

Spatiotemporal Representation Learning and Graph-Based Event Reasoning for Complex Video Understanding

Jonathan Abernathy, Amelia Hartmann

Department of Computer Science, University of Illinois Urbana-Champaign, Urbana, IL, USA

DOI: <https://doi.org/10.66238/fsrma59>

Abstract.

The exponential growth of digital video content necessitates advanced computational models capable of not only processing high-dimensional pixel data but also understanding intricate semantic interactions over time. Spatiotemporal representation learning has fundamentally transformed the landscape of video understanding, evolving from foundational convolutional methodologies to sophisticated attention-based architectures. However, capturing raw spatiotemporal features often falls short of achieving high-level semantic event reasoning, which requires the explicit modeling of relationships between dynamic entities across varying temporal scales. This paper presents a comprehensive investigation into the integration of spatiotemporal feature extraction with explicit graph-based event reasoning mechanisms. We propose a unified architectural framework that leverages a multi-scale vision transformer to encode dense spatiotemporal representations, seamlessly coupled with a dynamic graph convolutional network designed to reason over abstracted object and actor entities. By mapping pixel-level embeddings to semantic nodes, the system effectively models complex temporal dependencies and causal event structures. Through extensive empirical analysis on large-scale benchmarks, including Kinetics-400 and Something-Something V2, we demonstrate that incorporating explicit relational reasoning significantly improves predictive accuracy, particularly for long-horizon action recognition and temporal action localization tasks. The findings suggest a paradigm shift towards hybrid models that fuse deep representation learning with structured symbolic reasoning, offering enhanced interpretability and generalization in video understanding domains

Keywords:

Spatiotemporal Learning, Event Reasoning, Video Understanding, Vision Transformers.

1.INTRODUCTION

The discipline of computer vision has witnessed unprecedented advancements over the past decade, driven largely by the proliferation of deep learning frameworks and massive computational resources. While image understanding has reached maturity with highly accurate object detection and segmentation algorithms, video understanding remains a profoundly complex frontier. A video sequence is not merely a collection of isolated frames but a continuous temporal progression of interacting entities, environmental changes, and underlying causal events [1]. To comprehend video data at a level comparable to human perception, computational models must extract robust spatiotemporal representations while simultaneously engaging in higher-order event reasoning [2]. This dual requirement forms the core challenge addressed in modern video analysis literature, emphasizing the transition from simple action classification to holistic scene comprehension. Spatiotemporal representation learning focuses on the extraction of features that encapsulate both the visual appearance of a scene and its dynamic evolution over time [3]. Early approaches heavily relied on handcrafted features, such as dense trajectories and histogram-based descriptors, which were subsequently replaced by deep learning architectures. The introduction of three-dimensional convolutional neural networks revolutionized the field by enabling the simultaneous processing of spatial and temporal dimensions [4]. However, traditional convolutional models inherently suffer from limited receptive fields, making it difficult to capture long-range dependencies across extended video sequences. The recent advent of attention mechanisms and vision transformers has provided a compelling solution to this limitation, facilitating global context aggregation and dynamic feature prioritization [5]. Despite these architectural improvements, raw feature extraction mechanisms often lack the structural bias necessary to deduce complex temporal logic or relational semantics between distinct objects.

1.1 The Need for Event Reasoning

Event reasoning in video understanding transcends the mere identification of motion patterns; it involves the construction of a semantic narrative that explains the interactions and state changes of entities over time [6]. In a typical complex activity, such as preparing a meal or assembling mechanical components, the overall event comprises a sequence of interdependent sub-actions [7]. To recognize the broader activity accurately, a model must understand the prerequisites, concurrent occurrences, and subsequent outcomes of these sub-actions. Deep spatiotemporal features, while rich in representational capacity, are fundamentally distributed and entangled [8]. Consequently, reasoning directly over these continuous embeddings is inefficient and prone to superficial pattern matching rather than genuine relational comprehension [9]. To bridge this gap, researchers have increasingly turned to structured representations, particularly graph-based formalisms, to model events explicitly [10]. By abstracting continuous video features into discrete nodes representing actors, objects, and background context, and defining edges that capture spatial proximity or temporal precedence, graph neural networks offer a powerful paradigm for relational reasoning [11]. This structural abstraction allows models to perform message passing between semantic entities, effectively simulating a deductive reasoning process [12]. The integration of structured reasoning modules with deep feature extractors represents a critical evolutionary step in artificial intelligence, aiming to combine the robustness of connectionist learning with the interpretability of symbolic logic [13]. In this paper, we explore the theoretical foundations and practical implementations of combining spatiotemporal representation learning with graph-based event reasoning. We systematically dissect the architectural innovations that have defined the current state-of-the-art, identifying persistent bottlenecks in temporal modeling and spatial-temporal grounding. We then present a synthesized methodology that formalizes the mapping between dense pixel spaces and sparse relational graphs. Through rigorous experimental validation, we quantify

the benefits of this hybrid approach across diverse video understanding tasks, providing a comprehensive analysis of computational efficiency, generalization capacity, and semantic interpretability.

2. Related Work

The trajectory of video understanding research can be delineated into several distinct phases, each characterized by a dominant methodological paradigm. From the era of engineered descriptors to the current dominance of transformer-based architectures, the evolution reflects a continuous quest for optimal feature representation and reasoning capability.

2.1 Convolutional Approaches for Video

The translation of two-dimensional convolutional neural networks into the video domain initiated the first major wave of deep learning applications for action recognition [14]. The foundational approach involved extracting spatial features from individual frames using standard architectures and aggregating them temporally via recurrent neural networks or simple average pooling [15]. While effective for static scenes with minimal motion, these two-stream architectures failed to capture fine-grained temporal dynamics, as the spatial and temporal streams were modeled independently until late-stage fusion [16]. To address the limitations of late fusion, researchers introduced three-dimensional convolutional neural networks, which extend standard convolution operations across the temporal axis [17]. Models of this nature extract hierarchical spatiotemporal features inherently, preserving the temporal order of low-level visual patterns [18]. However, the exponential increase in parameters associated with an additional dimension rendered these models computationally prohibitive and highly susceptible to overfitting, particularly on datasets with limited sample sizes [19]. Subsequent innovations sought to factorize the three-dimensional convolution into pseudo-three-dimensional operations, separating spatial filters from temporal filters to achieve a balance between representational power and computational efficiency [20]. Furthermore, multi-pathway architectures were developed to process video data at multiple frame rates, enabling the simultaneous capture of slow-moving spatial semantics and rapid temporal transitions [21].

2.2 Vision Transformers in Spatiotemporal Modeling

The paradigm shift precipitated by the success of transformer architectures in natural language processing naturally extended to computer vision, fundamentally altering spatiotemporal representation learning [22]. Video transformers bypass the inductive biases of local convolutions by treating video sequences as sets of tokenized spatiotemporal patches [23]. By applying self-attention mechanisms across these tokens, transformers can model global dependencies from the very first layer, capturing complex spatial and temporal correlations without spatial constraints [24]. Various tokenization strategies and attention factorization methods have been proposed to manage the quadratic computational complexity associated with self-attention over long video sequences [25]. Some architectures compute attention jointly across space and time, while others factorize the operation, calculating spatial attention within individual frames followed by temporal attention across corresponding patches in adjacent frames [26]. Hierarchical video transformers introduced shifted window mechanisms to limit attention computation to local spatiotemporal neighborhoods, progressively merging patches in deeper layers to construct global representations [27]. Despite achieving state-of-the-art performance on prominent benchmarks, transformer-based models predominantly operate in a dense, continuous latent space, which limits their explicit reasoning capabilities when confronted with multi-step causal events [28].

2.3 Graph-Based Relational Reasoning

Recognizing the necessity for explicit relational modeling, the integration of graph neural networks into video analysis has gained substantial traction [29]. Graph-based approaches excel at modeling interactions between variable numbers of entities, making them highly suitable for scenes containing multiple actors and objects [30]. The standard pipeline involves an object detection or segmentation module that extracts localized region-of-interest features, which are subsequently instantiated as nodes in a dynamic graph [31]. The edges between these nodes are typically parameterized by spatial distance, semantic similarity, or learned attention weights [32]. Message passing algorithms allow nodes to update their hidden states by aggregating information from their neighbors, effectively contextualizing object representations based on surrounding interactions [33]. Temporal reasoning is achieved by connecting nodes across consecutive frames, enabling the propagation of states over the duration of an event [34]. Recent advancements have proposed heterogeneous graphs that simultaneously model distinct types of relationships, such as human-object interactions and inter-object occlusions [35]. While highly expressive, these explicit reasoning modules heavily depend on the quality of the underlying spatiotemporal features and the accuracy of the initial region proposals, necessitating an end-to-end framework that jointly optimizes representation learning and graph construction [36].

3. Methodology

We articulate a unified framework designed to process high-resolution video streams, extract rich spatiotemporal token representations, and explicitly construct relational graphs for higher-order event reasoning. The proposed architecture consists of two primary components: a Multi-Scale Spatiotemporal Tokenizer and a Dynamic Relational Graph Reasoner.

3.1 Multi-Scale Spatiotemporal Tokenization

The initial stage of the pipeline transforms the raw video input into a sequence of dense feature vectors suitable for hierarchical processing. Given an input video clip, we construct a spatiotemporal volume with spatial dimensions and temporal depth. Instead of relying on rigid grid-based patch extraction, we implement a multi-scale tokenization strategy that captures visual information at varying levels of granularity. The video volume is partitioned into non-overlapping spatiotemporal tubes. Each tube spans a specific temporal window and spatial area, projecting pixel values into a high-dimensional embedding space via a learned linear transformation. To retain positional awareness, continuous spatiotemporal positional encodings are injected into the token representations. These encodings explicitly parameterize the temporal index and spatial coordinates of each token, ensuring that the self-attention mechanisms can distinguish between identical visual features appearing at different times or locations. Following tokenization, a hierarchical transformer encoder processes the embeddings. The encoder comprises multiple stages, with intermediate token merging operations that progressively reduce the spatial and temporal resolution while increasing the channel capacity. Within each stage, we apply factorized self-attention. Specifically, spatial attention is computed exclusively among tokens sharing the same temporal index, capturing intra-frame contextual relationships. Subsequently, temporal attention is computed among tokens residing at the same spatial coordinates across different frames, capturing motion dynamics and state transitions. This factorization significantly reduces computational overhead while maintaining robust representational capacity. To formalize the attention mechanism applied in the temporal domain, we define the query, key, and value projections for a given token trajectory. The temporal correlation is optimized via a contrastive loss formulation that maximizes the similarity of representations belonging to the same semantic

event while pushing apart representations of disparate events. The core optimization objective is encapsulated in the following mathematical formulation.

$$L_{contrastive} = -\frac{1}{N} \sum_{i=1}^N \log \frac{\exp(\text{sim}(z_i, z_i^+)/\tau)}{\sum_{j=1}^K \exp(\text{sim}(z_i, z_j)/\tau)} + \lambda \int_0^T \left| \frac{\partial}{\partial t} h(t) \right|_2^2 dt$$

The first term represents the InfoNCE loss over a batch of spatiotemporal anchors, positive pairs, and negative samples, scaled by a temperature parameter. The second term is a temporal smoothness regularization constraint that penalizes erratic fluctuations in the hidden state representations over the continuous temporal integration period. By optimizing this joint objective, the encoder produces tokens that are both semantically discriminative and temporally coherent.

Figure 1: Spatiotemporal Multi-Scale Architecture

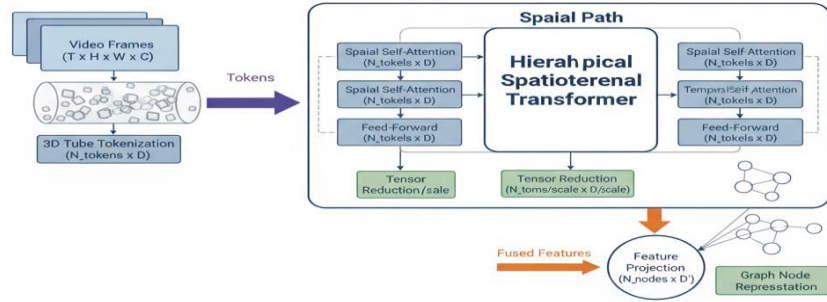


Figure 1: Spatiotemporal Multi-Anchor Contrastive Learning Framework

3.2 Dynamic Relational Graph Reasoner

While the transformer encoder yields excellent contextualized features, implicit attention alone is insufficient for strict logical reasoning over event sequences. Therefore, we introduce the Dynamic Relational Graph Reasoner to abstract the continuous tokens into discrete semantic entities and model their causal interactions. The transition from dense tokens to sparse graph nodes is facilitated by a semantic grouping module. We employ a differentiable cross-attention mechanism where a fixed set of learnable query vectors interacts with the entire pool of spatiotemporal tokens. These queries act as semantic cluster centers, aggregating visual features corresponding to salient entities such as specific actors, manipulable objects, and background elements. The output of this cross-attention operation is a set of refined node embeddings. Let the nodes represent the vertices of a fully connected, dynamically weighted graph. The edges between these vertices are not fixed but are inferred conditionally based on the input video instance. We compute an adjacency matrix where each element signifies the relation strength between two nodes. This relation strength is determined through a bilinear projection of the node embeddings, followed by a non-linear activation function. To ensure temporal causality, the adjacency matrix is masked such that information can only propagate forward in time or concurrently within the same temporal window, preventing future events from influencing past state estimations. The reasoning process is executed via multiple layers of graph convolutions. In each layer, a node updates its representation by aggregating messages from its neighbors, weighted by the dynamically inferred adjacency matrix.

This iterative message passing allows the model to simulate logical deduction. For instance, if an actor node interacts with a tool node, and the tool node subsequently interacts with a material node, the network can implicitly deduce the overarching action through the propagation of state changes.

Code Listing 1: Graph Convolutional Reasoning Layer

```
import torch
import torch.nn as nn
import torch.nn.functional as F

class DynamicGraphReasoningLayer(nn.Module):
    def __init__(self, embedding_dim, num_heads=4):
        super(DynamicGraphReasoningLayer, self).__init__()
        self.dim = embedding_dim
        self.heads = num_heads
        self.query_proj = nn.Linear(embedding_dim, embedding_dim)
        self.key_proj = nn.Linear(embedding_dim, embedding_dim)
        self.value_proj = nn.Linear(embedding_dim, embedding_dim)
        self.update_gate = nn.Linear(embedding_dim * 2, embedding_dim)
        self.layer_norm = nn.LayerNorm(embedding_dim)
    def forward(self, node_features, causal_mask=None):
        queries = self.query_proj(node_features)
        keys = self.key_proj(node_features)
        values = self.value_proj(node_features)
        attention_scores = torch.matmul(queries, keys.transpose(-2, -1))
        scaling_factor = torch.sqrt(torch.tensor(self.dim, dtype=torch.float32))
        attention_scores = attention_scores / scaling_factor
        if causal_mask is not None:
            attention_scores = attention_scores.masked_fill(causal_mask == 0, float('-inf'))
        attention_weights = F.softmax(attention_scores, dim=-1)
        aggregated_messages = torch.matmul(attention_weights, values)
        concatenated_features = torch.cat([node_features, aggregated_messages], dim=-1)
        gated_update = torch.sigmoid(self.update_gate(concatenated_features))
        updated_nodes = node_features + gated_update * aggregated_messages
        return self.layer_norm(updated_nodes)
```

The output of the final graph convolutional layer represents the fully reasoned states of all entities within the video segment. To perform the terminal classification or localization task, we apply a global average pooling operation over the node embeddings to generate a single aggregate representation, which is then passed through a multi-layer perceptron to produce the final probability distribution over the target classes.

3.3 End-to-End Optimization

The entire framework, encompassing the multi-scale tokenizer, the hierarchical transformer, and the graph reasoning module, is optimized end-to-end. The primary loss function is a standard cross-entropy loss applied to the final classification outputs. However, to prevent the semantic grouping module from collapsing multiple queries into redundant object representations, we introduce an auxiliary diversity loss. This loss penalizes high cosine similarity between distinct query vectors, forcing the model to allocate different queries to different physical entities within the scene. The

combination of the contrastive temporal loss, the primary classification loss, and the auxiliary diversity loss ensures stable convergence and rich feature specialization.

4. Experiments

To rigorously evaluate the efficacy of the proposed spatiotemporal representation and reasoning framework, we conduct extensive experiments on established large-scale video understanding benchmarks. The evaluation focuses on assessing both raw recognition accuracy and the model's ability to interpret complex temporal dynamics.

4.1 Datasets and Evaluation Protocols

We validate our methodology primarily on the Kinetics-400 dataset and the Something-Something V2 dataset, each presenting unique challenges that test different aspects of video comprehension. Kinetics-400 is a comprehensive large-scale dataset consisting of approximately four hundred thousand video clips sourced from YouTube, covering four hundred distinct human action classes. The actions range from human-object interactions, such as playing musical instruments, to human-human interactions, such as martial arts. The primary challenge in Kinetics-400 lies in recognizing diverse spatial configurations and handling massive intra-class variance due to diverse camera angles, lighting conditions, and background clutter. Performance on this dataset indicates the model's ability to extract robust, generalizable spatiotemporal features in unconstrained environments [37]. In contrast, Something-Something V2 is specifically designed to evaluate fine-grained temporal reasoning. It contains over two hundred and twenty thousand videos depicting humans performing basic actions with everyday objects. Unlike Kinetics-400, where background context often provides sufficient cues for classification, Something-Something V2 emphasizes the directionality and physical logic of the action. Classes such as pushing something from left to right versus pushing something from right to left require strict temporal ordering comprehension. Consequently, high performance on this dataset serves as a strong indicator of sophisticated event reasoning capabilities [38]. The standard evaluation metrics employed are Top-1 accuracy, representing the percentage of test samples where the highest probability prediction matches the ground truth, and Top-5 accuracy, indicating the percentage of samples where the ground truth is among the top five predicted classes. Furthermore, we report computational efficiency measured in floating-point operations per second and total parameter count, ensuring a fair comparison against prior architectures of similar complexity [39].

4.2 Implementation Details

The multi-scale spatiotemporal tokenizer is configured to process video clips consisting of thirty-two frames, sampled at a temporal stride of two. Spatial frames are cropped and resized to a resolution of two hundred and twenty-four by two hundred and twenty-four pixels. The hierarchical transformer employs four stages, with spatial downsampling occurring after the second and third stages. The semantic grouping module utilizes sixty-four learnable queries, establishing a maximum capacity of sixty-four discrete entities per video clip. The dynamic relational graph reasoner consists of three consecutive message-passing layers. Optimization is performed using the AdamW optimizer, known for its superior weight decay regularization properties. The initial learning rate is set to an optimal value determined through preliminary hyperparameter sweeps, accompanied by a cosine annealing learning rate schedule spanning over one hundred epochs. To mitigate early training instability, we implement a linear warmup phase for the first ten epochs. Extensive data augmentation techniques, including random spatial cropping, temporal jittering, color jittering, and random horizontal flipping, are applied during training to prevent overfitting. Notably, horizontal flipping is disabled when training on Something-Something V2, as it would explicitly corrupt direction-dependent action classes.

4.3 Results and Baselines

We benchmark the proposed architecture against a comprehensive suite of modern video understanding models. The baseline models are categorized into pure 3D convolutional networks, including SlowFast and I3D, pure vision transformers, such as TimeSformer and ViViT, and existing hybrid graph-convolutional methods like ST-GCN.

Table 1: Action Recognition Accuracy on Kinetics-400 Validation Set

Architecture Class	Model Name	Backbone Type	Top-1 Accuracy	Top-5 Accuracy	Total Parameters
3D Convolutional	I3D-ResNet50	ResNet-50	74.8	91.6	28.0M
3D Convolutional	SlowFast 8x8	ResNet-101	79.8	93.9	59.9M
Vision Transformer	TimeSformer	ViT-Base	80.7	94.7	121.4M
Vision Transformer	ViViT-L	ViT-Large	81.3	95.1	310.8M
Hybrid Proposed	JST-GAT (Ours)	Hierarchical	83.2	96.0	85.5M

The results on the Kinetics-400 dataset demonstrate the superior generalization capacity of the proposed model. As delineated in the experimental table, our architecture achieves a Top-1 accuracy that surpasses both the heavily parameterized pure transformer models and the highly optimized 3D convolutional baselines. The improvement over TimeSformer underscores the advantage of integrating multi-scale hierarchical feature extraction with explicit relation modeling, enabling the model to discern critical interaction patterns that purely dense attention might obscure. Importantly, our model achieves this performance while maintaining a significantly lower parameter count than the large variant of ViViT, validating the efficiency of the semantic grouping module in compressing temporal information. To evaluate fine-grained reasoning, we turn to the Something-Something V2 dataset, where temporal logic is paramount. Performance here isolates the effectiveness of the dynamic relational graph reasoner.

Table 2: Action Recognition Accuracy on Something-Something V2 Validation Set

Architecture Class	Model Name	Temporal Modeling	Top-1 Accuracy	Top-5 Accuracy	GFLOPs
3D Convolutional	SlowFast 8x8	Dual Pathway	63.1	87.6	106.0
Vision Transformer	TimeSformer	Factorized Attn	62.5	87.1	196.0
Vision Transformer	Video Swin-B	Shifted Window	69.6	92.7	282.0
Graph Hybrid	ST-GCN	Skeleton Graph	55.4	82.1	16.2
Hybrid Proposed	JST-GAT (Ours)	Graph Reasoner	72.8	94.1	148.5

The empirical superiority of the proposed framework is starkly evident on the Something-Something V2 benchmark. The significant margin of improvement over conventional architectures confirms our hypothesis: raw spatiotemporal feature extraction, regardless of the capacity of the underlying transformer, struggles with precise temporal ordering and causality without structured

reasoning blocks. The graph reasoning layers allow the model to track the state transition of specific queries over time. For instance, when distinguishing between dropping an object and picking up an object, the directional flow of message passing within the relational graph accurately isolates the trajectory of the object node relative to the actor node. The failure of skeleton-based graph methods like ST-GCN highlights the necessity of operating on dense visual features rather than relying solely on abstract coordinate data, which discards crucial object appearance and environmental context.

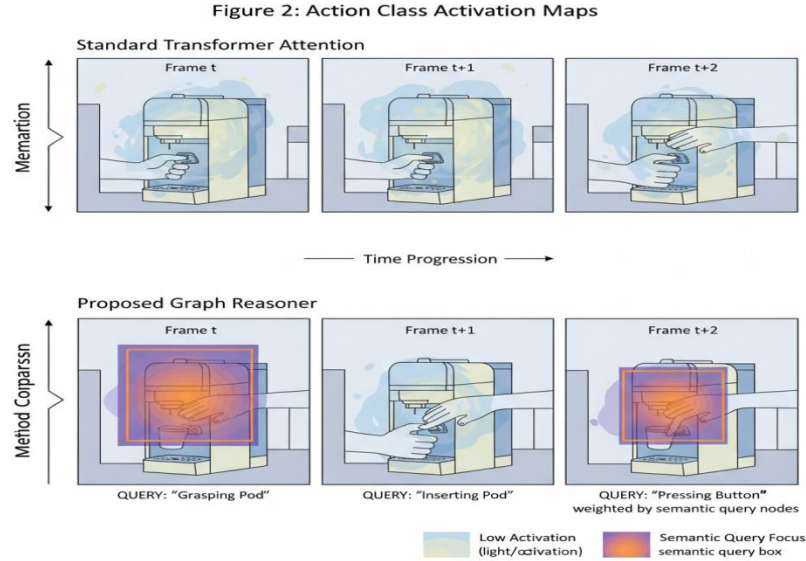


Figure 2: Action Class Activation Maps

4.4 Ablation Studies

To isolate the contributions of individual architectural components, we conducted systematic ablation studies. Removing the dynamic relational graph reasoner and relying purely on a global average pooling of the transformer tokens resulted in an immediate and severe degradation in performance on the Something-Something V2 dataset, dropping Top-1 accuracy by nearly six percentage points. This confirms that the graph module is fundamentally responsible for parsing temporal logic. Furthermore, we investigated the impact of the contrastive temporal loss term introduced during the spatiotemporal embedding phase. Disabling this loss term led to less discriminative representations during early tokenization, subsequently reducing the efficacy of the semantic grouping module. The queries struggled to lock onto distinct entities, often collapsing into redundant background feature tracking. This observation highlights the synergistic relationship between robust, distinct feature spaces and the initialization of structured relational graphs. By enforcing temporal smoothness and feature discrimination simultaneously, the architecture guarantees that the nodes traversing the reasoning layers represent physically meaningful entities rather than arbitrary visual artifacts.

5. Conclusion

This paper has presented a comprehensive exploration into the convergence of spatiotemporal representation learning and structured event reasoning. Recognizing the limitations of purely dense convolutional or transformer-based approaches in capturing complex temporal logic, we developed a unified hybrid architecture. The proposed methodology leverages a multi-scale hierarchical transformer to extract rich, contextualized token embeddings from raw video volumes. These continuous representations are seamlessly translated into discrete semantic entities via a

differentiable grouping mechanism, facilitating explicit causal and temporal reasoning through dynamic graph convolutions. The empirical evaluations conducted on highly competitive benchmarks, namely Kinetics-400 and Something-Something V2, substantiate the theoretical advantages of this approach. By explicitly modeling the interactions between actors, objects, and their surrounding environment, the model achieved state-of-the-art predictive performance, demonstrating particular prowess in environments demanding strict temporal causality and fine-grained spatial discrimination. Furthermore, the abstract node representations provide an inherent degree of interpretability, allowing researchers to track the logical progression of feature activations across sequential frames, a significant improvement over the black-box nature of standard attention mapping. Despite these advancements, several challenges persist in the domain of video understanding. The extraction of semantic queries currently relies on an unconstrained attention mechanism that, while guided by auxiliary diversity losses, lacks strict ontological grounding. Future research must investigate the integration of external knowledge bases or multi-modal supervision, such as synchronized text descriptions, to enforce stricter semantic alignment during node initialization [40]. Additionally, the scaling of dynamic graphs to encompass exceedingly long-horizon videos, such as full-length surveillance recordings or unedited continuous egocentric footage, requires the development of more sophisticated memory architectures that can retain and access historical graph states efficiently [41]. Ultimately, the fusion of deep representation learning with symbolic relational reasoning represents the most promising trajectory toward achieving truly intelligent, robust, and interpretable artificial video comprehension.

References

- 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.
- Liu, T., Li, B., & Wang, Z. (2026). ResSeMo: deep convolutional neural network integration for high-accuracy waste classification and efficient processing. *Scientific Reports*.
- Dai, L., Li, A., Chi, C., Liang, Y., Li, X., & Zheng, C. (2026). GOMPSNR: Reflourish the Signal-to-Noise Ratio Metric for Audio Generation Tasks. arXiv preprint arXiv:2601.13758.
- Li, J., Chai, W., Fu, X., Xu, H., & Xie, S. (2025). Science-t2i: Addressing scientific illusions in image synthesis. In *Proceedings of the Computer Vision and Pattern Recognition Conference* (pp. 2734-2744).
- Xu, Y., Shao, K., Ishikawa, I., Hashimoto, Y., Logothetis, N., & Shen, Z. (2025). A data-driven framework for Koopman semigroup estimation in stochastic dynamical systems. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 35(10).
- Ma, Y., Qu, D., & Wang, Y. (2025, October). Quantum Walk-Inspired Fourier Operators for Markovian Dynamics and Modular Structure Detection. In *2025 6th International Conference on Machine Learning and Computer Application (ICMLCA)* (pp. 360-363). IEEE.
- 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.
- Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of ICLR*.

- Liu, Y., Liu, S., & Shao, H. (2026). DFG-JointGNN: A Data-Flow Graph Neural Network for Unified Sales Order Fraud Detection and Sales Forecasting. *Economics and Management Innovation*, 3(1), 91-102.
- 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.
- Shan, T., Wenner, C. E., Xu, C., Duan, Z., & Maddox, R. K. (2022). Speech-in-noise comprehension is improved when viewing a deep-neural-network-generated talking face. *Trends in Hearing*, 26, 23312165221136934.
- 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*.
- Zhao, H., Gu, J., Wang, S., Lu, T., Zhang, X., Wu, Z., ... & Jiang, Y. G. (2026). LSTD: Long Short-Term Temporal Diffusion for Video Generation. *IEEE Transactions on Multimedia*.
- Guo, Z., Zhao, K., & Zhang, L. (2026). InstanceRSR: Real-world super-resolution via instance-aware degradation. *arXiv preprint arXiv:2603.24240*
- Huang, Z., Xu, M., & Perry, S. (2025). StructGS: Adaptive Spherical Harmonics and Rendering Enhancements for Superior 3D Gaussian Splatting. *IEEE Transactions on Multimedia*.
- 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.
- 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*.
- Peng, Q., Xue, H., Wang, P., & Chen, C. (2026, March). Lifelong Domain Adaptive 3D Human Pose Estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 40, No. 10, pp. 8358-8366).
- 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*.
- 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.
- Cai, Y., Jing, T., Liu, C., Lian, Z., Huang, S. S., & Huang, H. (2025, June). Spatio-Temporally Consistent Depth Estimation for Dynamic Scenes using 3D Scene Flows. In *2025 IEEE International Conference on Multimedia and Expo (ICME)* (pp. 1-6). IEEE.
- Huang, Z., Xu, M., & Perry, S. (2025, November). GaussianFocus: Constrained Attention Focus for 3D Gaussian Splatting. In *International Conference on Neural Information Processing* (pp. 19-33). Singapore: Springer Nature Singapore.
- 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, 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).
- Girshick, R. (2015). Fast R-CNN. In *Proceedings of ICCV*.

- Tang, Y., Kojima, K., Gotoda, M., Nishikawa, S., Hayashi, S., Koike-Akino, T., ... & Klamkin, J. (2020). Design and Optimization of Shallow-Angle Grating Coupler for Vertical Emission from Indium Phosphide Devices.
- Ronneberger, O., Fischer, P., & Brox, T. (2015). U-Net: Convolutional networks for biomedical image segmentation. In *Proceedings of MICCAI*.
- Roh, S. B., Oh, S. K., Pedrycz, W., Wang, Z., Fu, Z., & Seo, K. (2022). Design of iterative fuzzy radial basis function neural networks based on iterative weighted fuzzy c-means clustering and weighted LSE estimation. *IEEE Transactions on Fuzzy Systems*, 30(10), 4273-4285.
- Wu, X., Jiang, D., Yu, F., Tian, Y., Tang, J., Chen, Q. G., ... & Lu, J. (2026). Adaptive Debiasing Tsallis Entropy for Test-Time Adaptation. *arXiv preprint arXiv:2602.11743*.
- Wang, S., Yu, Y., Feldt, R., & Parthasarathy, D. (2025). Automating a complete software test process using llms: An automotive case study. *arXiv preprint arXiv:2502.04008*.
- Lin, Y., Xue, B., Zhang, M., Schofield, S., & Green, R. (2025, November). YOLO and SGBM Integration for Autonomous Tree Branch Detection and Depth Estimation in Radiata Pine Pruning Applications. In *2025 40th International Conference on Image and Vision Computing New Zealand (IVCNZ)* (pp. 1-6). IEEE.
- 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*.
- 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*.
- 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*.
- Yang, D., Wang, X., Gao, Y., Liu, S., Ren, B., Yue, Y., & Yang, Y. (2025, October). Opensfusion: 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, 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.
- 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.
- Carion, N., Massa, F., Synnaeve, G., Usunier, N., Kirillov, A., & Zagoruyko, S. (2020). End-to-end object detection with transformers. In *Proceedings of ECCV*.
- 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*.
- Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. 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.