model evaluation must include district and mode breakdowns; a statewide aggregate score can hide whether gains come from the many low-count hours or from the smaller set of high-impact pedestrian peaks.

Table VII. District and mode profile for the retained nodes

district	mode	nodes	total_count	mean_hourly_node_count	p95_hourly_node_count
3	Bicycle	1	2,645	0.301	1
3	Pedestrian	16	318,604	2.267	9
4	Bicycle	13	295,424	2.587	11
4	Pedestrian	13	797,147	6.981	29
5	Bicycle	1	84,424	9.611	35
5	Pedestrian	2	624,382	35.541	139

The main model comparison is shown in Table VIII for the one-hour horizon. Text-Transformer-GNN achieves the lowest one-hour MAE, RMSE, MASE, and R2 among all tested methods. Its MAE is 2.393 counts per node-hour, compared with 2.736 for persistence, 3.062 for the daily seasonal baseline, and 3.387 for the weekly seasonal baseline. The MAE reduction relative to the strongest one-hour baseline is 12.5%. The non-text Transformer-GNN also improves on persistence, but the text-augmented version lowers MAE by a further 2.7% relative to the non-text attention model. The sMAPE values for learned models are higher than those of seasonal baselines because many low-volume hours create large symmetric percentage penalties; however, the absolute-count metrics and R2 show that the calibrated attention models better track operational magnitude.

Table VIII. One-hour forecast comparison on the October-December test period

Model	MAE	RMSE	sMAPE	MASE	R2
Persistence	2.736	7.515	66.755	0.861	0.772
Daily seasonal	3.062	9.329	68.096	0.963	0.649
Weekly seasonal	3.387	11.956	69.736	1.065	0.424
Ridge-temporal	3.022	10.130	108.315	0.951	0.586
Ridge-graph	3.015	10.145	107.891	0.948	0.585
ElasticNet-text	2.985	9.976	105.799	0.939	0.599
Transformer-GNN	2.458	7.288	112.637	0.773	0.786
Text-Transformer-GNN	2.393	7.048	111.452	0.753	0.800

The one-hour result is especially notable because persistence is a difficult baseline at such a short horizon. When counts are low and slowly changing, simply carrying forward the last observed value is often hard to beat. The calibrated attention model improves on persistence because it does not abandon recent demand; the validation weights retain a persistence component while allowing text-conditioned attention and seasonal components to adjust the forecast when the target context differs from the most recent hour. This matters for transitions such as the morning ramp-up, late-evening taper, and weekend midday periods, where a pure persistence rule tends to lag behind the change in activity.

The RMSE and R2 columns reinforce the same conclusion. One-hour RMSE falls from 7.515 for persistence to 7.048 for Text-Transformer-GNN, and R2 rises from 0.772 to 0.800. The improvement is smaller in RMSE than in MAE because the largest pedestrian peaks remain difficult for all models, but it is still important because the calibrated attention forecast reduces routine errors without creating a large penalty at high-count hours. The percentage metric behaves differently. sMAPE is lower for seasonal baselines because zero-heavy series can punish small learned-model deviations severely. For operational count forecasting, this is why the discussion treats sMAPE as a diagnostic rather than as the primary ranking criterion.

The horizon-level comparison in Table IX shows that the proposed model remains strongest as the forecast horizon increases. At three hours, daily seasonal forecasting becomes the strongest simple baseline with MAE 3.062, but Text-Transformer-GNN lowers MAE to 2.613, a 14.7% reduction. At six hours, the same daily baseline remains strong, yet Text-Transformer-GNN obtains MAE 2.656, a 13.3% reduction. These improvements indicate that text-conditioned attention helps the model select relevant analogues from the previous week rather than relying on a single fixed prior-day count. The calibration weights also support this interpretation: the model component receives a weight of 0.40 at one hour and 0.55 at three and six hours, while daily and weekly seasonal components remain useful stabilizers. Reporting each horizon separately also follows the broader operational lesson that a forecast should be assessed at the decision scale where it will be used, whether for admission control [28], inventory planning [31], or early-warning intervention [32].

Table IX. Top three models by forecast horizon

Horizon_h	Model	MAE	RMSE	MASE	R2
1	Text-Transformer-GNN	2.393	7.048	0.753	0.800
1	Transformer-GNN	2.458	7.288	0.773	0.786

1	Persistence	2.736	7.515	0.861	0.772
3	Text-Transformer-GNN	2.613	8.317	0.822	0.721
3	Transformer-GNN	2.743	8.908	0.863	0.680
3	Daily seasonal	3.062	9.329	0.963	0.649
6	Text-Transformer-GNN	2.656	8.736	0.836	0.692
6	Transformer-GNN	2.764	9.081	0.870	0.668
6	Daily seasonal	3.062	9.329	0.963	0.649

The calibration weights reveal how the best model changes with horizon. At one hour, the text attention component receives weight 0.40, with the remaining mass divided among persistence, daily, and weekly references. At three and six hours, the attention component increases to 0.55 and persistence drops to zero, while daily and weekly seasonal information remains in the blend. This pattern matches the transportation intuition that the last observed count is useful for immediate continuity, whereas same-hour analogues from previous days become more valuable as the horizon extends. The weights are not hand selected; they are chosen on September validation data and then fixed for the test period.

The ablation in Table X isolates the value of graph attention and text. Ridge-temporal and Ridge-graph are close, suggesting that raw graph lags alone are not enough to solve the sparse active-count problem. The non-text Transformer-GNN reduces one-hour MAE from 3.015 for Ridge-graph to 2.458. Adding text tags reduces MAE further to 2.393 and increases R2 from 0.786 to 0.800. The graph and attention operator produce the largest absolute gain, while the text tags provide a smaller but consistent refinement. This pattern is reasonable because time-of-day and day-of-week already capture much of the recurrent structure, whereas text tags mainly disambiguate holidays, weekend recreation periods, overnight hours, and commute-like intervals.

Table X. One-hour ablation of graph, attention, and text components

Model	MAE	RMSE	MASE	R2
Ridge-temporal	3.022	10.130	0.951	0.586
Ridge-graph	3.015	10.145	0.948	0.585
Transformer-GNN	2.458	7.288	0.773	0.786
Text-Transformer-GNN	2.393	7.048	0.753	0.800

Mode-district residuals show where operational risk remains. Bicycle nodes in Districts 3 and 4 have low MAE because their volumes are small and peaks are modest. District 5 bicycle demand has higher error, reflecting a single higher-volume bicycle node. Pedestrian District 5 has the largest MAE at 15.25 counts per node-hour, but its R2 remains 0.781 because the model captures much of the large variation at those sites. This result cautions against interpreting a single global MAE as uniform performance. For active transportation programs, high-volume pedestrian nodes may require additional features, such as weather, school calendars, local events, or facility-specific metadata, if the goal is to explain extreme peaks rather than forecast typical operations.

Table XI. Text-Transformer-GNN one-hour error by mode and district

Mode	District	MAE	RMSE	MASE	R2	N
Bicycle	3	0.605	1.446	0.190	0.070	2,208
Bicycle	4	0.922	1.958	0.290	0.762	28,704
Bicycle	5	5.893	9.546	1.854	0.616	2,208
Pedestrian	3	1.889	5.560	0.594	0.294	35,328
Pedestrian	4	2.374	5.416	0.747	0.713	28,704
Pedestrian	5	15.252	25.161	4.798	0.781	4,416

The event-context table provides a second view of model behavior. Overnight hours have the lowest MAE because most counts are near zero, while weekday evening return and midday local activity have the highest errors because they include both regular demand and occasional sharp peaks. Weekend recreation contexts produce intermediate error and a mean observed count above seven, showing that weekend activity is not simply a low-volume condition. The text labels therefore serve two purposes: they improve the attention query, and they organize residual analysis in categories that traffic operators can understand without inspecting every site-hour prediction.

Table XII. One-hour Text-Transformer-GNN error by text-defined operating context

Event context	MAE	RMSE	R2	N	Mean observed
holiday schedule; overnight low demand	0.518	1.401	-0.140	1,104	0.312
overnight low demand	0.621	2.555	0.453	24,288	0.606
holiday schedule; late evening taper	0.959	3.125	0.715	736	1.230
late evening taper	1.505	5.406	0.769	16,192	2.334
holiday schedule	1.563	3.499	0.832	368	1.717
ordinary hourly demand	2.211	6.912	0.663	8,096	3.190
holiday schedule; weekday morning access	2.772	5.493	0.728	552	4.745
holiday schedule; weekday evening return	3.055	6.043	0.833	552	5.306
holiday schedule; midday local activity	3.182	5.873	0.819	1,104	6.758
weekday morning access	3.301	7.727	0.805	8,556	7.868

The figures reinforce the tabular findings. Fig. 1 shows the workflow from raw hourly counts to graph construction, attention forecasting, and explanation. Fig. 2 confirms that retained nodes span several California regions, while Fig. 3 shows clear diurnal structure for both modes. The model comparison in Fig. 4 highlights the MAE advantage of the text-augmented attention model. The one-week trace in Fig. 5 illustrates how the calibrated forecast follows day-to-day variation while avoiding the largest prior-day carryover errors. Fig. 6 shows that most error concentration occurs in high-volume pedestrian groups, and Fig. 7 shows that the most challenging contexts are active daytime periods rather than overnight hours.

Figure 1. Text-augmented Transformer-GNN workflow

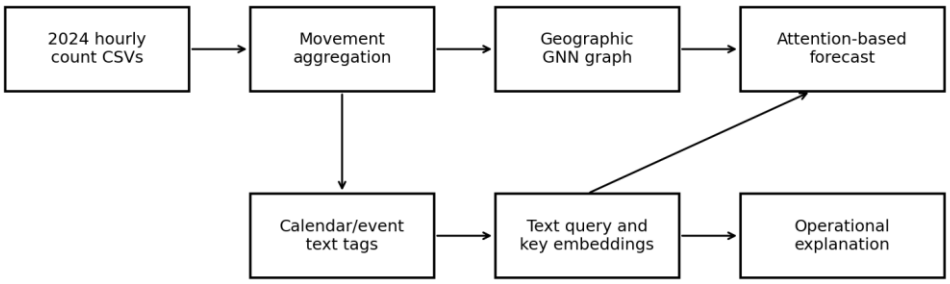

Fig. 1. Text-augmented Transformer-GNN workflow.

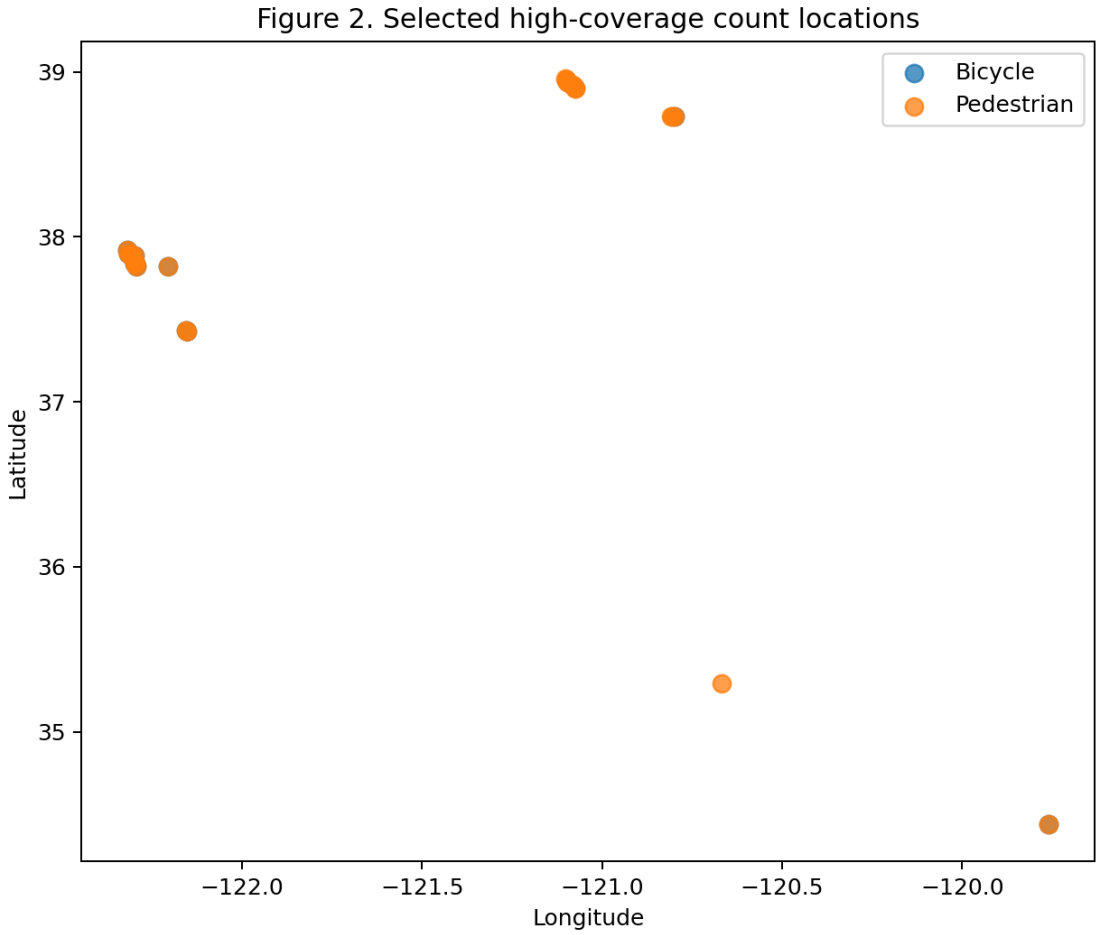

Fig. 2. Geographic distribution of the selected high-coverage count locations.

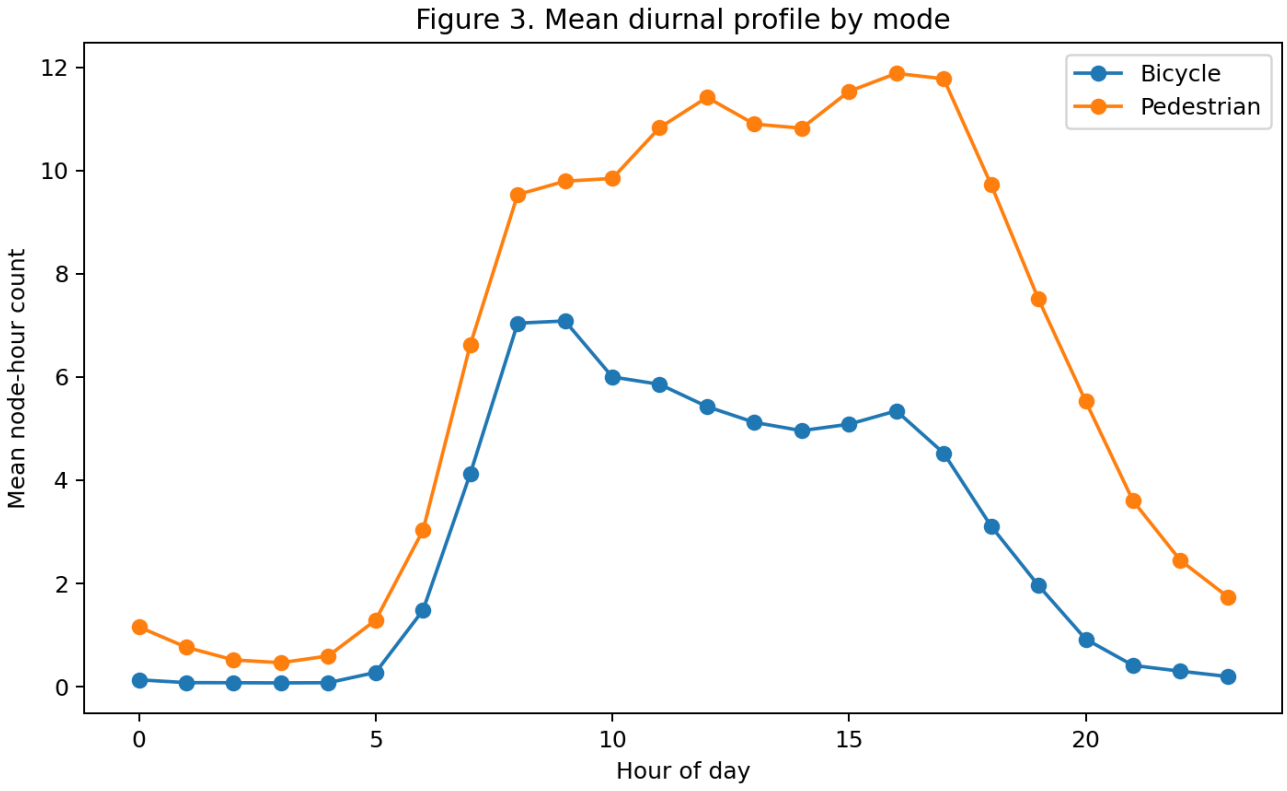

Fig. 3. Mean diurnal profile by mode.

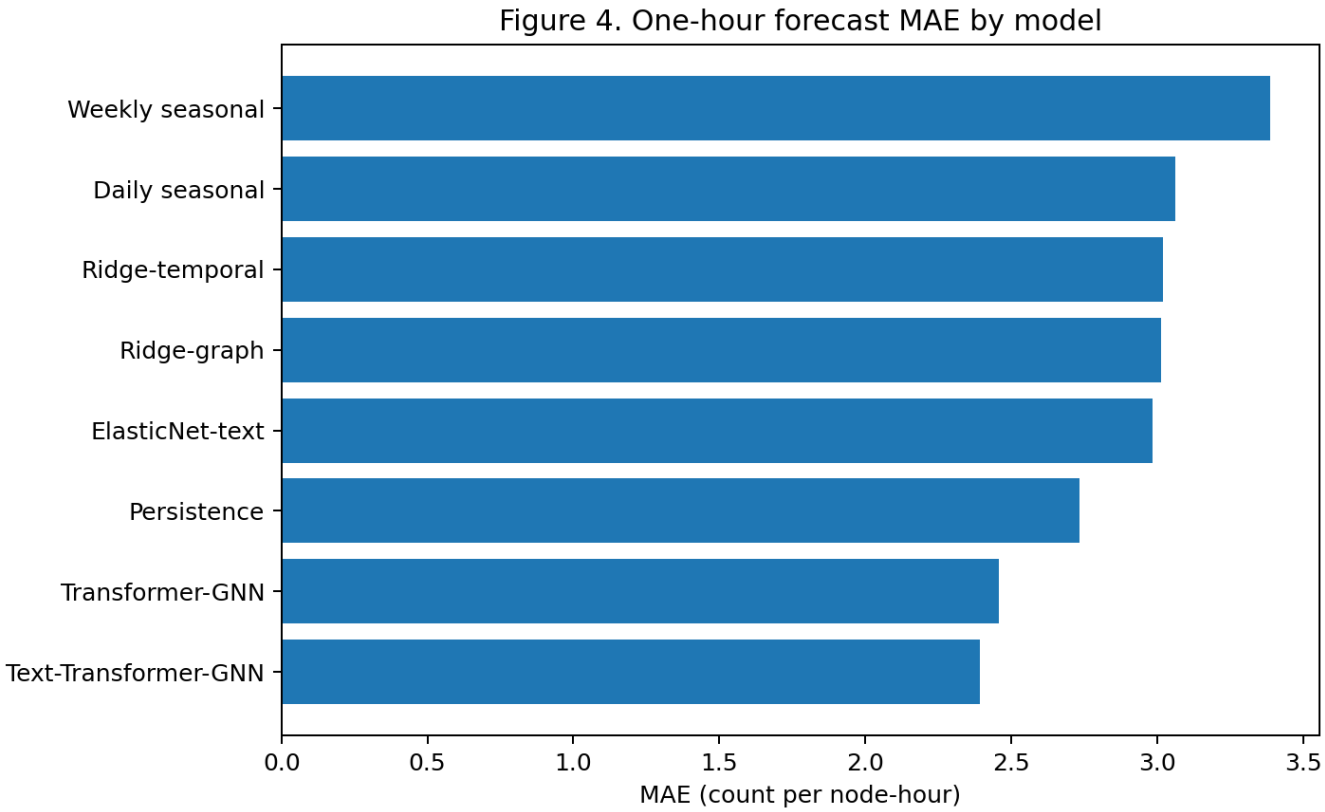

Fig. 4. One-hour MAE comparison across forecasting models.

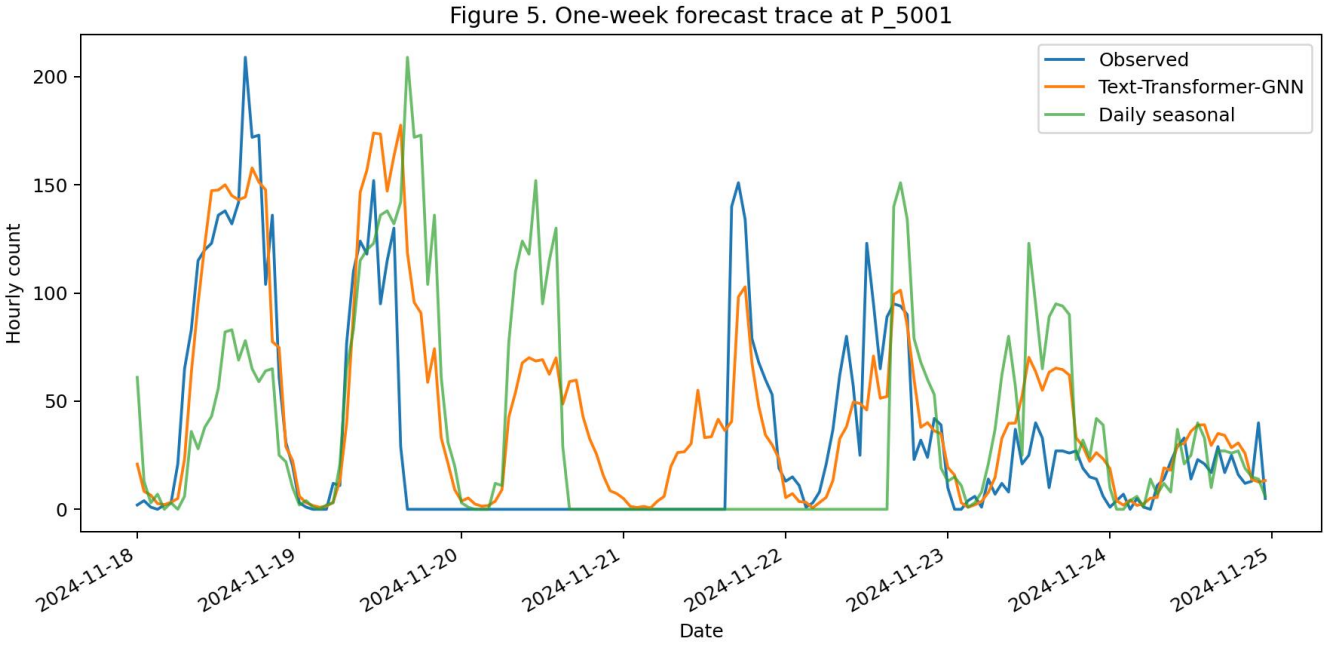

Fig. 5. One-week observed and predicted trace at the busiest retained node.

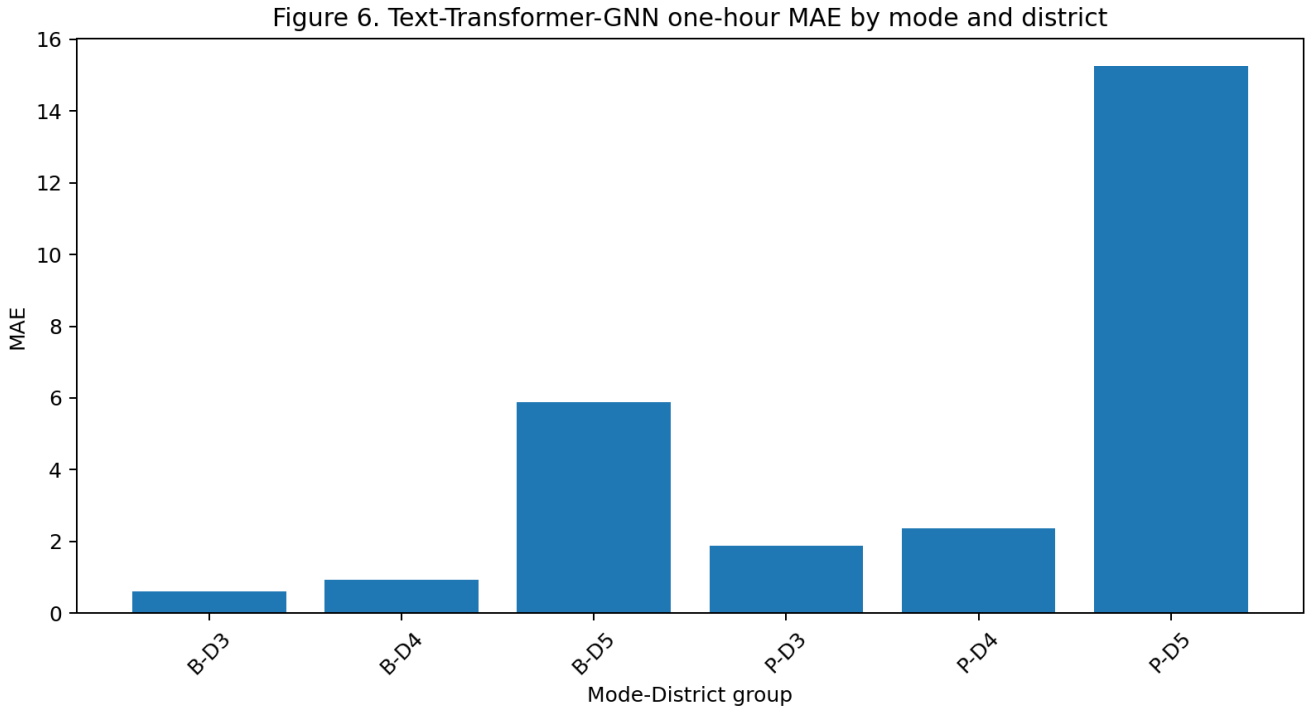

Fig. 6. One-hour Text-Transformer-GNN MAE by mode and district.

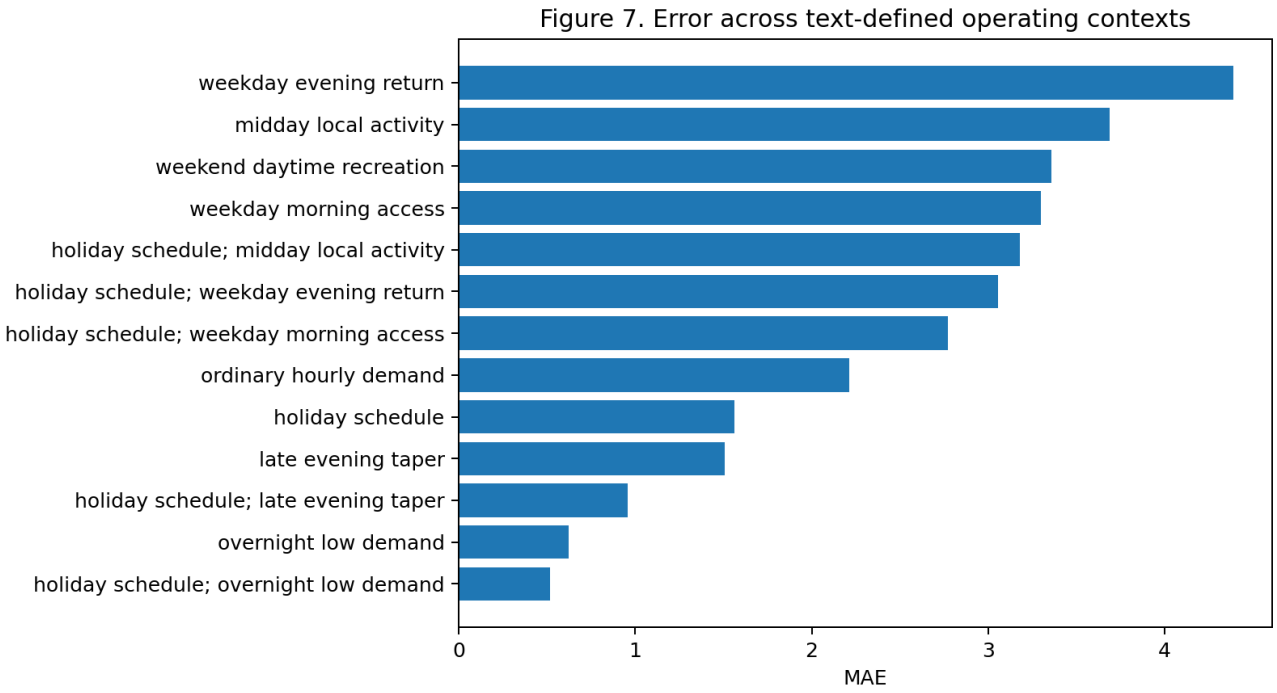

Fig. 7. One-hour Text-Transformer-GNN MAE across event-text contexts.

Operational explanations are most useful when they identify both success and residual uncertainty. Evidence-card and rationale interfaces likewise separate a headline judgment from inspectable support [55], [56], [57], while incident cards prioritize actionable evidence [59]. Table XIII includes representative test-period cases. Some high-volume pedestrian peaks are underestimated even after graph and text calibration, which is valuable information for operators: the model recognizes the context but lacks external signals that would distinguish an unusual local surge from a typical midday or commute period. Other cases show the model correcting a prior-day baseline that would have carried a high count into a weekend hour. The explanation text is concise, but it identifies the relevant context, whether the forecast moved above or below the daily baseline, and whether the forecast closely matched, underestimated, or overestimated the observed count.

Table XIII. Representative operational explanations for one-hour test forecasts

Date hour	Node	Mode	Observed	Forecast	Daily baseline	Context	Explanation
2024-11-29 12:00	P_4007	Pedestrian	360	3.470	7	midday local activity	midday local activity; graph-text forecast below daily; underpredicted.
2024-12-02 09:00	P_5001	Pedestrian	339	137.490	85	weekday morning access	weekday morning access; graph-text forecast above daily; underpredicted.
2024-12-02 17:00	P_5001	Pedestrian	320	161.090	136	weekday evening return	weekday evening return; graph-text forecast above daily; underpredicted.
2024-12-05 16:00	P_5001	Pedestrian	317	185.740	296	weekday evening return	weekday evening return; graph-text forecast below daily; underpredicted.
2024-12-03 16:00	P_5001	Pedestrian	315	184.840	266	weekday evening return	weekday evening return; graph-text forecast below daily; underpredicted.

Overall, the evidence supports three findings. First, active transportation forecasting benefits from combining persistence and seasonality rather than choosing one universal baseline. Persistence is powerful at one hour, while daily seasonality becomes stronger at longer horizons. Second, graph attention improves over linear lag models because it lets the forecast use a weighted set of recent analogues and neighboring node signals. Third, structured event text adds a small but meaningful improvement and creates a direct bridge from numerical prediction to operations-friendly language. This combination is especially suitable for agencies that need both near-term estimates and succinct explanations of why a forecast changed from the prior day or prior week. The connection from forecast to action is therefore explicit but bounded: the model estimates demand, while downstream resource or policy choices remain separate, as in constrained budgeting [29] and conservative off-policy evaluation [68].

The results also show that a good active-mobility model should be evaluated in more than one way. A single overall MAE could hide the fact that low-volume overnight hours are easy, while weekday evening and midday activity create a much larger share of operational uncertainty. Conversely, a high sMAPE can overstate poor performance during hours where both forecast and observation are near zero. The model comparison, ablation, mode-district table, and event-context table together provide a more reliable account of performance. They show where the proposed model improves the average forecast and where additional operational data would be needed to explain rare peaks. Future interface evaluation should similarly separate visual hierarchy [63], editable briefing structure [65], human critique alignment [66], and predictive accuracy.

Limitations

The study intentionally uses only variables present in the 2024 count files plus calendar-derived operating context. This keeps the experiment internally consistent, but it also limits what the model can know. Weather, school

schedules, transit disruptions, construction, major civic events, tourism pulses, and sensor maintenance logs are not represented. The largest residuals occur at high-volume pedestrian sites where such external factors are most likely to matter. Adding those feeds could improve peak-hour accuracy, but it would require a different data integration design and quality-control process.

The forecasting panel keeps nodes with at least 99% hourly coverage. This makes the evaluation clean and comparable, but it excludes partial-year sites that may be important for local planning. The full raw data summary remains part of the analysis, yet the predictive benchmark focuses on continuous operations. A future model could use masking, transfer learning, or hierarchical pooling to include low-coverage nodes without treating missing data as zero demand. Few-shot time-series modeling [26] and cross-dataset wearable transfer [35] provide two candidate design patterns, but they would require a dedicated cross-site validation protocol. Another limitation is that the graph is based on geographic proximity rather than actual walking and cycling network connectivity. Nearby sensors can share demand patterns, but true active-transportation flows depend on crossings, trails, land use, barriers, and facility type.

The explanation layer is grounded in structured tags and model components. That makes the text auditable, but it also means the explanations are intentionally conservative. They can say that a forecast was shaped by a weekend recreation context or graph-neighbor signal; they cannot claim that a specific festival, weather change, or incident caused a peak unless such data are added. The sMAPE metric also requires careful interpretation because sparse count data contain many hours where observed and predicted values are both near zero. For this reason, the discussion emphasizes MAE, RMSE, MASE, and R2 while still reporting sMAPE for completeness. If later versions ingest free-form reports or operator prompts, trust-calibrated multilingual retrieval [47], prompt-injection risk communication [67], and command-safety evaluation [69] would become relevant safeguards.

The model is also designed for short-term aggregate node forecasting rather than route choice, turning-movement assignment, or safety exposure estimation at every crossing leg. Directional labels are used in the raw aggregation step, but the predictive target is the total hourly mode-location count. A more detailed application could forecast each movement direction separately, incorporate facility geometry, or link counts to crash-risk models. Those extensions would require additional validation because movement-level counts are even sparser than the aggregated node series. The present workflow is best interpreted as a practical monitoring layer that can be expanded as richer active transportation data become available.

Conclusion

This study evaluated a text-augmented Transformer-GNN workflow for short-term bicycle and pedestrian forecasting over 2024 California hourly mobility counts. The method aggregates directional movements to mode-location nodes, builds a geographic graph, encodes calendar-derived event text, applies temporal attention over the previous week, and produces grounded operational explanations. On the October-December test period, Text-Transformer-GNN achieved the best MAE at 1, 3, and 6 hour horizons, improving over the strongest seasonal baseline by 12.5%, 14.7%, and 13.3%. Ablation results show that attention and graph features produce the main improvement, while event text further refines the forecast and organizes residual analysis. The method is therefore useful for active-mobility monitoring settings where agencies need not only accurate short-term counts but also concise explanations of the conditions that shaped the forecast.

References

[1] M. S. Ahmed and A. R. Cook, "Analysis of Freeway Traffic Time-Series Data by Using Box-Jenkins Techniques," Transportation Research Record, no. 722, pp. 1-9, 1979.

[2] B. L. Smith, B. M. Williams, and R. K. Oswald, "Comparison of Parametric and Nonparametric Models for Traffic Flow Forecasting," Transportation Research Part C: Emerging Technologies, vol. 10, no. 4, pp. 303-321, 2002.

[3] B. M. Williams and L. A. Hoel, "Modeling and Forecasting Vehicular Traffic Flow as a Seasonal ARIMA Process: Theoretical Basis and Empirical Results," Journal of Transportation Engineering, vol. 129, no. 6, pp. 664-672, 2003.

[4] Y. Lv, Y. Duan, W. Kang, Z. Li, and F.-Y. Wang, "Traffic Flow Prediction With Big Data: A Deep Learning Approach," IEEE Transactions on Intelligent Transportation Systems, vol. 16, no. 2, pp. 865-873, 2015.

[5] E. I. Vlahogianni, M. G. Karlaftis, and J. C. Golias, "Short-Term Traffic Forecasting: Where We Are and Where We're Going," Transportation Research Part C: Emerging Technologies, vol. 43, pp. 3-19, 2014.

[6] P. Ryus et al., Guidebook on Pedestrian and Bicycle Volume Data Collection. Washington, DC, USA: Transportation Research Board, NCHRP Report 797, 2014.

[7] Federal Highway Administration, Traffic Monitoring Guide. Washington, DC, USA: U.S. Department of Transportation, 2016.

[8] S. Hankey, G. Lindsey, and X. Wang, "Estimating Use of Non-Motorized Infrastructure: Models of Bicycle and Pedestrian Traffic in Minneapolis, MN," Landscape and Urban Planning, vol. 107, no. 3, pp. 307-316, 2012.

[9] R. Nosal and L. F. Miranda-Moreno, "Cycling and Weather: A Multi-City and Multi-Facility Study in North America," Transportation Research Record, vol. 2468, no. 1, pp. 74-82, 2014.

[10] R. J. Schneider, L. S. Arnold, and D. R. Ragland, "A Methodology for Counting Pedestrians at Intersections: Use of Automated Counters to Extrapolate Weekly Volumes From Short Manual Counts," Transportation Research Record, vol. 2140, no. 1, pp. 1-12, 2009.

[11] Q. Xin, "Probabilistic Bike-Sharing Demand Forecasting under Changing Weather and Seasonal Regimes with Transformer-Based Models, Transport Findings, Mar. 2026, doi: 10.32866/001c.157499.

[12] T. N. Kipf and M. Welling, "Semi-Supervised Classification with Graph Convolutional Networks," in Proc. International Conference on Learning Representations, 2017.

[13] Y. Li, R. Yu, C. Shahabi, and Y. Liu, "Diffusion Convolutional Recurrent Neural Network: Data-Driven Traffic Forecasting," in Proc. International Conference on Learning Representations, 2018.

[14] B. Yu, H. Yin, and Z. Zhu, "Spatio-Temporal Graph Convolutional Networks: A Deep Learning Framework for Traffic Forecasting," in Proc. International Joint Conference on Artificial Intelligence, 2018, pp. 3634-3640.

[15] Z. Wu, S. Pan, G. Long, J. Jiang, and C. Zhang, "Graph WaveNet for Deep Spatial-Temporal Graph Modeling," in Proc. International Joint Conference on Artificial Intelligence, 2019, pp. 1907-1913.

[16] S. Guo, Y. Lin, N. Feng, C. Song, and H. Wan, "Attention Based Spatial-Temporal Graph Convolutional Networks for Traffic Flow Forecasting," in Proc. AAAI Conference on Artificial Intelligence, vol. 33, no. 1, 2019, pp. 922-929.

[17] C. Zheng, X. Fan, C. Wang, and J. Qi, "GMAN: A Graph Multi-Attention Network for Traffic Prediction," in Proc. AAAI Conference on Artificial Intelligence, vol. 34, no. 1, 2020, pp. 1234-1241.

[18] A. Vaswani et al., "Attention Is All You Need," in Proc. Advances in Neural Information Processing Systems, 2017, pp. 5998-6008.

[19] B. Lim, S. O. Arik, N. Loeff, and T. Pfister, "Temporal Fusion Transformers for Interpretable Multi-horizon Time Series Forecasting," International Journal of Forecasting, vol. 37, no. 4, pp. 1748-1764, 2021.

[20] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," Neural Computation, vol. 9, no. 8, pp. 1735-1780, 1997.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in Proc. North American Chapter of the Association for Computational Linguistics, 2019, pp. 4171-4186.

[22] T. B. Brown et al., "Language Models are Few-Shot Learners," in Proc. Advances in Neural Information Processing Systems, 2020, pp. 1877-1901.

[23] S. M. Lundberg and S.-I. Lee, "A Unified Approach to Interpreting Model Predictions," in Proc. Advances in Neural Information Processing Systems, 2017, pp. 4765-4774.

[24] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why Should I Trust You? Explaining the Predictions of Any Classifier," in Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135-1144.

[25] L. Zhang, R. Ma, and P. Greg, "Digital-Twin Dispatching for Urban Mobility via Spatio-Temporal Transformers and Offline Reinforcement Learning," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 337-363, Aug. 2025, doi: 10.51903/jtie.v4i2.501.

[26] S. He, C. Li, and H. Rao, "Few-Shot Cold-Start Workload Forecasting for New AI Inference Tenants with Time-Series Foundation Models," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 306-324, Apr. 2025, doi: 10.51903/jtie.v4i1.546.

[27] H. Zhou and K. Zhang, "News-Based Uncertainty and Macro-Market Fusion for VIX Direction Forecasting: Evidence from 2015-2024 FRED Panel, J. Technol. Informatics Eng., vol. 4, no. 2, pp. 487-501, Aug. 2025, doi: 10.51903/jtie.v4i2.540.

[28] S. Zhao, Y. Ren, and X. Chang, "Profit-Aware Spot GPU Admission Control with Cost-Sensitive Loss and Evidence-Grounded Policy Memos for AI Workload Supply-Demand Matching," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 45-59, Jun. 2026, doi: 10.51903/jtie.v5i2.545.

[29] Y. Lu, H. Zhou, and Y. Zhang, "A Constrained, Data-Driven Budgeting Framework Integrating Macro Demand Forecasting and Marketing Response Modeling," J. Technol. Informatics Eng., vol. 4, no. 3, Dec. 2025, doi: 10.51903/jtie.v4i3.466.

[30] C. Wang, Z. Wen, R. Zhang, P. Xu, and Y. Jiang, "GPU Memory Requirement Prediction for Deep Learning Task Based on Bidirectional Gated Recurrent Unit

Optimization Transformer," in Proc. 5th Int. Conf. Artificial Intelligence, Virtual Reality and Visualization (AIVRV), Chengdu, China, 2025, doi: 10.1109/AIVRV67401.2025.11350369.

[31] S. He, J. Nie, and C. Li, "Power-Aware Inventory Planning for AI Infrastructure Using Job-Level Forecasting and LLM Workload Explanations," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 341-359, Apr. 2026, doi: 10.51903/jtie.v5i1.548.

[32] J. Zhang, "Early Warning, Grade Prediction, and Teacher-Facing LLM-Ready Explanations Toward an Open Volleyball Course: Reproducible Evidence from Four Public Education Datasets," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 20-44, Jun. 2026, doi: 10.51903/jtie.v5i2.525.

[33] J. Bai, H. Wang, Q. Wu, and B. Zhang, "Privacy-Robust Incrementality Estimation in Cookieless Settings via Uplift Modeling: Reproducible Evidence from the Hillstrom E-Mail Experiment, J. Technol. Informatics Eng., vol. 5, no. 1, pp. 17-38, Apr. 2026, doi: 10.51903/jtie.v5i1.468.

[34] G. Mi, T. Ye, and D. Wood, "A Lightweight Medical Foundation Model for Cross-Modal Multi-Task Pretraining and Parameter-Efficient Few-Shot Transfer on MedMNIST," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 572-589, Aug. 2025, doi: 10.51903/jtie.v4i3.492.

[35] J. Zhang, "From General Human Activity Recognition to Volleyball-Oriented Wearable Transfer Learning: Cross-Dataset Evidence from UCI HAR and WISDM for Domain Adaptation and Edge Deployment," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 263-283, Apr. 2025, doi: 10.51903/jtie.v4i1.524.

[36] Q. Wu, G. Mi, and D. Wood, "Calibration-Light Subject-Independent Motor Imagery BCI via Self-Supervised Pretraining and Conformer, J. Technol. Informatics Eng., vol. 4, no. 1, pp. 239-262, Apr. 2025, doi: 10.51903/jtie.v5i1.493.

[37] B. Zhou, H. Wang, and X. Chang, "Distilling VMAF into an Edge-Deployable Quality Predictor: A Pilot Shot-Level Proxy with LLM-Ready Quality Tokens," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 447-463, Aug. 2025, doi: 10.51903/jtie.v4i2.522.

[38] S. Lu and T. Zou, "Uncertainty-Aware Medical Vision-Language Classification on a Lightweight MedMNIST-Compatible Biomedical Patch Benchmark," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 1-19, Jun. 2026, doi: 10.51903/jtie.v5i2.530.

[39] Z. Ling, Q. Xin, Y. Lin, G. Su, and Z. Shui, "Optimization of Autonomous Driving Image Detection Based on RFAConv and Triplet Attention," in Proc. 2nd Int. Conf. Software Engineering and Machine Learning (SEML), 2024.

[40] Z. Zhong, M. Zheng, H. Mai, J. Zhao, and X. Liu, "Cancer Image Classification Based on DenseNet Model," J. Phys.: Conf. Ser., vol. 1651, no. 1, Art. no. 012143, Nov. 2020, doi: 10.1088/1742-6596/1651/1/012143.

[41] B. Wang, Y. He, Z. Shui, Q. Xin, and H. Lei, "Predictive Optimization of DDoS Attack Mitigation in Distributed Systems Using Machine Learning," in Proc. 6th Int. Conf. Computing and Data Science (CDS), 2024, pp. 89-94.

[42] Y. Li and S. Lu, "Language-Guided Feature Selection for DDoS and Intrusion Detection on CICIDS2017," J. Technol. Informatics Eng., vol. 4, no. 1, pp. 284-305, Apr. 2025, doi: 10.51903/jtie.v4i1.531.

[43] J. Bai, S. Chen, D. Zheng, and M.-J. Kuo, "Interpretable Attack-Chain Stage Detection from AWS CloudTrail Event Sequences via Linear Models and HMM Smoothing," Inf. Electr. Electron. Eng., vol. 6, no. 1, pp. 28-43, May 2026, doi: 10.33474/infotron.v6i1.24923.

[44] Z. S. Zhong, J. Chen, E. Zhong, and X. Sun, "Evidence-Calibrated RAG for Unanswerable Question Answering: Retrieval Coverage, Abstention Calibration, and Hallucination-Proxy Analysis on SQuAD 2.0," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 502-520, Aug. 2025, doi: 10.51903/jtie.v4i2.536.

[45] W. Su, S. Chen, and C. Zhao, "Budgeted Multi-Hop Retrieval Agent for Compositional Question Answering: A Retrieval-Policy Evaluation on the Official MultiHop-RAG Benchmark," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649-662, Dec. 2025, doi: 10.51903/jtie.v4i3.543.

[46] J. Jin, "Evidence-Chain Reliable RAG: Hallucination Detection, Source Attribution, and Deterministic Provenance Explanations," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 520-533, Aug. 2025, doi: 10.51903/jtie.v4i2.535.

[47] Y. Chen and H. Xu, "Trust-Calibrated Multilingual RAG for Humanitarian Information Platforms: Empirical Evaluation on OMoS-QA for Migration Information Access," Int. J. Graph. Des., vol. 4, no. 1, pp. 141-164, Apr. 2026, doi: 10.51903/ijgd.v4i1.3552.

[48] S. Meng, J. Chen, and I. Zheng, "LLM-Inspired Offline Reranking for Financial Search: Query Rewriting, Hybrid Retrieval, and Listwise Relevance Ranking on FiQA," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 361-378, Apr. 2026, doi: 10.51903/jtie.v5i1.537.

[49] Z. Li, K. Zhang, and A. Wong, "Numerical-Reasoning Guardrails for a Quant Research Assistant: A Compact

Reproducible Benchmark Using SEC and FRED Data," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 75-90, Jun. 2026, doi: 10.51903/jtie.v5i2.541.

[50] M.-J. Kuo, D. Zheng, and J. Hires, "Federated Topic-Preference Learning for Knowledge-Grounded Chat with Differential Privacy," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 385-401, Aug. 2025, doi: 10.51903/jtie.v4i2.502.

[51] R. Zhang, Z. Wen, C. Wang, C. Tang, P. Xu, and Y. Jiang, "Quality Analysis and Evaluation Prediction of RAG Retrieval Based on Machine Learning Algorithms," arXiv:2511.19481, 2025. [Online]. Available: https://doi.org/10.48550/arXiv.2511.19481

[52] Y. Chen, S. Zhou, and E. Lin, "Accounting-Aware Evidence Retrieval for Institutional Due Diligence of Tokenized Trade Receivable RWA, J. Technol. Informatics Eng., vol. 4, no. 3, pp. 649-663, Dec. 2025, doi: 10.51903/jtie.v4i3.542.

[53] J. Li and A. Zhou, "Multi-Regulation RAG for AI Product Counsel: A Legal Governance Framework for Cross-Border Digital Commerces, Rule of Law Stud. J., vol. 2, no. 2, pp. 105-123, Jun. 2026, doi: 10.64780/rolsj.v2i2.225.

[54] Z. S. Zhong and S. Ling, "Improved Theoretical Guarantee for Rank Aggregation via Spectral Method," Inf. Inference: J. IMA, vol. 13, no. 3, Art. no. iaae020, Sep. 2024, doi: 10.1093/imaiai/iaae020.

[55] J. Jin, "LLM-Style Evidence Cards for Scientific Search Interfaces: A UI/UX Design Framework for Retrieval Transparency, Ranking Trust, and Visual Evidence Hierarchy," Int. J. Graph. Des., vol. 3, no. 2, pp. 397-414, Oct. 2025, doi: 10.51903/ijgd.v3i2.3698.

[56] B. Zhang, Y. Ren, and J. Zou, "LLM-Style Explainable E-Commerce Recommendation Cards: A UI/UX Design Framework for Trust-Calibrated Product Recommendation," Int. J. Graph. Des., vol. 3, no. 2, pp. 381-396, Oct. 2025, doi: 10.51903/ijgd.v3i2.3697.

[57] Q. Wu, S. Meng, and J. Zhao, "Text-Grounded LLM-Assisted Design Rationale Interfaces: Turning Advertising Layout Metadata into Explainable UI/UX Decision Cards," Int. J. Graph. Des., vol. 3, no. 1, pp. 216-240, May 2025, doi: 10.51903/ijgd.v3i1.3713.

[58] Y. Zhang and H. Zhang, "A Therapist-Facing Session Copilot for Live Counseling Support: Reasoning-Guided Retrieval and Ranking from Multi-Turn Counseling Dialogues," J. Technol. Informatics Eng., vol. 4, no. 2, pp. 464-486, Aug. 2025, doi: 10.51903/jtie.v4i2.547.

[59] J. Nie, G. Liu, C. Li, and T. Zou, "Evidence-Constrained Incident Visualization Cards for Distributed Cloud Logs: A UI/UX Framework for Turning Hadoop, OpenStack, and ZooKeeper Logs into Actionable SRE Design Interfaces," Int. J. Graph. Des., vol. 4, no. 1, pp. 179-185, Apr. 2026, doi: 10.51903/ijgd.v4i1.3703.

[60] H. Tu, S. Zhao, and A. Zhou, "Visual Brief Cards for Advertising Design: A Structured UI/UX Framework for Turning Creative Intentions into Graphic Design Decisions," Int. J. Graph. Des., vol. 3, no. 1, pp. 210-226, May 2025, doi: 10.51903/ijgd.v3i1.3714.

[61] Y. Zhang and H. Zhang, "Visualizing the Right Counseling Support: Evidence-Linked Recommendation Cards for Explainable Mental Health Intake Interfaces," Int. J. Graph. Des., vol. 3, no. 1, pp. 214-229, May 2025, doi: 10.51903/ijgd.v3i1.3722.

[62] T. Ye, X. Chang, and E. Zhong, "Uncertainty-Aware Breast Ultrasound Explanation Cards: A Visual Communication Framework for Image-Based AI Diagnostic Support Using BreastMNIST_224," Int. J. Graph. Des., vol. 3, no. 2, pp. 365-380, Oct. 2025, doi: 10.51903/ijgd.v3i2.3701.

[63] Z. Li, S. Zhou, and Z. Zhou, "Financial Risk Dashboard Design for Institutional RWA Investors: Visual Hierarchy, Chart Comprehension, and Explainability in FinChart-Bench," Int. J. Graph. Des., vol. 3, no. 1, pp. 196-210, May 2025, doi: 10.51903/ijgd.v3i1.3715.

[64] H. Xu, Y. Chen, and A. Med, "Automatic Detection and Explanation of Dark Patterns from Interface Microcopy: Empirical Comparison of BERT-Style Encoders, RoBERTa-Style Encoders, and LLM-Style Decoders on the ec-darkpattern Dataset," J. Technol. Informatics Eng., vol. 4, no. 3, pp. 590-612, Dec. 2025, doi: 10.51903/jtie.v4i3.491.

[65] J. Mu, Y. Lu, and E. Hwang, "Structured Visual Brief Interfaces for Advertising Design: A UI/UX Framework for Turning Creative Intentions into Designer-Editable Graphic Design Cards," Int. J. Graph. Des., vol. 4, no. 1, pp. 192-208, Apr. 2026, doi: 10.51903/ijgd.v4i1.3702.

[66] Y. Li, S. Lu, and L. Zhao, "LLM-as-Design-Critic: Aligning AI-Generated UI Feedback with Human Graphic Design Judgment, Int. J. Graph. Des., vol. 3, no. 1, pp. 196-215, May 2025, doi: 10.51903/ijgd.v3i1.3661.

[67] W. Su, H. Rao, and E. Ma, "Privacy and Data-Integrity Risk Cards for LLM Agents: A UI/UX Design Framework for Secure Human Oversight under Prompt-Injection Attacks," Int. J. Graph. Des., vol. 4, no. 1, pp. 186-191, Apr. 2026, doi: 10.51903/ijgd.v4i1.3699.

[68] T. Ye, J. Mu, and J. Hunter, "Off-Policy Evaluation and Conservative Policy Selection for Slot-Level Dynamic Bidding and Ranking on the Open Bandit Dataset (Small)," J. Technol. Informatics Eng., vol. 5, no. 1, pp. 178-199,

Apr. 2026, doi: 10.51903/jtie.v5i1.503.

[69] B. Zhang, X. Sun, G. Liu, and B. Zhou, "LLM-Style DevOps Copilot for Cloud-Native Troubleshooting: Retrieval-Augmented Runbook Generation and Command-Safety Evaluation," J. Technol. Informatics Eng., vol. 5, no. 2, pp. 104-118, 2026, doi: 10.51903/jtie.v5i2.534.

[70] Z. S. Zhong and S. Ling, "Uncertainty Quantification of Spectral Estimator and MLE for Orthogonal Group Synchronization," arXiv:2408.05944, Aug. 2024.

[71] R. J. Hyndman and G. Athanasopoulos, Forecasting: Principles and Practice, 3rd ed. Melbourne, Australia: OTexts, 2021.