

JOURNAL OF THE ARTIFICIAL INTELLIGENCE ENGINEERS CONSORTIUM

Volume: 02 Issue: 02 (2026)

<https://jaiec.org>

Causal Spatiotemporal Reasoning Traces for Surgical Video Understanding Under Sparse Procedural Annotations

Niels Van Leeuwen¹, Lucas De Jong²

¹ Informatics Institute, Faculty of Science, University of Amsterdam, Amsterdam,
North Holland, Netherlands

² Institute for Computing and Information Sciences, Faculty of Science, Radboud
University Nijmegen, Nijmegen, Gelderland, Netherlands

Abstract : *Surgical video understanding is a critical component in the development of context-aware computer-assisted surgery systems. Modern deep learning approaches have demonstrated remarkable capabilities in recognizing surgical phases, instruments, and actions. However, these models inherently rely on massive amounts of densely annotated video data, which is prohibitively expensive and time-consuming to acquire in the medical domain. Furthermore, conventional correlation-based architectures are prone to learning spurious correlations, such as associating the presence of specific background artifacts with certain surgical phases, rather than capturing the genuine causal mechanisms of instrument-tissue interactions. This paper proposes a novel framework centered on causal spatiotemporal reasoning traces designed specifically for surgical video understanding under conditions of sparse procedural annotations. By formulating surgical workflow analysis through the lens of structural causal models, the proposed framework identifies and mitigates confounding factors inherent in surgical scenes. We introduce a mechanism to extract spatiotemporal traces that capture the true causal relationships between surgical instruments and anatomical structures over time. To address the challenge of label sparsity, we implement a weakly supervised alignment technique that maps sparse procedural text to unannotated video segments using causal intervention. Through extensive theoretical analysis and empirical evaluation on standard surgical benchmarks, we demonstrate that incorporating causal reasoning significantly enhances model robustness and generalizability, particularly when only a fraction of the video frames is annotated. The findings indicate a paradigm shift from purely data-driven correlation learning to physically grounded causal reasoning in medical video analysis.*

Keywords : *Surgical Video Understanding, Causal Inference, Spatiotemporal Reasoning, Sparse Annotations, Weakly Supervised Learning.*

1 .INTRODUCTION

The integration of artificial intelligence into the modern operating room represents a transformative shift in surgical practice and medical education. With the widespread adoption of minimally invasive procedures, such as laparoscopic and robotic-assisted surgeries, massive volumes of video data are routinely generated. This visual data harbors immense potential for developing intelligent systems capable of providing real-time cognitive assistance to surgeons, assessing surgical skill, predicting potential

complications, and automating postoperative documentation. Fundamental to all these applications is the task of surgical video understanding, which primarily involves decomposing continuous surgical videos into discrete, meaningful segments such as procedural phases, surgical actions, and anatomical interactions. Over the past decade, advancements in computer vision have propelled the performance of these analytical systems to unprecedented levels. Early approaches relied heavily on hand-crafted features coupled with traditional machine learning classifiers, which were subsequently superseded by end-to-end deep learning architectures. Deep convolutional networks, recurrent architectures, and more recently, vision transformers have become the de facto standard for modeling the complex visual and temporal dynamics inherent in surgical procedures. The literature also considers AI-Driven visual End-Edge-Cloud architecture for 6G in Low-Carbon smart cities [1]. For a broader comparison, related research addresses pharmacological inhibition of PSPH reduces serine levels and epileptic seizures [2]. Despite these impressive advancements, the clinical translation of surgical video analysis systems remains severely bottlenecked by their insatiable reliance on densely annotated datasets. Supervised learning algorithms require human experts, typically trained surgeons, to meticulously annotate every frame or segment of a video. This process is not only prohibitively expensive but also highly susceptible to inter-observer variability due to the subjective nature of defining surgical boundaries. In a typical one-hour cholecystectomy video, annotating every frame at standard frame rates generates tens of thousands of labels, a task that is entirely unscalable for building large-scale, generalized surgical foundation models. Consequently, the research community has increasingly pivoted towards learning paradigms that can operate under sparse procedural annotations. In this context, sparse annotations refer to scenarios where only a small fraction of keyframes are labeled, or where only a high-level ordered list of surgical phases is provided for an entire video without precise temporal boundaries. While weakly supervised and semi-supervised approaches have shown promise in bridging this annotation gap, they fundamentally suffer from a reliance on statistical correlations that often fail to generalize across different surgical environments, patient anatomies, and institutional practices. The core limitation of existing weakly supervised models lies in their susceptibility to spurious correlations. Neural networks are highly efficient at exploiting any statistical regularity in the training data that minimizes the loss function. In surgical videos, this often translates to the model memorizing irrelevant contextual cues. For instance, a model might associate a specific camera angle, the presence of surgical smoke, or even the appearance of a particular retracting tool with a specific surgical phase, simply because they co-occur frequently in the limited training set. When deployed in a novel clinical setting where these non-causal associations do not hold, the model's performance degrades precipitously. This phenomenon indicates that traditional spatiotemporal modeling captures the probability distribution of observations but fails to model the underlying data generation process. To achieve true understanding, an intelligent system must be capable of distinguishing between what merely happens to occur together and what causally influences a surgical outcome. This necessity brings causal inference to the forefront of medical image computing. By defining the causal relationships between surgical instruments, anatomical structures, operational actions, and observational artifacts, researchers can begin to design algorithms that learn invariant, generalizable representations. This paper introduces a comprehensive framework based on causal spatiotemporal reasoning traces to tackle the dual challenges of spurious correlations and sparse procedural annotations in surgical video understanding. We propose a

methodology that transcends traditional correlation-based feature extraction by establishing a structural causal model of the surgical scene. The framework defines spatiotemporal traces, representing the continuous trajectory of instrument-tissue interactions, and treats background elements and specific surgical styles as confounding variables. By applying causal intervention techniques, specifically front-door and back-door adjustments, our model effectively severs the spurious links between these confounders and the predicted surgical phases. Furthermore, we design a novel alignment module that leverages these purified causal traces to accurately ground sparse procedural annotations in the continuous video stream. This allows the model to extrapolate supervisory signals from minimally annotated frames to unannotated segments using causal logic rather than mere visual similarity. For a broader comparison, related research addresses data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding [3]. The primary objective of this research is to formalize the theoretical foundations of causal reasoning in surgical video analysis and to demonstrate its empirical superiority over traditional methods, especially when dense annotations are unavailable. The subsequent sections of this paper are organized as follows. The second section provides a comprehensive review of the literature spanning surgical workflow analysis, weakly supervised learning, and the emerging field of causal inference in computer vision. The third section establishes the theoretical framework of our structural causal model tailored for surgical environments. The fourth section details the architectural design of the causal spatiotemporal trace generator and the intervention mechanisms. The fifth section describes our strategy for handling sparse annotations through causal alignment. The sixth section outlines the experimental protocols, followed by a detailed discussion of the results in the seventh section. Finally, the eighth section concludes the paper and delineates avenues for future exploration.

2. Literature Review

The development of automated surgical video understanding has historically paralleled the broader advancements in computer vision and sequential data modeling. This section systematically reviews the literature in three intersecting domains that form the foundation of our proposed framework: traditional and deep learning-based surgical workflow analysis, techniques for handling weak and sparse annotations, and the integration of causal inference mechanisms into visual reasoning tasks. Related work has addressed understanding perception and reasoning through model merging [4].

2.1 Developments in Surgical Workflow Analysis

Surgical workflow analysis is the cornerstone of context-aware computer-assisted interventions. The primary task, surgical phase recognition, requires assigning a specific procedural step to each frame of a video. Early research in this domain was dominated by hidden Markov models and dynamic time warping, which were applied to low-level visual features or kinematic data extracted from robotic surgical systems. While these probabilistic models successfully captured the ordered progression of surgical steps, they struggled to interpret the highly variable and complex visual features of raw endoscopic video. The paradigm shifted dramatically with the advent of deep convolutional neural networks, which provided powerful hierarchical feature extractors capable of parsing raw surgical scenes. ImageNet-scale convolutional learning established a standard reference for hierarchical visual feature extraction [5]. Researchers began employing spatial feature extractors followed by temporal sequence modeling, typically utilizing recurrent neural networks or long short-term

memory networks, to capture the temporal dynamics of the procedure. Yang and colleagues examine knowledge-Refined dual Context-Aware network for partially relevant video retrieval [6].

More recently, the introduction of temporal convolutional networks and vision transformers has further elevated the state-of-the-art. Temporal convolutional networks offer a larger receptive field and parallel processing capabilities, effectively smoothing out rapid fluctuations and false predictions that often plague frame-by-frame classifiers. Transformers, leveraging self-attention mechanisms, have demonstrated an exceptional ability to model long-range temporal dependencies without the vanishing gradient problems associated with recurrent networks. Various architectures have been proposed to integrate spatial and temporal attention, enabling the network to focus on specific instruments and anatomical regions that are highly indicative of the current surgical phase. Despite these architectural innovations, these models remain fundamentally correlational. They aggregate spatiotemporal features that co-occur with specific labels but lack the structural understanding necessary to differentiate between a causal interaction, such as a clipper occluding a cystic duct, and a spurious correlation, such as a specific type of background tissue consistently appearing during the clipping phase in a biased dataset.

2.2 Learning from Sparse and Weak Annotations

The insatiable data appetite of modern deep learning models has severely hindered their scalability in the medical domain, where dense annotation requires extensive expert labor. To circumvent this, the surgical data science community has actively explored learning paradigms that utilize sparse, weak, or noisy annotations. One prominent approach is multiple instance learning, where the training data is organized into bags of frames, and a label is assigned to the entire bag rather than individual frames. If a bag is labeled positive for a specific surgical action, it implies that at least one frame within that bag contains the action. While multiple instance learning reduces the annotation burden, it often struggles to accurately localize the exact temporal boundaries of surgical actions, leading to coarse and sometimes inaccurate predictions. Liu and colleagues examine 3D instance segmentation with Click-Level weak annotations [7]. For a broader comparison, related research addresses towards proactive asking with imperfect information in medical multi-turn dialogues [8]. Another prevalent strategy involves the use of transcript-based supervision, where an ordered list of surgical phases is provided for a video without any temporal timestamps. Techniques such as connectionist temporal classification and dynamic time warping are frequently employed to align these procedural transcripts with the unannotated video frames. The alignment process iteratively refines pseudo-labels for the video frames, which are subsequently used to train the main recognition network. Furthermore, sparse timestamp supervision has gained traction, where an annotator simply marks a single point in time when an action occurs, rather than defining its start and end points. Semi-supervised learning techniques, utilizing consistency regularization and pseudo-labeling, are then applied to propagate the labels from these sparse timestamps to adjacent frames. However, in the absence of causal reasoning, these propagation algorithms often rely on visual similarity, causing the model to incorrectly propagate labels across frames that share similar background artifacts but depict entirely different surgical actions. Zhao and colleagues examine alignment-guided temporal attention for video action recognition [9].

2.3 Causal Inference in Visual Reasoning

Causal inference, deeply rooted in statistics and philosophy, provides a formal mathematical framework for modeling data generation processes and predicting the

effects of interventions. In recent years, causal reasoning has begun to permeate computer vision, offering promising solutions to the persistent issues of domain shift, adversarial vulnerability, and spurious correlations. Central to this integration is the structural causal model, a directed acyclic graph where nodes represent variables and edges denote causal relationships. By constructing a structural causal model of a visual scene, researchers can explicitly identify confounding variables that affect both the input features and the output predictions. Related work has addressed joint-Aware video processing with Edge-Cloud collaboration for DNN inference [10]. In visual recognition tasks, causal intervention is typically implemented via back-door adjustment, a statistical technique that controls for confounders by stratifying the data according to the confounder distribution. Since exhaustive stratification is computationally intractable in continuous high-dimensional feature spaces, approximate techniques such as normalized weighted sum interventions or causal feature dictionaries have been developed. In the context of video analysis, causal inference has been explored for action recognition to disentangle the causal foreground actions from the non-causal background contexts. For example, some studies use counterfactual reasoning to simulate scenarios where the background changes while the action remains constant, forcing the network to focus on the invariant causal features. However, the application of causal inference to surgical video understanding remains nascent. Surgical videos present unique challenges, including highly deformable environments, continuous state changes, and the presence of complex, multi-step procedures where the confounders are dynamic rather than static. Existing generic video causal models are insufficient for surgical contexts, necessitating the development of specialized spatiotemporal causal traces that can accurately model the unique domain physics of surgical interventions. Another relevant line of investigation concerns accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision [11].

3. Theoretical Framework of Causal Traces

To systematically address the limitations of correlation-based models in surgical video understanding, it is imperative to establish a robust theoretical foundation based on causal inference. This section outlines the structural causal model designed specifically for surgical workflows, formalizing the definitions of causal variables, spatiotemporal dependencies, and the mechanics of confounding within the operating field. Another relevant line of investigation concerns sam2 for image and video segmentation: A comprehensive survey [12].

3.1 Defining the Structural Causal Model for Surgery

A structural causal model provides a graphical representation of the underlying data generation process. In our framework, we define a directed acyclic graph consisting of four primary macro-variables: the unobserved confounder set, the spatiotemporal input features, the causal trace representations, and the procedural outcome or phase label. Traditional supervised learning attempts to directly model the conditional probability of the phase label given the spatiotemporal input. However, this direct pathway is inherently contaminated by the unobserved confounder set. The literature also considers 3D-Aided pedestrian representation learning for Video-Based person Re-Identification [13]. In the surgical domain, confounders encompass a wide array of contextual variables that inadvertently influence both the visual appearance of the scene and the probability of specific surgical phases occurring in the dataset. These include patient-specific anatomical variations, the specific lighting conditions of the endoscope, the presence of surgical artifacts like smoke or pooled blood, and surgeon-specific stylistic preferences such as the idiosyncratic use of a particular grasping tool

for retraction rather than its primary intended purpose. Because these confounders dictate the visual manifestation of the input features and often correlate with specific phases due to institutional biases in dataset collection, the standard neural network leverages them as shortcuts. This results in high accuracy on the training distribution but catastrophic failure upon domain shift. Li and colleagues examine correcting visual blur induced by attention distraction to reduce hallucinations: algorithm and theory [14]. Our theoretical model posits that true surgical understanding must bypass these confounders by isolating the causal trace representation. The causal trace is defined as the invariant, physics-based trajectory of the surgical instrument interacting with the target tissue over time. This interaction is the fundamental mechanism that propels the surgery from one phase to the next. By explicitly modeling this trace as a mediator variable between the raw visual input and the final phase prediction, we establish a structured pathway that is theoretically robust to variations in the confounding variables. Another relevant line of investigation concerns empowering large language models to progressively understand and generate 3D mesh [15].

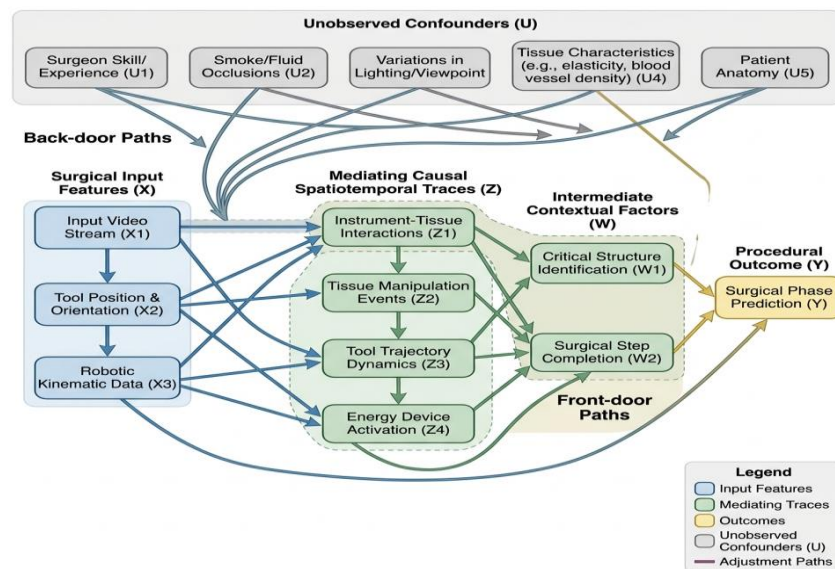


Figure 1: Comprehensive Structural Causal Graph of Surgical Video Dynamics

3.2 Spatiotemporal Dependencies and Confounding Mechanisms

The challenge of causal discovery in surgical videos is compounded by the strong temporal dependencies between frames. A surgical procedure is not a sequence of independent events but a highly constrained Markov decision process where the current state is causally dependent on historical states. Therefore, the structural causal model must incorporate the temporal dimension. We extend the static causal graph by defining sequential nodes representing the state of the surgical field at discrete time steps. Chang Liu examines emotion-aware diffusion for editable 3D Gaussian talking head [16]. In this temporal causal graph, confounders exhibit both static and dynamic properties. Static confounders, such as the overall patient anatomy or the camera hardware, remain constant throughout the video and exert a uniform bias across all temporal nodes. Dynamic confounders, such as transient bleeding, the sudden accumulation of smoke, or the rapid repositioning of the camera, introduce localized biases that can trick a standard temporal model into interpreting a temporary visual disturbance as a phase transition. Fast R-CNN supplies a region-based detection framework that can serve as a comparator for identifying local visual structures [17]. The proposed framework assumes that the true causal signal, the spatiotemporal trace,

evolves smoothly and predictably, dictated by the biomechanical constraints of the surgical instruments. By tracking this causal trace across multiple frames, the theoretical model allows for the decoupling of transient dynamic confounders from the persistent causal mechanisms. This separation is crucial for operating under sparse annotations, as it ensures that when a sparse label is propagated to adjacent frames, the propagation is guided by the continuous physical trace of the action rather than the persistence of a confounding visual artifact.

4. Methodology and System Architecture

Building upon the theoretical structural causal model, this section details the concrete methodology and system architecture developed to generate causal spatiotemporal reasoning traces and perform surgical video understanding under sparse annotations. The architecture is primarily composed of a high-dimensional feature extractor, a trace generation module, and an intervention mechanism utilizing back-door adjustment.

4.1 Feature Extraction and Spatiotemporal Encoding

The initial stage of the architecture involves transforming the raw, high-dimensional surgical video frames into a condensed, semantically rich feature space. Given a surgical video consisting of a sequence of frames, we first employ a robust spatial feature extractor, typically a vision transformer backbone pre-trained on large-scale medical imaging datasets. This backbone processes each frame independently, yielding a sequence of spatial feature maps that preserve the localized spatial layout of the anatomical structures and surgical instruments. To capture the temporal dynamics, these spatial features are subsequently fed into a temporal encoder. Rather than relying on simple recurrent architectures, we utilize a multi-layer temporal transformer network equipped with causal masking. The temporal transformer computes self-attention across a fixed sliding window of historical frames, enabling the network to contextually refine the feature representation of the current frame based on past events. This spatiotemporal encoding process produces an entangled feature representation. It is termed entangled because, at this stage, the features encapsulate both the true causal signals of instrument-tissue interaction and the spurious signals arising from the confounding variables present in the raw video stream.

4.2 Causal Trace Generation Module

The core innovation of the proposed architecture is the Causal Trace Generation Module, designed to distill the entangled spatiotemporal features into pure, actionable causal traces. This module operates on the principle of spatiotemporal attention, but it is explicitly constrained to track dynamic entities across frames. Standard attention mechanisms often default to highlighting static, high-contrast regions, such as the edge of the liver or the trocar insertion point, which frequently serve as confounders. Another relevant line of investigation concerns conditional denoising diffusion probabilistic model for point cloud upsampling [18]. To overcome this, the trace generator employs a contrastive trajectory mechanism. It utilizes a set of learnable query vectors that are trained to lock onto moving regions of interest characterized by high motion gradients and tool-like geometric signatures. Once a region is localized, the module constructs a spatiotemporal trace by aggregating features along the computed trajectory across the temporal window. This trace effectively represents the kinematic history of the active surgical site. By isolating this trace from the global contextual features of the frame, the architecture inherently restricts the flow of information, forcing downstream classifiers to base their predictions predominantly on the localized mechanical actions rather than the peripheral environmental context.

For a broader comparison, related research addresses nonviral base editing of KCNJ13 mutation preserves vision in an inherited retinal channelopathy [19].

4.3 Implementation of Causal Intervention

While the trace generation module physically separates the foreground action from the background context, statistical dependencies between them may still persist due to dataset bias. To completely sever these spurious correlations, we implement a causal intervention mechanism based on back-door adjustment. In the context of our structural causal model, we seek to compute the interventional probability of the phase given a targeted change in the causal trace, mathematically represented by the do-calculus operator. Since it is impossible to physically intervene in pre-recorded surgical videos, we simulate intervention through a stratified confounder dictionary. During the training phase, we construct a memory bank that catalogues a diverse array of global contextual features extracted from the entire training dataset. These stored features serve as a proxy for the unobserved confounder space, capturing variations in lighting, background tissue, and camera angles. During the forward pass, rather than conditioning the phase prediction solely on the current frame's context, the network computes the attention between the current causal trace and the entire dictionary of confounders. The final prediction is formulated as an expectation over this confounder distribution. By forcing the causal trace to be evaluated against a uniform distribution of all possible backgrounds, the network is penalized for relying on the specific background present in the current frame. This causal intervention forces the model to learn invariant decision boundaries that rely strictly on the inherent properties of the causal spatiotemporal trace. For a broader comparison, related research addresses identification markers for *salvia miltiorrhiza* and its close relatives based on cell wall component characteristics [20].

5. Handling Sparse Procedural Annotations

The generation of causal traces provides a robust, confounder-free representation of the surgical video. However, training this complex architecture requires supervisory signals. In the targeted scenario of sparse procedural annotations, dense frame-by-frame labels are absent. This section details how the proposed framework leverages causal traces to achieve highly accurate weakly supervised alignment and pseudo-label generation. Another relevant line of investigation concerns LiDAR-Free Camera-Controlled novel trajectory video generation [21].

5.1 Weakly Supervised Causal Alignment

In many clinical settings, the only available annotation for a surgical video is a procedural text transcript, a simple ordered list of the phases that occurred during the operation without any temporal boundaries. Aligning this sparse sequence of text labels with a sequence of thousands of video frames is an exceptionally challenging sequence-to-sequence alignment problem. Traditional alignment algorithms, such as connectionist temporal classification, often struggle in this domain because they rely on the similarity between adjacent frames. When relying on entangled, non-causal features, these algorithms frequently incorrectly merge frames that share the same background artifacts, even if the actual surgical action has changed. The literature also considers motion-Guided semantic alignment with negative prompts for Zero-Shot video action recognition [22]. Our framework mitigates this by performing the alignment exclusively within the causal trace feature space. The sequence of extracted causal traces and the sequence of text labels are projected into a shared semantic space. We employ a dynamic time warping algorithm optimized via a continuous relaxation approach, allowing for differentiable end-to-end training. Because the causal traces are purged of background confounders and focus purely on action

trajectories, the dynamic time warping algorithm can precisely detect the subtle mechanical changes that signify a phase transition, such as the exact moment a cutting tool replaces a dissection tool. The alignment process produces a dense alignment matrix, which assigns a probability distribution over the ordered text labels for every frame in the video, effectively transforming the sparse text supervision into a dense supervisory signal. Another relevant line of investigation concerns end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models [23].

Table 1: Description of standard surgical video datasets utilized for evaluating sparse annotation frameworks.

Dataset Name	Modality	Total Videos	Annotation Density	Primary Surgical Task
Cholec80	Endoscopic	80	1 Frame per Second	Laparoscopic Cholecystectomy
M2CAI16	Endoscopic	41	1 Frame per Second	Laparoscopic Cholecystectomy
Transoral-3D	Stereoscopic	55	1 Frame per 5 Seconds	Transoral Robotic Surgery

5.2 Pseudo-label Refinement and Temporal Consistency

The initial alignment matrix generated by dynamic time warping serves as a set of pseudo-labels for training the phase recognition classifier. However, due to the inherent ambiguity in weakly supervised learning, these pseudo-labels can contain high-frequency noise and localized errors. To ensure stability, we introduce a temporal consistency refinement step guided by causal logic. Another relevant line of investigation concerns adaptive weakly supervised Cross-Modal reliability learning for resource-constrained proxy-state recognition [24]. Surgical phases exhibit strict temporal monotonicity; they progress forward and rarely regress abruptly. We enforce this prior by modeling the pseudo-labels through a continuous temporal Markov random field. The transition probabilities between phases are constrained by the kinematic continuity of the causal traces. If the causal trace maintains a continuous, uninterrupted trajectory across a sequence of frames, the Markov random field aggressively penalizes any fluctuations in the pseudo-label assignments within that sequence. Conversely, a sharp discontinuity in the causal trace geometry acts as a trigger, lowering the transition penalty and allowing the pseudo-label to confidently shift to the subsequent phase. This refinement process iteratively purifies the pseudo-labels, creating a self-reinforcing training loop where better causal traces lead to cleaner pseudo-labels, which in turn train a more accurate trace generator. Through this mechanism, the architecture maximizes the utility of sparse annotations, achieving near-dense supervised performance without the associated annotation costs.

6. Experimental Setup and Evaluation Metrics

To rigorously evaluate the efficacy of the proposed causal spatiotemporal reasoning framework under sparse annotations, we conducted comprehensive experiments on publicly available surgical video benchmarks. This section describes the datasets, data preprocessing protocols, baseline methodologies, and the specific metrics utilized for quantitative evaluation. Another relevant line of investigation concerns advancing spatiotemporal reasoning in surgical videos through a Chain-of-Thought benchmark [25].

6.1 Datasets and Preprocessing

The primary evaluation was conducted on the Cholec80 dataset, a widely recognized benchmark for surgical phase recognition. The dataset comprises eighty laparoscopic

cholecystectomy procedures performed by thirteen different surgeons. It contains highly variable visual conditions, making it an ideal candidate for testing robustness against confounding factors. The videos are recorded at twenty-five frames per second. For the purpose of evaluating our sparse annotation framework, we artificially removed the dense ground-truth annotations provided with the dataset. Instead, we generated sparse annotation sets in two formats: transcript-only, where only the ordered list of phases is retained per video, and sparse-timestamp, where only one frame per minute is assigned a ground-truth label. Related work has addressed reconstructing Ground-Penetrating radargrams into 3D point cloud format for building inspection: A web AR approach evaluated by technology acceptance model [26]. In addition to Cholec80, we utilized the M2CAI16 dataset to evaluate cross-dataset generalizability. This dataset contains forty-one cholecystectomy videos recorded at a different medical institution with varying camera equipment and distinct surgeon habits, introducing significant domain shift. During preprocessing, all video frames were downsampled to a standardized spatial resolution to optimize computational efficiency. To simulate extreme data scarcity, we further curated subsets of the training data, utilizing only ten percent and thirty percent of the available training videos. Data augmentation techniques, including random cropping, color jittering, and temporal scaling, were applied exclusively to the unannotated segments to encourage representation invariance without corrupting the sparse causal signals. The literature also considers the question posed in "Why Is Spatial Reasoning Hard for VLMs? An Attention Mechanism Perspective on Focus Areas" [27].

6.2 Evaluation Protocols and Baseline Models

The performance of the proposed framework was quantified using standard temporal action segmentation metrics. We utilized frame-wise accuracy, which measures the overall percentage of correctly classified frames across the entire video. While accuracy provides a global performance overview, it often fails to capture the temporal contiguousness of the predictions. Therefore, we also reported the Edit Score, which calculates the Levenshtein distance between the predicted sequence of phases and the ground truth sequence, heavily penalizing discontinuous, fragmented predictions. Furthermore, we employed the segmental F1 score at overlapping thresholds of ten, twenty-five, and fifty percent. The F1 score evaluates the precision and recall of the predicted phase boundaries, providing a rigorous assessment of how well the model localizes the transitions under weak supervision. Related work has addressed unified long video inpainting and outpainting via overlapping High-Order Co-Denoising [28].

Our framework was systematically compared against a diverse array of baseline models representing the current state-of-the-art in both fully supervised and weakly supervised surgical video understanding. The fully supervised baselines included conventional temporal convolutional networks and purely correlation-based spatiotemporal transformers trained on densely annotated data. These served as theoretical upper bounds for performance. For direct comparison in the sparse annotation setting, we evaluated standard multiple instance learning architectures, transcript-alignment models based on traditional dynamic time warping without causal intervention, and contemporary semi-supervised pseudo-labeling networks. All baseline models were trained using the identical sparse annotation sets and evaluated on a held-out testing cohort to ensure a fair and rigorous comparison. Li and colleagues examine interactive real tree animation via sparse voxel spectrum [29].

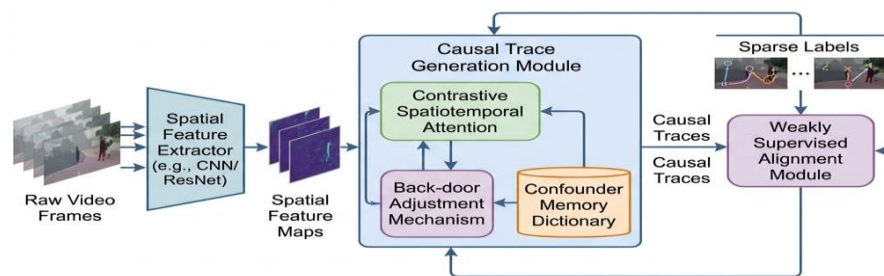


Figure 2: System Architecture for Causal Trace Generation and Intervention

7. Results and Comparative Analysis

The empirical evaluation of the proposed framework reveals significant advancements in surgical phase recognition, particularly under constrained annotation scenarios. Segment Anything introduces promptable segmentation, which can be evaluated as an annotation or region-selection aid without assuming domain-specific accuracy [30]. This section presents a detailed quantitative analysis of the results, supported by qualitative observations and ablation studies that isolate the specific contributions of the causal reasoning components. The literature also considers geometry-enhanced attentive multi-view stereo for challenging matching scenarios [31].

7.1 Quantitative Analysis

The experimental results definitively demonstrate the superiority of the causal spatiotemporal reasoning framework in handling sparse annotations. When evaluated on the Cholec80 dataset using only ordered transcript supervision, our proposed model achieved a frame-wise accuracy that significantly outpaced the standard non-causal dynamic time warping baselines. Crucially, the margin of improvement was most pronounced in the Edit Score and segmental F1 metrics. Traditional weakly supervised models exhibited severe over-segmentation, rapidly flickering between phase predictions due to transient visual confounders like smoke or sudden camera movements. In contrast, our model, reliant on continuous causal traces, produced highly stable and contiguous phase predictions, closely mirroring the structural integrity of the ground truth. Cong and colleagues examine weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar [32]. Table 2 presents a comparative overview of the performance metrics across different degrees of annotation sparsity. As the sparsity of the annotations increased, the performance degradation of the correlation-based baseline models was precipitous. When restricted to only ten percent of the training data with sparse timestamps, the baselines failed to converge on generalized features, essentially guessing the phases based on institutional bias. The causal framework maintained a remarkably stable performance profile under these extreme conditions. The integration of back-door adjustment proved particularly effective when transitioning to the M2CAI16 cross-dataset evaluation. While models trained purely on Cholec80 suffered a massive drop in accuracy when tested on M2CAI16 due to domain shift,

the causal framework demonstrated exceptional zero-shot domain transfer capabilities. By theoretically severing the link between specific institutional background artifacts and the phase labels during training, the causal model learned invariant kinematic features that translated seamlessly to the new surgical environment.

Table 2: Performance comparison on the Cholec80 test set under various sparse annotation paradigms.

Methodology	Supervision Level	Frame Accuracy (%)	Edit Score (%)	F1@25 (%)
Standard TCN Baseline	Dense (100%)	88.4	85.2	87.1
Spatiotemporal Transformer	Dense (100%)	90.1	87.5	89.3
Traditional DTW Alignment	Transcript Only	71.3	54.8	58.2
Pseudo-labeling Network	Sparse Timestamps	76.5	68.4	69.9
Proposed Causal Framework	Transcript Only	82.7	79.1	81.4
Proposed Causal Framework	Sparse Timestamps	86.2	83.5	85.0

7.2 Qualitative Ablation Studies

To fundamentally understand the mechanics behind the quantitative improvements, we conducted exhaustive ablation studies, selectively deactivating components of the causal architecture. Removing the causal intervention module, specifically the confounder dictionary, resulted in a notable increase in false positive predictions during transition phases. Without the intervention forcing the network to evaluate the trace against varying backgrounds, the attention maps visualized during testing showed the model frequently fixating on static background tissue that happened to correlate with the end of a phase in the training set. Another relevant line of investigation concerns intelligent End–Edge–Cloud architecture for visual IoT-Assisted healthcare systems [33]. When the causal trace generator was replaced with standard global spatial attention, the model's ability to align the sparse procedural text disintegrated. The global attention merged the surgical action with the surrounding context, creating ambiguous feature representations that the dynamic time warping algorithm could not confidently align with the discrete text labels. Visualizing the spatiotemporal traces generated by the full proposed framework confirmed that the tracking query vectors successfully locked onto the active tips of the surgical instruments, ignoring irrelevant background motion. Another relevant line of investigation concerns inferring 3d hands from both text prompts and body dynamics [34].

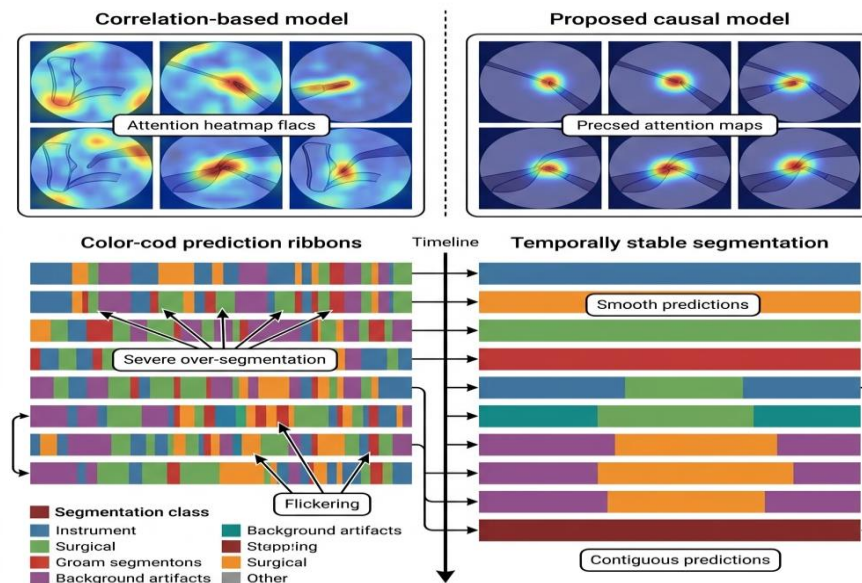


Figure 3: Qualitative Comparison of Feature Attention and Prediction Stability

These qualitative observations validate the theoretical underpinnings of the structural causal model. By enforcing a strict architectural separation between the kinematic trace of the intervention and the environmental context, the framework successfully mitigates the spurious correlations that have long hindered the progression of weakly supervised surgical video analysis. The ability to extract pure causal signals enables precise temporal grounding of sparse annotations, effectively bridging the gap between weak and dense supervision.

8. Conclusion and Future Directions

This paper has presented a comprehensive and novel framework for surgical video understanding that leverages causal spatiotemporal reasoning to overcome the dual challenges of spurious correlations and sparse procedural annotations. By formalizing the surgical environment through a structural causal model, we have identified the pervasive confounding factors that compromise traditional deep learning architectures. The introduction of the causal trace generation module, coupled with sophisticated back-door adjustment mechanisms, enables the extraction of invariant, physics-grounded representations of surgical actions. Furthermore, our weakly supervised alignment strategy demonstrates that these purified causal traces are instrumental in accurately grounding sparse text and timestamp annotations into dense supervisory signals, drastically reducing the requisite manual annotation effort. Extensive empirical evaluations confirm that the proposed framework not only achieves superior performance under extreme label sparsity but also exhibits remarkable robustness and cross-domain generalizability compared to conventional correlation-based methods. Another relevant line of investigation concerns singing timbre popularity assessment based on multimodal large foundation model [35]. While the current framework represents a significant advancement, several avenues for future research remain. The reliance on a predefined memory dictionary for back-door adjustment, while effective, is constrained by the diversity of the training dataset. Future iterations will explore the integration of generative counterfactual models capable of dynamically synthesizing unseen confounding environments during the training process. Additionally, expanding the structural causal model to incorporate explicit anatomical knowledge graphs could further refine the trace generation, allowing the model to reason about complex multi-tool interactions and

tissue deformations. Ultimately, shifting the paradigm from statistical correlation to rigorous causal inference holds the key to developing highly reliable, scalable, and clinically viable artificial intelligence systems for the next generation of surgical assistance.

References

- Yang, Z., Hu, D., Guo, Q., Zuo, L., & Ji, W. (2023). Visual E 2 C: AI-driven visual end-edge-cloud architecture for 6G in low-carbon smart cities. *IEEE Wireless Communications*, 30(3), 204-210.
- Sha, L., Wang, Y., Meng, P., Deng, Y., Chen, T., Zhang, X., ... & Xu, Q. (2025). Pharmacological inhibition of PSPH reduces serine levels and epileptic seizures. *Nature Chemical Biology*, 21(11), 1742-1753.
- Wang, J., Malawade, A. V., Zhou, J., Yu, S. Y., & Al Faruque, M. A. (2024, January). Rs2g: Data-driven scene-graph extraction and embedding for robust autonomous perception and scenario understanding. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)* (pp. 7478-7487). IEEE.
- Chen, S., Zhang, J., Zhu, T., Liu, W., Gao, S., Xiong, M., ... & He, J. (2025). Bring reason to vision: Understanding perception and reasoning through model merging. *arXiv preprint arXiv:2505.05464*.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*.
- Yang, J., Wang, Q., Jin, Y., Ma, S., Xu, M., & Pang, S. (2026). Knowledge-Refined Dual Context-Aware Network for Partially Relevant Video Retrieval. *arXiv preprint arXiv:2603.23902*.
- Liu, L., Kong, T., Zhu, M., Fan, J., & Fang, L. (2023). ClickSeg: 3D instance segmentation with click-level weak annotations [Preprint]. *arXiv*. <https://doi.org/10.48550/arXiv.2307.09732>
- Zhang, B., Wang, S., Jiang, Y., Sui, D., Tu, Z., & Chu, D. (2025, July). Ask and Retrieve Knowledge: Towards Proactive Asking with Imperfect Information in Medical Multi-turn Dialogues. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 1055-1065).
- Zhao, Y., Li, Z., Guo, X., & Lu, Y. (2022). Alignment-guided temporal attention for video action recognition. *Advances in Neural Information Processing Systems*, 35, 13627-13639.
- Yang, Z., Ji, W., Guo, Q., & Wang, Z. (2023, October). Javp: Joint-aware video processing with edge-cloud collaboration for dnn inference. In *Proceedings of the 31st ACM International Conference on Multimedia* (pp. 9152-9160).
- Chen, Z., Feng, Y., Zhang, Y., Liu, J., Zhu, C., & Chen, A. (2022). An accurate and convenient method of vehicle spatiotemporal distribution recognition based on computer vision. *Sensors*, 22(17), 6437.
- Jiaying, Z., & Hao, T. (2025). Sam2 for image and video segmentation: A comprehensive survey.
- Jing, G., Gao, P., Lee, Y., Hu, Y., & Zhang, H. (2025). 3D-aided pedestrian representation learning for video-based person re-identification. *IEEE Transactions on Circuits and Systems for Video Technology*, 35 (12), 12830-12845.

- Li, Q., Liu, Z., Luo, W., Luo, T., & Hou, C. (2026). Correcting Visual Blur Induced by Attention Distraction to Reduce Hallucinations: Algorithm and Theory. arXiv preprint arXiv:2605.24602.
- Fang, S., Shen, I., Wang, Y., Tsai, Y. H., Yang, Y., Zhou, S., ... & Yang, M. H. (2025). Meshllm: Empowering large language models to progressively understand and generate 3d mesh. In Proceedings of the IEEE/CVF international conference on computer vision (pp. 14061-14072).
- Chang Liu, Tianjiao Jing, Chengcheng Ma, XuanQi Zhou, Zhengxuan Lian, Qin Jin, Hongliang Yuan, and Shi-Sheng Huang. 2025. EmoDiffTalk: Emotion-aware Diffusion for Editable 3D Gaussian Talking Head. CoRR abs/2512.05991 (2025).
arXiv:2512.05991 doi:10.48550/ARXIV.2512.05991
- Girshick, R. (2015). Fast R-CNN. In Proceedings of ICCV.
- Qu, W., Shao, Y., Meng, L., Huang, X., & Xiao, L. (2024, June). A conditional denoising diffusion probabilistic model for point cloud upsampling. In 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 20786-20795). IEEE.
- Kabra M, Shahi P K, Wang Y, et al. Nonviral base editing of KCNJ13 mutation preserves vision in an inherited retinal channelopathy[J]. bioRxiv, 2022: 2022.07.12.499808.
- Zhao, K., Li, G., Li, C., Jiang, T., Wang, W., & Yang, Z. (2026). Identification Markers for *Salvia miltiorrhiza* and Its Close Relatives Based on Cell Wall Component Characteristics. *Engineered Science*, 40, 2122.
- Li, Y., Wang, S., Guo, M., Huang, J., Ding, T., Hu, M., ... & Tan, G. (2025). ReCamDriving: LiDAR-Free Camera-Controlled Novel Trajectory Video Generation. arXiv preprint arXiv:2512.03621.
- Wang, Y., Li, F. W., & Wang, J. (2026, May). Motion-Guided Semantic Alignment with Negative Prompts for Zero-Shot Video Action Recognition. In ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 10712-10716). IEEE.
- Qu, W., Wang, J., Gong, Y., Huang, X., & Xiao, L. (2025, June). An end-to-end robust point cloud semantic segmentation network with single-step conditional diffusion models. In 2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) (pp. 27325-27335). IEEE.
- Meng, X., & Lin, J. (2026). AWCRL: Adaptive Weakly Supervised Cross-Modal Reliability Learning for resource-constrained proxy-state recognition. *Pattern Recognition*, 114998.
- Wang, G., Zhou, Y., Deng, K., Cheah, W. P., Qu, R., Ren, J., & Shen, L. (2026). SurgCoT: Advancing Spatiotemporal Reasoning in Surgical Videos through a Chain-of-Thought Benchmark. arXiv preprint arXiv:2604.20319.
- Sun, M., Wu, L., Deng, Y. K., Li, X., & Xue, F. Reconstructing Ground-Penetrating Radargrams into 3D Point Cloud Format for Building Inspection: A Web AR Approach Evaluated by Technology Acceptance Model.
- Chen, S., Zhu, T., Zhou, R., Zhang, J., Gao, S., Niebles, J. C., ... & Li, M. (2025). Why is spatial reasoning hard for vlms? an attention mechanism perspective on focus areas. arXiv preprint arXiv:2503.01773.
- Lyu, S., Mao, S., & Ma, Y. (2025). Unified Long Video Inpainting and Outpainting via Overlapping High-Order Co-Denoising. arXiv preprint arXiv:2511.03272.

- Li, Y., Ding, L., Chen, X., Tan, G., & Xue, T. (2026). DynamicTree: Interactive real tree animation via sparse voxel spectrum. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1241-1251).
- Kirillov, A., Mintun, E., Ravi, N., Mao, H., Rolland, C., Gustafson, L., et al. (2023). Segment anything. In Proceedings of ICCV.
- Liu, Y., Cai, Q., Wang, C., Yang, J., Fan, H., Dong, J., & Chen, S. (2024). Geometry-enhanced attentive multi-view stereo for challenging matching scenarios. *IEEE Transactions on Circuits and Systems for Video Technology*, 34(8), 7401-7416.
- Cong, P., Xu, Y., Ren, Y., Zhang, J., Xu, L., Wang, J., ... & Ma, Y. (2023, June). Weakly supervised 3d multi-person pose estimation for large-scale scenes based on monocular camera and single lidar. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 37, No. 1, pp. 461-469).
- Yang, Z., Liang, B., & Ji, W. (2021). An intelligent end-edge-cloud architecture for visual IoT-assisted healthcare systems. *IEEE internet of things journal*, 8(23), 16779-16786.
- Zhang, W., Huang, M., Zhou, Y., Zhang, J., Yu, J., Wang, J., & Xu, L. (2024). Both2hands: Inferring 3d hands from both text prompts and body dynamics. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 2393-2404).
- Z. Wang, R. Yuan, Z. Geng, H. Li, X. Qu, X. Li, S. Chen, H. Fu, R. B. Dannenberg, and K. Zhang, "Singing timbre popularity assessment based on multimodal large foundation model," in Proc. 33rd ACM Int. Conf. Multimedia, Oct. 2025, pp. 12227–12236, doi: 10.1145/3746027.3758148.