TITAN: Test-tIme Training via counterfactuAl generation and casual interventioN for text-video retrieval

Mengze Li*
Zhejiang University
Zhejiang, China

Zihan Huang†
huanzh@zju.edu.cn
Zhejiang University
Zhejiang, China

Han Wang‡
Zhejiang University
Zhejiang, China

Jiaxu Miao
Zhejiang University
Zhejiang, China

Wenyan Fan
Zhejiang University
Zhejiang, China

Zhou Zhao
Zhejiang University
Zhejiang, China

Shengyu Zhang
Zhejiang University
Zhejiang, China

Wenming Tan
Zhejiang University
Zhejiang, China

Fei Wu
Zhejiang University
Zhejiang, China

ABSTRACT

Text-video retrieval aims to retrieve matching videos from the video pool based on the given language query. Despite significant progress, previous methods may not adapt to the constantly changing distribution in practical retrieval applications. Towards that, we introduce test-time training, which allows for efficient fine-tuning on the testing domain for model adjustment. A representative approach is to utilize the source domain model to produce pseudo-labels for the testing domain examples, leading to subsequent model refinement. However, when it comes to cross-modal retrieval, the generated pseudo-labels (i.e., the matched text-videos) present two critical issues: (A) *Alignment Imbalance*: in the text-video level, lots of videos from the testing domain do not match any queries and, as a result, do not contribute to the fine-tuning process; in the word-video level, certain words receive excessive attention, while others are unjustly neglected during alignment. (B) *Domain Biases*: the pseudo-labels generated suffer from confounding factors originating from the source domain, which differ from those inherent in the testing domain. This mismatch hampers effective test-time refinement. To address these two issues, we propose the **T**est-t**I**me **T**raining model via counterfactu**A**l generation and casual interventio**N** for text-video retrieval (**TITAN**), which includes two targeted modules: (1) The Counterfactual Hierarchical Rebalancing module re-matches text-video pairs by considering the matching priority bidirectionally between queries and videos, and corrects the word-video alignment by counterfactually suppressing the over-attended words in queries. (2) The Causal Bias Correction module employs the causal backdoor adjustment to effectively eliminate the distortions caused by the shifting of domain-specific confounders. Additionally, we propose a new cross-domain dataset as the testbed for model validation. Extensive experiments demonstrate the effectiveness of TITAN over the state-of-the-arts. The code is available [1] and our dataset will be released.

CCS CONCEPTS

• **Computing methodologies** → **Natural language processing**; **Computer vision**.

KEYWORDS

text-video retrieval, test-time training, multi-modal

ACM Reference Format:
Mengze Li, Zihan Huang, Han Wang, Jiaxu Miao, Wenyan Fan, Zhou Zhao, Shengyu Zhang, Wenming Tan, and Fei Wu. 2023. TITAN: Test-tIme Training via counterfactuAl generation and casual interventioN for text-video retrieval. In *Proceedings of Proceedings of the 31th ACM International Conference on Multimedia (MM '23).* ACM, New York, NY, USA, 10 pages. https://doi.org/none

1 INTRODUCTION

The text-video retrieval task aims to identify the target video from video pools according to a textual description specified by the user, which holds significant industrial application value [9, 27]. This field has attracted increasing attention from researchers, who have achieved promising results by utilizing neural networks and data-driven approaches. Yet, these achievements hinge on the assumption of consistent training and testing data distributions [9, 27]. In actual industrial applications, the data distributions for text-video retrieval tasks are subject to the whims of ever-evolving events and trends. For instance, when a war breaks out, there will be a significant increase in the retrieval of news videos; when a certain movie becomes a sensation, video retrievals related to that movie skyrocket. This incessant domain shift can drastically undermine the accuracy of models trained on previously defined source domains when tested on a constantly changing testing domain. To tackle this issue, we introduce test-time training into the text-video retrieval task, which allows the retrieval model that has been adequately trained on the source domain training set to efficiently and constantly adapt to the unannotated testing set through fine-tuning.

*Contributed equally to this research.
†Contributed equally to this research.
‡Contributed equally to this research.

MM '23, October 29–November 02, 2023, Ottawa, Canada

[1] https://anonymous.4open.science/r/AnonymousRepo-E9F1

Figure 1: A schematic diagram of the model trained and tested on two domains (real-life and action-movie) with different biases. During testing, if the visual information corresponding to the keyword "covered the face" in the query is ambiguous, the model may make incorrect predictions by relying on the bias learned from the real-life domain (which may not hold in the action-movie domain).

In light of this novel paradigm, a seemingly natural strategy is to generate pseudo-labels for testing samples, which in turn permit model fine-tuning [21, 26, 32]. However, when it comes to text-video retrieval, two primary issues emerge in relation to the pseudo-labels (matched test text-video pairs) generated:

- **Cross-modal Alignment Imbalance.** Our experimental findings reveal two pervasive issues of pseudo-label imbalance: (1) In the *text*-video level, an overwhelming number of queries are mismatched with a scant number of videos during the text-video retrieval process. Consequently, a considerable number of non-matched testing videos do not contribute to the fine-tuning process. (2) In the *word*-video level, a large number of queries matched with a few videos have certain words with extremely high frequency. In contrast, other query words, those forming crucial links with key visual elements in the videos, are overlooked. Training with imbalanced pseudo-labels can compromise the model's capability to match critical query words with corresponding visual elements in videos, thereby impacting the performance in cross-modal retrieval tasks.
- **Heterogeneous Domain Biases.** The nuanced differences in constantly changing testing data distribution introduce varying biases in cross-modal retrieval tasks. Figure 1 illustrates an example: in slice-of-life scenarios, a given query containing the keyword "covered the face" typically corresponds to a video of someone applying a facial mask before bedtime. Therefore, detecting signature visual elements (*e.g.*, pajama) may be the key to cross-modal matching. However, in some other scenes (*e.g.*, movies about robbers holding up a bank), the same keyword (*i.e.*, "covered the face") might typically correspond to a quite different visual element (*i.e.*, "gun" not pajamas). The crux of the problem lies in the fact that pseudo-labels, generated for testing data by a model trained intensively on the source domain, may carry source domain biases (*e.g.*, pajama is the critical visual element for query "covered the face"). These biases, while potentially beneficial in the source domain, may not be applicable in the ever-changing testing domain and can seriously undermine the model fine-tuning. In contrast, the introduction of *testing* domain bias could potentially benefit the model adjustment [45]. Thus, the challenge that we face is twofold: mitigating the influence of source domain bias while simultaneously capitalizing on the potential benefits of testing domain bias.

To address two issues, we propose the **T**est-t**I**me **T**raining model via counterfactu**AI** generation and casual interventi**ON** for text-video retrieval (**TITAN**). It consists of two targeted modules: **(1) Counterfactual Hierarchical Rebalancing.** At the text-video level, during generating pseudo-labels for the testing domain, we ensure that videos and texts form a mutually optimal match. This bidirectional matching strategy alleviates the issue of a large number of videos being left unmatched, a common pitfall of the conventional one-way text-to-video retrieval approach. At the word-video level, we identify and eliminate potentially over-attended words, resulting in the generation of counterfactual texts for fine-tuning. These counterfactual samples enhance the model's attention to other keywords that align with critical visual information. **(2) Causal Bias Correction.** Inspired by the backdoor adjustment theory in causal inference, this module is designed to correct biases in the pseudo-labels. We devise a debiasing mechanism that approximates the backdoor adjustment technique using neural networks, thus mitigating source domain biases by blocking the backdoor path. Additionally, we incorporate a testing domain bias enhancement mechanism into the module, leveraging mutual information to effectively capitalize on the potential advantages inherent in the backdoor path where *testing*-domain confounders are present.

To validate the effectiveness of the TITAN model, we conduct experiments on two widely used text-video retrieval benchmarks, each respectively corresponding to the training and testing domains. In addition, given the nascent state of exploration in test-time training for text-video retrieval, there exists a dearth of readily available datasets in this vein. To thoroughly evaluate the performance of our TITAN model, we propose a large-scale text-video retrieval dataset as another testbed for the evaluation of test-time training. This dataset,

carefully labeled by annotators, spans slice-of-life and movie domains, encapsulating 48,747 text-video examples. The extensive experiments on the two datasets demonstrated the robust performance of TITAN, reflected in notable enhancements in accuracy relative to the baseline models. Comprehensive analysis experiments, including ablation studies and qualitative analysis, provide deeper insights into TITAN's strengths.

Key contributions can be summarized as follows:

- We pioneer the exploration of test-time training in the context of text-video retrieval. To facilitate this investigation, we present a large-scale dataset tailored for this task.
- We propose the TITAN model, where the Counterfactual Hierarchical Rebalancing module employs cross-modal rematching and counterfactual generation to re-balance the pesudo-labels. The Causal Bias Correction module approximates the causal backdoor adjustment to rectify the training-distribution bias and leverage the testing-distribution bias.
- We provide extensive empirical validation of our model, demonstrating superior performance against state-of-the-art baselines. Through ablation studies and qualitative analysis, we further reveal the rationale behind TITAN.

2 RELATED WORK

Video-related Cross-modal Retrieval. In recent years, with the widespread application of videos, the demand for video-related retrieval has increased dramatically, and this field has attracted more and more attention from researchers [5, 24, 34, 36, 42, 43]. Text-video Retrieval [6, 9, 18, 38] is a widely researched and challenging branch of video-based retrieval tasks [10, 13, 37, 40, 44], which entail consolidating the input language query and retrieving the target video from the video set. Recently, some researchers attempt to extract more abundant information from training video data by mining latent video semantics to achieve enriched representations [7]. Several researchers also focus on enhancing cross-modal representations by integrating explicit high-level semantics into their models [35]. Other researchers have concentrated on identifying a fine-grained semantic correspondence between texts and videos or integrating coarse-grained and fine-grained representations to achieve multi-grained contrastive learning. [11] incorporates a pre-trained model and achieves excellent results using a very simple approach. To ensure stable performance when faced with shifting data distribution, we introduce the test-time training setting into the text-video retrieval task to fully exploit information in the testing domain and enhance the model's performance.

Test-time Training. Test-time Training (TTT) is a recently proposed paradigm that facilitates adaptation of a model trained on training domains to new test domains with different distribution, in an unsupervised manner and avoids re-visitation of the source data [4, 14, 14, 19, 29, 30]. It maximizes the utility of unlabeled data from the test domain to train the model, consequently exhibiting promising potential across diverse tasks for significantly enhancing model performance in domains with distribution shifts. The test-time training setting is introduced into several area of tasks recently, including single image dehazing across multi-domains, object detection, photorealistic style transfer, and image segmentation [15, 20, 23, 31]. Further yet, researchers combines TTT with 3D body reconstruction techniques, augmenting performance without adding or modifying modules [22]. [8] improves the performance of the test-time training method with the mask autoencoders and achieves excellent results. While experiment results demonstrate TTT's reliability in different tasks, few attempts have been made to apply TTT to text-video retrieval tasks, and further research is warranted. Our work combines TTT with text-video retrieval tasks, validating TTT's reliability in this task. To the best of our knowledge, our work is the early exploration to address the text-video retrieval task under the test-time training setting.

3 METHOD

In this section, we will provide a detailed description of our **T**est-t**I**me **T**raining model via counterfactu**A**l generation and casual interventio**N** for text-video retrieval (**TITAN**).

3.1 Overview

3.1.1 Task Formulation. **Text-video Retrieval.** The input to the system consists of a natural language query sequence denoted by $Q = \{Q^i\}_{i=1}^{N_Q}$ and a video set denoted by $\mathcal{V} = \{\mathcal{V}^i\}_{i=1}^{N_\mathcal{V}}$, where N_Q and $N_\mathcal{V}$ are the total numbers of the queries and the videos, respectively. The objective of the text-video retrieval task is to identify the corresponding video from the video set $\mathcal{V}$ for each language query in the query sequence Q. The optimization process ϵ for the retrieval model $\mathcal{M}$ is defined as:

$$\epsilon(\mathcal{V}, Q; \gamma) = \max_{\gamma} \xi(\mathcal{M}(\mathcal{V}, Q; \gamma), \delta(\mathcal{V}, Q)), \tag{1}$$

in which the function $\mathcal{M}(.)$ outputs the prediction of the video sequence $\mathcal{V}'$ corresponding one-to-one with the language query sequence Q and γ is the learnable parameter of the retrieval model $\mathcal{M}$. The function $\delta(.)$ represents the generation process of the ground truth. In addition, the function $\xi(.)$ optimizes the consistency between the two functions.

Test-time Training for Text-video Retrieval. The introduction of the test-time training setting helps improve the generalization of the model on the text-video retrieval task. In detail, given the source domain data $\mathcal{D}_s$, which includes the language query sequence Q_s and the video set $\mathcal{V}_s$, the model $\mathcal{M}$ is fully trained with them. Then, the test-time training setting requires the model $\mathcal{M}$ to learn the testing domain knowledge with the unpaired testing domain data $\mathcal{D}_t$, including Q_t and $\mathcal{V}_t$, thereby improving the generalization of the model $\mathcal{M}$ on the testing domain.

3.1.2 Model Pipeline. We adopt the strategy of finetuning the fully trained source retrieval model on the testing domain, to improve the model's generalization ability for the testing domain data $\mathcal{D}_t$. Specifically, the base retrieval model $\mathcal{M}$ contains a video encoder and a text encoder to generate multi-modal features, which is introduced in the appendix. After the model $\mathcal{M}$ fully trained with the source domain data $\mathcal{D}_s$, there are three steps to finetune it with the testing domain data $\mathcal{D}_t$, as shown in Figure 2:

(1) We generate the pseudo-labels for the testing domain data $\mathcal{D}_t$ using the retrieval modal $\mathcal{M}$.
(2) The Counterfactual Hierarchical Rebalancing module addresses the issue of imbalanced cross-modal matching at both the text-video level and the word-video level for the pseudo-labels. The

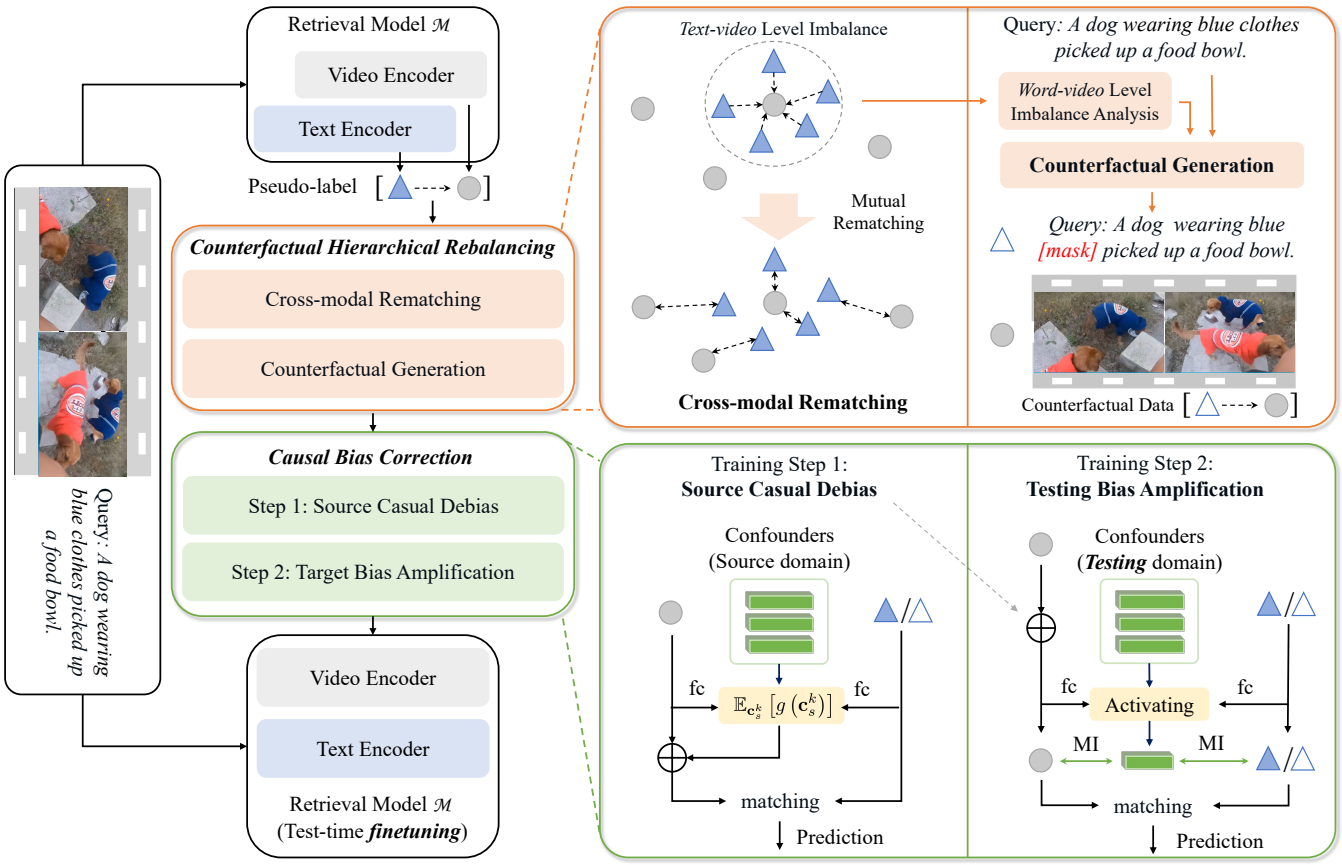

Figure 2: The finetuning process of the retrieval model $\mathcal{M}$ on the testing domain data with the pseudo-labels. Several modules corrects the pseudo-labels. In detail, (1) the Cross-modal Rematching alleviates lots of queries incorrectly matched to limited videos by re-matching. (2) The Counterfactual Generation outputs additional counterfactual training examples by masking over-attended words. (3) The Casual Bias Correction corrects the bias in the pseudo-labels, whose training process contains two steps: the Source Causal Debias and the Testing Bias Amplification.

Casual Bias Correction module alleviates the source domain bias in the pseudo labels and incorporates the testing domain bias to further improve accuracy.

(3) We finetune the retrieval model $\mathcal{M}$ with the testing domain data $\mathcal{D}_t$ and the pseudo-labels generated by step (1) and (2) to obtain the target model.

3.2 Counterfactual Hierarchical Rebalancing

In our observations, we identify two levels of imbalances within the pseudo-labels produced for the testing domain data $\mathcal{D}_t$ using the fully trained source model $\mathcal{M}$. These imbalances are present both at the text-video level and at the more granular word-video level. To combat these imbalances and enhance the performance of our model, we strategically introduce two specialized components to this module.

3.2.1 Cross-modal Rematching. For the text-video level, we observe that most language queries may wrongly be matched with a small number of videos and lots of videos are not matched, which makes insufficient videos involved in the model fine-tuning. To address the problem, we consider whether a given text and video are mutually the best match for each other, rather than just one-way text-to-video retrieval, during the process of generating pseudo-labels for the testing domain.

Specifically, given the i-th language query Q_t^i and the i-th video $\mathcal{V}_t^j$ belonging to the testing domain data $\mathcal{D}_t$, the retrieval model $\mathcal{M}$ generates the query feature $\mathbf{q}_t^i$ and the video feature $\mathbf{v}_t^j$ with the multi-modal encoders. Then, we calculate the similarity score s_i^j with the cosine similarity:

$$s_i^j = \frac{(\mathbf{q}_t^i)^T \cdot \mathbf{v}_t^j}{\|\mathbf{q}_t^i\|_1 * \|\mathbf{v}_t^j\|_1}. \tag{2}$$

For the i-th query Q_t^i, we calculate its similarity scores $\mathcal{S}^i = \{\mathcal{S}_j^i\}_{j=1}^{N_{\mathcal{V}_t}}$ with all the videos in the testing domain video set $\mathcal{V}_t$, where $N_{\mathcal{V}_t}$ is the video number in $\mathcal{V}_t$. We obtain the ranking of each score in the video score set $\mathcal{S}^i$ with the $ranking(.)$ function. The

score rankings of all videos are represented as $\mathcal{R}^i_{\mathcal{V}_t} = \{\mathcal{R}^i_{\mathcal{V}_t^j}\}_{j=1}^{N_{\mathcal{V}_t}}$, whose calculating process is formulated as:

$$\mathcal{R}^i_{\mathcal{V}_t} = ranking(\mathcal{S}^i). \tag{3}$$

Similarly, for the j-th video $\mathcal{V}_t^j$, its similarity scores $\mathcal{S}_j = \{\mathcal{S}_j^i\}_{i=1}^{N_{Q_t}}$ with all the queries in the query sequence Q_t. Then, we obtain the ranking of each score in the similarity score set $\mathcal{S}_j$ with the $ranking(.)$ function and represent these rankings as $\mathcal{R}^j_{Q_t} = \{\mathcal{R}^j_{Q_t^i}\}_{i=1}^{N_{Q_t}}$. This calculation process is formulated as:

$$\mathcal{R}^j_{Q_t} = ranking(\mathcal{S}_j). \tag{4}$$

When predicting the matching degree $\mathcal{M}_j^i$ between the i-th query Q_t^i and the j-th video $\mathcal{V}_t^j$, we take into account both two types of ranking obtained from the above calculations:

$$\mathcal{M}_j^i = \mathcal{R}^i_{\mathcal{V}_t^j} + \alpha * \mathcal{R}^j_{Q_t^i}, \tag{5}$$

where α is a hyper-parameter. We consider the k-th video $\mathcal{V}_t^k$ with the highest match degree $\mathcal{M}_k^i$ to the i-th query Q_t^i as the predicted retrieval result of the Cross-modal Rematching module.

3.2.2 Counterfactual Generation. At the word-video level, we observe a significant disparity in word frequency within the query-level matches. Certain words, predominantly found within a limited number of text-video matches, exhibit an extremely high frequency, while other query words, which are aligned with key visual elements within the videos, are significantly underrepresented. This imbalance, if left unchecked during the fine-tuning process, could potentially undermine the model's capacity to accurately match a broad spectrum of keywords with corresponding visual elements in the videos. To address this issue, we construct a set of counterfactual queries, paying more attention to low-frequency while suppressing over-attended high-frequency words. By fine-tuning the retrieval model $\mathcal{M}$ with these adjusted queries, we aim to rectify the imbalance and enhance the model's performance.

Specifically, before the Cross-modal Rematching, we choose all the videos $\mathcal{V}_c = \{\mathcal{V}_c^j\}_{j=1}^{N_{\mathcal{V}_c}}$ matched with more than β (a hyperparameter) queries. For any selected video $\mathcal{V}_c^j$, we count the frequency $\mathcal{F}_{\mathcal{V}_c^j} = \{\mathcal{F}^i_{\mathcal{V}_c^j}\}_{i=1}^{N_{\mathcal{F}}}$ of nouns, verbs, and adjectives in the matched queries $Q_{\mathcal{V}_c^j}$ with the statistical function $statistic(.)$, where $N_{\mathcal{F}}$ is the total number of words counted. This statistical process is represented as:

$$\mathcal{F}_{\mathcal{V}_c^j} = statistic(Q_{\mathcal{V}_c^j}). \tag{6}$$

For any word w^i, if its frequency $\mathcal{F}^i_{\mathcal{V}_c^j}$ is larger than the hyperparameter θ, we select all the queries Q_{w^i} containing the word w^i from the matched queries $Q_{\mathcal{V}_c^j}$. We mask the word w^i in the selected queries Q_{w^i} to generate counterfactual queries Q'_{w^i} using the $masking(.)$ function:

$$Q'_{w^i} = masking(Q_{w^i}). \tag{7}$$

We use our designed Cross-modal Rematching module to match these counterfactual queries Q'_{w^i} with the corresponding videos and add the matching pairs to the pseudo-labels.

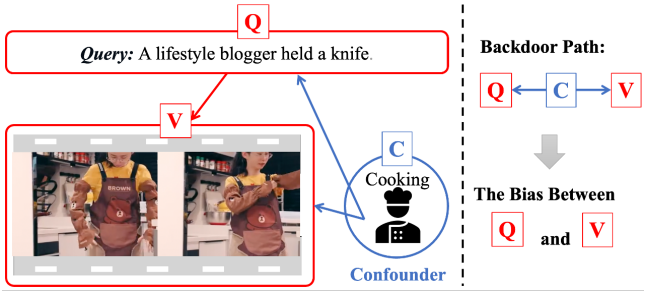

Figure 3: An intuitive illustration of the causal graph with an example. The arrow indicates the causal influence.

3.3 Casual Bias Correction

Source and testing domain biases are often at odds. When pseudo-labels for the testing domain carry the source domain bias, the accuracy of pseudo-labels is adversely affected. Conversely, incorporating testing domain bias can actually enhance the accuracy of the pseudo-labels. To address this issue, we introduce the Causal Bias Correction module, inspired by the principles of causal backdoor adjustment. This module is designed to diminish the impact of source domain bias in pseudo-labels, while simultaneously amplifying the influence of testing domain bias. The comprehensive analysis performed by the Causal Bias Correction module, which is based on query-video pairings, might be potentially computationally intensive, as it would require a search across the entirety of the video database for each query. To maintain computational efficiency, we only apply the bias correction to the top η most likely videos as identified by the retrieval model $\mathcal{M}$.

Before introducing the module, we first explain the causal graph involved in the text-video retrieval task about four casual variables: Query Q, Video V, and Confounder C. As shown in Figure 3, any arrow represents a cause-and-effect relationship between two nodes. This casual graph specifies the relationships between variables. Here, $Q \rightarrow V$ represents the process of retrieving the video V according to the query Q. In addition, the confounder C affects Q and V, which leads to a backdoor path $Q \leftarrow C \rightarrow V$ and the bias between Q and V. Based on the example in Figure 3, we provide a detailed explanation: **(1) Backdoor Path ($Q \leftarrow C \rightarrow V$).** Based on common sense, in daily cooking, people usually put on an apron and pick up a knife to cut vegetables, which leads to casual relationships $Q \leftarrow C \rightarrow V$. **(2) The Bias Between Q and V.** However, the node C can cause the occurrence of Q and V. It often leads to the synchronous occurrence of Q and V, resulting in the bias between them. The bias caused by the backdoor path $Q \leftarrow C \rightarrow V$ can be alleviated by casual backdoor adjustment [1].

3.3.1 Source Casual Debias. We need to train the Casual Bias Correction module on the source domain data $\mathcal{D}_s$ while ensuring not being affected by the source domain bias. Inspired by backdoor adjustment, we propose a targeted module design.

Specifically, given the i-th language query Q_s^i and the i-th video $\mathcal{V}_s^j$ belonging to the source domain data $\mathcal{D}_s$, the retrieval model $\mathcal{M}$ generates the query feature $\mathbf{q}_s^i$ and the video feature $\mathbf{v}_s^j$ with the multi-modal encoders. We extract the action features and the object

features from videos of the source domain data $\mathcal{D}_s$, to initialize the source domain confounder features $\mathbf{c}_s = \{\mathbf{c}_s^k\}_{k=1}^{N_{C_s}}$ with the pretrained detectors [25, 28, 39]. In it, N_{C_s} is the confounder number. We introduce the do-calculus to realize the backdoor adjustment for the casual graph in Figure 3. After derivation, we implement the backdoor adjustment with the extracted cross-modal features as:

$$P(V|do(Q)) := \mathbb{E}_{\mathbf{c}_s^k}[\sigma(f(\mathbf{q}_s^i, \mathbf{v}_s^j, \mathbf{c}_s^k))] \approx \sigma(\mathbb{E}_{\mathbf{c}_s^k}[f(\mathbf{q}_s^i, \mathbf{v}_s^j, \mathbf{c}_s^k)]). \tag{8}$$

The detailed derivation process is presented in the appendix. In it, $\sigma(.)$ is the sigmoid activation function, and $f(.)$ is a cross-modal matching network. $\mathbb{E}_{\mathbf{c}_s^k}[.]$ represents the expectation calculation. We implement the matching network $f(.)$ as:

$$f(\mathbf{q}_s^i, \mathbf{v}_s^j, \mathbf{c}_s^k) = MLP((\mathbf{v}_s^j + g(\mathbf{c}_s^k)) \odot \mathbf{q}_s^i), \tag{9}$$

where MLP is the multi-layer perception and $g(\mathbf{c}_s^k)$ is a feature transformation of the confounder feature $\mathbf{c}_s^k$. $\odot$ is the element-wise product. Then, we need to substitute Equation 9 into Equation 8. Based on the approximation $\mathbb{E}[Relu(x)] \approx Relu(\mathbb{E}[x])$ proposed by a theoretical analysis research [3], we can derive the following:

$$\sigma(\mathbb{E}_{\mathbf{c}_s^k}[f(\mathbf{q}_s^i, \mathbf{v}_s^j, \mathbf{c}_s^k)]) \approx \sigma(MLP((\mathbf{v}_s^j + \mathbb{E}_{\mathbf{c}_s^k}[g(\mathbf{c}_s^k)]) \cdot \mathbf{q}_s^i)). \tag{10}$$

Furthermore, we implement $\mathbb{E}_{\mathbf{c}_s^k}[g(\mathbf{c}_s^k)]$ as the scaled product attention [33] to dynamically allocate weights to different confounders in source domain confounder set $\mathbf{c}_s$ based on the cross-modal features $\mathbf{q}_s^i$ and $\mathbf{v}_s^j$:

$$\mathbb{E}_{\mathbf{c}_s^k}[g(\mathbf{c}_s^k)] = softmax((w_1\mathbf{q}_s^i + w_2\mathbf{v}_s^j) \cdot \mathbf{c}_s^T) \cdot (w_3\mathbf{c}_s), \tag{11}$$

where w_1, w_2, and w_3 are learnable parameters. The matching score between the i-th query and the j-th video is predicted as $\mathcal{S}_s = \sigma(\mathbb{E}_{\mathbf{c}_s^k}[f(\mathbf{q}_s^i, \mathbf{v}_s^j, \mathbf{c}_s^k)])$. We use the binary cross entropy loss to train this module on the source domain data $\mathcal{D}_s$.

3.3.2 Testing Bias Amplification. Due to interfering factors such as lighting and occlusion, the key video information corresponding to the language query may not be clear enough. In this case, inferring based on the testing domain bias can improve the accuracy of retrieval. Inspired by this idea, we train the Casual Bias Correction module on the testing domain data $\mathcal{D}_t$ and learn the testing domain bias to enhance the cross-modal retrieval ability. Analyzing from the perspective of causal graphs (shown in Figure 3), strengthening the backdoor path $V \leftarrow C \rightarrow Y$ can enhance the bias between the video casual variable V and the matching result variable Y. To achieve this goal, we enhance the correlation between V and C, and between C and Y by increasing their mutual information.

Following this idea, we modify our Casual Bias Correction module. Specifically, similar to the Source Casual Debias stage, we extract the action features and the object features from videos of the testing domain data $\mathcal{D}_t$ with the pretrained detectors [25, 28, 39], to initialize the testing domain confounder features $\mathbf{c}_t = \{\mathbf{c}_t^k\}_{k=1}^{N_{C_t}}$. Then, given the i-th language query Q_t^i and the j-th video $\mathcal{V}_t^j$ belonging to the testing domain data $\mathcal{D}_t$, the retrieval model $\mathcal{M}$ generates the query feature $\mathbf{q}_t^i$ and the video feature $(\mathbf{v}')_t^j$ with its cross-modal encoders. After alleviating the impact of the source domain bias, we get the updated video feature $\mathbf{v}_t^j$ by adding $\mathbb{E}_{\mathbf{c}_s^k}[g(\mathbf{c}_s^k)]$. Then, the cross-modal matching score $\mathcal{S}_t$ is calculated:

$$\mathcal{S}_t = MLP(\mathbf{v}_t^j \odot \mathbf{q}_t^i), \tag{12}$$

where MLP is the multi-layer perception and $\odot$ is the element-wise product. By integrating the analysis of cross-modal features ($\mathbf{q}_t^i$ and $\mathbf{v}_t^j$) and all the confounding features $\mathbf{c}_t$, we generate the corresponding confounder feature $\mathbf{c}_t^c$ for the current text-video pair:

$$\mathbf{c}_t^c = \sigma(\mathcal{K} \cdot \mathbf{c}_t^T) \cdot \mathbf{c}_t, \tag{13}$$

where σ is the sigmoid function and $\mathcal{K} = w_4\mathbf{v}_t^j + w_5\mathbf{q}_t^i$. w_4 and w_5 are learnable parameters.

Assuming that the pseudo-labels of the cross-modal matching score is y, the training loss l is calculated as:

$$l = -(ylog(\mathcal{S}_t) + (1-y)log(1-\mathcal{S}_t)) - I(\mathbf{v}_t^j; \mathbf{c}_t^c) - I(\mathbf{q}_t^i; \mathbf{c}_t^c), \tag{14}$$

where $I(.)$ represents the mutual information calculation. These mutual information terms can be estimated by the existing mutual information estimator [2].

4 EXPERIMENT

To thoroughly evaluate our TITAN model, we carry out comprehensive experiments across two distinct datasets.

4.1 Experiment Preparation

4.1.1 Datasets. Two pairs of domains are developed by us as the testbed for our TITAN model, including Video2Gif (normal web video-text pairs to gif-text pairs) and Life2Movie (real family-life scenes to movie scenes).

Specifically, for the Video2Gif dataset, it consists of the widely used MSR-VTT dataset [41] and the TGIF dataset [17], in which the MSR-VTT dataset is collected from the video website (e.g., Youtube) and the TGIF dataset is a large-scale retrieval dataset containing 120,000 GIF-query pairs. We follow the official dataset split on both of them. In addition, following the previous method [21], all models are trained on the training set of the normal web video-text domain (source domain). Then, they are finetuned and tested on the testing set of the gif-text domain (testing domain).

Due to the insufficient existing cross-domain retrieval datasets available for test-time training, we propose another dataset, Life2Movie, to facilitate a comprehensive validation of our TITAN model. In detail, the videos are collected from the real-life scenes and the movie scenes, which are carefully labeled by annotators. There are 38,800 examples in the real-life domain, and 9,947 exmples in the movie domain. Two-pass labeling methods are used for the annotation process. In addition to careful annotation by annotators, three inspectors participate in the review of each video-query example. If they all agree on the annotation, the example is retained. Otherwise, the example is re-annotated or discarded. More dataset details can be found in the appendix.

4.1.2 Evaluation Metrics. We follow a widely-adopted evaluation protocol in the experiments [12]. Specifically, the standard retrieval metrics, including Rank N (R@N, N=1,5,10) and median rank (MdR), are adopted. The Rank N measures the percentage of the test queries that have the target video item among the top-N retrieved results.

Table 1: Performance comparison with baselines on the Video2Gif and Life2Movie datasets. Overall 1^{st} and 2^{nd} best in red/blue.

Method	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
Video2Gif Dataset				
CMFG [46]	11.6	25.4	34.1	30
CRET [12]	17.4	34.6	43.4	17
CRET+TTT[21]	17.3	**35.2**	43.9	17
CRET+TTTFlow[26]	**17.9**	35.0	**44.0**	**16**
TITAN (Ours)	**20.4**	**38.8**	**47.8**	**12**
Life2Movie Dataset				
CMFG [46]	5.3	16.3	23.7	34
CRET [12]	6.0	18.6	28.1	32
CRET+TTT [21]	**6.3**	18.9	28.2	32
CRET+TTTFlow [26]	6.1	**19.0**	**28.7**	**31**
TITAN (Ours)	**7.4**	**22.1**	**32.8**	**26**

4.1.3 Implementation Detail. For model training, the batch size is set as 128 and the learning rate is $1e-5$. We implement our TITAN model with Pytorch 1.9 and train it on a Linux server with 8 3090 GPUs. About the hyperparameters of the Counterfactual Hierarchical Rebalancing module, α is set as 1 and β is defined as 30. About the hyperparameters of the Casual Bias Correction module, η is 10. In addition, our TITAN model adopts the online finetuning and prediction with testing data sequentially streamed and predicted, following the testing protocol of the previous method [16].

4.1.4 Baselines. Existing text-video retrieval methods are not directly designed for the test-time training setting. In order to comprehensively compare the performance of our model, we use several state-of-the-art models as baselines and extend these models to the test-time training setting: (1) text-video retrieval methods: **CRET** [12], **CMFG** [46]. (2) test-time training methods: **TTTFlow** [26] and **TTT** [21].

4.2 Performance Comparison

We compare our TITAN model and the baselines. The experiment results are shown in Table 1. Regarding state-of-the-art retrieval models, such as **CRET** and **CMFG**, they are initially trained on the source domain dataset. Subsequently, they undergo fine-tuning on the testing domain using model-generated pseudo-labels, with the online adaptation the same as our TITAN model. In addition, two advanced test-time training methods are added to the state-of-the-art retrieval method to compare with our TITAN model. Based on the tables, it can be observed that:

- The current state-of-the-art baselines for the text-video retrieval task, such as **CRET** and **CMFG**, do not achieve satisfactory performance. This is because these models are not specifically designed for test-time training and lack domain-specific knowledge. Incorporating targeted designs, such as **TTTFlow** and **TTT**, into these retrieval models improve their robustness to domain gaps and leads to better performance on the testing domain.
- The TITAN model achieves the best performance on both sets of domain pairs. Our TITAN achieves a significant performance boost for the "MdR" indicator decreasing it from 16/31 to 12/26 on two datasets, respectively, when compared to the baselines. This reflects the rationality of the TITAN design. Furthermore, the improvement is credited to the reasonably designed modules in the model. Specifically, the Counterfactual Hierarchical Rebalancing module effectively alleviates the imbalances coming from the text-video level and the word-video level. The Casual Bias Correction module effectively mitigates the source domain bias in pseudo labels and enhances the testing domain bias.

Table 2: Ablation study on the Life2Movie dataset. Overall 1^{st} and 2^{nd} best in red/blue.

#	CHR	CBC	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
1			5.7	20.4	28.5	31
2	✓		6.2	**21.3**	**31.7**	27
3		✓	**6.3**	21.2	30.1	**27**
4	✓	✓	**7.4**	**22.1**	**32.8**	**26**

4.3 Ablation Study

We are interested in analyzing the contribution of each module in the TITAN model. To analyze each module, we remove the key modules from the TITAN model and construct models with varying structures. The modules under investigation include the **C**ounterfactual **H**ierarchical **R**ebalancing and the **C**asual **B**ias **C**orrection. We represent them as CHR and CBC during the ablation study process, respectively. We evaluate the constructed ablation study models on the Life2Movie dataset. The experimental results are shown in Table 2. From the table, several observations can be made:

- Combining the key modules leads to better performance than using any single component alone, demonstrating that our proposed modules mutually benefit each other for the text-video retrieval task under the test-time training setting.
- The addition of any one of the two key modules results in improved performance. It highlights the necessity and effectiveness of these modules, including the Counterfactual Hierarchical Rebalancing module and the Casual Bias Correction module.

The ablation study results on the Video2Gif dataset are shown in the appendix.

4.4 Detailed Investigation

4.4.1 Hyperparameter Analysis of β. We investigate how varying the hyperparameter β of the Counterfactual Generation part affects the performance of the target model. The hyperparameter β

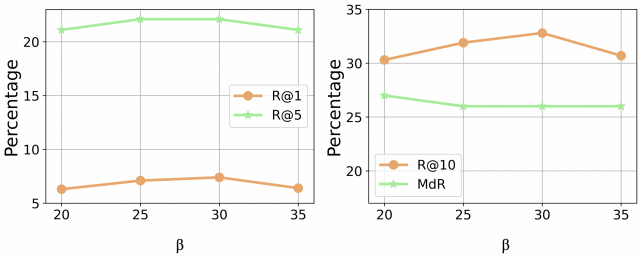

Figure 4: Effect of the different hyperparameter β of the Counterfactual Generation part on the Life2Movie dataset.

Table 3: Performance comparison with baselines on the Video2Gif dataset with the offline adaptation method.

Method	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓
CMFG [46]	12.3	25.6	35.8	28
CRET [12]	17.5	35.0	44.1	17
CRET+TTT[21]	17.6	35.2	44.5	16
CRET+TTTFlow[26]	**18.0**	**35.4**	**44.6**	**16**
TITAN (Ours)	**20.7**	**39.3**	**48.7**	**11**

controls which testing domain example the Counterfactual Generation module references to generate counterfactual data. The experiment results on the Life2Movie dataset are illustrated in Figure 4, which indicates the following discoveries:

- The accuracy changes of most metrics are within 1 as the hyperparameter β varies. This suggests that our model is relatively robust and not highly sensitive to the parameter.
- During the initial stage ($\beta \leq 30$), the model's accuracy gradually improves as the value of the hyperparameter β increases (except for the MdR metric, where a lower value indicates better performance). This indicates that as the number of matched language queries for a video increases, the model pays less attention to some keywords in these queries. Therefore, using counterfactual training data to augment such queries leads to a more significant improvement in model performance.
- When β is greater than 30, the number of generated counterfactual queries is not enough. This reduces the impact of counterfactual queries on improving the model's accuracy.

4.4.2 Offline Adaptation. We are interested in the performance of our model, when changing the adaptation from the online to the offline. Thus, under the offline-adaptation setting, we compare our TITAN model with the state-of-the-arts on the Video2Gif dataset. The experiment results are shown in Table 3. From the table, it can be found that our TITAN model performs better than all the baselines. This demonstrates the wide applicability of our TITAN model to different adaptation settings.

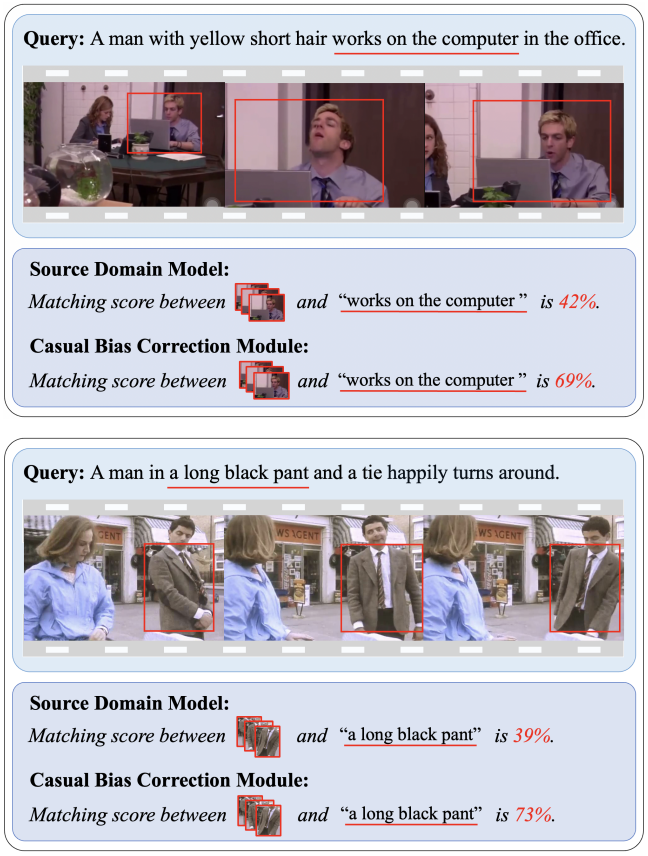

Figure 5: Two examples of the Casual Bias Correction module correcting the Source Domain model prediction.

4.5 Case Study

To underscore the effectiveness of our Causal Bias Correction module, we conduct a targeted case study. Specifically, we randomly select two examples from the testing domain. For each example, we manually extract visual and textual semantic units from the given video and the associated language query, respectively. Leveraging both the source domain model and our Causal Bias Correction module, we compute the correlation between these cross-modal semantic units. By scrutinizing the correction scores, we are able to discern the influence of either source domain bias or testing domain bias. Our experimental results, illustrated in Figure 5, yield the following observations:

- The first example reveals that the source domain model struggles with correctly aligning the given cross-modal information. This could be attributed to the fact that the source domain data, which was primarily collected from real family scenes, has engendered a certain bias. For instance, in these scenes, individuals working on computers are often seen wearing T-shirts. Consequently, the source domain model has learned to inaccurately associate the act of working on a computer with the absence of formal attire, such as a shirt and a tie. Our Causal Bias Correction module effectively diminishes

the impact of this source domain bias, resulting in an accurate alignment of cross-modal information.

- In the second example, only a portion of the actor's trousers is visible. Yet, based on the actor's consistent dressing habits, an inference can be made that when he wears a gray suit and tie for his upper body, it's likely accompanied by a pair of long black trousers for the lower body. The source domain model falls short in making this inference. However, the Causal Bias Correction module, having learned the nuances of the testing domain bias, can correctly deduce the correlation between the cross-modal information in this scenario.

For a more comprehensive comparison between the source domain model and our Causal Bias Correction module, we have included additional examples in the appendix. Furthermore, to gain a deeper understanding of our model's capabilities compared to existing baselines, we have also appended a variety of illustrative examples in the same section.

5 CONCLUSION

In this work, we set out to enhance model generalization across diverse test-time data distributions by introducing the test-time training paradigm to the text-video retrieval task. To this end, we proposed the TITAN model—specifically designed for this novel task setting—featuring two distinctive modules: the Counterfactual Hierarchical Rebalancing module and the Causal Bias Correction module. To further support our investigation, we introduced a large-scale cross-domain dataset for text-video retrieval, serving as a rigorous testbed for our model. Our thorough evaluations across two datasets provide compelling evidence of TITAN's superior performance, as it consistently outperforms baseline models by significant margins. This work paves the way for future research on leveraging test-time training for enhanced model generalization in the challenging field of text-video retrieval.

REFERENCES

[1] Riddhiman Adib, Paul Griffin, Sheikh Iqbal Ahamed, and Mohammad Adibuzzaman. 2020. A causally formulated hazard ratio estimation through backdoor adjustment on structural causal model. In *Machine Learning for Healthcare Conference*. PMLR, 376–396.

[2] Sungsoo Ahn, Shell Xu Hu, Andreas Damianou, Neil D Lawrence, and Zhenwen Dai. 2019. Variational information distillation for knowledge transfer. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 9163–9171.

[3] Pierre Baldi and Peter Sadowski. 2014. The dropout learning algorithm. *Artificial intelligence* 210 (2014), 78–122.

[4] Mohammad Zalbagi Darestani, Jiayu Liu, and Reinhard Heckel. 2022. Test-Time Training Can Close the Natural Distribution Shift Performance Gap in Deep Learning Based Compressed Sensing. In *International Conference on Machine Learning, ICML 2022, 17-23 July 2022, Baltimore, Maryland, USA (Proceedings of Machine Learning Research, Vol. 162)*. PMLR, 4754–4776.

[5] Jianfeng Dong, Xianke Chen, Minsong Zhang, Xun Yang, Shujie Chen, Xirong Li, and Xun Wang. 2022. Partially Relevant Video Retrieval. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 246–257.

[6] Alex Falcon, Giuseppe Serra, and Oswald Lanz. 2022. A Feature-space Multimodal Data Augmentation Technique for Text-video Retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4385–4394.

[7] Sheng Fang, Shuhui Wang, Junbao Zhuo, Qingming Huang, Bin Ma, Xiaoming Wei, and Xiaolin Wei. 2022. Concept propagation via attentional knowledge graph reasoning for video-text retrieval. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4789–4800.

[8] Yossi Gandelsman, Yu Sun, Xinlei Chen, and Alexei Efros. 2022. Test-time training with masked autoencoders. *Advances in Neural Information Processing Systems* 35 (2022), 29374–29385.

[9] Ning Han, Jingjing Chen, Guangyi Xiao, Hao Zhang, Yawen Zeng, and Hao Chen. 2021. Fine-grained Cross-modal Alignment Network for Text-Video Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*. ACM, 3826–3834.

[10] Zhijian Hou, Chong-Wah Ngo, and Wing Kwong Chan. 2021. CONQUER: Contextual Query-aware Ranking for Video Corpus Moment Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 3900–3908. https://doi.org/10.1145/3474085.3475281

[11] Fan Hu, Aozhu Chen, Ziyue Wang, Fangming Zhou, Jianfeng Dong, and Xirong Li. 2022. Lightweight Attentional Feature Fusion: A New Baseline for Text-to-Video Retrieval. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part XIV*. Springer, 444–461.

[12] Kaixiang Ji, Jiajia Liu, Weixiang Hong, Liheng Zhong, Jian Wang, Jingdong Chen, and Wei Chu. 2022. Cret: Cross-modal retrieval transformer for efficient text-video retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 949–959.

[13] Chen Jiang, Kaiming Huang, Sifeng He, Xudong Yang, Wei Zhang, Xiaobo Zhang, Yuan Cheng, Lei Yang, Qing Wang, Furong Xu, Tan Pan, and Wei Chu. 2021. Learning Segment Similarity and Alignment in Large-Scale Content Based Video Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 1618–1626. https://doi.org/10.1145/3474085.3475301

[14] Sunwoo Kim, Soohyun Kim, and Seungryong Kim. 2022. Deep Translation Prior: Test-Time Training for Photorealistic Style Transfer. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*. AAAI Press, 1183–1191.

[15] Sunwoo Kim, Soohyun Kim, and Seungryong Kim. 2022. Deep translation prior: Test-time training for photorealistic style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 1183–1191.

[16] Xiang Li, Kai Zhang, Bolun Li, Qiang Zhao, and Yuchao Dai. 2021. Test-Time Training for Single Image Dehazing and Beyond. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 15264–15274.

[17] Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. TGIF: A new dataset and benchmark on animated GIF description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4641–4650.

[18] Hongying Liu, Ruyi Luo, Fanhua Shang, Mantang Niu, and Yuanyuan Liu. 2021. Progressive Semantic Matching for Video-Text Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*. ACM, 5083–5091.

[19] Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. 2022. Towards Multi-domain Single Image Dehazing via Test-time Training. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 5821–5830.

[20] Huan Liu, Zijun Wu, Liangyan Li, Sadaf Salehkalaibar, Jun Chen, and Keyan Wang. 2022. Towards multi-domain single image dehazing via test-time training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 5831–5840.

[21] Yuejiang Liu, Parth Kothari, Bastien Van Delft, Baptiste Bellot-Gurlet, Taylor Mordan, and Alexandre Alahi. 2021. TTT++: When does self-supervised test-time training fail or thrive? *Advances in Neural Information Processing Systems* 34 (2021), 21808–21820.

[22] Jonathan Samuel Lumentut and In Kyu Park. 2022. 3D Body Reconstruction Revisited: Exploring the Test-time 3D Body Mesh Refinement Strategy via Surrogate Adaptation. In *Proceedings of the 30th ACM International Conference on Multimedia*. 5923–5933.

[23] Fei Lyu, Mang Ye, Andy J Ma, Terry Cheuk-Fung Yip, Grace Lai-Hung Wong, and Pong C Yuen. 2022. Learning from synthetic ct images via test-time training for liver tumor segmentation. *IEEE transactions on medical imaging* 41, 9 (2022), 2510–2520.

[24] Zhixin Ma and Chong-Wah Ngo. 2022. Interactive Video Corpus Moment Retrieval using Reinforcement Learning. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 296–306.

[25] Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. 2022. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 19023–19034.

[26] David Osowiechi, Gustavo A. Vargas Hakim, Mehrdad Noori, Milad Cheraghalikhani, Ismail Ben Ayed, and Christian Desrosiers. 2023. TTTFlow: Unsupervised Test-Time Training With Normalizing Flow. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*. 2126–2134.

[27] Yookoon Park, Mahmoud Azab, Seungwhan Moon, Bo Xiong, Florian Metze, Gourab Kundu, and Kirmani Ahmed. 2022. Normalized Contrastive Learning for Text-Video Retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

[28] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems* 28 (2015).

[29] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2022. Sketch3T: Test-Time Training for Zero-Shot SBIR. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*. IEEE, 7452–7461.

[30] Aneeshan Sain, Ayan Kumar Bhunia, Vaishnav Potlapalli, Pinaki Nath Chowdhury, Tao Xiang, and Yi-Zhe Song. 2022. Sketch3t: Test-time training for zero-shot sbir. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 7462–7471.

[31] Samarth Sinha, Peter Gehler, Francesco Locatello, and Bernt Schiele. 2023. TeST: Test-time Self-Training under Distribution Shift. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2759–2769.

[32] Yongyi Su, Xun Xu, and Kui Jia. 2022. Revisiting Realistic Test-Time Training: Sequential Inference and Adaptation by Anchored Clustering. *arXiv preprint arXiv:2206.02721* (2022).

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

[34] Guolong Wang, Xun Wu, Zhaoyuan Liu, and Junchi Yan. 2022. Prompt-based Zero-shot Video Moment Retrieval. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 413–421.

[35] Haoran Wang, Di Xu, Dongliang He, Fu Li, Zhong Ji, Jungong Han, and Errui Ding. 2022. Boosting Video-Text Retrieval with Explicit High-Level Semantics. In *Proceedings of the 30th ACM International Conference on Multimedia*. 4887–4898.

[36] Ziyue Wang, Aozhu Chen, Fan Hu, and Xirong Li. 2022. Learn to Understand Negation in Video Retrieval. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 434–443.

[37] Zheng Wang, Jingjing Chen, and Yu-Gang Jiang. 2021. Visual Co-Occurrence Alignment Learning for Weakly-Supervised Video Moment Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 1459–1468. https://doi.org/10.1145/3474085.3475278

[38] Peng Wu, Xiangteng He, Mingqian Tang, Yiliang Lv, and Jing Liu. 2021. HANet: Hierarchical Alignment Networks for Video-Text Retrieval. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*. ACM, 3518–3527.

[39] Wenhao Wu, Zhun Sun, and Wanli Ouyang. 2023. Revisiting classifier: Transferring vision-language models for video recognition. *Proceedings of the AAAI, Washington, DC, USA* (2023), 7–8.

[40] Chen-Wei Xie, Siyang Sun, Liming Zhao, Jianmin Wu, Dangwei Li, and Yun Zheng. 2022. Deep Video Understanding with a Unified Multi-Modal Retrieval Framework. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 7055–7059.

[41] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 5288–5296.

[42] Xun Yang, Fuli Feng, Wei Ji, Meng Wang, and Tat-Seng Chua. 2021. Deconfounded video moment retrieval with causal intervention. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1–10.

[43] Bolin Zhang, Chao Yang, Bin Jiang, and Xiaokang Zhou. 2022. Video Moment Retrieval with Hierarchical Contrastive Learning. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*. ACM, 346–355.

[44] Haotian Zhang, Allan D. Jepson, Iqbal Mohomed, Konstantinos G. Derpanis, Ran Zhang, and Afsaneh Fazly. 2021. Personalized Multi-modal Video Retrieval on Mobile Devices. In *MM '21: ACM Multimedia Conference, Virtual Event, China, October 20 - 24, 2021*, Heng Tao Shen, Yueting Zhuang, John R. Smith, Yang Yang, Pablo César, Florian Metze, and Balakrishnan Prabhakaran (Eds.). ACM, 1185–1191. https://doi.org/10.1145/3474085.3481545

[45] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. 2021. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 11–20.

[46] Shengwei Zhao, Yuying Liu, Shaoyi Du, Zhiqiang Tian, Ting Qu, and Linhai Xu. 2023. CMFG: Cross-Model Fine-Grained Feature Interaction for Text-Video Retrieval. In *MultiMedia Modeling: 29th International Conference, MMM 2023, Bergen, Norway, January 9–12, 2023, Proceedings, Part II*. Springer, 435–445.