Towards Vision-Spatiotemporal Fusion in Traffic Forecasting: A Survey on Cross-Modal Alignment

Anna Wang[1], **Chao Zhang**[1*], **Mingwei Lin**[2], **Junbo Zhang**[3], **Zeshui Xu**[4], **Wentao Li**[5], **Pengfei Zhang**[6] and **Oscar Castillo**[7]

[1]School of Computer and Information Technology, Shanxi University, China
[2]College of Computer and Cyber Security, Fujian Normal University, China
[3]JD iCity, JD Technology, China
[4]Business School, Sichuan University, China
[5]College of Artificial Intelligence, Southwest University, China
[6]School of Intelligent Medicine, Chengdu University of Traditional Chinese Medicine, China
[7]Division of Graduate Studies and Research, Tijuana Institute of Technology, Mexico
{wanganna1, czhang}@sxu.edu.cn, lmwfjnu@fjnu.edu.cn, msjunbozhang@outlook.com, xuzeshui@263.net, drliwentao@swu.edu.cn, zhangpengfei@cdutcm.edu.cn, ocastillo@tectijuana.mx

Abstract

Traffic forecasting is evolving, with world models emerging as a powerful framework applicable to tasks such as core state, trajectory, event, and demand forecasting. These tasks involve both visual and spatiotemporal data, yet most existing methods treat them separately, hindering a unified understanding of traffic scenes in both semantic meanings and spatiotemporal dynamics. The fusion of the two modalities is critical for building models that comprehend complex traffic scenarios. However, the fusion issue faces two fundamental misalignments: semantic, where pixels conflict with traffic concepts, and geometric, which requires spatial intelligence to map 2D inputs into 3D. This survey reframes vision-spatiotemporal fusion via the unique lens of cross-modal alignment, addressing semantic and geometric failures that limit forecasting reliability. First, we categorize existing methods into three paradigms: feature-level, semantic-level, and task-level. This reveals their progression from low-level feature manipulation to high-level architectural integration. Second, we synthesize representative techniques per paradigm, highlighting geometric challenges such as cross-view association and spatial mapping. Third, we examine current datasets and benchmarks, highlighting their deficiencies in evaluating alignment. Finally, we outline future directions, including spatiotemporal intelligence for robust perception and holistic traffic world models. The unified framework establishes a reference for robust and explainable forecasting systems.

*Corresponding authors.

1 Introduction

Traffic forecasting has evolved, where visual or spatiotemporal data alone can no longer fully perceive and forecast complex traffic systems [Li *et al.*, 2025c]. With the growth of multi-source sensor networks, tasks such as state analysis and trajectory forecasting require the grasp of both scene semantics and spatiotemporal dynamics, making deep integration inevitable [Palladin *et al.*, 2025]. World models, which build computational frameworks that internally represent and reason about environmental dynamics, offer a novel method for traffic forecasting. They aim to unify visual and spatiotemporal data, covering core states, trajectory, event, and demand patterns [Ding *et al.*, 2025].

However, merging these data streams faces a challenge deeper than just handling different formats. The core issue is a mismatch in how information is represented. Cameras deliver raw pixels, shaped by changing angles, light, and obstructions, governed by optical and geometric rules [Jain *et al.*, 2025]. Forecasting models, however, work with organized structures such as graphs or time series that summarize overall traffic dynamics. This fundamental divide creates two main problems. One is a semantic gap: connecting raw visual patterns to meaningful traffic concepts. The other is a geometric gap: fitting 2D visual data into a coherent 3D world model. Common fusion techniques, such as simple feature concatenation, often cannot bridge these gaps. As a result, model performance tends to stagnate, and adapting to new scenarios remains difficult [Wang *et al.*, 2026].

Investigations into these issues are dispersed among distinct research communities, operating without a unified framework. For instance, substantial effort in computer vision is devoted to parsing traffic camera footage. Concurrently, the vision-spatiotemporal forecasting field advances via architectures employing graph neural networks (GNNs) and Transformers [Peddi *et al.*, 2024]. Separately, some studies recast time series as visual representations for convolutional neural network (CNN) processing [Lv *et al.*,

2025]. These related endeavors nonetheless leave the primary obstacle unresolved: coherently integrating real-time visual scenes with spatiotemporal models. Such fragmentation underscores the necessity for a consolidated viewpoint on vision-spatiotemporal alignment [Fang *et al.*, 2026; Chen *et al.*, 2026].

Prevailing surveys in vision-spatiotemporal traffic forecasting commonly classify methods by fusion strategy or timing, frequently sidestepping the essential question of meaningful cross-modal alignment. In contrast, this survey adopts a comprehensive view of traffic forecasting, encompassing (i) core state (e.g., flow, speed), (ii) trajectory & behavior, (iii) event & anomaly, and (iv) demand & service forecasting. Despite their varied objectives, these tasks share a foundational challenge: fusing real-time visual streams with spatiotemporal signals. Our central premise is that alignment granularity, the representational level at which cross-modal correspondences are established, constitutes the fundamental challenge common to this entire problem class. Categorizing prior works via this lens reveals an evolution from superficial correlation to profound integration and yields a pragmatic framework for evaluating methodological efficacy. We present a systematic analysis of cross-modal alignment for vision-spatiotemporal traffic forecasting[1]. Aiming to facilitate the development of spatiotemporal-intelligent traffic world models, this survey makes three principal contributions:

- **A taxonomy built around alignment granularity charts the route towards traffic world models.** The proposed framework organizes existing methods according to where they perform alignment: on features, semantics, or the task itself. This scheme makes it easier to examine and compare different lines of research.
- **A dedicated review of alignment techniques considers geometric challenges as the basis for spatial intelligence.** This survey moves beyond basic format conversion, collecting methods that fit live visual scenes into structured forecasting models and separately dissecting core geometric issues such as linking across views and spatial mapping.
- **An integrated analysis forms the groundwork for robust, spatially-aware world models.** We evaluate current datasets, benchmarks, and evaluation practices, identifying their shortcomings. From this analysis, we recommend steps to develop more unified and effective traffic forecasting systems.

2 Preliminaries and Taxonomy

Vision-spatiotemporal fusion combines visual data with structured traffic signals. The main aim is to improve traffic forecasting, spanning core state, trajectory, event, and demand forecasting. Here, $\mathcal{V}$ represents the visual modality, for instance, video from urban cameras, and $\mathcal{S}$ represents spatiotemporal signals, such as traffic flow graphs or vehicle trajectories. A forecasting model learns a function $\mathcal{F}$ to produce the forecasting $\hat{\mathbf{y}}$ from both types of input:

$$\hat{\mathbf{y}} = \mathcal{F}(\mathbf{V}, \mathbf{S}; \mathbf{\Theta}), \quad \mathbf{V} \in \mathcal{V},\ \mathbf{S} \in \mathcal{S}, \tag{1}$$

where $\mathbf{\Theta}$ contains all model parameters. One main difficulty comes from the fact that visual and spatiotemporal data are very different in both meaning and structure. Bridging the semantic and geometric differences requires alignment, which is not straightforward in real-world urban traffic scenes.

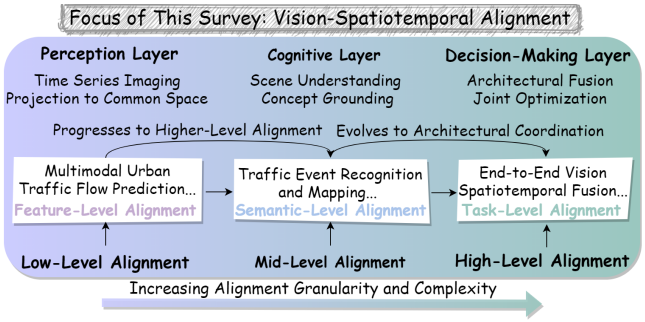

Figure 1: Levels of alignment for vision and spatiotemporal data. Methods are categorized as feature-level, semantic-level, or task-level alignment.

To organize diverse methods, we classify methods by the level at which alignment is carried out. This taxonomy focuses on three levels: feature, semantic, and task. Figure 1 shows this range, from methods that adjust low-level features to methods that coordinate full model modules. The figure also reflects a trend in practice: vision is increasingly treated as an integral part of forecasting, rather than just a source of engineered features.

At the feature level, visual and spatiotemporal inputs are mapped into a common latent space $\mathcal{Z}$ [Liu *et al.*, 2025]. The model learns $f_v : \mathcal{V} \to \mathcal{Z}$ and $f_s : \mathcal{S} \to \mathcal{Z}$ so that the resulting representations are comparable. Methods either transform one modality to resemble the other or embed both into a shared space. This level helps establish initial associations but often does not capture high-level traffic semantics or the geometric layout.

Semantic-level alignment establishes links between interpretable concepts across modalities [Shaw *et al.*, 2025]. Let $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ denote traffic concepts extracted from images via $g_v : \mathcal{V} \to \mathcal{C}$, and $\mathcal{D} = \{d_1, d_2, \ldots, d_L\}$ denote states in the spatiotemporal model. A mapping $h : \mathcal{C} \to \mathcal{D}$ connects visual concepts to model states. This method addresses the semantic gap, but it depends on accurate concept detection and proper geometric fitting.

Task-level alignment works differently, coordinating dedicated modules for visual and spatiotemporal data, $\mathcal{M}_v$ and $\mathcal{M}_s$, in an end-to-end system [Khan and Gauch, 2025]. Their outputs are combined using:

$$\hat{\mathbf{y}} = \mathcal{F}_{\text{fuse}}(\mathcal{M}_v(\mathbf{V}), \mathcal{M}_s(\mathbf{S}); \mathbf{\Theta}), \tag{2}$$

with all parameters $\mathbf{\Theta}$ learned jointly. Here, alignment emerges automatically during training, guided only by the final forecasting task. Key research questions include how

[1]https://github.com/Anna042023/Vision-Spatiotemporal-Fusion-in-TF

to make the modules interact effectively, how to avoid dominance by one modality, and how to maintain model interpretability.

The main difference between semantic and task-level alignment is the use of intermediate concepts. Semantic alignment explicitly defines interpretable traffic concepts, while task-level alignment integrates coordination directly into end-to-end learning. This distinction provides a straightforward way to classify the methods discussed in the following sections.

3 Vision-Spatiotemporal Alignment in Traffic Forecasting

Based on the alignment taxonomy, this section analyzes cross-modal alignment methods within traffic forecasting, encompassing core state, trajectory, event, and demand forecasting. We examine how current techniques operate at each granularity level, detailing their design and the challenges of aligning visual inputs with structured traffic representations. This progression from feature-level to semantic and task-level alignment marks a clear evolution towards more deeply integrated systems for complex urban environments.

3.1 Feature-Level Alignment

Feature-level alignment methods aim to integrate visual observations $\mathcal{V}$ with spatiotemporal traffic data $\mathcal{S}$ by projecting them into a shared feature space $\mathcal{Z}$. These techniques primarily rely on statistical objectives that encourage similarity between features, without explicitly incorporating traffic semantics [Gong *et al.*, 2022]. This paradigm provides a foundational, though often abstract, mechanism for merging heterogeneous data streams in multimodal forecasting.

A dominant strategy involves converting spatiotemporal traffic data into visual representations. Traffic time series $\mathbf{S} = \{s_t\}_{t=1}^{T}$ from roadway sensors are encoded into 2D image-like formats $\mathbf{I}_s$ using transformations such as heatmaps, Gramian angular fields (GAF), or recurrence plots (RP) [Ni *et al.*, 2025]. The GAF transformation, for instance, maps a normalized traffic time series $\tilde{\mathbf{s}} = \{\tilde{s}_t\}_{t=1}^{T}$ to a matrix $\mathbf{G}$ with elements defined as:

$$\mathbf{G}_{ij} = \cos(\phi_i + \phi_j), \quad \phi_i = \arccos(\tilde{s}_i),\ \tilde{s}_i \in [-1, 1], \quad (3)$$

which captures temporal correlations within the signal. This conversion enables the application of pre-trained convolutional neural networks, effectively framing traffic forecasting as an image-to-image translation problem [Ferreira *et al.*, 2025]. VisionTS [Chen *et al.*, 2025] exemplifies this method, transforming traffic flow and speed series into heatmaps to achieve strong zero-shot forecasting performance that rivals specialized models with minimal fine-tuning. A significant limitation, however, is the decoupling of the imaged traffic signal from concurrent visual streams $\mathbf{V}$ (e.g., live camera feeds). This ignores real-time semantic and geometric correspondences between the visual scene and the measured traffic states, reducing the task to signal processing rather than contextual reasoning, a critical shortcoming in dynamic urban settings where visual context directly influences traffic conditions [Zhang *et al.*, 2023].

An alternative strategy employs implicit encoding via dual-branch encoders that learn separate mappings $f_v : \mathcal{V} \to \mathcal{Z}$ and $f_s : \mathcal{S} \to \mathcal{Z}$. The alignment of their outputs, $\mathbf{z}_v = f_v(\mathbf{V})$ and $\mathbf{z}_s = f_s(\mathbf{S})$, is often driven by auxiliary contrastive learning objectives:

$$\mathcal{L}_{\text{cont}} = -\log \frac{\exp(\text{sim}(\mathbf{z}_v, \mathbf{z}_s)/\tau)}{\sum_{j \neq i} \exp(\text{sim}(\mathbf{z}_v^{(j)}, \mathbf{z}_s^{(i)})/\tau)}, \quad (4)$$

where $\text{sim}(\cdot, \cdot)$ denotes cosine similarity and τ is a temperature parameter [Huang *et al.*, 2024]. Although more flexible than explicit transformation, this paradigm frequently struggles with the inherent heterogeneity between CNN-processed images and GNN-processed traffic graphs. Minimizing feature distance alone may fail to establish semantically meaningful correspondences essential for traffic dynamics, such as associating a specific visual event (e.g., an accident) with its impact on network-wide flow.

The principal critique of feature-level alignment centers on its abstraction. By operating on low-level features, these methods obscure the high-level traffic semantics necessary for explainable forecasts. They also typically lack explicit mechanisms to resolve the geometric discrepancy between 2D image coordinates and 3D traffic network geometry. Consequently, the performance of such methods may saturate in complex scenarios where contextual understanding is as crucial as forecasting accuracy [He *et al.*, 2025].

3.2 Semantic-Level Alignment

Semantic-level alignment addresses feature-level limitations by establishing interpretable links between traffic concepts from visual streams $\mathcal{V}$ and states in structured models $\mathcal{S}$. A semantic extractor $g_v : \mathcal{V} \to \mathcal{C}$ serves as the core component, transforming raw pixels into traffic semantics $\mathcal{C} = \{c_1, c_2, \ldots, c_K\}$ such as accidents or lane blockages [Nie *et al.*, 2023]. This strategy aligns more effectively with human reasoning processes.

Early forecasting methods depend on pre-trained detectors to extract fixed attributes such as vehicle counts. Contemporary methods utilize vision-language models (VLMs) for open-vocabulary scene understanding relevant to traffic management [Lin *et al.*, 2025a]. Structured intermediate representations such as scene graphs further enhance this process by modeling relationships between traffic entities. This improves generalization in diverse conditions where predefined ontologies prove insufficient. Recent advances integrate VLMs with scene graphs to support this grounding process [Zeng *et al.*, 2025].

Extracted semantic concepts $\mathcal{C}$ require grounding to dynamic states $\mathcal{D} = \{d_1, d_2, \ldots, d_L\}$ within the spatiotemporal model, such as forecasted speed on graph edges. A grounding function $h : \mathcal{C} \to \mathcal{D}$ establishes this connection via learned or heuristic mapping. For instance, the concept "congestion at intersection X" (c_k) should associate with speed reductions (d_l) on corresponding road segments. This process inherently addresses both semantic and geometric gaps, typically involving spatial registration and temporal synchronization (sync.).

A standard formulation minimizes a grounding loss:

$$\mathcal{L}_{\text{ground}} = \sum_{k=1}^{K} \|\mathbf{r}(c_k) - \mathbf{q}(d_{l_k})\|^2, \quad \text{where } l_k = h(c_k), \quad (5)$$

where $\mathbf{r}(c_k)$ and $\mathbf{q}(d_{l_k})$ denote concept and state representations.

Primary challenges include the inherent ambiguity and compositionality of real-world traffic semantics, the temporal evolution of concepts, and limited availability of high-quality annotations for training robust extractors [Tang *et al.*, 2024; Wu *et al.*, 2023]. Given these constraints, developing mechanisms that extend beyond supervised learning on fixed labels remains difficult. Future directions may involve self-supervised semantic extraction from unlabeled footage and frameworks for compositional representations.

3.3 Task-Level Alignment

Task-level alignment represents the deepest integration form, embedding alignment objectives within end-to-end optimization of the complete forecasting architecture $\mathcal{F}$. A fusion function $\mathcal{F}_{\text{fuse}}$ coordinates a vision component $\mathcal{M}_v(\mathbf{V})$ and a spatiotemporal component $\mathcal{M}_s(\mathbf{S})$ directly for final forecasting performance, without explicit semantic supervision [Yi *et al.*, 2024]. This enables models to autonomously discover effective cross-modal interactions.

Common architectural designs incorporate attention-based mechanisms and gated fusion modules tailored for traffic data. Cross-modal attention layers compute relevance between a visual feature $\mathbf{v}_t$ and a traffic feature $\mathbf{s}_i$:

$$e_{t,i} = \frac{\mathbf{v}_t \mathbf{W}_q (\mathbf{s}_i \mathbf{W}_k)^\top}{\sqrt{d}}, \quad (6)$$

with corresponding attention weights:

$$\alpha_{t,i} = \frac{\exp(e_{t,i})}{\sum_{j=1}^{N} \exp(e_{t,j})}, \quad \hat{\mathbf{y}}_t = \sum_{i=1}^{N} \alpha_{t,i}, \mathbf{s}_i. \quad (7)$$

Gated fusion modules dynamically weight modal contributions. A gating vector $\mathbf{g} = \sigma(\mathbf{W}_v \mathbf{z}_v + \mathbf{W}_s \mathbf{z}_s + \mathbf{b})$ controls the fusion process: $\tilde{\mathbf{z}} = \mathbf{g} \odot \mathbf{z}_v + (\mathbf{1} - \mathbf{g}) \odot \mathbf{z}_s$, where $\mathbf{W}_*$ denotes learnable weights, σ represents the sigmoid function, $\mathbf{b}$ is a bias vector, and $\odot$ indicates element-wise multiplication. More recent methods explore unified transformer-based architectures that treat visual patches and traffic graph nodes as tokens for full cross-modal attention [Lee *et al.*, 2022].

The primary strength of task-level alignment its holistic, performance-driven optimization also constitutes its main weakness for trustworthy forecasting. The implicit alignment process creates opaque behavior, obscuring learned cross-modal correspondences and complicating debugging. Modal imbalance presents another critical issue. When historical traffic data dominate near-term forecasting, models may underutilize visual streams despite their value for anomaly anticipation [Dai *et al.*, 2025]. This poses risks in safety-critical applications requiring transparency. Research challenges therefore center on designing interaction mechanisms that balance expressiveness with interpretability while managing information imbalances [Wang *et al.*, 2025; Li *et al.*, 2025d].

The central challenge involves designing fusion mechanisms that balance forecasting accuracy with interpretability for practical deployment. Promising directions include visualizing cross-modal attention patterns in traffic scenarios, introducing lightweight semantic bottlenecks to guide fusion, and developing methods to actively balance modality contributions during training.

Table 1 systematically compares representative methods across these three granularity levels. This survey categorizes alignment granularity into feature-level, semantic-level, and task-level paradigms, revealing a trade-off between interpretability and holistic performance across core state, trajectory, event, and demand forecasting tasks. Feature-level methods offer transparent yet abstract transformations. Semantic-level methods align interpretable concepts but face extraction challenges. Task-level methods achieve high performance via end-to-end optimization but lack transparency. The distinction between semantic-level and task-level alignment is defined by the use of explicit intermediate semantic variables. This taxonomy clarifies the progression from correlation to deep integration and provides a framework for method evaluations.

4 The Geometry Gap in Vision-Spatiotemporal Alignment

Merging visual data $\mathcal{V}$ with structured spatiotemporal traffic data $\mathcal{S}$ faces a basic challenge that goes beyond semantic differences: the geometry gap. This problem arises from a mismatch between the 2D, projective nature of images and the 3D, metric world where traffic dynamics $\mathcal{S}$ occur. Semantic-level alignment interprets visual content but usually assumes spatial correspondence already exists. In reality, establishing this correspondence is a core geometric problem vital for fusion accuracy. Figure 2 shows how the geometry gap appears via two linked challenges that define spatial intelligence needs for world models: cross-view association and spatial coordinate mapping.

4.1 Cross-View Association

Employing distributed camera networks requires linking the same physical object across multiple, often non-overlapping views. This is essential for building consistent traffic flow pictures across a city [Yang *et al.*, 2021]. Formally, with cameras $\mathcal{C}_{\text{cam}} = \{C_1, C_2, \ldots, C_N\}$ and detected objects $\mathcal{O}_i$ in the view C_i, the goal is a matching function $\mathcal{M}$ that maps objects from different views to the same global entity if they are the same physical object.

Traditional methods use re-identification (ReID) models. These learn discriminative feature embeddings, often trained with a triplet loss:

$$\mathcal{L}_{\text{triplet}} = \max\big(0, \text{dist}(\mathbf{f}_a, \mathbf{f}_p) - \text{dist}(\mathbf{f}_a, \mathbf{f}_n) + \alpha\big), \quad (8)$$

where $\mathbf{f}_a$, $\mathbf{f}_p$, and $\mathbf{f}_n$ are feature embeddings for anchor, positive, and negative samples; $\text{dist}(\cdot, \cdot)$ is a distance metric; and α is a margin. Traffic scenes add tough challenges like viewpoint changes, lighting differences, and similar-looking vehi-

Method	Align.	Visual Input	S-T Input	Mechanism	Traffic Task	Key Insight
VisionTS [Chen *et al.*, 2025]	• (F)	Time-series image (GAF)	Traffic time series	Visual MAE pre-training	Zero-shot forecasting	Transforms signals to images; bridges vision pre-training to traffic forecasting.
Dual-Encoder [Gong *et al.*, 2022]	• (F)	Traffic camera image	Traffic graph nodes	Bi-level adversarial alignment	Multi-modal forecasting	Task-aware data and refined feature alignment for domain adaptation.
Semantic Extractor [Nie *et al.*, 2023]	◦ (S)	Traffic video	Traffic graph nodes	Pixel-to-concept parsing	Explainable event forecast	Guides reconstruction with traffic semantics (instances, events).
Scene Graph Anticipation [Peddi *et al.*, 2024]	◦ (S)	Traffic scene images	Graph-structured events	Neural ODE/SDE modeling	Scene understanding	Models continuous dynamics of traffic entity interactions.
Gated Fusion [Yi *et al.*, 2024]	▪ (T)	Visual feature sequence	Spatiotemporal graph features	Multi-dimensional spatiotemporal interaction	Multi-task forecasting	Captures cross-interactions for continuous multi-task learning.
Unified ViT-GNN Transformer [Lee *et al.*, 2022]	▪ (T)	Camera image patches	Traffic graph nodes	Cross-modality attention fusion	Pedestrian detection	Explores modality-specific features for traffic perception.

Legend: Align.: Alignment Granularity (•: Feature-level (F), ◦: Semantic-level (S), ▪: Task-level (T)). S-T Input: Spatiotemporal Input. The table summarizes core methodologies discussed in Section 3.

Table 1: Representative methods for vision-spatiotemporal alignment in traffic forecasting.

cles [Kannan and Min, 2025]. Newer methods add spatial-temporal constraints, using camera layout and travel time estimates to make association more reliable [Wang *et al.*, 2024].

Another strategy first locates objects in 3D before associating them, reducing the problem to applying distance thresholds in 3D space, which avoids appearance confusion but requires very accurate calibration and depth estimation [Jiang *et al.*, 2024]. Hybrid methods combine appearance features with geometric cues for better results [Sun *et al.*, 2024]. Lasting issues such as occlusions and calibration drift still hurt reliability.

4.2 Spatial Coordinate Mapping

Merging visual data with traffic models requires converting 2D image coordinates to 3D network locations. This spatial coordinate mapping is key for linking visual events to specific road segments or world coordinates.

The standard solution uses explicit geometric mapping via camera calibration. For a point on the road plane with world coordinates (X, Y), its projection to image coordinates (u, v) follows the camera projection model:

$$\gamma \begin{bmatrix} u \\ v \\ 1 \end{bmatrix} = \mathbf{K}[\mathbf{R}|\mathbf{t}] \begin{bmatrix} X \\ Y \\ 0 \\ 1 \end{bmatrix}, \tag{9}$$

where $\mathbf{K}$ is the intrinsic matrix, $[\mathbf{R}|\mathbf{t}]$ are extrinsic parameters, and γ is a scale factor. However, keeping accurate calibration across many cameras is challenging and error-prone, and works poorly in non-flat scenes [Kong *et al.*, 2024].

Learning-based methods are emerging to fix these issues. One method treats mapping as regression, training networks

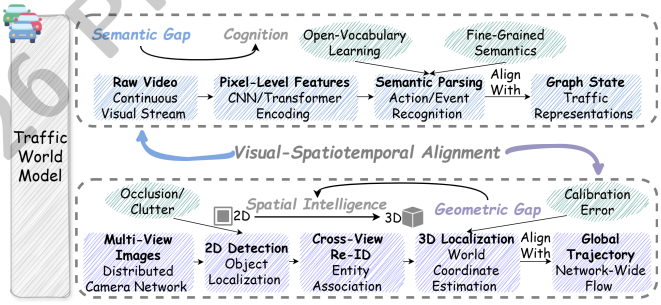

Figure 2: Dual pillars of a traffic world model: semantic and geometric gaps. The upper part illustrates the semantic gap, requiring parsing of raw video into traffic concepts aligned with spatiotemporal graph states. The lower part illustrates the geometry gap, necessitating cross-view association and 3D spatial mapping to align with global trajectories or road networks. Key technical hurdles are highlighted at each stage.

to forecast world coordinates straight from image features [Mirzaie and Rosenhahn, 2025]. Another promising direction uses neural rendering frameworks such as NeRF. These implicitly capture dense 3D geometry from multiple views [Hua and Wang, 2024]. Once trained, such models can provide accurate 3D localization, offering automatic spatial alignment that adapts to scene changes.

4.3 Frontiers in Geometric Learning

The field is moving towards end-to-end geometric learning. This makes spatial intelligence a core, learnable part of the world model, not a separate step. The goal is to learn geometric correspondences as latent representations useful for downstream tasks.

One frontier explores geometry-aware neural architectures. Here, cross-modal attention mechanisms use geometric priors [Shrout *et al.*, 2025]. Another area looks at integrating differentiable rendering [Zhou *et al.*, 2024]. This lets gradients flow via rendering processes to adjust geometric parameters. These methods hint at future self-calibrating vision-spatiotemporal systems.

A major remaining challenge is how to evaluate alignment quality. Current works mainly use final forecasting accuracy as an indirect measure [Wu *et al.*, 2025]. Developing direct metrics, such as cross-camera association accuracy and spatial mapping error, is vital for progress. Moreover, datasets with high-quality geometric annotations are still rare, limiting training and testing of advanced models.

5 Resources and Evaluations

This section reviews widely adopted datasets, common evaluation protocols, and challenges encountered during real-world deployment. The objective is to depict the field's present configuration and to reveal core limitations that hinder advancement in urban traffic management.

5.1 Datasets and Benchmarks

For alignment-driven methods to function robustly, multi-modal traffic datasets should ensure consistent synchronization across spatial and temporal axes to support comprehensive forecasting tasks, including core state, trajectory, event, and demand forecasting. As outlined in Table 2, prior studies rely primarily on three real-world benchmarks along with two supplementary data forms. Accordingly, the range of attainable investigations is largely determined by the alignment properties embedded within each dataset.

Three benchmarks are particularly influential: nuScenes [Caesar *et al.*, 2020], Waymo Open Dataset [Sun *et al.*, 2020], and CityFlow [Tang *et al.*, 2019]. Each offers synchronized video, LiDAR, and map resources, together with detailed object trajectories. Forecasting models leverage these extensive spatiotemporal annotations to derive traffic graphs or sequential representations. A central alignment difficulty concerns large-scale geometric registration, specifically linking 2D image pixels to 3D map coordinates and object trajectories across multiple sensors. Although their scale enables city-level analysis, the substantial annotation cost leads to sparser and less accurate labels for fine-grained components, such as lane-level traffic flow.

Other datasets fulfill more specialized roles. CARLA [Malik *et al.*, 2022] and comparable simulators generate synthetic data with exact ground truth, supporting analysis of rare or safety-critical scenarios despite an inherent sim-to-real discrepancy. Benchmarks constructed from time-series imagery permit evaluation of methods that encode traffic signals as images, yet they largely avoid the more demanding semantic and geometric alignment challenges posed by real-world camera streams.

Datasets explicitly intended to evaluate alignment quality in forecasting remain limited. Typically, benchmarks provide raw sensor measurements together with final task labels, but exclude validated intermediate correspondences, such as cross-camera vehicle identities or precise image-to-map alignments. This lack of intermediate ground truth complicates error analysis, making it difficult to pinpoint the causes of failure when fusion-based models are applied to real-world traffic environments.

5.2 Critical Evaluation of Evaluation Paradigms

Traffic forecasting models are typically examined via performance on the final forecasting task. For traffic speed and flow analysis, the most frequently adopted measures are Mean Absolute Error (MAE) and Root Mean Square Error (RMSE). They are expressed as $\text{MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$ and $\text{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2}$. In these formulations, $\hat{y}_i$ refers to the estimated traffic quantity, y_i corresponds to the observed ground truth, and N denotes the total number of samples. For vehicle trajectory forecasting, evaluation commonly relies on Average Displacement Error (ADE) and Final Displacement Error (FDE). They are defined as $\text{ADE} = \frac{1}{T}\sum_{t=1}^{T}\|\hat{\mathbf{p}}_t - \mathbf{p}_t\|$ and $\text{FDE} = \|\hat{\mathbf{p}}_T - \mathbf{p}_T\|$, where $\hat{\mathbf{p}}_t$ and $\mathbf{p}_t$ indicate forecasted and actual vehicle positions at time t, and T denotes the forecasting horizon. Although these indicators are essential for comparing the overall forecasting capability, they capture alignment fidelity only indirectly. Strong forecasting results can emerge from predominant reliance on the spatiotemporal modality $\mathcal{S}$, while the concrete role played by the visual information $\mathcal{V}$ in contextual reasoning remains ambiguous.

Placing primary emphasis on downstream accuracy therefore leaves a discernible evaluation gap. Accurately disentangling the effects of cross-modal fusion requires criteria that explicitly address alignment behavior. This requirement motivates complementary evaluation perspectives: at the feature level, through inspection of latent representation coherence; at the semantic level, through appraisal of traffic concept extraction and correspondence; and at the geometric level, through direct measurement of cross-camera association and spatial registration precision. The lack of benchmarks dedicated to these facets continues to restrict progress towards interpretable and dependable fusion designs.

Additionally, existing evaluations are predominantly conducted under near-ideal conditions. They seldom challenge models with realistic alignment disturbances, including sensor desynchronization, calibration drift, or changes in scene configuration. Within current traffic forecasting benchmarks, robustness to such factors is critically important yet remains insufficiently examined [Li *et al.*, 2025e].

6 Future Directions

The analysis has identified several research gaps. This section suggests potential paths for advancing spatial intelligence and building more capable traffic world models.

6.1 Advancing Spatial Intelligence for Robust Perception

Current perception systems often fail in novel or unexpected situations [Li *et al.*, 2025a]. Enhancing their robustness necessitates a tighter integration with traffic forecasting. Per-

Dataset	Visual Modality	Spatiotemporal Modality	Temporal Sync.	Primary Alignment Challenge
nuScenes [Caesar *et al.*, 2020]	Multi-camera, 360° views	3D object tracks, HD map elements	Yes	Large-scale geometric registration in complex intersections
Waymo Open [Sun *et al.*, 2020]	Multi-camera, LiDAR streams	Detailed vehicle trajectories, map data	Yes	Fine-grained spatial grounding for traffic participants
CityFlow [Tang *et al.*, 2019]	City-scale traffic camera network	Vehicle trajectories, ReID labels	Approximate	Cross-view association for network-wide traffic flow analysis
CARLA (Synthetic) [Malik *et al.*, 2022]	Rendered traffic camera views	Perfect vehicle state/logs	Perfect	Sim2Real domain gap for traffic scenarios
Time-Series Benchmarks	Generated images from traffic signals	Traffic flow/speed series	N/A	Representation conversion fidelity for traffic forecasting

Table 2: Core multimodal datasets for vision-spatiotemporal traffic forecasting research.

ception and forecasting form an interdependent cycle: reliable perceptual data is fundamental for accurate forecasting, while forecasting context can, in turn, guide perception, for instance by suggesting probable locations of occluded vehicles. Future architectures should therefore facilitate this joint reasoning within a unified representational space. Such a closed-loop design can support error correction: a perception error that leads to a flawed forecasting may be identified and subsequently rectified via the resulting inconsistency. Such deeply integrated perception-forecasting synergy is pivotal for developing the next generation of highly reliable traffic world models capable of adapting to urban dynamics.

6.2 Embodied and Interactive World Models

Embodied world models are poised to become a core component of next-generation systems. These models understand the world from the point of view of a specific agent, such as a single vehicle. They also consider that agent's possible actions. The world is not just something to observe but changes because of the agent's own decisions and how others react to them [Lin *et al.*, 2025b]. Future research needs frameworks to model this loop. Reinforcement learning or differentiable simulators are possible tools. The key difficulty is simulating how other agents will behave in response. The model should forecast how an action, such as changing lanes, influences the future behavior of nearby drivers. This moves forecasting from passive observation to interactive reasoning, requiring the simulation of action consequences rather than just forecasting a fixed future.

6.3 Towards Generalizable Traffic World Models

Traffic conditions differ greatly across cities and over time. This diversity requires world models with strong generalization ability [Vafa *et al.*, 2024]. A promising path adopts the foundation model paradigm, where a large model is pretrained on extensive and diverse traffic data. The goal is for it to learn common principles of traffic flow and visual scenes. Developing such a model faces significant hurdles. Researchers should compile varied datasets and devise self-supervised learning tasks. These tasks should enable the model to align visual and spatiotemporal data in a transferable manner for new environments. A successful general model can serve as a foundation for many downstream applications.

6.4 Trustworthy and Efficient World Models

For practical deployment, traffic world models should satisfy two critical requirements: trustworthiness and efficiency. Trustworthiness ensures that a model's forecasts are reliable and its decision-making process is transparent. This involves several key attributes: explainability to reveal the reasoning behind forecasting, uncertainty quantification to evaluate the confidence level of each forecast, and robustness to maintain performance when faced with unseen scenarios, such as unusual weather or traffic incidents [Li *et al.*, 2025b]. Efficiency is equally crucial for real-world applications. Models should be computationally lightweight and fast enough to run in real-time on embedded hardware within vehicles or roadside infrastructure. A central challenge arises because the pursuit of high forecasting accuracy often conflicts with these stringent efficiency needs.

7 Conclusion

Given the necessity of fusing visual and spatiotemporal data, this survey constructs a taxonomy for vision-spatiotemporal alignment covering core state, trajectory, event, and demand forecasting, and charts the path to a traffic world model by analyzing semantic alignment as a cognitive layer and spatial intelligence bridging the geometric gap. Critically reviewing datasets and evaluation paradigms highlights key bottlenecks. The proposed framework directly addresses the fusion challenge, offering a unified foundation for reliable, context-aware forecasting and guiding future research toward spatially-intelligent world models.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Nos. 62272284, 12201518, 62406044), the Science and Technology Cooperation and Exchange Special Projects of Shanxi (No. 202504041101007), the Central Government Guides Local Science and Technology Innovation (No. YDZJSX2024D015), the Sichuan Science and Technology Program (No. 2025ZNSFSC0806), and the Top Young Talents of Shanxi "Three Jin" Talents Program (No. SJYC2025334).

References

[Caesar *et al.*, 2020] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. Nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, pages 11621–11631, 2020.

[Chen *et al.*, 2025] Mouxiang Chen, Lefei Shen, Zhuo Li, Xiaoyun Joy Wang, Jianling Sun, and Chenghao Liu. Visionts: Visual masked autoencoders are free-lunch zero-shot time series forecasters. In *ICML*, 2025.

[Chen *et al.*, 2026] Jinwen Chen, Hao Miao, Chenxi Liu, Yan Zhao, and Kai Zheng. Visionst: Coordinating cross-modal traffic prediction with interactive geo-image encoding. In *WWW*, pages 7307–7317, 2026.

[Dai *et al.*, 2025] Ruiting Dai, Chenxi Li, Yandong Yan, Lisi Mo, Ke Qin, and Tao He. Unbiased missing-modality multimodal learning. In *ICCV*, pages 24507–24517, 2025.

[Ding *et al.*, 2025] Jingtao Ding, Yunke Zhang, Yu Shang, Yuheng Zhang, Zefang Zong, Jie Feng, Yuan Yuan, Hongyuan Su, Nian Li, Nicholas Sukiennik, et al. Understanding world or predicting future? A comprehensive survey of world models. *ACM Computing Surveys*, 58(3):57, 2025.

[Fang *et al.*, 2026] Yuchen Fang, Hao Miao, Yuxuan Liang, Liwei Deng, Yue Cui, Ximu Zeng, Yuyang Xia, Yan Zhao, Torben Bach Pedersen, Christian S Jensen, et al. Unraveling spatio-temporal foundation models via the pipeline lens: A comprehensive review. *TKDE*, 38(3):2040–2063, 2026.

[Ferreira *et al.*, 2025] Gabriel O Ferreira, Fabrizio Dabbene, Francesco Malandrino, and Chiara Ravazzi. Unraveling urban mobility: A domain knowledge-free trajectory classification using gramian angular fields. *TITS*, 2025.

[Gong *et al.*, 2022] Shenjian Gong, Shanshan Zhang, Jian Yang, Dengxin Dai, and Bernt Schiele. Bi-level alignment for cross-domain crowd counting. In *ICCV*, pages 7542–7550, 2022.

[He *et al.*, 2025] Zhu He, Mingwei Lin, Xin Luo, and Zeshui Xu. Structure-preserved self-attention for fusion image information in multiple color spaces. *TNNLS*, 36(7):13021–13035, 2025.

[Hua and Wang, 2024] Tongyan Hua and Lin Wang. Benchmarking implicit neural representation and geometric rendering in real-time rgb-d slam. In *CVPR*, pages 21346–21356, 2024.

[Huang *et al.*, 2024] Brandon Huang, Chancharik Mitra, Assaf Arbelle, Leonid Karlinsky, Trevor Darrell, and Roei Herzig. Multimodal task vectors enable many-shot multimodal in-context learning. In *NeurIPS*, pages 22124–22153, 2024.

[Jain *et al.*, 2025] Anubhooti Jain, Mayank Vatsa, and Richa Singh. Words over pixels? Rethinking vision in multimodal large language models. In *IJCAI*, pages 10481–10489, 2025.

[Jiang *et al.*, 2024] Xiaohui Jiang, Shuailin Li, Yingfei Liu, Shihao Wang, Fan Jia, Tiancai Wang, Lijin Han, and Xiangyu Zhang. Far3d: Expanding the horizon for surround-view 3d object detection. In *AAAI*, pages 2561–2569, 2024.

[Kannan and Min, 2025] Shyam Sundar Kannan and Byung-Cheol Min. Zeroscd: Zero-shot street scene change detection. In *ICRA*, pages 4665–4671, 2025.

[Khan and Gauch, 2025] Reeshad Khan and John Gauch. Tinybev: Cross-modal knowledge distillation for efficient multi-task bird's-eye-view perception and planning. In *ICCV*, pages 4577–4585, 2025.

[Kong *et al.*, 2024] Quan Kong, Yuki Kawana, Rajat Saini, Ashutosh Kumar, Jingjing Pan, Ta Gu, Yohei Ozao, Balazs Opra, Yoichi Sato, and Norimasa Kobori. Wts: A pedestrian-centric traffic video dataset for fine-grained spatial-temporal understanding. In *ECCV*, pages 1–18, 2024.

[Lee *et al.*, 2022] Wei-Yu Lee, Ljubomir Jovanov, and Wilfried Philips. Cross-modality attention and multimodal fusion transformer for pedestrian detection. In *ECCV*, pages 608–623, 2022.

[Li *et al.*, 2025a] Haoyang Li, Xin Wang, Ziwei Zhang, and Wenwu Zhu. Out-of-distribution generalization on graphs: A survey. *TPAMI*, 47(11):10490–10512, 2025.

[Li *et al.*, 2025b] Wentao Li, Bowen Yang, Witold Pedrycz, Chao Zhang, and Tao Zhan. Adaptive hyper-box granulation with justifiable granularity for feature selection. *TKDE*, 37(12):6847–6862, 2025.

[Li *et al.*, 2025c] Xiaoyu Li, Yitian Zhang, Guodong Long, Yupeng Hu, Wenpeng Lu, Meng Chen, Chengqi Zhang, and Yongshun Gong. Adaptive traffic forecasting on daily basis: A spatio-temporal context learning approach. *TKDE*, 37(8):4446–4459, 2025.

[Li *et al.*, 2025d] Yujie Li, Xiangkun Wang, Xin Yang, Marcello Bonsangue, Junbo Zhang, and Tianrui Li. Improving open-world continual learning under the constraints of scarce labeled data. In *SIGKDD*, pages 1647–1658, 2025.

[Li *et al.*, 2025e] Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S Jensen, et al. Tsfm-bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *SIGKDD*, pages 5595–5606, 2025.

[Lin *et al.*, 2025a] Jiawen Lin, Shiran Bian, Yihang Zhu, Wenbin Tan, Yachao Zhang, Yuan Xie, and Yanyun Qu. Seqvlm: Proposal-guided multi-view sequences reasoning via vlm for zero-shot 3d visual grounding. In *SIGMM*, pages 3094–3103, 2025.

[Lin *et al.*, 2025b] Mingwei Lin, Jiaqi Liu, Hong Chen, Xiuqin Xu, Xin Luo, and Zeshui Xu. A 3d convolution-incorporated dimension preserved decomposition model for traffic data prediction. *TITS*, 26(1):673–690, 2025.

[Liu *et al.*, 2025] Chenxi Liu, Shaowen Zhou, Qianxiong Xu, Hao Miao, Cheng Long, Ziyue Li, and Rui Zhao. Towards cross-modality modeling for time series analytics: A survey in the llm era. In *IJCAI*, pages 10564–10572, 2025.

[Lv *et al.*, 2025] Qinzhi Lv, Lijuan Liu, Ruotong Yang, and Yan Wang. Multimodal urban traffic flow prediction based on multi-scale time series imaging. *Pattern Recognition*, 164:111499, 2025.

[Malik *et al.*, 2022] Sumbal Malik, Manzoor Ahmed Khan, and Hesham El-Sayed. Carla: Car learning to act-an inside out. *Procedia Computer Science*, 198:742–749, 2022.

[Mirzaie and Rosenhahn, 2025] Mona Mirzaie and Bodo Rosenhahn. Interpretable decision-making for end-to-end autonomous driving. In *ICCV*, pages 794–804, 2025.

[Ni *et al.*, 2025] Jingchao Ni, Ziming Zhao, ChengAo Shen, Hanghang Tong, Dongjin Song, Wei Cheng, Dongsheng Luo, and Haifeng Chen. Harnessing vision models for time series analysis: A survey. In *IJCAI*, pages 10612–10620, 2025.

[Nie *et al.*, 2023] Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

[Palladin *et al.*, 2025] Edoardo Palladin, Samuel Brucker, Filippo Ghilotti, Praveen Narayanan, Mario Bijelic, and Felix Heide. Self-supervised sparse sensor fusion for long range perception. In *ICCV*, pages 27498–27509, 2025.

[Peddi *et al.*, 2024] Rohith Peddi, Saksham Singh, Saurabh, Parag Singla, and Vibhav Gogate. Towards scene graph anticipation. In *ECCV*, pages 159–175, 2024.

[Shaw *et al.*, 2025] Ankit Kumar Shaw, Chandan Kumar Sah, Xiaoli Lian, Arsalan Shahid Baig, Tuopu Wen, Kun Jiang, Mengmeng Yang, Diange Yang, and Li Zhang. Saferoute: Enhancing traffic scene understanding via a unified deep learning and multimodal llm. In *ICCV*, pages 4547–4556, 2025.

[Shrout *et al.*, 2025] Oren Shrout, Ori Nizan, Yizhak Ben-Shabat, and Ayellet Tal. Patchcontrast: Self-supervised pre-training for 3d object detection. In *CVPR*, pages 2456–2466, 2025.

[Sun *et al.*, 2020] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, pages 2446–2454, 2020.

[Sun *et al.*, 2024] Yubao Sun, Jihui Tang, Qingshan Liu, Zhendong Zhang, and Liang Huang. Highway visibility level prediction using geometric and visual features driven dual-branch fusion network. *TITS*, 25(8):9992–10004, 2024.

[Tang *et al.*, 2019] Zheng Tang, Milind Naphade, Ming-Yu Liu, Xiaodong Yang, Stan Birchfield, Shuo Wang, Ratnesh Kumar, David Anastasiu, and Jenq-Neng Hwang. Cityflow: A city-scale benchmark for multi-target multi-camera vehicle tracking and re-identification. In *CVPR*, pages 8797–8806, 2019.

[Tang *et al.*, 2024] Qi Tang, Yao Zhao, Meiqin Liu, and Chao Yao. Seeclear: Semantic distillation enhances pixel condensation for video super-resolution. In *NeurIPS*, pages 134902–134926, 2024.

[Vafa *et al.*, 2024] Keyon Vafa, Justin Y Chen, Ashesh Rambachan, Jon Kleinberg, and Sendhil Mullainathan. Evaluating the world model implicit in a generative model. In *NeurIPS*, pages 26941–26975, 2024.

[Wang *et al.*, 2024] Xiaoding Wang, Jianmin Liu, Hui Lin, Sahil Garg, and Mubarak Alrashoud. A multi-modal spatial-temporal model for accurate motion forecasting with visual fusion. *Information Fusion*, 102:102046, 2024.

[Wang *et al.*, 2025] Liying Wang, Xiaoli Zhang, Chuanmin Jia, and Siwei Ma. Mafs: Masked autoencoder for infrared-visible image fusion and semantic segmentation. *TIP*, 34:6490–6505, 2025.

[Wang *et al.*, 2026] Bingjie Wang, Chao Zhang, Wentao Li, and Deyu Li. Mf3: Multimodal federated learning with dual-path mamba-transformer for metro flow prediction. In *WWW*, pages 5451–5462, 2026.

[Wu *et al.*, 2023] Haixu Wu, Tengge Hu, Yong Liu, Hang Zhou, Jianmin Wang, and Mingsheng Long. Timesnet: Temporal 2d-variation modeling for general time series analysis. In *ICLR*, 2023.

[Wu *et al.*, 2025] Hsiu-Fu Wu, Ya-Ting Yang, Yung-Ter Chen, and I-Fan Chou. Trafficinternvl: Understanding traffic scenarios with vision-language models. In *ICCV*, pages 5229–5236, 2025.

[Yang *et al.*, 2021] Weixiang Yang, Qi Li, Wenxi Liu, Yuanlong Yu, Yuexin Ma, Shengfeng He, and Jia Pan. Projecting your view attentively: Monocular road scene layout estimation via cross-view transformation. In *ICCV*, pages 15536–15545, 2021.

[Yi *et al.*, 2024] Zhongchao Yi, Zhengyang Zhou, Qihe Huang, Yanjiang Chen, Liheng Yu, Xu Wang, and Yang Wang. Get rid of isolation: A continuous multi-task spatio-temporal learning framework. In *NeurIPS*, pages 136701–136726, 2024.

[Zeng *et al.*, 2025] Tong Zeng, Longfeng Wu, Liang Shi, Dawei Zhou, and Feng Guo. Are vision llms road-ready? A comprehensive benchmark for safety-critical driving video understanding. In *SIGKDD*, pages 5972–5983, 2025.

[Zhang *et al.*, 2023] Jiarui Zhang, Filip Ilievski, Kaixin Ma, Aravinda Kollaa, Jonathan Francis, and Alessandro Oltramari. A study of situational reasoning for traffic understanding. In *SIGKDD*, pages 3262–3272, 2023.

[Zhou *et al.*, 2024] Siwei Zhou, Youngha Chang, Nobuhiko Mukai, Hiroaki Santo, Fumio Okura, Yasuyuki Matsushita, and Shuang Zhao. Path-space differentiable rendering of implicit surfaces. In *SIGGRAPH*, page 19, 2024.