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

Shaowu Xu[1], Xibin Jia[1]*,Chao Fan[1], Junyu Gao[2], Jing Chang[3], Qianmei Sun[3]
1 College of Computer Science, Beijing University of Technology, China; 2 Institute of Automation, Chinese Academy of Sciences, China; 3 Beijing Chao-yang Hospital, Capital Medical University, China

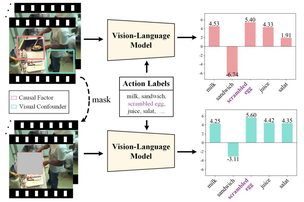
CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation: What Goes Wrong?

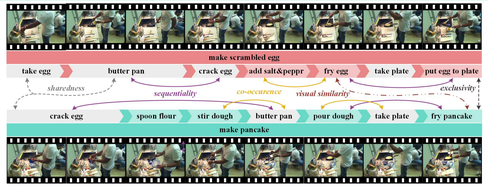
- LTAR must recognize long sequences of correlated atomic actions.
- VLM label supervision helps, but alignment can follow statistical shortcuts.
- **Three latent issues: co-occurrence hallucination (H), codependency illusion (I), visual confounders (C).**

Example: Co-occurrence Hallucination

Masking causal regions can increase the score for "scrambled egg", showing reliance on non-causal visual cues.



Example: Codependency Illusion



Isolated label modeling misses codependency cues among atomic actions, such as sharedness, exclusivity, etc.

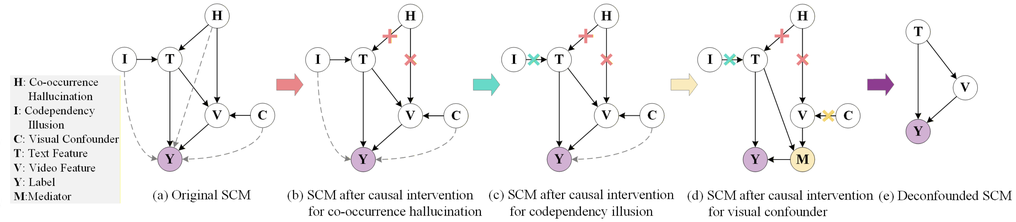
Contributions

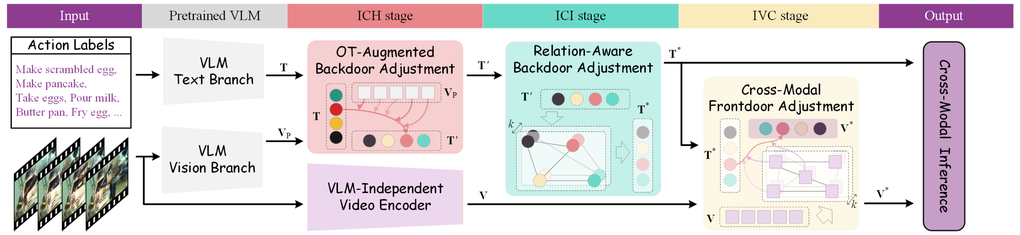
- A structural causal view of VLM-based LTAR with confounders H, I, and C.
- A progressive framework: ICH → ICI → IVC.
- SOTA performance on Breakfast, COIN, and Charades with lower inference cost.

Core idea: **make text causal first, then use it as a mediator to deconfound vision.**

Method: Progressive Causal Intervention under SCM

PCMCI transforms a confounded VLM-based LTAR model into a deconfounded cross-modal model through Intervention for Co-occurrence Hallucination (ICH), Intervention for Codependency Illusion (ICI), and Intervention for Visual Confounder (IVC).





Latent confounders induce back-door paths: $\{T,V\} \leftarrow H \rightarrow Y$ $\quad T \leftarrow I \rightarrow Y$ $\quad V \leftarrow C \rightarrow Y$

Goal: predict Y from V and T while blocking spurious effects from H, I, and C.

SCM Paths

ICH
Back-door adjustment for H

- VLM-independent encoder blocks $H \rightarrow V$
- OT-guided back-door adjustment blocks $H \rightarrow T$, yielding T'

$$S = \langle T, V^P \rangle$$

$$P^* = \underset{P \in U}{\operatorname{argmax}} \langle P, -log\ S \rangle + \lambda \mathcal{H}(P)$$

$$H = P^* \cdot V^P$$

$$P(Y \mid V, do(T)) \approx \sum_{h \in H} P(Y \mid V, do(T), h) P(h) \approx P(Y \mid V, [T, H] W^H) = P(Y \mid V, T')$$

ICI
Back-door adjustment for I

- Relational cues estimates surrogate of I
- Back-door adjustment deconfounds $T' \rightarrow T^*$, blocking $T \leftarrow I \rightarrow Y$

$$R(T') = (\bigoplus_{k=1}^{K} \mathcal{G}_{k(T';\Theta_k)}) W^R,$$

$$P(Y \mid V, do(T)) \approx \sum_{i \in I} P(Y \mid V, do(T'), i) P(i) \approx P(Y \mid V, [T', R(T')] W^I) = P(Y \mid V, T^*)$$

IVC
Front-door adjustment for C

- Deconfounded T^* serves as mediator M
- Mediator-guided visual refinement deconfounds $V \rightarrow V^*$, blocking $V \leftarrow C \rightarrow Y$

$$P(Y \mid do(V), T) = \sum_{v' \in V'} P(Y \mid do(M), v', T) P(v') = \sum_{m \in T^*} \sum_{v' \in V'} P(Y \mid m, v', T) P(m \mid V) P(v') \approx P(Y \mid [M, V'] W^C, T) = P(Y \mid V^*, T)$$

$$V' = \sigma(\mathcal{A}(M, \mathcal{F}(V); \Theta_M)) \mathcal{F}(V)$$

Experiments

97.46	90.51	86.54	53.3
Breakfast Acc	Breakfast mAP	COIN mAP	Charades mAP

PCMCI achieves competitive LTAR.

Main Comparison

Method	B-Acc	B-mAP	C-Acc	C-mAP	FLOPs	Params
Text4Vis[1]	95.49	60.49	91.98	63.39	835	287
MA-LMM [2]	93.00	71.84	93.20	67.67	29063	7526
HierarQ [3]	97.18	76.32	**94.78**	70.10	37877	7881
PCMCI	**97.46**	**90.51**	94.53	**86.54**	**650**	**211**

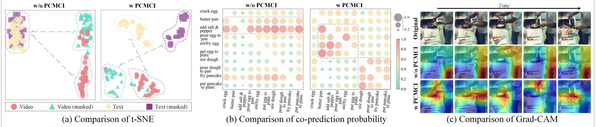
Breakfast (B) and COIN (C) results. PCMCI gains much higher mAP with far fewer FLOPs/parameters than LLM-based baselines.

Ablation + Order

Setup	ICH	ICI	IVC	B-Acc	B-mAP
Base	-	-	-	91.55	80.23
ICH+ICI	√	√	-	95.77	85.59
IVC only	-	-	√	93.24	88.31
PCMCI	√	√	√	**97.46**	**90.51**

	Intervention Order	B-Acc	B-mAP
Order 1	IVC → ICH → ICI	92.96	81.76
Order 2	IVC → ICI → ICH	92.39	80.94
Order 3	ICI → IVC → ICH	94.08	83.13
Order 4	ICI → ICH → IVC	95.49	87.43
Order 5	ICH → IVC → ICI	96.34	89.72
PCMCI	**ICH → ICI → IVC**	**97.46**	**90.51**

Visualization Evidence



(a) less hallucinated alignment; (b) better action co-dependency; (c) attention shifts from attire/background to causal regions.

Contact:
Shaowu Xu (Beijing University of Technology)
Email: swxu@emails.bjut.edu.cn; xswandsy@gmail.com
Website: https://github.com/xushaowu

References:
[1] Wu W, Sun Z, Song Y, Wang J, Ouyang W. Transferring vision-language models for visual recognition: A classifier perspective. International Journal of Computer Vision. 2024 Feb;132(2):392-409.
[2] He B, Li H, Jang YK, Jia M, Cao X, Shah A, Shrivastava A, Lim SN. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. InProceedings of the IEEE/CVF conference on computer vision and pattern recognition 2024 (pp. 13504-13514).
[3] Azad S, Vineet V, Rawat YS. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. InProceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition 2025 (pp. 8545-8556).