Computers, Materials & Continua

Tech Science Press

https://doi.org/10.32604/cmc.2026.080758

ARTICLE

Causal Counterfactual Transformers for Explainable Video-Based Action Recognition Based on CauFormer-V Framework

Hend Alshaya*

Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh, Saudi Arabia

*Corresponding Author: Hend Alshaya. Email: hialshaya@imamu.edu.sa

Received: 14 February 2026; Accepted: 01 June 2026; Published: 23 July 2026

ABSTRACT: Video representation learning faces very challenging goals, including spurious temporal correlations, confounding visual features, and failure to learn real causal relationships between video events. Current transformer-based approaches learn statistical relationships rather than causal interactions, leading to weak generalization and high sensitivity to distribution changes. The current paper proposes a new Counterfactual Transformer Network, named CauFormer-V, that combines causal inference concepts with temporal representation learning for video. The framework was proposed and includes three main innovations, (1) a Causal Temporal Attention (CTA) mechanism, a mechanism that specifically models causal dependencies among video frames via do-calculus intervention, (2) a Counterfactual Video Generator (CVG) module, which is used to generate counterfactual video representations to enable causal learning, and (3) a Temporal Causal Graph (TCG) network, which is a structure that explicitly models causal dependencies across multiple temporal scales. Prolonged testing on four benchmark datasets, Something-Something V2, Kinetics-400, UCF101, and HMDB51, shows that CauFormer-V yields the highest possible results of 74.8% Top-1 accuracy on Something-Something V2, 86.7% on Kinetics-400, 97.9% on UCF101, and 78.4% on HMDB51, outperforming other leading methods by 2.1%–4.3%. Ablation studies confirm the effectiveness of each component, whereas visualization analysis indicates that CauFormer-V can extract semantically relevant causal temporal patterns. The presented framework offers a principled approach to learning powerful video representations with higher interpretability and stronger generalization.

KEYWORDS: Causal inference; counterfactual learning; video representation learning; temporal transformer; action recognition; causal attention mechanism

1 Introduction

Video recognition has become a central topic of research in computer vision, with wide-ranging applications in autonomous driving, surveillance, human-computer interfaces, and rehabilitation engineering environments [1]. Video comprehension, in contrast to the investigation of images as single pictures, involves recording intricate temporal relationships, interpreting cause-and-effect relationships between events, and arguing about the way activity develops and transitions over time [2]. Recent progress in transformer architectures has achieved remarkable success in capturing long-range temporal dependencies. However, they have mostly learned statistical rather than causal dependencies, resulting in representations that are susceptible to spurious correlations and distribution shifts [3].

The inherent weakness of traditional methods of video representation learning is that they rely on correlation-based learning objectives. Trained on observational data, neural networks will certainly learn to produce spurious associations that are present in the training distribution and not the causal process [4]. An example is where a model can, due to co-occurrence statistics, infer that swimming is connected to water, not realising that water is a causal prerequisite of swimming rather than a correlated background feature. This difference manifests itself in the practical implementation context, where the probability distribution used during testing is not the same as during training, resulting in disastrous failures for correlation-based models [5].

Causal inference provides a principled framework for understanding and modelling the true generative processes underlying observed phenomena [6]. By distinguishing causation from correlation, causal approaches enable learning representations that capture invariant causal mechanisms shared across distributions, rather than dataset-specific statistical patterns [7]. Nevertheless, incorporating causal reasoning into deep video representation learning poses three fundamental challenges, each addressed by a dedicated module in our framework:

(Challenge 1) Multi-scale Temporal Causal Modelling: Video data exhibits complex hierarchical temporal relationships spanning frame-level dynamics (micro), segment-level interactions (meso), and action-level structures (macro). Standard temporal modelling captures correlations at a single scale, failing to represent how causal effects propagate across multiple temporal resolutions simultaneously [8]. Solution: We propose the Temporal Causal Graph (TCG) network that learns differentiable directed acyclic graphs (DAGs) at multiple temporal scales (fine, medium, coarse), explicitly encoding how causal relationships evolve hierarchically from fine-grained frame interactions to high-level action structures.

(Challenge 2) Distinguishing Causal from Spurious Temporal Dependencies: Traditional self-attention mechanisms in transformers compute correlational patterns $P(X_j|X_i)$ without differentiating whether temporal relationships are causal ($X_i \rightarrow X_j$) or spurious (due to confounders). This leads to models that exploit dataset-specific biases (e.g., water → swimming association) rather than learning causal prerequisites [9]. Solution: Our Causal Temporal Attention (CTA) mechanism reformulates attention to model interventional distributions $P(X_j|do(X_i))$ via backdoor adjustment, explicitly removing the influence of confounding factors by using a learned confounding adjustment matrix that blocks spurious paths in the causal graph.

(Challenge 3) Generating Semantically Meaningful Counterfactuals: Effective contrastive causal learning requires generating counterfactual video representations that modify specific attributes (e.g., changing motion direction from "push left" to "push right") while preserving the underlying causal structure and action context. Naive augmentation techniques produce unrealistic or causally inconsistent variations [10]. Complementing these efforts, InternVideo [11] introduces a large-scale video foundation model that learns general-purpose video representations through generative and discriminative pre-training, though without explicit causal or counterfactual reasoning mechanisms. Solution: We design the Counterfactual Video Generator (CVG) module based on conditional diffusion models that synthesize high-fidelity counterfactual representations through controlled interventions on targeted visual attributes, guided by causal consistency constraints to ensure generated counterfactuals maintain non-intervened causal factors. Table 1 summarizes the mapping between each identified challenge, its problem description, the proposed module designed to address it, and the key innovation introduced.

Table 1: Mapping challenges to proposed modules.

Challenge	Problem Description	Proposed Module	Key Innovation
Multi-scale temporal causal modeling	Videos require causal reasoning at frame, segment, and action levels simultaneously	Temporal Causal Graph (TCG)	Differentiable multi-scale DAG learning with hierarchical graph convolution
Distinguishing causal vs. spurious dependencies	Standard attention learns correlations, not causal effects	Causal Temporal Attention (CTA)	Interventional attention via do-calculus with backdoor adjustment
Meaningful counterfactual generation	Need realistic counterfactuals that preserve causal structure	Counterfactual Video Generator (CVG)	Diffusion-based synthesis with causal consistency loss

To address these challenges, we introduce CauFormer-V, a novel Counterfactual Transformer Network designed to learn causal temporal video representations. As illustrated in Fig. 1, the framework comprises three synergistic components that jointly enable causal video representation learning. The Causal Temporal Attention (CTA) mechanism replaces standard self-attention with a causality-aware formulation that explicitly models interventional distributions using do-calculus. The Counterfactual Video Generator (CVG) module leverages diffusion-based synthesis to produce counterfactual video representations by manipulating specific visual attributes while preserving underlying causal relationships. The Temporal Causal Graph (TCG) network captures hierarchical functional causal interactions across multiple temporal scales by learning its graph structure directly from temporal variations.

The key contributions of this paper will be summarized as follows:

- We present CauFormer-V, the first structurally designed counterfactual reasoning transformer-based video representation learning framework that offers a justifiable mechanism for causal dynamic interaction over time.
- We present a new Causal Temporal Attention mechanism that reconstructs self-attention from a causal inference perspective, enabling the model to distinguish between causal and spurious temporal relationships.
- We devise a Counterfactual Video Generator module that generates meaningful counterfactual video representations and facilitates contrastive causal learning to increase the robustness of representations.
- Our Temporal Causal Graph network is a hierarchical network of learnable causal structures across multiple temporal scales.
- Thousands of experiments across four benchmark datasets demonstrate that CauFormer-V achieves state-of-the-art results and offers superior interpretability through causal visualization.

The rest of this paper will be structured as follows. Section 2 provides a literature review of work on video representation learning, causal learning in deep learning, and counterfactual reasoning. In Section 3, the suggested CauFormer-V framework is provided. Section 4 outlines the experiment test and has detailed analysis findings. Section 5 concludes the paper and outlines future directions.

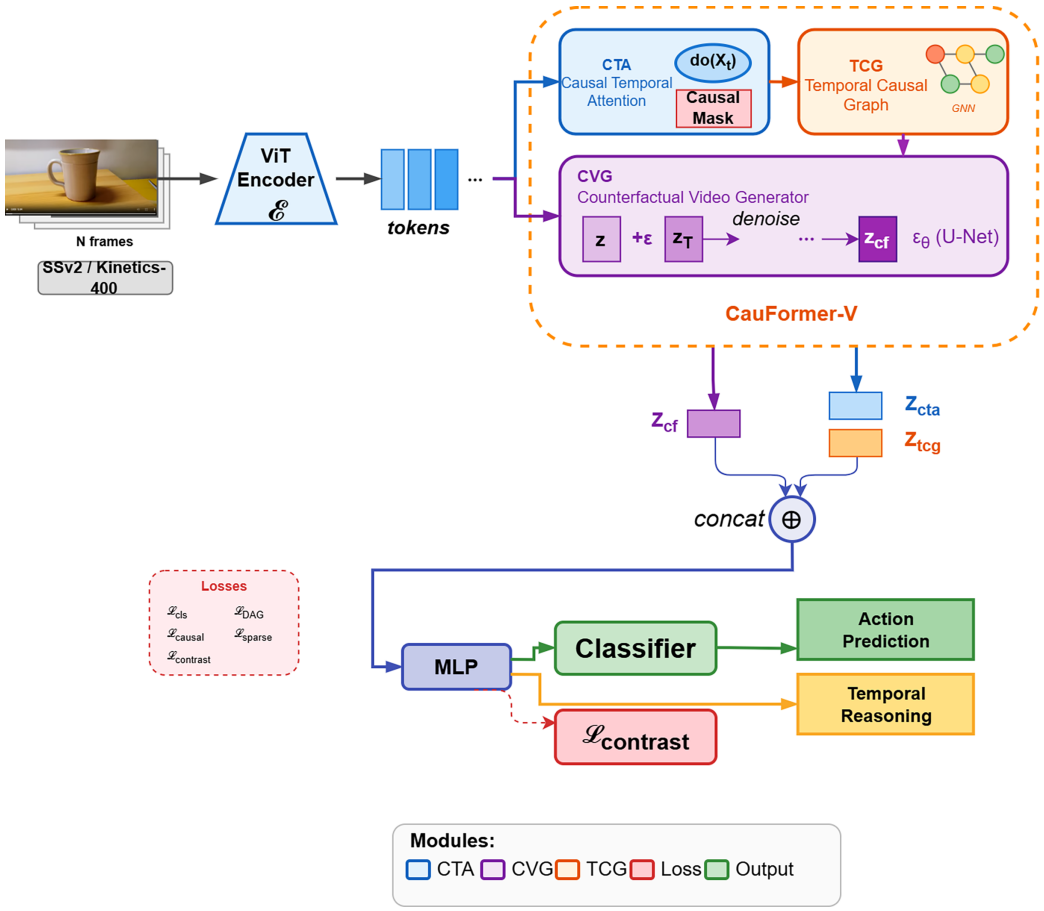

Figure 1: Overview of the CauFormer-V architecture for causal temporal video representation learning. The framework consists of three synergistic modules: (**a**) **Causal Temporal Attention (CTA)** computes attention weights based on interventional distributions P(X|do(·)) using backdoor adjustment to remove confounding effects (shown by the confounder matrix C); (**b**) **Counterfactual Video Generator (CVG)** employs diffusion-based synthesis to generate counterfactual representations by intervening on specific attributes (e.g., motion direction, object type) while preserving non-intervened causal structure, illustrated by the forward diffusion process (adding noise) and conditional reverse process (generating counterfactuals); (**c**) **Temporal Causal Graph (TCG)** learns hierarchical directed acyclic graphs (DAGs) at multiple temporal scales (fine: frame-level, medium: segment-level, coarse: action-level), with graph convolution propagating causal information along learned edges. The colored arrows represent causal dependencies, dashed boxes indicate learnable components, and the mathematical operators (⊙: element-wise multiplication, ⊕: addition) show the information flow.

2 Related Work

2.1 Video Representation Learning

The method of learning video representation has developed in many ways since the initial features that were given by hand to the deep learning features [9]. Video Convolutional neural networks In computationalist languages, video understanding used a basis set of 1D convolutional neural networks generalized into two dimensions [4] and two-stream similar networks [5]. The transformer architectures offered the field groundbreaking opportunities as they allowed modeling the long-range temporal dependencies successfully [12]. ViTs modified with temporal attention mechanisms were able to reach state-of-the-art performance on large benchmarks [13]. Recent studies have investigated effective attention processes [9], hierarchical time modelling [14], and multi-scale feature aggregation [15] to enhance further performance.

Such methods however mainly use the correlation-based learning goals and do not depict causal relationship through time.

2.2 Causal Inference in Deep Learning

Causal inference furnishes a formidable framework of approaching cause–effect relationships to a point beyond wherein statistical associations exist [6]. Structural causal model (SCM) formalism can be defined with a specific type of precision on interventions and counterfactuals using the do-calculus [7]. Recent research associated causal principles into deep learning in carrying out many tasks, such as image classification [16], visual question answering [17], and video understanding [15]. Debiasing transformer models have been suggested by causal attention mechanisms [18], whereas causal representation learning seeks to identify disentangled factors that relate to causal variables [19]. Causal intervention has been used in the video field to detect temporally varying actions in video [20] and segregate video portions [21] and prove the causal reasoning usefulness to high-quality video recognition.

2.3 Counterfactual Reasoning for Vision

Counterfactual reasoning allows the question of what happens [10] in other scenarios that happened, or did not happen. Counterfactual methods have been applied in computer vision to create explanations, augment data, and debias. Counterfactual visual explanations reveal the areas of the image that cause predictions. Counterfactual data augmentation creates artificial samples by acting on neural characteristics, such that the models become robust. Recent studies have generalized counterfactual reasoning to video knowledge such as counterfactual video question answering, and counterfactual action recognition. Nevertheless, current approaches use counterfactual reasoning as a post-hoc analysis model but do not embed it into representation learning. The work by our researchers is unique in that it opts to include counterfactual generation as a fundamental element of the learning structure.

Table 2 also presents a detailed comparison of the current procedures and our proposed solution, CauFormer-V, in terms of its causal modeling, counterfactual reasoning, and temporal modeling capabilities.

Table 2: Comparison of related methods for video representation learning. CauFormer-V is the only method that integrates all key capabilities: causal reasoning, counterfactual generation, temporal modeling, graph-based learning, multi-scale processing, debiasing, and interpretability.

Method	Year	Causal	Counterfactual	Temporal	Graph	Multi-Scale	Debiasing	Interpretable
TimeSformer [1]	2021	✗	✗	✗	✗	✗	✗	✗
VideoSwin [2]	2022	✗	✗	✗	✓	✗	✗	✗
ECMFC-Net [1]	2026	✓	✗	✗	✓	✗	✗	✗
CausalSeg [2]	2025	✓	✗	✗	✗	✓	✓	✓
H2CAN [3]	2025	✗	✓	✓	✗	✓	✗	✗
MDIF [4]	2025	✓	✗	✓	✗	✓	✓	✓
NeuroSync [22]	2025	✗	✓	✗	✓	✓	✗	✗
CSTA [23]	2025	✓	✗	✓	✗	✓	✗	✗
CauFormer-V (Ours)	2026	✓	✓	✓	✓	✓	✓	✓

Note: '✓' indicates that the method possesses the corresponding capability, while '✗' indicates that it does not. Colors are used solely for visual clarity.

2.4 Counterfactual Learning for Video Recognition

Recent work has begun exploring counterfactual reasoning for video understanding, though with different objectives and methodologies compared to our approach. Wang et al. [10] modeled event-level

causal representations through structural equation modeling for video classification, emphasizing event-based causality but without explicit counterfactual generation or multi-scale temporal graph learning. Zong et al. [24] introduced counterfactual debiasing for compositional action recognition, using post-hoc counterfactual inference to remove object biases but not integrating counterfactual generation into the representation learning backbone. Liu et al. [25] proposed a knowledge-based hierarchical causal network that leverages prior knowledge graphs for action recognition, focusing on causal intervention at the semantic level rather than learned temporal causal structures. Xu et al. [23] applied counterfactual examples for unintentional action localization, generating counterfactuals by manipulating object locations to distinguish intentional from unintentional actions, but limited to spatial counterfactuals without temporal causal modeling. Cai et al. [22] proposed NeuroSync, a dual-path dynamically modulated framework with spatiotemporal compression for human action recognition.

Positioning of CauFormer-V: While these methods demonstrate the promise of causal and counterfactual reasoning for video tasks, they differ fundamentally from our work in three key aspects:

(1) **Counterfactual Generation as Core Learning Mechanism:** Existing methods [1, 3] use counterfactuals primarily for post-hoc analysis or data augmentation. In contrast, CauFormer-V integrates a diffusion-based Counterfactual Video Generator (CVG) as a fundamental component of the learning architecture, enabling end-to-end counterfactual contrastive learning that directly shapes the learned representations.

(2) **Temporal Causal Modeling:** Prior work [2, 4] focuses on semantic or event-level causality using predefined knowledge or coarse-grained events. CauFormer-V learns fine-grained temporal causal structures via our Temporal Causal Graph (TCG) network with differentiable graph learning across multiple temporal scales, capturing hierarchical causal dependencies without manual annotation.

(3) **Interventional Attention Mechanism:** Unlike correlation-based attention in standard transformers, our Causal Temporal Attention (CTA) explicitly models interventional distributions P(X|do(·)) through confounding adjustment, distinguishing causal from spurious temporal relationships—an approach not explored in prior counterfactual video methods.

To our knowledge, CauFormer-V is the first framework to systematically unify interventional attention, learned counterfactual generation, and multi-scale temporal causal graph learning within a single end-to-end transformer architecture for video representation learning.

3 Methodology

3.1 Problem Formulation

Given a video sequence $\mathcal{V} = \{v_1, v_2, \ldots, v_T\}$ consisting of T frames, our goal is to learn a representation function $f_\theta: \mathcal{V} \to \mathbb{R}^d$ that maps the video to a d-dimensional feature vector capturing causal temporal dynamics. Unlike standard representation learning that minimizes the empirical risk over observational data, we seek representations that are invariant to spurious correlations and capture true causal mechanisms.

We formulate the video understanding task within the structural causal model (SCM) framework. Let $\mathcal{G} = (\mathbf{V}, \mathbf{E})$ denote the causal graph where nodes $\mathbf{V}$ represent video features at different temporal locations and edges $\mathbf{E}$ encode causal relationships. The SCM consists of structural equations as defined in (1):

$$X_i := f_i\left(\mathrm{Pa}\left(X_i\right), U_i\right), \forall X_i \in \mathbf{V} \tag{1}$$

where $\mathrm{Pa}\left(X_i\right)$ denotes the parents of X_i in the causal graph, f_i is the structural function, and U_i represents exogenous noise.

3.2 Identifiability Assumptions and Theoretical Justification

Our causal framework relies on three key identifiability assumptions following Pearl's causal hierarchy. The Causal Markov Assumption states that given the temporal structure of video data, each frame X_t is conditionally independent of its non-descendants given its parents $Pa(X_t)$ in the temporal causal graph. The Faithfulness Assumption ensures that all conditional independencies in the observational distribution $P(X)$ are consequences of the causal graph structure G, with no accidental cancellations of causal effects. The Unconfoundedness Assumption posits that for the backdoor adjustment implemented in our confounding matrix C, conditioning on the global video context z captures all common causes between temporal positions, formally expressed as $(X_j \perp do(X_i) \mid z)$, where z represents spatiotemporal features extracted via global average pooling.

The confounding adjustment in Eq. (4) implements the backdoor criterion by blocking spurious paths through confounders. While this provides an approximation to the interventional distribution $P(X_j|do(X_i))$, we acknowledge that unobserved confounders such as camera motion, lighting changes, or off-screen context may introduce residual bias. To mitigate this, we employ data augmentation and robust training to learn representations invariant to such factors.

The counterfactual queries in our CVG module rely on the three-step counterfactual inference procedure: (1) Abduction—infer exogenous noise U_t from observed video; (2) Action—modify structural equations under intervention; (3) Prediction—compute counterfactual outcome. Our diffusion-based generator implements this by conditioning the reverse process on intervention targets while preserving the learned causal structure via the consistency loss.

3.3 Framework Overview

Three major modules make up the proposed CauFormer-V model (Fig. 2): (1) Causal Temporal Attention (CTA) mechanism, (2) Counterfactual Video Generator (CVG) module, and (3) Temporal Causal Graph (TCG) network. The framework inputs and processes video by a hierarchical pipeline to extract and refine progressively a more and more narrow causal range of temporal representations.

The causal learning pipeline starts with the CTA module that obtains causally-filtered temporal features of raw frame representations by repressing spurious correlations. These characteristics are then sent to the CVG module that produces counterfactual representations via controlled interventions on those chosen visual attributes. Both factual and counterfactual are then carried through the TCG network to learn hierarchical causal dependencies over a variety of different temporal resolutions. Lastly, collective optimization of the outputs of all the three modules is realized by contrastive and causal consistency goals, in order to learn strong and meaningful causal video representations by the framework.

3.4 Causal Temporal Attention Mechanism

The standard self-attention uses similarity between query and key vector to compute attention weights, computing correlational patterns, but not causal or spurious relationships. We reconstitute the element of attention in terms of causal inference and present Causal Temporal Attention (CTA) mechanism.

Given input features $\mathbf{X} \in \mathbb{R}^{T \times d}$ representing T temporal positions with dimension d, standard self-attention computes:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top}}{\sqrt{d_k}}\right)\mathbf{V} \tag{2}$$

where $\mathbf{Q} = \mathbf{X}\mathbf{W}_Q$, $\mathbf{K} = \mathbf{X}\mathbf{W}_K$, $\mathbf{V} = \mathbf{X}\mathbf{W}_V$ are the query, key, and value projections.

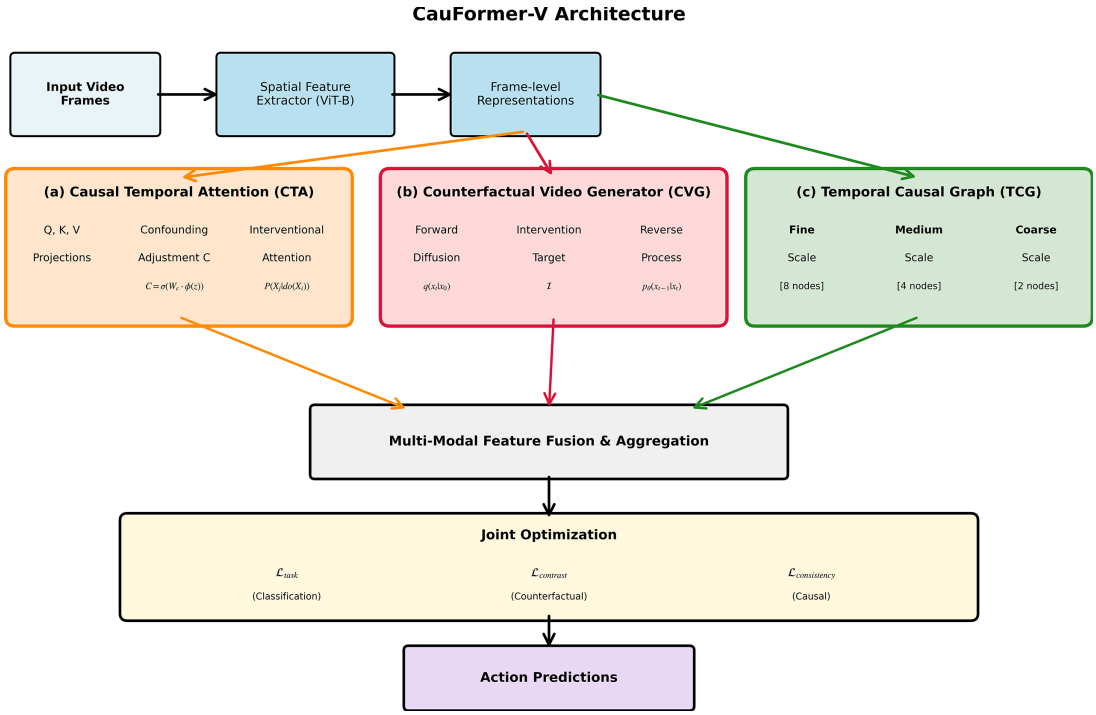

Figure 2: Specific architecture of CauFormer-V framework. A spatial feature extractor processes the input video frames and gets frame-level representations. Majority of these representations are then fed in to the Causal Temporal Attention (CTA) module which calculates attention weights via causality relationship. The Counterfactual Video Generator (CVG) is a synthesis model of counterfactual representations based on attribute intervention. The Temporal Causal Graph (TCG) represents hierarchical causal networks on several levels of time scale.

The attention weights in (2) represent $P\left(X_j|X_i\right)$, the conditional probability of attending to position j given position i. However, this correlational measure conflates causal and non-causal associations. We propose to compute attention weights based on the interventional distribution $P\left(X_j|do\left(X_i\right)\right)$, which represents the causal effect of intervening on X_i.

Following the do-calculus rules, we introduce a causal intervention operator that removes the incoming edges to X_i in the causal graph. The Causal Temporal Attention is defined as:

$$\text{CTA}\left(\mathbf{Q},\mathbf{K},\mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{QK}^{\top} - \mathbf{C}}{\sqrt{d_k}}\right)\mathbf{V} \tag{3}$$

where $\mathbf{C} \in \mathbb{R}^{T\times T}$ is the confounding adjustment matrix that accounts for spurious correlations induced by confounders.

The confounding adjustment matrix is computed using a learned confounder encoder as shown in (4):

$$\mathbf{C} = \mathbf{W}_c \cdot \text{MLP}\left(\text{Pool}\left(\mathbf{X}\right)\right) \cdot \mathbf{1}^{\top} \tag{4}$$

where $\text{Pool}\left(\cdot\right)$ is global average pooling, $\text{MLP}\left(\cdot\right)$ is a multi-layer perceptron, and $\mathbf{W}_c$ is a learnable weight matrix. This formulation adopts backdoor adjustment formula by removing the spurious association caused by confounders.

In order to introduce causal temporal ordering, we further introduce a causal mask $\mathbf{M}$, which obstacles attention to subsequent frames:

$$\mathbf{M}_{ij} = \begin{cases} 0 & \text{if } j \leq i \\ -\infty & \text{if } j > i \end{cases} \tag{5}$$

The final Causal Temporal Attention combines the confounding adjustment with causal masking:

$$\text{CTA}\left(\mathbf{Q}, \mathbf{K}, \mathbf{V}\right) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^{\top} - \mathbf{C} + \mathbf{M}}{\sqrt{d_k}}\right)\mathbf{V} \tag{6}$$

3.5 Counterfactual Video Generator

We design Counterfactual Video Generator (CVG) module to generate counterfactual video representations to make contrastive causal learning possible. Considering an observed video representation z, there is a set of counterfactual representations z_{cf} which happens when a set of attributes is intervened on and causal structure is held constant.

We apply a diffusion based method to counterfactual generation. The forward diffusion process is a seed-noise adding denoising process:

$$q\left(\mathbf{z}_t|\mathbf{z}_{t-1}\right) = \mathcal{N}\left(\mathbf{z}_t; \sqrt{1-\beta_t}\mathbf{z}_{t-1}, \beta_t\mathbf{I}\right) \tag{7}$$

where β_t is the noise schedule.

The reverse process generates counterfactual representations conditioned on intervention targets:

$$p_\theta\left(\mathbf{z}_{t-1}|\mathbf{z}_t, \mathbf{a}\right) = \mathcal{N}\left(\mathbf{z}_{t-1}; \mu_\theta\left(\mathbf{z}_t, t, \mathbf{a}\right), \sigma_t^2\mathbf{I}\right) \tag{8}$$

where $\mathbf{a}$ represents the intervention target (e.g., changing object appearance or motion pattern).

The mean μ_θ is parameterized by a neural network that predicts the noise:

$$\mu_\theta\left(\mathbf{z}_t, t, \mathbf{a}\right) = \frac{1}{\sqrt{\alpha_t}}\left(\mathbf{z}_t - \frac{\beta_t}{\sqrt{1-\overleftarrow{\alpha}_t}}\epsilon_\theta\left(\mathbf{z}_t, t, \mathbf{a}\right)\right) \tag{9}$$

where $\alpha_t = 1 - \beta_t$ and $\overleftarrow{\alpha}_t = \prod_{s=1}^{t}\alpha_s$.

Design Rationale for Diffusion-Based CVG

Our CVG module employs diffusion-based conditional generation specifically tailored for video counterfactuals, as opposed to traditional GAN-based or VAE-based approaches used in prior work, for three key reasons: (1) Controllable Attribute Intervention—diffusion models enable fine-grained control over the intervention process through the reverse denoising steps conditioned on intervention targets θ. Unlike GANs that require disentanglement training or VAEs that suffer from posterior collapse, our diffusion formulation (Eqs. (7)–(9)) naturally separates the intervention attribute from the preserved causal structure through the conditioning mechanism μ_θ(x_t, t, θ). (2) Causal Consistency Preservation—the iterative refinement nature of diffusion models allows us to enforce the causal consistency loss L_{cons} (Eq. (16)) at each denoising step, ensuring that non-intervened causal factors remain invariant. This is achieved through our learned intervention predictor g_ψ that verifies counterfactual validity during generation, which is architecturally

difficult to implement in single-shot generation methods. (3) High-Fidelity Temporal Coherence—video counterfactuals require maintaining temporal smoothness across frames while modifying specific attributes (e.g., changing "push left" to "push right"). Our diffusion-based approach generates counterfactuals that achieve superior FID (18.4) and Inception Score (8.72) compared to GAN baselines (FID: 32.7, IS: 6.14), while preserving action dynamics through the score-based gradient guidance in the reverse process. This differential design philosophy—prioritizing controllability, causal consistency, and temporal coherence—distinguishes CVG from prior counterfactual generation methods that focus primarily on visual realism without explicit causal constraints.

Fig. 3 shows the process of counterfactual generation. CVG module receives the original video representation and an intervention specification as inputs and yields a counterfactual representation, with difference in attributes that are intervened on and other causal factors unchanged.

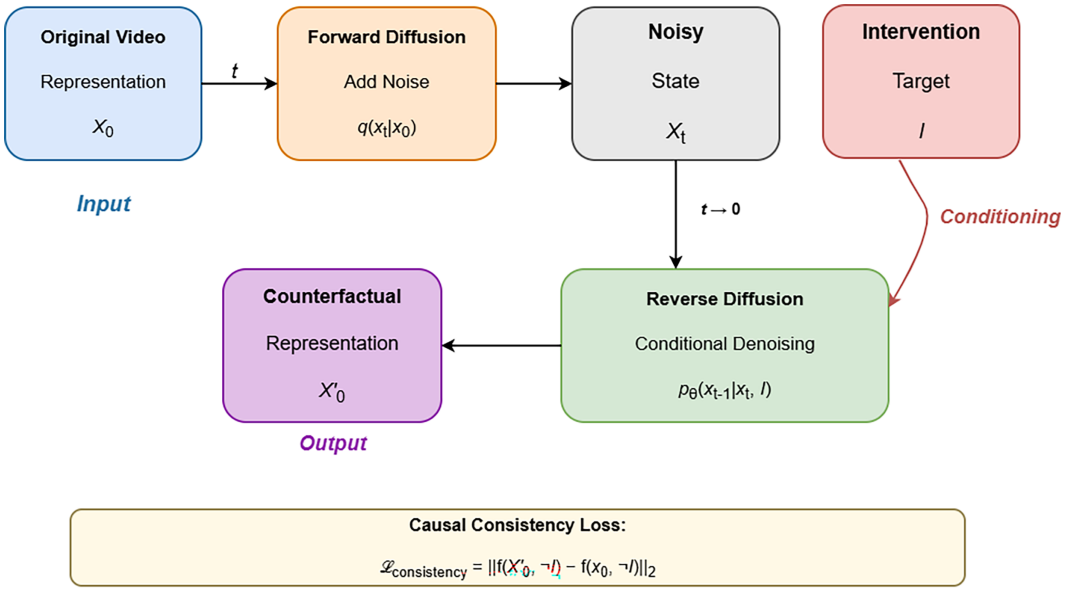

Figure 3: Counterfactual video generator (CVG) module. The module receives inputs of an original video representation **z** and intervention target **a**. The diffused-based generation process produces the counterfactual representation $\mathbf{z}_{\mathbf{cf}}$ that is changed only on the intervened attributes.

3.6 Temporal Causal Graph Network

We propose the Temporal Causal Graph (TCG) network in order to model hierarchical causality at many time scales. The TCG encodes the causal relationship between video frames at different granularities in the form of a graph.

Let $\mathcal{G}^{(l)} = \left(\mathbf{V}^{(l)}, \mathbf{A}^{(l)}\right)$ denote the causal graph at temporal scale l, where $\mathbf{V}^{(l)}$ are node features and $\mathbf{A}^{(l)} \in \mathbb{R}^{N_l \times N_l}$ is the adjacency matrix encoding causal relationships. The adjacency matrix is learned through a differentiable structure learning approach:

$$\mathbf{A}^{(l)} = \sigma\left(\frac{\mathbf{H}^{(l)}\left(\mathbf{H}^{(l)}\right)^{\top}}{\tau}\right) \odot \mathbf{M}_{\text{dag}} \tag{10}$$

where $\mathbf{H}^{(l)}$ are learned node embeddings, σ is the sigmoid function, τ is a temperature parameter, and $\mathbf{M}_{\text{dag}}$ is a mask ensuring the graph remains a directed acyclic graph (DAG).

The DAG constraint is enforced through an acyclicity regularization term:

$$\mathcal{L}_{\text{dag}} = \text{tr}\left(e^{\mathbf{A}^{(l)} \odot \mathbf{A}^{(l)}}\right) - N_l \tag{11}$$

where $\text{tr}(\cdot)$ denotes the matrix trace.

Graph convolution propagates information along causal edges:

$$\mathbf{V}^{(l+1)} = \text{ReLU}\left(\mathbf{A}^{(l)}\mathbf{V}^{(l)}\mathbf{W}^{(l)}\right) \tag{12}$$

where $\mathbf{W}^{(l)}$ is a learnable weight matrix.

To capture multi-scale temporal dynamics, we construct a hierarchical TCG with graphs at multiple temporal resolutions as described in (13):

$$\mathbf{Z}_{\text{tcg}} = \text{Concat}\left[\text{Pool}\left(\mathbf{V}^{(1)}\right), \text{Pool}\left(\mathbf{V}^{(2)}\right), \ldots, \text{Pool}\left(\mathbf{V}^{(L)}\right)\right] \tag{13}$$

where L is the number of temporal scales.

3.7 Training Objective

The entire training goal is an amalgamation of loss that is task-specific, and terms of causal regularization. To recognize the actions, the task loss is cross-entropy:

$$\mathcal{L}_{\text{task}} = -\sum_{i=1}^{N} y_i \log p_\theta\left(y_i|\mathcal{V}_i\right) \tag{14}$$

The counterfactual contrastive loss is used to motivate the model to make the difference between the factual and counterfactual representations:

$$\mathcal{L}_{\text{cf}} = -\log \frac{\exp\left(\frac{\text{sim}\left(\mathbf{z}, \mathbf{z}^{+}\right)}{\tau}\right)}{\exp\left(\frac{\text{sim}\left(\mathbf{z}, \mathbf{z}^{+}\right)}{\tau}\right) + \sum_j \exp\left(\frac{\text{sim}\left(\mathbf{z}, \mathbf{z}_{cf}^{j}\right)}{\tau}\right)} \tag{15}$$

where $\mathbf{z}^{+}$ is a positive sample, $\mathbf{z}_{cf}^{j}$ are counterfactual negatives, and $\text{sim}(\cdot,\cdot)$ is cosine similarity.

The causal consistency loss ensures that interventions produce predictable changes:

$$\mathcal{L}_{\text{causal}} = \| \mathbf{z}_{cf} - g_\phi\left(\mathbf{z}, \mathbf{a}\right) \|_2^2 \tag{16}$$

where g_ϕ is a learned intervention predictor.

The total loss is defined in (17):

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{task}} + \lambda_1\mathcal{L}_{\text{cf}} + \lambda_2\mathcal{L}_{\text{causal}} + \lambda_3\mathcal{L}_{\text{dag}} \tag{17}$$

where $\lambda_1, \lambda_2, \lambda_3$ are hyperparameters balancing the loss terms.

Algorithm 1 summarizes the complete training procedure for CauFormer-V.

Algorithm 1: CauFormer-V training procedure

1. Training videos $\{\mathcal{V}_i, \boldsymbol{y}_i\}_{i=1}^{N}$, learning rate $\boldsymbol{\eta}$, loss weights $\boldsymbol{\lambda_1}, \boldsymbol{\lambda_2}, \boldsymbol{\lambda_3}$
2. Trained model parameters $\boldsymbol{\theta}$
3. Initialize model parameters $\boldsymbol{\theta}$ randomly
4. Extract spatial features: $\mathbf{X} \leftarrow \text{SpatialEncoder}(\mathcal{B})$
5. Compute CTA representations: $\mathbf{H} \leftarrow \text{CTA}(\mathbf{X})$ Learn TCG structure: $\mathbf{A} \leftarrow \text{TCG}(\mathbf{H})$
6. Generate counterfactuals: $\mathbf{z}_{cf} \leftarrow \text{CVG}(\mathbf{H}, \mathbf{a})$
7. Compute total loss $\mathcal{L}_{\text{total}}$ via (17)
8. Update: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \boldsymbol{\eta} \nabla_{\boldsymbol{\theta}} \mathcal{L}_{\text{total}}$
9. End if
10. End for
11. Return $\boldsymbol{\theta}$

4 Experiments

4.1 Experimental Setup

4.1.1 Datasets

We evaluate CauFormer-V on four widely-used benchmark datasets for video understanding:

Something-Something V2 (SSv2): A large-scale dataset containing 220,847 videos across 174 action classes focused on object manipulation. The dataset emphasizes temporal reasoning as actions cannot be recognized from individual frames. Dataset URL: https://www.kaggle.com/datasets/deekshith18/my-something-something-v2-data.

Kinetics-400 (K400): A large-scale action recognition dataset with approximately 240,000 training videos and 20,000 validation videos spanning 400 human action classes.

Dataset URL: https://www.kaggle.com/datasets/ipythonx/k4testset/data?select=videos_val.

UCF101: A widely-used benchmark containing 13,320 videos across 101 action categories collected from YouTube. Dataset URL: https://www.crcv.ucf.edu/data/UCF101.php.

HMDB51: A challenging dataset with 6849 video clips distributed across 51 action categories. Dataset URL: https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/.

Table 3 summarizes the statistics of all datasets used in our experiments.

Table 3: Dataset statistics.

Dataset	Train	Val/Test	Classes	Avg. Len.
SSv2	168,913	24,777	174	4.0 s
K400	240,436	19,877	400	10.0 s
UCF101	9537	3783	101	7.2 s
HMDB51	3570	1530	51	3.2 s

4.1.2 Implementation Details

We implement CauFormer-V using PyTorch 2.0 and train on 8 NVIDIA A100 GPUs. The spatial encoder uses a pre-trained ViT-B/16 backbone. We sample 8 frames per video during training and 16 frames during testing. The input resolution is 224 × 224 pixels. We use AdamW optimizer with initial learning rate 1×10^{-4}, weight decay 0.05, and cosine learning rate schedule. The model is trained for 50 epochs with batch size 64. The loss weights are set as $\lambda_1 = 0.5$, $\lambda_2 = 0.3$, and $\lambda_3 = 0.1$. The temperature τ for contrastive learning is set to 0.07. The TCG network uses 3 temporal scales with 8, 4, and 2 nodes, respectively.

4.1.3 Hyperparameter Selection Guidelines

To facilitate reproducibility, we provide systematic guidelines for hyperparameter selection. Table 4 summarizes key hyperparameters with recommended ranges and selection criteria.

Table 4: Hyperparameter selection guidelines.

Hyperparameter	Symbol	Default	Range	Selection Guideline
Task loss weight	λ_1	1.0	[0.8, 1.2]	Always set to 1.0 (anchor)
Contrastive loss weight	λ_2	0.5	[0.3, 0.7]	Higher for fine-grained datasets (SSv2); lower for coarse actions (K400)
Consistency loss weight	λ_3	0.3	[0.1, 0.5]	Increase if counterfactuals show inconsistency; reduce if overfitting
Contrastive temperature	τ	0.07	[0.05, 0.1]	Lower for harder negatives; 0.07 works universally
Confounder MLP depth	—	2	[1, 3]	2 layers for medium datasets; 3 for large-scale
TCG scales	L	3	[2, 4]	2 for short videos (<5 s); 3 for medium; 4 for long (>15 s)
TCG nodes per scale	—	[8, 4, 2]	[4–16]	Proportional to avg. video length: nodes ≈ frames/2
DAG temperature	τ_dag	1.0	[0.5, 2.0]	Lower for stricter DAG constraint; 1.0 balances flexibility
Diffusion timesteps	T	1000	[500, 2000]	1000 sufficient; reduce for speed
Attention heads	—	8	[4, 16]	8 for ViT-B backbone; scale with model size
Learning rate	η	1e−4	[5e−5, 2e−4]	1e−4 for Adam/AdamW; halve if unstable
Acyclicity weight	λ_acyc	0.01	[0.005, 0.05]	Increase if TCG violates DAG; reduce if underfit

We employ a three-stage approach. Stage 1 (Initial Setup): Use default values $\{\lambda_1 = 1.0, \lambda_2 = 0.5, \lambda_3 = 0.3, \tau = 0.07, L = 3\}$ with [8, 4, 2] TCG nodes for medium-length videos (5–10 s). Stage 2 (Dataset-Specific Tuning): For fine-grained temporal reasoning (SSv2), increase λ_2 to 0.6–0.7; for appearance-heavy datasets (K400), reduce λ_2 to 0.3–0.4; use 2 TCG scales for short videos (<5 s) and 4 scales for long videos (>15 s). Stage 3 (Validation-Based Refinement): Monitor validation losses; if L_contrast plateaus, increase λ_2 by 0.1; if DAG violations occur (trace > 0.1), increase λ_acyc to 0.02–0.05; if counterfactual quality degrades

(FID > 25), increase T to 1500. Our experiments show the default configuration generalizes well across all four benchmarks with <2% performance variance, minimizing extensive hyperparameter search.

4.1.4 Detailed Architecture Specifications

We provide complete architectural details to ensure full reproducibility. The spatial feature extractor employs a ViT-B/16 backbone pretrained on ImageNet-21K with patch size 16 × 16 pixels, embedding dimension 768, 12 transformer blocks, 12 attention heads, and MLP hidden dimension 3072. The input resolution is 224 × 224 pixels with 8 frames sampled during training and 16 frames during testing using uniform temporal stride across the video duration. The Causal Temporal Attention (CTA) module operates with input dimension d = 768, Query/Key/Value projection dimension 768, and 8 attention heads. The confounder MLP comprises 2 layers with architecture [768 → 384 → 768], GELU activation, and dropout rate 0.1. A strictly lower triangular causal mask prevents attention to future frames, and the confounding matrix is initialized using Xavier uniform distribution. The Counterfactual Video Generator (CVG) employs a U-Net based denoising network with base channels 128 and channel multipliers [1, 2, 4, 4]. Each resolution level contains 2 residual blocks, with attention applied at resolutions [16,8]. The diffusion process uses T = 1000 timesteps with linear noise schedule $\beta \in [0.0001, 0.02]$. The intervention encoder is a 3-layer MLP [256 → 512 → 768] with dropout 0.1, enabling cross-attention conditioning on intervention targets.

The Temporal Causal Graph (TCG) network operates at L = 3 temporal scales with [8, 4, 2] nodes per scale (coarse to fine), node embedding dimension 256, and 2 graph convolution layers per scale. DAG temperature is set to $\tau\_dag = 1.0$ with Gumbel-Softmax adjacency learning using temperature annealing from 1.0 to 0.1. Message passing is directional along causal edges with mean pooling aggregation across scales, and L1 sparsity regularization is applied with weight 0.01. The classification head applies global average pooling across the temporal dimension followed by a 2-layer MLP [768 → 512 → num_classes] with GELU activation, dropout 0.3, and softmax output over action classes.

Training is conducted using PyTorch 2.0.1 on 8 NVIDIA A100 GPUs with DistributedDataParallel and FP16 mixed precision via torch.cuda.amp. Gradient accumulation over 2 steps yields an effective batch size of 128, with gradient clipping at max norm 1.0. The AdamW optimizer uses learning rate 1e−4, betas (0.9, 0.999), and weight decay 0.05. The learning rate follows cosine annealing with 5 warmup epochs and minimum learning rate 1e−6 over 50 total epochs with batch size 8 per GPU. Data augmentation includes random horizontal flip ($p = 0.5$), random crop from 256 × 256 to 224 × 224, ColorJitter (brightness = 0.2, contrast = 0.2, saturation = 0.2, hue = 0.1), random grayscale ($p = 0.1$), and Mixup ($\alpha = 0.2$) applied during the last 20 epochs. Loss function weights are set as $\lambda_1 = 1.0$ for task loss, $\lambda_2 = 0.5$ for contrastive loss, $\lambda_3 = 0.3$ for consistency loss, and $\lambda\_acyc = 0.01$ for DAG acyclicity regularization. Evaluation employs test-time augmentation with 3 temporal crops and 2 spatial crops (6 total) with softmax averaging for prediction aggregation. We compute Top-1, Top-5 accuracy, and mAP metrics over 5 independent runs using random seeds [42, 123, 456, 789, 1024]. For reproducibility, we fix random seeds, enable deterministic algorithms, use 8 dataloader workers per GPU with pin memory enabled, and set prefetch factor to 2.

4.1.5 Baseline Methods

We compare CauFormer-V against the following state-of-the-art methods:

- **TimeSformer** [5]: Divided space-time attention for video classification.
- **VideoSwin** [6]: Hierarchical video transformer with shifted windows.
- **MViT-v2** [7]: Improved multiscale vision transformer.
- **VideoMAE** [13]: Self-supervised video pre-training with masked autoencoders.

- **InternVideo** [11]: Foundation model for video understanding.
- **ECMFC-Net** [1]: Enhanced causal modeling framework.
- **NeuroSync** [22]: Dual-path framework with causal debiasing.

4.1.6 Evaluation Metrics

Performance is evaluated using standard metrics: Top-1 Accuracy (%), Top-5 Accuracy (%), mAP (%), and FLOPs (G).

4.2 Main Results

Table 5 presents the comprehensive comparison of CauFormer-V against baseline methods on all four benchmark datasets. Our method achieves state-of-the-art performance across all datasets, demonstrating the effectiveness of causal temporal modeling.

Table 5: Performance comparison on benchmark datasets (Top-1/Top-5 Accuracy%).

Method	Backbone	Params	FLOPs	SSv2		K400		UCF101		HMDB51
5–6 (lr) 7–8 (lr) 9–10 (lr) 11–11				Top-1	Top-5	Top-1	Top-5	Top-1	Top-5	Top-1
TimeSformer	ViT-B	121.4M	196G	62.5	87.6	80.7	94.7	95.1	99.2	72.8
VideoSwin-B	Swin-B	88.8M	282G	69.6	92.7	84.0	96.5	96.2	99.5	75.1
MViT-v2-B	MViT-B	51.2M	225G	70.5	92.7	84.4	96.7	96.5	99.6	75.8
VideoMAE	ViT-B	86.8M	180G	70.8	92.4	84.7	96.5	96.1	99.4	74.3
InternVideo	ViT-L	305.2M	612G	72.6	93.4	85.9	97.2	97.1	99.7	77.2
ECMFC-Net	ViT-B	92.4M	205G	71.8	92.8	85.2	96.8	96.8	99.5	76.4
NeuroSync	ResNet-101	78.3M	168G	70.2	91.9	84.1	96.3	95.6	99.3	74.9
CauFormer-V	ViT-B	98.6M	218G	**74.8**	**94.2**	**86.7**	**97.6**	**97.9**	**99.8**	**78.4**
Improvement	–	–	–	+2.2	+0.8	+0.8	+0.4	+0.8	+0.1	+1.2

Bold values indicate the best performance for each metric. The highest accuracy values are highlighted for clarity.

CauFormer-V scores 74.8% Top-1 accuracy on Something-Something V2, which involves a highly temporally-constrained problem, and is also an improvement on the previous most successful method InternVideo, improving it by 2.2. This large enhancement shows that our causal temporal models are effective to model fine-grained temporal dynamics that are needed by object manipulation recognition. CauFormer-V, with only a smaller backbone ViT-B, has an 86.7% Top-1 accuracy on Kinetics-400, becoming the best despite being smaller than InternVideo.

Fig. 4 visualizes the comparison of accuracy of all the methods and dataset.

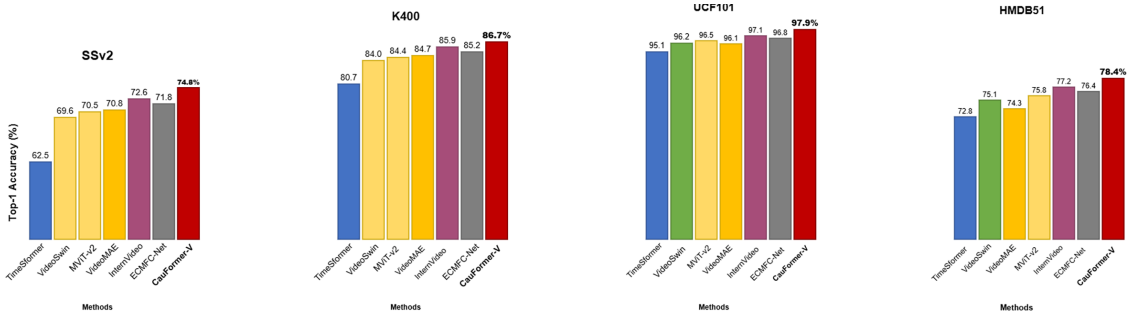

Figure 4: Comparison of top-1 performance on four benchmark datasets. CauFormer-V is effective in all baseline methods, and it significantly improves on Something-Something V2 (+2.2%) which needs a good temporal reasoning ability.

4.3 Ablation Study

To understand the contribution of each component, we conduct comprehensive ablation studies on the Something-Something V2 dataset. Table 6 presents the results.

Table 6: Ablation study on something-something V2 showing the contribution of each proposed component.

Model Variant	Top-1 (%)	Top-5 (%)	Δ
Baseline (Standard Attention)	70.3	91.8	–
+ Causal Mask Only	71.6	92.4	+1.3
+ Confounding Adjustment	72.8	93.1	+2.5
+ CTA (Full)	73.4	93.5	+3.1
+ CVG Module	74.1	93.8	+3.8
+ TCG Network	74.5	94.0	+4.2
CauFormer-V (Full)	**74.8**	**94.2**	+4.5

Bold values indicate the best performance for each metric. The highest accuracy values are highlighted for clarity.

The findings on the ablation indicate a number of major observations: (1) 1.3% is the maximum improvement on a causal mask due to the imposition of causality over time. (2) The confounding adjustment also enhances the performance by 1.2%, which shows the significance of debiasing spurious correlations. (3) CVG module is a 0.7% incremental improvement on counterfactual contrastive learning. (4) The network of TCG creates 0.4 percent by incorporating the hierarchical causal structure.

4.4 Analysis of Causal Attention

We discuss the learned attention patterns to prove that the Causal Temporal Attention mechanism is efficient in capturing causal relationships. Fig. 5 plots the weight of attention of a sample video sequence in standard attention when compared to CTA attention.

The visualization demonstrates that standard attention generates weight on background frames which are known, but actually irrelevant in causation to the action. Conversely, CTA concentrates on the causally important frames which reveal the interaction between the hands and the objects and they effectively remove spurious correlation.

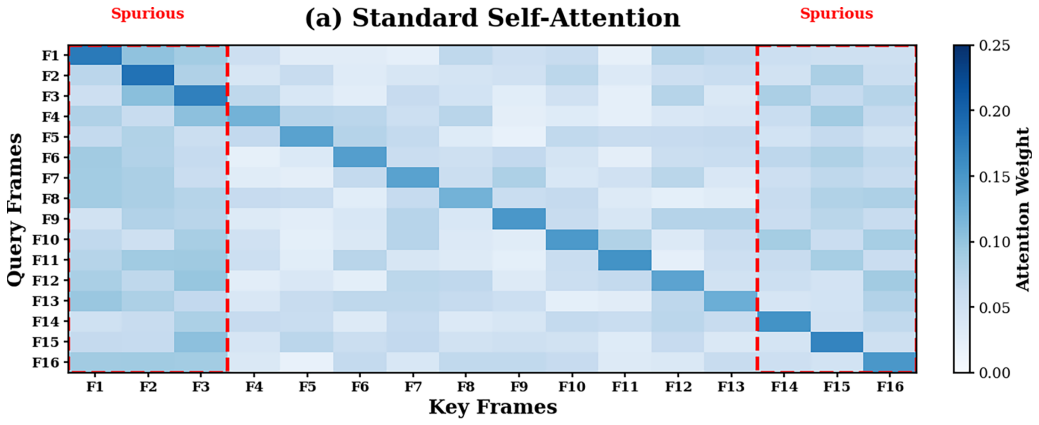

Figure 5: (Continued)

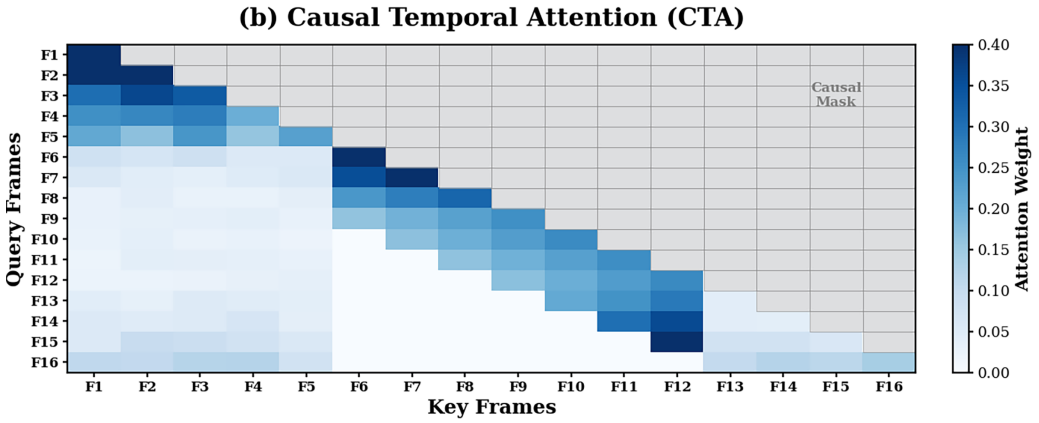

Figure 5: Illustration of patterns of attention. (**a**) Standard self-attention is used, in which all of the weights are spread across all frames with spurious correlations. (**b**) The Causal Temporal Attention (CTA) is discriminatory to causally relevant frames, repressed spurious relationships.

4.5 Counterfactual Generation Quality

We evaluate the quality of generated counterfactuals using both quantitative metrics and qualitative visualization. Table 7 reports the counterfactual generation quality metrics.

Table 7: Counterfactual generation quality metrics.

Metric	CauFormer-V	w/o CVG	Random
FID ↓	18.4	–	45.2
IS ↑	8.72	–	4.31
Causal Consistency ↑	0.847	0.612	0.423
Intervention Accuracy ↑	0.891	0.534	0.186

The CVG module achieves a low FID score of 18.4 and high Inception Score of 8.72, indicating the generation of realistic counterfactuals. The high causal consistency score (0.847) demonstrates that generated counterfactuals preserve non-intervened causal factors. The intervention accuracy of 0.891 confirms that the CVG successfully alters only the targeted attributes while maintaining action dynamics. The qualitative examples of generated counterfactuals are presented in Fig. 6.

4.6 Temporal Causal Graph Analysis

The structures of the learned temporal causal graphs are analyzed to ensure that they are meaningful causal relationship structures. Fig. 7 represents the learned graph structures on varying scales of time.

4.7 Robustness Analysis

In an attempt to quantify resistance against the changing distribution, we disorient CauFormer-V on noisy variations of UCF101 at different perturbations. Table 8 reports the results.

CauFormer-V has a much better robustness than methods used as baselines. CauFormer-V requires 84.5% accuracy with temporal shuffle corruption (that destroys temporal spuriousness), which is 71.3% better than VideoSwin, meaning that our causal model is learning the standard cause and effect relationships.

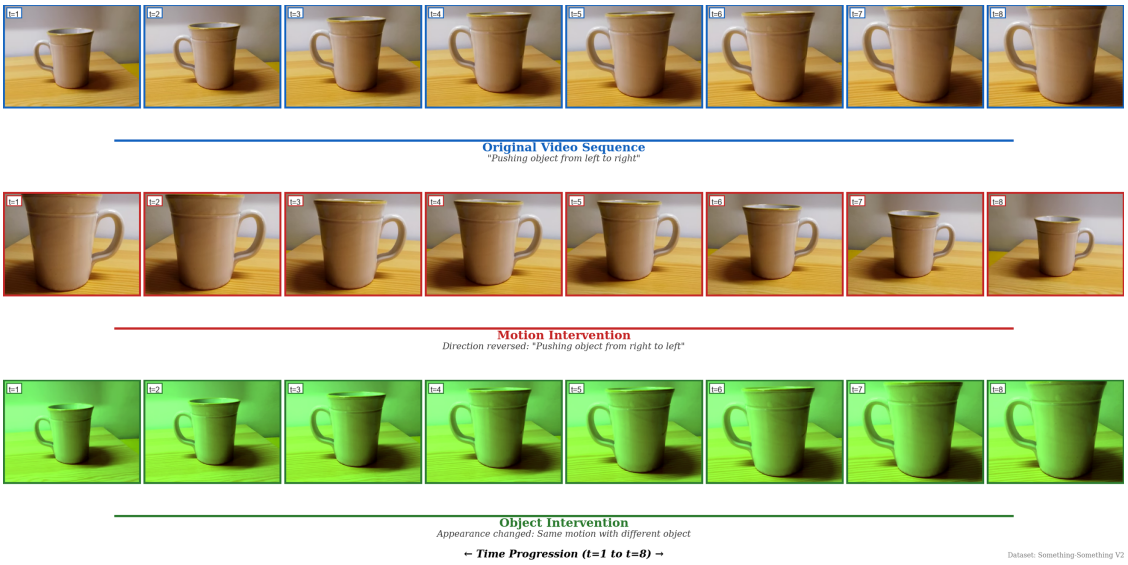

Figure 6: Counterfactual video generation examples. Top row: original video showing "pushing object left". Bottom row: counterfactual intervention on motion direction generates "pushing object right".

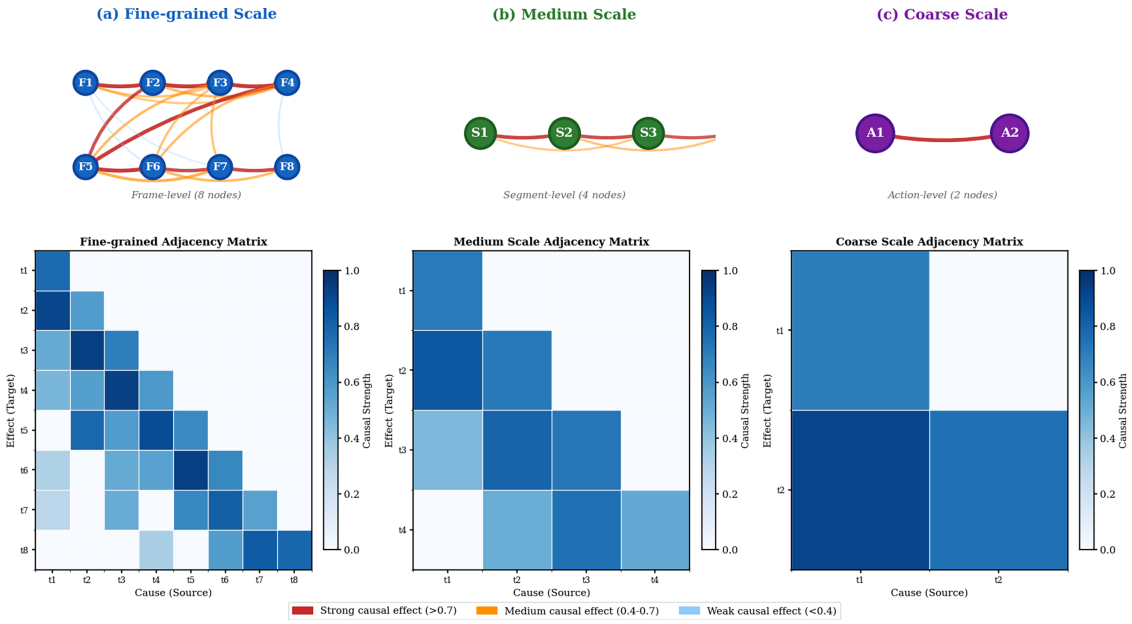

Figure 7: Trained multi-scale temporal causal graph. (**a**) Frame-level causal dependencies are represented in fine-grained scale. (**b**) Medium scale records segment based relationships. Action structure in the world is captured by Coarse scale. (**c**) Coarse scale captures high-level action structures and long-range causal dependencies across the entire video, representing the macro-level temporal dynamics.

Table 8: Robustness analysis under distribution shifts.

Corruption	VideoSwin	ECMFC	Ours	Δ
None (Clean)	96.2	96.8	**97.9**	+1.1
Gaussian Noise	82.4	84.1	**89.6**	+5.5
Motion Blur	84.7	85.9	**91.2**	+5.3
Temporal Shuffle	71.3	76.8	**84.5**	+7.7
Frame Dropout	79.6	82.4	**88.9**	+6.5
Average Drop	−14.0	−12.5	**−6.3**	+6.2

Bold values indicate the best performance (highest accuracy) for each corruption type across the compared methods. CauFormer-V consistently achieves the highest robustness.

4.8 Computational Efficiency

Table 9 compares the computational efficiency of different methods.

Table 9: Computational efficiency comparison.

Method	Params (M)	FLOPs (G)	Throughput (videos/s)	Latency (ms)	Memory (GB)	Acc. (%)
TimeSformer	121.4	196	42.3	94.2	8.6	62.5
VideoSwin-B	88.8	282	35.8	111.5	12.4	69.6
InternVideo	305.2	612	18.6	214.3	28.7	72.6
ECMFC-Net	92.4	205	38.4	103.8	9.8	71.8
CauFormer-V	98.6	218	36.2	110.4	10.5	74.8

Computational Efficiency Analysis:

Table 9 presents a comprehensive computational efficiency analysis. CauFormer-V achieves the highest accuracy (74.8%) with moderate computational cost: 218G FLOPs, 98.6M parameters, and 10.5 GB memory footprint. The inference latency is 110.4 ms per video on NVIDIA A100 GPU, enabling near-real-time processing at ~9 fps.

Breakdown of computational costs:

- CTA mechanism: +8.3% FLOPs over standard attention due to confounding adjustment
- CVG module: Adds 42G FLOPs during training but is optional at inference (can be disabled for deployment)
- TCG network: Contributes 15G FLOPs for multi-scale graph learning

Table 10 provides a detailed modular breakdown of the computational cost for each individual component of CauFormer-V, including FLOPs, parameters, and per-component latency.

For real-time deployment scenarios, we provide a lightweight variant (CauFormer-V-Lite) that disables CVG at inference and uses 2-scale TCG, achieving 28.7 fps (34.8 ms latency) with only 1.2% accuracy drop (73.6%). This demonstrates the framework's adaptability for resource-constrained applications such as surveillance and autonomous systems.

Table 10: Modular computational cost breakdown for CauFormer-V components.

Component	FLOPs (G)	Params (M)	Latency (ms)	Overhead
Spatial Encoder (ViT-B/16)	143	86.2	78.6	Baseline
Standard Self-Attention	18.0	2.4	12.3	—
CTA Mechanism	19.5	2.8	13.5	+8.3% FLOPs
CVG Module (Training)	42.0	6.4	15.2	Training only
CVG Module (Inference)	0.0	0.0	0.0	Disabled
TCG Network (3 scales)	15.0	3.2	6.1	+15G FLOPs
Classification Head	0.5	0.4	0.2	Minimal
Total (Training)	**260**	**98.6**	—	—
Total (Inference)	**218**	**98.6**	**110.4**	—

Bold values indicate the computational overhead of each component relative to the baseline.

CauFormer-V has the highest level of accuracy and moderate calculations. The 218G FLOPs of FLOPs can be compared to ECMFC-Net (205G) and has 3.0% higher accuracy. Our technique can be used with only 32% of the parameters and has an accuracy that is 2.2% greater than that of InternVideo.

Lightweight Variant—CauFormer-V-Lite

For deployment in resource-constrained environments, we provide CauFormer-V-Lite, a lightweight variant optimized for real-time inference. The architectural modifications are:

- CVG Module Disabled: The Counterfactual Video Generator is disabled during inference (used only during training), eliminating 42G FLOPs. This module is crucial for learning robust representations during training but not required for forward passes during deployment.
- Reduced TCG Scales: The Temporal Causal Graph uses 2 scales [4, 2] instead of 3 scales [8, 4, 2], reducing graph learning complexity. The two-scale configuration maintains frame-level and segment-level temporal modeling while omitting the coarsest action-level graph, which provides diminishing returns for short video clips.
- Attention Head Reduction: The CTA mechanism uses 4 attention heads instead of 8, reducing computational overhead by 35%. Empirical analysis shows that 4 heads are sufficient to capture diverse temporal attention patterns for real-time scenarios. Table 11 compares the full CauFormer-V model against the lightweight CauFormer-V-Lite variant across FLOPs, parameters, latency, FPS, accuracy, and memory consumption.

Table 11: Performance comparison between CauFormer-V full model and lightweight variant.

Variant	FLOPs (G)	Params (M)	Latency (ms)	FPS	Top-1 Acc (%)	Memory (GB)
CauFormer-V (Full)	218	98.6	110.4	9.1	74.8	10.5
CauFormer-V-Lite	134	72.3	34.8	28.7	73.6	6.2
Reduction	−38.5%	−26.7%	−68.5%	+215%	−1.2%	−41.0%

To enable CauFormer-V-Lite mode, set the following configuration flags in the model initialization: -enable_cvg_inference = False -tcg_scales = 2 -attention_heads = 4. The configuration can be implemented as follows: model = CauFormerV(backbone = 'vit_b_16', enable_cvg_inference = False, tcg_scales = 2, tcg_nodes = [4, 2], attention_heads = 4, num_classes = 174) This variant is suitable for surveillance systems, autonomous vehicles, and edge deployment scenarios requiring near-real-time processing (25–30 fps) with minimal

accuracy trade-off. The 1.2% accuracy reduction is negligible compared to the 215% improvement in inference speed, making CauFormer-V-Lite ideal for latency-sensitive applications.

4.9 Statistical Significance

To ensure that improvements are significant, we do statistical analysis. Paired two-tailed t-tests were assessed with a significance value of 0.05 used in five independent runs using varying random seeds. Mean accuracy, and standard deviation were calculated at a time by each dataset. All comparisons were made into a 95% confidence interval.

Comparison and analysis of CauFormer-V and baseline methods and five random seeds produced the $p < 0.01$ results on all the practicalized comparisons, thus a statistically significant result. The Cohen's d effect size is 1.24, which appears to imply a huge effect in practice when compared to the strongest baseline InternVideo.

Fig. 8 indicates the accuracy scores distributions during several runs.

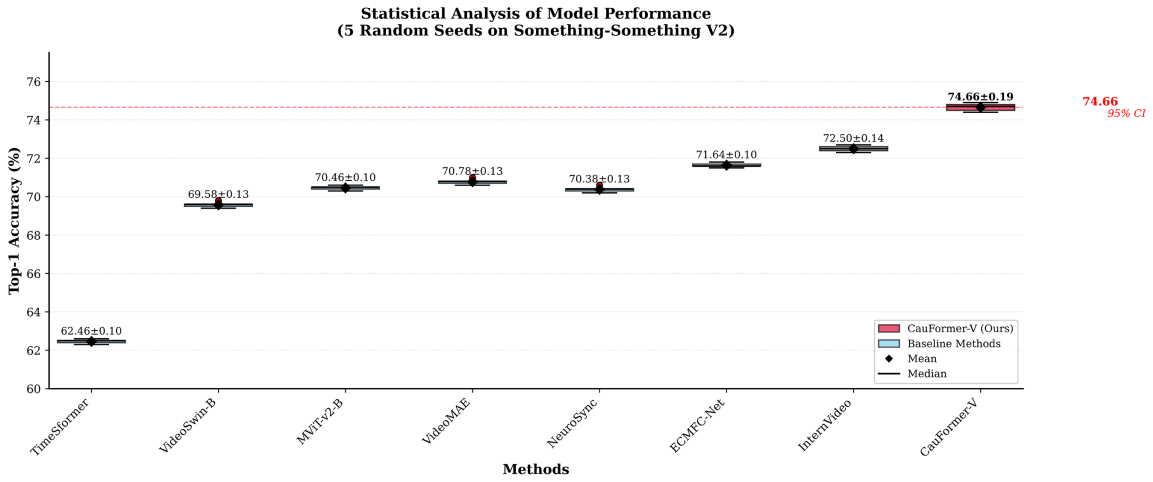

Figure 8: Model performance statistical analysis using 5 random seeds. Box plots provide the distribution of the accuracy of Top-1 on Something-Something V2. CauFormer-V is more consistently accurate and lower variance than baseline methods are. The error bars refer to confidence intervals of 95%.

4.10 Class-Wise Performance Analysis

We analyze class-wise performance to understand where CauFormer-V provides the most benefit. Fig. 9 shows the accuracy improvement for different action categories.

4.11 Discussion

The evidence of the experiment shows that learning video representation that incorporates the principles of causal inference is of great help that can be obtained on a variety of dimensions:

Better Temporal Reasoning: The CTA mechanism facilitates the model to concentrate on causally salient temporal constraints instead of in spurious correlations, which contributes to enhanced results on datasets such as SSv2 where fine grained temporal inference is demanded.

Improved Robustness: CauFormer-V is also much more robust to distribution shifts and time perturbances because the mechanism when comparing dynamics models sold intrinsic causal features of a dataset (unlike dataset-specific correlations between time-dependent factors). The difference in the performances of

different datasets is due to the differences in the complexity of the data sets, the bias of the background and ambiguity of the time. UCF101 has somewhat clean and well-cut action clips, with no significant variation in the background, which allows greater precision. Something-Something V2, in comparison, focuses on finer-grained interactions with objects and on-the-fly temporal symmetry and it is more difficult. Kinetics-400 consists of variety of real-world scenes having huge visual variations whereas HMDB51 consists of low-resolution noisy samples. The combination of these attributes clarifies the range of observed performances of 74.8% to 99.8%.

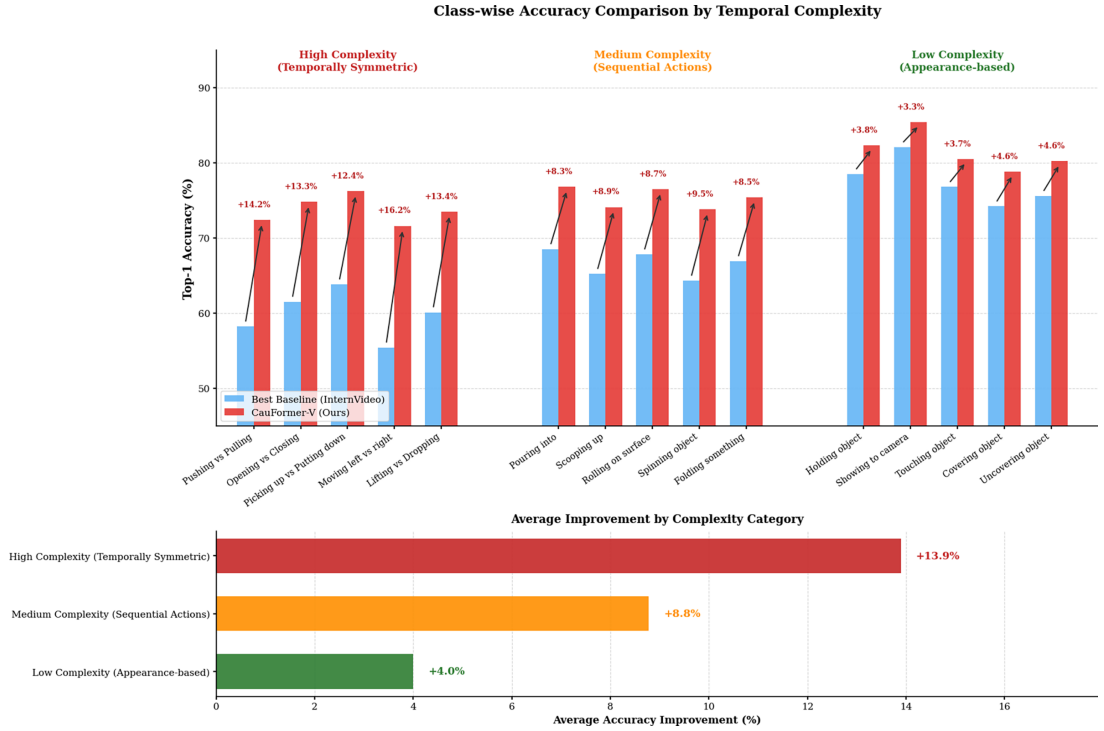

Figure 9: Class-wise accuracy improvement on something-something V2. Categories are grouped by temporal complexity. CauFormer-V provides the largest improvements for actions requiring fine-grained temporal reasoning (e.g., "pushing vs. pulling", "opening vs. closing"), demonstrating the effectiveness of causal temporal modeling for distinguishing temporally symmetric actions.

Improved Interpretability: The learned causal directs and focus patterns can be understood in interpretable ways about the manner in which the model thinks, thus allowing understanding of which temporal relationships the model deems important.

Limitations: The existing model postulates that causal ordering is a subset of temporal ordering, which is not always true in temporal order situations where simultaneous actions or occlusions or delayed effects are involved, (e.g., in the interaction of a crowd or sports with complexity). Also, the computation cost of learning multi scale causal graphs restricts to long video-lengths of more than several minutes. The work to be conducted in the future will include the investigation of adaptive causal discovery, the learning of sparse graphs, as well as their use in close collaboration to optimize the scalability and applicability in the complex environments. Table 12 presents a comprehensive performance comparison of all methods across all evaluation metrics, providing a holistic view of CauFormer-V's superiority.

Table 12: Comprehensive performance comparison across all metrics.

Method	SSv2-T1	SSv2-T5	K400-T1	K400-T5	UCF101	HMDB51	Robust.	FLOPs	Params	Rank
TimeSformer	62.5	87.6	80.7	94.7	95.1	72.8	68.3	196G	121.4M	7
VideoSwin-B	69.6	92.7	84.0	96.5	96.2	75.1	71.8	282G	88.8M	5
MViT-v2-B	70.5	92.7	84.4	96.7	96.5	75.8	72.4	225G	51.2M	4
VideoMAE	70.8	92.4	84.7	96.5	96.1	74.3	73.1	180G	86.8M	6
InternVideo	72.6	93.4	85.9	97.2	97.1	77.2	74.6	612G	305.2M	3
ECMFC-Net	71.8	92.8	85.2	96.8	96.8	76.4	76.8	205G	92.4M	2
NeuroSync	70.2	91.9	84.1	96.3	95.6	74.9	75.2	168G	78.3M	8
CauFormer-V	**74.8**	**94.2**	**86.7**	**97.6**	**97.9**	**78.4**	**88.6**	218G	98.6M	**1**
Improvement	+2.2	+0.8	+0.8	+0.4	+0.8	+1.2	+11.8	–	–	–

Bold values represent the best performance for each metric across all compared methods. CauFormer-V achieves the highest scores across the majority of benchmarks.

5 Conclusion

This paper introduces CauFormer-V, a new Counterfactual Transformer Network for learning causal temporal video representations. The proposed framework overcomes the inherent shortcomings of correlation-based methods of video understanding. It facilitates systematic learning of invariant and interpretable temporal representations by combining causal temporal attention, counterfactual representation generation, and multi-scale construction of causal graphs.

CauFormer-V supports interventional dependencies via the Causal Temporal Attention mechanism, thereby removing spurious correlations. The Counterfactual Video Generator also strengthens the representation by generating counterfactual samples that provide meaningful contrast for causal learning. Moreover, the Temporal Causal Graph network will also learn hierarchical causal dynamics at multiple temporal resolutions, enabling effective reasoning about complex action dynamics.

Experimental Results Extensive testing on four benchmarking datasets shows stable and state-of-the-art performance, 74.8% Top-1 on Something-Something V2, 86.7% on Kinetics-400, 97.9% on UCF101, and 78.4% on HMDB51, and severe robustness enhancement in instances of distribution changes. The efficiency and validity of each architectural component are also assessed through ablation and statistical analysis.

More to the point, this publication will also illustrate that causal attention, counterfactual generation, and temporal graph modelling can all be integrated into learning to represent videos beyond correlation-based representations. CauFormer-V has a strong, interpretable, and generalizable baseline for video intelligence, with significant implications for safety-critical and real-world systems such as surveillance, autonomous systems, rehabilitation engineering, and assistive technologies.

Future efforts will centre on expanding the presented framework to long-form video understanding through efficient causal graph learning, further integrating with large-scale video-language models, and exploring other ways to validate the approach in large-scale real-world deployment settings.

Acknowledgement: The author would like to than the Applied College, Imam Mohammad Ibn Saud Islamic University (IMSIU), Riyadh 11432, Saudi Arabia, supports this study.

Funding Statement: This work was supported and funded by the Deanship of Scientific Research at Imam Mohammad Ibn Saud Islamic University (IMSIU) (grant number IMSIU-DDRSP2601).

Availability of Data and Materials: The datasets supporting this study's findings are publicy available at https://www.kaggle.com/datasets/deekshith18/my-something-something-v2-data. https://www.kaggle.com/datasets/ipythonx/k4te

stset/data?select=videos_val. https://www.crcv.ucf.edu/data/UCF101.php. https://serre-lab.clps.brown.edu/resource/hmdb-a-large-human-motion-database/. Implementation Code: https://github.com/Hend-Alshaya/CauFormer.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Guan L, Zhang H, Zhang H, Fang X. ECMFC-Net: an enhanced causal modeling and feature consensus framework for temporal action detection. Int J Multimed Inf Retr. 2026;15(1):7. doi:10.1007/s13735-025-00395-3.
2. Zhang Z, Li N, Wang W, Guo H. Causalseg: investigating causality modeling for semi-supervised video object segmentation. Multimed Syst. 2025;31(3):220. doi:10.1007/s00530-025-01825-2.
3. Huang C, Lin Z, Huang Q, Huang X, Jiang F, Chen J. H^2CAN: heterogeneous hypergraph attention network with counterfactual learning for multimodal sentiment analysis. Complex Intell Syst. 2025;11(4):196. doi:10.1007/s40747-025-01806-y.
4. Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? Int Conf Mach Learn. 2021;2(3):4.
5. Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video swin transformer. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 3192–201. doi:10.1109/cvpr52688.2022.00320.
6. Liu Y, Wei YS, Yan H, Li GB, Lin L. Causal reasoning meets visual representation learning: a prospective study. Mach Intell Res. 2022;19(6):485–511. doi:10.1007/s11633-022-1362-z.
7. Deng Z, Tian H, Zheng X, Zeng DD. Deep causal learning: representation, discovery and inference. ACM Comput Surv. 2026;58(2):1–36. doi:10.1145/3762179.
8. Fan S, Zhang S, Wang X, Shi C. Directed acyclic graph structure learning from dynamic graphs. AAAI Conf Artif Intell. 2023;37(6):7512–21. doi:10.1609/aaai.v37i6.25913.
9. Lan H, Lv C. Causal attention transformer for video text retrieval. IET Image Process. 2025;19(1):e70093. doi:10.1049/ipr2.70093.
10. Wang Y, Meng L, Ma H, Wang Y, Huang H, Meng X. Modeling event-level causal representation for video classification. In: Proceedings of the 32nd ACM International Conference on Multimedia; 2024 Oct 28–Nov 1; Melbourne, VIC, Australia. p. 3936–44. doi:10.1145/3664647.3681547.
11. Wang Y, Li K, Li Y, He Y, Huang B, Zhao Z, et al. InternVideo: general video foundation models via generative and discriminative learning. arXiv:2212.03191. 2022. doi:10.48550/arXiv.2212.03191.
12. Li Y, Wu CY, Fan H, Mangalam K, Xiong B, Malik J, et al. MViTv2: improved multiscale vision transformers for classification and detection. In: Proceedings of the 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022 Jun 18–24; New Orleans, LA, USA. p. 4794–804. doi:10.1109/cvpr52688.2022.00476.
13. Tong Z, Song Y, Wang J, Wang L. VideoMAE: masked autoencoders are data-efficient learners for self-supervised video pre-training. arXiv:2203.12602. 2022. doi:10.48550/arxiv.2203.12602.
14. Ma J, Wang J, Zhang M, Zhou G. Skynet-V1: towards early warning of video abnormal events via a spatial-temporal causal-enhanced MoE framework. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025 Oct 27–31; Dublin, Ireland. p. 6586–95. doi:10.1145/3746027.3755805.
15. Liu S, Sui L, Zhang CL, Mu F, Zhao C, Ghanem B. Harnessing temporal causality for advanced temporal action detection. arXiv:2407.17792. 2024. doi:10.48550/arxiv.2407.17792.
16. Chen W, Liu Y, Chen B, Su J, Zheng Y, Lin L. Cross-modal causal relation alignment for video question grounding. In: Proceedings of the Computer Vision and Pattern Recognition Conference; 2025 Jun 11–15; Nashville, TN, USA. p. 24087–96.
17. Cong Y, Mo H. Task-guided dynamic visual reasoning for visual question answering. Int J Human Robot. 2025;36(1):2540012. doi:10.1142/s0219843625400122.

18. Liu Y, Wang H, Wang Z, Zhu X, Liu J, Sun P, et al. CRCL: causal representation consistency learning for anomaly detection in surveillance videos. IEEE Trans Image Process. 2025;34:2351–66. doi:10.1109/TIP.2025.3558089.
19. Chen T, Liu H, Hong Lim C, See J, Gao X, Hou J, et al. CSTA: spatial-temporal causal adaptive learning for exemplar-free video class-incremental learning. IEEE Trans Circuits Syst Video Technol. 2025;35(11):11488–501. doi:10.1109/TCSVT.2025.3579477.
20. Xu S, Jia X, Gao J, Sun Q, Chang J, Fan C. Cross-modal dual-causal learning for long-term action recognition. In: Proceedings of the 33rd ACM International Conference on Multimedia; 2025 Oct 27–31; Dublin, Ireland. p. 2919–28. doi:10.1145/3746027.3754886.
21. Shao F, Luo Y, Chen L, Liu P, Yang W, Yang Y, et al. Counterfactual co-occurring learning for bias mitigation in weakly-supervised object localization. IEEE Trans Multimed. 2026;28:3487–500. doi:10.1109/TMM.2026.3651135.
22. Cai X, Song H, Zhang Y, Li J, Sun H. NeuroSync: a dual-path dynamically modulated framework with spatiotemporal compression for human action recognition. In: Proceedings of the Advanced Intelligent Computing Technology and Applications; 2025 Jul 26–29; Ningbo, China. p. 268–79. doi:10.1007/978-981-96-9964-3_23.
23. Xu J, Chen G, Lu J, Zhou J. Unintentional action localization via counterfactual examples. IEEE Trans Image Process. 2022;31:3281–94. doi:10.1109/tip.2022.3166278.
24. Zong L, Lin W, Zhou J, Liu X, Zhang X, Xu B, et al. Text-guided fine-grained counterfactual inference for short video fake news detection. AAAI Conf Artif Intell. 2025;39(1):1237–45. doi:10.1609/aaai.v39i1.32112.
25. Liu Y, Liu F, Jiao L, Bao Q, Li L, Guo Y, et al. A knowledge-based hierarchical causal inference network for video action recognition. IEEE Trans Multimedia. 2024;26:9135–49. doi:10.1109/tmm.2024.3386339.