Progressive Cross-Modal Causal Intervention for Long-Term Action Recognition

Supplementary Material

6. Causal Formulation and Interpretation

6.1. Scope of the Structural Causal Model

The structural causal model (SCM) in the main paper explicitly represents three dominant confounding mechanisms, namely co-occurrence hallucination (H), codependency illusion (I), and visual confounders (C). These factors are identified based on recurring failure modes observed in VLM-based LTAR systems. Co-occurrence hallucination arises from biased cross-modal alignment, codependency illusion stems from the lack of explicit relational modeling among actions, and visual confounders originate from persistent non-causal visual patterns.

In our framework, causal adjustment is applied only to the confounding mechanisms that are explicitly modeled in the SCM. PCMCI therefore focuses on these three factors because they are both empirically prominent and structurally interpretable. Although these factors may appear predictive under the training distribution, they do not necessarily provide stable cues across distributions and may degrade generalization under distribution shift. This motivates explicit deconfounding rather than relying solely on implicit regularization.

6.2. Back-door Adjustment via Surrogate Variables

In the ICH and ICI stages, the true confounders (H and I) are not directly observable. To address this, we construct surrogate variables from learned feature embeddings and approximate back-door adjustment as

$$P(Y \mid V, do(T)) \approx \sum_{h} P(Y \mid V, T, h) P(h). \tag{11}$$

The surrogate variables capture the dominant variation induced by latent confounding factors. Although they do not exactly recover the underlying latent variables, they provide a practical approximation that allows the model to suppress unstable dependencies in the learned embeddings.

We adopt the Normalized Weighted Geometric Mean (NWGM) for efficient marginalization, avoiding sampling-based estimation while preserving the relative contribution of different confounding components. In practice, this formulation provides a stable approximation of interventional inference and integrates naturally into end-to-end optimization. In our implementation, the weights involved in NWGM aggregation are normalized with a softmax operation before combination. This normalization ensures that the resulting coefficients form a valid distribution over surrogate confounding components, which stabilizes optimization and makes the approximation more consistent with the probabilistic interpretation of marginalization.

6.3. Front-door Adjustment with Textual Mediator

In the IVC stage, we use the deconfounded text embedding T^* as mediator M. This choice is supported by two considerations. First, T^* provides a compact and semantically informative embedding of the target action. Second, after the preceding deconfounding stages, T^* becomes less entangled with non-causal correlations and is therefore more suitable for mediating the influence from visual evidence to prediction.

Consequently, T^* serves as a relatively stable intermediate embedding linking visual input to the target label, reducing direct reliance on confounded visual patterns. This design is relevant in LTAR since visual observations often contain persistent non-causal cues, such as scene context, actor appearance, or recurring objects.

6.4. Interpretation of d-separation

Exact d-separation is difficult to guarantee in high-dimensional representation learning. Instead, PCMCI adopts a progressive approximation strategy. More specifically, ICH suppresses cross-modal spurious alignment, ICI reduces incorrect relational dependencies among actions, and IVC further mitigates residual visual confounding.

Through this staged design, the mediator becomes increasingly less dependent on the modeled confounding factors, which makes the front-door approximation more reliable in practice. This perspective is consistent with causal representation learning settings in which exact identifiability is rarely achievable and the primary objective is to reduce reliance on unstable cues rather than to recover the full data-generating process.

7. Codependency Illusion: Motivation and Quantification

This section clarifies why textual codependency illusion is not a dataset-specific artifact but an intrinsic challenge in LTAR, and introduces a quantitative proxy for measuring it.

Long-term actions are typically composed of multiple atomic actions with non-trivial dependency structure, including shared sub-actions, exclusive sub-actions, and order-sensitive transitions. When action labels are encoded independently by a pretrained VLM, such structural dependencies are not explicitly represented. As a result, the model may fail to distinguish actions that are visually similar but semantically differentiated by their relational context, or

may incorrectly associate actions that frequently co-occur but do not belong to the same procedural pattern.

The ICI stage is introduced precisely to address this gap. Rather than assuming that these dependencies will be learned implicitly from downstream supervision, PCMCI explicitly treats them as a source of bias in textual embeddings. This is particularly relevant in LTAR, where action identity is often determined not only by local visual evidence but also by compositional context across atomic actions.

To quantitatively evaluate codependency illusion, we define a metric based on the co-prediction matrix $Q \in \mathbb{R}^{N\times N}$, where

$$Q_{ij} = P(\text{action}_i \mid \text{action}_j). \tag{12}$$

This matrix reflects how strongly the prediction of one action is associated with another. Ideally, high values should mainly appear for semantically or procedurally related action pairs.

We measure the density of implausible action associations using the off-diagonal density

$$\text{ODD}_\tau(Q) = \frac{1}{N(N-1)} \sum_{i\neq j} \mathbb{1}[Q_{ij} > \tau], \tag{13}$$

where we set $\tau = 0.6$ in all experiments.

A higher value of ODD_τ indicates stronger codependency illusion, as it means that the model more frequently assigns high conditional probability to semantically implausible action pairs. In our experiments, PCMCI reduces off-diagonal density by approximately 14%, demonstrating its effectiveness in suppressing spurious relational dependencies. This result is consistent with the qualitative visualization in the main paper, where PCMCI produces more semantically coherent action co-prediction patterns.

8. Design Choices and Additional Studies

8.1. Why Optimal Transport in ICH

Co-occurrence hallucination arises from many-to-many cross-modal over-alignment, where multiple visual patterns repeatedly co-occur with a given action label. As a result, naive similarity weighting may over-emphasize recurrent but non-causal correspondences.

Optimal Transport (OT) is well-suited for this scenario because it enforces global matching constraints under marginal distributions. Unlike independent similarity weighting, OT jointly optimizes the alignment structure over the entire similarity matrix, allowing it to identify disproportionately strong correspondences that are indicative of hallucinated associations. This property is desirable because it refines cross-modal similarity before constructing the surrogate confounder representation.

8.2. Why ICH Precedes ICI

An important design question is why the order between ICH and ICI matters, although both operate on textual representations. The key reason is that the two stages address different sources of bias.

ICH first suppresses cross-modal hallucinated alignments inherited from the pretrained VLM. If ICI is applied before ICH, the relation-aware module operates on text embeddings that are still contaminated by cross-modal bias, and the learned inter-action dependencies may partially fit these spurious alignments rather than genuine semantic structure. In contrast, when ICH is performed first, ICI receives a cleaner intermediate embedding and can focus more effectively on modeling true action dependencies.

Therefore, although both stages target biases associated with textual embeddings, their order is not interchangeable in practice. The intervention order influences the quality of the mediator used later in IVC, which explains the performance gap between alternative orderings observed in the ablation study.

8.3. Choice of Video Encoder

A critical design choice in PCMCI is the selection of the VLM-independent video encoder. Since PCMCI explicitly performs causal adjustment on learned feature embeddings, the quality of these embeddings directly affects the effectiveness of downstream deconfounding. In particular, the encoder should preserve temporally structured information, since long-term actions are defined not only by appearance cues but also by the ordering and composition of atomic actions over time.

We compare multiple architectures with different temporal modeling capacities:

Table 7. Ablation on video encoders (Acc %).

	X3D	ViT	TimeSformer	VideoMamba
Breakfast	90.42	95.21	97.46	97.75
COIN	88.49	92.60	94.53	91.13

The results suggest that temporal modeling is important, but stability across datasets also matters. In particular, TimeSformer consistently outperforms purely spatial modeling (ViT) and lightweight temporal modeling (X3D), indicating that capturing long-range temporal dependencies is important for LTAR. Although VideoMamba achieves competitive performance on Breakfast, its performance is less stable across datasets, suggesting that its advantage may be more dataset-dependent.

Based on this analysis, we adopt TimeSformer as a balanced choice. It provides strong temporal modeling capacity while maintaining stable performance across datasets, which is desirable for a causal deconfounding framework

whose downstream modules rely on structurally reliable video embeddings.

8.4. Parameter-efficient Adaptation

We further evaluate PCMCI under a parameter-efficient setting:

Table 8. Parameter-efficient adaptation on Breakfast.

Tunable Modules	Params (%)	Acc (%)
Text + Classifier	5.41	86.76
+ Visual Module	8.44	90.70

The results show that PCMCI maintains strong performance even when only a small fraction of parameters is updated. This suggests that PCMCI can be adapted efficiently under constrained tuning budgets, without requiring full end-to-end optimization.

9. Implementation Details and Algorithm

9.1. Overall Pipeline

PCMCI processes untrimmed videos and performs video-level prediction. The framework follows a progressive deconfounding strategy across three stages, corresponding to co-occurrence hallucination suppression, codependency correction, and visual confounder mitigation. The complete pipeline is summarized in Algorithm 1.

Importantly, PCMCI does not perform explicit atomic-action localization or action-window selection. Instead, the entire input video is fed into both the frozen VLM branch and the trainable video encoder, and all subsequent stages operate on embeddings extracted from the full video sequence. In this sense, the temporal aspect of the method is modeled implicitly through the video encoder and the relation-aware transformations, rather than by explicitly cropping temporal windows for individual atomic actions.

More specifically, the frozen VLM provides the initial cross-modal priors, while the trainable video encoder extracts VLM-independent visual features from the input sequence. The model then progressively refines the textual and visual embeddings through the ICH, ICI, and IVC stages before performing cross-modal inference on the resulting deconfounded features.

9.2. Inference Module

We use a 1-layer cross-attention Transformer decoder with hidden dimension 512 and 8 attention heads. This lightweight design balances performance and computational efficiency, and is sufficient because the main representational refinement is already handled by the causal deconfounding stages.

Algorithm 1 Progressive Cross-Modal Causal Intervention (PCMCI)

Require: Input video X, action text labels
Require: Frozen VLM encoder $\mathcal{E}_{vlm}$, trainable video encoder $\mathcal{E}_v$
Extract VLM features:
  $V^P \leftarrow \mathcal{E}_{vlm}^{vis}(X)$
  $T \leftarrow \mathcal{E}_{vlm}^{txt}(\text{labels})$
Extract VLM-independent features:
  $V \leftarrow \mathcal{E}_v(X)$
// Stage 1: ICH
$S \leftarrow \langle T, V^P \rangle$
$P^* \leftarrow \arg\min_{P \in U} \langle P, -\log S \rangle + \lambda \mathcal{H}(P)$
$\tilde{S} \leftarrow f(S, P^*)$
$H \leftarrow \tilde{S} V^P$
$T' \leftarrow [T, H] W^H$
// Stage 2: ICI
$R \leftarrow \mathcal{R}(T')$
$T^* \leftarrow [T', R] W^I$
// Stage 3: IVC
$M \leftarrow T^*$
$V' \leftarrow \mathcal{T}(V, M)$
$V^* \leftarrow [M, V'] W^C$
// Inference
$Z \leftarrow \text{CrossAttention}(Q = T^*, K = V^*, V = V^*)$
$\hat{Y} \leftarrow \text{Classifier}(Z)$
return $\hat{Y}$

10. Failure Mode Analysis and Limitations

Although PCMCI substantially improves robustness to spurious correlations, the results on COIN also reveal an important limitation of the proposed framework. As shown in the main paper, PCMCI achieves the best mAP on COIN (86.54), substantially outperforming prior methods, while its Acc (94.53) is slightly lower than HierarQ (94.78). This gap is small in absolute value, but it is still informative because it reflects a systematic trade-off introduced by causal deconfounding.

COIN is particularly challenging because it spans 180 tasks from 12 domains and consists of diverse YouTube videos with considerable variation in scene layout, object appearance, recording style, and procedural context. Under such conditions, some contextual signals that are not strictly causal may nevertheless become highly predictive of the dominant action label within the observed distribution. Such signals may include recurring backgrounds, domain-specific tools, or characteristic visual contexts that frequently accompany a specific task. A model that exploits such signals can benefit in top-1 classification, even if those cues do not reflect the essential action mechanism.

PCMCI is explicitly designed to reduce reliance on these unstable correlations and instead emphasize invariant action semantics and inter-action structure. This design leads to clear gains in mAP, which is more sensitive to whether

the model captures the broader dependency structure among atomic actions. At the same time, this design may reduce the advantage of exploiting highly predictive contextual cues on individual samples. This likely explains why PCMCI does not obtain the highest Acc on COIN, despite achieving a large margin on mAP.

Therefore, an important limitation of PCMCI is that it may sacrifice a small amount of single-label accuracy when unstable contextual evidence is unusually strong and happens to align with the target label in the test distribution. In such cases, models that more aggressively exploit dataset-specific regularities may still achieve slightly better top-level classification performance. This behavior is not inconsistent with the causal objective of PCMCI, but instead reflects the practical cost of preferring robustness over opportunistic exploitation of unstable cues.

This observation also suggests several directions for future work. First, it would be valuable to develop adaptive deconfounding mechanisms that can better balance robustness and discriminative utility, rather than treating all shortcut-like signals uniformly. Second, future work could investigate confidence-aware or sample-dependent adjustment strategies, so that the degree of deconfounding can be adapted according to the reliability of the contextual evidence. Third, richer temporal reasoning may further improve performance on datasets such as COIN, where tasks often exhibit large procedural diversity and loose temporal structure. More broadly, our results suggest that causal deconfounding is highly effective for structured recognition, but further advances are needed to reconcile robustness with maximal top-level accuracy in highly diverse real-world video domains.