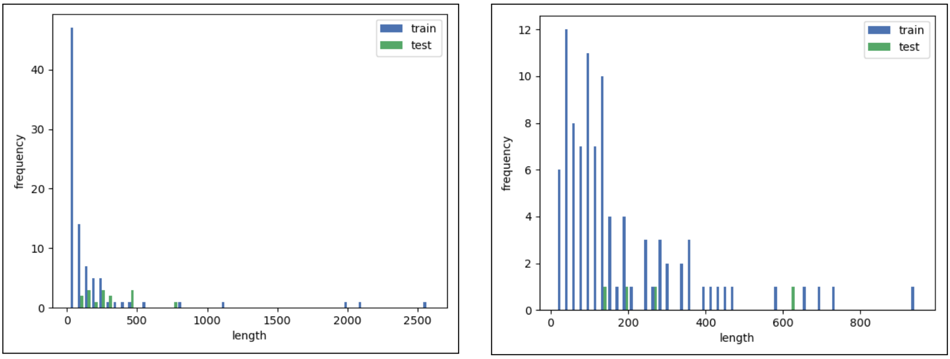
Appendix

This paper presents the issue of frame length bias in text-video retrieval and proposes a causal intervention method to alleviate the length bias. In this Supplementary Material, we provide more details regarding the following topics. Section A gives more visual clues to verify the bias for each dataset. Section B shares more information about the baseline debiasing method. Section C gives more information regarding evaluation metrics, the results and the ablation study.

A Bias Verification

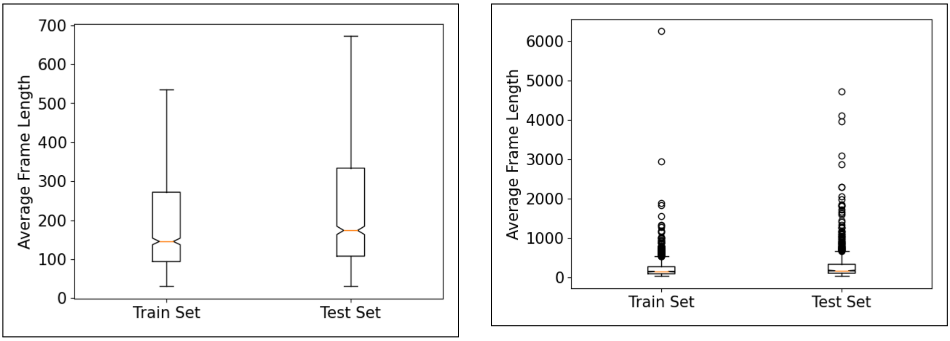
We first provide more information that supports and verifies the bias for each dataset. For Epic-Kitchens-100, Figure 6 presents the disparity for two cases. [2]



(a) Caption: *'put-down mozzarella'* (b) Caption: *'pick up rubbish'*.

Figure 6: The histogram shows the discrepancy between two <verb, noun> pairs (classes) in the Epic-Kitchens-100 dataset. The GT Recall is at 152nd rank and 169th rank, respectively.

Figure 7 depicts the discrepancy that occurs among many classes (semantic pairs of verb and noun) between the train/test set.



(a) w/o Outliers (b) w Outliers

Figure 7: Average frame length comparison between the training and test set in the Epic-Kitchens-100 dataset. It shows that clips in the test set are longer than the training set.

[2] We thank Mr. Haoxin Li for the initial discussion on verifying the bias.

We also share another version of Figure 2a in Figure 8. We get the absolute value of the discrepancy among the <verb, noun> pairs in the training and test sets and sort it in descending order, showing a long-tailed distribution. While 559 videos out of 1114 have a disparity of more than 60, 212 of them have a disparity of at least 200, showing a crucial amount of biased pairs in the dataset.

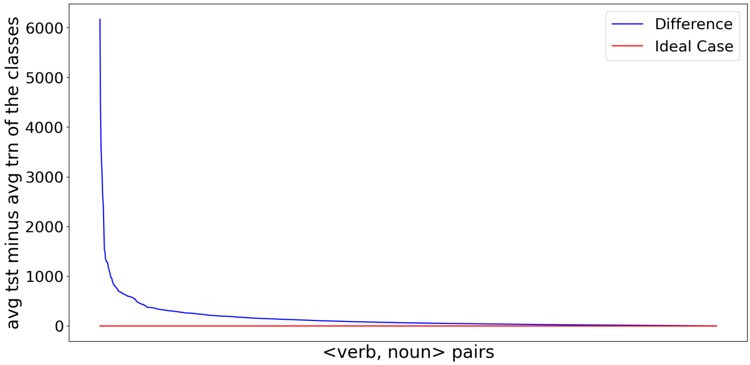
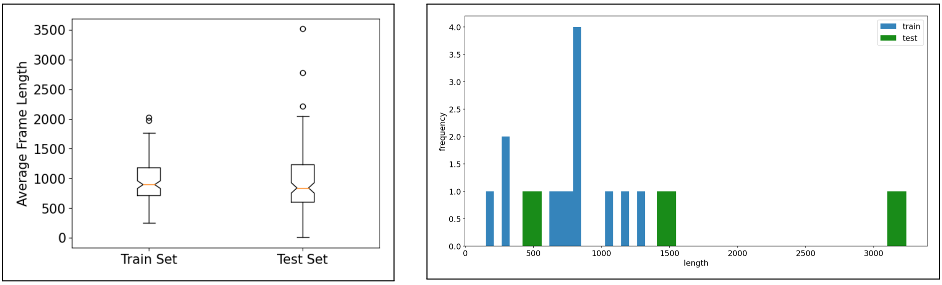


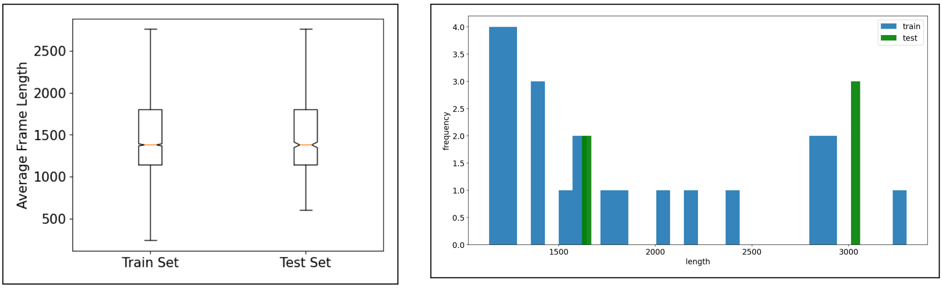
Figure 8: Long-tailed distribution among the training and test set regarding the average frame length of the classes for the Epic-Kitchens-100 dataset.

The same trend can be seen for YouCook2 in Figure 9. We note that the complex structure of the actions in YouCook2 affects the figures. For instance, 75% of the video clips have more than one action; thus, we do not know which specific actions last how long, whereas we have this information in the Epic-Kitchens-100 dataset. Thus, we assume that each action in the same video clip lasts equally with one another. MSR-VTT also includes complex actions, as it happens in YouCook2. Besides, the MSR-VTT dataset contains coarse-grained features where the captions and visual clues define more generic actions, such as walking and playing the guitar. Thus, spatial clues such as background or scene in the dataset may bring biases. These factors limit our observation to visualise the frame length bias for the whole dataset. However, we can still see a similar trend in support of proving the bias as shown in Figure 10. We argue that due to their complex and coarse-grained nature, the quantitative results are not as high as in EK-100.



(a) Comparison w Outliers (b) Caption: *'mix salad'*

Figure 9: a) The average frame length comparison between the training and test set in the YouCook2 dataset shows that clips in the test set are longer than the training set. b) An example of this disparity.



(a) Comparison w/o Outliers (b) Caption: *'make origami'*

Figure 10: a) The average frame length comparison between the training and test set in the MSR-VTT dataset shows that clips in the training set are more dense than the test set. However, the discrepancy is relatively lower in the MSR-VTT compared to other datasets. b) An example of this disparity.

B Baseline Debiasing Method

We scrutinise the effect of the frame length bias on the Epic-Kitchens-100 dataset by examining the failure cases in the text-to-video retrieval, which are the ones whose retrievals are ranked over ten. However, we need to exclude some of the failure cases to eliminate any confounders. We exclude the ones **i)** if its caption includes either a tail verb or a tail noun so that we can be sure that this failure does not happen because of the long-tail distribution, **ii)** if the average frame length difference of the class between the train and test set is lower than 60, ensure that there is enough discrepancy, **iii)** if the average frame length of retrieved top 20 video clips is closer to the average frame length of the class in the test set than the average frame length of the training set. This yields 244 samples that may be affected by this bias, making the amount non-trivial. Figure 6, 7, and 8 depict the discrepancy in terms of frame length between training and testing.

On top of the insight above, we turn captions into classes consisting of verb and noun classes by following [6] as shown in Figure 2 to depict this gap for the whole dataset. The difference between the average frame length of a class among training/test sets yields the discrepancy. We then visualise the discrepancy. One class includes a verb and noun describing the action. We also observe that the average frame length in the test set is higher than the one in the training set in each dataset.

These insights shape our motivation to address the bias. To this end, we first start with a baseline debiasing method. Table 8 shows the effect of the *RmvOne* naive method on an individual sample. It indicates that if the gap between the training and test set regarding average frame length decreases, 1) top retrieved video clips increase, and 2) ground truth value at Recall improves. As shown in the Table, if we remove the shortest video clips in this class in such a way that the average frame length between training and set sets gets similar, we realise that the retrieval result gets better as well as the average frame length of the top retrieval clips converges to average frame length of the class in the test set. Besides, when we remove the longest video clips, the opposite also happens.

Since it is impractical to see the effect of *RmvOne* on many samples, we share how *RmvAll* naive method affects the individual samples. After applying *RmvAll*, we choose 60 captions randomly out of 244 suspected samples described above. We realise that 41 out of 60 samples get affected positively, while 19 out of them get negatively. When we compare

	Avg frame length		**# of**	**GT rank**	**Avg frame length of**
Method	**Train set**	**Test set**	**samples**	**recall**	**the top 10 retrieval**
Baseline + RmvOne: Remove longest clips	**64.69**	305.5	43	261	131.2
Baseline	**177.11**	305.5	95	169	147.4
Baseline + RmvOne: Remove shortest clips	**302.32**	305.5	43	14	244.5

Table 8: Caption: 'pick up rubbish' for *RmvOne* naive method. While the red colour is directly proportional to the average frame length in the train set, the blue is inversely proportional.

the total difference between the negative and positive sides, we see a higher increase on the positive side, namely 13,321 versus 2,274. This observation double confirms the bias. We see the same trend for the other datasets as well.

Lastly, we implement our causal approach by using backdoor adjustment. It is based on the law of iterated expectations. According to this statement, the expected value of a random variable can be obtained by summing the expected values of that random variable when it is conditioned on a second random variable. To perform better in applying this method, we split the dataset into subsets so that the discrepancy in the subsets would be as much lower as possible. On the contrary, we may divide the dataset into equal splits through which various subsets could still contain high discrepancies. We follow the Algorithm 2 when the training set includes more short video clips than long ones.

Algorithm 2 Divide the dataset into M subsets in an adjusted way:

```
$V \leftarrow$ videos in ascending order based on the frame length
$M \leftarrow$ number of splits
$i = 1$
$0 < th < 1$                                   ▷ Threshold to decide the amount for the last two splits
while $M \geq 1$ do
    $M_i = V$[:len($V$)/2]
    if $M == 2$ then
        $M_i = V$[:th*len($V$)]
        $M_(i+1) = V$[th*len($V$):]
        break;                                 ▷ The algorithm always finishes here
    end if
    $i$+=1
    $M$-=1
    $V = V$[len($V$)/2:]
end while
```

C Results

C.1 Implementation Details

By following the algorithm above, Epic-Kitchens-100's split includes 50,282 video clips, while the second has 16,935 video clips when M is selected as two. In YouCook2, the first split includes 8,000 video clips, while the second has 2,337 of them. In MSR-VTT, the first split includes 5,950, while the second has 3,050 video clips.

Another thing is that our causal approach is faster than the baseline approach. One epoch takes 305 seconds in the baseline method on the Epic-Kitchens-100 dataset. However, it takes 244 seconds in our method, while the first split and the second split take 175 seconds and 69 seconds, respectively. We use one NVIDIA RTX 2080 Ti for our experiments. In terms of sampling the positive and negative pairs, we follow the baseline study.

We assign only one relevant video to the query for text-to-video retrieval tasks using conventional evaluation metrics, such as Recall and mAP. We assign all the other videos completely irrelevant, even if there could be somewhat relevant videos. To this end, we use the nDCG metric by calculating a relevancy matrix between the captions. This calculation is done via Formula 6, setting non-binary relevancy. For instance, we assume that captions are 'cut tomato', 'cut chicken' and 'take plate'. If we have a query called 'cut tomato', we calculate R as the list of [1, 0.5, 0] when we apply Formula to these captions. However, it would be [1, 0, 0] for the conventional evaluation metrics.

C.2 Quantitative Results

Table 9 shows that our causal method overpasses the baseline and SOTA methods for MSR-VTT full split, as well.

Method	**Recall (T2V)**						**nDCG**		
	R@1↑	**R@5↑**	**R@10↑**	**MedR↓**	**MnR↓**	**Rsum↑**	**V2T↑**	**T2V↑**	**AVG↑**
MSR-VTT Full Split									
Baseline	9.39	27.60	39.01	20	138.74	76	26.31	24.87	25.59
Baseline + Ours	**10.42** **(+1.03)**	**29.51** **(+1.91)**	**41.68** **(+2.67)**	**16** **(-4)**	**84.28** **(-54.46)**	**81.61** **(+5.61)**	**28.27** **(+1.96)**	**26.35** **(+1.48)**	**27.31** **(+1.72)**
RAN	9.80	27.20	38.28	20	133.53	75.28	26.88	25.73	26.31
RAN + Ours	*10.90* *(+1.10)*	*30.11* *(+2.91)*	*42.20* *(+3.92)*	*16* *(-4)*	*84.68* *(-48.85)*	*83.21* *(+7.93)*	*28.73* *(+1.85)*	*26.61* *(+0.88)*	*27.67* *(+1.36)*
RANP	9.83	27.74	39	20	134.82	76.57	26.94	25.37	26.16
RANP + Ours	*10.69* *(+0.86)*	*30.18* *(+2.44)*	*42.50* *(+3.50)*	*15* *(-5)*	*87.06* *(-47.76)*	*83.37* *(+6.80)*	*28.38* *(+1.44)*	*26.36* *(+0.99)*	*27.37* *(+1.21)*

Table 9: Baseline and SOTA comparison on text-video retrieval for MSR-VTT full split. The lower, the better for MedR and MnR metrics; the higher, the better for the rest.

Regarding the models' effect on transformer-based models. Due to our limited computation resources, we are unable to experiment with recent transformer-based methods based on heavy backbones. Even so, we argue that it would not be a fair comparison as spatial bias would heavily affect them during the finetuning, as suggested by a recent study [34]. Moreover, another recent paper [43] claims that they address spatial bias by using a sampling policy similar to RAN and RANP methods. Thus, our method can address the temporal bias even after mitigating the spatial bias, which would address the impact of the bias. Nevertheless, we plan to investigate the relationship between spatial and temporal biases more.

C.3 Qualitative Results

Figure 11 shows the failure cases on our causal method, which could occur due to various factors. For instance, considering the caption in the Epic-Kitchens-100 dataset (t2v), we see worse performance than the baseline, although the outcome is reasonable. However, we note that the caption contains a tail noun, *oven mitt*. While a split may learn this feature, another split may not learn it due to fewer examples. Moreover, just summing up the similarity feature may not guarantee to preservation of the learned feature regarding various actions.

If we examine another caption in the YouCook2 dataset (t2v), we can see that the causal approach reach better result within the first ten retrieved video clips. However, the baseline method gets a higher result within fifty retrieved clips. We assume that the ambiguity of the noun, *oil*, may affect the result. In other words, the difference in vector space regarding the noun could be high among the splits.

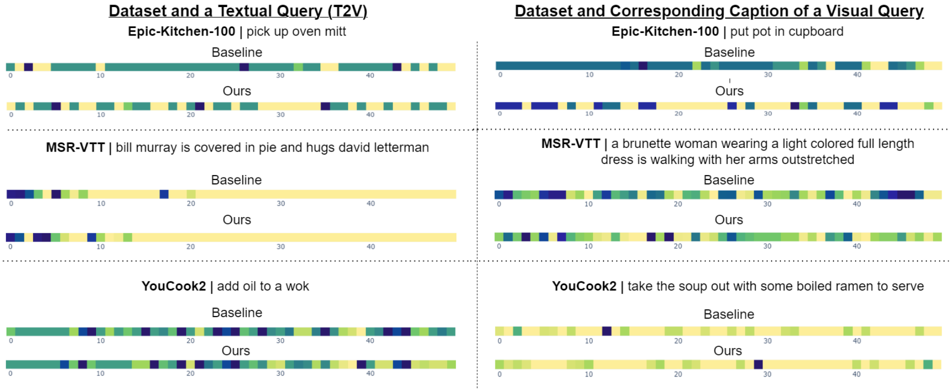


Figure 11: Qualitative results for text-video retrieval showing the failure cases. The semantic relevancy, calculated based on nDCG, of the top 50 retrievals given a query from each dataset. The darker the colour, the more relevant retrievals to the query, varying from 0 to 1. While the left side is for T2V and the right is for V2T. Best viewed in colour.

C.4 Ablation Study

Table 10 shares the result for an ablation study between an ensemble method and our causal approach in each dataset.

Method	Epic-Kitchens-100	YouCook2	MSR-VTT
	nDCG (avg)		
Baseline	39.15	49.56	60.30
Baseline + Ensemble	39.76	49.80	60.60
Baseline + Ours	**41.67**	**51.65**	**62.50**

Table 10: Ablation study for the ensemble method.

A failure method. We also try another approach; however, it does not bring any sharp increase in Precision or Recall as it gets sharply lower results in the nDCG metric. In this approach, we first train a model on the full dataset. Then, we finetune it on the data splits. We assume that finetuning works as parameter sharing, which harms the backdoor adjustment; in other words, it re-creates the link between L and V, which is shown in Figure 3.

Future Work. We aim to address the same bias in related tasks such as video corpus moment retrieval, video moment retrieval [13] and video localisation [33]. Besides, we plan to address various biases in a collaborative method.