Towards Debiasing Frame Length Bias in Text-Video Retrieval via Causal Intervention

Burak Satar[1,2]
burak001@e.ntu.edu.sg
Hongyuan Zhu[1]
zhuh@i2r.a-star.edu.sg
Hanwang Zhang[2]
hanwangzhang@ntu.edu.sg
Joo Hwee Lim[1,2]
joohwee@i2r.a-star.edu.sg

[1] Institute for Infocomm Research
A*STAR
Singapore

[2] School of Computer Science and Engineering
Nanyang Technological University
Singapore

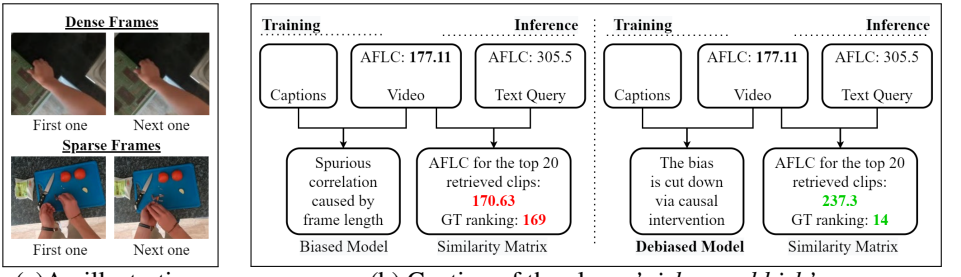
Abstract

Many studies focus on improving pretraining or developing new backbones in text-video retrieval. However, existing methods may suffer from the learning and inference bias issue, as recent research suggests in other text-video-related tasks. For instance, spatial appearance features on action recognition or temporal object co-occurrences on video scene graph generation could induce spurious correlations. In this work, we present a unique and systematic study of a temporal bias due to frame length discrepancy between training and test sets of trimmed video clips, which is the first such attempt for a text-video retrieval task, to the best of our knowledge. We first hypothesise and verify the bias on how it would affect the model illustrated with a baseline study. Then, we propose a causal debiasing approach and perform extensive experiments and ablation studies on the Epic-Kitchens-100, YouCook2, and MSR-VTT datasets. Our model overpasses the baseline and SOTA on nDCG, a semantic-relevancy-focused evaluation metric which proves the bias is mitigated, as well as on the other conventional metrics.[1]

1 Introduction

In text-video retrieval, nowadays, the state-of-the-art models [9, 19, 26, 27, 28, 44] can achieve promising performance on famous benchmarks [25, 35, 46]. However, recent studies [29, 31, 39, 40] demonstrate that many existing visual-text models are overly affected by superficial correlations. For instance, some works [2, 11, 15] address the static appearance bias for action recognition. While [3] focuses on object co-occurrences that bring spurious correlations specifically in the spatial domain, [23, 37] examine the same topic in the temporal domain. Some other works reveal the correlation between the start-end time of the actions and the actions themselves in untrimmed videos on video moment retrieval [39, 40] and temporal sentence grounding [18, 42] tasks. Unlike these studies, we focus on a temporal bias that has yet to be addressed in text-video-related tasks. Frame length discrepancy between training and test sets of trimmed video clips causes non-relevant retrieved items.

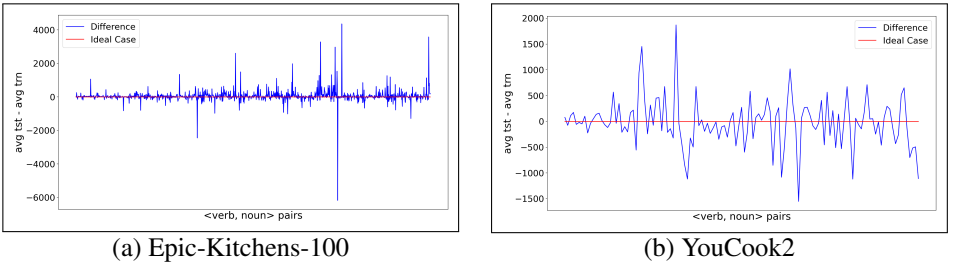
[1] https://buraksatar.github.io/FrameLengthBias/



(a)An illustration. (b) Caption of the class: *'pick up rubbish'*.

Figure 1: a) Motion semantics may differ between long and short video clips when the frames are uniformly sampled. If it is unbalanced between training/test sets, this may introduce frame length bias. b) AFLC denotes the average frame length of a class, meaning <verb, noun> pairs (classes) which affect the retrieved clips. We propose a novel causal intervention method to remove this spurious correlation.

For example, in the case of text-to-video retrieval, as shown in Figure 1, the top twenty retrieved clips' average frame length is similar to the training class's average frame length, stating that some irrelevant clips are retrieved just because of the bias coming from the discrepancy. We refer to 'class' as a joint combination of 'verb class' and 'noun class' by considering the verb and noun tokens together. Some classes can be semantically similar. For instance, 'take' and 'pick up' would be in the same verb class. Thus, we use the notion of class to calculate a matrix, measuring semantic relevancy among verbs and nouns. In addition, we utilise it to identify the biases in Figure 2, showing the discrepancy. Only a recent work [41] closer to our approach attempts to mitigate video duration bias in watch-time prediction for video recommendation. However, the proposed model uses the video duration as textual input and does not consider any visual feature via any visual sampling method. They apply causal inference based on video duration while the discrepancy in the video duration is not considered, and it is followed by a pre-text task of watch-time prediction to increase the effect.



(a) Epic-Kitchens-100 (b) YouCook2

Figure 2: The figures show the discrepancy among all the <verb, noun> pairs (classes) in each dataset, which is calculated by the average frame length difference between the training and test sets. The number of pairs in the X-axis is 1,144 and 125, respectively. See the Supplementary Material for more details regarding verifying the bias in the three datasets.

To address overlooked frame length bias in text-video retrieval, we first apply baseline debiasing methods, which delete either the shortest or longest video clips in a class to reduce the discrepancy between train and test sets. However, the effect is limited. Then, we intervene in the causal graph to remove the frame length's unwanted impact by applying

the backdoor adjustment principle [24]. Specifically, we divide the training data into splits regarding frame length; for each split, we learn a similarity matrix using the same text-video retrieval model. Then, we sum the similarity matrices. Note that we also consider the discrepancy within the splits regarding frame length to increase the debiasing effect. The contributions of this paper are threefold: **i)** To the best of our knowledge, we are the first to address a temporal bias in text-video retrieval tasks and also the first to address frame length bias in any text-video-related tasks. We verify the bias illustrated with various methods. **ii)** We propose a causal inference approach via backdoor adjustment to mitigate the frame length bias. **iii)** The experiments and ablation study verify the advantages of the proposed approach over the baseline and SOTA studies by evaluating retrieved clips semantically via Discounted Cumulative Gain (nDCG) as well as Recall and mAP.

2 Related Work

Text-Video Retrieval. In text-video retrieval, which aims to rank samples in a modality given another modality, deep learning-based approaches have emerged as promising techniques due to their ability to learn high-level features directly from the data. One popular method is to encode text and video features into a common space [20, 21, 22], where the similarity can be measured using various distance metrics. Another approach is to utilise the semantic relationships between text and video features [5, 7, 26]. For instance, Chen *et al*. [5] use semantic role labelling to capture the relationship between verbs and nouns in text and actions and objects in videos. Besides, Falcon *et al*. [7] implements a positive and negative sampling strategy based on semantic similarities between verb and noun pairs. Recent models based on visual transformers have shown promising results with the help of pre-training on giant datasets [21]. For example, Bain *et al*. [1] use raw video frames rather than extracted features and apply attention mechanisms for pre-training on various exocentric video datasets. On top of this work, Lin *et al*. [17] pre-train the modified model on an enormous egocentric dataset curated from Ego4D [10]. Nevertheless, further research is needed to address existing biases in the task.

Biases in Video-Language. Recent studies have highlighted the presence of biases in video-language tasks, which can affect the performance of models since the models can rely on spurious correlations in the data rather than genuine causal relationships. For instance, temporal [37, 40] and spatial [2, 3, 11, 15] biases may arise due to the nature of the data collection process, where certain activities or scenes may be over-represented or under-represented [31, 39, 42]. In addition, certain words or phrases may be over-represented in the captions, leading to a bias towards those concepts [18, 23]. However, various biases are overlooked, and it is crucial to understand the sources of these biases. In this respect, we address the frame length bias.

3 Method

3.1 Base Model

Given its state-of-the-art performance, we follow *Chen et al.* [5] for the baseline work. The model contains two encoders, one for text and, one for video, and a text-video matching part.

Textual Encoding. We disentangle the textual features hierarchically by utilising a pre-existing semantic role labelling tool [30] to comply with disentangled video features. For

example, whereas a sentence could define global features, local features are represented by words that refer to actions and entities. We establish the connection between actions and entities as r_{ij}, where i denotes action nodes and j denotes entity nodes. Subsequently, the semantic role matrix W_r, which is designed to accommodate various semantic roles, is multiplied with initialised node embeddings $g_i^0 = g_i \odot W_r r_{ij}$ such that $g_i \, \varepsilon \, \{g_e, g_a, g_o\}$. The one-hot vector r_{ij} indicates the edge type from node i to node j, while $\odot$ signifies element-wise multiplication. Then, a graph-attention network is employed to process adjacent nodes. W_t matrix, which is utilised for all relationship varieties, exploits attended nodes, as shown in Eq. 1. When attention is applied to each node, the result is referred to as β. Once these formulas are applied, we obtain the textual representation for global and local features $c_i \, \varepsilon \, \{c_e, c_a, c_o\}$; for sentence node, verbs, and words, respectively.

$$g_i^{l+1} = g_i^l + W_t^{l+1} \sum_{j \varepsilon N_i} \beta_{ij}(g_j^l) \tag{1}$$

Video Encoding. Disentangling videos into hierarchical features can be challenging, although it is comparatively simple to parse language queries into hierarchical features. To this end, we employ three distinct video embeddings that concentrate on various levels of video aspects. Given a video, denoted as V, represented as a sequence of frame-wise features $\sum_{i=1}^{M} f_i \, \{f_1, ..., f_M\}$, we apply different weights to generate embeddings for three different levels, which are then incorporated with a soft attention mechanism.

$$v_{x,i} = \sum_{i=1}^{M} W_x^v f_i, \qquad x \, \varepsilon \{e, a, o\} \tag{2}$$

Text-Video Matching. The matching score is computed by averaging the cosine similarity with the video and textual embeddings. We use the contrastive ranking loss [4] by attempting to have positive and negative pairs larger than a predetermined margin in training. Suppose v and c symbolise visual and textual representations; positive and negative pairs can be formulated as (v_p, c_p) and (v_p, c_n) / (v_n, c_p), respectively. A pre-set margin named *Delta* is used to determine contrastive loss.

$$s(V,C) = \sum_{i=1}^{3} \frac{< v_i, c_i >}{||v_i||_2 ||c_i||_2} \tag{3}$$

$$L(v_p, c_p) = [\Delta + s(v_p, c_n) - s(v_p, c_p)] + [\Delta + s(v_n, c_p) - s(v_p, c_p)] \tag{4}$$

3.2 Baseline Debiasing Method

Algorithm 1 Delete the shortest clips. For each class in the common class set:

```
V ← videos in ascending order based on the frame length
x ← avg frame length of class for the training set & y ← avg frame length of class for the test set
while y ≥ x + δ do
    Delete the first clip v_0 from the training set
    if len(V) ≤ α then
        break;
    end if
end while
```

We can naively remove this bias by following two methods. In the first method, *RmvOne*, we delete the shortest and longest class samples so that the training set's average frame length

becomes similar to the test set for only one class. Note that the class notion refers to <verb, noun> pairs to group the captions semantically. However, this method does not affect the evaluation metrics, but only a few samples. Thus, another simple method, *RmvAll*, can be suggested. We do the same as in *RmvOne*, but considering all classes such that the high discrepancy will be reduced in the whole dataset to a pre-set margin δ. We set the minimum number of video clips of a class in training as α so that there are enough samples for each class. It aims to reduce discrepancies between training and test sets for the same classes. Specifically, Algorithm 1 presents the way if the average test set is higher than the average train set and removes the shortest clips. The same logic applies when the situation is the opposite, which deletes the longest clips.

3.3 Method with Causal Intervention

Many works use extracted features that are uniformly sampled in order to remove the effect of frame length. However, these features may still contain bias due to the sparsity or density of the sampled frames, as shown in Figure 1. Thus, the model learns that action should be dense or sparse rather than motion semantics. The ideal case would be to have all the video clips at the same length. It is not just impractical but would also not reflect real-world applications. For example, while some actions take more time, others take less time intrinsically. Thus, while we need to keep this natural connection, we should remove the spurious correlation on video features that would occur because of the discrepancy in terms of frame length. Figure 3 shows our structural causal model (SCM) to illustrate how our model works. V, Q, Y and L denote video representation, textual representation, text-video matching and frame length, respectively. The link from (V, Q) to Y is for capturing the similarity between the textual and visual features. The link from L to Y signifies the frame length effect on similarity, suggesting that while some actions can take less time, others would take more time. Moreover, the link from L to V implies that frame length would affect the video encoder such that various videos could be retrieved not because of their semantic similarity to the query but instead of their frame length. If this bias is not addressed, densely sampled video features would be memorised in case the training set contains mostly shorter clips than the testing set.

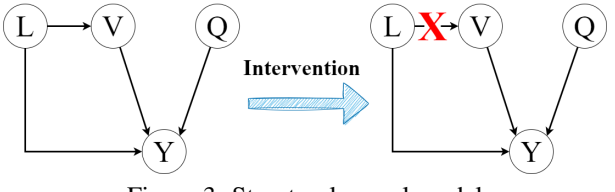


Figure 3: Structural causal model.

Figure 4 shows the implementation of two splits for a dataset based on the frame length. As a high-level idea, we follow the principle of backdoor adjustment to remove bias by splitting the dataset based on frame length. We formalise our causal method in Formula 5 by using the law of iterated expectations. Note that L becomes independent via interventions. As shown in the last row of the formula, the final estimation can be created by individually estimating $P(L)$ and $E[Y|V,Q,L]$ and then combining those estimates. We divide the training samples into M equal portions based on frame length to cut off the link, discretising the $P(L)$ distribution into separate components. These frame length groups are denoted by

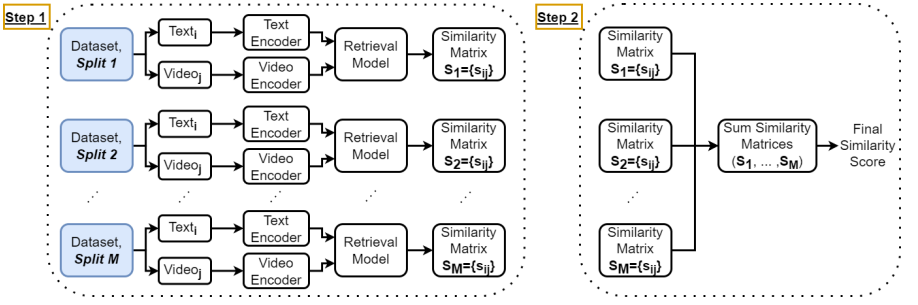


Figure 4: The implementation of the causal model for training. Similarity matrices are constructed using the same retrieval model on different splits that are arranged with a causal perspective to mitigate frame length bias. Then, they are summed up. No change is needed for the inference.

$\{L_k\}_{k=1}^{M}$. We estimate the deconfounded model via this approximation. Note that $f_k(v,q)$ is the similarity score for each frame length group L_k.

$$\begin{aligned} E[Y|do(V,Q)] &= \sum_l P(L=l|V,Q)E[Y|V,Q,L=l] \\ &= \sum_l P(L=l)E[Y|V,Q,L=l] \\ &\approx \sum_{k=1}^{M} (L_k)E[Y|V,Q,L\in L_k] \\ &\triangleq \sum_{k=1}^{M} (L_k)f_k\{V,Q\} \end{aligned} \tag{5}$$

4 Experiments

Datasets. We use three datasets for our experiments. **i)** *Epic-Kitchens-100 (EK-100)* [6], a collection of unscripted egocentric action data gathered worldwide using wearable cameras. The annotated videos display diverse daily kitchen activities, accompanied by captions provided by human annotators that include at least one verb and one or more nouns. The dataset comprises 67,217 training and 9,668 test set pairs. **ii)** *YouCook2* [45, 46], which is from cooking-related videos via third-person viewpoint collected from YouTube with 89 different recipes. The video clips are recorded from a third-person viewpoint within diverse kitchen settings. Imperative English sentences and temporal boundaries referencing the actions are used to label the video clips, and human annotators are used. There are 10,337 pairs in the training set and 3,492 pairs in the test set. **iii)** *MSR-VTT dataset* [36] comprises 10,000 video clips with 20 descriptions for each video, a combination of human annotation and a commercial video search engine. The dataset offers several train/test splits, with one of the most popular ones being the 1k-A split, consisting of 9,000 clips for training and 1,000 clips for testing. The full split, which consists of 6,513 video clips for training, 2,990 video clips for testing, and 497 video clips for validation, is another often-used split.

Implementation details. We use the video features that TBN [14] has extracted for Epic-Kitchens-100. Each video clip included RGB, flow, and audio features. Note that we

use the frame itself rather than using extracted features for replicating a sota work, EgoVLP. We utilise S3D features from *Li et al.* [16] pretrained on HowTo100M[21] for YouCook2. Since the test set is not made available to the public, we feed our model with the validation dataset for evaluation in accordance with other studies. For MSR-VTT, appearance level features of the ResNet-152 model provided by Chen *et al.* [4] are implemented. The epoch is chosen as 100 for all. Δ is determined as 0.2 by following the baseline model. δ and α are chosen as 10 and 60fps, respectively, for our baseline debiasing method. For SOTA methods RAN and RANP, negative and positive sampling thresholds are selected as 0.75 and 0.20, respectively. We report the best results out of three repetitions.

Evaluation metrics. We use the nDCG [12] by considering non-binary similarity to show how the bias affects various retrieved video clips and how the causal model mitigates the bias. Given a caption query q_i and a ranked list of video clips X_r, it [32] is defined as $nDCG(q_i,X_r) = \frac{DCG(q_i,X_r)}{IDCG(q_i,X_r)}$. Then, DCG is calculated as $DCG(q_i,X_r) = \sum_{j=1}^{N_r} \frac{R(q_i,x_j)}{log_2(j+1)}$, and the ranking list only considers the first N_r items, while x_j is the j-th item in the list X_r. *IDCG* is calculated via *nDCG* and is the ideal case where X_r is ordered by relevancy. R, the relevancy matrix, is between 0 and 1 and represents the mean Intersection over Union (IoU) for the verb and noun classes. We follow [6] to define the R matrix as between a caption q_i and a video x_j by averaging the IoU of verb and noun classes. While q_i^v refers to the collection of verb classes in the caption, x_k^N denotes the set of noun classes in the video clip.

$$R(q_i,x_j) = \frac{1}{2}\left(\frac{|q_i^v \cap x_j^v|}{|q_i^V \cup x_j^v|} + \frac{|q_i^N \cap x_j^N|}{|q_i^N \cap x_j^N|}\right) \tag{6}$$

By using the same logic, *nDCG* is defined for a query video x_i and a set of captions C_r. We follow the scripts provided by [32] to create the relevancy matrices for the datasets. We utilise the mean average precision (mAP) and Recall (R@k) for a fair comparison.

4.1 Results

Quantitative Results. The first baseline debiasing method, *RmvOne*, is impractical to repeat for all video classes, although it works for many examples. Table 1 shows a result on the following baseline debiasing method, *RmvAll*, for Epic-Kitchens-100. Specifically, we delete 2,392 clips from 164 classes, equivalent to 3.6% of all the data, applying Algorithm 1. It reaches marginally higher results on nDCG, even though it uses fewer data for training. Considering that we lose some information for many classes, such as diverse and complex visual cues, it is reasonable not to see a sharp increase by this naive method. We also compare it to the model that randomly deletes the same amount of video clips called *RmvRand*, showing that knowing which clips to remove is essential. Although the ensemble approach overpasses the baseline, it is still lower than our method. Besides, its training takes three times more than ours; more importantly, its nDCG score is much lower than our approach, showing that the ensemble method does not address the bias as much as our causal model.

Tables 1-3 show the results of the causal method when M is chosen as 2. Specifically, the dataset is divided into two splits based on the frame length by considering the distribution of the dataset. Rather than having equal splits, we make one split that has more video clips than the other to have less discrepancy within the splits in terms of frame length. We choose the mean length of the test set as a threshold for splitting. Considering the baseline comparison, the average scores for nDCG increase by more than 2 points in each dataset. We see a similar trend for Recall and mAP metrics. A reasonable increase is observed when we apply

our method to SOTA methods. Since these methods implement a specific scheme for positive and negative sampling, they force fewer pairs to match anchor samples when the dataset is split, which may limit the increase. Note that 'T2V' refers to text-to-video, and 'V2T' refers to video-to-text. Refer to the Supplementary Material to see the results of the causal method on the MSR-VTT's full split and more detail on baseline debiasing experiments.

Method	nDCG			mAP		
	V2T	T2V	AVG	V2T	T2V	AVG
Epic-Kitchens-100						
Baseline	39.40	38.91	39.15	40.47	36.60	38.54
Baseline + RmvRand	39.69	38.42	39.06	40.37	35.7	38.04
Baseline + RmvAll	40.06	38.82	39.44	41.01	36.34	38.67
Baseline + Ensemble	40.38	39.15	39.76	43.17	38.80	40.98
Baseline + Ours	**42.73 (+3.33)**	**40.61 (+1.70)**	**41.67 (+2.52)**	**45.36 (+4.89)**	**37.80 (+1.20)**	**41.58 (+3.04)**

Table 1: Baseline comparison on text-video retrieval for Epic-Kitchens-100.

Method	nDCG (AVG)	mAP (AVG)
Epic-Kitchens-100		
RAN	41.06	39.46
RAN + Ours	*41.84 (+0.78)*	*41.24 (+1.78)*
RANP	43.14	43.77
RANP + Ours	*43.80 (+0.66)*	*44.12 (+0.35)*

Table 2: SOTA comparison on text-video retrieval for Epic-Kitchens-100.

Method	Recall (T2V)						nDCG		
	R@1↑	R@5↑	R@10↑	MedR↓	MnR↓	Rsum↑	V2T↑	T2V↑	AVG↑
YouCook2									
Baseline	13.17	36.31	50.74	10	66.47	100.23	49.42	49.70	49.56
Baseline + Ours	**14.60 (+1.43)**	**37.80 (+1.49)**	**51.58 (+0.84)**	10	**63.18 (-3.29)**	**103.98 (+3.75)**	**51.92 (+2.50)**	**51.39 (+1.69)**	**51.65 (+2.09)**
RAN	13.29	36.37	50.40	10	64.85	100.06	50.17	50.35	50.26
RAN + Ours	*14.92 (+1.63)*	*37.37 (+1.00)*	*50.86 (+0.46)*	10	*63.78 (-1.07)*	*103.15 (+3.09)*	*50.97 (+0.80)*	*51.25 (+0.90)*	*51.11 (+0.85)*
RANP	13.63	35.65	50.32	10	64.34	99.60	50.49	50.19	50.34
RANP + Ours	*15.23 (+1.60)*	*37.60 (+1.95)*	*51.58 (+1.26)*	10	*61.34 (-3.00)*	*104.41 (+4.81)*	*51.53 (+1.08)*	*51.05 (+0.86)*	*51.29 (+0.95)*
MSR-VTT 1kA Split									
Baseline	20.76	47.29	59.92	6	41.10	127.97	59.77	60.84	60.30
Baseline + Ours	**24.64 (+3.88)**	**52.99 (+5.70)**	**66.09 (+6.17)**	**5 (-1)**	**26.26 (-14.84)**	**143.72 (+15.75)**	**62.67 (+2.90)**	**62.33 (+1.49)**	**62.50 (+2.20)**
RAN	21.08	47.98	60.95	6	42.28	130.01	59.49	60.15	59.82
RAN + Ours	*24.54 (+3.46)*	*53.50 (+5.52)*	*66.70 (+5.75)*	*5 (-1)*	*26.91 (-15.37)*	*144.74 (+14.73)*	*60.95 (+1.46)*	*61.86 (+1.71)*	*61.41 (+1.59)*
RANP	21.14	47.72	60.32	6	41.66	129.18	59.94	60.55	60.25
RANP + Ours	*24.03 (+2.89)*	*53.24 (+5.52)*	*66.53 (+6.21)*	*5 (-1)*	*27.35 (-14.31)*	*143.81 (+14.63)*	*61.54 (+1.60)*	*61.58 (+1.03)*	*61.56 (+1.31)*

Table 3: Baseline and SOTA comparison on text-video retrieval for YouCook2 and MSR-VTT. The lower, the better for MedR and MnR metrics; the higher, the better for the rest.

Qualitative Results. Figure 5 shows qualitative examples, proving that the bias is mitigated. We utilise the nDCG metric, knowing that we cannot examine this by using only Recall or mAP metrics due to their nature of binary similarity. Regarding text-to-video retrieval on the left side of the figure, the top retrieved video clips and the neighbour clips become more relevant than the baseline in the first example. In the second example, the top retrieved clip is already related to the query in the baseline model; however, the causal model eliminates most of the unrelated clips and provides more relevant clips in total. The third example's query is complex, but our approach still outperforms the baseline. On the right side, for video-to-text retrieval, queries are videos, and we retrieve the textual queries. However, for simplicity, we report their corresponding captions. Darker colours refer to higher relevancy. Please refer to the Supplementary Material for more analysis.

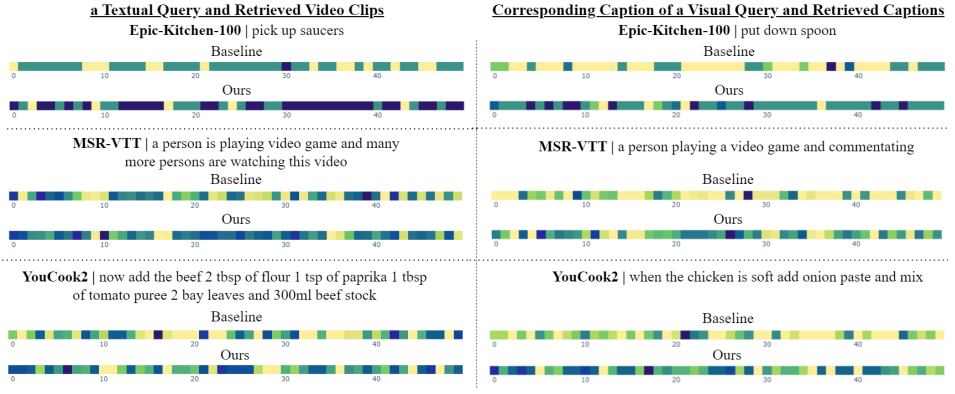


Figure 5: Qualitative results for text-video retrieval. The semantic relevancy, calculated based on nDCG, of the top 50 retrievals given a query from each dataset. The darker the colour, the more relevant retrievals to the query, varying from 0 to 1. While the left side is for T2V, the right side is for V2T. Best viewed in colour.

4.2 Analysis

Ablation Study. Table 4 examines three questions: **i)** *How to split the dataset?* When we adjust the splits based on the frame length distribution of the dataset rather than dividing them into two equal splits, we reach a higher result. Adjusted splits have higher entropy, bringing better cooperation between splits. **ii)** *Which split effects more?* When the splits are adjusted according to the frame length distribution, they share a similar score in nDCG, even though the second split has fewer video clips. Also, the first split brings a higher score in mAP/Recall, as expected. **iii)** *How many splits do we need?* The more splits we have, the lower the scores we get. To have the adjusted splits when $M > 2$, we first put the videos in ascending order according to the length of the frame, then divide them into two and continue to divide the remainder until we reach enough splits for the experiments.

Method	Epic-Kitchens-100		YouCook2			MSR-VTT		
	nDCG (avg)↑	mAP (avg)↑	nDCG (avg)↑	R@10↑	MnR↓	nDCG (avg)↑	R@10↑	MnR↓
Baseline	39.15	38.54	49.56	50.74	66.47	60.30	59.92	41.10
Baseline + Ours (Equal 2 Splits)	41.19	41.07	51.01	51.15	66.22	62.18	66.98	26.65
Baseline + Ours (Adjusted 2 Splits)	41.67	41.58	51.65	51.58	63.18	62.50	66.09	26.26
Baseline + Ours (Adjusted 3 Splits)	41.06	39.48	51.45	49.28	68.94	62.48	64.56	30.54
Baseline + Ours (Adjusted 4 Splits)	39.89	37.54	51.64	47.11	74.33	62.23	61.93	33.74
First Split Only (Adjusted)	37.24	38.59	48.86	48.42	80.44	61.02	52.79	50.71
Second Split Only (Adjusted)	38.00	34.07	50.48	36.77	115.87	59.75	53.72	50.17

Table 4: Ablation study for the causal method.

Computational Analysis. Table 5 presents the computational cost breakdown, implemented by THOP library [47], on the YouCook2 dataset where the dimensions of video embedding and batch size are 1024 and 64, respectively. Considering the causal method reaches better results with two splits, we highlight two points: **i)** If it is used sequentially, there is no need for extra resources compared to the baseline method. The advantage of the

causal method is that run time takes 20% less for training. **ii)** If the splits are trained simultaneously, the run time can drop 50% by doubling parameters and GFLOPs in return. iii) We get a similar trend in all parameters for EK-100 and MSR-VTT datasets. The only difference is that the parameters and GFLOPs are proportional to the dimension. For instance, the visual feature dimensions in EK-100 and MSR-VTT are 2048 and 3072, respectively. Thus, our approach provides a faster run time without extra resources and latency.

	Text Enc	Video Enc
Parameter (M)	10.25	3.15
GFLOPs (per clip)	11.93	4.03

	Run Time (s)
Baseline	35
Causal (Split1/Split2)	18 / 10

Table 5: Computational cost breakdown on YouCook2 dataset.

Method	nDCG (avg)	mAP (avg)
Epic-Kitchens-100		
Baseline	38.12	39.79
w/o audio	36.55	37.77
w/o spatial	36.63	35.63
w/o temporal	32.56	32.50

Table 6: Comparison between spatial and temporal features.

Spatial vs Temporal features. Table 6 shows the importance of temporal features in the Epic-Kitchens-100 dataset such that removing them affects the result drastically. While we specifically focus on an overlooked temporal bias in this study, we note that biases in both domains should be addressed in the ideal case even though no study has achieved it yet.

The models' effect on transformer-based models. Noting that our approach is model-agnostic, Table 7 shows the results of its implementation to transformer-based models. While limited computation resources led our model not to converge on the EgoVLP experiment, other modalities may affect our approach to the MMT experiment. Either way, we notice that the method's effect becomes limited on transformer-based models, and we share our related assertions in the Supplementary Material which could be related to the spatial biases.

Epic-Kitchens-100		YouCook2		MSR-VTT	
Method	**AVG**	**Method**	**AVG**	**Method**	**AVG**
EgoVLP [17]	12.53	TACo [38]	53.53	MMT [8]	63.79
EgoVLP + Ours	13.06	TACo + Ours	54.03	MMT + Ours	63.94

Table 7: Comparison with transformer-based models on text-video retrieval on nDCG metric.

5 Conclusion

To the best of our knowledge, this is the first attempt to study the effect of a temporal bias caused by a frame length mismatch between training and test sets of trimmed video clips and show improvement with debiasing on the text-video retrieval task. We then discuss detailed experiments and ablation studies using our causal approach on the Epic-Kitchens-100, YouCook2 and MSR-VTT datasets. Benchmark using the nDCG metric demonstrates that the bias has been reduced. We reckon the following limitations for future works: **i)** Long video clips may contain ambiguity, including various actions irrelevant to the annotated action. **ii)** Other temporal biases may still affect the model, e.g. the order of the actions.

Acknowledgements

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project A18A2b0046).

References

[1] Max Bain, Arsha Nagrani, Gul Varol, and Andrew Zisserman. Frozen in time: A joint video and image encoder for end-to-end retrieval. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, page 1708–1718, Montreal, QC, Canada, Oct 2021. IEEE. ISBN 978-1-66542-812-5. doi: 10.1109/ICCV48922.2021.00175. URL https://ieeexplore.ieee.org/document/9711165/.

[2] Sofia Broomé, Ernest Pokropek, Boyu Li, and Hedvig Kjellström. Recur, attend or convolve? on whether temporal modeling matters for cross-domain robustness in action recognition. In *2023 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 4188–4198, 2023. doi: 10.1109/WACV56688.2023.00418.

[3] Aman Chadha, Gurneet Arora, and Navpreet Kaloty. iPerceive: Applying common-sense reasoning to multi-modal dense video captioning and video question answering. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1–13, 2021.

[4] S. Chen, Y. Zhao, Q. Jin, and Q. Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *CVPR*, 2020.

[5] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. Fine-grained video-text retrieval with hierarchical graph reasoning. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10635–10644, 2020. doi: 10.1109/CVPR42600.2020.01065.

[6] Dima Damen, Hazel Doughty, Giovanni Maria Farinella, , Antonino Furnari, Jian Ma, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. Rescaling egocentric vision: Collection, pipeline and challenges for epic-kitchens-100. *International Journal of Computer Vision (IJCV)*, 130:33–55, 2022. URL https://doi.org/10.1007/s11263-021-01531-2.

[7] Alex Falcon, Giuseppe Serra, and Oswald Lanz. Learning video retrieval models with relevance-aware online mining. In *International Conference on Image Analysis and Processing*, pages 182–194. Springer, 2022.

[8] Valentin Gabeur, Chen Sun, Karteek Alahari, and Cordelia Schmid. Multi-modal Transformer for Video Retrieval. In *European Conference on Computer Vision (ECCV)*, 2020.

[9] Satya Krishna Gorti, Noel Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. X-pool: Cross-modal language-video attention for text-video retrieval. In *CVPR*. arXiv, Mar 2022. URL http://arxiv.org/abs/2203.15086. arXiv:2203.15086 [cs].

[10] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, Chen Zhao, Siddhant Bansal, Dhruv Batra, Vincent Cartillier, Sean Crane, Tien Do, Morrie Doulaty, Akshay Erapalli, Christoph Feichtenhofer, Adriano Fragomeni, Qichen Fu,

Abrham Gebreselasie, Cristina Gonzalez, James Hillis, Xuhua Huang, Yifei Huang, Wenqi Jia, Weslie Khoo, Jachym Kolar, Satwik Kottur, Anurag Kumar, Federico Landini, Chao Li, Yanghao Li, Zhenqiang Li, Karttikeya Mangalam, Raghava Modhugu, Jonathan Munro, Tullie Murrell, Takumi Nishiyasu, Will Price, Paola Ruiz Puentes, Merey Ramazanova, Leda Sari, Kiran Somasundaram, Audrey Southerland, Yusuke Sugano, Ruijie Tao, Minh Vo, Yuchen Wang, Xindi Wu, Takuma Yagi, Ziwei Zhao, Yunyi Zhu, Pablo Arbelaez, David Crandall, Dima Damen, Giovanni Maria Farinella, Christian Fuegen, Bernard Ghanem, Vamsi Krishna Ithapu, C. V. Jawahar, Hanbyul Joo, Kris Kitani, Haizhou Li, Richard Newcombe, Aude Oliva, Hyun Soo Park, James M. Rehg, Yoichi Sato, Jianbo Shi, Mike Zheng Shou, Antonio Torralba, Lorenzo Torresani, Mingfei Yan, and Jitendra Malik. Ego4d: Around the world in 3,000 hours of egocentric video. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, page 18973–18990, New Orleans, LA, USA, Jun 2022. IEEE. ISBN 978-1-66546-946-3. doi: 10.1109/CVPR52688.2022.01842. URL https://ieeexplore.ieee.org/document/9879279/.

[11] Kensho Hara, Yuchi Ishikawa, and Hirokatsu Kataoka. Rethinking training data for mitigating representation biases in action recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 3349–3353, June 2021.

[12] Kalervo Järvelin and Jaana Kekäläinen. Cumulated gain-based evaluation of ir techniques. *ACM Trans. Inf. Syst.*, 20(4):422–446, oct 2002. ISSN 1046-8188. doi: 10.1145/582415.582418. URL https://doi.org/10.1145/582415.582418.

[13] Wei Ji, Renjie Liang, Zhedong Zheng, Wenqiao Zhang, Shengyu Zhang, Juncheng Li, Mengze Li, and Tat-seng Chua. Are binary annotations sufficient? video moment retrieval via hierarchical uncertainty-based active learning. In *CVPR*, 2023.

[14] Evangelos Kazakos, Arsha Nagrani, Andrew Zisserman, and Dima Damen. Epic-fusion: Audio-visual temporal binding for egocentric action recognition. (arXiv:1908.08498), Aug 2019. URL http://arxiv.org/abs/1908.08498. arXiv:1908.08498 [cs].

[15] Jie Lei, Tamara L. Berg, and Mohit Bansal. Revealing single frame bias for video-and-language learning, 2022. URL https://arxiv.org/abs/2206.03428.

[16] Linjie Li, Jie Lei, Zhe Gan, Licheng Yu, Yen-Chun Chen, Rohit Pillai, Yu Cheng, Luowei Zhou, Xin Eric Wang, William Yang Wang, et al. Value: A multi-task benchmark for video-and-language understanding evaluation. In *35th Conference on Neural Information Processing Systems (NeurIPS 2021) Track on Datasets and Benchmarks*, 2021.

[17] Kevin Qinghong Lin, Alex Jinpeng Wang, Mattia Soldan, Michael Wray, Rui Yan, Eric Zhongcong Xu, Difei Gao, Rongcheng Tu, Wenzhe Zhao, Weijie Kong, Chengfei Cai, Hongfa Wang, Dima Damen, Bernard Ghanem, Wei Liu, and Mike Zheng Shou. Egocentric video-language pretraining. In *NeurIPS*. arXiv, Oct 2022. URL http://arxiv.org/abs/2206.01670. arXiv:2206.01670 [cs].

[18] Daizong Liu, Xiaoye Qu, and Wei Hu. Reducing the vision and language bias for temporal sentence grounding. In *Proceedings of the 30th ACM International Conference on Multimedia*, MM '22, page 4092–4101, New York, NY, USA, 2022. Association for

Computing Machinery. ISBN 9781450392037. doi: 10.1145/3503161.3547969. URL https://doi.org/10.1145/3503161.3547969.

[19] Yu Liu, Huai Chen, Lianghua Huang, Di Chen, Bin Wang, Pan Pan, and Lisheng Wang. Animating images to transfer clip for video-text retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 6, Madrid Spain, Jul 2022. ACM. ISBN 978-1-4503-8732-3. doi: 10.1145/3477495.3531776. URL https://dl.acm.org/doi/10.1145/3477495.3531776.

[20] A. Miech, I. Laptev, and J. Sivic. Learning a text-video embedding from incomplete and heterogeneous data. *arXiv:1804.02516*, 2018.

[21] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, and J. Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *ICCV*, 2019.

[22] A. Miech, J. B. Alayrac, L. Smaira, I. Laptev, J. Sivic, and A. Zisserman. End-to-end learning of visual representations from uncurated instructional videos. In *CVPR*, 2020.

[23] Guoshun Nan, Rui Qiao, Yao Xiao, Jun Liu, Sicong Leng, Hao Zhang, and Wei Lu. Interventional video grounding with dual contrastive learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2765–2775, June 2021.

[24] Judea Pearl. The do-calculus revisited, 2012. URL https://arxiv.org/abs/1210.4852.

[25] Will Price, Carl Vondrick, and Dima Damen. Unweavenet: Unweaving activity stories. (arXiv:2112.10194), Apr 2022. URL http://arxiv.org/abs/2112.10194. arXiv:2112.10194 [cs].

[26] Burak Satar, Zhu Hongyuan, Xavier Bresson, and Joo Hwee Lim. Semantic role aware correlation transformer for text to video retrieval. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1334–1338, 2021. doi: 10.1109/ICIP42928.2021.9506267.

[27] Burak Satar, Hongyuan Zhu, Hanwang Zhang, and Joo Hwee Lim. Exploiting semantic role contextualized video features for multi-instance text-video retrieval epic-kitchens-100 multi-instance retrieval challenge 2022, 2022.

[28] Burak Satar, Hongyuan Zhu, Hanwang Zhang, and Joo Hwee Lim. Rome: Role-aware mixture-of-expert transformer for text-to-video retrieval, 2022.

[29] Burak Satar, Hongyuan Zhu, Hanwang Zhang, and Joo Hwee Lim. An overview of challenges in egocentric text-video retrieval, 2023.

[30] P. Shi and J. Lin. Simple bert models for relation extraction and semantic role labeling, 2019.

[31] Xun Wang, Bingqing Ke, Xuanping Li, Fangyu Liu, Mingyu Zhang, Xiao Liang, and Qiushi Xiao. Modality-balanced embedding for video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 2578–2582, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450387323. doi: 10.1145/3477495.3531899. URL https://doi.org/10.1145/3477495.3531899.

[32] Michael Wray, Hazel Doughty, and Dima Damen. On semantic similarity in video retrieval. In *CVPR*, 2021.

[33] Shaoning Xiao, Long Chen, Songyang Zhang, Wei Ji, Jian Shao, Lu Ye, and Jun Xiao. Boundary proposal network for two-stage natural language video localization. In *AAAI*. arXiv, 2021. URL http://arxiv.org/abs/2103.08109. arXiv:2103.08109 [cs].

[34] Yao Xiao, Ziyi Tang, Pengxu Wei, Cong Liu, and Liang Lin. Masked images are counterfactual samples for robust fine-tuning. In *CVPR*, 2023.

[35] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), Jun 2016.

[36] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. Msr-vtt: A large video description dataset for bridging video and language. IEEE International Conference on Computer Vision and Pattern Recognition (CVPR), June 2016.

[37] Li Xu, Haoxuan Qu, Jason Kuen, Jiuxiang Gu, and Jun Liu. Meta spatio-temporal debiasing for video scene graph generation. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, page 374–390, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-19812-0.

[38] Jianwei Yang, Yonatan Bisk, and Jianfeng Gao. Taco: Token-aware cascade contrastive learning for video-text alignment. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11562–11572, October 2021.

[39] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 1–10. Association for Computing Machinery, 2021. doi: 10.1145/3404835.3462823. URL https://doi.org/10.1145/3404835.3462823.

[40] Sunjae Yoon, Ji Woo Hong, Eunseop Yoon, Dahyun Kim, Junyeong Kim, Hee Suk Yoon, and Chang D. Yoo. Selective query-guided debiasing for video corpus moment retrieval. In Shai Avidan, Gabriel Brostow, Moustapha Cissé, Giovanni Maria Farinella, and Tal Hassner, editors, *Computer Vision – ECCV 2022*, page 185–200, Cham, 2022. Springer Nature Switzerland. ISBN 978-3-031-20059-5.

[41] Ruohan Zhan, Changhua Pei, Qiang Su, Jianfeng Wen, Xueliang Wang, Guanyu Mu, Dong Zheng, and Peng Jiang. Deconfounding duration bias in watch-time prediction for video recommendation. In *SIGKDD*. arXiv, Jun 2022. URL http://arxiv.org/abs/2206.06003. arXiv:2206.06003 [cs].

[42] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Towards debiasing temporal sentence grounding in video, 2021. URL https://arxiv.org/abs/2111.04321.

[43] Huaiwen Zhang, Yang Yang, Fan Qi, Shengsheng Qian, and Changsheng Xu. Debiased video-text retrieval via soft positive sample calibration. *IEEE Transactions on Circuits and Systems for Video Technology*, page 1–1, 2023. ISSN 1051-8215, 1558-2205. doi: 10.1109/TCSVT.2023.3248873.

[44] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. Centerclip: Token clustering for efficient text-video retrieval. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, July 11–15, 2022, Madrid, Spain*, 2022.

[45] L. Zhou, C. Xu, and J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*, pages 7590–7598, 2018. URL https://www.aaai.org/ocs/index.php/AAAI/AAAI18/paper/view/17344.

[46] Luowei Zhou, Chenliang Xu, and Jason J. Corso. Towards automatic learning of procedures from web instructional videos. In *AAAI*. arXiv, Nov 2017. URL http://arxiv.org/abs/1703.09788. arXiv:1703.09788 [cs].

[47] Ligeng Zhu. THOP: PyTorch-OpCounter, 2022. URL https://github.com/Lyken17/pytorch-OpCounter.