# BiMa: Towards Biases Mitigation for Text-Video Retrieval via Scene Element Guidance

Huy Le
huylda1@fpt.com
FPT Software AI Center
Hanoi, Vietnam

Nhat Chung
nhatcm3@fpt.com
FPT Software AI Center
Hanoi, Vietnam

Tung Kieu
tungkvt@cs.aau.dk
Aalborg University
Aalborg, Denmark
Pioneer Centre for AI
Copenhagen, Denmark

Anh Nguyen
anh.nguyen@liverpool.ac.uk
University of Liverpool
Liverpool, UK

Ngan Le
thile@uark.edu
University of Arkansas
Fayetteville, USA

## Abstract

Text-video retrieval (TVR) systems often suffer from visual-linguistic biases present in datasets, which cause pre-trained vision-language models to overlook key details. To address this, we propose `BiMa`, a novel framework designed to mitigate biases in both visual and textual representations. Our approach begins by generating scene elements that characterize each video by identifying relevant entities/objects and activities. For visual debiasing, we integrate these scene elements into the video embeddings, enhancing them to emphasize fine-grained and salient details. For textual debiasing, we introduce a mechanism to disentangle text features into content and bias components, enabling the model to focus on meaningful content while separately handling biased information. Extensive experiments and ablation studies across five major TVR benchmarks (*i.e.*, MSR-VTT, MSVD, LSMDC, ActivityNet, and DiDeMo) demonstrate the competitive performance of `BiMa`. Additionally, the model's bias mitigation capability is consistently validated by its strong results on out-of-distribution retrieval tasks. The code is available at: https://github.com/Fsoft-AIC/BiMa

## CCS Concepts

• **Computing methodologies** → *Matching*; *Scene understanding*;
• **Information systems** → *Similarity measures*.

## Keywords

Multi-modal Learning, Text-Video Retrieval, Bias Mitigation

**ACM Reference Format:**
Huy Le, Nhat Chung, Tung Kieu, Anh Nguyen, and Ngan Le. 2025. BiMa: Towards Biases Mitigation for Text-Video Retrieval via Scene Element Guidance. In *Proceedings of the 33rd ACM International Conference on Multimedia (MM '25), October 27–31, 2025, Dublin, Ireland*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3746027.3754833


*MM '25, Dublin, Ireland*

ACM ISBN 979-8-4007-2035-2/2025/10
https://doi.org/10.1145/3746027.3754833

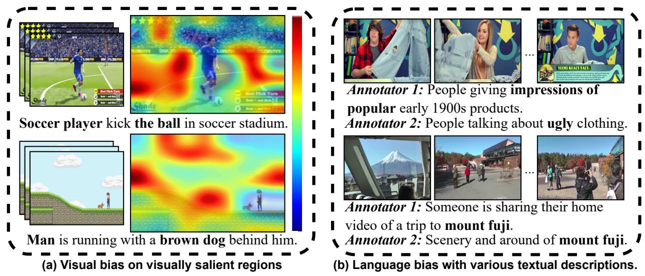


**Figure 1: Illustration of Visual-Linguistic Bias in TVR. (a) Visual Bias: For each example, video (left of the 1st row), the associated visual feature map (right of the 1st row) extracted by pre-trained `CLIP` [30] and its corresponding Textual query (2nd row). This highlights a bias toward visually dominant regions while overlooking main actors/objects and activities when they occupy smaller regions in the scene. (b) Textual Bias: For each example, video (1st row) and their corresponding textual descriptions (2nd row) from various annotators, capturing emotional responses and personal perspectives.**

## 1 Introduction

The task of retrieving videos based on textual queries and vice versa, known as text-video retrieval (TVR), has rapidly evolved within multimedia information retrieval due to significant advancements in large-scale pre-trained Vision-Language Models (VLMs) such as `CLIP` [40] and `BLIP` [29]. Despite remarkable progress, existing TVR frameworks, including `TeachCLIP` [47], `TextProxy` [54], and `NarVid` [15], largely ignore the underlying visual-linguistic representation biases intrinsic to both the training data and pre-trained VLMs [33, 43, 44]. Representation biases are systematic deviations in datasets causing models to overemphasize specific features or patterns rather than generalizable and task-relevant aspects [36, 44]. *Liu et al.* [36] demonstrate that models tend to depend excessively on prominent visual concepts and dataset-specific textual patterns, resulting in representations that are tuned to the dataset rather than capturing robust, semantic-rich features. This bias limits the models' ability to generalize to diverse, unseen scenarios. Moreover, the precise nature of the biases learned by neural networks remains largely unclear—some may even contain generalizable and transferable patterns that are not immediately apparent to human observers. *Shvetsova et al.* [44] also confirm that most video datasets are heavily focused on visually salient concepts, such as object bias. Thus, The absence of bias mitigation within the existing learned TVR frameworks can lead to limitations, resulting in suboptimal performance when exposed to unseen data. We proceed to further describe the bias problem in TVR.

**Visual-Linguistic Biases.** Widely-use TVR datasets such as MSR-VTT [55], MSVD [3], LSMDC [41], ActivityNet [24], and DiDeMo [12] often present practical challenges stemming from what we identify as visual-linguistic biases. These biases arise because each dataset is typically created with specific objectives that align with particular research goals, applications, or target users. These biases are variations in visual and textual representations that can skew pre-trained models towards focusing on subjective or dataset-specific features, potentially causing them to overlook essential factual information. The visual-linguistic bias problem can be partitioned into two sub-problems: visual bias and textual bias, as follows.

Visual bias primarily arises due to coarse-grained annotations that often omit critical details such as key actors, objects, or their interactions. Consequently, visual embeddings produced by pre-trained encoders frequently emphasize visually dominant areas, neglecting smaller but semantically crucial components. Figure 1(a) illustrates this issue, showing how pre-trained CLIP disproportionately focuses on the larger environmental context, marginalizing important yet smaller elements such as the "man" and "brown dog," crucial to understanding the scene.

Textual bias occurs when annotators' subjective interpretations, cultural perspectives, emotional states, or language usage differences produce varying textual descriptions for identical video scenes. This variability causes models to capture subjective rather than objective semantic content. As depicted in Figure 1(b), distinct annotators produce emotionally and culturally varied descriptions, demonstrating the need for effective textual bias neutralization.

To address the aforementioned visual-linguistic bias problem, we propose Bias Mitigation Text-Video Retrieval (BiMa). The BiMa's objectives are twofold: (i) neutralize visual-linguistic biases and (ii) enhance the model's focus on relevant features. To obtain these goals, BiMa is equipped with three key modules–(i) *Scene Element Construction*, (ii) *Visual Scene Debias*, and (iii) *Textual Content Debias*. The *Scene Element Construction* module aggregates fine-grained entities and actions, providing structured semantic guidance. Subsequently, the *Visual Scene Debias* module leverages these elements to reorient visual attention toward essential components, thus reducing visual bias. The *Textual Content Debias* module employs a novel disentanglement mechanism to separate textual content from bias-induced variations, ensuring semantic consistency across diverse annotations. Critically, this mechanism is self-supervised, obviating the need for expensive additional annotations. Finally, to further evaluate our proposed bias mitigation framework, out-of-domain retrieval experiments are leveraged to validate the efficacy and assess the generalization of models beyond training-specific biases by evaluating performance across distinctly different datasets or scenarios, such as training on MSR-VTT and testing on ActivityNet Captions. In summary, our contributions are:

- Explicitly identifying and defining visual-linguistic biases in TVR task, drawing upon recent bias representation findings [36, 44].
- Proposing BiMa, a systematic framework for bias mitigation with three integrated modules: *Scene Element Construction*, *Visual Scene Debias*, and *Textual Content Debias*.
- Achieving state-of-the-art performance and robust generalization across multiple TVR benchmarks (MSR-VTT, MSVD, LSMDC, ActivityNet Captions, DiDeMo), supported by comprehensive ablation studies and significant bias reduction in out-of-domain retrieval settings.

## 2 Problem Definition

**Text-Video Retrieval.** Given a text query $\mathbf{T}^{(i)} = \langle \mathbf{T}_1^{(i)}, \mathbf{T}_2^{(i)}, \ldots, \mathbf{T}_{N_t}^{(i)} \rangle$ with $N_t$ word tokens and a video $\mathbf{V}^{(i)} = \langle \mathbf{V}_1^{(i)}, \mathbf{V}_2^{(i)}, \ldots, \mathbf{V}_{N_f}^{(i)} \rangle$ with $N_f$ frames. A Text Encoder encodes $\mathbf{T}^{(i)}$ into a sequence of $N_t + 1$ word embeddings $\mathbf{t}^{(i)} = \langle \mathbf{t}_1^{(i)}, \mathbf{t}_2^{(i)}, \ldots, \mathbf{t}_{N_t}^{(i)}, \mathbf{t}_{\texttt{[CLS]}}^{(i)} \rangle$, where $\mathbf{t}_{\texttt{[CLS]}}^{(i)}$ is the global textual embedding. A Video Encoder encodes $\mathbf{V}^{(i)}$ into a sequence of $N_f$ frame embedding $\mathbf{v}^{(i)} = \langle \mathbf{v}_1^{(i)}, \mathbf{v}_2^{(i)}, \ldots, \mathbf{v}_{N_f}^{(i)} \rangle$. We aim to learn a cross-modality similarity measure that assigns a high similarity score to $\mathbf{V}^{(i)}$ and $\mathbf{T}^{(i)}$ and a low similarity score to $\mathbf{V}^{(j)}$ (with $j \neq i$) and $\mathbf{T}^{(i)}$. Equation 1 formally defines the TVR problem. For brevity, we only focus on the Text-to-Video (**T2V**). The Video-to-Text (**V2T**) can be similarly defined.

$$\max s(\mathbf{V}^{(i)}, \mathbf{T}^{(i)}) = \max \mathbb{P}(\mathbf{V}^{(i)} | \mathbf{T}^{(i)}) = \max \mathbb{P}(\mathbf{v}^{(i)} | \mathbf{t}^{(i)}) \quad (1)$$

Here, $s(\cdot, \cdot)$ is a *cosine* similarity and $\mathbb{P}(\cdot|\cdot)$ indicates the conditional probability.

**Visual-Linguistic Biases.** When *visual bias* occurs, the visual embedding features of a video $\mathbf{V}^{(i)}$ differ from the matching visual embedding features generated for a corresponding query $\mathbf{T}^{(i)}$. We define the visual bias as follows.

$$\mathbb{P}(\mathbf{v}^{(i)} | \mathbf{V}^{(i)}) \neq \mathbb{P}(\mathbf{v}^{(i)} | \mathbf{T}^{(i)}) \Rightarrow \mathbb{P}(\mathbf{v}^{(i)} | \mathbf{V}^{(i)}) \neq \mathbb{P}(\mathbf{v}^{(i)} | \mathbf{t}^{(i)}) \quad (2)$$

Equation 2 indicates that the Video Encoder does not provide the matching visual embedding features. When a pre-trained model is used as the Video Encoder, the bias can lean towards the pre-trained features.

When *textual bias* occurs, the probability of the visual embedding features given the textual biases ($\tilde{\mathbf{t}}^{(i)}$) is higher than the visual embedding given the true semantic textual features ($\hat{\mathbf{t}}^{(i)}$). We define the textual bias as follows.

$$\mathbb{P}(\mathbf{v}^{(i)} | \hat{\mathbf{t}}^{(i)} + \tilde{\mathbf{t}}^{(i)}) > \mathbb{P}(\mathbf{v}^{(i)} | \hat{\mathbf{t}}^{(i)}) \Rightarrow \mathbb{P}(\mathbf{v}^{(i)} | \mathbf{t}^{(i)}) \neq \mathbb{P}(\mathbf{v}^{(i)} | \hat{\mathbf{t}}^{(i)}) \quad (3)$$

Equation 3 indicates that the similarity between the visual embedding features and the textual biases is higher than the similarity between the visual embedding features and the true semantic textual features.

## 3 Methodology

Figure 2 shows the the framework overview. comprising of three main modules: (i) *Scene Element Construction*, (ii) *Visual Scene Debias*, and (iii) *Textual Content Debias*. We proceed to introduce these modules in the following sections.

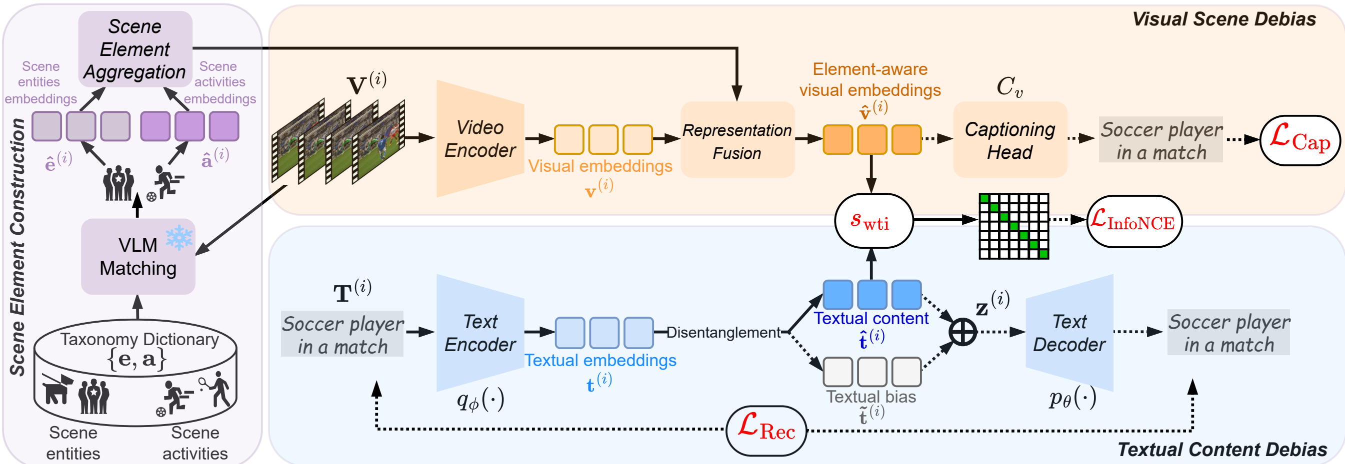


**Figure 2: Overview of BiMa. The dashed line represents the model flow used exclusively during training, while the solid line indicates usage in both training and inference.**

## 3.1 Scene Element Construction

*Scene Element Construction* aims to generate scene elements for each video and the corresponding embeddings. In our proposed framework, scene elements play a central role in debiasing across different modalities. Scene elements consist of scene entities (noun phrases representing actors or objects) and scene activities (verb phrases representing behaviors or actions), which are drawn from a comprehensive scene taxonomy dictionary (detailed implementation is presented in the **Supplementary Material**). Formally, scene elements are represented as a set of nouns and verb phrases $\{\mathbf{e}, \mathbf{a}\}$ that is extracted from large-scale text-video descriptions.

**Scene Elements Generation.** Given a video $\mathbf{V}^{(i)}$, we leverage the zero-shot capability of a pre-trained CLIP model with frozen weights to perform text-video matching to identify the most relevant scene elements. In particular, we use CLIP's textual encoder to obtain global textual tokens including a set of $N_e$ scene entities $\mathbf{e} = \{\mathbf{e}^{(i)}\}_{i=1}^{N_e}$, and a set of $N_a$ scene activities $\mathbf{a} = \{\mathbf{a}^{(i)}\}_{i=1}^{N_a}$. Simultaneously, we feed $\mathbf{V}^{(i)}$ into CLIP's visual encoder to obtain frame-level visual feature embeddings $\mathbf{v}^{(i)}$. The embeddings $\mathbf{v}^{(i)}$ are then temporally aggregated through mean pooling to yield a video-level representation $\bar{\mathbf{v}}^{(i)}$. Then, we calculate the similarity between $\bar{\mathbf{v}}^{(i)}$ with scene entities dictionary $\mathbf{e}$ and scene activities dictionary $\mathbf{a}$. After that, we select the top-$\kappa$ relevant scene entities $\hat{\mathbf{e}}^{(i)} = \langle \hat{\mathbf{e}}_1^{(i)}, \hat{\mathbf{e}}_2^{(i)}, \ldots, \hat{\mathbf{e}}_\kappa^{(i)} \rangle$ and top-$\kappa$ scene activities $\hat{\mathbf{a}}^{(i)} = \langle \hat{\mathbf{a}}_1^{(i)}, \hat{\mathbf{a}}_2^{(i)}, \ldots, \hat{\mathbf{a}}_\kappa^{(i)} \rangle$ with the highest similarity scores as "scene elements" of $\mathbf{V}^{(i)}$.

**Scene Elements Aggregation.** To enhance the discriminative features corresponding to $\mathbf{V}^{(i)}$, we combine scene entities and scene activities into a unified scene element embedding $\mathbf{c}^{(i)}$. Specifically, we introduce a balancing coefficient $g$ that adjusts the relative importance of scene entities $\hat{\mathbf{e}}^{(i)}$ and scene activities $\hat{\mathbf{a}}^{(i)}$ for each video $\mathbf{V}^{(i)}$. This coefficient $g$ supports a smooth aggregation of embeddings by establishing an association between each feature token and its most relevant counterpart among the other feature tokens. The formulation of $g$ is provided in Equation 4.

$$g_{(\hat{\mathbf{a}}^{(i)}, \hat{\mathbf{e}}^{(i)})} = 1 - \frac{1}{\kappa} \sum_{l}^{\kappa} \max_{j} [s(\hat{\mathbf{a}}_l^{(i)}, \hat{\mathbf{e}}_j^{(i)})_{j=1}^{\kappa}] \tag{4}$$

Here, $s(\cdot)$ indicates the *cosine* similarity. The coefficient $g$ is then used to aggregate $\hat{\mathbf{e}}^{(i)}$ and $\hat{\mathbf{a}}^{(i)}$ to obtain scene element embeddings $\mathbf{c}^{(i)}$ as follows.

$$\mathbf{c}^{(i)} = \hat{\mathbf{a}}^{(i)} \oplus g_{(\hat{\mathbf{a}}^{(i)}, \hat{\mathbf{e}}^{(i)})} \cdot \hat{\mathbf{e}}^{(i)} \tag{5}$$

Here, $\oplus$ indicates the element-wise summation operation.

## 3.2 Visual Scene Debias

*Visual Scene Debias* module aims to mitigate visual biases by explicitly emphasizing fine-grained semantic contents in the visual representation. Particularly, it leverages scene elements to produce element-aware visual scene embeddings. Formally, we aim to mitigate the problem $\mathbb{P}(\mathbf{v}^{(i)}|\mathbf{V}^{(i)}) \neq \mathbb{P}(\mathbf{v}^{(i)}|\mathbf{t}^{(i)})$ (see Equation 2) by augmenting the visual embedding $\mathbf{v}^{(i)}$ with the scene element $\mathbf{c}^{(i)}$ to obtain the augmented visual feature $\hat{\mathbf{v}}^{(i)}$. Then, $\hat{\mathbf{v}}^{(i)}$ is employed as the feature for matching such that $\mathbb{P}(\hat{\mathbf{v}}^{(i)}|\mathbf{V}^{(i)}) \approx \mathbb{P}(\hat{\mathbf{v}}^{(i)}|\mathbf{t}^{(i)})$. We elaborate on this in the following sections.

**Cross-modality Debias via Representation Fusion.** Debiasing visual representations involves generating element-aware visual scene embeddings that emphasize relevant information at a lower-dimensional latent space, while capturing interactions between cross-modal features. More specifically, we aim to align the video embeddings $\mathbf{v}^{(i)}$ and scene element embeddings $\mathbf{c}^{(i)}$ to highlight the relevant information $\hat{\mathbf{v}}^{(i)}$ of the video $\mathbf{V}^{(i)}$. First, we compute an attention map between a query $\mathbf{q}_v$, a key $\mathbf{k}_c$, and a value $\mathbf{v}_c$. Here, $\mathbf{q}_v$ represents video embeddings $\mathbf{v}^{(i)}$, and both $\mathbf{k}_c$ and $\mathbf{v}_c$ represent scene element embeddings $\mathbf{c}^{(i)}$. Then, element-aware visual scene features $\hat{\mathbf{v}}^{(i)}$ are computed through a cross-attention layer [50] as shown in Equation 6.

$$\mathbf{q}_v = \text{Linear}(\mathbf{v}^{(i)}), \mathbf{k}_c = \text{Linear}(\mathbf{c}^{(i)}), \mathbf{v}_c = \text{Linear}(\mathbf{c}^{(i)}),$$
$$\hat{\mathbf{v}}^{(i)} = \text{softmax}\left(\frac{\mathbf{q}_v \mathbf{k}_c^\top}{\sqrt{d}}\right) \mathbf{v}_c.$$

**Fine-grained Semantic Learning with Captioning Head.** Inspired by `CoCa` [59] that leverages captioning for salient semantic pretraining, we propose to additionally adopt a *Captioning Head* $C_v(\cdot)$ parameterized by a stack of $M$ transformer decoders to further learn cross-modal details within element-aware visual scene embeddings $\hat{\mathbf{v}}^{(i)}$. In other words, on top of contrastive matching with textual embeddings, the element-aware visual scene embeddings are learned via a text-generative decoding process. At each decoding step, $C_v(\cdot)$ is trained to predict the subsequent text token $\mathbf{T}_l^{(i)}$ with the highest log-likelihood. Thus, it learns to auto-regressively maximize the log-likelihood of predicting the input's textual token $\mathbf{T}_l^{(i)}$ based on previous tokens $\mathbf{T}_{0:l-1}^{(i)}$. As $C_v(\cdot)$ relies on $\hat{\mathbf{v}}^{(i)}$, the embeddings $\hat{\mathbf{v}}^{(i)}$ can be established at a deeper alignment of cross-modal representation via Equation 6.

$$\mathcal{L}_{\text{Cap}} = \sum_{l=1}^{N_t} -\log C_v(\mathbf{T}_l^{(i)} | \mathbf{T}_{0:l-1}^{(i)}, \hat{\mathbf{v}}^{(i)}). \tag{6}$$

## 3.3 Textual Content Debias

*Textual Content Debias* aims to detach the textual bias from the description and encourage the model to capture the true semantic features via a disentanglement process. Formally, we aim to solve the problem $\mathbb{P}(\mathbf{v}^{(i)}|\mathbf{t}^{(i)}) \neq \mathbb{P}(\mathbf{v}^{(i)}|\hat{\mathbf{t}}^{(i)})$ ( Equation 3) by repurposing the *Text Encoder* to decompose every textual description $\mathbf{T}^{(i)}$ into a true semantic component $\hat{\mathbf{t}}^{(i)}$ and a textual bias component $\tilde{\mathbf{t}}^{(i)}$. Then, $\hat{\mathbf{t}}^{(i)}$ is employed as the textual feature for matching with the augmented element-aware visual scene features $\hat{\mathbf{v}}^{(i)}$ which is obtained from the *Visual Scene Debias*, such that $\mathbb{P}(\mathbf{v}^{(i)}|\mathbf{t}^{(i)}) \approx \mathbb{P}(\hat{\mathbf{v}}^{(i)}|\hat{\mathbf{t}}^{(i)})$. We elaborate on this in the below sections.

**Content-Bias Representation Disentanglement.** Motivated by the disentanglement ability of $\beta$-`VAE` framework [13], we propose to use *Text Encoder* to decompose the original textual content $\mathbf{T}^{(i)}$ into two parts in the latent space: a textual content component $\hat{\mathbf{t}}^{(i)}$ and the textual bias component $\tilde{\mathbf{t}}^{(i)}$. These two parts are learned under completely different constraints, thereby separating their representations. The former is aligned with the element-aware visual scene features $\hat{\mathbf{v}}^{(i)}$ to enable content matching. The latter is modeled as a stochastic representation following a Gaussian distribution with the mean (*mu*) and variance ($\sigma^2$). By doing this, we aim to handle biases as Gaussian noise. Combining both components will result in the latent representation $\mathbf{z}^{(i)} = \hat{\mathbf{t}}^{(i)} + \tilde{\mathbf{t}}^{(i)}$. Then, the obtained latent representation $\mathbf{z}^{(i)}$ is employed to reconstruct the original textual description $\mathbf{T}^{(i)}$. By reconstructing the two separate components, we ensure the complementary property of each component. In other words, we ensure the combination resulting in $\mathbf{z}^{(i)}$ is meaningful. The reconstruction closely follows the ELBO formulation from conventional `VAE` framework [23], which optimizes the variational lower bound on data log-likelihood as shown in Equation 7.

$$\begin{aligned}\mathcal{L}_{\text{VAE}} = & \underbrace{-\mathbb{E}_{\mathbf{z}^{(i)} \sim q_\phi(\mathbf{z}^{(i)}|\mathbf{T}^{(i)})}\left[\log p_\theta(\mathbf{T}^{(i)}|\mathbf{z}^{(i)})\right]}_{\mathcal{L}_{\text{Rec}}} \\ & + \underbrace{D_{\mathbb{KL}}\left(q_\phi(\mathbf{z}^{(i)}|\mathbf{T}^{(i)}) || p_\theta(\mathbf{z}^{(i)})\right)}_{\mathcal{L}_{\mathbb{KL}}}\end{aligned} \tag{7}$$

Here, $p_\theta(\mathbf{z}^{(i)})$ is a prior modeled as $\mathcal{N}(\mathbf{0}, \mathbf{I})$. The conditional probability distributions $q_\phi(\mathbf{z}^{(i)}|\mathbf{T}^{(i)}), p_\theta(\mathbf{T}^{(i)}|\mathbf{z}^{(i)})$ are estimated by the *Text Encoder* and *Text Decoder* networks, respectively. Here, $\mathbf{z}^{(i)}$ plays the role of textual embedding, i.e., $\mathbf{z}^{(i)} = \mathbf{t}^{(i)}$. Next, $D_{\mathbb{KL}}(\cdot, \cdot)$ is the Kullback-Leibler (KL) divergence aiming to enforce the posterior and the prior distribution to be close to each other. The KL divergence plays as a disentanglement factor to enforce the disentanglement process of the bias from the original textual features. Also, it serves as a regularizer that prevents the latent representation from collapsing to zero. The reconstruction term $\mathcal{L}_{\text{Rec}}$ in the `VAE` loss measures the log-likelihood of *Text Decoder* $p_\theta$ to auto-regressively reconstruct the ground-truth input's textual description $\mathbf{T}^{(i)}$ as defined in Equation 8.

$$\mathcal{L}_{\text{Rec}} = \sum_{l=1}^{N_t} -\log p_\theta(\mathbf{T}_l^{(i)} | \mathbf{T}_{0:l-1}^{(i)}, \mathbf{z}^{(i)}). \tag{8}$$

To capture the unpredictable variations of the textual biases, we model $\tilde{\mathbf{t}}^{(i)}$ as a probabilistic representation. While these features are not explicitly matched to visual representations, they act as distractors for TVR and thus must be modeled to facilitate effective disentanglement. Following this insight, we model the latent representation $\tilde{\mathbf{t}}^{(i)}$ as a multivariate Gaussian distribution.

$$q_\phi(\tilde{\mathbf{t}}^{(i)}|\mathbf{T}^{(i)}) \sim \mathcal{N}(\mu, \text{diag}(\sigma^2)) \tag{9}$$

To ensure stable training, we apply the reparameterization trick as formulated in Equation 10.

$$q_\phi(\tilde{\mathbf{t}}^{(i)}|\mathbf{T}^{(i)}) = \mu + \sigma \cdot \epsilon, \tag{10}$$

where $\epsilon$ is an auxiliary noise variables and $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

After modeling the latent representation of textual biases as a distribution $q_\phi(\tilde{\mathbf{t}}^{(i)}|\mathbf{T}^{(i)})$, we generate textual bias embedding $\tilde{\mathbf{t}}^{(i)}$ by sampling $K$ instances from this distribution during the training process as shown in Equation 11. This process allows for gradient propagation through the sampled embeddings of bias modeling.

$$\tilde{\mathbf{t}}^{(i)} = \langle \tilde{\mathbf{t}}_1^{(i)}, \cdots, \tilde{\mathbf{t}}_2^{(i)}, \tilde{\mathbf{t}}_K^{(i)} \rangle \sim q_\phi(\tilde{\mathbf{t}}^{(i)}|\mathbf{T}^{(i)}) \tag{11}$$

**Element-Aware Content Learning by Alignment.** Our framework performs disentanglement between content and bias representations by explicitly matching the content embeddings with element-aware visual embeddings, and performing reconstruction of textual disentanglement. In particular, our model is trained via contrastive alignment between element-aware visual embeddings $\hat{\mathbf{v}}^{(i)}$ and textual content embeddings $\hat{\mathbf{t}}^{(i)}$ to optimize their cross-modal similarity (see Equation 12). Meanwhile, the original texts $\mathbf{t}^{(i)}$ are disentangled as content $\hat{\mathbf{t}}^{(i)}$ and bias $\tilde{\mathbf{t}}^{(i)}$, and reconstructed via $\mathbf{z}^{(i)}$ and $\mathcal{L}_{\text{VAE}}$. Thus, by the process of elimination, we capture bias representations $\tilde{\mathbf{t}}^{(i)}$ as residual information that is less relevant to element-aware visual embeddings $\hat{\mathbf{v}}^{(i)}$, yet are part of the original information $\mathbf{t}^{(i)}$.

## 3.4 Objective Function

To enhance the model's ability to discern relevant features within the data, we replace mean pooling with an advanced weighted pooling mechanism termed Weighted Token Interaction (WTI) [34] in order to learn the token-level interaction between cross-modal embeddings. The token-level matching function is denoted as $s_{\text{wti}}(\cdot)$.

For contrastive learning, we adopt the InfoNCE loss function [48] to optimize cross-modal similarity between $\hat{\mathbf{v}}^{(i)}$ and $\hat{\mathbf{t}}^{(j)}$ of two sample $i$ and $j$ as shown in Equation 12.

$$\begin{aligned}\mathcal{L}_{\text{InfoNCE}} = &-\frac{1}{2B}\sum_{i=1}^{B}\log\frac{\exp(s_{\text{wti}}(\hat{\mathbf{v}}^{(i)},\hat{\mathbf{t}}^{(i)})\cdot\tau)}{\sum_{j=1}^{B}\exp(s_{\text{wti}}(\hat{\mathbf{v}}^{(j)},\hat{\mathbf{t}}^{(i)})\cdot\tau)} \\ &-\frac{1}{2B}\sum_{i=1}^{B}\log\frac{\exp(s_{\text{wti}}(\hat{\mathbf{v}}^{(i)},\hat{\mathbf{t}}^{(i)})\cdot\tau)}{\sum_{j=1}^{B}\exp(s_{\text{wti}}(\hat{\mathbf{v}}^{(i)},\hat{\mathbf{t}}^{(j)})\cdot\tau)},\end{aligned} \tag{12}$$

where $\tau$ is a learnable scaling factor and $B$ is the batch size.

Our process to mitigate visual-linguistic biases exhibits during both training and inference by our end-to-end architecture. However, bias modeling is not required during the inference. The overall objective is defined in Equation 13.

$$\mathcal{L} = \mathcal{L}_{\text{InfoNCE}} + \lambda_{\text{Cap}}\cdot\mathcal{L}_{\text{Cap}} + \lambda_{\text{Rec}}\cdot\mathcal{L}_{\text{Rec}} + \lambda_{\mathbb{KL}}\cdot\mathcal{L}_{\mathbb{KL}} \tag{13}$$

Here, $\lambda_{\texttt{cap}}$, $\lambda_{\texttt{rec}}$, and $\lambda_{\mathbb{KL}}$ are the hyperparameters that control the trade-off among three loss terms.

# 4 Experiments

## 4.1 Experimental Setups.

**Datasets.** We conduct experiments on five major TVR benchmarking datasets: MSR-VTT [55], MSVD [3], LSMDC [41], ActivityNet [24] and DiDeMo [12]. These datasets vary in video duration, research goals, target users, and text annotations, providing a comprehensive evaluation of different methods. We evaluate the Text→Video (**T2V**) performance and provide additional results of Video→Text (**V2T**) performance on standard rank-based metrics *i.e.* Recall at top {1, 5, 10} (recall at rank 1, 5, 10), Rsum (R@1 + R@5 + R@10).

**Implementation Details.** We use `CLIP` ViT/B-32 as the backbone for both *Video Encoder* and *Text Encoder*. We set $N_t = 32$ and $N_f = 32$ as the number of word tokens and video frames for all datasets except DiDeMo and ActivityNet, where $N_t$ and $N_f$ are set to 64. We train with batch size of 128 for 5 epochs, except for DiDeMo with 10 epochs and ActivityNet with 20 epochs. We use the Adam [22] as the optimizer. The learning rate follows the cosine schedule with a linear warmup strategy [10]. For Equation 4, we set $\kappa = 20$. For Equation 13, We set $\lambda_{\texttt{cap}} = 0.3$, $\lambda_{\texttt{rec}} = 0.5$, and $\lambda_{\mathbb{KL}} = 1e^{-4}$.

## 4.2 Quantitative Results

**Comparison with SOTA.** We compare the proposed `BiMa` with SOTA methods. Table 1 shows the retrieval results on the MSR-VTT dataset, where our proposed method attains SOTA on both **T2V** and **V2T** tasks. We achieve 53.5 in the **T2V** task, outperforming the runner-up `TextProxy` [54] by 1.2 and the third best `DITS` [52] by 1.6 w.rt. R@1. On other metrics, `BiMa` also significantly outperforms the runner-up results. In **V2T** task, we achieve 52.2, outperforming the runner-up `NarVid` [15] by 2.2 w.rt. R@1. Table 2 shows the retrieval results on **T2V**, where `BiMa` consistently achieves the best results on all the metrics on all four datasets MSVD, LSMDC, ActivityNet, and DiDeMo. It demonstrates that our strategy can work well across different domains and different text-video data variations, thus underscoring the efficacy of `BiMa` through the "Visual-Linguistic Bias Mitigation" technique. For example, on MSVD, our approach at 55.7 has outperformed the recent SOTA method `NarVid` [15] by 2.6 w.r.t R@1. Similarly on ActivityNet, `BiMa` significantly outperforms the runner-up `TextProxy` by 2.4 and w.r.t R@1, achieving SOTA performance of 55.4.

**Generalization to Unseen Domains.** Our mitigation of visual-linguistic biases is geared towards maximizing the quality of cross-modal semantic representations during training, thereby capturing better underlying patterns that can generalize to unseen data. Hence, to evaluate `BiMa`'s debias capability, we evaluate `BiMa` on out-of-distribution retrieval settings [4] (denoted as A → B) where the model is trained on dataset A and benchmarked on dataset B, which is unseen during training. In Table 3, we compare our `BiMa` with recent SOTAs and the baseline `CLIP4Clip` in three OOD retrieval benchmarks (MSR-VTT → LSMDC, MSR-VTT → ActivityNet, and MSR-VTT → DiDeMo) where `BiMa` significantly outperforms others on the three benchmarks. These results suggest that our bias mitigation capability supports model generalization and transfer to unseen data.

**Hyperparameter Sensitivity.** In Figure 3, we evaluate hyperparameters $\lambda_{\text{Cap}} \in [0.1, 0.5]$, $\lambda_{\text{Rec}} \in [0.3, 0.7]$ and $\lambda_{\mathbb{KL}} \in [0.0001, 0.0005]$. In Figure 3(left), `BiMa` achieves highest R@1 score when $\lambda_{\text{Cap}} = 0.3$ for **T2V**. As a result, we select $\lambda_{\text{Cap}} = 0.3$. In Figure 3(middle), the best is achieved with $\lambda_{\text{Rec}} = 0.5$, so we set $\lambda_{\text{Rec}} = 0.5$ as the default. In Figure 3(right), we show that $\lambda_{\mathbb{KL}}$ is highly sensitive and leads to a trade-off between **T2V** and **V2T** performance when $\lambda_{\mathbb{KL}}$ change from 0.0001 to 0.0003. Thus, we set $\lambda_{\mathbb{KL}} = 0.0001$ to balance between **T2V** and **V2T**.

**Effect of the Visual Scene Debias.** Table 4 shows that scene element features can improve performance through (i) *Scene Entities* (see Exp #2) or *Scene Activities* (see Exp #3):—either when only one type of scene element is fused with video representation, the R@1 score improved by a large margin of 1.4 (only scene entities) and 1.7 (only scene activities) compared to Baseline;—or, through *Scene Elements Aggregation* (see Exp #4), these features can be combined to enable an improvement over single feature utilization, with R@1 increased to 51.5 from 49.2 / 49.4. Furthermore, through (ii) *Captioning Head* (see Exp #5), better alignment between element-aware visual scene features and textual features can also be achieved, hinting that visual embeddings are enabled with finer-grained cross-modal features as R@1 increased from 51.5 to 52.1.

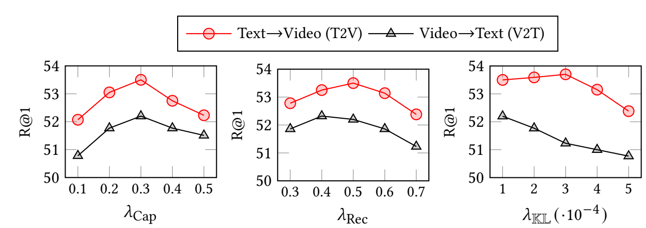


**Figure 3: Hyperparameter sensitivity study on MSR-VTT.**

**Effect of Textual Content Debias.** Table 4 shows the effect of the *Textual Content Debias* module. From Exp #5 and #7, we observe that the efficacy of the *Text Content Debias* and *Visual Scene Debias*

**Table 1: T2V and V2T performance comparisons on MSR-VTT dataset. The best and second best are bold and <u>underlined</u>. Two-stage methods are marked with †.**

| PT | Methods | T2V | | | | V2T | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | R@1↑ | R@5↑ | R@10↑ | Rsum↑ | R@1↑ | R@5↑ | R@10↑ | Rsum↑ |
| ✓ | ClipBERT [27] CVPR'20 | 22.0 | 46.8 | 59.9 | 128.7 | - | - | - | - |
| | SupportSet [38] ICLR'21 | 30.1 | 58.5 | 69.3 | 157.9 | 28.5 | 58.6 | 71.6 | 158.7 |
| | Frozen [1] ICCV'21 | 32.5 | 61.5 | 71.2 | 165.2 | - | - | - | - |
| | TMVM [32] NeurIPS'22 | 36.2 | 64.2 | 75.7 | 176.1 | 34.8 | 63.8 | 73.7 | 172.3 |
| | RegionLearner [57] AAAI'23 | 36.3 | 63.9 | 72.5 | 172.7 | 34.3 | 63.5 | 73.2 | 171.0 |
| | In-Style [43] ICCV'23 | 36.2 | 61.8 | 71.9 | 169.9 | - | - | - | - |
| ✗ | CenterCLIP [61] SIGIR'22 | 44.2 | 71.6 | 82.1 | 197.9 | 42.8 | 71.7 | 82.2 | 196.7 |
| | CLIP4Clip [37] Neurocomputing'22 | 44.5 | 71.4 | 81.6 | 197.5 | 42.7 | 70.9 | 80.6 | 194.2 |
| | EMCL-Net [17] NeurIPS'22 | 46.8 | 73.1 | 83.1 | 203.0 | 46.5 | 73.5 | 83.5 | 203.5 |
| | X-Pool [9] CVPR'22 | 46.9 | 72.8 | 82.2 | 201.9 | - | - | - | - |
| | TS2-Net [35] ECCV'22 | 47.0 | 74.5 | 83.8 | 205.3 | 45.3 | 74.1 | 83.7 | 203.1 |
| | HBI [18] CVPR'23 † | 48.6 | 74.6 | 83.4 | 206.6 | 46.8 | 74.3 | 84.3 | 205.4 |
| | DiCoSA [19] IJCAI'23 | 47.5 | 74.7 | 83.8 | 206.0 | 46.7 | 75.2 | 84.3 | 206.2 |
| | UATVR [7] ICCV'23 | 47.5 | 73.9 | 83.5 | 204.9 | 46.9 | 73.8 | 83.8 | 204.5 |
| | DiffusionRet [20] ICCV'23 † | 49.0 | 75.2 | 82.7 | 206.9 | 47.7 | 73.8 | 84.5 | 206.0 |
| | PAU [20] NeurIPS'23 | 48.5 | 72.7 | 82.5 | 203.7 | 48.3 | 73.0 | 83.2 | 204.5 |
| | DGL [58] AAAI'24 | 45.8 | 69.3 | 79.4 | 194.5 | - | - | - | - |
| | EERCF [6] AAAI'24 | 47.8 | 74.1 | 84.1 | 206.0 | 44.7 | 74.2 | 83.9 | 202.8 |
| | TeachCLIP [47] CVPR'24 | 46.8 | 74.3 | 82.6 | 203.7 | - | - | - | - |
| | DITS [52] NeurIPS'24 | 51.9 | 75.7 | 84.6 | 212.2 | - | - | - | - |
| | TextProxy [54] AAAI'25 | <u>52.3</u> | <u>77.8</u> | <u>85.8</u> | <u>215.9</u> | - | - | - | - |
| | TempMe [42] ICLR'25 | 46.1 | 71.8 | 80.7 | 198.6 | 45.6 | 72.4 | 81.2 | 199.2 |
| | NarVid [15] CVPR'25 | 51.0 | 76.4 | 85.2 | 212.6 | <u>50.0</u> | <u>75.4</u> | <u>83.8</u> | <u>209.2</u> |
| | **BiMa (ours)** | **53.5** | **78.6** | **86.5** | **218.6** | **52.2** | **77.1** | **85.3** | **214.6** |

**Table 2: T2V comparisons on MSVD, LSMDC, ActivityNet, and DiDeMo datasets. The best and second best are bold and <u>underlined</u>.**

| Methods | MSVD | | | | LSMDC | | | | ActivityNet | | | | DiDeMo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1↑ | R@5↑ | R@10↑ | Rsum↑ | R@1↑ | R@5↑ | R@10↑ | Rsum↑ | R@1↑ | R@↑↑ | R@10↑ | Rsum↑ | R@1↑ | R@5↑ | R@10↑ | Rsum↑ |
| ClipBERT | - | - | - | - | - | - | - | - | 21.3 | 49.0 | 63.5 | 133.8 | 20.4 | 48.0 | 60.8 | 129.2 |
| SupportSet | 28.4 | 60.0 | 72.9 | 161.3 | - | - | - | - | 29.2 | 61.6 | 94.7 | 185.5 | - | - | - | - |
| Frozen | 33.7 | 64.7 | 76.3 | 174.7 | 15.0 | 30.8 | 40.3 | 86.1 | - | - | - | - | 34.6 | 65.0 | 74.7 | 174.3 |
| TMVM | 36.7 | 67.4 | 81.3 | 185.4 | 17.8 | 37.1 | 45.9 | 100.8 | - | - | - | - | 36.5 | 64.9 | 75.4 | 176.8 |
| RegionLearner | 44.0 | 74.9 | 84.3 | 203.2 | 17.1 | 32.5 | 41.5 | 91.1 | - | - | - | - | 32.5 | 60.8 | 72.3 | 165.6 |
| In-Style | 44.8 | 72.5 | 81.2 | 198.5 | 16.1 | 33.6 | 39.7 | 89.4 | - | - | - | - | 32.1 | 61.9 | 71.2 | 165.2 |
| CenterCLIP | 45.2 | 75.5 | 84.3 | 205.0 | 22.6 | 41.0 | 49.1 | 112.7 | 40.5 | 72.4 | 83.6 | 196.5 | 42.8 | 68.5 | 79.2 | 190.5 |
| CLIP4Clip | 45.2 | 75.5 | 84.3 | 205.0 | 22.6 | 41.0 | 49.1 | 112.7 | 40.5 | 72.4 | 83.6 | 196.5 | 42.8 | 68.5 | 79.2 | 190.5 |
| EMCL-Net | 42.1 | 71.3 | 81.1 | 194.5 | 23.9 | 42.4 | 50.9 | 117.2 | 41.2 | 72.7 | 83.6 | 197.5 | 45.3 | 74.2 | 82.3 | 201.8 |
| X-Pool | 47.2 | 77.4 | 86.0 | 210.6 | 25.2 | 43.7 | 53.5 | 122.4 | - | - | - | - | - | - | - | - |
| TS2-Net | - | - | - | - | 23.4 | 42.3 | 50.9 | 116.6 | 41.0 | 73.6 | 84.5 | 199.1 | 41.8 | 71.6 | 82.0 | 195.4 |
| HBI | - | - | - | - | - | - | - | - | 42.2 | 73.0 | 84.6 | 199.8 | 46.9 | 74.9 | 82.7 | 204.5 |
| DiCoSa | 47.4 | 76.8 | 86.0 | 210.2 | 25.4 | 43.6 | 54.0 | 123.0 | 42.1 | 73.6 | 84.6 | 200.3 | 45.7 | 74.6 | 83.5 | 203.8 |
| UATVR | 46.0 | 76.3 | 85.1 | 207.4 | - | - | - | - | - | - | - | - | 43.1 | 71.8 | 82.3 | 197.2 |
| DiffusionRet | 46.6 | 75.9 | 84.1 | 206.6 | 24.4 | 43.1 | 54.3 | 121.8 | 45.8 | 75.6 | 86.3 | 207.7 | 46.7 | 74.7 | 82.7 | 204.1 |
| PAU | 47.3 | 77.4 | 85.5 | 210.2 | - | - | - | - | - | - | - | - | 48.6 | 76.0 | 84.5 | 209.1 |
| EERCF | 47.0 | 77.5 | 85.4 | 209.9 | - | - | - | - | 43.1 | 74.5 | 86.0 | 203.6 | - | - | - | - |
| DGL | - | - | - | - | 21.4 | 39.4 | 48.4 | 109.2 | 38.6 | 69.2 | 81.6 | 189.4 | - | - | - | - |
| TeachCLIP | 47.4 | 77.3 | 85.5 | 210.2 | - | - | - | - | 42.2 | 72.7 | 85.2 | 200.1 | 43.7 | 71.2 | 81.1 | 196.0 |
| DITS | - | - | - | - | <u>28.2</u> | <u>47.3</u> | <u>56.6</u> | <u>132.1</u> | - | - | - | - | 51.1 | 77.9 | 85.8 | 214.8 |
| TextProxy | - | - | - | - | - | - | - | - | <u>53.0</u> | <u>80.9</u> | <u>89.6</u> | <u>223.5</u> | 50.6 | 76.9 | 86.0 | 213.5 |
| TempMe | - | - | - | - | 23.5 | 41.7 | 51.8 | 117.0 | 44.9 | 75.2 | 85.5 | 205.6 | 48.0 | 72.4 | 81.8 | 202.2 |
| NarVid | <u>53.1</u> | <u>81.4</u> | <u>88.8</u> | <u>223.3</u> | - | - | - | - | - | - | - | - | <u>53.4</u> | <u>79.1</u> | <u>86.3</u> | <u>218.8</u> |
| **BiMa (ours)** | **55.7** | **83.2** | **91.2** | **230.1** | **35.6** | **56.5** | **68.1** | **160.2** | **55.4** | **83.4** | **92.4** | **231.2** | **56.0** | **81.9** | **87.8** | **225.7** |

**Table 3: Out-of-distribution performance of T2V models on LSMDC, ActivityNet, and DiDeMo. The best and second-best results are highlighted in bold and <u>underlined</u>, respectively.**

| Methods | MSR-VTT→ LSMDC | | | | MSR-VTT → ActivityNet | | | | MSR-VTT → DiDeMo | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | R@1↑ | R@5↑ | R@10↑ | Rsum↑ | R@1↑ | R@5↑ | R@10↑ | Rsum↑ | R@1↑ | R@5↑ | R@10↑ | Rsum↑ |
| CLIP4Clip | 15.3 | 31.3 | 40.5 | 87.1 | 29.1 | 58.3 | 72.1 | 159.5 | 31.8 | 57.0 | 66.1 | 154.9 |
| EMCL-Net | 16.6 | 29.3 | 36.5 | 82.4 | 28.7 | 56.8 | 70.6 | 156.1 | 30.0 | 56.1 | 65.8 | 151.9 |
| DiffusionRet | 17.1 | 32.4 | 41.0 | 90.5 | 31.5 | 60.0 | 73.8 | 165.3 | 33.2 | 59.3 | 68.4 | 160.9 |
| **BiMa (ours)** | **24.0** | **45.0** | **56.8** | **125.8** | **35.2** | **64.8** | **89.0** | **189.0** | **36.7** | **62.6** | **71.2** | **170.5** |

improves R@1 from 52.1 to 53.5. This suggests that the *Textual Content Debias* module is complementary with *Visual Scene Debias* to improve performance.

**Table 4: T2V ablation studies for network designs on MSR-VTT. Exp #1 is our baseline which is `CLIP4clip` with WTI matching.**

| Exp | Visual Scene Debias | | | Textual Content Debias | Performance | | |
|---|---|---|---|---|---|---|---|
| | Scene Element | | Captioning Head | Content-Bias Representation Disentanglement | | | |
| | Scene Entities | Scene Activities | | | R@1↑ | R@5↑ | R@10↑ |
| #1 | ✗ | ✗ | ✗ | ✗ | 45.7 | 73.0 | 82.6 |
| #2 | ✓ | ✗ | ✗ | ✗ | 49.2 | 75.3 | 83.0 |
| #3 | ✗ | ✓ | ✗ | ✗ | 49.4 | 75.7 | 83.4 |
| #4 | ✓ | ✓ | ✗ | ✗ | 51.5 | 76.6 | 84.8 |
| #5 | ✓ | ✓ | ✓ | ✗ | 52.1 | 77.2 | 86.0 |
| #6 | ✗ | ✗ | ✗ | ✓ | 47.0 | 75.8 | 84.0 |
| #7 | ✓ | ✓ | ✓ | ✓ | **53.5** | **78.6** | **86.5** |

**Table 5: Ablation study between content features $\hat{\mathbf{t}}$ or bias features $\tilde{\mathbf{t}}$ as textual features for T2V task on MSR-VTT.**

| Methods | R@1↑ | R@5↑ | R@10↑ |
|---|---|---|---|
| *using* $\hat{\mathbf{t}}$ (content) | **53.5** | **78.6** | **86.5** |
| *using* $\tilde{\mathbf{t}}$ (bias) | 1.3 | 10.2 | 18.9 |

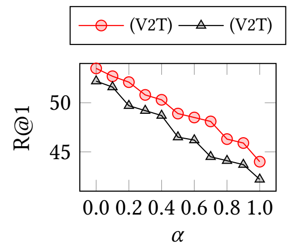


**Figure 4: Effect of bias features on MSR-VTT.**

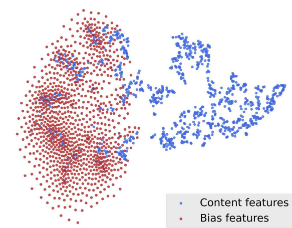


**Figure 5: t-SNE [49] visualization between content and bias embeddings on MSR-VTT test-set.**

**Effect of Textual Bias Features.** To evaluate the impact of bias on textual representations in retrieval tasks, we conduct a controlled analysis by fusing content and bias features using a weighted sum: $\hat{\mathbf{t}}^{(i)} + \alpha \cdot \tilde{\mathbf{t}}^{(i)}$, where $\hat{\mathbf{t}}^{(i)}$ denotes the content embedding and $\tilde{\mathbf{t}}^{(i)}$ the bias embedding. The scalar parameter $\alpha$ modulates the degree of bias injected into the content representation. By varying $\alpha$, we systematically assess how retrieval performance degrades as bias intensity increases, thereby quantifying the trade-off between preserving semantic fidelity and suppressing bias-induced noise.

Quantitatively, in Figure 4, we illustrate the impact of varying $\alpha$ on retrieval performance. As $\alpha$ transitions from content-only to a mix of content and bias features, we observe that introducing more bias into the content increases noise during retrieval, which degrades performance. This demonstrates how the presence of bias disturbs the retrieval process.

As shown in Table 5, we further validate the importance of disentangling content features from bias features in the TVR task. Specifically, the use of content features $\hat{\mathbf{t}}^{(i)}$ achieves significantly higher retrieval performance compared to the use of bias features $\tilde{\mathbf{t}}^{(i)}$ across all recall metrics. When relying on $\hat{\mathbf{t}}^{(i)}$ (content features), the model attains an R@1 of 53.5%, R@5 of 78.6%, and R@10 of 86.5%. In stark contrast, using $\tilde{\mathbf{t}}^{(i)}$ (bias features) alone results in extremely poor performance, with R@1 dropping to 1.3%, R@5 to 10.2%, and R@10 to 18.9%. This substantial gap clearly illustrates that bias features, when isolated, fail to provide sufficient semantic grounding for accurate retrieval. Instead, they appear to introduce noise that misleads the model away from the true video content, corroborating our earlier hypothesis that biases embedded in the linguistic patterns can harm retrieval accuracy if not properly handled. Ultimately, our analysis underscores that effective bias mitigation, *e.g.* using our proposed strategy, is essential for generalizable TVR.

Qualitatively, in Figure 5, we visualize the content and bias feature embeddings of 1,000 samples from the MSR-VTT test set using t-SNE [49]. The results show that our disentanglement process effectively separates content and bias components within the textual features. However, we also observe that some bias features remain partially entangled with the content representation. This highlights the inherent difficulty in completely disentangling bias information and suggests a promising direction for future investigation.

Additional ablation studies on **Effect of the Number of Scene Elements**, **Computational Cost**, and **Effect of Coefficient** $g$ are included in the **Supplementary Material**, respectively.

## 4.3 Qualitative Results

Figure 6 qualitatively demonstrates the performance of `BiMa` on MSR-VTT testset. We observe that `BiMa` is able to focus on relevant scene entities and scene activities, shown by the top three attention heatmaps per video, highlighting key elements in the scene. Further, `BiMa` can focus on small yet important actors and objects *e.g.*, the "alarm clock" in the 1st example and the "animal" in the 4th example. Next, `BiMa` also can focus on behaviors *e.g.*, "throwing" in the 4th example. This suggests that `BiMa` effectively integrates scene elements, which contributes to its improved retrieval accuracy over `CLIP4Clip`. In contrast, `CLIP4Clip` is unable to identify important scene elements and retrieves unrelated results. In summary, the experimental results highlight the advantage of `BiMa` in associating visual content with textual descriptions, showing its enhanced performance in capturing nuanced visual features relevant to the query. Additional qualitative results are presented in the **Supplementary Material**.

# 5 Related Work

**Text-Video Retrieval.** TVR is a cross-modal retrieval task that aims to match videos with text descriptions [6, 7, 18, 20, 25, 35, 47, 57]. Existing TVR methods typically leverage pre-trained VLMs like `CLIP` [40], `BLIP` [29] and adopt contrastive learning on pairs of samples to enhance their cross-modal representations across vision and language domains. Recent advancements, such as `DiffusionRet` [20], introduce multi-stage training mechanisms that use both discriminative and generative models to address challenges in out-of-domain retrieval tasks. In cases of ambiguous matching, where videos have multiple valid captions, uncertainty-based frameworks have emerged to handle this by learning a joint embedding space with probabilistic distance metrics. For example, `UATVR` [7] aligns cross-modal features as a distribution-matching procedure, while `PAU` [28] introduces semantic prototypes to capture ambiguous semantics within an uncertainty-based framework. Although our work also addresses data ambiguities, we distinguish between "bias" and "uncertainty" in terms of both definition and objective. We define "bias" as overlooked features stemming from coarse-grained data and annotator subjectivity, whereas "uncertainty" refers to the

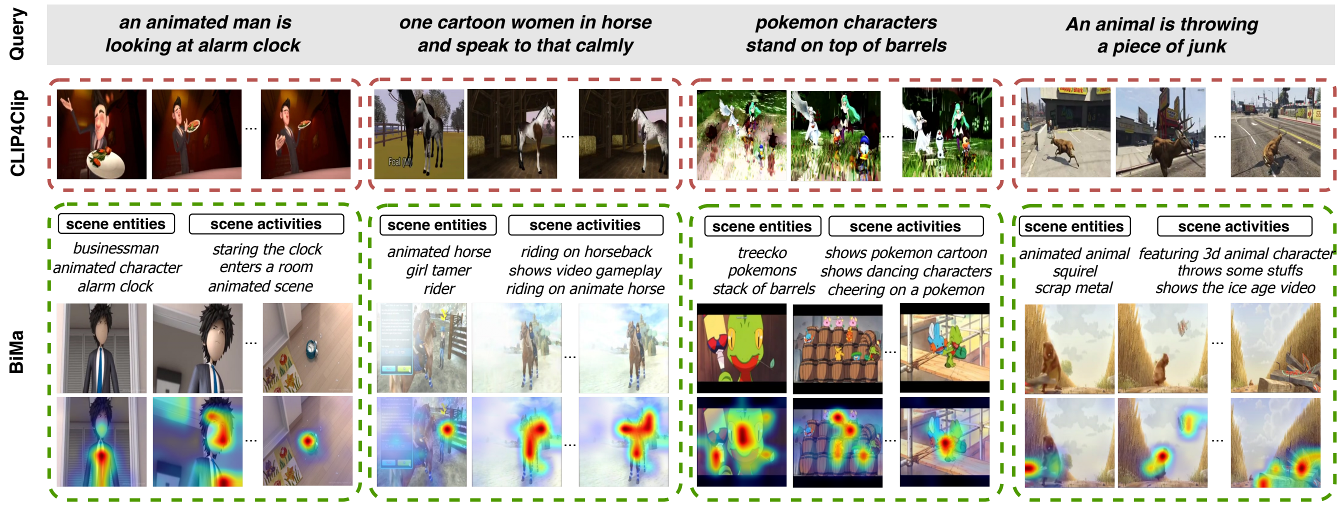


**Figure 6: TVR qualitative Results. Videos in green boxes are true top-1 videos retrieved from `BiMa`, and in red boxes are false top-1 videos retrieved from `CLIP4Clip`. We also provide attention visualizations and the top 3 relevant scene elements for each video to show where `BiMa` is focused on.**

diversity within representations due to the variation of annotators. Our approach focuses on mitigating biases by recalibrating cross-modal features in the latent space and treating bias components as noise. In contrast, uncertainty-based methods attempt to model representation diversity by treating matched visual-linguistic representations as stochastic variables, and they do not directly alleviate visual-linguistic biases.

**Bias in Vision-Language Models.** Bias in VLMs has become a growing concern as these models, widely used in real-world applications, are often pre-trained on large internet-scale datasets. While this provides VLMs with extensive knowledge, it also makes them susceptible to inheriting biases present in the underlying data, such as cultural stereotypes, racial biases, and gender imbalances [2, 8, 31, 45]. For instance, `CLIP` and `BLIP` tend to amplify societal biases [53]. Recent literature explicitly categorizes these biases as representation biases—systematic deviations or skews within datasets, causing models to disproportionately rely on certain dataset-specific patterns rather than generalizable, task-relevant features [36, 44]. These biases typically manifest as concept bias (where models rely heavily on prominent visual concepts), temporal bias (where temporal aspects are inadequately represented), and textual bias (arising from subjective annotator interpretations or emotionally charged annotations) [44]. Additionally, recent empirical evaluations reveal that even modern datasets, designed with significant diversification efforts, still contain intrinsic dataset-specific biases, limiting models' generalizability and robustness [36]. As VLMs become more prominent, addressing and mitigating these biases is critical for ensuring ethical and fair AI applications. For visual bias, existing approaches have relied on costly solutions such as simulators [26] or crowdsourcing [16, 39] to annotate visual features. These approaches are not scalable and lack universal applicability. Recent studies [51, 56] utilize scene factorization to eliminate biases in visual representation by focusing on relevant scene entities. However, many of these approaches rely heavily on computationally intensive object detectors, restricting their applicability to unseen scenarios. For textual bias, existing studies address biases in sentence embeddings by introducing training constraints [14] or directly modifying datasets [60]. Recent Autoencoder-based methods [5, 11, 21, 46] aim to detach the domain bias from textual representations. However, these methods typically rely on supervised learning using additional bias annotations, limiting their generalizability and scalability. To the best of our knowledge, `BiMa` represents a pioneering effort in systematically addressing these identified visual and textual representation biases in VLMs through the integration of scene element aggregation, visual attention redirection, and textual disentanglement. Crucially, our method is self-supervised and does not require additional human annotations, effectively addressing the limitations noted in prior studies [36, 44].

# 6 Conclusion

**Conclusion.** In this study, we addressed visual-linguistic bias challenges in the TVR task, drawing upon recent bias representation findings [36, 44] to address an unexplored area in the TVR literature. We proposed `BiMa`, a novel framework to mitigate biases in both visual and textual representations. Our proposed framework incorporates three modules– *Scene Element Construction*, *Visual Scene Debias*, and *Textual Content Debias*. Through extensive experiments and ablation studies across five major TVR benchmark datasets, we demonstrated the remarkable performance of our approach, surpassing existing SOTA TVR methods. Additionally, our `BiMa` exhibited remarkable performance in handling out-of-distribution retrieval problems showing the bias mitigation and generalization capabilities on unseen data variations.

**Limitation.** Although our taxonomy dictionary is constructed from diverse sources containing a comprehensive set of scene elements, some inherent biases may remain unmitigated. Addressing these challenges remains an avenue for future research.

## References


[1] Max Bain, Arsha Nagrani, Gül Varol, and Andrew Zisserman. 2021. Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 1708–1718.

[2] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Annual Conference on Neural Information Processing Systems (NeurIPS)* (2016), 4349–4357.

[3] David L. Chen and William B. Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*. 190–200.

[4] Shizhe Chen, Yida Zhao, Qin Jin, and Qi Wu. 2020. Fine-Grained Video-Text Retrieval With Hierarchical Graph Reasoning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 10635–10644.

[5] Tianlang Chen, Zhongping Zhang, Quanzeng You, Chen Fang, Zhaowen Wang, Hailin Jin, and Jiebo Luo. 2018. "Factual" or "Emotional": Stylized Image Captioning with Adaptive Learning and Attention. In *European Conference on Computer Vision (ECCV)*. 527–543.

[6] Yizhen Chen, Jie Wang, Lijian Lin, Zhongang Qi, Jin Ma, and Ying Shan. 2023. Tagging before Alignment: Integrating Multi-Modal Tags for Video-Text Retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*. 396–404.

[7] Bo Fang, Wenhao Wu, Chang Liu, Yu Zhou, Yuxin Song, Weiping Wang, Xiangbo Shu, Xiangyang Ji, and Jingdong Wang. 2023. UATVR: Uncertainty-Adaptive Text-Video Retrieval. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 13677–13687.

[8] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), 3635–3644.

[9] Satya Krishna Gorti, Noël Vouitsis, Junwei Ma, Keyvan Golestan, Maksims Volkovs, Animesh Garg, and Guangwei Yu. 2022. X-Pool: Cross-Modal Language-Video Attention for Text-Video Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4996–5005.

[10] Priya Goyal, Piotr Dollár, Ross B. Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. 2017. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour. *arXiv preprint arXiv:706.02677* (2017).

[11] Longteng Guo, Jing Liu, Peng Yao, Jiangwei Li, and Hanqing Lu. 2019. MSCap: Multi-Style Image Captioning With Unpaired Stylized Text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 4204–4213.

[12] Lisa Anne Hendricks, Oliver Wang, Eli Shechtman, Josef Sivic, Trevor Darrell, and Bryan C. Russell. 2017. Localizing Moments in Video with Natural Language. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 5804–5813.

[13] Irina Higgins, Loïc Matthey, Arka Pal, Christopher P. Burgess, Xavier Glorot, Matthew M. Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-VAE: Learning Basic Visual Concepts with a Constrained Variational Framework. In *International Conference on Learning Representations (ICLR)*.

[14] Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. 2020. Reducing Sentiment Bias in Language Models via Counterfactual Evaluation. In *Findings of Conference on Empirical Methods in Natural Language Processing (EMLP)*. 65–83.

[15] Chan Hur, Jeong hun Hong, Dong hun Lee, Dabin Kang, Semin Myeong, Sang hyo Park, and Hyeyoung Park. 2025. Narrating the Video: Boosting Text-Video Retrieval via Comprehensive Utilization of Frame-Level Captions. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

[16] Badr Youbi Idrissi, Diane Bouchacourt, Randall Balestriero, Ivan Evtimov, Caner Hazirbas, Nicolas Ballas, Pascal Vincent, Michal Drozdzal, David Lopez-Paz, and Mark Ibrahim. 2023. ImageNet-X: Understanding Model Mistakes with Factor of Variation Annotations. In *International Conference on Learning Representation (ICLR)*.

[17] Peng Jin, Jinfa Huang, Fenglin Liu, Xian Wu, Shen Ge, Guoli Song, David A. Clifton, and Jie Chen. 2022. Expectation-Maximization Contrastive Learning for Compact Video-and-Language Representations. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 30291–30306.

[18] Peng Jin, Jinfa Huang, Pengfei Xiong, Shangxuan Tian, Chang Liu, Xiangyang Ji, Li Yuan, and Jie Chen. 2023. Video-Text as Game Players: Hierarchical Banzhaf Interaction for Cross-Modal Representation Learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2472–2482.

[19] Peng Jin, Hao Li, Zesen Cheng, Jinfa Huang, Zhennan Wang, Li Yuan, Chang Liu, and Jie Chen. 2023. Text-Video Retrieval with Disentangled Conceptualization and Set-to-Set Alignment. In *International Joint Conference on Artificial Intelligence (IJCAI)*. 938–946.

[20] Peng Jin, Hao Li, Zesen Cheng, Kehan Li, Xiangyang Ji, Chang Liu, Li Yuan, and Jie Chen. 2023. DiffusionRet: Generative Text-Video Retrieval with Diffusion Model. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 2470–2481.

[21] Vineet John, Lili Mou, Hareesh Bahuleyan, and Olga Vechtomova. 2019. Disentangled Representation Learning for Non-Parallel Text Style Transfer. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*. 424–434.

[22] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *International Conference on Learning Representation (ICLR)*.

[23] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representation (ICLR)*.

[24] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017. Dense-Captioning Events in Videos. In *IEEE/CVF International Conference on Computer Vision (ICCV)*. 706–715.

[25] Huy Le, Tung Kieu, Anh Nguyen, and Ngan Le. 2024. WAVER: Writing-Style Agnostic Text-Video Retrieval Via Distilling Vision-Language Models Through Open-Vocabulary Knowledge. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*.

[26] Guillaume Leclerc, Hadi Salman, Andrew Ilyas, Sai Vemprala, Logan Engstrom, Vibhav Vineet, Kai Xiao, Pengchuan Zhang, Shibani Santurkar, Greg Yang, et al. 2022. 3db: A framework for debugging computer vision models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 8498–8511.

[27] Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L. Berg, Mohit Bansal, and Jingjing Liu. 2021. Less Is More: ClipBERT for Video-and-Language Learning via Sparse Sampling. In *IEEE/CVF International Conference on Computer Vision and Pattern Recognition (CVPR)*. 7331–7341.

[28] Hao Li, Jingkuan Song, Lianli Gao, Xiaosu Zhu, and Hengtao Shen. 2023. Prototype-based Aleatoric Uncertainty Quantification for Cross-modal Retrieval. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 24564–2458.

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning (ICML)*. 12888–12900.

[30] Manling Li, Ruochen Xu, Shuohang Wang, Luowei Zhou, Xudong Lin, Chenguang Zhu, Michael Zeng, Heng Ji, and Shih-Fu Chang. 2022. Clip-event: Connecting text and images with event structures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 16420–16429.

[31] Paul Pu Liang, Irene Mengze Li, Emily Zheng, Yao Chong Lim, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2020. Towards Debiasing Sentence Representations. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*.

[32] Chengzhi Lin, Ancong Wu, Junwei Liang, Jun Zhang, Wenhang Ge, Wei-Shi Zheng, and Chunhua Shen. 2022. Text-adaptive multiple visual prototype matching for video-text retrieval. *Annual Conference on Neural Information Processing Systems (NeurIPS)* (2022), 38655–38666.

[33] Daizong Liu, Xiaoye Qu, and Wei Hu. 2022. Reducing the Vision and Language Bias for Temporal Sentence Grounding. In *ACM International Conference on Multimedia (MM)*. 4092–4101.

[34] Fan Liu, Huilin Chen, Zhiyong Cheng, Anan Liu, Liqiang Nie, and Mohan Kankanhalli. 2022. Disentangled multimodal representation learning for recommendation. *IEEE Transactions on Multimedia* 25 (2022), 7149–7159.

[35] Yuqi Liu, Pengfei Xiong, Luhui Xu, Shengming Cao, and Qin Jin. 2022. Ts2-net: Token shift and selection transformer for text-video retrieval. In *European Conference on Computer Vision (ECCV)*. 319–335.

[36] Zhuang Liu and Kaiming He. 2025. A Decade's Battle on Dataset Bias: Are We There Yet?. In *ICLR*.

[37] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: An empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing* 508 (2022), 293–304.

[38] Mandela Patrick, Po-Yao Huang, Yuki Markus Asano, Florian Metze, Alexander G. Hauptmann, João F. Henriques, and Andrea Vedaldi. 2021. Support-Set Bottlenecks for Video-Text Representation Learning. In *International Conference on Learning Representations (ICLR)*.

[39] Gregory Plumb, Marco Túlio Ribeiro, and Ameet Talwalkar. 2022. Finding and Fixing Spurious Patterns with Explanations. *Transactions on Machine Learning Research* (2022).

[40] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*. 8748–8763.

[41] Anna Rohrbach, Atousa Torabi, Marcus Rohrbach, Niket Tandon, Christopher Pal, Hugo Larochelle, Aaron Courville, and Bernt Schiele. 2017. Movie description. *International Journal of Computer Vision* 123 (2017), 94–120.

[42] Leqi Shen, Tianxiang Hao, Tao He, Sicheng Zhao, Yifeng Zhang, pengzhang liu, Yongjun Bao, and Guiguang Ding. 2025. TempMe: Video Temporal Token Merging for Efficient Text-Video Retrieval. In *ICLR*.

[43] Nina Shvetsova, Anna Kukleva, Bernt Schiele, and Hilde Kuehne. 2023. In-Style: Bridging Text and Uncurated Videos with Style Transfer for Text-Video Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 21981–21992.

[44] Nina Shvetsova, Arsha Nagrani, Bernt Schiele, Hilde Kuehne, and Christian Rupprecht. 2025. Unbiasing through Textual Descriptions: Mitigating Representation

Bias in Video Benchmarks. (2025).

[45] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth M. Belding, Kai-Wei Chang, and William Yang Wang. 2019. Mitigating Gender Bias in Natural Language Processing: Literature Review. In *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL)*. 1630–1640.

[46] Yutong Tan, Zheng Lin, Peng Fu, Mingyu Zheng, Lanrui Wang, Yanan Cao, and Weiping Wang. 2022. Detach and Attach: Stylized Image Captioning without Paired Stylized Dataset. In *ACM International Conference on Multimedia (MM)*. 4733–4741.

[47] Kaibin Tian, Ruixiang Zhao, Zijie Xin, Bangxiang Lan, and Xirong Li. 2024. Holistic Features are almost Sufficient for Text-to-Video Retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 17138–17147.

[48] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation Learning with Contrastive Predictive Coding. *arXiv preprint arXiv:1807.03748* (2018).

[49] Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research* 9 (2008), 2579–2605.

[50] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 5998–6008.

[51] Khoa Vo, Sang Truong, Kashu Yamazaki, Bhiksha Raj, Minh-Triet Tran, and Ngan Le. 2023. Aoe-net: Entities interactions modeling with adaptive attention mechanism for temporal action proposals generation. *International Journal of Computer Vision* 131, 1 (2023), 302–323.

[52] Jiamian Wang, Pichao Wang, Dongfang Liu, Qiang Guan, Sohail A. Dianat, Majid Rabbani, Raghuveer Rao, and Zhiqiang Tao. 2024. Diffusion-Inspired Truncated Sampler for Text-Video Retrieval. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024 NeurIPS 2024)*.

[53] Zeyu Wang, Klint Qinami, Ioannis Christos Karakozis, Kyle Genova, Prem Nair, Kenji Hata, and Olga Russakovsky. 2020. Towards Fairness in Visual Recognition: Effective Strategies for Bias Mitigation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 8916–8925.

[54] Jian Xiao, Zhenzhen Hu, Jia Li, and Richang Hong. 2025. Text Proxy: Decomposing Retrieval from a 1-to-N Relationship into N 1-to-1 Relationships for Text-Video Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

[55] Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 5288–5296.

[56] Kashu Yamazaki, Khoa Vo, Quang Sang Truong, Bhiksha Raj, and Ngan Le. 2023. VLTinT: visual-linguistic transformer-in-transformer for coherent video paragraph captioning. In *AAAI Conference on Artificial Intelligence (AAAI)*. 3081–3090.

[57] Rui Yan, Mike Zheng Shou, Yixiao Ge, Jinpeng Wang, Xudong Lin, Guanyu Cai, and Jinhui Tang. 2023. Video-text pre-training with learned regions for retrieval. In *AAAI Conference on Artificial Intelligence (AAAI)*, Vol. 37. 3100–3108.

[58] Xiangpeng Yang, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2024. DGL: Dynamic Global-Local Prompt Tuning for Text-Video Retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 6540–6548.

[59] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research* (2022).

[60] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Ryan Cotterell, Vicente Ordonez, and Kai-Wei Chang. 2019. Gender Bias in Contextualized Word Embeddings. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*. 629–634.

[61] Shuai Zhao, Linchao Zhu, Xiaohan Wang, and Yi Yang. 2022. Centerclip: Token clustering for efficient text-video retrieval. In *ACM SIGIR International Conference on Research and Development in Information Retrieva (SIGIR)*. 970–981.