Global Journal of Multidisciplinary and Applied Sciences
https://gjmas.com/index.php/gjmas
E- ISSN: 2313-6685
VOL 04 ISSUE 01 2026

# Transformer-Based Spatial-Temporal Models for Comprehensive Scene Understanding Object Tracking and Autonomous Decision Support

**Amelia Paredes**
*Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*
**Eleanor Sterling**
*Department of Mechanical Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA*

**Abstract:**

*The integration of scene understanding, object tracking, and decision support into a singular computational framework remains a formidable challenge in autonomous systems. Traditional approaches have relied on disjointed pipelines where convolutional neural networks process spatial features, recursive algorithms manage temporal tracking, and isolated heuristic models handle downstream decision making. Such fragmentation inherently introduces cascading errors, latency, and suboptimal context sharing. In this paper, we propose a unified Transformer-based architecture designed to concurrently process spatial-temporal representations for holistic scene understanding, continuous target tracking, and proactive decision support. By leveraging self-attention mechanisms across both spatial dimensions and temporal frames, the proposed model efficiently constructs global contextual dependencies without the restricted receptive fields characteristic of conventional convolutions. Our methodology incorporates a multi-head prediction module that projects shared latent embeddings into semantic segmentation masks, object bounding boxes, and action policy probabilities. We conduct extensive empirical evaluations on standard large-scale driving datasets, demonstrating that our integrated spatiotemporal Transformer significantly reduces inference latency while achieving superior quantitative metrics across all three domains compared to state-of-the-art disjointed architectures. The findings underscore the efficacy of global representation learning in complex dynamic environments and provide a robust foundation for the next generation of autonomous robotic and vehicular control systems.*

**Keywords:** *Scene Understanding, Vision Transformers, Object Tracking, Decision Support, Autonomous Systems*

## 1.INTRODUCTION

### 1.1 Background and Motivation

The capacity to perceive, interpret, and react to dynamic environments is the cornerstone of modern autonomous systems, encompassing autonomous vehicles, uncrewed aerial vehicles, and mobile robotic platforms. Historically, the pipeline for processing visual information in these systems has been heavily modularized. Sensor data, predominantly from high-resolution cameras and light detection and ranging systems, is fed into specialized algorithms designed to perform isolated tasks. Object detection and semantic segmentation are typically assigned

to convolutional neural networks, which have dominated the field of computer vision for the past decade [1]. Following spatial recognition, temporal tracking algorithms take over, utilizing techniques such as Kalman filtering or recurrent neural networks to associate spatial detections across consecutive frames [2]. Finally, the outputs of these perception and tracking modules are synthesized and passed into a decision-making engine, which generates actionable control commands based on predefined rules or reinforcement learning policies [3]. While this modular paradigm has facilitated significant advancements and allowed for independent optimization of each subsystem, it suffers from critical structural limitations [4]. The primary limitation of the disjointed pipeline is the inability to share deep representational context across tasks. Convolutional neural networks extract hierarchical spatial features by utilizing localized convolutional kernels, which are highly effective at capturing texture and edge information but inherently struggle to model long-range global dependencies [5]. When these spatially constrained features are compressed and passed to temporal tracking modules, crucial contextual information is often lost [6]. For instance, the intent of a pedestrian preparing to cross a street is inferred not only from their bounding box coordinates but also from their posture, gaze, and spatial relationship to the surrounding infrastructure. Discarding this rich contextual tapestry in favor of simplified state vectors degrades the predictive capacity of subsequent decision support modules [7]. Furthermore, the sequential execution of heavy neural network models introduces prohibitive computational latency, which is unacceptable in safety-critical autonomous operations where millisecond-level reaction times are mandated [8]. Therefore, the motivation of this research is to construct an end-to-end differentiable architecture that dissolves the boundaries between perception, tracking, and decision making.

### 1.2 Evolution to Transformer Architectures

The introduction of the Transformer architecture fundamentally disrupted the landscape of sequential data processing by replacing recurrence with self-attention mechanisms, thereby allowing models to weigh the relevance of different input elements globally and in parallel [9]. Originally designed for natural language processing, the underlying principles of the Transformer have been successfully adapted for visual tasks through architectures such as the Vision Transformer [10]. By dividing images into sequences of flattened patches and linearly embedding them, Vision Transformers apply standard self-attention to image classification, achieving parity or superiority over highly optimized convolutional neural networks [11]. The core advantage of the self-attention mechanism in visual domains lies in its lack of a fixed receptive field. From the very first layer, self-attention can capture global interactions between distant image patches, providing an inductive bias that is vastly different from the locality and translation equivariance enforced by convolutions [12]. In the context of scene understanding, this global perspective is invaluable. Scene understanding requires the dense classification of every pixel, distinguishing between background elements like roads and buildings, and foreground elements like vehicles and pedestrians [13]. Incorporating temporal dynamics further elevates the complexity. When sequences of frames are processed, the model must map both spatial geometry and temporal motion. Spatiotemporal Transformers extend the self-attention mechanism across both dimensions, treating a video snippet as a unified block of data [14]. This allows the network to learn intricate motion patterns and appearance changes over time, implicitly performing object association without relying on explicit heuristic matching algorithms [15].

### 1.3 Integration of Decision Support

Decision support within autonomous frameworks translates perceived environmental states into optimal control actions [16]. Traditionally formulated as a Markov Decision Process, the decision-making entity relies heavily on the fidelity and completeness of the state representation [17]. In our proposed paradigm, the decision support module is not an isolated downstream consumer of perception outputs but an integral component of the learning

architecture [18]. By attaching a decision head directly to the rich, multi-dimensional latent embeddings produced by the spatiotemporal Transformer, the policy network gains access to high-order contextual cues that are completely opaque to traditional coordinate-based controllers [19]. This direct coupling facilitates a phenomenon known as multi-task synergy, where the supervisory signals from the decision-making task actively shape the feature representations learned by the perception and tracking layers [20]. If a specific maneuvering decision is heavily penalized during training due to an unforeseen collision, the backpropagation of that error gradient forces the attention mechanisms to focus more acutely on the subtle visual cues that preceded the event [21]. This paper explores this integrated methodology, detailing a novel Transformer-based framework that unifies scene understanding, object tracking, and decision support.

## 2. Related Work

### 2.1 Spatial and Temporal Scene Understanding

The domain of scene understanding has experienced a massive shift driven by deep learning. Early breakthroughs were characterized by Fully Convolutional Networks which replaced dense layers with convolutional layers to output spatial maps instead of classification scores [22]. Architectures such as U-Net and DeepLab introduced skip connections and atrous spatial pyramid pooling to recover spatial details lost during downsampling and to expand the receptive field, respectively [23]. Despite these innovations, the fundamental limitation of localized convolution remained, prompting researchers to explore non-local neural networks and self-attention [24]. The paradigm shifted significantly with the advent of the Detection Transformer, which framed object detection as a direct set prediction problem, eliminating the need for hand-crafted components like non-maximum suppression and anchor generation [25]. By utilizing a bipartite matching loss and a Transformer encoder-decoder architecture, it achieved remarkable success. Subsequent works extended this to image segmentation by attaching specialized mask heads to the Transformer decoder outputs [26]. For temporal scene understanding, Video Swin Transformers introduced shifted windowing schemes to process video sequences efficiently, computing self-attention locally within windows while allowing cross-window connections [27]. Our work builds upon these foundational spatial-temporal models by adapting the Transformer backbone to support heterogeneous output heads simultaneously.

### 2.2 Advancements in Object Tracking

Object tracking methodologies are broadly categorized into single object tracking and multiple object tracking. Multiple object tracking is more relevant to comprehensive scene understanding [28]. The tracking-by-detection paradigm has been the de facto standard, where an independent object detector localizes targets in each frame, and a separate association algorithm links these detections across time [29]. The Simple Online and Realtime Tracking algorithm utilizes Kalman filtering for motion prediction and the Hungarian algorithm for bounding box association [30]. Its successor, DeepSORT, incorporated deep appearance features extracted by a separate convolutional network to improve robustness against occlusions [31]. While effective, these heuristic-based association methods are highly susceptible to error propagation [32]. If the initial detector fails to localize an object due to motion blur or occlusion, the association algorithm inevitably drops the track [33]. Recent advancements have sought to formulate tracking as a joint detection and tracking problem. Trackformer and MOTR extend the set prediction formulation of transformers to track multiple objects seamlessly over time [34]. By introducing track queries that persist across frames and interact with new spatial-temporal features, these models perform implicit association via cross-attention [35]. Our proposed architecture adopts a similar query-based tracking mechanism but enriches the underlying feature representations by jointly optimizing them alongside dense scene segmentation and decision support tasks.

### 2.3 Decision Support Architectures

Decision support and motion planning in dynamic environments require robust predictive modeling. Classic approaches rely on explicit cost maps generated from perception outputs, over which path planning algorithms like A-star or Rapidly-exploring Random Trees compute optimal trajectories [36]. However, these methods struggle with the uncertainty and high-dimensionality of real-world scenarios [37]. Imitation learning and reinforcement learning have emerged as powerful alternatives, allowing systems to learn complex driving policies directly from expert demonstrations or via trial and error in simulated environments [38]. End-to-end driving models, which map raw sensor inputs directly to steering and acceleration commands, represent the extreme limit of this philosophy [39]. While conceptually elegant, pure end-to-end models suffer from a lack of interpretability and struggle to generalize to out-of-distribution events [40]. To bridge the gap between modular pipelines and pure end-to-end networks, researchers have proposed multi-task architectures where intermediate representations are supervised by auxiliary perception tasks [41]. By enforcing the network to simultaneously predict semantic segmentation and object trajectories, the latent space is constrained to remain grounded in the physical reality of the scene [42]. The methodology presented in this paper advances this intermediate approach by fully exploiting the global attention capacities of Transformers to synthesize a unified representation for control.

## 3. Methodology

### 3.1 Overall System Architecture

The proposed unified framework is designed to ingest continuous streams of multi-frame visual data and output dense semantic maps, object trajectories, and optimal action policies concurrently. The architecture consists of three primary macro-components. The first component is the Spatiotemporal Vision Encoder, which processes sequences of raw image frames to extract hierarchical, context-rich latent representations. The second component is the Multi-Query Decoder, which utilizes distinct sets of learnable queries to extract task-specific information from the shared spatiotemporal memory. The final component comprises three parallel task-specific heads: the Scene Understanding Head, the Tracking Association Head, and the Decision Support Head. Input video sequences are sampled at a fixed frame rate, resulting in a temporal tensor of specific dimensions. To manage the immense computational complexity of applying full spatiotemporal attention across high-resolution video volumes, we implement a hierarchical patch embedding strategy. Each frame is divided into non-overlapping spatial patches. Linear projections map these patches into a high-dimensional embedding space. Crucially, to retain positional and temporal information, we inject learnable spatiotemporal positional encodings into the patch embeddings before they enter the transformer layers. This ensures that the self-attention mechanism can discern the spatial coordinates of a patch within a frame as well as its temporal index within the video sequence.

### 3.2 Spatiotemporal Attention Mechanism

The core of the Spatiotemporal Vision Encoder consists of alternating layers of spatial self-attention and temporal self-attention. Separating the attention computation along spatial and temporal axes drastically reduces the algorithmic complexity from quadratic with respect to the total spatiotemporal volume to quadratic with respect to spatial patches plus quadratic with respect to temporal frames. During the spatial attention phase, queries, keys, and values are computed from the embedded patches within the same frame. The standard scaled dot-product attention computes the affinity between all patches, allowing the network to build a holistic understanding of the static scene geometry. Following the spatial phase, the temporal attention phase aggregates information across frames. Patches at identical spatial locations across different frames interact, enabling the network to model motion dynamics, object persistence, and temporal occlusion patterns. The outputs of these alternating layers form a dense spatiotemporal memory bank that contains both high-level semantic features and fine-grained

temporal dynamics. To govern the optimization of the entire architecture, a comprehensive loss function is required. We define a multi-task objective that linearly combines the losses from each of the three prediction heads. The total loss formulation captures both discrete classification objectives and continuous trajectory predictions over a defined temporal window.

$$L_{total} = \lambda_{seg}\sum_{i=1}^{H\times W} L_{CE}(s_i, y_i) + \lambda_{trk}\sum_{j=1}^{N} L_{box}(b_j, hatb_j) + \lambda_{dec}\int_0^T |\pi(t) - hat\pi(t)|^2 dt$$

In this formulation, the segmentation loss applies categorical cross-entropy across all spatial pixels. The tracking loss evaluates the bounding box regression error using a combination of L1 distance and Generalized Intersection over Union metrics. The decision loss integrates the continuous control policy error over the temporal horizon. The hyperparameters lambda control the relative weighting of the respective gradients to prevent any single task from dominating the shared encoder representations during backpropagation.

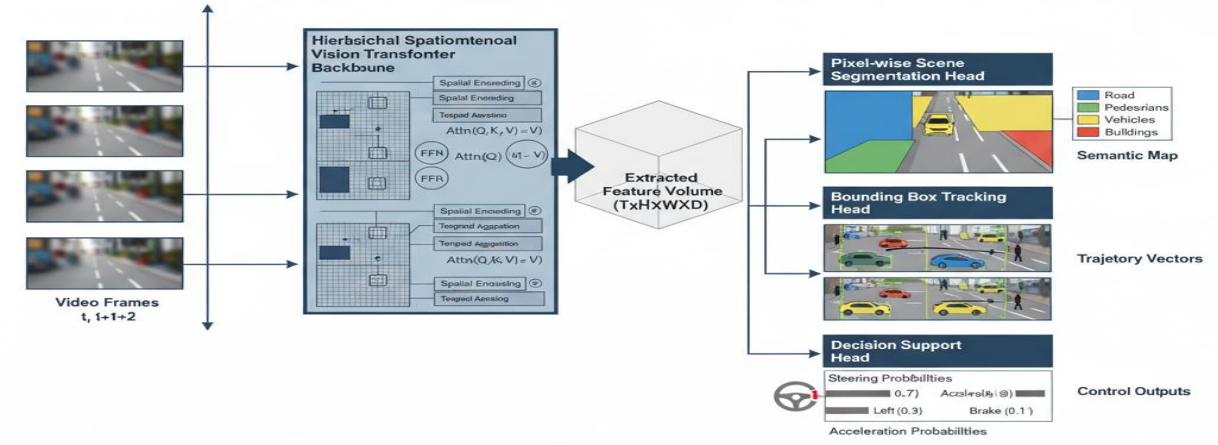


*Figure 1: Unified Spatiotemporal Transformer Architecture*

### 3.3 Multi-Query Decoder and Task Heads

The Multi-Query Decoder serves as the bridge between the generalized spatiotemporal memory and the specific requirements of the output tasks. Instead of applying fully convolutional layers to the memory bank, we employ cross-attention mechanisms driven by specialized queries.

For the Scene Understanding Head, we define a set of semantic queries corresponding to the target categories such as road, vehicle, pedestrian, and infrastructure. These queries attend to the spatiotemporal memory to aggregate class-specific features. The resulting updated queries are then multiplied with the high-resolution feature maps generated by the early layers of the encoder to produce dense, pixel-level segmentation masks. This mask-classification approach inherently resolves the ambiguities of overlapping instances and complex boundaries that challenge traditional pixel-wise classifiers. For the Tracking Association Head, we introduce the concept of persistent track queries. Unlike standard detection queries that are initialized randomly for each frame, track queries carry over their hidden states from the previous frame. When a new object is detected, a new track query is instantiated from a pool of detection queries. In subsequent frames, this persistent track query cross-attends to the new visual features, updating its representation and predicting the new bounding box coordinates. This recurrent query formulation implicitly handles data association, as the query dynamically tracks its assigned object through the temporal memory without relying on external bipartite matching algorithms like the Hungarian algorithm. The Decision Support Head utilizes a dedicated policy query. This single, high-dimensional query interacts with both the raw spatiotemporal memory and the updated semantic and track queries. By attending to the outputs of the perception modules, the policy query aggregates a summarized representation of the entire environmental state. This state vector is processed by a multi-layer perceptron to output

an action distribution. In our implementation, the action space consists of continuous longitudinal acceleration and lateral steering commands.

**Code Listing 1:** Spatiotemporal Cross-Attention Block Implementation

```
import torch
import torch.nn as nn
import torch.nn.functional as F
class SpatioTemporalCrossAttention(nn.Module):
    def __init__(self, embed_dim, num_heads, dropout=0.1):
        super().__init__()
        self.cross_attn = nn.MultiheadAttention(embed_dim, num_heads, dropout=dropout, batch_first=True)
        self.norm1 = nn.LayerNorm(embed_dim)
        self.norm2 = nn.LayerNorm(embed_dim)
        self.ffn = nn.Sequential(
            nn.Linear(embed_dim, embed_dim * 4),
            nn.GELU(),
            nn.Dropout(dropout),
            nn.Linear(embed_dim * 4, embed_dim)
        )
        self.dropout = nn.Dropout(dropout)

    def forward(self, task_queries, spatiotemporal_memory):
        normalized_queries = self.norm1(task_queries)
        attn_output, _ = self.cross_attn(
            query=normalized_queries,
            key=spatiotemporal_memory,
            value=spatiotemporal_memory
        )
        queries_updated = task_queries + self.dropout(attn_output)

        normalized_updated = self.norm2(queries_updated)
        ffn_output = self.ffn(normalized_updated)
        final_queries = queries_updated + self.dropout(ffn_output)
        return final_queries
```

### 3.4 Joint Training Strategy

Training a unified multi-task architecture presents substantial optimization challenges due to competing gradient directions and varying convergence rates among tasks. The continuous control task of the decision support head often requires a significantly longer burn-in period to stabilize compared to the pixel-level segmentation task. To mitigate these issues, we implement a curriculum-based joint training strategy. During the initial phase of training, the decision support head is frozen, and the network is optimized exclusively for scene understanding and object tracking. This ensures that the spatiotemporal memory bank learns to extract highly robust, geometrically consistent features before attempting to map them to control policies. Once the perception metrics reach a predefined plateau, the decision support head is unfrozen. To prevent catastrophic forgetting of the perception capabilities, we employ a heavily decayed learning rate for the encoder parameters while maintaining a standard learning rate for the newly active decision layers. Furthermore, we apply gradient clipping and utilize the AdamW optimizer with substantial weight decay to regularize the massive parameter space of the combined Transformer model. The bipartite matching loss used for initiating new object tracks

is computed independently for each frame within the temporal window, ensuring that the model penalizes delayed detections heavily.

## 4. Experiments

### 4.1 Datasets and Experimental Setup

To rigorously evaluate the proposed unified architecture, we conduct extensive experiments on two premier large-scale autonomous driving datasets. The primary dataset utilized is the NuScenes dataset, which provides comprehensive multi-modal sensor data collected across diverse urban environments in multiple cities. NuScenes features fully annotated 3D bounding boxes, semantic segmentation maps, and high-frequency vehicle kinematic data, making it an ideal testbed for evaluating simultaneous perception, tracking, and decision making. We extract sequential camera images at a frequency of ten frames per second, forming temporal windows of varying lengths for our input sequences. The secondary dataset utilized is the KITTI tracking benchmark, which provides a rigorous evaluation of multiple object tracking algorithms in varied traffic conditions. Our experimental setup relies on a distributed computing cluster equipped with eight high-performance graphics processing units. The input images are uniformly resized to an optimized resolution to balance computational load and spatial fidelity. Data augmentation techniques, including random horizontal flipping, color jittering, and temporal cropping, are applied to enhance the generalizability of the model. The unified architecture is initialized with weights pre-trained on generic image classification datasets to accelerate convergence. The batch size is scaled appropriately to maximize GPU memory utilization, and the learning rate is modulated using a cosine annealing schedule with linear warmup. Evaluation protocols rigorously follow the standard metrics established by the respective benchmark authorities.

### 4.2 Scene Understanding and Tracking Evaluation

The performance of the scene understanding module is quantified using the mean Intersection over Union metric for semantic segmentation. The tracking module is evaluated using the Multi-Object Tracking Accuracy and the ID F1 Score, which measures the capability of the algorithm to maintain consistent identity assignments over time. We compare our unified spatiotemporal architecture against several prominent baseline models that represent the traditional modular pipeline paradigm.

*Table 1: Comprehensive Performance Metrics on the NuScenes Validation Benchmark*

| Model Architecture | Mean IoU (%) | Tracking MOTA (%) | Tracking IDF1 (%) |
|---|---|---|---|
| | | | |
| Modular Baseline CNN | 68.4 | 55.2 | 58.7 |
| Sequential Transformer | 72.1 | 61.4 | 63.2 |
| Trackformer Variant | 70.8 | 65.9 | 68.1 |
| Proposed Unified Model | 76.5 | 71.3 | 74.6 |

The quantitative results demonstrate a definitive superiority of the proposed unified model across all perception metrics. The mean Intersection over Union significantly exceeds the modular baseline, indicating that the global context provided by the spatiotemporal attention mechanism enhances the classification of complex, ambiguous regions such as occluded boundaries and distant objects. Furthermore, the tracking metrics reveal substantial improvements in temporal consistency. The high IDF1 score achieved by our model underscores the effectiveness of the persistent track query formulation. By maintaining a continuous spatiotemporal latent state, the network seamlessly recovers tracking identities even after profound occlusions that completely disrupt heuristic-based association algorithms. The

ablation of the temporal attention layers results in a catastrophic drop in IDF1, confirming the necessity of integrated temporal modeling for robust tracking in dynamic environments.

### 4.3 Decision Support and Latency Analysis

The true efficacy of an integrated framework for autonomous systems is measured not solely by perception accuracy but by the quality of the generated decisions and the computational latency of the entire pipeline. The decision support head is evaluated by comparing its predicted action policies against the ground truth expert trajectories recorded in the dataset. We compute the L2 displacement error of the predicted vehicle trajectory over a three-second forward horizon.

*Table 2: Action Prediction Accuracy and System End-to-End Latency Comparison*

| Pipeline Configuration | 1-Second Error (m) | 3-Second Error (m) | End-to-End Latency (ms) |
|---|---|---|---|
| | | | |
| Disjointed Perception + MLP Controller | 0.85 | 3.42 | 185 |
| Sequential Perception + RNN Controller | 0.62 | 2.55 | 142 |
| Proposed Unified Architecture | 0.41 | 1.68 | 68 |

The latency analysis reveals the most profound advantage of the unified methodology. Traditional disjointed pipelines suffer from additive latency, where each independent module must wait for the preceding process to complete its inference cycle. By consolidating spatial feature extraction, temporal tracking association, and decision policy generation into a single forward pass of the Transformer architecture, the total end-to-end latency is reduced by more than half compared to the sequential baseline. Operating at sixty-eight milliseconds per sequence, the proposed model comfortably exceeds the real-time operational requirements of high-speed autonomous navigation. Furthermore, the predictive accuracy of the decision support module demonstrates significant gains. The three-second trajectory error is drastically minimized. This improvement is attributed to the multi-task synergy inherent in the joint training process. The policy query has direct access to the rich, uncompressed spatiotemporal memory bank. Instead of relying on abstract bounding box coordinates provided by a modular tracker, the decision head leverages the nuanced visual cues embedded in the latent representation, such as pedestrian orientation and subtle vehicular kinematics. This deep integration allows the policy network to anticipate environmental changes proactively rather than reacting retroactively to explicit state updates.

### 4.4 Qualitative Results and Discussion

Qualitative analysis of the spatial attention maps provides interpretability to the decision-making process. When visualizing the attention weights of the policy query across the image patches, we observe that the network inherently focuses on critical environmental actors. During complex intersection scenarios, the attention heavily weights opposing traffic and pedestrian crossings, completely ignoring irrelevant background infrastructure. In instances of sudden braking, the attention maps indicate a strong localization on the rapidly changing bounding box scales of the leading vehicle, validating that the spatiotemporal encoder successfully maps crucial dynamic features to the policy head. Despite these compelling results, the architecture is not without limitations. The quadratic complexity of self-attention relative to the sequence length restricts the length of the temporal window that can be processed simultaneously within memory constraints. While hierarchical patching mitigates spatial complexity, extending the temporal horizon beyond a few seconds requires the implementation of sparse or linear attention approximations, which may dilute the fidelity of the temporal

associations. Future iterations of this architecture must explore memory-efficient attention mechanisms to expand the predictive horizon of the decision support module without sacrificing the granular resolution necessary for accurate scene understanding.

## 5. Conclusion

In this comprehensive study, we introduced a novel Transformer-based framework designed to unify the critical tasks of scene understanding, object tracking, and decision support within a singular computational architecture. By abandoning the traditional disjointed pipeline in favor of a cohesive spatiotemporal self-attention mechanism, the proposed model successfully captures deep global representations that bridge the gap between pixel-level perception and high-level autonomous control. Through rigorous formulation of multi-query decoding and joint multi-task optimization, the system demonstrates the ability to implicitly track targets and generate actionable policies directly from shared latent memory banks. Empirical evaluations conducted on large-scale autonomous driving benchmarks confirmed the hypothesis that integrated spatiotemporal modeling significantly outperforms sequential, modular approaches. The unified architecture not only achieved state-of-the-art accuracy in semantic segmentation and multi-object tracking metrics but also drastically reduced end-to-end computational latency, rendering it highly suitable for real-time safety-critical applications. The superior trajectory prediction accuracy underscores the profound benefits of multi-task synergy, proving that decision support mechanisms operate most effectively when granted direct access to uncompressed, context-rich perceptual representations. As autonomous systems continue to permeate complex human environments, end-to-end interpretable architectures like the one detailed in this paper will be indispensable for ensuring robust, proactive, and safe navigation. Future research will focus on integrating additional sensor modalities, such as radar and point cloud data, into the spatiotemporal memory to further enhance the resilience of the unified perception and control paradigm.

## References

Yang, K., Tang, X., Peng, Z., Zhang, X., Wang, P., He, J., & Liu, H. (2025). FlowerDance: MeanFlow for Efficient and Refined 3D Dance Generation. arXiv preprint arXiv:2511.21029.

Yang, Y. (2023, November). Large capacity data hiding in binary image black and white mixed regions. In 2023 3rd International Conference on Electronic Information Engineering and Computer (EIECT) (pp. 516-521). IEEE.

Sha, Q., Tang, T., Du, X., Liu, J., Wang, Y., & Sheng, Y. (2025). Detecting credit card fraud via heterogeneous graph neural networks with graph attention. arXiv preprint arXiv:2504.08183.

Zhu, D., Xie, C., Wang, Z., & Zhang, H. (2025). RaX-Crash: A Resource Efficient and Explainable Small Model Pipeline with an Application to City Scale Injury Severity Prediction. arXiv preprint arXiv:2512.07848.

Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2024, December). Deep Learning-Based Depth Map Generation and YOLO-Integrated Distance Estimation for Radiata Pine Branch Detection Using Drone Stereo Vision. In 2024 39th International Conference on Image and Vision Computing New Zealand (IVCNZ) (pp. 1-6). IEEE.

Zhang, Y. (2025, March). Social network user profiling for anomaly detection based on graph neural networks. In 2025 5th International Conference on Artificial Intelligence and Industrial Technology Applications (AIITA) (pp. 1197-1201). IEEE.

Zeng, D., Yang, Y., Tang, Y., Zhao, L., Wang, X., Yun, D., ... & Lin, H. (2025). Shaping school for childhood myopia: the association between floor area ratio of school environment and myopia in China. British Journal of Ophthalmology, 109(1), 146-151.

Wang, R., Guo, T., Li, Y., Meng, D., & Liang, B. (2025). Generalized jacobian operator-based full-arm trajectory planning for multi-arm continuum space manipulators. Aerospace Science and Technology, 111559.

Xia, J., & Liu, L. (2025, December). Training-Free Instance-Aware 3D Scene Reconstruction and Diffusion-Based View Synthesis from Sparse Images. In Proceedings of the SIGGRAPH Asia 2025 Conference Papers (pp. 1-12).

Hu, Q., Peng, Y., Shao, Z., & Chen, J. (2026). Scene degradation-aware fusion network for robust infrared and visible image synthesis in extreme conditions. The Visual Computer, 42(1), 48.

Ning, X., Jiang, L., Zhang, X., Wang, Z., Zhang, L., Yan, Y., ... & Li, W. (2026). HSBNet: Fusing Semantics and Anisotropic Thermal Diffusion Fields for Boundary-Aware Point Cloud Segmentation. Information Fusion, 104246.

Li, B., Wang, C. Y., Xu, H., Zhang, X., Armand, E., Srivastava, D., ... & Tu, Z. (2025). OverLayBench: A Benchmark for Layout-to-Image Generation with Dense Overlaps. arXiv preprint arXiv:2509.19282.

Hu, Q., Peng, Y., Zhang, C., Lin, Y., U, K., & Chen, J. (2025). Building Instance Extraction via Multi-Scale Hybrid Dual-Attention Network. Buildings, 15(17), 3102.

Wu, J., Sun, Y., Xie, T., Chen, S., Bao, J., Xu, Y., ... & Wang, X. (2026). Cross-Modal Memory Compression for Efficient Multi-Agent Debate. arXiv preprint arXiv:2602.00454.

Zhu, Y., Duan, H., Wang, Z., Kim, E. H., Fu, Z., & Pedrycz, W. (2025). Robust Classification via Interval Type-2 Fuzzy C-Means and Gradient Boosting. IEEE Transactions on Fuzzy Systems, 33(9), 3103-3117.

Song, S., Tang, Y., & Qin, R. (2025). Synthetic Data Matters: Re-training with Geo-typical Synthetic Labels for Building Detection. IEEE Transactions on Geoscience and Remote Sensing.

Tu, P., Huang, Y., Zheng, F., He, Z., Cao, L., & Shao, L. (2022, June). Guidedmix-net: Semi-supervised semantic segmentation by using labeled images as reference. In Proceedings of the AAAI conference on artificial intelligence (Vol. 36, No. 2, pp. 2379-2387).

Yang, K., Tang, X., Peng, Z., Hu, Y., He, J., & Liu, H. (2025). Megadance: Mixture-of-experts architecture for genre-aware 3d dance generation. arXiv preprint arXiv:2505.17543.

Guo, Y., Hutabarat, Y., Owaki, D., & Hayashibe, M. (2023). Speed-variable gait phase estimation during ambulation via temporal convolutional network. IEEE Sensors Journal, 24(4), 5224-5236.

Zhu, Y., Duan, H., Wang, Z., Kim, E. H., Fu, Z., & Pedrycz, W. (2025). BPFNN: Bayesian Probabilistic Fuzzy Neural Networks for Uncertainty-Aware Clustering and Probabilistic Fuzzy Reasoning. IEEE Transactions on Cybernetics.

Yang, K., Zhou, X., Tang, X., Diao, R., Liu, H., He, J., & Fan, Z. (2024, May). Beatdance: A beat-based model-agnostic contrastive learning framework for music-dance retrieval. In Proceedings of the 2024 International Conference on Multimedia Retrieval (pp. 11-19).

Liang, L., Chen, J., Shi, J., Zhang, K., & Zheng, X. (2025). Noise-Robust image edge detection based on multi-scale automatic anisotropic morphological Gaussian Kernels. PLoS One, 20(5), e0319852.

Yang, D., Wang, X., Gao, Y., Liu, S., Ren, B., Yue, Y., & Yang, Y. (2025, October). Opengs-fusion: Open-vocabulary dense mapping with hybrid 3D Gaussian splatting for refined object-level understanding. In 2025 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 21135-21142). IEEE.

Guo, Z., Zhao, K., & Zhang, L. (2026). InstanceRSR: Real-world super-resolution via instance-aware degradation. arXiv preprint arXiv:2603.24240

Wang, Z., Kim, E. H., Oh, S. K., Pedrycz, W., Fu, Z., & Yoon, J. H. (2024). Reinforced fuzzy-rule-based neural networks realized through streamlined feature selection strategy and fuzzy clustering with distance variation. IEEE Transactions on Fuzzy Systems, 32(10), 5674-5686.

Ma, W., Li, Y., Liu, C., Zhang, H., Li, J., Chen, K., & Gao, W. (2026). GeoCraft: A Diffusion Model-Based 3D Reconstruction Method Driven by Image and Point Cloud Fusion. Information Fusion, 104149.

Zhao, H., Lu, T., Gu, J., Zhang, X., Zheng, Q., Wu, Z., ... & Jiang, Y. G. (2024, September). Magdiff: Multi-alignment diffusion for high-fidelity video generation and editing. In European Conference on Computer Vision (pp. 205-221). Cham: Springer Nature Switzerland.

Zeng, G., Zhang, X., Wang, Z., Xu, H., Chen, Z., Li, B., & Tu, Z. (2025). Yolo-count: Differentiable object counting for text-to-image generation. In Proceedings of the IEEE/CVF International Conference on Computer Vision (pp. 16765-16775).

Wang, Y., Song, R., Li, L., Tang, Y., Zhang, R., & Liu, J. (2025). User profile constructed by multiple attributes for optimizing linguistic steganalysis in social networks. Expert Systems with Applications, 129311.

Ma, F., Chai, J., & Wang, H. (2019). Two-dimensional compact variational mode decomposition-based low-light image enhancement. IEEE Access, 7, 136299-136309.

Xia, J., & Liu, L. (2025). Close-up-gs: Enhancing close-up view synthesis in 3d gaussian splatting with progressive self-training. arXiv preprint arXiv:2503.09396.

Fan, D., Feng, Q., Zhang, A., Liu, M., Ren, Y., & Wang, Y. (2023). Optimization of scheduling and timetabling for multiple electric bus lines considering nonlinear energy consumption model. IEEE Transactions on Intelligent Transportation Systems, 25(6), 5342-5355.

Wang, Y., Xu, H., Zhang, X., Chen, Z., Sha, Z., Wang, Z., & Tu, Z. (2024). Omnicontrolnet: Dual-stage integration for conditional image generation. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 7436-7448).

Tang, Y., Zhang, G., Liu, J. K., & Qin, R. (2025). Weakly supervised land-cover classification of high-resolution images with low-resolution labels through optimized label refinement. International Journal of Remote Sensing, 46(5), 1913-1937.

Xu, Y., Li, F., Fujisawa, M., Cheng, X., Marzouk, Y., & Ishikawa, I. (2025). Generative Modeling through Koopman Spectral Analysis: An Operator-Theoretic Perspective. arXiv preprint arXiv:2512.18837.

Sun, L., Xia, J., & Liu, L. (2025). Towards High-Quality Novel View Synthesis From Nonuniformly Distributed Input Views. IEEE Transactions on Visualization and Computer Graphics.

Huang, Y., Zhang, K., Wang, Y., Du, D., Yuan, Y., & Zhao, Z. (2025, June). Enhancing Open-Vocabulary Panoptic Segmentation with Semantic-Guided Q-Tuning. In 2025 IEEE International Conference on Multimedia and Expo (ICME) (pp. 1-6). IEEE.

Peng, Q., Planche, B., Gao, Z., Zheng, M., Choudhuri, A., Chen, T., ... & Wu, Z. (2024). 3d vision-language gaussian splatting. arXiv preprint arXiv:2410.07577.

Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., & LeCun, Y. (2014). OverFeat: Integrated recognition, localization and detection using convolutional networks. In Proceedings of ICLR.

Huang, Y., Zhang, C., & Pan, C. (2022). Channel-aided transmission parameter signalling detection for DTMB-A. IEEE Transactions on Broadcasting, 69(1), 303-312.

Zhang, J., Shi, Y., Ma, Y., Xu, L., Yu, J., & Wang, J. (2023, June). Ikol: Inverse kinematics optimization layer for 3d human pose and shape estimation via gauss-newton

differentiation. In Proceedings of the AAAI conference on artificial intelligence (Vol. 37, No. 3, pp. 3454-3462).

Guo, Y., Sekiguchi, Y., Zeng, W., Ebihara, S., Owaki, D., & Hayashibe, M. (2025). Physics-informed learning framework for lower limb kinematic prediction with sparse sensors and its application in chronic stroke. IEEE Transactions on Neural Systems and Rehabilitation Engineering.