Introducing Decomposed Causality with Spatiotemporal Object-Centric Representation for Video Classification

Yachong Zhang[1], **Lei Meng**[1*], **Shuo Xu**[1], **Zhuang Qi**[1], **Wei Wu**[1], **Lei Wu**[1], **Xiangxu Meng**[1]

[1]School of Software, Shandong University, Jinan, China
{zhangyachong, shuo.xu, z_qi, wu_wei}@mail.sdu.edu.cn, {lmeng, i_lily, mxx}@sdu.edu.cn

Abstract

Video classification requires event-level representations of objects and their interactions. Existing methods typically rely on data-driven approaches, which either learn such features from whole frames or object-centric visual regions. Therefore, the modeling of spatiotemporal interactions among objects is usually overlooked. To address this issue, this paper presents a Decomposition of Synergistic, Unique, and Redundant Causal Representations Learning (SurdCRL) model for video classification, which introduces a newly-proposed SURD causal theory to model the spatiotemporal features of both object dynamics and their in- and cross-frame interactions. Specifically, SurdCRL employs three modules to model the object-centric spatiotemporal dynamics using distinct types of causal components, where the first module **Spatial-Temporal Entity Modeling** decouples the frame into object and context entities, and employs a temporal message passing block to capture object state changes over time, generating spatiotemporal features as basic causal variables. Second, the **Dual-Path Causal Inference** module mitigates confounders among causal variables by front-door and back-door interventions, thus enabling the subsequent causal components to reflect their intrinsic effects. Finally, the **Causal Composition and Selection** module employs the compositional structure-aware attention to project the causal variables and their high-order interactions into the synergistic, unique, and redundant components. Experiments on two benchmarking datasets verify that SurdCRL better captures event-relevant object-centric representation by decomposing spatiotemporal object interactions into three types of causal components.

Introduction

With the continuous advancement of video representation learning techniques, such as spatiotemporal convolutional networks and Video Transformers, video classification has achieved remarkable progress. Unlike a mere collection of static images, video captures rich temporal dynamics and evolving semantic interactions. To better model such complexities, recent research has gradually shifted from scene-level modeling to a more fine-grained, object-centric perspective, focusing on how different entities interact over time. However, this modeling process often faces challenges such as visual ambiguity and contextual interference, which make models more prone to focusing on patterns that are irrelevant to video understanding. These limitations hinder the development of representations that are both discriminative and generalizable. Thus, effectively identifying object-centric representation from complex spatiotemporal dynamics remains a critical challenge in video classification.

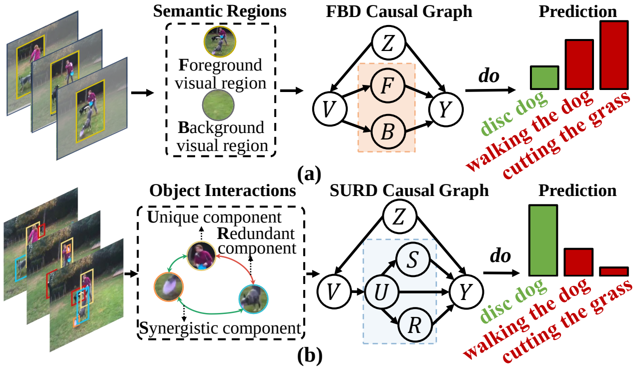

Figure 1: By introducing SURD theory, SurdCRL integrates object-centric causal representation. The conventional causal graph (a) focuses on foreground-background decomposition. In contrast, the surd causal graph (b) captures object interactions as distinct components and mitigate the bias caused by confounders Z to enable precise inference of Y.

To better represent the spatiotemporal dynamics in videos, existing methods can be broadly categorized into three types. The first leverages 3D convolutions or spatiotemporal attention to globally model frame dependencies, but often overlooks key semantic objects and their interactions. The second type shifts toward object-centric models (Liu et al. 2025), where semantic entities are extracted using external detectors, alleviating semantic dilution in global encoding. However, due to data biases introduced by external detectors and spurious correlations from statistical bias in the training data, the discriminative capability of the model is inevitably compromised. The third type incorporates causal inference (Pearl et al. 2016) to uncover structural dependencies and eliminate confounders. For instance, counterfactual reasoning and causal intervention have been applied to reduce bias in multimodal tasks like VQA (Antol

[*]Corresponding Author

et al. 2015). However, most of these methods rely on external textual knowledge. Although recent studies have explored causal modeling in unimodal video understanding, as shown in Figure 1(a), they typically construct a causal graph by decomposing frames into foreground and background, which results in limited semantic expressiveness.

The recently proposed SURD causal theory (Martínez-Sánchez, Arranz, and Lozano-Durán 2024) decomposes causality into synergistic, unique, and redundant components. Inspired by this theory, this paper presents a Decomposition of Synergistic, Unique, and Redundant Causal Representations Learning (SurdCRL) method for video classification. As shown in Figure 1(b), SurdCRL decomposes spatiotemporal object-centric interactions into distinct causal components and applies causal inference by means of the do-operator in the early modeling stage to suppress spurious patterns induced by confounders. Specifically, SurdCRL first employs the Spatial-Temporal Entity Modeling (STEM) module to disentangle each frame into object and context entities and model their temporal transitions, generating spatiotemporal features that serve as basic causal nodes in the structural causal model. Next, the Dual-Path Causal Inference (DPCI) module leverages these features from STEM and applies front-door and back-door interventions to block confounding paths toward object and context nodes, mitigating data biases from unobservable confounders and suppressing spurious correlations caused by observable background noise. Finally, the Causal Composition and Selection (CCS) module models spatiotemporal interactions among entities using structure-aware attention and a CDF-based sampling strategy, projecting the object-centric representation into synergistic, unique, and redundant components.

To validate the effectiveness of SurdCRL, we conduct extensive experiments on two benchmark datasets, including performance comparison, ablation studies, in-depth analysis, and case studies. The results show that SurdCRL effectively models object representations by incorporating causal theory. In summary, the main contributions of this paper are:

- This paper presents a SurdCRL causal model based on SURD theory which decomposes causality into synergistic, unique, and redundant components. To the best of our knowledge, this is the first causal model that models the object-centric representation in video classification.
- SurdCRL mitigates bias from confounders among causal variables via causal inference, enabling better modeling of high-order interactions and allowing each causal component to contribute its true effect to classification.
- Experimental results verify that SurdCRL refines the structural causal model via decomposed causality, enabling the model to focus on event-relevant object-centric interactions and achieve precise video understanding.

Related Work

Video Classification Video classification has been extensively explored through various architectural paradigms. Early CNN-based methods such as SlowFast (Feichtenhofer et al. 2019), X3D (Feichtenhofer 2020), and TSM (Lin, Gan, and Han 2019) utilize 2D/3D convolutions with multi-path or shift mechanisms for better temporal modeling. Transformer-based models like ViViT (Arnab et al. 2021), VideoMAE-v2 (Wang et al. 2023), InternVideo-v2 (Wang et al. 2024b) and Uniformerv2 (Li et al. 2023a) further enhance long-range temporal modeling via attention mechanisms, masked pretraining, and hybrid designs. More recently, state-space models such as VideoMamba (Li et al. 2024) provide efficient solutions for modeling temporal dependencies with lower computational cost. However, these models often overlook confounding biases and fail to capture fine-grained object-centric representation.

Object-Centric Video Representation Learning Object-centric video representation learning focuses on constructing robust object-level features for tasks like action localization and recognition (Xu et al. 2023; Qu et al. 2025). Early works (Zhang et al. 2019; Materzynska et al. 2020) enhance CNNs with tracking or spatial interaction modules. Transformer-based models such as ORViT (Herzig et al. 2022) and DAIR (Li et al. 2025a) introduce object-aware attention for temporal reasoning, while others leverage trajectories (Zhang et al. 2024), vision-language pretraining (Zhang et al. 2023), or slot attention (Qian, Ding, and Lin 2024; Didolkar et al. 2025) to enable scalable object discovery without dense labels. However, most methods depend on external detectors or priors, making them sensitive to detection noise and spurious correlations. Recent works (Li et al. 2023b; Huang et al. 2025) attempt to reduce appearance and background bias as well as cross-modal redundancy through spatial priors and adaptive sampling, but still lack explicit causal modeling to systematically mitigate such biases.

Causal Inference in Video Understanding Compared to traditional debiasing techniques (Wang et al. 2020; Qi et al. 2023), causal inference has shown promise in reducing spurious correlations and disentangling model effects (Wang et al. 2022a,b; Qi et al. 2025b; Meng et al. 2025). In multimodal video tasks, such as Video Question Answering, counterfactual reasoning (Niu et al. 2021; Vosoughi et al. 2024; Wang et al. 2025) and interventions (Liu, Li, and Lin 2023; Chen et al. 2025, 2024; Wang et al. 2024a) have been applied to alleviate vision-language bias and capture cross-modal structures. However, most methods rely on textual supervision. Though recent efforts (Wang et al. 2024c; Liu et al. 2024) explore causal modeling in purely visual modalities, they often construct causal graphs through coarse foreground, background, and motion decomposition, limiting semantic expressiveness and interaction understanding.

Problem Formulation

This study investigates feature representation in video classification. Given a dataset $\mathcal{D} = \{V_i \mid i = 1, \dots, N\}$ with labels $Y = \{y_i \mid i = 1, \dots, J\}$, conventional methods extract frame-level features $F_v = \mathcal{M}(V)$ using a holistic encoder $\mathcal{M}$, and predict labels via category mapping $\mathcal{P}(F_v) \rightarrow Y$. Object-Centric Representation Learning (OCRL) methods decompose frames into object patches $O = \{O_i \mid O_1, O_2, \dots, O_K\}$ and a context patch C, and extract features F_{oc} via an object-centric encoder $\mathcal{B}(\cdot)$:

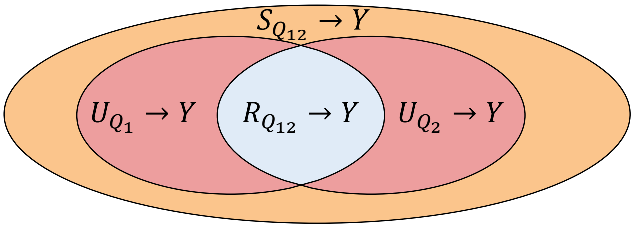

Figure 2: Diagram of the decomposition of causal dependencies between a vector of observed variables $Q = \{Q1, Q2\}$ and a target variable Y into their synergistic (S), unique (U) and redundant (R) components.

$F_{oc} = \mathcal{B}(O, C)$. The above methods model the likelihood $P(Y \mid V)$. In contrast, causal representation learning methods introduce the do-operator and extract deconfounded features $F_s = \mathcal{I}(\mathrm{do}(V))$, where $\mathcal{I}$ is a structural causal model under the distribution $P(Y \mid \mathrm{do}(V))$.

Our proposed SurdCRL is inspired by the SURD causal theory, which decomposes causal effects into synergistic, unique, and redundant components. We first extract object O and context C entities using a tracker, their features are then obtained as $F_o = \mathcal{T}_o(O)$ and $F_c = \mathcal{T}_c(C)$, where $\mathcal{T}_o$ and $\mathcal{T}_c$ denote the spatial-temporal feature extraction network. To mitigate the effect of confounders, we extend beyond conventional OCRL methods by applying front-door $\mathcal{FI}(\cdot)$ and back-door $\mathcal{BI}(\cdot)$ interventions to F_o and F_c respectively. Further extending beyond simple foreground-background causal decoupling, we introduce a composition and sampling module $\mathcal{H}(\cdot)$ to model entity interactions and derive the synergistic, unique and redundant causal components: $F_s, F_u, F_r = \mathcal{H}(F_o, F_c)$. These components are fused for the final prediction: $\mathcal{P}(F_u, F_s, F_r) \rightarrow Y$.

Preliminary on SURD Causal Theory

The SURD theory (Martínez-Sánchez, Arranz, and Lozano-Durán 2024) decomposes causal effects among observed variables $Q = \{Q_1, Q_2, \ldots, Q_N\}$ on a future outcome Q_j^+ into three components: Unique causality from Q_i to Q_j^+ that cannot be obtained from any other individual variable $Q_k \neq Q_i$. Redundant causality from $\mathbf{Q_i} = \{Q_{i_1}, Q_{i_2}, \ldots\}$ to Q_j^+ refers to the shared causal influence collectively contributed by all elements in the subset $\mathbf{Q_i} \subseteq Q$. Synergistic causality from $\mathbf{Q_i} = \{Q_{i_1}, Q_{i_2}, \ldots\}$ to Q_j^+ arises from the joint effect of the variables in $\mathbf{Q_i}$.

Although the SURD theory provides a principled way to decompose causal effects, it is not directly applicable to video classification due to the lack of clearly defined variables. To bridge this gap, our study formalizes this theory and creates a trainable model. Specifically, as illustrated in Figure 2, we provide the following definitions: (1) **Observed variables** Q: Q denotes the object-context patches extracted from input video V. (2) **Target variable** Q_j^+: Q_j^+ denotes the prediction logits Y. (3) **Unique component** U: $U = \{U_{Q_1}, U_{Q_2}, \ldots\}$, where each $U_{Q_i} = T(Q_i)$ represents the patch-level feature of Q_i extracted by the spatiotemporal modeling network $T(\cdot)$, providing information for predicting Y that cannot be substituted by other variables. (4) **Redundant component** R: $R = \{R_{\mathbf{Q_1}}, R_{\mathbf{Q_2}}, \ldots\}$, where each $R_{\mathbf{Q_i}} = RS(\mathbf{Q_i})$ denotes the overlapping information among variable combinations that are already captured by individual variables in relation to predicting Y, and $RS(\cdot)$ is a redundant sampler. (5) **Synergistic component** S: $S = \{S_{\mathbf{Q_1}}, S_{\mathbf{Q_2}}, \ldots\}$, where each $S_{\mathbf{Q_i}} = SS(\mathbf{Q_i})$ denotes the joint effect of variables in $\mathbf{Q_i}$, providing additional discriminative power for predicting Y beyond any single variable, and $SS(\cdot)$ is a synergistic sampler.

Methodology

This paper presents a SurdCRL causal model, as shown in Figure 3, SurdCRL includes three modules: Spatial-Temporal Entity Modeling (STEM) for extracting object and context spatiotemporal features as grounded causal nodes in the causal graph. Dual-Path Causal Inference (DPCI) applies front-door and back-door interventions to remove confounders, and Causal Composition and Selection (CCS) for composing and selecting higher-order interactions, disentangling synergistic, unique and redundant components.

Spatial-Temporal Entity Modeling Module

The Spatial-Temporal Entity Modeling (STEM) module constructs spatial-temporal causal nodes from input video by decomposing frames into object O_i and context C entities, i.e., $V \rightarrow O_i, V \rightarrow C$. These nodes serve as foundational representations for subsequent causal reasoning.

Object-Centric Representation Learning Given video frames, a ViT-style encoder extracts patch features $X \in \mathbb{R}^{THW \times d}$, and object tracking boxes $B \in \mathbb{R}^{TO \times 4}$ are obtained via SAM2 (Ravi et al. 2024). RoIAlign and spatial positional encoder $\mathcal{D}$ are then applied to obtain spatial object representations, which are further processed by a video encoder $\varphi(\cdot)$ to produce object tokens $X_o \in \mathbb{R}^{TO \times d}$:

$$X_o = \varphi(\mathrm{RoIAlign}(X, B) + \mathcal{D}(B)) \tag{1}$$

where T is the number of frames, O the number of objects, HW the spatial resolution, and d the feature dimension.

To model object dynamics, we design a Temporal Message Passing (TMP) Block, where each object at each frame is a node v with latent state h. The message passing process includes two steps (Gilmer et al. 2017): message computation and state update. First, at each iteration i, differences between latent states are measured by various distance metrics, and the message from temporal neighbors $u \in N_v$ (i.e., the same object across adjacent frames) is computed as:

$$m_v^{i+1} = \Phi(\mathrm{distance}(h_v^i, h_u^i)) \tag{2}$$

where Φ instantiates a multi-layer perceptron. Next, the state update formulated as $h_v^{i+1} = h_v^i + m_v^{i+1}$ captures temporal state transitions. Iterating this process over all frames yields object state change features $F_{os} \in \mathbb{R}^{TO \times d}$.

To further capture global dynamics, we introduce a Global Cross Multi-Head Relation Aggregation (MHRA) to obtain the global temporal object features $F_o \in \mathbb{R}^{O \times d}$:

$$F_o = \mathrm{Softmax}\left(\frac{q_o W_q (F_{os} W_k)^\top}{\sqrt{d}}\right)(F_{os} W_v) \tag{3}$$

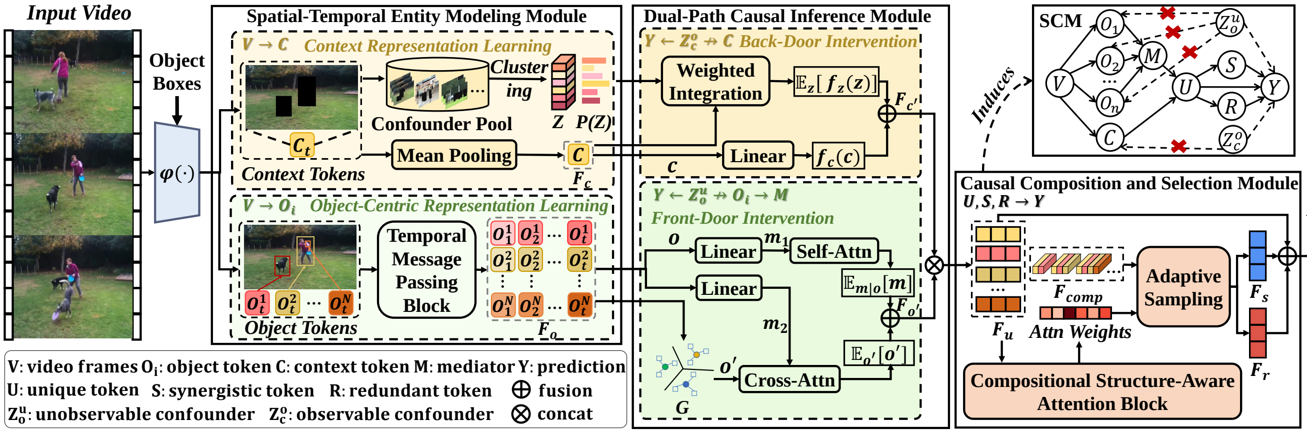
Input Video
Object Boxes
$\varphi(\cdot)$
Spatial-Temporal Entity Modeling Module
$V \rightarrow C$ Context Representation Learning
Cluster ing
Confounder Pool
Z $P(Z)$
C_t
Context Tokens
Mean Pooling
C
F_c
$V \rightarrow O_i$ Object-Centric Representation Learning
Temporal Message Passing Block
O_t^1 O_t^2 ⋯ O_t^N
Object Tokens
F_o
Dual-Path Causal Inference Module
$Y \leftarrow Z_c^o \nrightarrow C$ Back-Door Intervention
Weighted Integration
$\mathbb{E}_z[f_z(z)]$
$F_{c'}$
c
Linear
$f_c(c)$
$Y \leftarrow Z_o^u \nrightarrow O_i \rightarrow M$
Front-Door Intervention
o
Linear
m_1
Self-Attn
Linear
$\mathbb{E}_{m|o}[m]$
m_2
$F_{o'}$
$\mathbb{E}_{o'}[o']$
o'
Cross-Attn
G
Induces
SCM
O_1 O_2 O_n V C M U S R Y Z_o^u Z_c^o
Causal Composition and Selection Module
$U, S, R \rightarrow Y$
F_{comp}
Attn Weights
F_u
Adaptive Sampling
F_s
F_r
Prob
Compositional Structure-Aware Attention Block
V: video frames O_i: object token C: context token M: mediator Y: prediction
U: unique token S: synergistic token R: redundant token ⊕ fusion
Z_o^u: unobservable confounder Z_c^o: observable confounder ⊗ concat

Figure 3: Illustration of the proposed SurdCRL framework, which consists of three core modules: STEM, DPCI and CCS. Each module plays a distinct role in constructing the causal graph by modeling nodes and edges, and their collaboration models the synergistic, unique, and redundant causal components, leading to the construction of a structural causal model (SCM).

where $q_o \in \mathbb{R}^{1\times d}$ represent a learnable global temporal token, W_q, W_k, W_v are learnable projection matrices.

Context Representation Learning To obtain complementary semantics for better video understanding, we mask object box regions in the original frames and encode the remaining context using $\varphi(\cdot)$, followed by mean pooling to obtain the aggregated context feature $F_c \in \mathbb{R}^d$.

Dual-Path Causal Inference Module

In video representation learning, the typical observational likelihood is modeled using Bayes' rule as $P(Y|V) = \sum_z P(Y|V,z)P(z|V)$, where $P(z|V)$ may introduce biased weights. However, training a video classification model aims to capture the true causal effect. To achieve this, the Dual-Path Causal Inference (DPCI) Module decomposes $P(Y|V)$ into $P(Y|C)$ and $P(Y|O)$, and applies causal interventions in the causal graph to block confounding paths $C \leftarrow Z_c^o \rightarrow Y$ and $O_i \leftarrow Z_o^u \rightarrow Y$, enabling the model to reason from simpler semantic cues and capture the true synergistic, unique, and redundant causal effects in the later stages of modeling the structural causal model.

Back-Door Intervention Observable confounders Z_c^o arise from contextual statistical biases, causing the model to prioritize high-frequency background visual patterns, which may lead to spurious associations between context features C and prediction Y. To mitigate the effect of such confounders Z_c^o, we adopt a back-door intervention strategy based on do-calculus, blocking the confounding path $Z_c^o \rightarrow Y$. The causal effect is estimated as:

$$P(Y|\mathrm{do}(C)) = \sum_z P(Y|C = R_c(V,z))P(z) \tag{4}$$

where $R_c(\cdot)$ is a context encoding function. To reduce the computational cost of multiple passes over all z, we use the Normalized Weighted Geometric Mean (NWGM) to approximate the results expected from the above feature layers:

$$P(Y|\mathrm{do}(C)) \overset{\text{NWGM}}{\approx} P(Y|C = \sum_z R_c(V,z)P(z)) \tag{5}$$

We parameterize the network to approximate the conditional probability in Eq. 5 (Yang et al. 2023) as follows:

$$P(Y|\mathrm{do}(C)) = W_a f_c(c) + W_b \mathbb{E}_z[f_z(z)] \tag{6}$$

where W_a and W_b are learnable parameters. To obtain $\mathbb{E}_z[f_z(z)]$, we approximate it as a weighted integration of all background prototypes, i.e., $\mathbb{E}_z[f_z(z)] = \sum_{i=1}^N \mu_i z_i P(z_i)$, where μ_i measures the relevance of each prototype z_i to the context feature C, and $P(z_i)$ reflects its empirical frequency. In practice, we apply dot-product attention to compute μ_i, using $\mathrm{softmax}\left[(W_c C)^\top (W_d z_i)/\sqrt{d}\right]$. The term $f_c(c)$ is computed via a linear projection and fused with $\mathbb{E}_z[f_z(z)]$ to yield the debiased context representation $F_{c'}$.

Confounder Dictionary Z Due to the absence of ground-truth context labels, we construct a confounder dictionary $Z = [z_1^o, \ldots, z_L^o]$ from observable background patterns. Specifically, we apply masking to frames and extract features via a pretrained backbone, forming a confounder pool P. K-Means clustering with PCA is then applied to derive Z, where each prototype z_i^o is the centroid of a cluster in P.

Front-Door Intervention Besides Z_c^o, object sequences O may embed unobservable confounders Z_o^u, which are hard to model directly because detector limitations can cause causal targets to be missed or introduce confounding targets. These confounders affect the prediction Y via the path $O \leftarrow Z_o^u \rightarrow Y$. To address this, we adopt a front-door adjustment by introducing an intermediate variable M along the causal path $O \rightarrow M \rightarrow Y$. This forms a two-stage process: a selector that distills task-relevant cues ($O \rightarrow M$), and a predictor that generates outputs based on the mediated features ($M \rightarrow Y$). By using this path, we block the influence of $Z_o^u \rightarrow O$. The standard likelihood can be written as:

$$P(Y|O) = \sum_m P(M = m|O)P(Y|M = m) \tag{7}$$

To remove the influence of spurious correlations caused by the unobservable confounders Z_o^u, we apply the do-operator to both O and M, yielding:

$$P(Y|\text{do}(O)) = \sum_{o'} P(o') \sum_{m} P(Y|m, o')P(m|O) \quad (8)$$

$$= \mathbb{E}_{o'}\, \mathbb{E}_{m|o}[P(Y|o', m)] \quad (9)$$

where o' denotes input samples of the whole object representation space. Inspired by (Wang et al. 2024a), base on the linear mapping model, Eq. 9 becomes $\mathbb{E}_{\mathbf{m}|o}[\mathbf{m}] + \mathbb{E}_{o'}[\mathbf{o}']$, where the bold symbol $\mathbf{m}$ are intermediate features extracted from O, and $\mathbf{o}'$ to mean the object features randomly sampled by the K-means from the entire training samples.

To implement the above expectations, we apply two linear layers to F_o to generate $\mathbf{m}_1$ and $\mathbf{m}_2$. $\mathbf{m}_1$ is passed through self-attention to produce local features F_L, which represent $\mathbb{E}_{m|o}[\mathbf{m}]$. Meanwhile, $\mathbf{m}_2$ serves as key/value in cross-attention, where the queries $\mathbf{o}'$ are randomly sampled from an object global clustering dictionary G to yield global features F_G, approximating $\mathbb{E}_{o'}[\mathbf{o}']$. The final debiased object representation $F_{o'}$ is obtained by fusing F_L and F_G.

Causal Composition and Selection Module

This module constructs synergistic, unique, and redundant causal nodes via compositional structure-aware attention and adaptive sampling. Starting from unique component U of individual variables, it builds higher-order interactions and disentangles them into S and R, forming the path $U, S, R \rightarrow Y$ and yielding the structural causal model.

Compositional Structure-Aware Attention Block We first concatenate the debiased features $F_{o'}$ and $F_{c'}$ to obtain the unique feature F_u. As shown in Figure 4(a), we perform hierarchical composition on F_u to enumerate all entity combinations from second to the nth order, forming a sequence of compositional features F_h. Each combination is encoded via an MLP and enriched with positional embeddings.

The resulting F_h is then processed by a structure-aware self-attention layer with a dedicated attention mask, producing compositional features F_{comp} and attention weights w:

$$F_{\text{comp}}, w = \text{Self-Attention}(\text{Norm}(F_h),\ \text{Mask}) + F_h \quad (10)$$

where the attention mask encodes compositional priors: (1) Cross-order subset visibility: higher-order tokens attend to all their subsets; (2) Same-order intersection visibility: tokens of the same order interact only if they share common objects; (3) Global CLS visibility: the CLS token interacts with all tokens. This structured attention ensures information flows in a semantically consistent and compositionally aware manner. Its effectiveness is validated in the supplementary material. Finally, a feed-forward network (FFN) with GeLU activation further refines F_{comp}.

Adaptive Sampling To avoid early elimination of informative but low-confidence synergistic causality, we replace top-N selection with inverse transform sampling. As shown in Figure 4(b), we compute the cumulative distribution function (CDF) from attention weights w: $\text{CDF}_k = \sum_1^k w_k$, where k denotes the index of the $\sum_{i=2}^{n} \binom{n}{i}$ combinations.

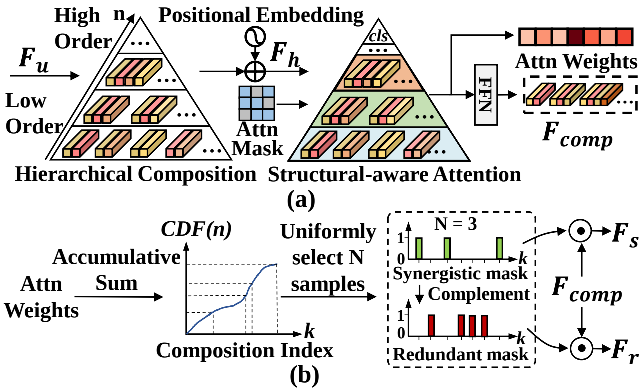

Figure 4: Illustration of the two key components in the CCS module: (a) compositional structure-aware attention block; (b) CDF-based adaptive sampling strategy.

Then, we uniformly sample N values in $[0, 1]$ and use the inverse CDF to obtain real-valued indices, which are mapped to the nearest tokens from cumulative scores. The selected positions define the binary synergistic mask, while their complements form the redundant mask. Both are applied to F_{comp} via dot-product to yield the synergistic causal feature F_s and redundant causal feature F_r, respectively.

Training Strategies

The training of SurdCRL consists of two stages. First, the STEM module is trained to produce stable representations of basic causal nodes. Then, DPCI and CCS module are jointly fine-tuned to extract synergistic, unique, and redundant causal representations under intervention, whose respective classification logits are summed to obtain the final prediction. Both stages minimize the prediction loss $\mathcal{L}_{ce} = CE(p, p')$, where CE is the cross-entropy loss, p is the predicted result, and p' is the ground truth label.

Experiments

Experiment Settings

Datasets We conduct extensive experiments on two benchmark video classification datasets: MSR-VTT (Xu et al. 2016) and ActivityNet (Caba Heilbron et al. 2015). Following the same data preprocessing protocols as in prior works, we split MSR-VTT into 7,010 videos for training and 2,990 for testing. For ActivityNet, the dataset is divided into 10,009 videos for training and 4,515 for testing.

Evaluation Protocol Consistent with prior video classification approaches, we employ Top-k (where $k = 1, 5$) accuracy to measure the effectiveness of classification results.

Implementation Details We sample 8 frames per video following the standard protocol. Pretrained Uniformerv2-B or InternVideo2-B serve as video encoders with a hidden size of 768. A SAM2 tracker selects the top 4 consistently tracked objects. Manhattan distance is used as the similarity metric in STEM. The global clustering dictionary G is fixed at 512, while the confounder dictionary Z is set to 64 for MSR-VTT and 256 for ActivityNet. In adaptive sampling, we set N=10. Training uses AdamW with cosine de-

Arch	Model	MSR-VTT		ActivityNet	
		Top-1	Top-5	Top-1	Top-5
CNN	GC-TDN	52.17	82.14	75.42	93.79
	TANet	53.47	81.20	76.13	94.06
SSM	VideoMamba-M	57.13	83.52	84.38	96.36
Trans.	ViViT-B	55.79	84.95	80.96	95.46
	VideoSwin-B	56.42	84.75	83.80	96.37
	VideoMAEv2-B	58.10	86.50	88.96	97.87
	UMT-B	59.92	86.17	89.48	97.64
	Internvideo2-B	60.85	87.25	90.37	97.80
CNN + Trans.	MViTv2-S	55.14	83.85	83.26	96.83
	VideoFocalNet-B	56.93	83.97	83.90	96.65
	Uniformerv2-B	60.73	85.75	86.77	96.94
Trans.	ORViT(Internvideo2-B)	61.24	86.80	90.79	98.22
	ECRL(Internvideo2-B)	61.36	87.42	91.04	98.47
	SurdCRL(Internvideo2-B)	62.17	**88.73**	**92.05**	**98.64**
CNN + Trans.	ORViT(Uniformerv2-B)	62.07	85.54	88.21	97.82
	ECRL(Uniformerv2-B)	63.26	86.33	88.74	97.52
	SurdCRL(Uniformerv2-B)	**63.82**	86.62	89.95	98.01

Table 1: Performance comparison of SurdCRL and baselines with different backbone architectures (Arch = Architecture, SSM = State Space Model, Trans. = Transformer).

Model	MSR-VTT		ActivityNet	
	Top-1	Top-5	Top-1	Top-5
Base	60.73	85.75	86.77	96.94
+ STEM	62.23	86.00	88.58	97.05
+ STEM + CSA	62.50	86.23	88.94	97.83
+ STEM + CSA + AS	62.80	86.10	89.10	97.79
+ STEM + CSA + AS + BDI	63.51	86.39	89.65	98.01
+ STEM + CSA + BDI + FDI	63.18	86.36	89.44	97.83
+ STEM + CSA + AS + BDI + FDI	**63.82**	**86.62**	**89.95**	**98.01**

Table 2: Ablation study of SurdCRL with Uniformerv2.

Setting	MSR-VTT		ActivityNet	
	Top-1	Top-5	Top-1	Top-5
Zero Mask	63.16	86.30	88.95	97.85
Random Mask	63.26	86.30	89.29	97.95
Segment Mask	63.31	86.10	89.69	**98.12**
Boxes Mask	63.54	86.13	89.85	98.06
Saliency Mask	**63.82**	**86.62**	**89.95**	98.01

Table 3: Impact of different mask types on the construction of the confounder dictionary Z in the BDI.

cay scheduling, where the learning rate ranges from 1e-7 to 1e-5. The first stage trains 25 epochs with a batch size of 2; the second trains 15 epochs with a batch size of 32.

Performance Comparison

We compare SurdCRL with 11 video classification methods, including GC-TDN (Hao et al. 2022), TANet (Liu et al. 2021), VideoMamba (Li et al. 2024), MViTv2 (Li et al. 2022), Uniformerv2 (Li et al. 2023a), ViViT (Arnab et al. 2021), VideoSwim (Liu et al. 2022b), VideoFocalNet (Wasim et al. 2023), UMT (Liu et al. 2022a), VideoMAEv2 (Wang et al. 2023), and InternVideo2 (Wang et al. 2024b). We select Uniformerv2 and InternVideo2 as strong baselines representing CNN+Transformer and pure Transformer architectures, respectively. We also compare against ORViT (Herzig et al. 2022), a plug-and-play object-centric model, and ECRL (Wang et al. 2024c), an event-level causal model. Results are summarized in Table 1.

- **Transformer-based models offer superior modeling capacity.** Transformer and CNN+Transformer models outperform CNNs and SSMs by better capturing long-range temporal and fine-grained spatial features.
- **SurdCRL generalizes well across diverse architectures.** It improves Top-1 accuracy by 3.09% on MSR-VTT and 3.18% on ActivityNet with Uniformerv2, and further reaches 92.05% on ActivityNet with InternVideo2, demonstrating robust spatial-temporal semantic modeling.
- **Causal decomposition gives SurdCRL a distinct advantage.** Compared to ORViT, it performs better across both backbone types by decomposing spatio-temporal object interactions into three causal components and mitigating confounders through causal inference.
- **Object-centric causal modeling proves superior.** Unlike ECRL, which builds region-level causal graphs, SurdCRL achieves significantly better results on ActivityNet, highlighting the benefits of fine-grained object-centric causal reasoning in complex interactions.

Ablation Study

In this section, we analyze the contribution of each module in our SurdCRL framework, as presented in Table 2. The following findings can be observed:

- **Spatial-Temporal Entity Modeling enhances representation.** The Spatial-Temporal Entity Modeling module (+STEM) boosts performance by disentangling fine-grained visual elements and modeling object-centric representation over time to capture core event dynamics.
- **Compositional attention and sampling aid causal discrimination.** Adding CSA (+STEM+CSA) and Adaptive Sampling (+STEM+CSA+AS) improves accuracy via high-order interaction modeling and causal effect isolation. Removing sampling (+STEM+CSA+BDI+FDI) slightly drops performance, showing adaptive selection's role in refining causal features.
- **Dual-path interventions provide robust debiasing.** Incorporating back-door (+STEM+CSA+AS+BDI) and front-door (+STEM+CSA+AS+BDI+FDI) interventions improves performance. These methods alleviate the effects of confounders introduced by observable background interference and unobservable data biases.

In-depth Analyses

Effectiveness Analyses of the Background Confounders Dictionary Z We construct the background confounder dictionary Z in BDI using different spatial masking strategies, as shown in Table 3. The Saliency (Liu et al. 2019) Mask performs best, as it effectively excludes salient objects and highlights background confounders. In contrast, Zero Mask and Random Mask perform poorly due to their lack of semantic guidance or the introduction of noise. Segment (Kirillov et al. 2023) Mask and Boxes (Zhang et al. 2022)

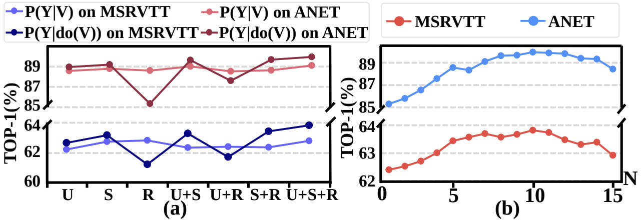

Figure 5: (a) Comparison of different causal components before and after intervention. (b) Performance trends under different settings of hyper-parameter N in adaptive sampling.

Mask yield improvements by removing object regions, but still lag behind Saliency Mask, because they also mask out some background regions that are visually relevant for modeling confounders. These results highlight the importance of isolating background regions that introduce spurious correlations for effective backdoor intervention.

Causal Effect Analysis of Unique, Synergistic, and Redundant Components We analyze the roles of unique (U), synergistic (S), and redundant (R) causalities on both MSR-VTT and ActivityNet (ANET) datasets before and after intervention, as shown in Figure 5(a). Prior to intervention, all components show similar performance, indicating that confounders hinder the model's ability to distinguish their causal roles. After applying our Dual-Path Causal Intervention, overall performance improves. We also observe that S outperforms U, indicating variable combinations offer extra discriminative cues. However, R and U+R show slight declines, as R begins to reflect its true function of capturing redundant cues, which reduces its contribution to correct classification. In contrast, S+R improves, indicating that when the expressive power of S is limited, R can provide useful co-occurrence cues to complement the representation. These results confirm that our intervention enables each causal component to better express its true effect.

Effectiveness Analyses of Hyper-parameter N in Adaptive Sampling We analyze the impact of the hyper-parameter N in adaptive sampling on MSR-VTT and ActivityNet, as shown in Figure 5(b). As N increases from 1, accuracy on both datasets steadily improves and peaks around $N = 10$, indicating that larger N helps better distinguish synergistic and redundant causalities. However, further increasing N causes performance drops. This is because excessive sampling blurs the boundary between synergistic and redundant causalities, thereby hindering the model's ability to accurately capture their respective causal effects. Notably, ActivityNet exhibits more pronounced fluctuations when N is small (up to 5% variation). This suggests that in object-centric datasets, small N may fail to capture discriminative object interactions necessary for modeling causality.

Case Study

Visualization of Decision Boundaries under Different Modeling Paradigms We use t-SNE to compare feature distributions and decision boundaries (Chen et al. 2023; Yan et al. 2025; Li et al. 2025b) across three confusing MSR-VTT categories, as shown in Figure 6. The base model yields cluttered features and overlapping boundaries. Modeling three causalities via STEM and CCS offers structural gains but remains relatively noisy due to the presence of confounders. In contrast, incorporating intervention through DPCI produces compact and more discriminative clusters.

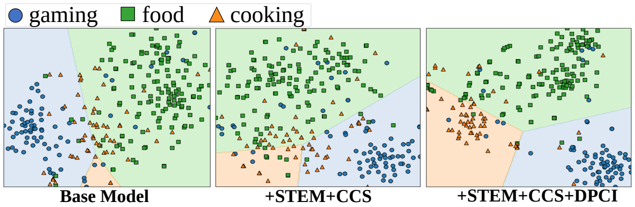

Figure 6: Visualization results on representation distribution under different module combinations.

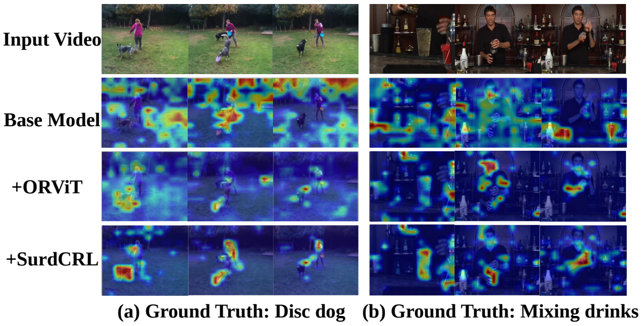

Figure 7: Visualization of attention on sampled frames, where ORViT is a transformer-based object-centric method, "+" indicates the method applied on the base model.

Visualization of the Causal Representation by SurdCRL To evaluate the efficacy of SurdCRL, we compare it against Uniformerv2 and ORViT on the ActivityNet validation set using Grad-CAM (Meng et al. 2020; Qi et al. 2025a), as shown in Figure 7. Results reveal clear differences in attention to visual cues. The base model attends to background, leading to ambiguous focus. ORViT shifts attention to relevant areas but still overlooks key interactions. SurdCRL, by contrast, highlights event-relevant cues like person-dog-disc engagement in (a) and hand-tool manipulation in (b), demonstrating stronger causal grounding.

Conclusion

This paper presents SurdCRL, a causal model that leverages object-centric interactions to construct an SCM, a structural causal model that decomposes causality into synergistic, unique, and redundant components. Experimental results demonstrate the effectiveness of SurdCRL in focusing on event-relevant object-centric interactions and achieving precise video understanding. In the future, how to leverage the multimodal information in video sequences to construct a spatiotemporal causal representation with aligned text–visual semantics will be one of our research directions.

Acknowledgements

This work is supported in part by the Key Research and Development Program of Shandong Province (Grant No. 2024TSGC0667) and the Shandong Province Youth Entrepreneurship Technology Support Program for Higher Education Institutions (Grant No. 2022KJN028).

References

Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C. L.; and Parikh, D. 2015. Vqa: Visual question answering. In *ICCV*, 2425–2433.

Arnab, A.; Dehghani, M.; Heigold, G.; Sun, C.; Lučić, M.; and Schmid, C. 2021. Vivit: A video vision transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, 6836–6846.

Caba Heilbron, F.; Escorcia, V.; Ghanem, B.; and Carlos Niebles, J. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 961–970.

Chen, T.; Liu, H.; He, T.; Chen, Y.; Gan, C.; Ma, X.; Zhong, C.; Zhang, Y.; Wang, Y.; Lin, H.; et al. 2024. MECD: Unlocking multi-event causal discovery in video reasoning. *Advances in Neural Information Processing Systems*, 37: 92554–92580.

Chen, W.; Liu, Y.; Chen, B.; Su, J.; Zheng, Y.; and Lin, L. 2025. Cross-modal causal relation alignment for video question grounding. In *CVPR*, 24087–24096.

Chen, Z.; Qi, Z.; Cao, X.; Li, X.; Meng, X.; and Meng, L. 2023. Class-level structural relation modeling and smoothing for visual representation learning. In *Proceedings of the 31st ACM International Conference on Multimedia*, 2964–2972.

Didolkar, A.; Zadaianchuk, A.; Awal, R.; Seitzer, M.; Gavves, E.; and Agrawal, A. 2025. Ctrl-o: language-controllable object-centric visual representation learning. In *CVPR*, 29523–29533.

Feichtenhofer, C. 2020. X3d: Expanding architectures for efficient video recognition. In *CVPR*, 203–213.

Feichtenhofer, C.; Fan, H.; Malik, J.; and He, K. 2019. Slowfast networks for video recognition. In *ICCV*, 6202–6211.

Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; and Dahl, G. E. 2017. Neural message passing for quantum chemistry. In *ICML*, 1263–1272. Pmlr.

Hao, Y.; Zhang, H.; Ngo, C.-W.; and He, X. 2022. Group contextualization for video recognition. In *CVPR*, 928–938.

Herzig, R.; Ben-Avraham, E.; Mangalam, K.; Bar, A.; Chechik, G.; Rohrbach, A.; Darrell, T.; and Globerson, A. 2022. Object-region video transformers. In *CVPR*, 3148–3159.

Huang, P.; Shu, X.; Yan, R.; Tu, Z.; and Tang, J. 2025. Appearance-Agnostic Representation Learning for Compositional Action Recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(4): 3039–3053.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *ICCV*, 4015–4026.

Li, K.; Li, X.; Wang, Y.; He, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2024. Videomamba: State space model for efficient video understanding. In *ECCV*, 237–255. Springer.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Wang, L.; and Qiao, Y. 2023a. Uniformerv2: Unlocking the potential of image vits for video understanding. In *ICCV*, 1632–1643.

Li, X.; Sun, P.; Liu, Y.; Duan, L.; and Li, W. 2025a. Simultaneous Detection and Interaction Reasoning for Object-Centric Action Recognition. *IEEE Transactions on Multimedia*, 27: 5283–5295.

Li, Y.; Wu, C.-Y.; Fan, H.; Mangalam, K.; Xiong, B.; Malik, J.; and Feichtenhofer, C. 2022. Mvitv2: Improved multiscale vision transformers for classification and detection. In *CVPR*, 4804–4814.

Li, Y.; Yang, X.; Zhang, A.; Feng, C.; Wang, X.; and Chua, T.-S. 2023b. Redundancy-aware transformer for video question answering. In *ACM MM*, 3172–3180.

Li, Z.; Meng, L.; Chao, G.; Wu, W.; Yan, X.; Yang, Y.; Qi, Z.; and Meng, X. 2025b. Semantic-Space-Intervened Diffusive Alignment for Visual Classification. *arXiv preprint arXiv:2505.05721*.

Lin, J.; Gan, C.; and Han, S. 2019. Tsm: Temporal shift module for efficient video understanding. In *ICCV*, 7083–7093.

Liu, J.-J.; Hou, Q.; Cheng, M.-M.; Feng, J.; and Jiang, J. 2019. A simple pooling-based design for real-time salient object detection. In *CVPR*, 3917–3926.

Liu, Y.; Li, G.; and Lin, L. 2023. Cross-modal causal relational reasoning for event-level visual question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(10): 11624–11641.

Liu, Y.; Li, S.; Wu, Y.; Chen, C.-W.; Shan, Y.; and Qie, X. 2022a. Umt: Unified multi-modal transformers for joint video moment retrieval and highlight detection. In *CVPR*, 3042–3051.

Liu, Y.; Liu, F.; Jiao, L.; Bao, Q.; Li, L.; Guo, Y.; and Chen, P. 2024. A Knowledge-Based Hierarchical Causal Inference Network for Video Action Recognition. *IEEE Transactions on Multimedia*, 26: 9135–9149.

Liu, Y.; Liu, F.; Jiao, L.; Bao, Q.; Li, S.; Li, L.; and Liu, X. 2025. Knowledge-driven compositional action recognition. *Pattern Recognition*, 163: 111452.

Liu, Z.; Ning, J.; Cao, Y.; Wei, Y.; Zhang, Z.; Lin, S.; and Hu, H. 2022b. Video swin transformer. In *CVPR*, 3202–3211.

Liu, Z.; Wang, L.; Wu, W.; Qian, C.; and Lu, T. 2021. Tam: Temporal adaptive module for video recognition. In *ICCV*, 13708–13718.

Martínez-Sánchez, Á.; Arranz, G.; and Lozano-Durán, A. 2024. Decomposing causality into its synergistic, unique, and redundant components. *Nature Communications*, 15(1): 9296.

Materzynska, J.; Xiao, T.; Herzig, R.; Xu, H.; Wang, X.; and Darrell, T. 2020. Something-else: Compositional action recognition with spatial-temporal interaction networks. In *CVPR*, 1049–1059.

Meng, L.; Feng, F.; He, X.; Gao, X.; and Chua, T.-S. 2020. Heterogeneous fusion of semantic and collaborative information for visually-aware food recommendation. In *Proceedings of the 28th ACM international conference on multimedia*, 3460–3468.

Meng, L.; Li, X.; Yan, X.; Ma, H.; Qi, Z.; Wu, W.; and Meng, X. 2025. Causal Inference over Visual-Semantic-Aligned Graph for Image Classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 19449–19457.

Niu, Y.; Tang, K.; Zhang, H.; Lu, Z.; Hua, X.-S.; and Wen, J.-R. 2021. Counterfactual vqa: A cause-effect look at language bias. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12700–12710.

Pearl, J.; et al. 2016. *Causal inference in statistics: A primer*. John Wiley & Sons.

Qi, Z.; Meng, L.; Chen, Z.; Hu, H.; Lin, H.; and Meng, X. 2023. Cross-silo prototypical calibration for federated learning with non-iid data. In *ACM MM*, 3099–3107.

Qi, Z.; Meng, L.; Li, Z.; Hu, H.; and Meng, X. 2025a. Cross-Silo Feature Space Alignment for Federated Learning on Clients with Imbalanced Data. In *AAAI*, 19986–19994.

Qi, Z.; Zhou, S.; Meng, L.; Hu, H.; Yu, H.; and Meng, X. 2025b. Federated Deconfounding and Debiasing Learning for Out-of-Distribution Generalization. *arXiv preprint arXiv:2505.04979*.

Qian, R.; Ding, S.; and Lin, D. 2024. Rethinking image-to-video adaptation: An object-centric perspective. In *ECCV*, 329–348. Springer.

Qu, H.; Yan, R.; Shu, X.; Gao, H.; Huang, P.; and Xie, G. 2025. MVP-Shot: Multi-Velocity Progressive-Alignment Framework for Few-Shot Action Recognition. *IEEE Transactions on Multimedia*, 27: 6593–6605.

Ravi, N.; Gabeur, V.; Hu, Y.-T.; Hu, R.; Ryali, C.; Ma, T.; Khedr, H.; Rädle, R.; Rolland, C.; Gustafson, L.; Mintun, E.; Pan, J.; Alwala, K. V.; Carion, N.; Wu, C.-Y.; Girshick, R.; Dollár, P.; and Feichtenhofer, C. 2024. SAM 2: Segment Anything in Images and Videos. *arXiv preprint arXiv:2408.00714*.

Vosoughi, A.; Deng, S.; Zhang, S.; Tian, Y.; Xu, C.; and Luo, J. 2024. Cross Modality Bias in Visual Question Answering: A Causal View With Possible Worlds VQA. *IEEE Transactions on Multimedia*, 26: 8609–8624.

Wang, B.; Ju, X.; Gao, J.; Li, X.; Hu, Y.; and Yin, B. 2025. Counterfactual Dual-Bias VQA: A Multimodality Debias Learning for Robust Visual Question Answering. *IEEE Transactions on Neural Networks and Learning Systems*, 36(9): 16366–16378.

Wang, L.; He, Z.; Dang, R.; Shen, M.; Liu, C.; and Chen, Q. 2024a. Vision-and-language navigation via causal learning. In *CVPR*, 13139–13150.

Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; and Qiao, Y. 2023. Videomae v2: Scaling video masked autoencoders with dual masking. In *CVPR*, 14549–14560.

Wang, T.; Li, Y.; Kang, B.; Li, J.; Liew, J.; Tang, S.; Hoi, S.; and Feng, J. 2020. The devil is in classification: A simple framework for long-tail instance segmentation. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XIV 16*, 728–744. Springer.

Wang, Y.; Li, K.; Li, X.; Yu, J.; He, Y.; Chen, G.; Pei, B.; Zheng, R.; Wang, Z.; Shi, Y.; et al. 2024b. Internvideo2: Scaling foundation models for multimodal video understanding. In *ECCV*, 396–416. Springer.

Wang, Y.; Li, X.; Ma, H.; Qi, Z.; Meng, X.; and Meng, L. 2022a. Causal inference with sample balancing for out-of-distribution detection in visual classification. In *CAAI International Conference on Artificial Intelligence*, 572–583. Springer.

Wang, Y.; Li, X.; Qi, Z.; Li, J.; Li, X.; Meng, X.; and Meng, L. 2022b. Meta-causal feature learning for out-of-distribution generalization. In *ECCV*, 530–545. Springer.

Wang, Y.; Meng, L.; Ma, H.; Wang, Y.; Huang, H.; and Meng, X. 2024c. Modeling event-level causal representation for video classification. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 3936–3944.

Wasim, S. T.; Khattak, M. U.; Naseer, M.; Khan, S.; Shah, M.; and Khan, F. S. 2023. Video-focalnets: Spatio-temporal focal modulation for video action recognition. In *ICCV*, 13778–13789.

Xu, B.; Shu, X.; Zhang, J.; Dai, G.; and Song, Y. 2023. Spatiotemporal decouple-and-squeeze contrastive learning for semisupervised skeleton-based action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 35(8): 11035–11048.

Xu, J.; Mei, T.; Yao, T.; and Rui, Y. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *CVPR*, 5288–5296.

Yan, X.; Li, Z.; Meng, L.; Qi, Z.; Wu, W.; Li, Z.; and Meng, X. 2025. Empowering Vision Transformers with Multi-Scale Causal Intervention for Long-Tailed Image Classification. *arXiv preprint arXiv:2505.08173*.

Yang, D.; Chen, Z.; Wang, Y.; Wang, S.; Li, M.; Liu, S.; Zhao, X.; Huang, S.; Dong, Z.; Zhai, P.; et al. 2023. Context de-confounded emotion recognition. In *CVPR*, 19005–19015.

Zhang, C.; Fu, C.; Wang, S.; Agarwal, N.; Lee, K.; Choi, C.; and Sun, C. 2024. Object-centric video representation for long-term action anticipation. In *WACV*, 6751–6761.

Zhang, C.; Gupta, A.; Zisserman, A.; et al. 2023. Helping hands: An object-aware ego-centric video recognition model. In *ICCV*, 13901–13912.

Zhang, H.; Li, F.; Liu, S.; Zhang, L.; Su, H.; Zhu, J.; Ni, L. M.; and Shum, H.-Y. 2022. Dino: Detr with improved denoising anchor boxes for end-to-end object detection. *arXiv preprint arXiv:2203.03605*.

Zhang, Y.; Tokmakov, P.; Hebert, M.; and Schmid, C. 2019. A structured model for action detection. In *CVPR*, 9975–9984.