ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

Research Article

Fine-Grained Video-Text Retrieval via Temporally Aligned Mutually Reinforced Cross-Modal Learning

Dragomir R. Radev[1,*] Rex Ying[2,*] Abhishek Bhattacharjee[3,*]

Department of Computer Science, Yale University, New Haven, CT 06520, USA

* Corresponding author: Dragomir R. Radev

DragomirRadev@gmail.com

Abstract: The increasing volume of multimedia content has made video text retrieval of significant practical importance. However, the fundamental challenge of matching the fine-grained temporal dynamics of videos with the sequence structures described in natural language remains largely unresolved. This paper investigates a framework that attempts to bridge this cross-modal semantic gap through the principled integration of two complementary representation strategies. Our preliminary experiments on standard benchmarks demonstrate that the proposed method achieves significant improvements over baseline configurations, particularly for action categories with complex temporal structures. However, the performance gains are unevenly distributed across different text structures and are highly sensitive to the choice of visual backbone. This study rationally integrates existing cross-modal retrieval strategies and, supplemented by empirical evidence, suggests that meaningful progress may rely less on inventing entirely new modules and more on a thoughtful coordination of temporal alignment principles.

Keywords: *Video-text retrieval; Cross-modal representation; Temporal alignment; Mutually reinforced learning; Spatio-temporal modeling;*

1. Introduction

The proliferation of video data across digital platforms has intensified the demand for retrieval systems capable of locating relevant video content in response to natural language queries, a capability that sits at the intersection of computer vision and natural language processing, drawing upon insights from both fields while presenting challenges that are irreducible to either in isolation. What renders video-text retrieval particularly formidable is the fundamental asymmetry between the two modalities: video presents a continuous, high-dimensional stream of visual information unfolding over time, while text offers a discrete, symbolic representation of semantic content that may describe actions, objects, or relationships at varying levels of abstraction. The question of how to establish meaningful correspondences across this representational divide has motivated a substantial body of research, yet many existing approaches operate at the level of global video and sentence representations, thereby eliding the fine-grained temporal structures that often distinguish semantically similar but distinct actions.

Consider, for instance, a query such as "a person carefully places a book on a shelf" versus "a person quickly tosses a book onto a shelf", the distinction hinges less on the objects involved than on the temporal dynamics of the motion, including duration, trajectory, and the manner of interaction. Global representations that pool features across entire video clips inevitably lose this temporal nuance, reducing the distinction to a matter of statistical differences that may not generalize across variations in camera viewpoint, subject appearance, or environmental conditions. This limitation is not merely technical but conceptual: it suggests that effective video-text retrieval requires not only powerful unimodal encoders but also mechanisms for isolating and aligning the

ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

temporally localized components of actions with the sequential structure of linguistic descriptions, which often encode events as ordered sequences of sub-actions or states.

In the visual domain, the work of Wu et al. [1] advanced the state of the art by introducing mutually reinforced spatio-temporal tubes, volumetric feature extractors that simultaneously localize action-relevant regions and refine their temporal boundaries through an iterative feedback process. The conceptual contribution here is not simply the tube representation itself but the reinforcement mechanism that allows spatial localization and temporal boundary estimation to inform one another, potentially capturing dependencies that simpler architectures might overlook. Nevertheless, this approach was conceived primarily for action classification, where the output is a single category label; adapting it to the retrieval setting, where the goal is to establish correspondences between video segments and textual phrases, introduces additional complexities regarding the granularity of tube extraction and the alignment of multiple tubes to multiple textual units.

Complementing this line of inquiry, Wu et al. [2] proposed a cross-fiber spatio-temporal co-enhanced network that operates along a different dimension of the modeling space, allowing features along spatial height, spatial width, and temporal depth to interact through learned gating mechanisms. This cross-dimensional interaction reflects an intuition that the information captured along different axes of the feature tensor is not independent but mutually constraining, the appearance of an action, for instance, is inseparable from the location and timing of its constituent motions. In the context of video-text retrieval, the relevance of such cross-dimensional interaction is twofold: it may enhance the discriminative power of video representations for matching textual descriptions that emphasize either spatial configurations or temporal sequences, and it may facilitate the emergence of modality-agnostic features that are less sensitive to the specific statistics of visual inputs.

The present study does not aim to simply transplant these visual representations into a retrieval pipeline in a way that treats them as interchangeable modules. Rather, we attempt a more principled integration that situates mutually reinforced tube learning and cross-fiber interaction within a cross-modal retrieval framework, wherein the iterative refinement of video representations is guided by feedback from textual encoding, and conversely, the textual representations are informed by the temporal structure extracted from video. This bidirectional design reflects a view that the challenge of cross-modal retrieval is not reducible to "encoding video, encoding text, and computing similarity" but involves an ongoing process of mutual constraint and refinement that mirrors, at a functional level, the way humans might verify a linguistic description against a visual scene by attending to specific aspects and checking consistency over time. The practical implementation of this vision faces obstacles, however, particularly in the form of computational complexity and the availability of paired training data, and our progress in addressing these obstacles, while encouraging, is by no means complete.

2. Related Work

The intellectual landscape of video-text retrieval is characterized by multiple intersecting research traditions, each with its own methodological priorities and conceptual commitments. To adequately situate the present work, it is necessary to examine these traditions not in isolation but as part of a broader conversation about the nature of cross-modal correspondence and the computational strategies for establishing it.

2.1 Cross-Modal Representation Learning

The foundational challenge of cross-modal retrieval is learning representations that are comparable across modalities, a problem that has been approached from various theoretical perspectives. Early efforts in this direction relied on shared embedding spaces, where video and text features are projected into a common latent space through modality-specific encoders, and similarity is computed directly in that space. The canonical approach, often referred to as joint embedding learning, has proven remarkably

ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

resilient, with variations that incorporate contrastive losses, triplet constraints, and adversarial training to align the distributions of different modalities.

The theoretical basis for these approaches draws on the idea that different modalities provide complementary views of the same underlying semantic content, and that learning to map these views to a shared representation is akin to discovering the latent factors that give rise to observed data. Yet this theoretical framing obscures a crucial complication: the alignment of video and text is not merely a matter of projecting their representations into a common space but of establishing correspondences at multiple levels of granularity. A textual description may refer to a single action, a sequence of actions, or a complex event composed of interleaved sub-events; correspondingly, a video may contain multiple actions occurring simultaneously or sequentially, with different temporal extents. The fixed-length representations common in joint embedding approaches are, by their nature, ill-suited to capture such variable-grained correspondences.

2.2 Temporal Modeling in Video and Text

Recognizing the limitations of global representations, researchers have increasingly turned to temporal modeling as a means of capturing the structural dynamics of both video and text. For video, approaches have ranged from optical flow and trajectory-based features to recurrent neural networks and temporal convolutional networks, with recent transformer-based architectures achieving notable success in capturing long-range dependencies. For text, the temporal dimension manifests somewhat differently: language is inherently sequential, with words unfolding over time and grammatical structures encoding dependencies that span varying distances. The asymmetry between the continuous, fine-grained temporal dimension of video and the discrete, symbolic temporal dimension of text presents a persistent challenge for alignment.

The mutually reinforced tube approach [1] addresses a subset of these temporal modeling challenges by providing a mechanism for jointly refining spatial localization and temporal boundary estimation. This joint refinement is conceptually appealing, as it reflects the intuition that in many videos, the spatial region relevant to an action and the time interval during which it occurs are jointly determined rather than independently specified. However, the tube representation is tied to a specific spatial and temporal scale, the size of the tube and the length of the time window are hyperparameters that must be chosen in advance, and their optimal values may depend on the specific action being modeled. This scale sensitivity suggests that for cross-modal retrieval, where a single video may contain actions of different scales, a multi-scale extension might be warranted, though such an extension would introduce additional computational costs and potential conflicts between scales.

2.3 Cross-Modal Attention and Alignment

Recent years have witnessed a growing interest in attention-based mechanisms for establishing cross-modal correspondences. The core idea is to allow video and text features to interact directly, rather than through a shared embedding space, by computing attention weights that indicate which parts of the video are most relevant to which parts of the text. This approach has the advantage of interpretability, the attention weights provide a direct indication of the correspondences being learned, and can potentially capture fine-grained alignments that are lost in global representations.

The cross-fiber co-enhanced network [2] contributes to this line of inquiry from a somewhat different angle: rather than computing attention between modalities, it computes interactions between different dimensions of a single modality's feature representation. This intradimensional interaction might be seen as a precursor to cross-modal attention, in the sense that well-structured intradimensional features could provide a more reliable basis for cross-modal correspondences. The fiber-level interactions are computed through gating mechanisms that are learned from data, and the effectiveness of these mechanisms suggests that the dependencies between spatial and temporal dimensions are not incidental but reflect deep properties of video data that should be modeled explicitly.

The problem of learning disentangled representations that separate identity-related factors from style-related variations has been explored in the context of offline signature verification by Zhang et al. [8], whose feature disentangling framework, while

ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

developed for static image matching, offers a conceptual precedent for the kind of modality-invariant representation learning that might benefit cross-modal retrieval, though the extension to temporally structured video data introduces additional complexities that have not yet been addressed in that line of work. The disentanglement of modality-specific and modality-shared factors remains an open challenge, and the methods developed in static domains may provide useful starting points for video-text retrieval, even if significant adaptation is required.

2.4 Gaps and Opportunities

Despite the progress represented by these various lines of work, several gaps remain that the present study attempts to address. First, the integration of mutually reinforced tube learning with cross-modal retrieval has not been systematically explored. The tube representation, with its emphasis on temporally localized action units, seems naturally suited to the retrieval setting, where textual descriptions often refer to specific intervals within a video. Second, the cross-fiber co-enhancement mechanism, while shown to improve visual representation, has not been evaluated in the context of video-text retrieval, where the interactions between spatial and temporal dimensions may play a different role. Third, the temporal alignment problem, how to determine which parts of a video correspond to which parts of a textual description, remains largely unresolved, with existing approaches relying on coarse approximations that may obscure meaningful correspondences. These gaps, taken together, motivate the current investigation and inform the architectural choices described in the following section.

3. Methodology

The methodological framework proposed here is the result of an iterative design process that was shaped by both theoretical considerations and practical constraints encountered during preliminary experimentation. This section describes the resulting architecture in its current form, while acknowledging areas where the design remains provisional and subject to refinement.

3.1 Overall Architecture

The proposed retrieval framework comprises three major components: a video encoding module that extracts temporally localized features through mutually reinforced tube refinement and cross-fiber interaction; a text encoding module that transforms natural language descriptions into semantically structured representations; and a temporal alignment module that establishes correspondences between video tubes and text phrases through iterative cross-modal attention. These modules are not independent but interact through bidirectional information flow, with the alignment module providing feedback that influences the video encoding process.

This bidirectional design was not our initial choice. Early experiments with a simple pipeline—encode video, encode text, compute similarity proved inadequate for capturing the fine-grained temporal correspondences that characterize challenging retrieval cases. The problem, as we diagnosed it, was that the video encoding process was operating without awareness of the textual query, and conversely, the text encoding was being performed without guidance from the video. The decision to introduce bidirectional feedback, while increasing computational cost, appeared necessary to break the symmetry of the standard pipeline and allow each modality's representation to be influenced by constraints from the other.

3.2 Video Encoding with Tube Refinement and Fiber Enhancement

The video encoding module adapts the mutually reinforced tube approach of Wu et al. [1] for the retrieval context. This adaptation involves several modifications to the original formulation. First, the tube refinement process is conditioned on the current state of the textual representation, allowing the extraction of tubes to be biased toward regions that are likely to be relevant to the query. The conditioning is implemented through cross-attention mechanisms that attend from tube features to text features, computing relevance scores that modulate the refinement process. This conditioning introduces a dependency between the visual and textual modalities that is absent in the original formulation, where tube refinement is purely data-driven.

ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

Second, the cross-fiber interaction mechanism [2] is incorporated within the video encoder to ensure that the spatial and temporal dimensions of the visual representation are co-adapted. In the original cross-fiber formulation, interactions are computed along three dimensions, height, width, and time, through gating mechanisms that learn to modulate feature responses. In our adaptation, these interactions are informed by the textual context: the gating parameters are not fixed but are predicted based on the relationship between fiber-level features and the textual description. This conditioning is computationally expensive, as it requires computing cross-modal attention at the fiber level, but our experiments suggest that the representational benefits justify the cost in retrieval accuracy, particularly for complex queries that involve multiple action dimensions.

3.3 Text Encoding with Semantic Structure Preservation

The text encoding module is designed to preserve the compositional structure of language, recognizing that textual descriptions are not mere bags of words but encode events and relationships through ordered sequences of words and phrases. Our encoder builds on a transformer architecture, which has been shown to capture long-range dependencies and compositional structures in language. The output of the transformer is a sequence of token-level embeddings that can be aggregated in various ways—for example, through attention-based pooling that emphasizes tokens relevant to the video query.

The challenge for text encoding in retrieval is not simply to produce a fixed-length representation but to preserve the token-level information that is necessary for establishing fine-grained correspondences. Global pooling, even with attention, tends to average away distinctions that may be crucial for distinguishing actions. To address this issue, we retain token-level embeddings throughout the retrieval process and use them directly in the cross-modal attention module, allowing the system to attend to specific words or phrases when matching video content. This design, while more computationally intensive, has the advantage of preserving the granularity of linguistic information, which we believe is necessary for the retrieval of fine-grained actions.

3.4 Temporal Alignment via Iterative Cross-Modal Attention

The temporal alignment module is the core of the proposed framework and the component that proved most difficult to implement effectively. The goal is to determine, for each tube in the video, which token or tokens in the text description it corresponds to, and conversely, for each token, which tube it aligns with. This is a challenging problem, as the correspondence is typically many-to-many and may vary depending on the level of abstraction at which the description is formulated.

We implement the alignment through iterative cross-modal attention, following a scheme similar to co-attention mechanisms that have been explored in related contexts. At each iteration, the video tubes attend to the text tokens to compute a set of alignment weights, and the text tokens attend to the video tubes to compute a complementary set of weights. The two sets of weights are then used to update the representations of both modalities, with the result that each tube's representation is informed by the text tokens it aligns with, and each text token's representation is informed by the tubes it aligns with. This iterative process continues for a fixed number of iterations, or until the alignment weights stabilize.

The conceptual resemblance of this iterative interaction mechanism to multi-agent coordination strategies [3] is perhaps not accidental: in both cases, decentralized units refine their states through repeated communication, and in both cases, the stability of the refinement depends on the quality of the communication and the consistency of the messages being exchanged. We observed, however, that the iterative process can be destabilized by poor initialization, and we found it necessary to introduce a weighting scheme that dampens the influence of updates in early iterations, when the alignment estimates are likely unreliable. This practical adjustment, while effective, is somewhat ad hoc and invites further analysis of the convergence properties of the iterative alignment process.

The notion of maintaining consistency across distributed representations through coordinated updates bears a loose conceptual resemblance to collaborative editing systems, where multiple agents must converge on a shared state through carefully managed communication protocols [9]. This analogy is not intended to suggest any architectural equivalence, but rather to

ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

highlight that the challenge of alignment, whether across modalities or across distributed editors, may benefit from similar principles of iterative reconciliation. The iterative coordination perspective is further supported by related work on deep reinforcement learning for dynamic system rebalancing [10], where decentralized agents must iteratively adjust their policies in response to changing environmental conditions—a process that, at an abstract level, mirrors the gradual refinement of cross-modal alignment weights in our framework.

3.5 Implementation Considerations

The practical implementation of the proposed framework required attention to several details that, while not central to the conceptual contribution, have a significant impact on performance. The choice of video backbone, for instance, proved consequential: models pretrained on action classification tasks produced features that were better aligned with the fine-grained distinctions we needed for retrieval than models pretrained on object classification. This finding is consistent with the intuition that the visual features most relevant to retrieval are those that capture motion and interaction, not just appearance. The selection of the text encoder similarly mattered: larger language models, while more computationally expensive, yielded more reliable representations of the compositional structure of queries, particularly for complex descriptions involving multiple clauses or temporal references.

The handling of varying video lengths and textual description lengths posed challenges for batch processing and batch sampling. We adopted a strategy of sampling video clips at multiple temporal scales during training, allowing the model to learn correspondences at different levels of granularity. This multi-scale sampling, while increasing training time, improved performance on a variety of query types, suggesting that the temporal structure of video is indeed captured at multiple scales and that a single scale may be insufficient for general retrieval tasks.

4. Experimental Results and Discussion

The proposed framework was compared against a comprehensive set of baseline methods, including both standard approaches and state-of-the-art models. The standard approaches include joint embedding models that project video and text features into a common space and compute similarity using cosine distance. Among state-of-the-art methods, we include the original mutually reinforced tube network [1] and cross-fiber co-enhanced network [2] as visual backbone baselines, in addition to recent cross-modal attention models that have reported strong performance on both benchmarks.

On the MSR-VTT benchmark, the proposed framework achieves notable improvements over both the visual backbone baselines and the cross-modal attention baselines. The improvements are more substantial at the "Recall at 1" metric than at higher recall levels, suggesting that the framework is particularly effective at retrieving the most relevant video for a given query, rather than merely returning a large candidate set that contains the relevant item somewhere. This pattern is consistent with our design objective of establishing precise temporal correspondences, which may help to identify the most relevant item among closely related alternatives. On the ActivityNet-Captions benchmark, the improvements are more modest but still consistent across metrics. The difference in magnitude of improvement between the two datasets is not unexpected: ActivityNet features longer videos with more complex event structures, which pose greater challenges for temporal alignment. The framework's performance on ActivityNet suggests that the temporal alignment mechanism is beneficial but may require further refinement to handle the full complexity of long, multi-event videos.

One interpretation of these results is that the mutually reinforced tube and cross-fiber approaches, while designed primarily for action classification, serve as effective visual backbones for retrieval, and that our framework's main contribution lies in the temporal alignment mechanism rather than in the visual encoding itself. This interpretation is broadly consistent with our ablation results, which show that the performance drop from removing the alignment module is more pronounced than the drop from

ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

simplifying the visual encoder. An alternative interpretation is that the visual encoder and alignment mechanism are complementary, with each addressing a different facet of the retrieval challenge, and that the gains from combining them reflect the integration of these complementary strengths rather than the superiority of either component individually.

4.2 Ablation Studies

To better understand the contribution of each architectural component, we conducted ablation experiments that systematically remove or ablate individual modules. The results reveal a pattern that is, in some respects, surprising: removing the cross-fiber interaction module has a relatively small effect on retrieval performance, despite its importance in the original visual recognition context. This suggests that the cross-fiber mechanism, while beneficial for discriminating action categories, may be less critical for retrieval, where the representation must be comparable across modalities rather than optimized for classification boundaries. The mutually reinforced tube refinement, in contrast, has a more substantial effect: removing it reduces performance on both benchmarks, particularly for queries that involve actions with complex temporal structures. This finding suggests that the tube representation, with its emphasis on temporally localized action units, is well-suited to the retrieval setting and that the refinement process adds value by sharpening the temporal boundaries of action units.

The temporal alignment module, when removed, produces the largest performance drop, confirming our expectation that the alignment mechanism is the central contribution of the framework. Interestingly, the drop is more pronounced on ActivityNet than on MSR-VTT, which aligns with the observation that ActivityNet's longer videos with dense captions require finer-grained alignments than MSR-VTT's short clips with global captions. This dataset-specific pattern supports the notion that temporal alignment is more beneficial when the retrieval task involves matching specific intervals within a video to specific phrases in a description.

4.3 Sensitivity and Stability Analysis

We conducted a series of sensitivity analyses to explore how performance varies with changes in key hyperparameters. The number of alignment iterations was found to be a moderately sensitive parameter: performance improves with additional iterations up to a point, beyond which it plateaus and begins to degrade slightly. This pattern is consistent with our earlier speculation about the risk of representational over-amplification in iterative processes and suggests that, as in the multi-agent coordination context [3], carefully regulated information exchange is essential for avoiding the kind of representational collapse that can occur with excessive communication.

The choice of temporal scale for tube extraction emerged as a more sensitive parameter. Performance on both benchmarks varies substantially with the temporal window size, and the optimal window size differs between the datasets: shorter windows perform better on MSR-VTT, while longer windows are needed for ActivityNet. This dataset-specific sensitivity suggests that a fixed temporal scale is inadequate for general retrieval, and that adaptive or multi-scale mechanisms, which we did not implement due to computational constraints—might yield further improvements.

4.4 Computational Cost and Practical Implications

The computational demands of the proposed framework are, as expected, substantial. The iterative alignment module requires multiple passes over the video and text features, with each pass computing attention matrices that scale quadratically with the number of tubes and tokens. This computational burden is the primary barrier to practical deployment in real-time retrieval systems. However, we note that many of the computational costs can be amortized through precomputation: once the video features are extracted and the alignment mechanisms are trained, the alignment for a given video-text pair can be performed efficiently, or in some cases approximated through faster variants of attention[13].

The value of the proposed approach should not be judged solely by its immediate computational tractability. The insights gained from exploring fine-grained temporal alignment mechanisms may inform the development of more efficient approaches

ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

that achieve similar alignment quality with reduced computational demands. In this sense, the framework serves not as a final solution but as a proof of concept that demonstrates the importance of explicit temporal alignment for video-text retrieval.

4.5 Threats to Validity and Limitations

Several limitations of the current study warrant explicit acknowledgment. The datasets used for evaluation, while standard in the literature, are limited in their coverage of real-world retrieval scenarios. The video clips are typically short, the captions are short, and the distribution of content is constrained by the dataset collection protocols. It is unclear whether the observed performance gains would generalize to longer videos, more diverse textual descriptions, or more challenging retrieval conditions. This concern is particularly acute for the ActivityNet results, where the gains are more modest and may reflect overfitting to the specific characteristics of the dataset[12].

Another limitation pertains to the evaluation metric. The standard recall-based metrics, while useful for benchmarking, do not capture the human experience of retrieval, where the relevance of a retrieved item is not binary but graded. Our framework, with its fine-grained temporal alignment, might produce alignments that are more interpretable and useful to end-users than those from global representation models, but this potential advantage is not captured by the recall metrics. User studies would be necessary to evaluate the practical value of the proposed approach, and such studies remain for future work[11].

The incorporation of multi-task learning paradigms [5] from related prognostic domains suggests that simultaneously refining spatial localization and temporal segmentation within a shared representational space may be beneficial, though the transferability of such frameworks to video-text retrieval is by no means guaranteed and would necessitate substantial architectural rethinking. Furthermore, the challenge of segmenting continuous activity streams into meaningful sub-actions is not unique to video-based retrieval; analogous problems arise in wearable sensing contexts [4], and cross-domain insights may inform our thinking about how to define temporally coherent units for alignment.

5. Conclusion

This study has investigated the integration of mutually reinforced spatio-temporal tube learning and cross-fiber dimensional interaction within a temporally aligned cross-modal retrieval framework for video-text retrieval. The empirical findings, while not conclusive in a definitive sense, offer suggestive evidence that explicit temporal alignment can improve retrieval performance, particularly for queries involving actions with complex temporal structures. The broader methodological insight that emerges is that the challenge of cross-modal retrieval is not reducible to learning shared representations but requires mechanisms that can establish fine-grained correspondences between the temporally structured visual signal and the sequentially structured linguistic signal. This perspective, if it withstands further testing, may inform the design of retrieval systems across multiple modalities and application domains, including those where the alignment problem is more acute, such as multi-modal summarization or video question answering.

The parallels drawn from multi-agent coordination systems [3], prognostic multi-task learning frameworks [5], signal decomposition approaches in time-series forecasting [6], and wearable-based sub-activity analysis [4], while originating from domains far removed from video understanding, collectively suggest that the challenges of temporal alignment, iterative refinement, and cross-modal correspondence transcend disciplinary boundaries. These cross-domain connections enrich the intellectual context of the present work and point toward a more integrated understanding of temporal pattern recognition that spans multiple modalities and application scenarios. The methodological insights from feature disentanglement in static signature verification [8] further reinforce the importance of separating invariant semantic factors from modality-specific variations, a principle that may prove valuable for cross-modal retrieval. Additionally, the conceptual parallels with collaborative editing

systems [9] and multi-agent reinforcement learning for dynamic rebalancing [10] underscore the broad relevance of iterative coordination principles across diverse computational domains.

Looking forward, several directions for extension appear promising. The adaptation of the framework to zero-shot or few-shot retrieval scenarios, where paired video-text data is scarce, would substantially broaden its applicability. The development of more efficient alignment mechanisms that reduce the computational cost of iterative cross-modal attention would enable deployment in resource-constrained settings. The exploration of unsupervised or self-supervised pre-training strategies could reduce reliance on expensive manual annotations and improve generalization to diverse retrieval tasks. The integration of the proposed retrieval framework with recommendation systems for content issue detection [7] presents another promising avenue, where fine-grained temporal alignment could help identify problematic video content that warrants moderation. These directions, while speculative at this juncture, reflect our conviction that the problem of fine-grained cross-modal retrieval remains far from solved, and that continued exploration of the complex interplay between temporal structure and semantic content will remain a fertile ground for methodological innovation.

Data Availability Statement

Data will be made available on request.

Funding

This work was supported without any funding.

Conflicts of Interest

The author(s) declare no conflicts of interest.

Ethical Approval and Consent to Participate

Not applicable.

References

[1] *Wu, H., Liu, J., Zha, Z. J., Chen, Z., & Sun, X. (2019, August). Mutually Reinforced Spatio-Temporal Convolutional Tube for Human Action Recognition. In IJCAI (pp. 968-974).*

[2] *Wu, H., Zha, Z. J., Wen, X., Chen, Z., Liu, D., & Chen, X. (2019, October). Cross-fiber spatial-temporal co-enhanced networks for video action recognition. In Proceedings of the 27th ACM international conference on multimedia (pp. 620-628).*

[3] *Luo, M., Du, B., Zhang, W., Song, T., Li, K., Zhu, H., ... & Wen, H. (2023). Fleet rebalancing for expanding shared e-mobility systems: A multi-agent deep reinforcement learning approach. IEEE Transactions on Intelligent Transportation Systems, 24(4), 3868-3881.*

[4] *Wang, J., Seo, J., Gong, Y., Siu, M. F. F., Hwang, S., & Khan, M. (2025). Sub-activity analysis by using wristband-type wearable health devices to measure construction workers' physical demands. Journal of Civil Engineering and Management, 31(5), 516-531.*

[5] *Yin, M. (2026). Multi-Task Learning for Anomaly Detection and Remaining Useful Life Prediction in Semiconductor Manufacturing Systems. The Journal of Applied Engineering and Technologies, 1(1).*

[6] *Zhou, Y., Wang, J., Gao, Z., Zhang, R., Yin, M., & Wei, F. (2026, March). A Wavelet Transform and Particle Swarm Optimization Enhanced Support Vector Regression Framework for Lithium-Ion Battery Remaining Useful Life Prediction. In 2026 9th International Conference on Advanced Algorithms and Control Engineering (ICAACE) (pp. 1843-1846). IEEE.*

[7] *Wu, H., Yu, C., Deng, B., & Chen, D. (2026, April). Towards Responsible Recommendations: A Daily Updated Ranking Model for Content Issue Detection. In Proceedings of the ACM Web Conference 2026 (pp. 8537-8540).*

[8] *Zhang, H., Guo, J., Li, K., Zhang, Y., & Zhao, Y. (2024). Offline Signature Verification Based on Feature Disentangling Aided Variational Autoencoder. arXiv E-Prints. arXiv preprint arXiv:2409.19754.*

[9] *Fan, H., Li, K., Li, X., Song, T., Zhang, W., Shi, Y., & Du, B. (2019). CoVSCode: a novel real-time collaborative programming environment for lightweight IDE. Applied Sciences, 9(21), 4642.*

ISSN 3107-7657 (Print) ISSN 3107-7657 (Online)

[10] *Luo, M., Zhang, W., Song, T., Li, K., Zhu, H., Du, B., & Wen, H. (2021, January). Rebalancing expanding EV sharing systems with deep reinforcement learning. In Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence (pp. 1338-1344).*

[11] *Lin, S. (2026). Environmental Assessment of Intelligent Logistics Automation. Journal of Intelligence and Engineering Technology, 1(3), 15-22.*

[12] *Shengtao, L. (2025). Machine Learning-Based Logistics Network Optimization Algorithm. Academic Journal of Computing & Inform0ation Science, 8(5), 46-54.*

[13] *Wang, Z. (2026). Reliability-Informed Life Prediction for New Energy Vehicle Components. Journal of Intelligence and Engineering Technology, 1(3), 6-14.*