electronics

MDPI

Article

Causal Visual–Semantic Enhancement for Video-Text Retrieval

Hua Lan [1] and Chaohui Lv [1,2,*]

[1] School of Information and Communication Engineering, Communication University of China, Beijing 100000, China; lanhua@cuc.edu.cn
[2] Key Laboratory of Acoustic Visual Technology and Intelligent Control System, Ministry of Culture and Tourism, Beijing 100000, China
* Correspondence: llvch@cuc.edu.cn

Abstract

The core challenge in video-text retrieval lies in measuring the cross-modal gap between visual and linguistic representations. While mainstream methods leverage deep learning to project video and text features into a shared space for similarity matching, they often rely on superficial statistical correlations. This limits their ability to model underlying causal relationships or capture high-level semantics, resulting in poor interpretability. To address this, we analyze the biases existing in video-text retrieval tasks and introduce a Causal Visual–Semantic Enhancement (CVSE) method that integrates causal inference with deep learning. Our approach applies causal intervention to mitigate bias from contextual confounders and uses textual descriptions as a condition to guide frame aggregation, emphasizing semantically relevant frames while suppressing redundant ones. Experiments on MSR-VTT, MSVD, and LSMDC demonstrate that the proposed method outperforms state-of-the-art retrieval models, validating its superior performance.

Keywords: visual–semantic; causal inference; computer vision; video-text retrieval

1. Introduction

Cross-modal video-text retrieval [1–8] has extensive applications in fields such as social media video search, intelligent security, content retrieval, and e-commerce product search. Most current video-text retrieval [9–11] studies leverage the powerful feature extraction capabilities of deep learning to learn multimodal feature representations, then map them into a common space for similarity calculation. For example, the studies reported in [12–15] employ convolutional neural networks, recurrent neural networks, or expert networks to extract video and text features. Although these methods have achieved promising results, their performance is hindered by the challenges of end-to-end optimization. Many researchers have leveraged the semantic extraction capabilities brought about by the breakthrough development of the pre-trained CLIP model [16–22] for image-text retrieval—learned from 400 million image-text pairs—to achieve end-to-end video-text retrieval via fine tuning. However, most deep learning-based video-text retrieval methods rely on superficial data correlations, lacking the modeling of underlying causal relationships. Moreover, their over-reliance on the distribution of the training data limits their generalizability. The black-box nature of deep learning also makes such model predictions uninterpretable to humans.

Although correlation-based deep learning methods have improved performance to some extent, correlation does not imply causation. The features acquired through deep learning typically can only capture the low-level semantic information of video text, such

Check for updates

Academic Editor: Stefanos Kollias

Received: 23 December 2025
Revised: 21 January 2026
Accepted: 26 January 2026
Published: 9 February 2026

Copyright: © 2026 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license.

as "what" and "where", but fail to capture higher-level contextual semantic information, such as "why". In addition, the redundancy of video frames or textual tokens, combined with inherent dataset biases and over-optimization, can lead to spurious correlations during training. Moreover, prior knowledge-based debiasing methods and fine tuning-based debiasing methods are often inadequate for current video-text models. Therefore, learning the causal relationships hidden behind the correlations is the key to improving performance and ensuring model robustness. Causal inference [23] provides a principled framework for this purpose. It can identify true cause–effect relationships and eliminate the influence of confounding factors, thereby mitigating spurious correlations arising from issues like data imbalance and bias. Among various causal inference techniques, back-door adjustment is used primarily for observable confounders, whereas front-door adjustment handles unobservable confounders. As a more advanced form of reasoning, counterfactual inference enables "what-if" hypothetical analysis. Studies have shown that incorporating causal inference into vision–language tasks can significantly enhance model performance and robustness [24–29].

To this end, this paper integrates causal inference theory with deep learning to propose a visual–semantic enhanced video-text retrieval method. By analyzing the video-text retrieval process, we observe that videos contain not only contextual observational co-occurrence biases but also common-sense causal relationships within visual semantics. We utilize causal intervention to eliminate the harmful effects of contextual observational co-occurrence bias information, excavate common-sense causal visual features, and obtain more causal contextual semantic information. Additionally, we use text information as a condition for frame aggregation to enhance the training of frames most semantically similar to the text description and weaken redundant frames. This enables the model to better leverage visual–semantic causal relationships to learn video feature representations.

The main contributions of this paper are outlined as follows:

(1) We analyze and identify the existence of contextual observational co-occurrence bias in video-text retrieval tasks, as well as the common-sense causal relationships that are implicitly embedded within visual semantics.
(2) To eliminate the harmful effects of contextual observational co-occurrence bias, we construct a causal graph and adopt a back-door intervention to extract common-sense causal visual features. Furthermore, we leverage the text as a conditional guide for frame aggregation, which enhances the representation of semantically relevant frames while suppressing redundant ones.
(3) We design experiments to verify the effectiveness of our method. Experimental results on large-scale datasets (MSR-VTT, MSVD, and LSMDC) show that the proposed method significantly outperforms state-of-the-art retrieval models in retrieval performance.

2. Related Work

Video-text retrieval: Video-text retrieval aims to achieve cross-modal matching by computing similarity scores between feature representations extracted from video and text data via deep learning. This task primarily involves three components: the extraction of video and text feature representations, feature embedding and matching, and objective function optimization. In recent years, due to core challenges such as the learning of efficient spatiotemporal video representations, bridging of the heterogeneity gap between modalities, and optimization of model parameters, video-text retrieval remains a challenging and highly active research field. Based on different feature extraction and alignment approaches, existing video-text retrieval methods can be categorized into three groups: (1) Coarse-grained feature-matching methods: These methods extract holistic se-

mantic information from video–text pairs, map them to a common space, and employ loss functions to optimize cross-modal semantic alignment. (2) Fine-grained feature-matching methods: These methods focus on aligning fine-grained units (e.g., video frames) with textual words or phrases, then aggregate local alignments to derive an overall video-text similarity measure. (3) Multi-grained feature-matching methods: These methods employ multi-level interactive matching between fine-grained and coarse-grained features, aggregating global–local, local–local, and global–global alignments to compute the final similarity score. Methods in category (1) (e.g., the work reported in [19,30,31]) often overlook fine-grained correspondences between video frames and textual words, making it difficult to capture local semantic associations and limiting their adaptability in complex retrieval scenarios. Methods in category (2) (e.g., [6,32,33]) achieve accurate local semantic mapping by annotating frame–word correspondences, but they heavily depend on high-quality, fine-grained annotations. Building on (1) and (2), methods in category (3) (e.g., [22,34–36]) introduce interactive alignment across different granularities, enhancing retrieval accuracy through multi-level feature interaction. However, these models fundamentally learn data correlations rather than true causal relationships, making them susceptible to spurious correlations caused by dataset biases. For instance, during dataset collection, certain scenes or activities may be over- or under-represented, while manual annotations may introduce linguistic biases through overused words or phrases. In summary, despite significant progress, current deep learning-based video-text retrieval methods primarily model data correlations, lacking the ability to infer robust causal relationships. Challenges in terms of model interpretability and generalization also persist.

Causal inference: Causal inference methods [23,37–41] have been successfully applied across diverse fields, including psychology, economics, medicine, and artificial intelligence. Unlike deep learning algorithms that predominantly capture correlational patterns, causal inference seeks to uncover causal effects among variables by removing confounding factors through interventions—such as front-door and back-door adjustments—along with counterfactual analysis. Causal relationships provide more stable foundations for decision making compared to the often-limited and weakly generalizable correlations identified by deep learning models. Furthermore, although effective, deep learning frequently operates as an opaque "black box", limiting interpretability. Consequently, the integration of causal inference into deep learning has attracted growing research interest, aiming to enhance both generalization ability and model interpretability.

This integrative trend is evident in various vision–language tasks. For example, in object detection, visual question answering (VQA), and image captioning, a causal reasoning framework [42] systematically addresses the out-of-distribution (OOD) problem in synthetic aperture radar (SAR) target recognition via causal graph modeling and intervention analysis. Its core idea is to explicitly separate background biases through causal intervention, thereby guiding models to focus on intrinsic target features. The study reported in [25] introduces a failure cause analysis (FCA) method for object detection algorithms based on causal inference and random walks. To eliminate implicit bias in multimodal fake-news detection, ref. [26] proposes a novel debiasing framework termed CCD (Causal Intervention and Counterfactual Reasoning Debiasing), which formulates the detection process as a causal graph. For explainable VQA, ref. [27] presents a Variational Causal Inference Network (VCIN) equipped with a multimodal explanation gating transformer. In the medical domain, ref. [28] designs a counterfactual causal effect intervention framework to mitigate language bias and improve interpretability in Medical VQA (VQA-Med). For text-based person retrieval, ref. [29] approaches the task from a causal perspective to extract robust, text-critical visual representations and establish domain-invariant cross-modal

correlations. Additionally, ref. [43] proposes a vision- and language-based framework for common-sense visual causal reasoning from still images.

In video-text retrieval, ref. [31] designs a semantic causal reasoning module that enables video and text input to adaptively capture full-memory contextual causal correlations within their respective feature sequences. Aligned with this research direction, our work introduces causal intervention mechanisms into the video-text retrieval task. Specifically, to address the key limitation of existing models only capturing observational co-occurrence information within limited visual contexts while neglecting implicit observational biases and visual–semantic causal relationships (e.g., causal dependencies between scenes and actions), we propose the elimination of interference from contextual co-occurrence biases through causal intervention. Simultaneously, our approach mines common-sense causal visual features to derive more causally coherent contextual visual–semantic representations.

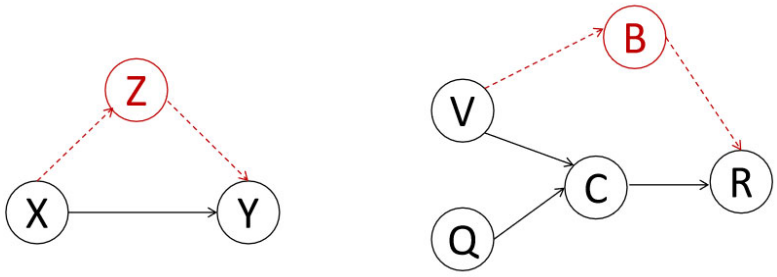
3. Proposed Method

In this section, we first introduce causal graphs and causal interventions. We then analyze the retrieval process of the baseline video-text retrieval model, pointing out that the encoders in existing video-text retrieval models tend to capture contextual co-occurrence information, along with observational biases or irrelevant details, while overlooking the implicit common-sense causal relationships present in visual semantics. To mitigate the influence of such biases, we construct a causal graph, apply back-door adjustment to uncover latent common-sense causal relations, and leverage textual information as a conditioning signal for frame aggregation. This enhances the training of frames that are semantically closest to the text description and suppresses redundant frames.

3.1. Causal Graph and Causal Intervention

A causal graph is a directed acyclic graph (DAG) that describes the interactions between variables and serves as a probabilistic model for analyzing causal relationships among variables. As shown in Figure 1a, nodes X, Y, and Z represent three interacting variables, where X is the cause variable, Y is the effect variable, and $X \rightarrow Y$ indicates that X is the cause of Y. Z is a confounder that influences the $X \rightarrow Y$ relationship. In the presence of the confounder (Z), $p(Y|X)$ can only learn the correlation between variables X and Y. Therefore, to explore the true causal effect of variable X on variable Y, it is necessary to eliminate the influence of confounders through intervention (do-calculus). Main causal interventions in causal inference include back-door adjustment and front-door adjustment, which involve intervening to ensure that confounders no longer affect the effect variable (Y). This paper assumes that the confounder is the observational co-occurrence bias in video-frame contexts. We adopt the back-door adjustment method by truncating the back-door path in the causal graph through deletion of the incoming edges of X. According to Bayes' theorem, the post-intervention probability distribution can be derived as

$$p_{do}(Y|X) = \sum_{Z} p_{do}(Y|X,Z)p_{do}(Z|X) \tag{1}$$



(**a**) A causal graph (**b**) Causal diagram for video-text retrieval

Figure 1. Interventional causal graph. (**b**) From the perspective of causal inference, identify the formation mechanism of biases in video-text retrieval. The results of video-text retrieval are affected by two paths—namely, the paths expressed as $(V, Q) \rightarrow C \rightarrow R$ and $V \rightarrow B \rightarrow R$.

3.2. Common-Sense Causal Relationships in Visual Semantics

All existing video-text datasets rely on manual annotation. As shown in Figure 2, variations in individual cognition and emotional preferences lead annotators to focus on different video frames or regions when describing the same video. In some cases, textual descriptions of identical content may even contradict one another. Moreover, videos typically contain richer information than their corresponding textual descriptions. Weak semantic captions and limited annotated data not only fail to capture complete visual semantics but may also cause models to learn spurious correlations and irrelevant features during pre-training for video-text alignment, ultimately degrading retrieval performance on test sets. Additionally, the spatiotemporal nature of videos introduces common-sense causal relationships within visual contexts. As illustrated in Figure 3, given the caption, "A person is cooking in the kitchen", the visual scene might show a person in a kitchen peeling an apple with a knife. One can infer that the person entered the kitchen earlier and took out an apple; their current intention could be to eat the peeled apple or to later use it to bake an apple pie. In such everyday scenarios, humans can readily reason about the underlying causal relationships linking visual and textual content. For machines, however, understanding these cross-modal semantics remains challenging due to the modality gap, which hinders genuine comprehension of semantic connections.

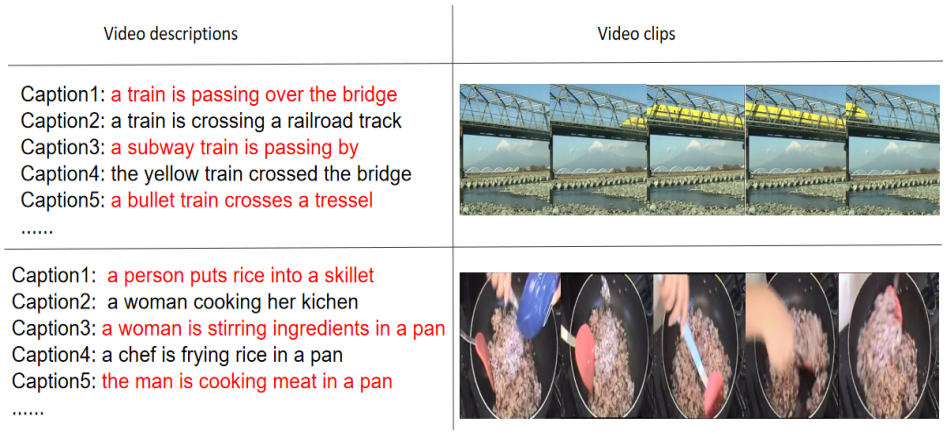


Figure 2. Two examples from the MSR-VTT dataset.

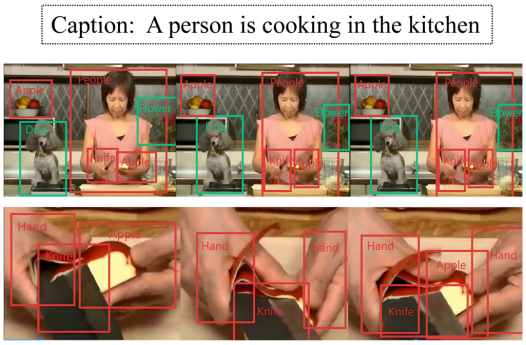


Figure 3. There exists causal information rooted in visual–semantic common sense and observational co-occurrence bias information in the video context. The information in the red boxes represents causal information rooted in visual–semantic common sense, while the information in the green boxes represents observational co-occurrence or fake information.

Current computational approaches to video-text understanding are largely confined to the learning of co-occurrence correlations. For instance, if a video segment simultaneously contains "person", "dog", and "knife", an encoder tends to learn context-dependent co-occurrence patterns with observational bias, i.e., it implicitly assumes that these elements are interrelated. Consequently, querying "dog" may retrieve videos featuring "knife" and "person". Although such statistically correlated co-occurrence signals can improve in-domain retrieval performance, the underlying visual–semantic relationships are often semantically unreasonable. They lack higher-order cognition of common-sense causality and generalize poorly beyond the training distribution.

To address this issue, we incorporate causal inference theory and employ causal intervention to mitigate the influence of context-dependent observational co-occurrence bias. This enables the model to better leverage contextual visual–semantic causal structures for the learning of video representations.

Specifically, for all instances in all datasets, let $V = \{v_i\}_{i=1}^{n_v}$ be the input video sequence, n_v represent the number of videos, $Q = \{q_i\}_{i=1}^{n_q}$ be the text sequence, and n_q denote the number of text descriptions. Assuming a given text (q_i), retrieval of video (v_i) from the video dataset is influenced by two factors with respect to video features: one is the semantic relationship between the video (v_i) and text (q_i), and the other is the bias relationship describing the superficial statistical correlation between v_i and q_i (i.e., observational bias information in the context), since semantic relationships exhibit causal invariance and remain constant across all instances, while bias relationships lack stability and general applicability. To prevent the training model from relying on bias relationships for prediction while ignoring common-sense causal relationships (as shown in Figure 1b), we constructed a causal graph for the video-text retrieval task. Nodes V, Q, and C represent video features, text features, and multimodal features for video–text matching, respectively. Node R denotes the video–text matching ranking. As mentioned above, assuming that observational biases between video and text all originate from contextual video features, B refers to the observational bias information in the context during training. According to the structural causal graph, the final video–text matching ranking (R) is influenced by two branches: The influence of the semantic relationship features of inputs V and Q on R is expressed as $(V, Q) \rightarrow C \rightarrow R$, and the influence of observational co-occurrence bias in the video context is expressed as $V \rightarrow B \rightarrow R$. Current video-text retrieval models learn the explicit co-occurrence correlation ($p = (R|(V, Q))$) between videos and texts. However, due to their neglect of implicit causal relationships, these models exhibit weak generalizability and cannot handle datasets with complex and variable scenarios. To address this issue, we emphasize the learning of video–text semantic relationships in the $(V, Q) \rightarrow C \rightarrow R$ branch.

Specifically, we use text information as a condition for frame aggregation, strengthen the training of frames semantically most similar to the text description, weaken the learning of redundant frames, and achieve causal visual–semantic alignment between videos and texts. Additionally, in the $V \to B \to R$ branch, the model is trained on the probability ($p = (R|(V, Q))$) of learning the observational bias information (B) in the context. To sever the $V \to B$ connection and eliminate the influence of confounding factors (observational bias information in the context), we adopt causal intervention via back-door adjustment. The training probability of the model after intervention is expressed as follows:

$$p_{do}(R|V) = \sum_{B} p_{do}(R|V, B) p_{do}(B|V) \tag{2}$$

After back-door adjustment, the input variable (V) and the confounding factor (B) become independent of each other, and $p_{do}(B|V)$ is equivalent to the marginal distribution ($p_{do}(B)$). Since the confounding factor (B) has no incoming edges, the marginal distribution remains consistent with the original distribution after intervention. The conditional probability of R, conditioned on the input variable (V) and the confounding factor (B), remains unchanged after intervention—identical to the original conditional probability. Therefore, the conditional probability after intervention (do-calculus) can compute an approximate conditional probability of the original distribution during training, denoted as (3).

$$p(R|do(V)) = p_{do}(R|V) = \sum_{B} p(R|V, B) p(B) \tag{3}$$

where $p(R|V, B)$ is the prediction result of the training model based on the observational co-occurrence bias information in the video context.

3.3. Causal Visual–Semantic Enhancement for Video-Text Retrieval

The core task of video-text retrieval is to retrieve semantically matched videos (v_i) from a candidate video pool based on a given textual description (q_i). To achieve more robust retrieval performance, we propose a method based on causal visual–semantic enhancement: by eliminating interfering factors in visual contexts and deeply exploring the causal semantic correlations between videos and texts, we further improve the accuracy of cross-modal matching. The overall framework is illustrated in Figure 4. The video-text retrieval model based on causal visual semantics consists of a video encoder, a text encoder, and a feature embedding and matching component.

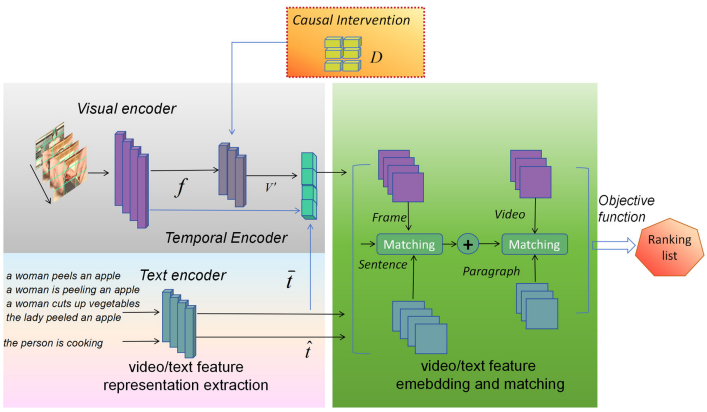


Figure 4. Overall framework of causal visual–semantic enhancement for video-text retrieval. Specifically, we introduce the Visual Genome dataset into the causal intervention module and leverage its rich visual object relationship information to analyze each video frame, thereby reducing the model's reliance on visual context co-occurrence information and alleviating contextual bias during training.

For text representation, we employ the multi-layer bidirectional transformer in the Bert model as the text encoder to extract fine-grained hierarchical features of the text. First, to aggregate frames with the most comprehensive visual semantics, we merge all annotations of a video into a paragraph; input the paragraph into the text encoder for learning; and output a paragraph-level text representation ($\bar{t}$) to describe the video, marked with the [CLS] token. Finally, we re-input each sentential annotation of the video into the text encoder, where the [CLS] token output for each sentence is used to construct the sentential representation of the video, that is, the text representation of the video can be denoted as $t = \left\{q_1, q_2, q_3, \cdots, q_{n_q}\right\}$, where n_q represents the number of text sentences. We randomly select N sentences and use $\hat{t} = \{t_1, t_2, t_3, \cdots, t_N\}$ as the sentence-level features. We input a given video (v_i) into a ViT/32 video encoder. First, we uniformly sample the video using a sampling method, taking one frame per second, then split the sampled frames into multiple regions. To obtain more visual–semantic information, we add [CLS] tokens and position tokens to each region and use the [CLS] token extracted by the last transformer layer as the frame-level feature ($f = \{f_1, f_2, f_3, \cdots .f_{n_{v_i}}\}$, where n_{v_i} represents the number of video frames in video v_i). Additionally, considering the temporal sequence of the video, we adopt the temporal transformer in Clip4clip to mine the temporal relationships between video frames. Each video-frame feature is first appended with a position token before being input into the temporal transformer. Finally, the output values are subjected to average pooling to obtain the video-level feature (V'), denoted as follows:

$$\hat{f}_i = TransEnc(f_i + \delta_i) \tag{4}$$

$$V' = \frac{1}{n_{v_i}} \sum_{i}^{n_{v_i}} \hat{f}_i \tag{5}$$

where $TransEnc(.)$ represents the transformer encoder, $\hat{f}_i$ is the output of video frame f_i after passing through the temporal transformer, and δ_i denotes the position marker added to video frame f_i.

The video-text retrieval method based on visual–semantic enhancement proposed in this paper consists of two key steps: (1) In the $(V, Q) \rightarrow C \rightarrow R$ branch, taking text information as the condition for frame aggregation, it achieves causal visual–semantic alignment between video and text by enhancing the training of frames with semantics most similar to the text description and weakening the learning of redundant frames. (2) In the $V \rightarrow B \rightarrow R$ branch, harmful effects of confounding factors (observational bias information in the context) are eliminated through causal intervention on the input (V).

To capture more comprehensive and refined semantic information from video–text pairs, we introduce a text-guided multi-head attention mechanism in the $(V, Q) \rightarrow C \rightarrow R$ branch. Specifically, we use text features as Queries (Q), while video features serve as both Keys (K) and Values (V). When presented with both frame-level local features and video-level global features of a video, we leverage their complementary nature to strengthen the learning of video–text semantic relationships. More specifically, first, we aggregate the video-level global features and frame-level local features through matrix operations. The aggregated video features ($\hat{v}$) are then processed via a linear transformation using a fully connected layer. This step maps both video and text features into a shared semantic space. Finally, interactive video features are derived through a residual network structure, which helps preserve and enhance the cross-modal information flow.

$$\hat{V} = Attention\left(\bar{t}, \hat{v}, \hat{v}\right) \tag{6}$$

$$\overline{v} = FFN\left(LN\left(Pool(f) + Pool\left(\hat{V}\right)\right)\right) \tag{7}$$

where $LN(\cdot)$ denotes layer normalization, $Pool(\cdot)$ represents average pooling, and $FFN(\cdot)$ stands for feed-forward network.

In the $V \rightarrow B \rightarrow R$ branch, the confounding factor (B) can be regarded as a collection of observational co-occurrence information in the video context. Extracting observational contextual information without semantic relationships from local regions of video frames influences the input video representation (V), thereby affecting the matching and ranking of the retrieval model simultaneously. The optimal solution for handling confounding factors in causal inference is to average different confounding factors, that is, to sum and average each contextual confounding factor and assign the maximum weight to the more reliable signal with fewer error signals. Given the video representation (V) and the confounding-factor representation (B), according to Equation (3), $p(R|do(V))$ is the expectation of the conditional probability ($p(R|V, B = b)$) with respect to the distribution of B.

$$p(R|do(V)) = \mathrm{E}_b[p(R|V, B = b)] = \mathrm{E}_b[\delta(\phi(v, b))] \tag{8}$$

where $\phi(\cdot)$ represents a training model based on the observational co-occurrence bias information of the video context, while $\delta(\cdot)$ denotes the output mapping of the retrieval model with respect to $\phi(\cdot)$. Our model ($\delta(\cdot)$) is defined as $Softmax(.)$.

In the video-text retrieval task, each regional feature ($f_{n_{vi}}$) can be regarded as a value of the confounding factor (B); thus, the model needs to calculate the expectation ($\mathrm{E}_b[\phi(v, b)]$). $\mathrm{E}_b[\phi(v, b)]$ learns the probability weight of each visual region acting as a confounding factor. Based on this, we approximate the expected value in the above formula using the Normalized Weighted Geometric Mean (NWGM) method by introducing an external expected value. Specifically, NWGM assumes that the Softmax function can be approximately swapped inside and outside the expectation. Therefore, we denote this as $p(R|do(V)) \approx \delta(\mathrm{E}_b[\phi(v, b)])$. Subsequently, we approximate the conditional probability using a linear model.

$$\phi(v, b) = W_v v + W_b \cdot \ell(b) \tag{9}$$

$$\mathrm{E}_b[\phi(v, b)] = W_v v + W_b \cdot \mathrm{E}_b[\ell(b)] \tag{10}$$

where W_v and W_b are learnable weight parameters.

To compute $\mathrm{E}_b[\ell(b)]$, we introduce the Visual Genome dataset [44] and obtain an external knowledge set (D) from it. We treat B as a set of fixed confounding factors ($D_b = \{b_1, b_2, b_3, \cdots, b_n\}$, where n denotes the feature dimension).

$$\mathrm{E}_b[\ell(b)] = \sum_b \left[softmax\left(\frac{Q^T K}{\sqrt{N_m}}\right) \Theta D_b \right] p(b) \tag{11}$$

For Equation (11), we employ the scaled dot-product attention mechanism ($Q = W_1 v$, $K = W_2 D_b$, where W_1 and W_2 are learnable weight parameters). Θ and $p(b)$ denote element-wise product and prior statistical probability, respectively.

3.4. Model Training

Based on the aforementioned overall framework, we proceed to elaborate the model's training objectives, feature-matching strategies, and optimization methods in detail. In this paper, we adopt cosine similarity to measure the semantic relevance between videos and texts, employ symmetric cross-entropy loss as the objective function, an adaptive learning rate optimizer. At the level of feature embedding and matching, we further introduce two

techniques—frame–sentence matching and video–paragraph matching—to enhance the granularity and robustness of cross-modal alignment.

For frame–sentence matching, we employ the following equations:

$$L_{V2T} = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp\left(s\left(f_i, \hat{t}_i\right)\right)}{\sum_{j=1}^{n} \exp\left(s\left(f_i, \hat{t}_j\right)\right)} \tag{12}$$

$$L_{T2V} = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp\left(s\left(\hat{t}_i, f_i\right)\right)}{\sum_{j=1}^{n} \exp\left(s\left(\hat{t}_i, f_j\right)\right)} \tag{13}$$

$$L_{FS} = \frac{1}{2}\left(L_{V2T} + L_{T2V}\right) \tag{14}$$

For video–paragraph matching, we employ the following equations:

$$l_{V2T} = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp\left(s\left(\overline{v}_i, \overline{t}_i\right)\right)}{\sum_{j=1}^{n} \exp\left(s\left(\overline{v}_i, \overline{t}_j\right)\right)} \tag{15}$$

$$l_{T2V} = \frac{1}{n}\sum_{i=1}^{n} \frac{\exp\left(s\left(\overline{t}_i, \overline{v}_i\right)\right)}{\sum_{j=1}^{n} \exp\left(s\left(\overline{t}_i, \overline{v}_j\right)\right)} \tag{16}$$

$$l_{VP} = \frac{1}{2}\left(l_{V2T} + l_{T2V}\right) \tag{17}$$

where n denotes the batch size and the similarity function ($s(\cdot)$) calculates the cosine similarity between video and text. L_{V2T} ensures that the similarity ($s\left(f_i, \hat{t}_j\right)$) between the video frame (f_i) and the text-description sentence ($\hat{t}_i$) from the same video–text pair is greater than the similarity ($s\left(f_i, \hat{t}_j\right)$) between F and any other text-description sentence ($\hat{t}_j$) from the video frame–text pairs. The same logic applies to L_{T2V}, L_{T2V}, and l_{T2V}. Therefore, by minimizing the sum of the two loss functions, the model is prompted to learn the regions where video and text semantics are most similar.

$$L = \mu L_{FS} + \lambda l_{VP} \tag{18}$$

where μ and λ are parameters within the range of 0 to 1.

4. Experiments and Results

To verify the effectiveness of the proposed method, this study conducts experiments on three cross-modal public video-text retrieval datasets: MSR-VTT [45], MSVD [46], and LSMDC [47]. First, we compare our method with state-of-the-art video-text retrieval techniques. Then, we validate the performance of the proposed method from two perspectives: strategy ablation and alignment ablation. Finally, to provide an intuitive demonstration, we present the visualization results of partial retrieval.

4.1. Experimental Data and Evaluation Metrics

The following is a detailed description of common datasets containing large numbers of video clips and their corresponding annotation descriptions, with comparisons shown in Table 1.

The MSR-VTT [45] dataset contains 10,000 video clips selected from 7180 videos on YouTube, covering 20 different domains, such as family, anime, cooking, TV interviews, and sports events. Each video clip was manually annotated with 20 different text descriptions by staff, and each video has a duration of $10 \sim 32$ s. The total duration of the entire dataset exceeds 40 h. Additionally, there are three types of data-splitting methods for test/training division: full split, 'Training-7K', and 'Training-9K'. The full split uses 7000 videos for

training and 3000 videos for testing. This paper adopts 'Training-9K', where 9000 videos and their corresponding text descriptions are used as the training dataset. Since each video corresponds to 20 text descriptions, the actual training set contains 180,000 video–text pairs. The test set consists of the remaining 1000 videos and one corresponding text for each, forming 1000 video–text pairs.

Table 1. Datasets for video–text retrieval.

Datasets	Clips	Captions	Source	Caption Source	Length	Year
MSR-VTT [45]	10,000	200,000	YouTube	Manual	10~30 s	2016
MSVD [46]	1970	78,800	YouTube	Manual	1~62 s	2011
LSMDC [47]	118,081	118,081	Movies	Manual	2~30 s	2015

The MSVD [46] dataset contains 1970 videos, each with a duration of $1 \sim 62$ s. The number of manually written text descriptions for each video clip varies, with an average of approximately 40 text descriptions per video. In this paper, 1200 video–text pairs are used as the training set, and the remaining 100 and 670 pairs are used as the validation set and test set, respectively.

The LSMDC [47] dataset contains 118,081 video–text pairs selected from 202 different movies, with a total duration of approximately 158 h. The dataset is formed by merging the M-VAD and MPII-MD datasets, where each video corresponds to one text description and each video has a duration of $2 \sim 30$ s. In this paper, 101,079 video–text pairs are used as the training set, 7408 pairs as the validation set, and the remaining 1000 pairs as the test set.

This paper draws R@K, MdR, and MnR as the evaluation metrics for video-text retrieval. R@K represents the ratio of relevant query instances in the top K retrieval results to the total number of query instances. The higher the R@K value, the better the retrieval performance. In this paper, we mainly focus on R@1, R@5, and R@10. MdR represents the median of the ranking position of relevant instances in the retrieval results, with lower values indicating better results. MnR represents the mean of the ranking position of relevant instances in the retrieval results, with lower values indicating better retrieval performance.

$$R@K = \frac{correct \quad matches}{top\text{-}k \quad results} \tag{19}$$

$$MdR = median(rank(y_i)) \tag{20}$$

$$MnR = \frac{1}{|Q|}\sum_{q \in Q} rank(y_q) \tag{21}$$

where $rank(y_i)$ represents the ranking of the correctly matched result (y_i) in the model's ranking results and Q is the set of queries.

4.2. Experimental Details

We trained the model on eight NVIDIA A100 GPUs and used PyTorch (1.13.1) software to edit the program code. To reduce computational and memory requirements, this paper adopted a downsampling process of one frame per second for videos in the MSVD, MSR-VTT, and LSMDC datasets. Taking the Clip4clip model as the backbone, this paper used a 12-layer ViT/32 video encoder and the multi-layer bidirectional transformer in the Bert model as the text encoder, with the initial value of the dynamic learning rate set to 0.00001. The text-description length of all datasets was set to 32, and the video length was set to 12. During training, the number of samples per batch was 32, and the AdamW optimizer was used to train for 200,000 rounds.

4.3. Experimental Results

To evaluate the effectiveness of the method proposed in this paper, we conducted experiments on the MSR-VTT, MSVD, and LSMDC datasets and compared them with previous state-of-the-art (SOTA) methods. We performed text-to-video retrieval and video-to-text retrieval experiments on the MSR-VTT dataset and compared them with DFLB [30], Memory Enhancing [31], CLIP2Video [20], X-CLIP [21], TS2-Net [18], X-Pool [19], Prompt Switch [17], CLIP4Clip [22], MSIA [36], LSECA [48], DGL [35], and TempMe [3]. The results are shown in Table 2.

Table 2. Comparison of the retrieval results of state-of-the-art techniques on the MSR-VTT dataset.

Method		Text to Video					Video to Text				
		R@1	R@5	R@10	MnR	MdR	R@1	R@5	R@10	MnR	MdR
NO CLIP AND CAUSAL INFERENCE	DFLB [30]	24.64	52.99	66.09	26.26	5.0	-	-	-	-	-
	Memory Enhancing [31]	**24.8**	**51.6**	**64.9**	-	-	-	-	-	-	-
CLIP-BASED AND CROSS-MODAL TEMPORAL FUSION	CLIP2Video [20]	45.6	72.6	81.7	14.6	-	43.5	72.3	82.1	10.2	2.0
	X-CLIP [21]	46.1	73.0	83.1	13.2		46.8	73.3	84.0	9.1	2.0
	TS2-Net [18]	46.5	73.6	**83.3**	13.9	-	45.3	74.1	83.7	9.2	2.0
	X-Pool [19]	46.9	72.8	82.2	14.3	2.0	44.4	73.3	84.0	9.0	2.0
	Prompt Switch [17]	**47.8**	**73.9**	82.2	14.1	-	**46.0**	**74.3**	**84.8**	8.5	-
CLIP-BASED AND COARSE-TO-FINE REPRESENTATION LEARNING	CLIP4Clip [22]	44.5	71.4	81.6	15.3	2.0	42.7	70.9	80.6	11.6	2.0
	MSIA [36]	**47.2**	73.8	**84.1**	-	2.0	-	-	-	-	-
	LSECA [48]	47.1	**74.9**	82.8	14.9	2.0	**47.5**	**75.4**	**83.4**	12.3	2.0
CLIP-BASED AND PARAMETER/INFERENCE-EFFICIENT	DGL [35]	45.8	69.3	79.4	16.3	-	**43.5**	**70.5**	**80.7**	13.1	-
	TempMe [3]	**46.1**	**71.8**	**80.7**	14.8	-	-	-	-	-	-
CLIP-BASED AND CAUSAL INFERENCE	**Ours**	**49.0**	**74.1**	**83.3**	12.4	2.0	**45.8**	**74.4**	**83.4**	11.5	2.0

Note: For each category (e.g., "NO CLIP AND CAUSAL INFERENCE," "CLIP-BASED AND CROSS-MODAL TEMPORAL FUSION," etc.), the highest value among all evaluation metrics (R@1, R@5, R@10) for the "Text to Video" and "Video to Text" tasks within that category is selected and bolded. If there are multiple models under the same classification, the comparison is made among all models in that category.

As can be seen from Table 2, our method performs excellently on various indicators and also shows outstanding performance in competition with the SOTA models of other methods. Specifically, compared with the DFLB [30] and Memory Enhancing [31] methods, which use causal inference but do not use the pre-trained Clip model, our proposed model achieves significant improvements in all evaluation indicators. In particular, our method improves the R@1 indicator by 24.2% compared with Memory Enhancing [31], which reflects the excellent performance of the pre-trained Clip model in multimodal feature extraction and representation and also proves the effectiveness of using Clip in our proposed method. When compared with methods using Clip and cross-modal temporal features, such as CLIP2Video [20], X-CLIP [21], TS2-Net [18], X-Pool [19], and Prompt Switch [17], our method captures the temporal features of videos and uses text information as the condition for frame fusion, which is somewhat similar to them. As can be seen from most of the indicators in Table 2, our method only improves the R@1 indicator by 1.2% and the R@5 indicator by 0.2% compared with Prompt Switch [17]. In addition, when compared with methods using Clip and learning coarse and fine feature representations, such as CLIP4Clip [22], MSIA [36], and LSECA [48], our method improves the R@1 by 1.8% compared with MSIA [36], and other indicators are also improved. We also compared out method with methods using Clip and learning efficient parameters or efficient inference, such as DGL [35] and TempMe [3]. Our method outperforms TempMe [3] in most indicators and reaches 49.0 in R@1, corresponding to an increase of 2.9%.

Similarly, our method was validated on the MSVD and LSMDC datasets. Tables 3 and 4 show the experimental results and comparisons with other methods. On the MSVD dataset, we compared two types of methods: those using Clip with cross-modal temporal features (CLIP2Video [20], X-CLIP [21], X-Pool [19], and Prompt Switch [17]) and those

using Clip with coarse–fine feature learning (CLIP4Clip [22], MSIA [36], and LSECA [48]). Our method improved all indicators. Meanwhile, on the LSMDC dataset, when compared with X-CLIP [21], TS2-Net [18], X-Pool [19], Prompt Switch [17], CLIP4Clip [22], MSIA [36], LSECA [48], and TempMe [3], our method also demonstrated significant performance improvements, verifying the effectiveness and generalization ability of our proposed approach.

Table 3. Comparison of the retrieval results of state-of-the-art techniques on the MSVD dataset.

Method		Text to Video					Video to Text				
		R@1	R@5	R@10	MnR	MdR	R@1	R@5	R@10	MnR	MdR
CLIP-BASED AND CROSS-MODAL TEMPORAL FUSION	CLIP2Video [20]	47.0	76.8	85.9	9.6	-	58.7	85.6	91.6	4.3	-
	X-CLIP [21]	47.1	**77.8**	-	9.5	-	60.9	87.8	-	4.7	-
	X-Pool [19]	**47.2**	77.4	86.0	9.3	-	66.4	90.0	94.2	3.3	-
	Prompt Switch [17]	47.1	76.9	**86.1**	9.5	-	**68.5**	**91.8**	**95.6**	2.8	-
CLIP-BASED AND COARSE-TO-FINE REPRESENTATION LEARNING	CLIP4Clip [22]	45.2	75.5	84.3	10.3	2.0	**62.0**	**87.3**	**92.6**	4.3	1.0
	MSIA [36]	45.4	74.4	83.3	-	2.0	-	-	-	-	-
	LSECA [48]	**46.9**	**76.8**	**85.7**	9.9	2.0	-	-	-	-	-
CLIP-BASED AND CAUSAL INFERENCE	**Ours**	**49.4**	**77.3**	**86.0**	9.3	2.0	**68.5**	**91.2**	**95.0**	3.0	1.0

Note: For each category (e.g., "NO CLIP AND CAUSAL INFERENCE," "CLIP-BASED AND CROSS-MODAL TEMPORAL FUSION," etc.), the highest value among all evaluation metrics (R@1, R@5, R@10) for the "Text to Video" and "Video to Text" tasks within that category is selected and bolded. If there are multiple models under the same classification, the comparison is made among all models in that category.

Table 4. Comparison of the retrieval results of state-of-the-art techniques on the LSMDC dataset.

Method		Text to Video					Video to Text				
		R@1	R@5	R@10	MnR	MdR	R@1	R@5	R@10	MnR	MdR
CLIP-BASED AND CROSS-MODAL TEMPORAL FUSION	X-CLIP [21]	23.3	43.0	-	56.0	-	22.5	42.2	-	50.7	-
	TS2-Net [18]	23.4	42.3	50.1	56.9	-	-	-	-	-	-
	X-Pool [19]	**25.2**	**43.7**	**53.5**	53.2	-	**22.7**	**42.6**	**51.2**	47.4	-
	Prompt Switch [17]	23.1	41.7	50.5	56.8	-	22.0	40.8	50.3	51.0	-
CLIP-BASED AND COARSE-TO-FINE REPRESENTATION LEARNING	CLIP4Clip [22]	22.6	41.0	49.1	61.0	11.0	**20.8**	**39.0**	**48.6**	54.2	12.0
	MSIA [36]	19.7	38.1	47.5	-	12.0	-	-	-	-	-
	LSECA [48]	**23.4**	**43.1**	**50.4**	56.0	10.0	-	-	-	-	-
CLIP-BASED AND PARAMETER/INFERENCE-EFFICIENT	DGL [35]	21.4	39.4	48.4	64.3	-	-	-	-	-	-
	TempMe [3]	**23.5**	**41.7**	**51.8**	53.5	-	-	-	-	-	-
CLIP-BASED AND CAUSAL INFERENCE	**ours**	**26.0**	**43.3**	**54.3**	**53.5**	10.0	**22.9**	**42.9**	**50.7**	50.0	12.0

Note: For each category (e.g., "NO CLIP AND CAUSAL INFERENCE," "CLIP-BASED AND CROSS-MODAL TEMPORAL FUSION," etc.), the highest value among all evaluation metrics (R@1, R@5, R@10) for the "Text to Video" and "Video to Text" tasks within that category is selected and bolded. If there are multiple models under the same classification, the comparison is made among all models in that category.

4.4. Ablation Study

In this section, we conduct ablation studies to evaluate the effectiveness of the proposed strategies on the MSR-VTT dataset.

Strategy Ablation: To obtain more visual–semantic features, we adopt the strategies of causal intervention (CI) and text-conditioned video-frame aggregation. Table 5 presents the comparative results of the model with and without the two strategies. By comparing the performance of the No CI + No text-conditioned method with that of the CI method and the text-conditioned method, it can be observed that both the causal intervention strategy and the text-conditioned video-frame aggregation strategy can effectively improve the model's accuracy when applied independently. This is because causal intervention can eliminate the negative impact of confounding factors on model training, while the text-conditioned video-frame aggregation strategy can guide the model to learn more semantically informative

features. In addition, the data shows that the CI + text-conditioned method achieves the most optimal performance among all the methods. This result directly demonstrates that the combination of causal inference and text-conditioned constraints can yield a synergistic effect, which serves as an effective strategy for enhancing the performance of text-to-video retrieval. To verify that the performance improvement of the CI + text-conditioned method stems from the causal inference mechanism rather than random experimental variation, we conducted an empirical bootstrapping test on the retrieval performance metrics. Appendix A elaborates on the test design and implementation procedure and provides a detailed analysis of the results.

Table 5. Experimental results on the MSR-VTT dataset.

Method	Text to Video		
	R@1	R@5	R@10
No CI + No text-conditioned	44.7	71.2	80.0
CI	47.1	72.8	81.7
Text-conditioned	47.6	73.2	82.1
CI +text-conditioned	49.0	74.1	83.3

Alignment Ablation: To demonstrate the effectiveness of the alignment approach, Figure 5 shows a comparison results between single-level alignment and multi-level alignment. According to the figure, we can conclude that using multi-level alignment (including video–paragraph or frame–sentence alignment) is more accurate than single-level alignment. This is because single-level alignment only represents local or global features, while multi-level alignment can capture more complete local and global features.

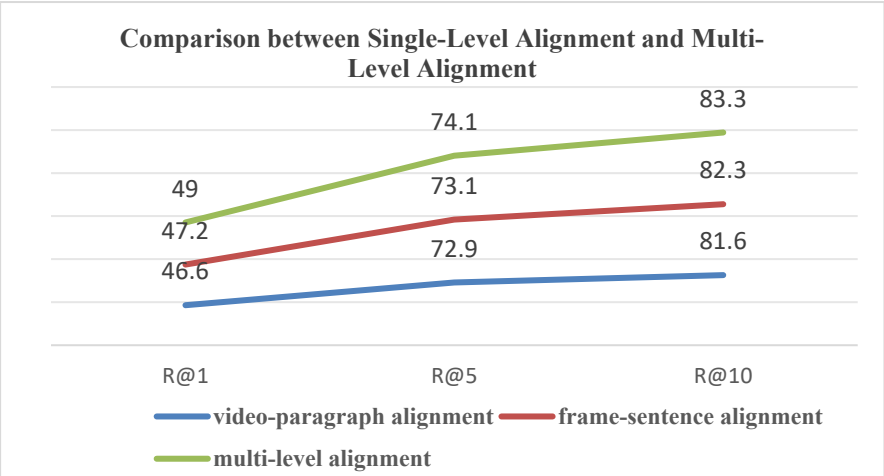


Figure 5. Comparison results between single-level alignment and multi-level alignment.

4.5. Qualitative Analyses

Figures 6 and 7 show example visualizations of retrieval results for the text-to-video and video-to-text tasks, respectively. To verify the effectiveness of causal intervention, we compared two strategies on the MSR-VTT dataset: the baseline text-conditioned approach and our proposed CI + text-conditioned method. Results show that incorporating causal intervention enables the model to more accurately capture vision–semantic causality. This leads to significant improvements not only in cross-modal retrieval performance but also in result stability, confirming the critical role of causal intervention in enhancing match quality.

Figure 6. Examples of text-to-video retrieval. The retrieval results of the text-conditioned strategy ignore the semantic information associated with the term "fighting".

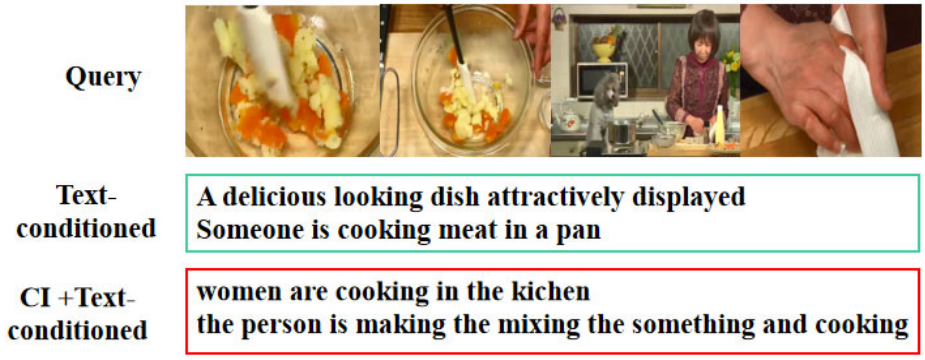


Figure 7. Examples of video-to-text retrieval.

5. Conclusions

Existing deep learning-based video-text retrieval models are typically trained to capture correlational patterns, which often fail to uncover the implicit causal relationships underlying the data. Moreover, due to biases present in visual–language datasets, these models tend to learn spurious correlations during training. To address these issues, this paper introduces causal intervention into video-text retrieval. By applying back-door adjustment to learn the causal relationships embedded in video contexts and performing text-conditioned frame aggregation, our approach not only strengthens the logical connections between video and text semantics but also accounts for both intra-modal and cross-modal relationships. Experiments on the MSR-VTT, MSVD, and LSMDC datasets show that the proposed method outperforms previous state-of-the-art approaches, improving both accuracy and generalization.

Since this paper uses the Visual Genome dataset to represent confounders, future work could explore more precise ways to characterize such confounders. Additionally, the current method only applies causal intervention to the visual feature extraction module, without modeling causal relationships between words in the text. Thus, another direction for future research is to investigate the implicit causal structures within each modality and between the two modalities.

Author Contributions: Conceptualization, H.L. and C.L.; methodology, H.L. and C.L.; software, H.L. and C.L.; validation, H.L. and C.L.; formal analysis, H.L.; investigation, H.L.; resources, H.L. and C.L.; data curation, H.L. and C.L.; writing—original draft preparation, H.L.; writing—review and editing, H.L. and C.L.; visualization, H.L.; supervision, C.L.; project administration, H.L.; funding acquisition, H.L. and C.L. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: The data used to support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest: The authors declare that they have no conflicts of interest.

Appendix A

To verify that the performance improvement achieved by the CI + text-conditioned method is, indeed, derived from the causal inference mechanism rather than random fluctuations in the experiment, the text-conditioned method was adopted as the baseline, with comparisons conducted based on the retrieval performance metrics (R@1, R@5, and R@10). For the assessment of statistical significance, the empirical bootstrapping method was employed for hypothesis testing. As a non-parametric testing approach, it is primarily utilized for evaluating significance at the test-set sample level—specifically, through resampling, with replacement performed on the test set, followed by the calculation of the sampling distribution of performance differences between the two methods across all metrics.

In the experimental setup, the number of resampling iterations was set to 10,000, and 10 independent repeated experiments were conducted for each method. Different random seeds were utilized in each iteration to control for randomness, and the significance level of the test was set to $\alpha = 0.05$. Table A1 presents the p value of each retrieval metric calculated via the empirical bootstrapping method. The results demonstrate that the p values of all three metrics are less than 0.05, thereby statistically and fully validating the effectiveness of the proposed causal inference mechanism.

Table A1. Empirical bootstrapping test results on the MSR-VTT dataset.

Metric	Text to Video (α=0.05)
	p
R@1	0.028
R@5	0.017
R@10	0.012

References

1. Omama, M.; Li, P.H.; Chinchali, S.P. Exploiting Distribution Constraints for Scalable and Efficient Image Retrieval. *arXiv* **2025**, arXiv:2410.07022.
2. Tian, K.; Cheng, Y.; Liu, Y.; Hou, X.; Chen, Q.; Li, H. Towards Efficient and Effective Text-to-Video Retrieval with Coarse-to-Fine Visual Representation Learning. *arXiv* **2024**, arXiv:2401.00701. [CrossRef]
3. Shen, L.; Hao, T.; He, T.; Zhao, S.; Zhang, Y.; Liu, P.; Bao, Y.; Ding, G. TempMe: Video Temporal Token Merging for Efficient Text-Video Retrieval. *arXiv* **2025**, arXiv:2409.01156.
4. Bai, Z.; Xiao, T.; He, T.; Wang, P.; Zhang, Z.; Brox, T.; Shou, M.Z. Bridging Information Asymmetry in Text-video Retrieval: A Data-centric Approach. *arXiv* **2025**, arXiv:2408.07249.
5. Muhammad, M.A.A.Z.; Rasheed, H.; Khan, S.; Khan, F.S. Video-ChatGPT:Towards detailed video understanding via large vision and language models. *arXiv* **2023**, arXiv:2306.05424.
6. Zhang, S.; Mu, H.; Li, Q.; Xiao, C.; Liu, T. Fine-Grained Features Alignment and Fusion for Text-Video Cross-Modal Retrieval. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024. [CrossRef]

7. Zhu, C.; Jia, Q.; Chen, W.; Guo, Y.; Liu, Y. Deep learning for video-text retrieval: A review. *Int. J. Multimed. Inf. Retr.* **2023**, *12*, 3. [CrossRef]
8. Yu, K.P.; Zhang, Z.; Hu, F.; Chai, J. Efficient in-context learning in vision-language models for egocentric videos. *arXiv* **2023**, arXiv:2311.17041.
9. Dong, J.; Li, X.; Xu, C.; Yang, X.; Yang, G.; Wang, X.; Wang, M. Dual encoding for video retrieval by text. *IEEE Trans. Pattern Anal. Mach. Intell.* **2021**, *44*, 4065–4080. [CrossRef]
10. Wen, J.; Chen, Y.; Shi, R.; Ji, W.; Yang, M.; Gao, D.; Yuan, J.; Zimmermann, R. HOVER: Hyperbolic Video-Text Retrieval. *IEEE Trans. Image Process.* **2025**, *34*, 6192–6203. [CrossRef]
11. Chen, X.; Liu, D.; Yang, X.; Li, X.; Dong, J.; Wang, M.; Wang, X. PRVR: Partially relevant video retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2025**, *48*, 1262–1277. [CrossRef]
12. Wang, X.; Zhu, L.; Yang, Y. T2vlad: Global-local sequence alignment for text-video retrieval. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 5079–5088.
13. Bain, M.; Nagrani, A.; Varol, G.; Zisserman, A. Frozen in time: A joint video and image encoder for end-to-end retrieval. In Proceedings of the IEEE/CVF international conference on computer vision, Montreal, QC, Canada, 10–17 October 2021; pp. 1728–1738.
14. Liu, S.; Fan, H.; Qian, S.; Chen, Y.; Ding, W.; Wang, Z. HiT: Hierarchical Transformer with Momentum Contrast for Video-Text Retrieval. In Proceedings of the 2021 IEEE International Conferenceon Computer Vision (ICCV), Montreal, QC, Canada, 10–17 October 2021.
15. Dzabraev, M.; Kalashnikov, M.; Komkov, S.; Petiushko, A. MDMMT: Multidomain multimodal transformer for video retrieval. *arXiv* **2021**, 3354–3363. [CrossRef]
16. Wang, J.; Wang, C.; Huang, K.; Huang, J.; Jin, L. VideoCLIP-XL: Advancing Long Description Understanding for Video CLIP Models. *arXiv* **2024**, arXiv:2410.00741.
17. Deng, C.; Chen, Q.; Qin, P.; Chen, D.; Wu, Q. Prompt Switch: Efficient CLIP Adaptation for Text-Video Retrieval. *arXiv* **2023**. [CrossRef]
18. Liu, Y.; Xiong, P.; Xu, L.; Cao, S.; Jin, Q. Ts2-net: Token shift and selection transformer for text-video retrieval. In Proceedings of the European Conference on Computer Vision, Tel Aviv, Israel, 23–27 October 2022; pp. 319–335.
19. Fang, C.; Zhang, D.; Wang, L.; Zhang, Y.; Cheng, L.; Han, J. X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval. In Proceedings of the Computer Vision and Pattern Recognition (CVPR), Lisboa, Portugal, 10–14 October 2022; pp. 5006–5015. [CrossRef]
20. Fang, H.; Xiong, P.; Xu, L.; Chen, Y. CLIP2Video: Mastering Video-Text Retrieval via Image CLIP. *arXiv* **2021**. [CrossRef]
21. Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; Ji, R. X-clip: Endto-end multi-grained contrastive learning for video-text retrieval. *arXiv* **2022**, arXiv:2207.07285.
22. Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; Li, T. CLIP4Clip: An Empirical Study of CLIP for End to End Video ClipRetrieval. *arXiv* **2021**. [CrossRef]
23. Judea, P. *Causality: Models, Reasoning and Inference*; Cambridge University Press: Cambridge, UK, 2009; Volume 40, p. 478.
24. Park, J.J.; Choi, S.J. Bridging Vision and Language: Modeling Causality and Temporality in Video Narratives. *arXiv* **2024**, arXiv:2412.10720. [CrossRef]
25. Li, Y.; Li, R.; Ma, Y.; Xue, Y.; Meng, L. FCA: A Causal Inference Based Method for Analyzing the Failure Causes of Object Detection Algorithms. In Proceedings of the 2023 IEEE 23rd International Conference on Software Quality, Reliability, and Security Companion (QRS-C), Chiang Mai, Thailand, 22–26 October 2023; pp. 2693–9371.
26. Chen, Z.; Hu, L.; Li, W.; Shao, Y.; Nie, L. Causal Intervention and Counterfactual Reasoning for Multi-modal Fake News Detection. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics, Toronto, ON, Canada, 9–14 July 2023; pp. 627–638.
27. Xue, D.; Qian, S.; Xu, C. Variational Causal Inference Network for Explanatory Visual Question Answering. In Proceedings of the 2023 IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 1–6 October 2023. [CrossRef]
28. Cai, L.; Fang, H.; Xu, N.; Ren, B. Counterfactual Causal-Effect Intervention for Interpretable Medical Visual Question Answering. *IEEE Trans. Med. Imaging* **2024**, *43*, 4430–4441. [CrossRef]
29. Liu, Y.; Qin, G.; Chen, H.; Cheng, Z.; Yang, X. Causality-Inspired Invariant Representation Learning for Text-Based Person Retrieval. *Proc. AAAI Conf. Artif. Intell.* **2024**, *38*, 14052–14060. [CrossRef]
30. BSatar, B.; Zhu, H.; Zhang, H.; Lim, J.H. Towards Debiasing Frame Length Bias in Text-Video Retrieval via Causal Intervention. *arXiv* **2023**, arXiv:2309.09311. [CrossRef]
31. Cheng, D.; Kong, S.; Wang, W.; Qu, M.; Jiang, B. Long Term Memory-Enhanced Via Causal Reasoning for Text-To-Video Retrieval. In Proceedings of the 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Republic of Korea, 14–19 April 2024; pp. 1520–6149.

32. Yue, W.U.; Qi, Z.; Wu, Y.; Sun, J.; Wang, Y.; Wang, S. Learning Fine-Grained Representations through Textual Token Disentanglement. 2025. Available online: https://openreview.net/forum?id=wGa2plE8ka (accessed on 27 May 2025).
33. Zou, X.; Wu, C.; Cheng, L.; Wang, Z. TokenFlow: Rethinking Fine-grained Cross-modal Alignment in Vision-Language Retrieval. *arXiv* **2022**, arXiv:2209.13822.
34. Wang, Z.; Sung, Y.L.; Cheng, F.; Bertasius, G.; Bansal, M. Unified Coarse-to-Fine Alignment for Video-Text Retrieval. In Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), Paris, France, 2–3 October 2023; Volume 34, pp. 1747–1756. Available online: https://openaccess.thecvf.com/content/ICCV2023/html/Wang_Unified_Coarse-to-Fine_Alignment_for_Video-Text_Retrieval_ICCV_2023_paper.html (accessed on 25 January 2026).
35. Yang, X.; Zhu, L.; Wang, X.; Yang, Y. DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval. *arXiv* **2024**, arXiv:2401.10588. [CrossRef]
36. Chen, L.; Deng, Z.; Liu, L.; Yin, S. Multilevel semantic interaction alignment for video-text cross-modal retrieval. *IEEE Trans. Circuits Syst. Video Technol.* **2024**, *34*, 6559–6575. [CrossRef]
37. Chen, W.; Liu, Y.; Wang, C.; Zhu, J.; Zhao, S.; Li, G.; Liu, C.L.; Lin, L. Cross-Modal Causal Intervention for Medical Report Generation. *arXiv* **2024**, arXiv:2303.09117. [CrossRef]
38. Rehman, M.U.; Nizami, I.F.; Ullah, F.; Hussain, I. IQA Vision Transformed: A Survey of Transformer Architectures in Perceptual Image Quality Assessment. *IEEE Access* **2024**, *12*, 83369–183393. [CrossRef]
39. Li, W.; Li, Z.; Yang, X.; Ma, H. Causal-ViT: Robust Vision Transformer by causal intervention. *Eng. Appl. Artif. Intell.* **2023**, *126*, 107123. [CrossRef]
40. Li, Z.; Wang, H.; Liu, D.; Zhang, C.; Ma, A.; Long, J.; Cai, W. Multimodal Causal Reasoning Benchmark: Challenging Vision Large Language Models to Infer Causal Links Between Siamese Images. Computer Vision and Pattern Recognition. *arXiv* **2024**, arXiv:2408.08105.
41. Chen, C.; Merullo, J.; Eickhoff, C. Axiomatic Causal Interventions for Reverse Engineering Relevance Computation in Neural Retrieval Models. In Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, Washington, DC, USA, 14–18 July 2024. [CrossRef]
42. Liu, J.; Liu, Z.; Zhang, Z.; Wang, L.; Liu, M. A New Causal Inference Framework for SAR Target Recognition. *IEEE Trans. Artif. Intell.* **2024**, *5*, 4042–4057. [CrossRef]
43. Wu, X.; Guo, R.; Li, Q.; Zhu, N. Visual Commonsense Causal Reasoning from A Still Image. *IEEE Access* **2025**, *13*, 85084–85097. [CrossRef]
44. Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.J.; Shamma, D.A.; et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *arXiv* **2017**, arXiv:1602.07332. [CrossRef]
45. Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A large video description dataset for bridging video and language. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 27–30 June 2016; pp. 5288–5296. [CrossRef]
46. Chen, D.; Dolan, W.B. Collecting highly parallel data for paraphrase evaluation. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Portland, OR, USA, 19–24 June 2011; pp. 190–200.
47. Rohrbach, A.; Rohrbach, M.; Tandon, N.; Schiele, B. A dataset for movie description. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3202–3212.
48. Wang, Z.; Zhang, D.; Hu, Z. LSECA: Local semantic enhancement and cross aggregation for video-text retrieval. *Int. J. Multimed. Inf. Retr.* **2024**, *13*, 30. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.