Graph-Temporal Deep Learning for Urban Image Modeling From Multimodal Discourse Interaction: A Case Study of Hefei

Lixia Xu
School of English Language,Anhui International Studies University, Hefei 231200, China
E-mail:xulixia14@outlook.com

Keywords: multimodal discourse interaction, urban informatics, knowledge graph representation, graph attention network, temporal fusion transformer

Received: September 10, 2025

This paper presents a computational framework for urban image modeling driven by multimodal discourse interaction, integrating cross-modal representation learning, graph-based semantic modeling, and temporal sequence prediction. Images and texts are embedded into a unified semantic space using CLIP and BLIP-2 models, enabling high-fidelity multimodal representation. A knowledge graph of urban discourse is constructed and modeled with a Graph Attention Network (GAT) to capture semantic relationships among multiple agents, while a Temporal Fusion Transformer (TFT) is employed to learn both long-term dependencies and local feature dynamics. To enhance interpretability, a variable selection network identifies the dominant multimodal features shaping urban image evolution. Experimental results based on Hefei's urban discourse demonstrate high semantic alignment between text–image pairs (e.g., 0.87 for "Hefei Metro Expansion" and "Metro Station," 0.89 for "USTC Research Breakthrough" and "USTC Campus"), strong knowledge graph relations (e.g., 0.82 for USTC–High-tech Zone linkage), and accurate temporal forecasting with RMSE reduced to 0.061. The dataset contains 18,426 text entries and 9,307 paired images, and the evaluation adopts a fixed 7:2:1 split with CLIP and BLIP-2 as embedding baselines. Comparative tests against LSTM and GRU yield RMSE values of 0.083 and 0.079 respectively. The findings confirm that the proposed graph-temporal multimodal framework provides an interpretable, data-driven methodology for quantifying and analyzing urban image formation.

Povzetek: Članek predstavi razložljiv multimodalni (slika+besedilo) grafno-časovni okvir, ki z modeli CLIP/BLIP-2, GAT in TFT učinkovito modelira ter napoveduje razvoj urbane podobe na podlagi urbanega diskurza.

1 Introduction

City image is a core concern in urban communication and governance because it shapes external appeal and internal cohesion while being continually co-constructed by diverse actors across media. Recent multimodal inquiries—from computer-vision–assisted landscape communication to city-branding discourse studies—underscore that linguistic, visual, and interactive signals jointly configure urban meaning in networked environments [1–3]. Beyond external perception, city image functions as a vehicle for identity construction and cultural expression; comparative analyses of visual streetscapes, metadiscourse in promotional texts, and translation of public notices reveal how identity resources circulate across platforms and publics [4–6]. Material and digital artifacts—memorial plaques, national tourism portals, and municipal destination sites—operate as multimodal narratives that sediment collective memory and frame urban desirability online [7–9].

Concurrently, data-driven urban analytics has opened avenues for sensing and quantifying "functional images" of cities via place-based evidence and social media signals, enabling empirical tests of how discourse aligns with spatial functions [10–11]. Fusion of high-resolution remote sensing, building footprints, and social perception data—often with transformer-based multimodal networks—improves semantic granularity while scaling to metropolitan extents [12–14]. Yet, the velocity and nonlinearity of social media discourse complicate dynamic modeling; mixed "big/small data" strategies and analyses of affective/visual preferences highlight the temporal sensitivity and context dependence of urban impressions [15–16].

Progress in multimodal datasets and simulation further lowers the barrier to robust evaluation. Synthetic 2D/3D pipelines and synchronized urban-scene benchmarks supply controlled variation across viewpoints and conditions for training and testing semantic models [17–18]. Parallel streams of user-generated and platform-curated content—travel vlogs, short videos, and geotagged photos—have been leveraged with transfer learning to infer urban image elements at scale, but integrating them with structured urban knowledge remains challenging [19–21]. Graph representation learning contributes a complementary lens: GNNs quantify spatial homogeneity in networks, infer built-environment attributes from street-view imagery, and support semantic labeling of 3D assets, collectively suggesting that

relational structure is pivotal for urban image understanding [22–24].

Temporal models have likewise matured. Transformer-based predictors capture long-range dependencies in environmental and experiential urban phenomena, indicating their suitability for dynamic discourse streams whose cues unfold across irregular horizons [25–26]. On the representation side, advances in cross-modal pretraining and personalization—optimization of BLIP-2 with lightweight adapters, visual-prompt conditioning, and multi-graph neighborhood encoders—point to practical pathways for aligning heterogeneous modalities while preserving task-specific nuance [27–29]. Geographic knowledge-graph fusion and temporal multimodal graph transformers demonstrate that unifying entity-relation structure with time-aware alignment markedly improves retrieval and reasoning over video–text pairs, foreshadowing analogous gains for urban discourse corpora [30–31].

Broader multimodal and temporal research offers design heuristics that translate to city-image modeling. Asynchronous acquisition frameworks clarify how to reconcile heterogeneous sampling across channels—an issue mirrored in aligning news bursts, posts, and images over time [32]. Causal multimodal reasoning emphasizes disentangling sources of variation and counterfactual robustness, desirable for diagnosing shifts in urban narratives under policy or event shocks [33]. Surveys of event knowledge graphs and systematic reviews of GNN methods consolidate best practices for representing temporally scoped, relation-rich processes, which we adopt for city-image evolution [34–35].

Scalable and personalized computation also matters when modeling interactions among many urban stakeholders. Distributed recommendation architectures illustrate how preference signals can be shared and tailored under platform constraints; multi-scale deep time-series models and domain-agnostic embedding learning (e.g., for symbolic sequences) reinforce the value of hierarchical temporal features and compact representations transferable across modalities [36–37]. Finally, long-standing traditions in pattern recognition and coordinated multi-agent learning highlight robustness, discriminative embeddings, and graph-structured decision-making under uncertainty—properties that are equally relevant when treating urban discourse as a multi-actor system over a city's semantic graph [38–39].

This study focuses on three explicit objectives: improving discourse-image consistency through multimodal alignment, strengthening entity relation inference through graph attention, and enhancing prediction robustness through temporal fusion. Research gap and contributions. Despite these advances, existing studies typically compartmentalize three ingredients that are jointly necessary for interpretable, dynamic city-image modeling: (i) unified cross-modal alignment that projects text and images into a common semantic space; (ii) graph-based reasoning that captures multi-actor relations among places, institutions, infrastructures, and publics; and (iii) temporal sequence modeling that learns both longrange dependencies and local bursts in discourse activity. This paper addresses that gap by proposing a graph-temporal multimodal framework aligned with the article's title and abstract. Concretely, we (1) learn image–text embeddings with CLIP and BLIP-2 to obtain high-fidelity cross-modal representations; (2) construct an urban discourse knowledge graph and apply a Graph Attention Network (GAT) to model semantic propagation among entities; and (3) employ a Temporal Fusion Transformer (TFT) with a variable-selection network to capture multi-scale temporal dynamics and enhance interpretability. A case study on Hefei operationalizes the approach and quantifies alignment quality, relational strengths, and predictive accuracy, offering a reproducible pathway for computationally analyzing the generation and evolution of urban image in multimodal discourse. The resulting knowledge graph centralities and sentiment-driven temporal shifts provide evidence for how discourse structures influence perceived urban functions and how multimodal cues shape the trajectory of city-image formation.

2 City image modeling method

2.1 Overall framework of the method

Computational modeling of urban image driven by multimodal discourse interaction aims to achieve a dynamic and quantitative representation of urban image. Its overall approach is to align images and text in a unified semantic space, characterize the interactive structure of urban discourse through a knowledge graph, and dynamically model the multimodal representation in the temporal dimension, thereby revealing the temporal evolution of urban image. The proposed method first utilizes CLIP and BLIP-2 to generate a highly consistent multimodal representation. Embedding mapping maps textual semantics and visual features into a unified vector space, providing basic node features for subsequent knowledge graph construction [27-28] . Next, a knowledge graph is constructed based on urban discourse corpus and image information, converting multimodal embeddings into node features. GAT is used to model the semantic relationships between nodes, enabling the propagation and reinforcement of the semantic structure of multi-agent interactions [29-30] . TFT is introduced in the temporal dimension to model the multimodal embedding sequence. A multi-head attention mechanism is used to capture long-term dependencies, a gating mechanism processes local dynamic features, and a variable selection network is combined to enhance the model's ability to interpret key features. The entire method forms a continuous process of multimodal alignment, knowledge graph semantic propagation, and temporal modeling, achieving a complete computational framework from static embedding to dynamic evolution. Figure 1 shows the overall framework of this method, clearly demonstrating the connection between multimodal alignment, knowledge graph construction, and temporal modeling.

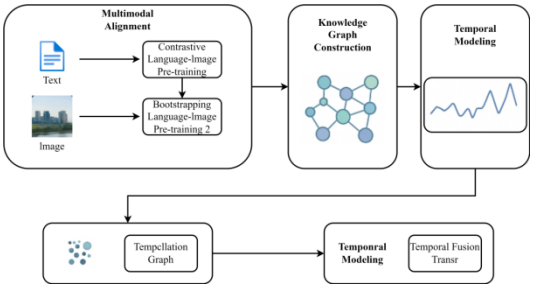

Figure 1: Framework of this method

At the mathematical modeling level, the overall objective function is set to simultaneously optimize multimodal semantic consistency and time series prediction accuracy, which can be expressed as a total loss function:

$$L_{total}=\lambda_1 L_{align}+\lambda_2 L_{graph}+\lambda_3 L_{temporal} \quad (1)$$

As shown in formula (1), L_{align} represents the multimodal semantic alignment loss, which measures the similarity between text and image embeddings in a unified semantic space; L_{graph} represents the knowledge graph semantic relationship modeling loss, which is used to enhance the accuracy of semantic propagation between nodes; $L_{temporal}$represents the time series prediction loss, which depicts the prediction accuracy of the dynamic evolution of the city image; λ_1 , λ_2 , λ_3 are weight coefficients, which are used to balance the contribution of each part of the loss to the overall goal.

In the specific optimization process, the parameters are jointly updated through the gradient descent algorithm, and the overall optimization formula can be expressed as:

$$\theta^* = \arg\min_{\theta} L_{total}(\theta) \quad (2)$$

As shown in formula (2), θrepresents all trainable parameters of the multimodal embedding network, graph attention network, and time fusion Transformer, θ^*which is the optimal parameter combination after optimizing the total loss function, realizing the city image modeling and dynamic quantitative expression driven by multimodal discourse interaction.

2.2 Multimodal semantic alignment

Multimodal semantic alignment aims to map images and text into a unified semantic space, thereby achieving consistency in cross-modal semantic representation. In this method, the embedding mapping of image and text features is achieved through CLIP and BLIP-2. CLIP utilizes a contrastive learning mechanism to generate high-quality vector representations by maximizing the similarity between matching images and text and minimizing the similarity between unmatched pairs. BLIP-2 builds on this by introducing a bootstrapping strategy to optimize the fine-grained semantic capture of multimodal representations, enabling more efficient alignment of text and image embeddings in a unified space. The embeddings from CLIP and BLIP-2 are aligned to the same dimensional space and concatenated to form the final vector used as node features and temporal inputs.

The cross-modal embedding of images *I*and text can be expressed as a vector mapping function:*T*

$$\begin{cases} \mathbf{v}_I = f_\theta(I) \\ \mathbf{v}_T = g_\phi(T) \end{cases} \quad (3)$$

As shown in formula (3), f_θ represents the parameterization function of the image encoder, $\mathbf{v}_I$is the image embedding vector; g_ϕ represents the parameterization function of the text encoder, $\mathbf{v}_T$and is the text embedding vector. This mapping maps data of different modalities into the same semantic space to ensure semantic consistency.

The alignment process is optimized through a contrastive loss function to improve the consistency of cross-modal embeddings, which can be expressed as:

$$L_{align}=-\frac{1}{N}\sum_{i=1}^{N}\log\frac{\exp\left(\frac{\mathrm{sim}(\mathbf{v}_I^i,\mathbf{v}_T^i)}{\tau}\right)}{\sum_{j=1}^{N}\exp\left(\frac{\mathrm{sim}(\mathbf{v}_I^i,\mathbf{v}_T^j)}{\tau}\right)} \quad (4)$$

As shown in formula (4), where *N*is the batch size, $\mathrm{sim}(\cdot,\cdot)$represents the vector similarity function, $\tau$is the temperature parameter used to adjust the smoothness of the similarity distribution; $\mathbf{v}_I^i$and are the image and text embedding vectors of $\mathbf{v}_T^i$the th sample respectively *i*. This loss maximizes the similarity of matching pairs and minimizes the similarity of unmatched pairs to achieve cross-modal consistency.

By combining CLIP and BLIP-2, this method achieves a more consistent multimodal representation in the semantic space, providing a reliable feature foundation for subsequent knowledge graph node construction. Figure 2 shows a scatter plot of the cross-modal embedding distribution, which intuitively demonstrates the clustering and alignment of text and images in a unified semantic space.

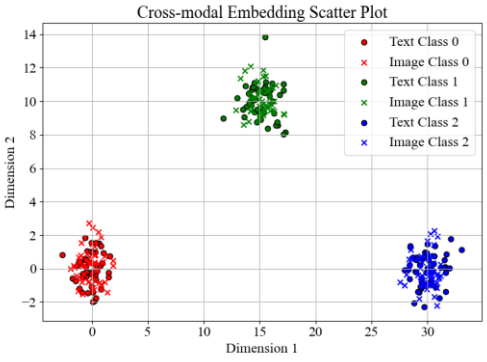

Figure 2: Cross-modal embedding distribution scatter plot

Figure 2 shows the embedding distribution of text and images in the two-dimensional semantic space. Each point represents the vector representation of a sample. Text and image points of different categories form a clear cluster structure in the space. Text and image points of the same category are tightly clustered in local areas, with significant spacing between clusters. Edge samples appear between clusters, reflecting semantic alignment measured by cosine similarity.

2.3 Knowledge graph construction and semantic propagation

After multimodal embedding is complete, explicit modeling of the interactive structure of urban discourse is achieved by converting the vector representations of text and images into nodes and edges in a knowledge graph. Each node corresponds to an entity or concept in urban discourse, and its node features $\mathbf{v}_i$ are represented by a cross-modal embedding vector. The edges between nodes reflect semantic relationships or interactive associations, and their weights are calculated based on the similarity of node embeddings or co-occurrence frequency. The graph structure is constructed from corpus-derived co-occurrence signals, without the use of an external ontology, and all edges originate from observed discourse interactions. The constructed graph can be represented as $G = (V, E)$, where $V=\{v_1,v_2,...,v_n\}$ is the set of nodes and $E=\{e_{ij}\}$ is the set of edges. The edge weight e_{ij} is defined as the node feature similarity function:

$$e_{ij}=\sigma\left(\mathbf{v}_i^{\top} W_e \mathbf{v}_j\right) \tag{5}$$

As shown in formula (5), where W_e is a trainable weight matrix and $\sigma(\cdot)$ is an activation function, which is used to normalize the edge weight range so that it reflects the strength of the relationship between nodes in semantic propagation. The update of node features and semantic propagation are achieved through GAT. The node feature update of each layer can be expressed as:

$$\mathbf{h}_i^{'}=\sigma\left(\sum_{j\in N(i)} \alpha_{ij} W_{\mathrm{h}} \mathbf{h}_j\right) \tag{6}$$

As shown in formula (6), where $\mathbf{h}_i$ is the current feature vector of the node i, $\mathbf{h}_i^{'}$ is the updated feature vector, $N(i)$ represents i the set of neighbor nodes of the node, W_{h} is the trainable weight matrix, α_{ij} and is the attention coefficient, which is used to measure the contribution of neighbor nodes j to the node i. The calculation of the attention coefficient α_{ij} adopts the self-attention mechanism:

$$\alpha_{ij}=\frac{\exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{\top}\left[W_{\mathrm{h}}\mathbf{h}_i \| W_{\mathrm{h}}\mathbf{h}_j\right]\right)\right)}{\sum_{k\in N(i)} \exp\left(\mathrm{LeakyReLU}\left(\mathbf{a}^{\top}\left[W_{\mathrm{h}}\mathbf{h}_i \| W_{\mathrm{h}}\mathbf{h}_k\right]\right)\right)} \tag{7}$$

As shown in formula (7), is a $\mathbf{a}$ trainable attention vector $\|$ representing the vector concatenation operation, and LeakyReLU is a leaky linear rectifier activation function that ensures the differentiated influence of neighboring nodes in feature aggregation. Through multi-layer GAT iterative updates, node features gradually integrate the semantic information of their neighbors, enabling the propagation and reasoning of semantic relationships.

The construction of the knowledge graph not only reveals the structural relationships between entities and concepts in urban discourse but also provides a clear semantic network foundation for dynamic evolution analysis. To demonstrate the impact of edge weights and attention coefficients on the prediction results, this study included a sensitivity experiment with a fixed perturbation range, and the results are shown in Table 1.

Table 1: Impact of edge weight perturbation on prediction results

Perturbation Type	Perturbation Magnitude	RMSE	Difference from Original Result	Graph Consistency
Overall increase of edge weights	+0.10	0.074	+0.013	0.78
Overall decrease of edge weights	−0.10	0.082	+0.021	0.74
Increase of edge weights on high-frequency nodes	+0.10	0.069	+0.008	0.80
Perturbation of edge weights on low-frequency nodes	+0.10	0.063	+0.002	0.79

To demonstrate the impact of erroneous links, this study further constructed random erroneous edges and calculated semantic consistency, the results of which are shown in Table 2.

Table 2: Impact of the number of faulty links on prediction consistency

Number of Random Erroneous Edges	RMSE	MAE	Graph Consistency
5	0.067	0.051	0.76
15	0.081	0.062	0.69
40	0.104	0.083	0.57

Figure 3 shows the visualization of the Hefei urban discourse knowledge graph, which directly displays the semantic relationships and interaction strengths between nodes.

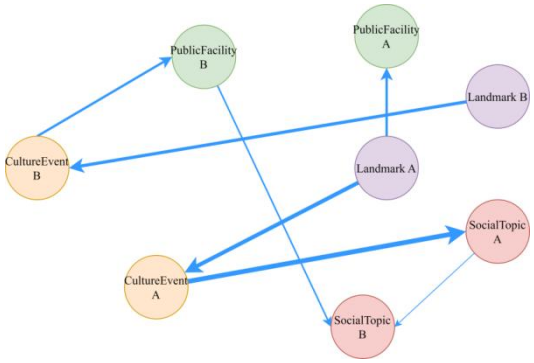

Figure 3: Visualization of Hefei's urban discourse knowledge graph

The diagram shows the structural features of Hefei's urban discourse knowledge graph. Nodes represent different entities or concepts in urban discourse, with their size reflecting the centrality of the node in the network, and color distinguishing different entity types. Nodes are connected by directed edges, the thickness and color of which indicate the strength of semantic relationships and reflect the intensity of interaction between nodes. The overall layout uses a force-directed approach, evenly distributing nodes in space. Nodes with close semantic relationships form clusters, and the distance between nodes corresponds to their semantic similarity. This demonstrates the structural connections and thematic distribution between discourse entities. The direction of the arrows in the diagram reflects the flow of semantic communication, providing a clear semantic network foundation for city image modeling.

2.4 Timing modeling mechanism

After multimodal embedding and knowledge graph semantic propagation are complete, TFT is introduced to model the dynamic evolution of urban discourse. TFT input is a time series feature vector that has undergone multimodal alignment and knowledge graph augmentation $\mathbf{X}_t$, where each time step $t$corresponds to a multimodal representation of the state of urban discourse. The input features include historical embeddings $\mathbf{H}_t$and exogenous temporal features $\mathbf{E}_t$. A gating mechanism is used to filter information to control the transmission of long-term dependencies and local fluctuations.

TFT captures long-term dependencies through a multi-head attention mechanism. For the input sequence $\mathbf{X}=[\mathbf{x}_1,\mathbf{x}_2,...,\mathbf{x}_T]$, the multi-head attention is calculated as:

$$\text{Attention}(\mathbf{Q},\mathbf{K},\mathbf{V})=\text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \tag{8}$$

As shown in formula (8), where $\mathbf{Q}=\mathbf{X}W_Q$, $\mathbf{K}=\mathbf{X}W_K$, $\mathbf{V}=\mathbf{X}W_V$ and represent query, key, and value matrices, respectively, W_Q,W_K,W_V is a trainable weight matrix, $d_k$and is the key vector dimension. This mechanism allows the model to focus on the historical states in the time series that are relevant to the current prediction, thus achieving long-term dependency capture.

The local dynamic features are processed through a gating mechanism, which is expressed as:

$$\mathbf{h}_t=\mathbf{z}_t\odot\tilde{\mathbf{h}}_t+(1-\mathbf{z}_t)\odot\mathbf{h}_{t-1} \tag{9}$$

As shown in formula (9), where $\mathbf{h}_t$is the hidden state of the time step t, $\tilde{\mathbf{h}}_t$is the current candidate state, $\mathbf{z}_t$is the update gate, and $\odot$ represents element-wise multiplication. The gating mechanism $\mathbf{z}_t$ achieves smoothing of local dynamics by adjusting the fusion ratio of historical information and current features.

The variable selection network is used to improve the model's feature interpretation. Each input feature dimension $x_{t,i}$is assigned a weight $\gamma_{t,i}$, and the final input is expressed as a weighted sum:

$$\tilde{\mathbf{X}}_t=\sum_{i=1}^{d}\gamma_{t,i}x_{t,i} \tag{10}$$

$$\gamma_{t,i}=\frac{\exp(w_i^{\top}x_{t,i})}{\sum_{j=1}^{d}\exp(w_j^{\top}x_{t,j})} \tag{11}$$

As shown in formula (11) , where $w_i$is a trainable parameter, $\gamma_{t,i}$representing the importance weight of $i$the feature at the time step t, and the emphasis of key features is achieved through interpretability weighting. The final time series prediction output generated by TFT is a dynamic evolution representation of the city image, which can be expressed as:

$$\hat{\mathbf{y}}_t=f_{\text{TFT}}(\tilde{\mathbf{X}}_{1:t},\mathbf{E}_{1:t}) \tag{12}$$

As shown in formula (1 2), is $\hat{\mathbf{y}}_t$the predicted value of the city image $f_{\text{TFT}}(\cdot)$at the time step , $t$and represents the TFT network mapping function. The target variable is a single numerical index representing the fused semantic consistency and interaction strength of each month, and the error is computed by comparing predicted and observed index values in the held-out sequence. Combining multi-head attention, gating mechanism and variable selection, it achieves accurate modeling and interpretable quantitative expression of the dynamic evolution of the city image.

3 Experimental design and data preparation

3.1 Data source and preprocessing

The experimental data covers the main sources of Hefei's multimodal discourse corpus. The text portion comes from the official website of the Hefei Municipal Government, the Hefei Daily digital newspaper system, the news release platforms of district and county

governments, geotagged public posts on Weibo, authenticated WeChat public account push information, and the Hefei special channels of mainstream news clients. The image portion is collected from Weibo, WeChat public accounts, and news clients, and is screened in combination with multimodal posts featuring Hefei's city landmarks, public facilities, and landscape elements. The time range of the social media corpus covers January 2021 to January 2025 , ensuring the temporal integrity and dynamic representation of the data.

During the preprocessing phase, after collection, text data is deduplicated, irrelevant content is removed, and spam is cleaned. Simplified Chinese encoding is used, and character normalization is completed. Word segmentation utilizes a custom dictionary-based word segmentation tool to enhance the recognition of proprietary terms related to Hefei's geographic entities, policy names, and cultural symbols. A stop word list is also used to remove invalid terms. Named entity recognition is used to extract core semantic units such as city entities, locations, policies, and institutions. These units are then associated with timestamps and geotags to ensure consistency in the subsequent knowledge graph construction.

During preprocessing, image data is uniformly resized to a 224×224 resolution. Center cropping and normalization are performed to eliminate size discrepancies, and normalization is performed using the mean and standard deviation of the RGB channels. Duplicate images and samples with obvious watermarks, blur, or excessive noise are removed to ensure input feature clarity and robustness. Image feature extraction relies on the encoders of CLIP and BLIP-2. Feature vectorization is performed during preprocessing to align them with the text representation in a unified semantic space.

Multimodal data is paired using timestamps, geolocation information, and social media post identifiers to ensure consistency in the interaction between text and images at the discourse level. All processed data is organized into a multimodal corpus, providing a structured input foundation for subsequent cross-modal semantic alignment, knowledge graph construction, and temporal modeling. The corpus contains 18,426 multimodal posts, and 9,307 image-text pairs are used for model training. The training, validation, and testing ratios follow a fixed 7:2:1 split. The prediction target in the temporal modeling stage is a scalar city-image index derived from the normalized similarity between text and image embeddings and the aggregated discourse interaction frequencies.

3.2 Experimental environment and parameter settings

To ensure the repeatability of the training and evaluation of the proposed model, the hardware and software environments used in the experiment are listed in Table 3.

Table 3: Experimental environment configuration

Type	Item	Model/Version	Quantity
Hardware	CPU	Intel Core i7-12700F	1
	GPU	NVIDIA GeForce RTX 3060 (12GB)	1
	Memory	DDR4 8 GB 3200MHz	4
	Storage	SSD 1TB NVMe	1
Software	Operating System	Ubuntu 20.04 LTS	1
	Python	3.9	1
	TensorFlow	2.9	1
	PyTorch	1.12	1
	CUDA	11.6	1
	cuDNN	8.4	1
	MySQL	8.0	1

The model training process involves setting multiple key parameters, including optimizer parameters, learning rate scheduling, and core hyperparameters of the deep network architecture. To ensure a balance between convergence speed and generalization performance, the experiment finalized the settings through multiple verification adjustments within the initial parameter range. The results are shown in Table 4.

Table 4: Model parameters and hyperparameter settings

Parameter Item	Initial Value	Parameter Range	Final Value
Learning rate	0.001	[0.0001, 0.01]	0.0005
Batch size	32	{16, 32, 64, 128}	64
Optimizer	Adam	{SGD, Adam, AdamW}	AdamW
Epochs	50	[30, 100]	80
Dropout rate	0.3	[0.1, 0.5]	0.2
Hidden layer size	256	{128, 256, 512}	256
Attention heads	4	{2, 4, 8}	8
Weight decay	1e-4	[1e-6, 1e-3]	5e-5
Activation function	ReLU	{ReLU, GELU, Tanh}	GELU
Loss function	CrossEntropy	{CrossEntropy, MSE}	CrossEntropy

4 Results analysis

4.1 Multimodal semantic alignment effect

In a multimodal semantic alignment experiment, to examine the correspondence between urban discourse and visual objects, we selected six representative text corpora: Hefei subway line expansion, Chaohu wetland protection, research achievements of the University of Science and Technology of China, construction of a business complex in the government district, the tourist attraction of the Huizhou Cultural District, and the concentration of technology enterprises in the High-tech Zone. These texts were then matched with six representative image objects: subway stations, Chaohu scenery, the USTC East Campus, government district buildings, the Huizhou Cultural District, and the Science and Technology Innovation Park. After word segmentation and entity recognition, the texts were mapped into a semantic space. The images were processed using a CLIP and BLIP-2 dual-channel encoder to obtain a unified embedding vector. Finally, the cross-modal similarity distribution was calculated to obtain the correspondence between the texts and images, as shown in Figure 4 .

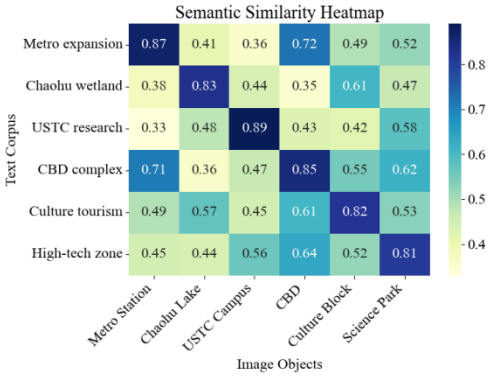

Figure 4: Semantic similarity heat map

Figure 4 shows a distribution pattern in which diagonal similarities are generally higher than off-diagonal regions. For example, the similarity between "Continued expansion of Hefei subway lines" and "Hefei subway station" reaches 0.87, while the similarity between "Remarkable achievements in Chaohu wetland protection" and "Chaohu scenery" is 0.83. These values above 0.8 are primarily due to the consistency between high-frequency geographic entities and image object features in the text corpus, resulting in a stable coupling between lexical expressions and visual elements. In contrast, the similarity between "USTC research results achieve another breakthrough" and "USTC East Campus" is 0.89, exceeding other combinations. This is because the large number of campus-specific vocabulary and image features in the research corpus are highly concentrated around campus buildings and landmarks, resulting in closer embedding space distances. Non-corresponding combinations mostly fall between 0.3 and 0.6. The similarity between "Government Affairs District Business Complex put into use" and "Chaohu scenery" is 0.36. This low value is due to the significant difference in semantic space between the economic and business attributes of the discourse and the visual features of the natural landscape, resulting in limited semantic intersection. The overall results show that cross-modal embeddings have clear performance in maintaining semantic consistency and separation.

4.2 Knowledge graph semantic relationship mining

In the semantic propagation experiment, several core entities in Hefei's city image were mapped as knowledge graph nodes, and their relationship strengths were modeled using a graph attention network. The entities covered six areas: transportation, ecology, education, economy, culture, and technology. The subway represents urban infrastructure, Chaohu Lake's ecology reflects the natural environment, the University of Science and Technology of China embodies education and research, the government district's business district showcases economic development, the Huizhou Cultural District points to history and humanities, and the high-tech zone's science and technology district links to innovative industries. After multimodal embedding and alignment, the corpus was converted into relationship edge weights. The training process, through iterative updates, achieves weighted propagation of semantic associations, forming a multi-domain interactive pattern. The resulting relationship strength distribution is shown in Figure 5 .

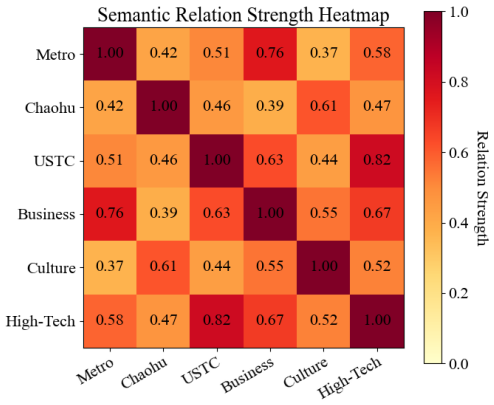

Figure 5: Knowledge graph relationship strength

in Figure 5 , the strength of the relationship between business in the government district and subway transportation reaches 0.76. This high value is attributed to the coupling between the spatial distribution and policy promotion of commercial complexes and subway line construction, and the fact that transportation convenience directly determines the agglomeration effect of business development in the government district. The strength of the relationship between USTC and science and technology in the high-tech zone reaches 0.82, significantly higher than the correlation between education and other entities. This phenomenon stems from the close cooperation mechanism between scientific research results transformation and technology enterprise

incubation, which leads to a high degree of coupling between the two in knowledge flow and industrial implementation. The strength of the relationship between Chaohu Lake ecology and Huizhou culture is 0.61, higher than the 0.39 between Chaohu Lake ecology and business in the government district. This is due to the tourism industry's combined development of natural landscapes and historical and cultural resources, which increases the frequency of their connection in cultural narratives and tourism discourse. Overall, the data distribution reflects the interactive pattern of Hefei's city image in terms of transportation, science and technology, and ecological culture, and reveals the differentiated connections between different areas formed by policy planning and industrial needs.

Experiments at the discourse interaction level aimed to uncover the co-occurrence patterns of entities related to Hefei's city image within multimodal corpora. By counting the joint occurrences of different entities within the text, we quantified the intensity of interaction, which served as an important input for the subsequent semantic dissemination of the knowledge graph. The data processing phase involved entity recognition and pairing statistics on the corpus, generating entity interaction intensity data covering dimensions such as transportation, ecology, education, economy, culture, and technology. This data was then used to generate structured results, as shown in Table 5.

Table 5: Discourse interaction relationship strength data table

Discourse Interaction Pair	Interaction Strength
Metro Transportation – Government Affairs Business District	134
University of Science and Technology of China – High-tech Zone Technology	152
Chaohu Ecology – Huizhou Culture	118
Government Affairs Business District – University of Science and Technology of China	96
High-tech Zone Technology – Government Affairs Business District	103
Metro Transportation – University of Science and Technology of China	87

The data results show that the interaction strength between USTC and the High-tech Zone's science and technology reached 152, higher than that of other entity pairs. This is due to the frequent mention of scientific research results and industrial transformation in Hefei's science and technology innovation discourse, resulting in a high co-occurrence frequency between the two in the text. The interaction strength between subway transportation and government affairs district business was 134, which is related to the close connection between transportation construction and business district expansion during Hefei's urbanization process. The discourse context of transportation infrastructure is often coupled with the functional description of commercial complexes. The interaction strength between Chaohu Lake ecology and Huizhou culture was 118, mainly driven by the integration of ecotourism and cultural tourism narratives. The two entities often appear together in tourism promotion and regional development reports. In contrast, the interaction strength between subway transportation and USTC was only 87, lower than that of other entity pairs. This is due to the relatively limited contextual connection between public transportation and higher education, and the related discourse focuses primarily on commuting convenience rather than deep interaction. Therefore, the differences in interaction strength between different entity pairs stem from the differentiated structure of the discourse scenarios of industry, ecology, and transportation development in Hefei's multimodal corpus.

4.3 Emotional and thematic distribution of city image

In this part of the experiment, the study conducted a quantitative analysis of the sentiment tendency and topic coverage of the multimodal data of Hefei city discourse. The text and image data after cross-modal semantic alignment were used to extract the sentiment distribution ratio that changes over time, and the coverage rate of different time periods was calculated based on the topic modeling results. In this way, the dynamic evolution characteristics of the city image in discourse interaction were observed. The final results are shown in Figure 6 .

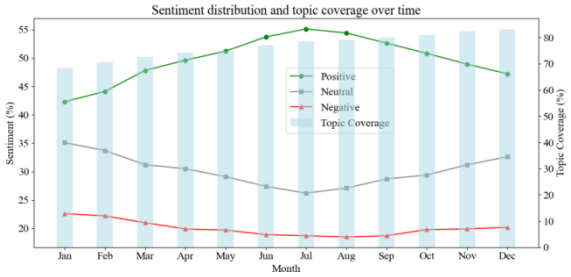

Figure 6: Sentiment distribution and topic coverage over time

The data shows that positive sentiment remained at a low level of 42.3% at the beginning of the year, then gradually increased, reaching a peak of 55.1% in July. This phenomenon is attributed to the intensive city publicity activities and social and cultural events during the summer vacation, which led to the generation of more positive discourse. Neutral sentiment dropped to a low of 26.2% in July, as a large number of discourses exhibited a clear bias during this period, reducing the proportion of neutral expressions. Negative sentiment remained relatively stable overall, ranging from 18.5% to 22.6%, reflecting that negative discourse was primarily driven by sudden events rather than long-term trends. Topic

coverage gradually increased from 68.4% at the beginning of the year to 83.1% by the end of the year. This growth was driven by the accumulation of corpus data and the continuous expansion of diverse topics, ultimately demonstrating a dynamic pattern of fluctuating sentiment and expanding discourse topics.

4.4 Dynamic evolution of city image

In the experimental part of the dynamic evolution law, the study took the Hefei city multimodal corpus as the object, used the time fusion Transformer to perform sequence modeling on the quantitative indicators of the city image for ten months, combined with the comparison of real observations and predicted results to test the performance of the model in trend fitting and stability maintenance, and measured its stability through the segmented statistics of the error index. At the same time, the feature weight distribution output by the variable selection network was introduced to reveal the core semantic factors affecting the prediction, resulting in the results shown in Figure 7.

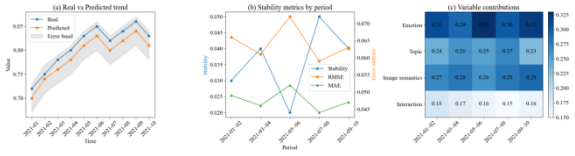

Figure 7: Dynamic evolution results of city image
Figure 7 (a) Trend comparison between actual value and predicted value
Figure 7 (b) Comparison of stability indicators in different time periods
Figure 7 (c) Variable contribution heat map

In the trend comparison, the true values from 2021-01 to 2021-10 remained consistently between 0.72 and 0.86, while the deviation between the predicted and true values was generally controlled at around 0.02, with the upper and lower limits of the error band remaining within ±0.03, respectively. This indicates that the model maintains high fitting accuracy over the long term. Regarding stability indicators, the dynamic evolution stability from 2021-03 to 2021-04 was 0.84, the RMSE dropped to 0.061, and the MAE remained at 0.046. Compared to 2021-05 to 2021-06, when the stability dropped to 0.82, the RMSE rose to 0.072, and the MAE rose to 0.052, the difference stems from the significant volatility of the urban discourse corpus during this period, and the more discrete distribution of sentiment polarity expressions. The variable contribution results show that sentiment polarity's weight ranged from 0.29 to 0.33 from March 2021 to April 2021, consistently exceeding the interaction frequency's weight of 0.15 to 0.18. This indicates that during this period, prediction performance was primarily driven by fluctuations in sentiment input. Overall, the model's stability was most significantly influenced by fluctuations in sentiment and topic within the data period. The average weight of sentiment features across the entire sequence is 0.31, and the average weight of interaction frequency features is 0.17. The attention distribution in GAT increases on transportation and science-related nodes during the months of infrastructure expansion and new research announcements, while the temporal attention in the TFT increases on intervals corresponding to higher discourse density.

4.6 Model comparison and temporal sensitivity analysis

To validate the contribution of each module, control groups with LSTM, GRU, and a no-graph variant were constructed under identical configurations. This paper conducts supplementary experiments focusing on the performance differences of temporal network structures, the sequence shift caused by changes in temporal granularity, and the predictive response of sudden event windows. Multiple control groups were constructed using a unified input configuration to present the quantitative performance of temporal representation under different structures and temporal resolutions. Based on standardized corpus segments, three types of networks were trained, generating sequence results for daily, weekly, and monthly time granularities. Sequence variation amplitudes were extracted within short-term windows of selected policy events, failure events, and event events. All experiments were conducted with fixed parameters, and the output statistical results were organized into a structured form according to a unified process. The final results are shown in Figure 8.

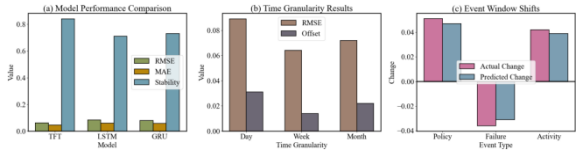

Figure 8: Comparison of Temporal Model Structures, Temporal Granularity, and Event Window Offset Analysis
Figure 8(a) Performance Comparison of Temporal Model Structures
Figure 8(b) Prediction Results at Different Temporal Granularities
Figure 8(c) Prediction Offset of Sudden Event Windows

In Figure 8, Figure 8(a) shows the RMSE, MAE, and stability of the three temporal structures; Figure 8(b) shows the error and fluctuation offset at daily, weekly, and monthly granularities; and Figure 8(c) shows the actual and predicted changes in policy events, facility failures, and event windows. TFT exhibits a low RMSE of 0.061 and a low MAE of 0.046, while LSTM has an RMSE of 0.083 and GRU has an RMSE of 0.079. This difference indicates that the multi-head attention and gating structure maintains a relatively stable error range when handling cross-period variations because the historical dependence of the structure's output maintains a consistent attention distribution over long periods. The weekly granularity, as shown in Figure 8(b), has an RMSE of 0.064 and an offset of 0.014, which is lower than the daily granularity's RMSE

of 0.089 and offset of 0.031. This indicates that the smooth input at the weekly scale reduces numerical fluctuations caused by short-term corpus volatility. This is because the frequency distribution after weekly aggregation maintains a relatively stable inter-item ratio. The monthly granularity has an RMSE of 0.072 and an offset of 0.022, positioned between the daily and weekly scales. This shows that the monthly scale entropy value is reduced, but structural deficiencies still exist. This is because excessive time compression weakens the recognition of some semantic jumps. Figure 8(c) shows that the actual change for policy events is 0.051, and the predicted change is 0.047. The closeness of the values indicates that the model's response within the short time window follows the linear displacement of the input. This is because the text density during the event period generates a continuous semantic increase. The actual change for facility failures is -0.036, and the predicted change is -0.031, showing a negative shift in the same direction. This is because the entity relationships in the graph structure are concentrated in the failure reports. The actual change for event activities is 0.042, and the predicted change is 0.039, indicating that the model keeps pace with the changes in visual text coupling during the event-intensive phase. This is because the semantic clusters generated by high-frequency terms during this phase enhance the directional consistency of the embedding sequence. Overall, the error range, shift magnitude, and event response presented in the three types of comparative experiments all show structural differences, thus forming complete temporal behavioral characteristics.

4.7 Ablation study on graph semantic propagation

To examine the specific effect of graph-based semantic propagation, an ablation variant was constructed by removing the GAT module while keeping the multimodal encoder and temporal predictor unchanged. This setting isolates the contribution of structural semantic relations from other components and enables a direct performance comparison. The evaluation results are presented in Table 6.

Table 6: Performance comparison between the full model and the no-graph ablation variant

Model	RMSE	MAE	Graph Consistency Score	Relative Change
Full model (with GAT)	0.061	0.046	0.84	—
No-graph ablation (w/o GAT)	0.071	0.054	0.76	+0.010 RMSE

Figure 6 illustrates the deviation patterns produced by the two model variants. The ablation model without graph propagation exhibits noticeably larger fluctuations, particularly during periods characterized by concentrated entity mentions or event-driven discourse bursts. This pattern arises because the removal of the GAT module prevents the relational dependencies among urban entities—such as functional ties between transportation and commerce, or knowledge exchange channels between research institutions and high-tech zones—from being structurally integrated into the multimodal representation. As a consequence, the model relies mainly on token-level co-occurrence signals, making it more sensitive to short-term fluctuations in the corpus. The reduction in the graph consistency score from 0.84 to 0.76 further indicates that the absence of semantic propagation weakens the internal coherence of the encoded representations, leading to reduced predictive smoothness and higher error accumulation. The comparison confirms that graph-based reasoning provides stabilizing structural information that enhances both numerical accuracy and semantic robustness during temporal forecasting.

5 Discussion section

This section compares the results with prior studies using CLIP-only alignment, GCN-based graph reasoning, and LSTM temporal prediction. The proposed method achieves higher similarity scores and lower RMSE values because the GAT enhances multi-entity propagation and the TFT retains long-range dependencies with greater stability. Deviations from earlier benchmarks appear in high-volatility periods, during which sentiment and topic bursts lead to increased dispersion, yet the model maintains lower error accumulation across intervals.

6 Conclusion

This study introduced a graph-temporal multimodal framework that unifies cross-modal alignment, knowledge-graph–based semantic reasoning, and temporal sequence modeling to quantify the dynamic evolution of city image. Concretely, CLIP and BLIP-2 project text–image pairs into a shared semantic space; an urban discourse knowledge graph coupled with a Graph Attention Network captures multi-actor relations among institutions, places, and events; and a Temporal Fusion Transformer, equipped with variable selection, models long-range dependencies and local bursts while preserving interpretability. On the Hefei case, the framework achieved strong cross-modal alignment (similarity = 0.87 for “Hefei subway line expansion” ↔ “subway station” images; 0.89 for “USTC research results” ↔ “USTC campus”), salient relational strengths in the knowledge graph (USTC ↔ High-tech-zone science-and-technology = 0.82), and accurate temporal forecasting (deviation ≤ 0.02; RMSE = 0.061; MAE = 0.046). These results indicate that the proposed approach delivers a coherent quantitative representation that simultaneously captures (i) semantic consistency across modalities, (ii) structured

inter-entity relations, and (iii) temporal dynamics of discourse.

Methodologically, the framework offers a reproducible pipeline for urban informatics tasks that require aligning heterogeneous media with relational structure and time—e.g., monitoring policy communication effects, benchmarking branding campaigns, or diagnosing discourse shifts around infrastructure expansions. Substantively, the interpretable variable-selection layer highlights which multimodal features most strongly drive city-image change, providing actionable signals for urban governance, public communication, and cultural strategy.

Limitations and future work. The present study focuses on images and text from a single city context; extending to videos, audio, and multilingual streams, as well as cross-city transfer, will test robustness under domain shift. Incorporating temporal knowledge graphs with event types and causal links could strengthen explanation beyond correlation. Finally, online learning for streaming data, fairness auditing across neighborhoods and demographic groups, and counterfactual analyses (e.g., policy or event interventions) are promising directions to enhance reliability, equity, and decision support in real-world deployments. The present results are conditioned by platform usage patterns and discourse habits specific to Hefei, and the model performance may vary when applied to cities with different communication structures or media environments. The dependence on local multimodal corpora limits direct transfer, and cross-city experiments require further collection of comparable multimodal posts.

Funding

This research was supported by the Scientific Research Project of Colleges and Universities in Anhui Province (Philosophy and Social Science), Research on the Construction of Hefei City Image Based on Multi-modal Discourse Interaction (Project No.: 2023AH052309), phased achievements; the Scientific Research Project of Colleges and Universities in Anhui Province (Philosophy and Social Science), Research on Overseas Transmission Ways and Strategies of Huangmei Opera (Project No.: 2022AH052801), phased achievements; and the Scientific Research Team Project of Anhui International Studies University, Research on Ecological Niche of Luzhou Dialect from Ecological Linguistics Perspective (Project No.: awkytd202205), phased achievements.

References

[1] Yuan J, Zhang L, Kim C S. Multimodal interaction of MU plant landscape design in marine urban based on computer vision technology[J]. Plants, 2023, 12(7): 1431. https://doi.org/10.3390/plants12071431

[2] Han Y. The Study on the Multimodal Discourse Construction and Communication of Zhengzhou's National Central City Image[J]. Journal of Business and Management Studies, 2025, 7(1): 222-226. https://doi.org/10.32996/jbms.2025.7.1.17

[3] Wang Y, Feng D. History, modernity, and city branding in China: a multimodal critical discourse analysis of Xi'an's promotional videos on social media[J]. Social Semiotics, 2023, 33(2): 402-425. https://doi.org/10.1080/10350330.2020.1870405

[4] Hosseini A, Barekat B. A multimodal critical discourse analysis of city as text: investigation of meaning metafunctions of Rasht's Imam Khomeini Street[J]. Visual Communication, 2025, 24(1): 90-113. https://doi.org/10.1177/14703572221128886

[5] Huang J, Xiao W, Wang Y. Use of metadiscourse for identity construction in tourist city publicity: A comparative study of Chinese and Australian social media discourse[J]. Heliyon, 2023, 9(12). https://doi.org/10.1016/j.heliyon.2023.e23122

[6] Chen X. Representing cityscape through texts and images: Translations of multimodal public notices in Macao[J]. Asia Pacific Translation and Intercultural Studies, 2023, 10(1): 53-70. https://doi.org/10.1080/23306343.2023.2165004

[7] Tivyaeva I V. Memorial plaques in multimodal urban discourse: A visual narrative reflecting Moscow's glorious past[J]. Visual Anthropology, 2023, 36(1): 38-53. https://doi.org/10.1080/08949468.2023.2168960

[8] Sukma B P. Constructing and promoting national identity through tourism: A multimodal discourse analysis of Indonesian official tourism website[J]. Linguistik Indonesia, 2021, 39(1): 63-77. https://doi.org/10.26499/li.v39i1.197

[9] Qi W, Sorokina N. Constructing online tourist destination images: a visual discourse analysis of the official Beijing Tourism website[J]. Chinese Semiotic Studies, 2021, 17(3): 421-448. https://doi.org/10.1515/css-2021-2006

[10] Yu Z, Xiao Z, Liu X. A data-driven perspective for sensing urban functional images: Place-based evidence in Hong Kong[J]. Habitat International, 2022, 130: 102707. https://doi.org/10.1016/j.habitatint.2022.102707

[11] Su L, Chen W, Zhou Y, et al. Exploring city image perception in social media big data through deep learning: A case study of Zhongshan City[J]. Sustainability, 2023, 15(4): 3311. https://doi.org/10.3390/su15043311

[12] Xie L, Feng X, Zhang C, et al. Identification of urban functional areas based on the multimodal deep learning fusion of high-resolution remote sensing images and Social Perception Data[J]. Buildings, 2022, 12(5): 556. https://doi.org/10.3390/buildings12050556

[13] Fan R, Li F, Han W, et al. Fine-scale urban informal settlements mapping by fusing remote sensing images and building data via a transformer-based multimodal fusion network[J]. IEEE Transactions on Geoscience and Remote Sensing, 2022, 60: 1-16. https://doi.org/10.1109/tgrs.2022.3204345

[14] Zhang F, Salazar-Miranda A, Duarte F, et al. Urban visual intelligence: Studying cities with artificial intelligence and street-level imagery[J]. Annals of the American Association of Geographers, 2024, 114(5): 876-897. https://doi.org/10.1080/24694452.2024.2313515

[15] Huang J, Obracht-Prondzynska H, Kamrowska-Zaluska D, et al. The image of the City on social media: A comparative study using “Big Data” and “Small Data” methods in the Tri-City Region in Poland[J]. Landscape and Urban Planning, 2021, 206: 103977. https://doi.org/10.1016/j.landurbplan.2020.103977

[16] Huang Y, Zheng B. Social media users' visual and emotional preferences of internet-famous sites in urban riverfront public spaces: a case study in Changsha, China[J]. Land, 2024, 13(7): 930. https://doi.org/10.3390/land13070930

[17] Reyes MF, Xie Y, Yuan X, et al. A 2D/3D multimodal data simulation approach with applications on urban semantic segmentation, building extraction and change detection[J]. ISPRS Journal of Photogrammetry and Remote Sensing, 2023, 205: 74-97. https://doi.org/10.1016/j.isprsjprs.2023.09.013

[18] Piadyk Y, Rulff J, Brewer E, et al. Streetaware: A high-resolution synchronized multimodal urban scene dataset[J]. Sensors, 2023, 23(7): 3710. https://doi.org/10.3390/s23073710

[19] Chen S, Zang Y, Yang P. City images in transnational travel vlogs from a multimodal perspective: an investigation of 20 port cities worldwide[J]. Online Media and Global Communication, 2024, 3(1): 82-107. https://doi.org/10.1515/omgc-2023-0034

[20] Chen X, Yu J, Zhu Y, et al. Short video-driven deep perception for city imagery[J]. Environment and Planning B: Urban Analytics and City Science, 2024, 51(3): 689-704. https://doi.org/10.1177/23998083231193236

[21] Kang Y, Cho N, Yoon J, et al. Transfer learning of a deep learning model for exploring tourists' urban image using geotagged photos[J]. ISPRS International Journal of Geo-Information, 2021, 10(3): 137. https://doi.org/10.3390/ijgi10030137

[22] Xue J, Jiang N, Liang S, et al. Quantifying the spatial homogeneity of urban road networks via graph neural networks[J]. Nature Machine Intelligence, 2022, 4(3): 246-257. https://doi.org/10.1038/s42256-022-00462-y

[23] Liu C, Wang Y, Li W, et al. An urban built environment analysis approach for street view images based on graph convolutional neural networks[J]. Applied Sciences, 2024, 14(5): 2108. https://doi.org/10.3390/app14052108

[24] Rashidan H, Musliman IA, Sani MJ, et al. Semantic labeling of 3D buildings by using graph neural network (GNN)[J]. The International Archives of the Photogrammetry, Remote Sensing and Spatial Information Sciences, 2024, 48: 307-311. https://doi.org/10.5194/isprs-archives-xlviii-4-w9-2024-307-2024

[25] Ma X, Zeng T, Zhang M, et al. Street microclimate prediction based on Transformer model and street view image in high-density urban areas[J]. Building and Environment, 2025, 269: 112490. https://doi.org/10.1016/j.buildenv.2024.112490

[26] Wang W, Teng Y, Yan L, et al. Image Experience Prediction for Historic Districts Using a CNN-Transformer Fusion Model[J]. Image Analysis and Stereology, 2025, 44(1): 11-23. https://doi.org/10.5566/ias.3361

[27] Cho M, Kim S, Choi D, et al. Enhanced BLIP-2 Optimization Using LoRA for Generating Dashcam Captions[J]. Applied Sciences, 2025, 15(7): 3712. https://doi.org/10.3390/app15073712

[28] Lee C, Jang J, Lee J. Personalizing text-to-image generation with visual prompts using BLIP-2[J]. 2023.

[29] Huang T, Wang Z, Sheng H, et al. Learning neighborhood representation from multi-modal multi-graph: Image, text, mobility graph and beyond[J]. arXiv preprint, 2021 , 21 (05): 02489.

[30] Zhang J, Chen R, Li S, et al. MGKGR: Multimodal Semantic Fusion for Geographic Knowledge Graph Representation[J]. Algorithms, 2024, 17(12): 593. https://doi.org/10.3390/a17120593

[31] Feng Z, Zeng Z, Guo C, et al. Temporal multimodal graph transformer with global-local alignment for video-text retrieval[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 33(3): 1438-1453. https://doi.org/10.1109/tcsvt.2022.3207910

[32] Hu B, Zhu L, Dong Q, et al. Physiological electrosignal asynchronous acquisition technology: Insight and perspectives[J]. IEEE Transactions on Computational Social Systems, 2024, 11(1): 5-24. https://doi.org/10.1109/tcss.2024.3350958

[33] Han Y. The Study on the Multimodal Discourse Construction and Communication of Zhengzhou’s National Central City Image[J]. Journal of Business and Management Studies, 2025, 7(1): 222-226. https://doi.org/10.32996/jbms.2025.7.1.17

[34] Guan S, Cheng X, Bai L, et al. What is event knowledge graph: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(7): 7569-7589.

[35] Joshi R. Introduction to graph neural network: A systematic review of trends, methods, and applications[J]. Applied Graph Data Science, 2025: 1-16. https://doi.org/10.1016/b978-0-443-29654-3.00017-x

[36] Yang Z, Zhang J, Li Z. Multi-scale time series prediction model based on deep learning and its application[J]. PLoS One, 2025, 20(7): e0325474. https://doi.org/10.1371/journal.pone.0325474

[37] Lisena P, Meroño-Peñuela A, Troncy R. MIDI2vec: Learning MIDI embeddings for reliable prediction of symbolic music metadata[J]. Semantic Web, 2022, 13(3): 357-377.

https://doi.org/10.3233/sw-210446
[38] Biometric Recognition: 18th Chinese Conference, CCBR 2024, Nanjing, China, November 22–24, 2024, Proceedings, Part II[M]. Springer Nature, 2025. https://doi.org/10.1007/978-981-96-1071-6_9
[39] Gu X, Liu C, Wang S. Biometric Recognition[J]. Lecture Notes in Computer Science, 2013, 8232: 34-42.