Visual Intelligence

RESEARCH Open Access

Efficient text-to-video retrieval via multi-modal multi-tagger derived pre-screening

Yingjia Xu[1], Mengxia Wu[1], Zixin Guo[2], Min Cao[1*], Mang Ye[3] and Jorma Laaksonen[2]

Abstract

Text-to-video retrieval (TVR) has made significant progress with advances in vision and language representation learning. Most existing methods use real-valued and hash-based embeddings to represent the video and text, allowing retrieval by computing their similarities. However, these methods are often inefficient for large volumes of video, and require significant storage and computing resources. In this work, we present a plug-and-play multi-modal multi-tagger-driven pre-screening framework, which pre-screens a substantial number of videos before applying any TVR algorithms, thereby efficiently reducing the search space of videos. We predict discrete semantic tags for video and text with our proposed multi-modal multi-tagger module, and then leverage an inverted index for space-efficient and fast tag matching to filter out irrelevant videos. To avoid filtering out relevant videos for text queries due to inconsistent tags, we utilize contrastive learning to align video and text embeddings, which are then fed into a shared multi-tag head. Extensive experimental results demonstrate that our proposed method significantly accelerates the TVR process while maintaining high retrieval accuracy on various TVR datasets.

Keywords: Text-to-video retrieval (TVR), Inverted index, Pre-screening, Contrastive learning (CL)

1 Introduction

With the explosive growth of video data on the Internet, the efficient retrieval of relevant videos from massive videos based on textual queries, referred to as text-to-video retrieval (TVR) [1–3], has become increasingly significant in people's daily lives. Recently, vision-language pre-training technology [4–6] has flourished, and TVR has experienced remarkable improvement in accuracy [7, 8]. However, the processing of video data, which typically contain abundant temporal information, is significantly more time-consuming than that of image data. This presents substantial challenges in achieving high-efficiency TVR and limiting its real-world application. Therefore, developing a highly efficient and effective framework that caters to the practical needs of TVR is crucial.

Existing TVR methods explore a latent semantic space to represent both video galleries and query text and then compute similarities between their representations, enabling the identification of semantically relevant videos for the given query text. These methods can be broadly grouped into two categories: real-valued methods and Hash-based methods. Real-valued methods [9–12] represent data in a continuous semantic space, and the similarity between video and text is computed either through a simple inner product between their representations (known as a two-stream approach) or by designing intricate interaction modules for deeper representation fusion across modalities (known as a one-stream approach). Hash-based methods [13–16] encode the data into a discrete representation space with binary Hash codes and subsequently compute the Hamming distance to measure similarities. The former typically achieves superior accuracy when

*Correspondence: mcao@suda.edu.cn
[1]School of Computer Science & Technology, Soochow University, Suzhou, 215006, China
Full list of author information is available at the end of the article

Springer

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

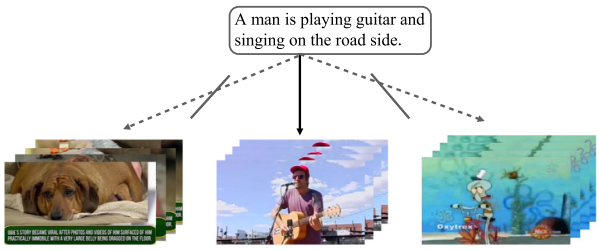


Figure 1 Explanation of the rationality of pre-screening prior to text-to-video retrieval (TVR) algorithms. The videos with the dotted arrow are irrelevant to the text query at the semantic level and should be pre-screened out in advance. The frames are sourced from the MSR-VTT [24] dataset

the real-valued method is compared with the Hash-based method. This is because the discrete representation space used in the Hash-based method tends to damage semantic information, thus affecting performance, whereas the continuous space in the real-valued method preserves it. Nevertheless, the real-valued method has certain drawbacks in terms of efficiency, particularly due to the more intricate computations involved in floating-point operations than in Hash value calculations, which can be more time-consuming. Furthermore, the interaction across modalities in its one-stream approach further exacerbates online time consumption.

In pursuit of superior performance and high efficiency, researchers have explored advancements in real value-based methods. Some studies [17–20] specifically focus on minimizing redundant content within videos to expedite video processing. Others [21–23] develop a two-step strategy (denoted as two-then-one): first, a large number of videos that have low similarity with the query text via similarity measurement in the two-stream approach are screened out, and then cross-modal interaction computations are performed in the one-stream approach, thus accelerating TVR. However, the screening process typically incurs a search time complexity of O(N), where N represents the number of videos. This poses a significant hindrance to practical applications, which usually involve processing a massive quantity of videos (i.e., a large value of N).

To target more practical TVR, we introduce a plug-and-play multi-modal multi-tagger-driven pre-screening (M3T-PS) framework. Specifically, we convert both videos and texts into discrete semantic tags via tag prediction and then perform tag matching to screen out videos that are inconsistent with the query text based on their tags. This process is intuitive and logical. For instance, if the query text revolves around a person, there is no need to search through video galleries related to animals, as illustrated in Fig. 1. Tags effectively convey thematic information of data. Unlike traditional methods [25–27], which compute similarity in continuous representation space, resulting in linear screening time complexity, the proposed pre-screening framework operates within a discrete tag space. Coupled with the inverted index technique [28–30], it can achieve a querying time complexity of O(1) while demanding minimal storage space. Moreover, the framework is designed to be plug-and-play, allowing seamless integration with common TVR methods to increase their efficiency.

M3T-PS draws inspiration from advancements in efficient image-text retrieval [31] while introducing a unique challenge. The effectiveness of efficient cross-modal retrieval relies on tag prediction, where accurate pre-screening can be guaranteed only when consistent tags are present across semantically relevant video and text. However, achieving such consistency in TVR poses a challenge because of its pronounced information asymmetry compared with image-text retrieval. Text often concisely summarizes key information in videos, whereas video contains rich content spread across frames, unlike single image. Consequently, directly predicting tags for video and text separately, similar to the method in efficient image-text retrieval [31], typically fails to produce consistent tags. In response, we develop the unified multi-modal multi-tagger module in M3T-PS, which fosters more consistent tags across modalities, as demonstrated in Sect. 3.1.

We conclude our contributions as follows:

1) With the aim of optimizing TVR efficiency while preserving performance, we propose a straightforward yet effective multi-modal multi-tagger-driven pre-screening called the M3T-PS framework.
2) To increase the framework's effectiveness, we introduce a multi-modal multi-tagger module to strengthen the consistency of the predicted tags between semantically relevant videos and text.
3) The proposed framework is designed to be plug-and-play, allowing for easy integration into existing TVR methods. Through extensive experiments conducted on four benchmarks, we demonstrate that this framework seamlessly enhances the efficiency of these methods while incurring minimal costs.

2 Related works

TVR aims to search for the most relevant video related to the semantics in a given sentence [17, 19, 20]. The research efforts are categorized into two groups: 1) real-valued methods, and 2) Hash-based methods. These research directions are outlined as follows.

Real-valued methods Real-valued methods [9–12] focus on finely capturing and aligning information between video and text. Wang et al. [32] integrated local alignment via transformer decoders and a captioning head for fine-grained word-level supervision, enhancing representation learning and reducing the cross-modal gap. Wu et al. [33] employed an enhanced self-attention mechanism within a transformer architecture to effectively fuse video and text features, and introduced combined similarity and consistency tasks to optimize the shared semantic space of text and video. Ge et al. [34] utilized contrastive learning to capture multi-grained feature contrasts both between and within text and video modalities, and designed an attention module to integrate these diverse similarities into a final similarity score. These approaches often necessitate complex architectures and substantial computational resources while significantly enhancing retrieval accuracy. This complexity further limits their practical application because of efficiency concerns.

Our M3T-PS aligns with real-valued methods in relying on high-dimensional, fine-grained feature representations for retrieval tasks. However, we also introduce a pre-screening mechanism that reduces the video search space, significantly enhancing the efficiency bypassing exhaustive similarity calculations, as performed in real-valued methods.

Hash-based methods Hash-based TVR methods [13–16] have demonstrated advantages in terms of low storage requirements and high computational efficiency. These methods aim to map high-dimensional data of text and video, into a common low-dimensional binary Hash space, making the retrieval process fast and storage-efficient. Li et al. [16] utilized weakly-supervised tags to learn image binary features, combining Hash learning, semantic mining, and data structure discovery to enhance retrieval and feature learning compatibility. Wang et al. [15] used high-level semantic labels and low-level features to preserve fine-grained similarities with sparse Hash codes, improving retrieval performance for large datasets through discrete optimization. Huo et al. [13] introduced a Hashing framework to capture global information and embed multi-modal, multi-label data into a discrete space, enhancing semantic relation preservation. Despite the advantages in terms of speed and efficiency, these approaches often result in compromises in retrieval accuracy.

Our M3T-PS also employs a discrete label space that boosts storage and query efficiency akin to that of Hash-based methods, while it preserves more required semantic details through consistency encouragement and further enhances retrieval accuracy by integrating real-valued methods after pre-screening.

3 Method

TVR aims to retrieve a ranked list of videos from a video gallery containing N videos, given a text query. Existing TVR methods suffer from the time complexity of $O(N)$, implying that the computational time increases linearly with the number of videos. This inefficiency hinders the real-world applications of such TVR methods where datasets often comprise a tremendous number of videos. To address this inefficiency, we introduce the M3T-PS, a plug-and-play framework designed to pre-screen the video gallery, reducing its volume from N to M ($M \ll N$) before applying any standard TVR methods [35–37].

Figure 2 illustrates the two key components in our proposed M3T-PS: tag prediction and tag matching. In Sect. 3.1, we introduce the tag prediction, where the text query and videos in the gallery are mapped as discrete semantic tags via our proposed multi-modal multi-tagger module. In Sect. 3.2, we present the tag matching, where the video gallery is effectively pre-screened to retain videos that share overlapping tags with the text queries. In Sect. 3.4, we comprehensively analyze the complexity of M3T-PS.

3.1 Tag prediction

The prior study on image-text retrieval [31] treats tag prediction as a multi-classification task, integrating distinct image and text multi-tag modules into the existing retrieval network. However, when directly applied to TVR, this design often leads to inconsistent tags across modalities because of the substantial information gap between video and text compared with image-text retrieval scenarios. To improve tag consistency between semantically relevant video and text, we propose a multi-modal multi-tagger module that aligns video and text representations and predicts multiple tags on the aligned video and text modalities. Figure 3 demonstrates the multi-modal multi-tagger module, which is introduced as follows: 1) data preparation for training the module, 2) the module architecture with the training objectives, and 3) the inference process of the module.

Data preparation The multi-modal multi-tagger module is trained in a supervised manner to process video-text pairs and generate both video tags and text tags. During training, the video-text pairs are provided by the TVR datasets, while the corresponding tags for both video and text are not offered. We extract tags from the texts and

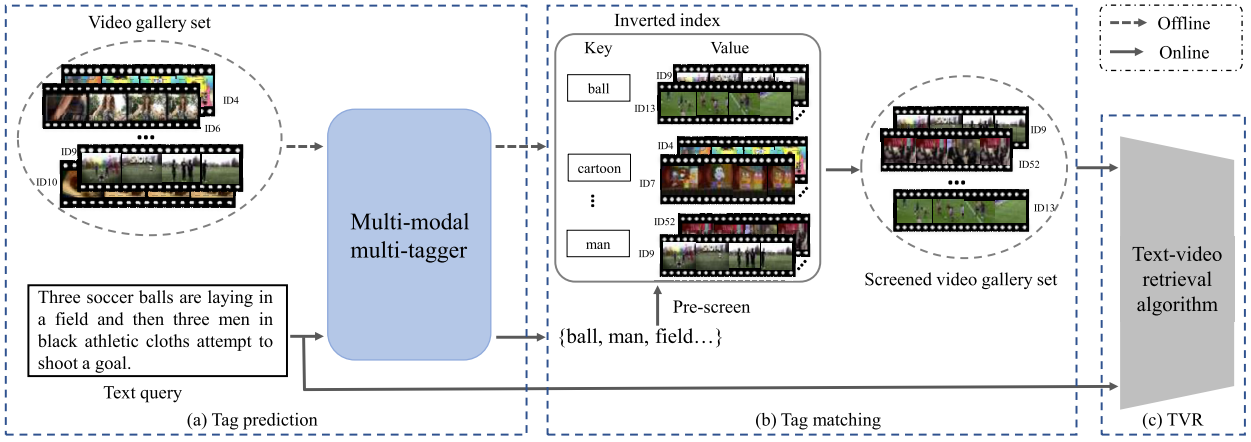


Figure 2 Illustration of the proposed multi-modal multi-tagger-driven pre-screening framework (M3T-PS) for plug-and-play pre-screening in TVR. It comprises two key components: tag prediction, which generates video tags offline and text tags online, and tag matching, which pre-screens videos having tags inconsistent with the texts. The proposed M3T-PS allows the seamlessly integration of any existing TVR algorithm. The frames are sourced from the MSR-VTT dataset

share them with their paired videos to create the tags required for the training process.

First, given a text T, we identify nouns as tags within the text and collect them into a tag set $\mathcal{A} = \{a_1, a_2, \ldots\}$, where a_i represents the i-th noun tag. After collecting the tag set for each text in the training dataset, we obtain a comprehensive set of noun tags, denoted as $\mathcal{K}$. However, noun tags within this set often exhibit an imbalanced distribution, with some nouns occurring far more frequently than others. We implement a selective retention mechanism for the set $\mathcal{K}$ to mitigate potential overfitting because of this imbalance. Specifically, we retain those that occur within a predefined range bounded by L and H, where L is the minimum frequency threshold and H denotes the number of top-frequency noun tags to exclude. Following this retention mechanism, a subset of $\mathcal{K}$ is obtained, denoted as $\mathcal{K}' = \{k_1, k_2, \ldots, k_K\} \in \mathcal{K}$, where K is the size of the set. The noun tags in the set $\mathcal{K}'$ serve as the tag labels for the subsequent tag prediction task. Similarly, for the extracted tag set $\mathcal{A}$ from a text, we refine it by retaining the tag a_i presented in the tag set $\mathcal{K}'$, and producing the refined tag set $\mathcal{A}' = \mathcal{A} \cap \mathcal{K}'$. Finally, each text T and its relevant video V in the training dataset are paired with the refined tag set $\mathcal{A}'$, resulting in the pairs $\{T, \mathcal{A}'\}$ and $\{V, \mathcal{A}'\}$.

We use the natural language toolkit (NLTK) [38] to extract nouns from the text as tags. As an alternative, we also experiment with Qwen [39] for noun extraction. Compared with the use of using Qwen, the use of NLTK for noun extraction yield superior results. The comparison is illustrated in Sect. 4.3.

Module architecture Figure 3 presents the multi-modal multi-tagger module. This module contains a video encoder and a text encoder for extracting video and text features, respectively. These features are subsequently processed through a parameter-shared multi-tag head, which predicts tags for both modalities.

Concretely, we employ a 12-layer visual transformer ViT-B/16 [40] as the video encoder. When given a video V, we first uniformly sample C frames from it, simplifying the video contents into $V = \{F_1, F_2, \ldots, F_C\}$. Each frame F_i is then processed by ViT-B/16, resulting in a representation $f_{F_i} = \{f_{F_i}^{\text{cls}}, f_{F_i}^1, f_{F_i}^2, \ldots, f_{F_i}^P\}$, where $f_{F_i}^{\text{cls}} \in \mathbb{R}^D$ is the global representation of the frame F_i with an input dimension of D, and P is the number of frame patches. Finally, we compute the video representation f_V by averaging the global representations of all sampled frames.

$$f_V = \frac{1}{C}\sum_{i=1}^{C} f_{F_i}^{\text{cls}}, \quad f_V \in \mathbb{R}^D. \tag{1}$$

For the text encoder, we use a 9-layer transformer, which transforms a text input T into the set of representations $g_T = \{f_T^{\text{cls}}, f_T^1, f_T^2, \ldots, f_T^S\}$, where $f_T^{\text{cls}} \in \mathbb{R}^D$ is the global representation of text and S is the token length of the text. Notably, we use the global text representation f_T^{cls} as the text representation f_T, denoted as $f_T = f_T^{\text{cls}}$. The parameter-shared multi-tag head consists of a 1-layer linear layer with a weight matrix $\boldsymbol{W} \in \mathbb{R}^{D\times K}$ and a bias vector $\boldsymbol{b} \in \mathbb{R}^K$, where K is the size of the tag set $\mathcal{K}'$. The video and text representations are fed into the multi-tag head to obtain their respective multi-tag probabilities p_V and p_T, respectively. This process is formulated as

$$p_V = f_V \boldsymbol{W} + \boldsymbol{b}, \tag{2}$$

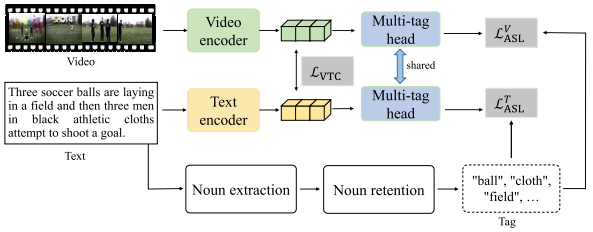


Figure 3 Illustration of the proposed multi-model multi-tagger module. Video and text embeddings are first aligned via video-text contrastive (VTC) loss, denoted as $\mathcal{L}_{\text{VTC}}$. These embeddings are then passed through a shared multi-tagging head, which is optimized via asymmetric loss (ASL). The ASL losses for text and video are represented as $\mathcal{L}^T_{\text{ASL}}$ and $\mathcal{L}^V_{\text{ASL}}$, respectively

$$p_T = f_T \boldsymbol{W} + \boldsymbol{b}. \tag{3}$$

Training objectives The training of multi-modal multi-tagger module is guided by three primary training objectives: video multi-tag loss $\mathcal{L}^V_{\text{ASL}}$, text multi-tag loss $\mathcal{L}^T_{\text{ASL}}$, and video-text contrastive (VTC) loss $\mathcal{L}_{\text{VTC}}$. The objective $\mathcal{L}$ is defined as

$$\mathcal{L} = \mathcal{L}^V_{\text{ASL}} + \mathcal{L}^T_{\text{ASL}} + \mathcal{L}_{\text{VTC}}, \tag{4}$$

where all loss terms contribute equally to the total loss of $\mathcal{L}$. The individual loss terms are detailed below.

The video and text multi-tag losses align the predicted tag probabilities with ground-truth tags. Specifically, we adopt asymmetric loss (ASL) [41] for highly imbalanced frequency of tags. Taking the video multi-tag as an example, its ASL loss is defined as

$$\mathcal{L}^V_{\text{ASL}} = \sum_{k=1}^{K} -y_k \mathcal{L}^+_V - (1-y_k)\mathcal{L}^-_V, \tag{5}$$

where $y_k = 1$ if the k-th tag of the tag list $\mathcal{K}'$ is in the video tag set $\mathcal{A}'$, otherwise $y_k = 0$, and

$$\begin{cases} \mathcal{L}^+_V = (1-p^k_V)^{\alpha_+} \log p^k_V, \\ \mathcal{L}^-_V = (\tilde{p}^k_V)^{\alpha_-} \log(1-\tilde{p}^k_V), \end{cases} \tag{6}$$

where p^k_V is the predicted video multi-tag probability for the k-th tag, and $\tilde{p}^k_V = \max(p^k_V - \delta, 0)$ is a shifted multi-tag probability. The negative sample is excluded when $\tilde{p}^k_V \le \delta$ and δ is a predefined scalar between 0 and 1. α_+ and α_- are the positive and negative focusing parameters, respectively. With a high α_+ (or α_-), the contribution from easy positives (or easy negatives) with $p^k_V \gg 0.5$ (or $p^k_V \ll 0.5$) to the loss is weakened, leading to focus more on more challenging samples during training. Similarly, we compute the multi-tag loss for the text, denoted as $\mathcal{L}^T_{\text{ASL}}$.

Video-text contrastive loss aligns video and text representations, enhancing the ability of video and text encoders to capture consistent semantics across modalities. Additionally, it aids the multi-tag head in generating consistent tags for both video and text by using these aligned representations as inputs. The loss $\mathcal{L}_{\text{VTC}}$ is formulated as

$$\mathcal{L}_{\text{VTC}} = \frac{1}{2}(\mathcal{L}_{\text{t2v}} + \mathcal{L}_{\text{v2t}}), \tag{7}$$

where $\mathcal{L}_{\text{t2v}}$ and $\mathcal{L}_{\text{v2t}}$ represent text-to-video and video-to-text contrastive loss, respectively, formulated as

$$\mathcal{L}_{\text{t2v}} = -\frac{1}{B}\sum_{i}^{B} \log \frac{e^{f^i_T \cdot f^i_V/\tau}}{\sum_{j=1}^{B} e^{f^i_T \cdot f^j_V/\tau}}, \tag{8}$$

$$\mathcal{L}_{\text{v2t}} = -\frac{1}{B}\sum_{i}^{B} \log \frac{e^{f^i_V \cdot f^i_T/\tau}}{\sum_{j=1}^{B} e^{f^i_V \cdot f^j_T/\tau}}, \tag{9}$$

where B is the batch size and τ is a learnable temperature parameter.

Inference For each video V in the gallery, we employ the multi-modal multi-tagger module to produce its multi-tag probability. The tags with the top-R_V probabilities in p_V are selected as final video tags for the subsequent matching process, represented as $\boldsymbol{G}_V = \{G^1_V, G^2_V, \ldots, G^{R_V}_V\}$. Similarly, the text query undergoes tag prediction by the the multi-modal multi-tagger module, resulting in $\boldsymbol{G}_T = \{G^1_T, G^2_T, \ldots, G^{R_T}_T\}$, where R_T denotes the number of text tags.

3.2 Tag matching

Following the tag prediction process, which predicts discrete tags for both videos and the text query, the tag matching process aims to pre-screen videos from the gallery that share no overlapping tags with the query text. To this end, a straightforward procedure is involved to traverse each video in the gallery and compare its tags with those of the text query. Although effective, this procedure becomes inefficient for practical applications when the number of videos in the gallery, denoted as N, is massive, resulting in a significant time complexity of O(N) for traversing the gallery.

To address this inefficiency, we adopt the inverted index technique, inspired by its proven efficiency in search engines [29]. Technically, we apply a strategy for constructing an inverted index that links tags to the corresponding video IDs to which they are assigned. For a specified text query, we leverage the extracted tags to directly query the inverted index, yielding a list of video IDs linked to those tags. This results in an efficient O(1) time complexity for selecting the video IDs containing overlapping tags with the text query.

Figure 2 summarizes the tag matching process, which consists of two steps: inverted list creation and video pre-screen. These steps are detailed as follows:

1) Inverted index creation. An inverted index is constructed offline to map each tag to its corresponding video ID in the gallery. This process involves scanning all videos in the gallery, extracting their tags, and creating an index where each tag serves as a key pointing to a list of video ID as values associated with that tag. Since video tags are discrete and finite in number, this inverted index can be stored efficiently, occupying minimal space. This structure ensures that the index is quickly accessible, enabling rapid video searches in response to text queries.
2) Video pre-screen. When performing a video pre-screen via a text query, we extract tags from the query and use these tags to look up the inverted index. For each tag in the query, the inverted index returns a list of video IDs that are associated with that tag. We then merge the lists of video IDs from all the tags to create a set of unique video IDs that have overlapping tags with the query. This pre-screened set of videos forms a candidate gallery, which is then used as input, along with the query text, for subsequent TVR algorithms.

3.3 TVR algorithm with M3T-PS

Figure 2 illustrates the integration of our proposed M3T-PS with the TVR algorithm. Specifically, M3T-PS employs a multi-modal multi-tagger module to predict tags for each video in the gallery during offline processing. An inverted index, also created offline, maps these tags to their corresponding video IDs. For any incoming text query, its tags are predicted online via the multi-modal multi-tagger, followed by a quick match with video IDs that share overlapping tags by querying the inverted index. This matching process pre-screens irrelevant videos, leaving a smaller set of retained videos than the entire gallery does. Standard TVR algorithms are then applied to these retained videos, rather than the entire gallery, enabling smaller computational load.

3.4 Complexity analysis and discuss

By employing a pre-screening process before TVR methods, we effectively filter out videos semantically irrelevant to the text query in the gallery, improving the efficiency of TVR methods. Although M3T-PS introduces some additional time and space consumption inevitably, these overheads are well justified given the notable efficiency improvement for TVR methods. We illustrate its time and space complexities in Sect. 4.

The online time consumption of M3T-PS encompasses tag prediction for text queries and video pre-screen during tag matching. Text query processing in tag prediction is trivially fast compared with video processing, making its time consumption negligible. The time complexity of M3T-PS is primarily dominated by video pre-screen, including the querying and merging steps. Querying operates in constant time O(1) because of the use of inverted index that maps tags to video IDs directly. The merging step, which consolidates the video IDs from the query results into a candidate set of size M, is fast enough to be considered negligible. Therefore, the overall time complexity of using M3T-PS followed by TVR can be expressed as $O(1) + O(M)$, where M is significantly smaller than N ($M \ll N$), the number of videos in the gallery. This combined time complexity is much less than $O(N)$, the time complexity of typical TVR methods operating directly on the full gallery.

In terms of space complexity, the M3T-PS comprises a fixed size with 174.2 M parameters. The only additional space requirement that scales with the size of the video gallery is the inverted index, which is represented by ID numbers and highly storage-efficient. Consequently, the additional space consumption for M3T-PS is minimal, even as the number of videos in the gallery increases.

4 Experiments

4.1 Experimental setting

Dataset We evaluate the effectiveness of the proposed M3T-PS on four common TVR datasets: MSR-VTT [24], MSVD [42], VATEX[1] [43], and DiDeMo [44]. Following Ref. [45], we conduct a sentence-to-video retrieval task for the first three datasets and conduct a paragraph-to-video retrieval task by concatenating all captions of one video as a single query for the last dataset.

Evaluation metrics Following Ref. [35], we adopt recall at rank k (R@k), which higher is better, and median rank (MdR), which lower is better, as retrieval performance metrics. We report R@1, R@5, and R@10. In addition, to evaluate the effectiveness of the proposed M3T-PS, we report the online running time of all queries and the speedup ratio for retrieval.

Implementation details All the experiments are conducted on four GeForce RTX 3090 GPUs with 24 GB of memory. For the multimodal multi-tagger module, we initialize the parameters of the video and text encoders via the public checkpoint from UMT-B/25M [37]. The initial learning rates are 1.0×10^{-5} for the two unimodal encoders, and 5.0×10^{-4} for the multi-tag head. We train the module for 5 epochs with a batch size of 32, using the Adam optimizer [46] and a cosine schedule [47] strategy for learning rate decay. For ASL, we set $\alpha_+ = 0$, $\alpha_- = 3$, and

[1] Because of the unavailability of some video clips, we use splits of 21,385, 2487, and 5151 for training, validation, and testing, respectively.

Table 1 Comparison between our proposed multi-modal multi-tagger-driven pre-screening (M3T-PS) framework and baseline methods on text-to-video retrieval (TVR) tasks on the MSR-VTT [24] and MSVD [42] datasets. The symbol x indicates the speedup ratio of the current line compared with the one above it. The symbol + indicates the additional overhead required for external procedures in the current row, e.g., text tag prediction and matching for pre-screening videos in M3T-PS, relative to the one above it. The symbol ↑ indicates that the higher the corresponding metric, the better the model performs. Conversely, the symbol ↓ indicates that the lower the corresponding metric, the better the model performs. MdR denotes median rank and Para (M) denotes the model parameters

Genre	Method	MSR-VTT						MSVD						Para (M)
		R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	Time (s) ↓	Speedup(%)	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	Time (s) ↓	Speedup(%)	
One-stream	X-pool [36]	46.6	73.0	82.9	2.0	1244.0	—	46.8	77.3	85.6	2.0	33093.4	—	152.5
	+ M3T-PS(Ours)	46.5	73.0	82.9	2.0	+6.2	1.8x	46.7	77.0	85.1	2.0	+28.0	2.0x	+174.2
Two-stream	CLIP4Clip [35]	43.0	70.5	79.8	2.0	18.8	—	46.0	75.7	84.5	2.0	417.5	—	151.2
	+ M3T-PS	43.2	70.9	79.7	2.0	+6.2	1.8x	45.9	75.3	83.9	2.0	+28.0	2.0x	+174.2
	X-CLIP [48]	47.0	73.5	82.7	2.0	21.0	—	46.8	76.1	84.9	2.0	542.6	—	164.4
	+ M3T-PS	46.9	73.5	82.8	2.0	+6.2	1.8x	46.7	75.9	84.5	2.0	+28.0	2.0x	+174.2
Two-then-one	BLIP [6]	51.7	76.6	84.7	1.0	748.9	—	47.3	77.1	85.2	2.0	27302.3	—	223.4
	+ M3T-PS	52.0	77.1	84.9	1.0	+6.2	1.8x	47.3	76.9	84.8	2.0	+28.0	2.0x	+174.2
	UMT-B [37]	51.0	76.5	84.2	1.0	262.5	—	50.8	79.7	86.2	1.0	9570.2	—	202.4
	+ M3T-PS	51.1	76.3	84.1	1.0	+6.2	1.8x	50.6	79.1	85.5	1.0	+28.0	2.0x	+174.2

δ = 0.05. The frame length is set to 8, and the maximum text token lengths are 32 for MSR-VTT, MSVD, and VATEX, and 64 for DiDeMo. In terms of tag annotation, we retain tags with the following frequency thresholds: L=100, H=1 for MSR-VTT; L=50, H=5 for MSVD; L=100, H=20 for VATEX; and L=20, H=15 for DiDeMo, resulting in tag lists with lengths of K = 414, 257, 663, and 345, respectively. During inference, we set R_V = 10 and R_T = 3 for video and text tag predictions, respectively.

Baselines The proposed M3T-PS is a plug-and-play framework that can be seamlessly integrated into common TVR baselines, categorized by one-stream approach, two-stream approach, and two-then-one, for pre-screening a video gallery and accelerating the TVR process. Specifically, the one-stream approach performs cross-modal attention mechanisms to directly integrate video and text features during the encoding process, enabling to predict their similarity score directly, including the X-pool [36]. The two-stream approach independently encodes video and text into dense embeddings without text-video integrations and then adopts similarity functions, including CLIP4Clip [35] and X-CLIP [48]. The two-then-one approach combines the two-stream and one-stream approaches, where the two-stream approach is employed to filter out substantial irrelevant videos, followed by a one-stream approach to integrate the texts with the remaining videos, balancing computational efficiency and retrieval effectiveness. The baselines of such methods include BLIP [6] and UMT-B [37].

4.2 Main results

Tables 1 and 2 present the TVR results of incorporating M3T-PS with the aforementioned baselines, compared with using the baselines alone. This finding demonstrates that the incorporation of M3T-PS consistently accelerates the retrieval process across all baselines, while not only maintaining but also in some cases, improving retrieval performance. For instance, when applied on MSVD, integrating M3T-PS can reduce the search scope to approximately half of the video gallery for each text query, eliminating the necessity of traversing the entire gallery. This results in a theoretical speedup ratio of 2.0 across all baseline methods on MSVD. Moreover, the R@1 accuracy is maintained when M3T-PS is applied with BLIP, further confirming the effectiveness of M3T-PS.

Table 3 presents a detailed comparison of BLIP utilizing the two-then-one and a variant, BLIP_{all}, which applies a one-stream approach involving cross-attention between the text query and all videos. Compared to BLIP_{all}, BLIP results in a decrease in TVR accuracy, e.g. decreasing from 52.1 to 51.7 in R@1 on the MSR-VTT dataset. Despite this reduction in accuracy, BLIP achieves a significant 7.8 times speedup on MSR-VTT, without the necessity for cross-attention between the text query and all videos. Furthermore, compared with BLIP_{all}, the additional steps introduced by BLIP, i.e. video-text similarity scoring and ranking for video screening, take only 0.1 s, highlighting the efficiency of the two-then-one over the one-stream method. Once M3T-PS is integrated with BLIP, an additional 1.8 times acceleration is achieved on MSR-VTT, because many irrelevant videos are screened out prior to using BLIP. Notably, the retrieval accuracy of M3T-PS is greater than that of BLIP alone in terms of R@1 on the MSR-VTT dataset. These gains are realized with tiny overhead, as evidenced by the reported time required for external procedures in M3T-PS, i.e., text tag prediction and matching processes for pre-screening videos, as shown in Tables 1-3. Furthermore, the M3T-PS module has a fixed parameter count of 174.2 M, as reported in Tables 1-2, ensuring that resource consumption remains manageable even as dataset size scales. These results underscore the

Table 2 Comparison between our proposed M3T-PS and baseline methods on TVR tasks on the VATEX [43] and DiDeMo [44] datasets

Genre	Method	VATEX						DiDeMo						Para (M)
		R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	Time (s) ↓	Speedup(%)	R@1 ↑	R@5 ↑	R@10 ↑	MdR ↓	Time (s) ↓	Speedup(%)	
One-stream	X-pool	51.3	84.8	92.0	1.0	35,176.8	—	43.7	70.8	79.5	2.0	1521.3	—	152.5
	+ M3T-PS	51.2	84.7	91.8	1.0	+25.1	2.3x	43.6	70.8	79.2	2.0	+6.4	1.6x	+174.2
Two-stream	CLIP4Clip	50.6	83.5	90.9	1.0	396.5	—	41.1	68.1	77.8	2.0	20.6	—	151.2
	+ M3T-PS	50.5	83.4	90.8	1.0	+25.1	2.3x	40.8	68.1	77.9	2.0	+6.4	1.6x	+174.2
	X-CLIP	52.2	85.1	92.0	1.0	452.9	—	42.3	70.6	80.7	2.0	24.7	—	164.4
	+ M3T-PS	52.2	85.0	91.8	1.0	+25.1	2.3x	44.2	72.1	81.8	2.0	+6.4	1.6x	+174.2
Two-then-one	BLIP	59.7	88.4	94.0	1.0	27,001.0	—	56.8	81.7	87.1	1.0	897.6	—	223.4
	+ M3T-PS	59.6	88.1	93.8	1.0	+25.1	2.3x	56.5	81.3	86.7	1.0	+6.4	1.6x	+174.2
	UMT-B	61.7	90.3	95.4	1.0	13,249.1	—	61.6	86.8	91.5	1.0	295.4	—	202.4
	+ M3T-PS	61.6	90.1	95.1	1.0	+25.1	2.3x	61.9	86.6	91.2	1.0	+6.4	1.6x	+174.2

Table 3 Comparison of retrieval accuracy and efficiency between BLIP and $BLIP_{all}$ with M3T-PS enhancement. The symbol xx indicates the speedup ratio of the current row compared with the one above it. $BLIP_{all}$ represents the BLIP utilizing the two-then-one and a variant. The symbol ++ indicates the additional time required for external procedures in the current row, relative to the one above it

	MSR-VTT						MSVD						Didemo					
	R@1	R@5	R@10	MdR	Time (s)	Speedup	R@1	R@5	R@10	MdR	Time (s)	Speedup	R@1	R@5	R@10	MdR	Time (s)	Speedup
$BLIP_{all}$	52.1	77.1	84.8	1.0	5747.0	—	47.6	77.2	85.1	2.0	154,576.0	—	57.6	83.5	89.8	1.0	17,180.0	—
BLIP	51.7	76.6	84.7	1.0	+0.1	7.8x	47.3	77.1	85.2	2.0	+1.5	5.2x	56.8	81.7	87.1	1.0	+0.5	8.0x
+M3T-PS	52.0	77.1	84.9	1.0	++6.2	1.8xx	47.3	76.9	84.8	2.0	++28.0	2.0xx	56.5	81.3	86.7	1.0	++6.4	1.6xx

Table 4 Ablation study on structural design. SHARE stands for a shared multi-tag head. Recall indicates the rate at which relevant videos are correctly identified among the retained videos. ASL and VTC denote asymmetric loss and video-text contrastive loss, respectively. AvgLen denotes the average length of the retained videos per text query. The symbol ✓represents an applied setting

ID	ASL	VTC	SHARE	Retrieval				Pre-screen	
				R@1	R@5	R@10	MdR	Recall	AvgLen
1	✓			50.5	75.6	83.0	1.0	98.1	520
2	✓	✓		51.2	76.3	84.1	1.0	99.0	578
3	✓		✓	50.9	75.7	83.3	1.0	98.6	520
4	✓	✓	✓	51.1	76.3	84.1	1.0	99.0	569

versatility of the M3T-PS framework, particularly in scenarios where multiple retrieval methods are employed, highlighting its potential in TVR optimization.

4.3 Ablation studies

We examine the impact of the following four components on our proposed M3T-PS model: 1) the structural design of the multi-modal multi-tagger module, 2) various loss functions used in training this module, 3) different tag types and tag extractors in training, and 4) different strategies for generating video and text in testing. Unless otherwise noted, the study is conducted on the MSR-VTT dataset, which involves the use of M3T-PS in pre-screening videos irrelevant to the text query, followed by UMT-B. The evaluation covers retrieval metrics that indicate accuracy, along with the pre-screen metric, recall of the videos retained by M3T-PS and the average number of videos retained per text query.

Effect of structural design choices Table 4 shows the effects of ASL and VTC losses, and the impact of sharing the multi-tag head during the training of the multi-modal multi-tagger module in M3T-PS. The results show that, irrespective of whether the multi-tag head is shared, incorporating the ASL and VTC loss terms yields the highest recall score of 99.0. This suggests that using both loss terms successfully retains the most relevant videos for each text query, although with an increased average length of the retained video gallery from 520 videos. When the multi-tag head is shared, the retained gallery's size can be reduced from 578 to 569 videos, while maintaining its recall rate. This demonstrates the efficiency of sharing the multi-tag head in retaining relevant videos for each text query. In terms of the retrieval accuracy metrics, the results indicate that employing ASL and VTC losses enhances the accuracy, outperforming scenarios without these losses. This highlights the efficacy of integrating ASL and VTC loss components, along with a shared multi-tag head, to achieve a balance between superior retrieval accuracy and maintaining a manageable video length.

Effect of multi-tag loss variations Table 5 summarizes the comparison of utilizing various loss terms, including cross-entropy loss and focal loss [49], to replace the ASL loss for both texts and videos. This finding demonstrates that employing the ASL loss outperforms using the other two losses on most metrics, while also resulting in the shortest average length of the retained video gallery. This indicates that employing the ASL loss term leads to a more accurate tag prediction for both video and text.

Table 5 Ablation study on multi-tag loss variations

Loss type	Retrieval				Pre-screen	
	R@1	R@5	R@10	MdR	Recall	AvgLen
Cross entropy	50.7	76.1	84.0	1.0	98.8	576
Focal loss	50.8	76.2	84.0	1.0	99.1	639
ASL	51.1	76.3	84.1	1.0	99.0	569

Effect of filtering intensity Table 6 explores how varying the filtering intensity, which is determined by the number of selected tags R_V for videos and R_T for texts, affects retrieval accuracy and performance. The results indicate that increasing R_V or R_T allows more relevant gallery samples to pass through the pre-screening stage, which leads to higher recall scores. However, this improvement has minimal impact on the final retrieval performance and comes at the expense of reducing efficiency, as reflected in the increased average length of the retained gallery. To achieve a balance between accuracy and efficiency, we set $R_V = 10$ and $R_T = 3$ in experiments, ensuring an optimal trade-off.

Effect of tag type Table 7 compares the impact of using different types of tags, specifically nouns and verbs, extracted from texts with NLTK, for training the multi-modal multi-tagger module. The results demonstrate that using nouns as tags tends to yield better overall performance in terms of retrieval accuracy, than does using verbs. This suggests that nouns, being more concrete and descriptive, are more effective in capturing the semantic content required for retrieval.

Effect of tag extractor Table 7 explores the effect of using different tag extraction tools, specifically NLTK and Qwen, for extracting tags from texts in module training. The results show that NLTK performs comparably to Qwen across various retrieval metrics, whereas resulting in a shorter average length of the retained video gallery. This indicates that NLTK remains an efficient choice for tag extraction.

Effect of tag generation strategies Table 7 explores the effects of different strategies for generating both text and video tags during testing. Specifically, for both video and text tags, two strategies are employed. One strategy relies on the trained multi-modal multi-tagger for predicting tags. To generate text tags, another strategy uses word part-of-speech tagging with NLTK. To generate video tags, a well-pretrained object detector [50] is alternatively used. It implies that utilizing the trained multi-modal multi-tagger to generate video and text tags is more efficient than the other strategies. This is proven by the highest retrieval accuracy and the small average length of the retained videos produced by this strategy.

Table 6 Ablation study on filtering intensity. R_V denotes the number of selected tags for videos, reflecting video filtering intensity, and R_T represents the number of selected tags for texts, indicating text filtering intensity

R_V/R_T	Retrieval				Pre-screen	
	R@1	R@5	R@10	MdR	Recall	AvgLen
Varying R_V and fixing $R_T = 3$.						
7	50.3	76.0	83.5	1.0	97.6	492
8	50.7	76.1	83.9	1.0	98.2	520
9	51.0	76.3	84.0	1.0	98.8	544
10	51.1	76.3	84.1	1.0	99.0	569
11	51.1	76.2	83.9	1.0	99.1	589
12	51.0	76.2	83.9	1.0	99.2	608
13	50.9	76.2	84.0	1.0	99.4	627
Varying R_T and fixing $R_V = 10$.						
2	50.3	75.4	82.9	1.0	97.2	492
3	51.1	76.3	84.1	1.0	99.0	569
4	51.1	76.5	84.3	1.0	99.6	656
5	51.0	76.5	84.2	1.0	99.7	715

Effect of M3T-PS architecture setup We have compared our M3T-PS with a simplified setup, corresponding to ID 1 in Table 4, which removes the shared multi-tag head and contrastive loss, thereby simplifying the M3T-PS structure into independent classifiers for text and video. In this simplified setup, we initially maintain the temporal modeling method in M3T-PS for the video branch, which averages the frame information, as shown in Eq. (1). In Table 8, we further compare our M3T-PS architecture with simplified setups consisting of independent classifiers for text and video. Specifically, we explored alternative temporal modeling methods, replacing the averaging approach, with vanilla attention and multi-layer perceptron (MLP), respectively. Experimental results demonstrate that our M3T-PS architecture substantially improves tag prediction consistency, which enhances pre-screening and retrieval performance, without introducing additional parameters. This improvement helps maintain retrieval accuracy while achieving higher efficiency, and underscoring the advantages of our proposed tagger design over simpler baseline methods.

4.4 Practical applications

The model's ability to generalize to unseen data is a critical factor for its broad applicability in real-world scenarios. Inevitably, models trained on a specific dataset may struggle when confronted with new test data that exhibit domain discrepancies from the training data. To evaluate the domain adaptability of our proposed multi-modal multi-tagger module, we apply the module, which is trained on MSR-VTT, to the MSVD, VATEX, and DiDeMo datasets, and report the retrieval performance of our proposed M3T-PS in Table 9. When trained and tested on different datasets, our M3T-PS experiences a minor performance drop across various metrics compared with trained

Table 7 Ablation study on different tag configurations: tag types (nouns vs. verbs), tag extractors (natural language toolkit(NLTK) vs. Qwen [39]), and tag generation strategies for both text and video. Components are separated by dashed lines. CLS refers to classification performed by the multi-modal multi-tagger module, whereas TKN indicates tokenization performed with NLTK. In addition, OD represents object detection

Test				Train				Retrieval				Pre-screen	
Text tag		Video tag		Tag type		Tag extractor							
CLS	TKN	CLS	OD	Noun	Verb	NLTK	Qwen	R@1	R@5	R@10	MdR	Recall	AvgLen
✓		✓			✓	✓		51.1	76.2	83.8	1.0	98.7	690
✓		✓		✓		✓		51.1	76.3	84.1	1.0	99.0	569
✓		✓		✓			✓	50.8	76.4	83.7	1.0	99.3	632
✓		✓		✓		✓		51.1	76.3	84.1	1.0	99.0	569
	✓	✓		✓		✓		45.3	69.0	75.5	2.0	77.8	370
✓	✓	✓		✓		✓		51.0	76.1	84.1	1.0	99.3	604
✓		✓	✓	✓		✓		51.1	76.3	84.1	1.0	99.2	617
✓		✓		✓		✓		51.1	76.3	84.1	1.0	99.0	569

Table 8 Ablation study on M3T-PS architecture setup. Avg, attention, and MLP represent three temporal modeling methods based on the simplified setup, corresponding to averaging frame information, using vanilla attention mechanism, and employing multi-layer perceptron, respectively

	Retrieval				Pre-screen		Para (M)
	R@1	R@5	R@10	MdR	Recall	AvgLen	
Avg	50.5	75.6	83.0	1.0	98.1	520	174.2
Attention	50.4	75.3	82.9	1.0	97.8	534	175.5
MLP	50.2	75.4	83.1	1.0	98.2	532	176.8
Ours	51.1	76.3	84.1	1.0	99.0	569	174.2

Table 9 Impact of domain transfer on retrieval performance. The results compare models trained on MSR-VTT and applied to other datasets (denoted with an arrow, e.g., MSR-VTT → MSVD) versus models trained directly on the target datasets (e.g., MSVD → MSVD)

domain transfer	Retrieval				Pre-screen	
	R@1	R@5	R@10	MdR	Recall	AvgLen
MSVD → MSVD	50.5	79.0	85.5	1.0	98.4	332
MSR-VTT → MSVD	50.2	78.5	84.8	1.0	98.0	426
VATEX → VATEX	61.5	90.0	95.1	1.0	99.4	1073
MSR-VTT → VATEX	61.4	89.7	94.7	1.0	99.0	1622
DiDeMo → DiDeMo	61.8	86.5	91.2	1.0	99.4	640
MSR-VTT → DiDeMo	60.0	84.1	89.1	1.0	97.0	623

and tested with the same dataset, thereby underscoring the robustness of our approach. For example, when training on the MSR-VTT dataset and testing on the VATEX dataset, the R@1 metric decreases by only 0.1%, alongside a minimal decline in other pre-screening metrics, such as recall. Similar trends are observed when transferring to the MSVD dataset. Notably, when transferring from the MSR-VTT to DiDeMo dataset, it exhibits a modest performance drop, decreasing from 61.8 to 60.0 in R@1. This decrease can be attributed to differences in annotation style between the datasets, particularly in the length and detail of the video captions. The captions in the DiDeMo dataset are, on average, four times longer than those in the MSR-VTT dataset, introducing challenges in aligning tags across modalities and accounting for the observed drop in performance. Nevertheless, the generalization capability of our model remains solid across domains. Benefiting from the efficient tag matching strategy, which works even on a single overlapping tag, our proposed M3T-PS demonstrates robustness when applied across different domains.

5 Visualization

To demonstrate the effectiveness of our proposed M3T-PS, we present the results after pre-screening given text queries across four datasets. In Fig. 4, each dataset shows the screened out results for one representative text query, where videos semantically unrelated to the text query are effectively filtered out.

6 Limitations and discussion

We discuss the limitations of the proposed M3T-PS framework, which relies on tag consistency to achieve efficient pre-screening. This approach is sensitive to the information disparity between videos, which consist of multiple frames and contain far more data, and concise text descriptions. Although M3T-PS promotes alignment of the predicted tags across modalities, through the video-text contrastive loss and shared multi-tag head, and relaxes matching criteria to count any overlapping tag as a match, it remains influenced by this inherent data imbalance. Additionally, to balance efficiency and accuracy, we filter the predicted tags to retain only the top n with the highest confidence scores, which may also lead to occasional mismatches. Future enhancements could involve refining the keyword matching strategy to incorporate semantic similarity, thereby increasing pre-screening accuracy.

Moreover, M3T-PS lacks advanced temporal modeling, which is essential for capturing rich temporal dynamics in complex videos. Recent approaches, such as those proposed by Wang et al. [51] and Papalampidi et al. [52], effectively handle these dynamics through decoupled pretrain-

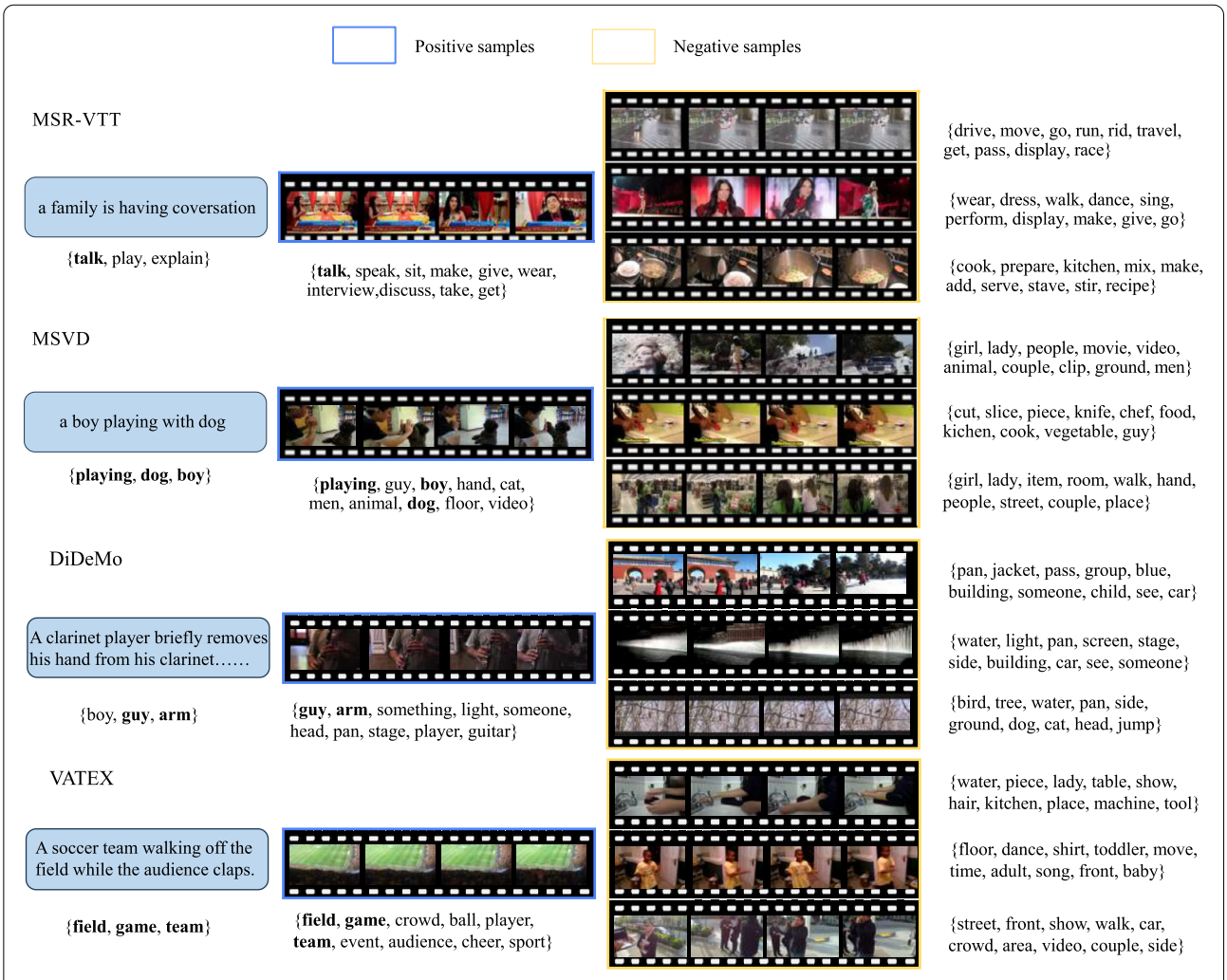


Figure 4 Visualization of M3T-PS pre-screening results for a single text query sampled from four different datasets. The tags associated with the text query and videos are shown in braces, with matching tags highlighted in bold. The frames are respectively sourced from the MSR-VTT, MSVD [42], DiDeMo [44], and VATEX [43] datasets, respectively

ing or memory-efficient spatiotemporal attention. Wang et al. [51] introduced a strategy that applied spatial features learned from image-text data to video-text tasks and then focused on temporal dynamics in video pretraining, achieving performance improvements across both image and video tasks. On the other hand, Papalampidi et al. [52] optimized long-video modeling by employing early spatiotemporal attention and efficient masking techniques. Both approaches highlight opportunities for potential advancements in M3T-PS. Integrating similar temporal modeling techniques further refine M3T-PS's capabilities for large-scale video retrieval tasks.

7 Conclusion

In this work, we investigate how to enhance the efficiency of the text-to-video retrieval called the TVR process. We propose the plug-and-play multi-modal multi-tagger-driven pre-screening framework termed as M3T-PS for filtering out videos irrelevant to the text query prior to employing any existing TVR algorithms. Our framework introduces a multi-modal multi-tagger module to predict discrete semantic tags for video and text, followed by an inverted index to quickly collect video IDs that share tags with the text query, significantly reducing the number of videos to search. Within training the multi-modal multi-tagger module, video and text embeddings are aligned through contrastive learning, and then fed into a shared multi-tag head, thereby enhancing the module's capacity of producing consistent tags for semantically relevant video and text queries. Extensive experiments on the TVR task demonstrate the effectiveness and efficiency of our

method in accelerating the retrieval process while maintaining retrieval accuracy.

Abbreviations

ASL, asymmetric loss; AvgLen, average length; CLS, classification; M3T-PS, multi-modal multi-tagger-driven pre-screening framework; MdR, median rank; MLP, multi-layer perceptron; NLTK, natural language toolkit; TVR, text-to-video retrieval; VTC loss, video-text contrastive loss.

Author contributions

All the authors contributed to the conceptualization and design of this study. Data collection, coding, and analysis were performed by YX and MW. The initial draft of the manuscript was finished by YX, MW, and ZG, and all the authors collaborated in enhancing the manuscript. All the authors reviewed and approved the final submitted version.

Funding

This work was supported by the National Natural Science Foundation of China (No. 62476188), the Open Projects Program of the State Key Laboratory of Multimodal Artificial Intelligence Systems, and the Academy of Finland in the USSEE Project (No. 345791).

Data availability

The datasets used in this study are publicly available at: MSR-VTT: https://drive.google.com/drive/folders/1LYVUCPRxpKMRjCSfB_Gz-ugQa88FqDu_; MSVD: https://drive.google.com/drive/folders/18EXLWvCCQMRBd7-n6uznBUHdP4uC6Q15; DiDeMo:https://pan.baidu.com/share/init?surl=Tsy9nb1hWzeXaZ4xr7qoTg&pwd=c842; VATEX: https://www.kaggle.com/datasets/khaledatef1/vatex0110/.

Declarations

Competing interests

The authors have no competing interests to declare that they are relevant to the content of this article.

Author details

[1]School of Computer Science & Technology, Soochow University, Suzhou, 215006, China. [2]Department of Computer Science, Aalto University, Espoo, 02150, Finland. [3]School of Computer Science, Wuhan University, Wuhan, 430072, China.

Received: 3 September 2024 Revised: 16 February 2025
Accepted: 17 February 2025 Published online: 06 March 2025

References

1. Gabeur, V., Sun, C., Alahari, K., & Schmid, C. (2020). Multi-modal transformer for video retrieval. In A. Vedaldi, H. Bischof, T. Brox, et al. (Eds.), *Proceedings of the 16th European conference on computer vision* (pp. 214–229). Cham: Springer.
2. Pereira, J. C., Coviello, E., Doyle, G., Rasiwasia, N., Lanckriet, G. R. G., Levy, R., et al. (2013). On the role of correlation and abstraction in cross-modal multimedia retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *36*(3), 521–535.
3. Chen, S., Zhao, Y., Jin, Q., & Wu, Q. (2020). Fine-grained video-text retrieval with hierarchical graph reasoning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10638–10647). Piscataway: IEEE.
4. Chen, F.-L., Zhang, D.-Z., Han, M.-L., Chen, X.-Y., Shi, J., Xu, S., et al. (2023). VLP: a survey on vision-language pre-training. *Machine Intelligence Research*, *20*(1), 38–56.
5. Yang, J., Duan, J., Tran, S., Xu, Y., Chanda, S., Chen, L., et al. (2022). Vision-language pre-training with triple contrastive learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 15671–15680). Piscataway: IEEE.
6. Li, J., Li, D., Xiong, C., & Hoi, S. C. H. (2022). BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In K. Chaudhuri, S. Jegelka, & L. Song (Eds.), *Proceedings of the 39th international conference on machine learning* (pp. 12888–12900). Retrieved July 17, 2024, from https://proceedings.mlr.press/v162/li22n.html.
7. Bain, M., Nagrani, A., Varol, G., & Zisserman, A. (2021). 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* (pp. 1728–1738). Piscataway: IEEE.
8. Lei, J., Li, L., Zhou, L., Gan, Z., Berg, T. L., Bansal, M., et al. (2021). Less is more: CLIPBERT for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 7331–7341). Piscataway: IEEE.
9. Liu, Y., Albanie, S., Nagrani, A., & Zisserman, A. (2019). Use what you have: video retrieval using representations from collaborative experts. In *Proceedings of the 30th British machine vision conference* (p. 279). Swansea: BMVA Press.
10. Li, X., Xu, C., Yang, G., Chen, Z., & Dong, J. (2019). W2VV++ fully deep learning for ad-hoc video search. In L. Amsaleg, B. Huet, M. A. Larson, et al. (Eds.), *Proceedings of the 27th ACM international conference on multimedia* (pp. 1786–1794). New York: ACM.
11. Yang, X., Dong, J., Cao, Y., Wang, X., Wang, M., & Chua, T.-S. (2020). Tree-augmented cross-modal encoding for complex-query video retrieval. In J. X. Huang, Y. Chang, X. Cheng, et al. (Eds.), *Proceedings of the 43rd international ACM SIGIR conference on research and development in information retrieval* (pp. 1339–1348). New York: ACM.
12. Dong, J., Li, X., & Snoek, C. G. M. (2018). Predicting visual features from text for image and video caption retrieval. *IEEE Transactions on Multimedia*, *20*(12), 3377–3388.
13. Huo, Y., Qin, Q., Dai, J., Wang, L., Zhang, W., Huang, L., et al. (2024). Deep semantic-aware proxy hashing for multi-label cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(1), 576–589.
14. Yu, J., Wu, X.-J., & Zhang, D. (2022). Unsupervised multi-modal hashing for cross-modal retrieval. *Cognitive Computation*, *14*(3), 1159–1171.
15. Wang, Y., Chen, Z.-D., Luo, X., & Xu, X.-S. (2022). A high-dimensional sparse hashing framework for cross-modal retrieval. *IEEE Transactions on Circuits and Systems for Video Technology*, *32*(12), 8822–8836.
16. Li, Z., Tang, J., Zhang, L., & Yang, J. (2020). Weakly-supervised semantic guided hashing for social image retrieval. *International Journal of Computer Vision*, *128*, 2265–2278.
17. Wu, X., Gao, C., Lin, Z., Wang, Z., Han, J., & Hu, S. (2022). RaP: redundancy-aware video-language pre-training for text-video retrieval. In Y. Goldberg, Z. Kozareva, & Y. Zhang (Eds.), *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 3036–3047). Stroudsburg: ACL.
18. Cao, M., Tang, H., Huang, J., Jin, P., Zhang, C., Liu, R., et al. (2024). Efficient text-video retrieval with sparse-and-correlated adapter. arXiv preprint. arXiv:2405.19465.
19. Li, P., Xie, C.-W., Zhao, L., Xie, H., Ge, J., Zheng, Y., et al. (2023). Progressive spatio-temporal prototype matching for text-video retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 4100–4110). Piscataway: IEEE.
20. Zhao, S., Zhu, L., Wang, X., & Yang, Y. (2022). Centerclip: token clustering for efficient text-video retrieval. In E. Amigó, P. Castells, J. Gonzalo, et al. (Eds.), *Proceedings of the 45th international ACM SIGIR conference on research and development in information retrieval* (pp. 970–981). New York: ACM.
21. Wang, Z., Sung, Y.-L., Cheng, F., Bertasius, G., & Bansal, M. (2023). Unified coarse-to-fine alignment for video-text retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 2816–2827). Piscataway: IEEE.
22. Tian, K., Cheng, Y., Liu, Y., Hou, X., Chen, Q., & Li, H. (2024). Towards efficient and effective text-to-video retrieval with coarse-to-fine visual representation learning. In M. J. Wooldridge, J. G. Dy, & S. Natarajan (Eds.), *Proceedings of the AAAI conference on artificial intelligence* (pp. 5207–5214). Palo Alto: AAAI Press.
23. Liu, S., Fan, H., Qian, S., Chen, Y., Ding, W., & Wang, Z. (2021). HIT: hierarchical transformer with momentum contrast for video-text retrieval. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 11915–11925). Piscataway: IEEE.
24. Xu, J., Mei, T., Yao, T., & Rui, Y. (2016). MSR-VTT: a large video description dataset for bridging video and language. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5288–5296). Piscataway: IEEE.
25. Wray, M., Doughty, H., & Damen, D. (2021). On semantic similarity in video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 3650–3660). Piscataway: IEEE.

26. Patrick, M., Huang, P.-Y., Asano, Y. M., Metze, F., Hauptmann, A. G., Henriques, J. F., & Vedaldi, A. (2021). Support-set bottlenecks for video-text representation learning. In *Proceedings of the 9th international conference on learning representations*. Retrieved May 3, 2024, from https://openreview.net/forum?id=EqoXe2zmhrh.
27. Wray, M., Larlus, D., Csurka, G., & Damen, D. (2019). Fine-grained action retrieval through multiple parts-of-speech embeddings. In *Proceedings of the IEEE international conference on computer vision* (pp. 450–459). Piscataway: IEEE.
28. Schütze, H., Manning, C. D., & Raghavan, P. (2008). *Introduction to information retrieval*. Cambridge: Cambridge University Press.
29. Büttcher, S., Clarke, C. L. A., & Cormack, G. V. (2016). *Information retrieval: implementing and evaluating search engines*. Cambridge: MIT Press.
30. Zobel, J., & Moffat, A. (2006). Inverted files for text search engines. *ACM Computing Surveys*, *38*(2), 6.
31. Cao, M., Bai, Y., Cao, Z., Nie, L., & Zhang, M. (2023). Efficient image-text retrieval via keyword-guided pre-screening. *IEEE Transactions on Circuits and Systems for Video Technology*, *34*(6), 5132–5145.
32. Wang, X., Zhu, L., Zheng, Z., Xu, M., & Yang, Y. (2022). Align and tell: boosting text-video retrieval with local alignment and fine-grained supervision. *IEEE Transactions on Multimedia*, *25*, 6079–6089.
33. Wu, X., Qian, J., & Wang, T. (2023). Text-video retrieval method based on enhanced self-attention and multi-task learning. *Multimedia Tools and Applications*, *82*(16), 24387–24406.
34. Ge, M., Li, Y., Wu, H., & Li, M. (2024). JM-CLIP: a joint modal similarity contrastive learning model for video-text retrieval. In *Proceedings of the IEEE international conference on acoustics, speech and signal processing* (pp. 3010–3014). Piscataway: IEEE.
35. Luo, H., Ji, L., Zhong, M., Chen, Y., Lei, W., Duan, N., et al. (2022). CLIP4Clip: an empirical study of CLIP for end-to-end video clip retrieval and captioning. *Neurocomputing*, *508*, 293–304.
36. Gorti, S. K., Vouitsis, N., Ma, J., Golestan, K., Volkovs, M., Garg, A., et al. (2022). X-pool: cross-modal language-video attention for text-video retrieval. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 5006–5015). Piscataway: IEEE.
37. Li, K., Wang, Y., Li, Y., Wang, Y., He, Y., Wang, L., et al. (2023). Unmasked teacher: towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 19948–19960). Piscataway: IEEE.
38. Bird, S. (2006). NLTK: the natural language toolkit. In N. Calzolari, C. Cardie, & P. Isabelle (Eds.), *Proceedings of the 21st international conference on computational linguistics and 44th annual meeting of the Association for Computational Linguistics* (pp. 141–148). Stroudsburg: ACL.
39. Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., et al. (2023). Qwen technical report. arXiv preprint. arXiv:2309.16609.
40. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., et al. (2020). An image is worth 16x16 words: transformers for image recognition at scale. In *Proceedings of the 9th international conference on learning representations*. Retrieved May 3, 2024, fromhttps://openreview.net/forum?id=YicbFdNTTy.
41. Ridnik, T., Ben-Baruch, E., Zamir, N., Noy, A., Friedman, I., Protter, M., et al. (2021). Asymmetric loss for multi-label classification. In *Proceedings of the IEEE/CVF international conference on computer vision* (pp. 82–91). Piscataway: IEEE.
42. Wu, Z., Yao, T., Fu, Y., & Jiang, Y.-G. (2017). Deep learning for video classification and captioning. In *Frontiers of multimedia research* (pp. 3–29). New York: ACM.
43. Wang, X., Wu, J., Chen, J., Li, L., Wang, Y.-F., & Wang, W. Y. (2019). Vatex: a large-scale, high-quality multilingual dataset for video-and-language research. In *Proceedings of the IEEE international conference on computer vision* (pp. 4581–4591). Piscataway: IEEE.
44. Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3202–3212). Piscataway: IEEE.
45. Wu, W., Luo, H., Fang, B., Wang, J., & Ouyang, W. (2023). Cap4video: what can auxiliary captions do for text-video retrieval?. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10704–10713). Piscataway: IEEE.
46. Kingma, D. P., & Ba, J. (2015). Adam: a method for stochastic optimization. In Y. Bengio & Y. LeCun (Eds.), *Proceedings of the 3rd international conference on learning representations*. Retrieved May 7, 2014, from http://arxiv.org/abs/1412.6980.
47. Smith, L. N. (2017). Cyclical learning rates for training neural networks. In *Proceedings of the IEEE winter conference on applications of computer vision* (pp. 464–472). Piscataway: IEEE.
48. Ma, Y., Xu, G., Sun, X., Yan, M., Zhang, J., & Ji, R. (2022). X-clip: end-to-end multi-grained contrastive learning for video-text retrieval. In J. Magalhães, A. Del Bimbo, S. Satoh, et al. (Eds.), *Proceedings of the 30th ACM international conference on multimedia* (pp. 638–647). New York: ACM.
49. Lin, T.-Y., Goyal, P., Girshick, R., He, K., & Dollár, P. (2017). Focal loss for dense object detection. In *Proceedings of the IEEE international conference on computer vision* (pp. 2980–2988). Piscataway: IEEE.
50. Anderson, P., He, X., Buehler, C., Teney, D., Johnson, M., Gould, S., et al. (2018). Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 6077–6086). Piscataway: IEEE.
51. Wang, J., Chen, D., Wu, Z., Luo, C., Zhou, L., Zhao, Y., et al. (2022). OmniVL: one foundation model for image-language and video-language tasks. In S. Koyejo, S. Mohamed, A. Agarwal, et al. (Eds.), *Proceedings of the 36th international conference on neural information processing systems* (pp. 5696–5710). Red Hook: Curran Associates.
52. Papalampidi, P., Koppula, S., Pathak, S., Chiu, J., Heyward, J., Patraucean, V., et al. (2024). A simple recipe for contrastively pre-training video-first encoders beyond 16 frames. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 14386–14397). Piscataway: IEEE.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen journal and benefit from:

- ► Convenient online submission
- ► Rigorous peer review
- ► Open access: articles freely available online
- ► High visibility within the field
- ► Retaining the copyright to your article

Submit your next manuscript at ► springeropen.com