Supplementary Experimental: Introducing Decomposed Causality With Spatiotemporal Object-Centric Representation For Video Classification

Anonymous submission

Setting	MSR-VTT		ActivityNet	
	Top-1	Top-5	Top-1	Top-5
Subtract	61.80	85.97	88.42	**97.20**
Euclidean	62.07	**86.19**	88.40	97.19
Cosine	61.86	85.87	88.29	97.08
Mahalanobis	62.13	86.16	88.46	97.02
Manhattan	**62.23**	86.00	**88.58**	97.05

Table 1: Performance comparison of different distance metrics in the STEM module.

Setting	MSR-VTT		ActivityNet	
	Top-1	Top-5	Top-1	Top-5
Full Frame	63.29	86.07	89.26	97.83
Foreground Frame	63.54	86.33	89.57	97.92
Random Object Boxes	63.41	86.00	89.43	97.85
Tracked Object Boxes	**63.82**	**86.62**	**89.95**	**98.01**

Table 2: Comparison of different construction strategies for the global confounder dictionary G used in FDI.

Setting	MSR-VTT		ActivityNet	
	Top-1	Top-5	Top-1	Top-5
w/o sv & iv	63.51	86.33	89.63	**98.16**
w/o sv	63.67	86.46	89.69	98.14
w/o iv	63.64	86.56	89.75	98.12
CSA	**63.82**	**86.62**	**89.95**	98.01

Table 3: Comparison of different masking components in CSA, where sv represents subset visibility and iv represents intersection visibility.

Setting	MSR-VTT		ActivityNet	
	Top-1	Top-5	Top-1	Top-5
1	63.04	86.46	89.47	98.04
2	63.31	86.24	89.71	**98.16**
4	**63.82**	**86.62**	**89.95**	98.01
8	63.46	86.26	89.63	98.12

Table 4: Performance comparison with different numbers of CSA layers.

Supplementary Experimental

To further validate the effectiveness and robustness of our proposed framework, we conduct a series of in-depth analyses. These supplementary experiments provide insights from multiple perspectives: (1) the choice of distance metrics in STEM, (2) the granularity and construction strategy of the global dictionary G, (3) the functional roles of different visibility masks in CSA, (4) the impact of CSA layer depth, and (5) the sensitivity of clustering size hyperparameters in both the confounder dictionary Z and global dictionary G. Together, these analyses confirm the generalizability and interpretability of our design, revealing optimal configurations that enhance causal reasoning performance across diverse video datasets. All algorithms and experimental code will be released upon acceptance.

Supplementary In-depth Analyses

Validation of the Effectiveness of Different Distance Metrics in STEM As shown in Table 1, We validate the effectiveness of integrating STEM with different distance metrics into the base model UniFormerV2. Manhattan distance achieves the best overall performance, suggesting its strength in capturing fine-grained differences. Mahalanobis and Euclidean distances also perform competitively, while Cosine shows slight degradation, possibly due to its insensitivity to feature magnitude. Overall, STEM remains robust across distance choices, with Manhattan selected as default.

Effectiveness Analysis of the Global Clustering Dictionary G As delineated in Table 2, we construct global dictionaries G with varying granularity. Results show that using tracked object boxes, which align object-level features across videos into a shared semantic space, provides robust reference prototypes that mitigate the impact of unobserved confounders. In contrast, using full frames, foreground frames, or random object boxes leads to performance degradation, as these inputs introduce noisy pseudo-prototypes that lack consistent semantics, potentially misleading the causal learning process after intervention.

Effectiveness Analysis of Different CSA Components As shown in Table 3, we ablate the two masking components in CSA. Removing either subset visibility (sv) or intersection visibility (iv) slightly degrades performance, while removing both leads to further drops. These results confirm that sv and iv jointly facilitate structured token interactions, promoting more effective compositional reasoning. Specif-

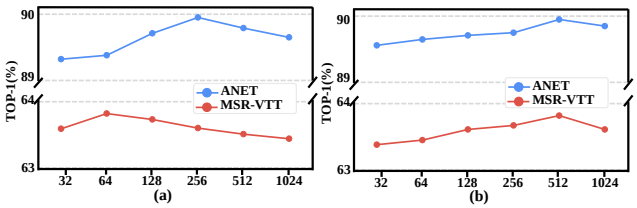

Figure 1: (a) Performance with Different Clustering Sizes of the Confounder Dictionary Z. (b) Performance with Different Clustering Sizes of the Global Clustering Dictionary G.

ically, iv ensures that only combinations sharing common objects can interact, while sv allows higher-order combinations to access their own subsets, enabling them to build semantics from simpler constituents.

Sensitivity Analysis of the CSA Layer Number Hyperparameter As shown in Table 4, we evaluate the impact of varying the number of CSA layers. Performance improves as the depth increases from 1 to 4, suggesting deeper compositional reasoning enhances representation learning. However, using 8 layers leads to slight drops, indicating that excessive depth may introduce redundancy or overfitting. We adopt 4 layers as the optimal setting.

Sensitivity Analysis of the Clustering Size Hyperparameter in Z and G As illustrated in Figure 1, we analyze the sensitivity of clustering size in both the confounder dictionary Z and the global dictionary G. For Z, performance on ActivityNet improves steadily with larger cluster sizes, peaking at 256, On MSR-VTT, however, the best performance is achieved when the cluster size is 64, after which accuracy slightly declines. This suggests that finer-grained confounder grouping benefits more complex datasets. For G, which is constructed by clustering all object-level features from the training set, similar trends are observed: both ActivityNet and MSR-VTT achieve the best performance when the number of clusters is set to 512. This indicates that clustering objects into 512 representative prototypes captures the global object patterns more effectively. From the perspective of front-door intervention, a well-structured global dictionary facilitates the construction of a robust mediator M, which serves as a pathway to mitigate spurious correlations between input features and predictions.