Supplementary

1. Visualizing the $Do_{Fourier}$ operator

Figure 2 shows the mechanism of Fourier foreground extraction. As we can see from this figure, for a non-shaking video, $Do_{Fourier}$ can be as effective as segmentation. Note that when we refer to the 'background,' it encompasses not just the traditional background scene but also clothing and other static objects. These elements can indeed convey critical insights into the action in question, which is a spurious correlation. While our segmentation method primarily isolates the human subject, it ensures that other static components remain unaltered. In contrast, the Fourier approach intrinsically filters out this static information from the video. The closer we get to retaining minimal information from the video non-action parts, the nearer we approach its purely causal form.

2. Further Ablations

Effect of Test-time Sampling Backgrounds from Source or Target Domains. We test the impact of estimating $P(BG = bg)$ in the source domain by sampling using target domain backgrounds in Table 1 for the Kinetics2Mimetics benchmark. We observe that there is actually an increase when we sample the backgrounds from the target domain instead of the source domain. We also noticed that this increase is correlated with the amount of background leakage for each operator. So, for Mask, this increase is negligible, while for Box, this increase is noticeable. This is because when there is background leakage, the model still might use leakage cues as a shortcut. So, for the bounding box, having backgrounds from the source dataset provides more cues about the wrong action.

Table 1: **Effect of sampling backgrounds from source or target domain at test time** Sampling backgrounds from both source and target domain can lead to improved domain generalization accuracy over a standard ERM baseline.

Kinetics2Mimetics	Source Domain Backgrounds	Target Domain Backgrounds
ERM baseline	31.7	31.7
$Do_{Fourier}$	41.8	42.6
Do_{Mask}	46.0	46.3
Do_{Box}	33.1	39.8

Required time for foreground extraction for different Do-operators. In Table 2, we demonstrate the time required per sample for foreground extraction. Importantly, this extraction is a one-time pre-processing step executed prior to the initiation of training. The reported metrics are based on performance using an RTX 3090 GPU. Upon obtaining the pre-extracted foregrounds, it is worth noting that the time duration for the intervention step remains consistent across all three evaluated methods. Specifically, this intervention occurs at each training epoch and takes approximately 1.8 seconds per sample when utilizing the aforementioned RTX 3090 GPU.

Table 2: **Required time for foreground extraction for different *Do*-operators** The reported metrics are in seconds using an RTX 3090 GPU.

Do-operator	Do_{Mask}	Do_{Box}	$Do_{Fourier}$
Foreground extraction time (s)	16.67	14.71	4.53

Removing Background vs. Do_{Mask} Effect. We compare the effect of removing background with our proposed Do_{Mask}. To this end, we do an intervention involving the removal of the background, subsequently replacing it with black pixels. As presented in Table 3, our findings shed light on a crucial aspect of this approach: while background removal does exhibit potential in augmenting domain generalization, it concurrently introduces a confounding factor, namely, the presence of black pixels. This implies that the number of black pixels can convey information about actions, for instance, "surfing" vs "writing." This effect is particularly pronounced in the HMDB2UCF and UCF2HMDB benchmark datasets, highlighting the persistence of residual background information despite the complete removal of the background itself. In Table 3, we compare the effect of Do_{Mask} with an intervention where the background is removed and replaced by black pixels.

Table 3: **Removing Background vs. Do_{Mask} Effect** We compare the performance of our proposed Do_{Mask} with an intervention involving the removal of the background and subsequently replacing it with black pixels.

Intervention	Kinetics2Mimetics	HMDB2UCF	UCF2HMDB
Background Removal	34.6	42.6	33.4
Do_{Mask}	38.2	59.2	44.1

Effect of Different Frequency Bands for $Do_{Fourier}$ Operator. Intuitively, frequency bands that exhibit a stronger correlation with the background inherently contain more background-related information. In the paper, we intervened on the zero frequency in time, but this intervention can expand to different frequency ranges. Table 4 shows the results from these interventions which provides insights into this relationship.

Table 4: **Effect of Different Frequency Bands for $Do_{Fourier}$ Operator** We compare the performance of our proposed $Do_{Fourier}$ with intervening on different frequency bands.

Frequency Band Range	1-end (used in Paper)	1-20	2-end
Top 1 Accuracy	41.8	33.1	38.2

Further Qualitative Results. Figure 1 shows qualitative results for the HMDB2UCF benchmark. For each example, we show the predictions of the baseline model on the original video and the prediction of our model on counterfactual videos trained with the Do_{Box}, Do_{Mask} and $Do_{Fourier}$ operators. We see that the baseline is often reliant on the background to predict the action. For instance, it can correctly predict 'shoot ball' when there is a basketball background (row 4) but struggles to predict the 'punch' action in row 2. On the other hand, our approach can predict the correct action when the background is irrelevant to the target action. However, there is still room for improvement in our model; for instance, in HMDB2UCF there are times when the baseline's reliance on the background can be useful for the model to predict the correct action (e.g., 'kick ball'). There are also cases where the background introduces objects relevant to a confounding action, for instance, the bow in row 3. There are also cases where a necessary object is removed. For example, in HMDB2UCF row 4, the action part of the video does not include the ball, meaning our model predicts the action as 'walk' for all three *Do*-operators. This suggests further