From Correlation to Causation: Counterfactual Bi-Directional Alignment for Robust Text–Video Retrieval

Wenbin Meng * and Ming Xu

Posted Date: 28 May 2026

doi: 10.20944/preprints202605.1948.v1

Keywords: text–video retrieval; cross-modal learning; adaptive feature decoupling; attention interaction; semantic alignment; efficient multimodal models

 doi:10.20944/preprints202605.1948.v1

Article

From Correlation to Causation: Counterfactual Bi-Directional Alignment for Robust Text–Video Retrieval

Wenbin Meng * and Ming Xu

Ningbo University, College of Information Science and Engineering, 315000, Zhejiang, China
* Correspondence: 15258117347@163.com

Abstract

Precise semantic matching between natural language queries and unconstrained videos remains a fundamental yet unresolved challenge in multimedia retrieval. Although recent transformer-based dual encoders and CLIP-style contrastive frameworks have improved global text–video alignment, they still struggle in complex scenes where (i) spatiotemporal cues are highly entangled among objects, motion patterns, and background context, and (ii) cross-modal interactions are easily biased by spurious correlations, resulting in brittle retrieval performance under compositional or ambiguous language. To overcome these limitations, we propose a unified framework that enhances text–video correspondence through three closely coupled components: *Query-adaptive Semantic Routing* (QSR), *Counterfactual Bi-directional Alignment* (CBA), and *Temporal Causal Regularization* (TCR). QSR introduces a query-conditioned routing mechanism that decomposes video representations into multiple semantic experts and dynamically assigns token-level relevance, allowing the model to selectively emphasize appearance, motion, and contextual cues according to the textual query. Based on the routed representations, CBA performs reciprocal attention in both text-to-video and video-to-text directions, while introducing a counterfactual alignment branch to suppress background-driven shortcuts; this encourages robust matching based on *causal* evidence rather than incidental correlations. Finally, TCR imposes temporal causality-aware consistency by penalizing alignment instability under lightweight temporal perturbations, thereby improving motion sensitivity without requiring dense frame sampling. For scalable deployment, we further incorporate parameter sharing across experts and quantization-friendly projections, achieving a favorable accuracy–latency trade-off. Experiments on MSR-VTT, MSVD, and VATEX demonstrate consistent improvements over strong baselines, achieving Recall@1 scores of **55.0%**, **60.3%**, and **68.5%**, respectively, while maintaining high inference efficiency.

Keywords: text–video retrieval; cross-modal learning; adaptive feature decoupling; attention interaction; semantic alignment; efficient multimodal models

1. Introduction

The exponential growth of video content across streaming platforms, social media networks, and surveillance infrastructures has created an urgent need for robust text–video semantic matching, which serves as a cornerstone task in multimedia retrieval and video understanding. As intelligent systems are increasingly expected to interpret complex visual dynamics through natural language, bridging the semantic gap between these heterogeneous modalities has become a critical research challenge. Early deep learning methods mainly relied on joint embedding spaces to associate textual queries with visual content, whereas the emergence of large-scale foundation models has substantially reshaped this research landscape. Transformer-based architectures have shown remarkable capability in modeling long-range dependencies [1,2]. Recent studies have further extended these paradigms to the multimodal domain, where large-scale pre-training frameworks have established strong benchmarks for generalized cross-modal representation learning [3,4]. Despite these advances, effectively

 doi:10.20944/preprints202605.1948.v1

capturing the multi-granular and context-dependent semantics of video content remains an open problem, especially when dynamic visual scenes must be aligned with natural language descriptions that are inherently ambiguous and compositionally diverse [5,6].

Existing studies have explored a wide range of architectures to reduce semantic discrepancies between vision and language. Early methods such as VideoBERT [7] and UniVL [8] adapted masked language modeling to video-language tasks, while subsequent approaches, including CLIP4Clip [9] and X-CLIP [10], leveraged the strong contrastive priors of CLIP to improve retrieval performance. More recent models, such as UMT [11], InternVideo [3], and various video-oriented large language models [12,13], have further enhanced temporal reasoning and large-scale multimodal pre-training. Nevertheless, existing methods still face several intrinsic limitations. First, many architectures predominantly depend on global representations, such as [CLS] token embeddings, which may obscure fine-grained or entangled semantic cues that are essential for distinguishing visually similar video segments [14]. Second, standard cross-modal attention mechanisms often fuse heterogeneous features without sufficient semantic disentanglement, causing interference when scenes involve dense interactions among multiple objects, actions, and contextual factors. Third, the high computational cost of recent Large Multimodal Models (LMMs) limits their practicality in real-time and large-scale retrieval scenarios. Therefore, achieving robust, fine-grained, and efficient text–video alignment remains a challenging and unresolved objective.

However, developing such a robust text–video retrieval framework is non-trivial and faces three key challenges. First, video representations are inherently entangled, since appearance, motion, object interactions, and background context are often encoded jointly in the same latent space, making it difficult to identify the visual evidence that is truly relevant to a given query. Second, cross-modal alignment is vulnerable to spurious correlations, where models may match textual descriptions with incidental background cues or dataset-specific co-occurrence patterns rather than causal semantic evidence. Third, improving temporal sensitivity usually requires dense frame sampling or heavy interaction modules, which increases computational cost and limits scalability in large-scale retrieval scenarios. These challenges require a retrieval framework that can adaptively disentangle query-relevant semantics, suppress non-causal alignment shortcuts, and maintain temporal robustness under efficient sampling conditions.

To address these limitations, we propose a unified text–video retrieval framework that targets *query-conditioned semantic disentanglement*, *counterfactual alignment robustness*, and *temporal causality-aware stability*, as illustrated in Figure 1. The framework contains three tightly coupled modules. (1) **Query-adaptive Semantic Routing (QSR)** routes video tokens into multiple semantic experts conditioned on the textual query, enabling adaptive emphasis on query-relevant appearance, motion, and contextual evidence instead of relying on fixed global pooling or query-agnostic aggregation. (2) **Counterfactual Bi-directional Alignment (CBA)** performs reciprocal text–video attention over the routed representations and introduces a counterfactual branch to suppress background-dominated shortcuts, penalizing alignments that remain overly confident after causal evidence is removed. (3) **Temporal Causal Regularization (TCR)** imposes consistency under lightweight temporal perturbations, encouraging reliance on stable motion-related cues rather than incidental frame artifacts and improving robustness under sparse sampling. Importantly, these innovations are designed under efficiency constraints. We share parameters across experts, adopt quantization-friendly projection heads, and keep the interaction module lightweight, enabling scalable deployment. Extensive experiments on MSR-VTT, MSVD, and VATEX validate that our method consistently improves top-rank accuracy and ranking stability over strong baselines.

 doi:10.20944/preprints202605.1948.v1

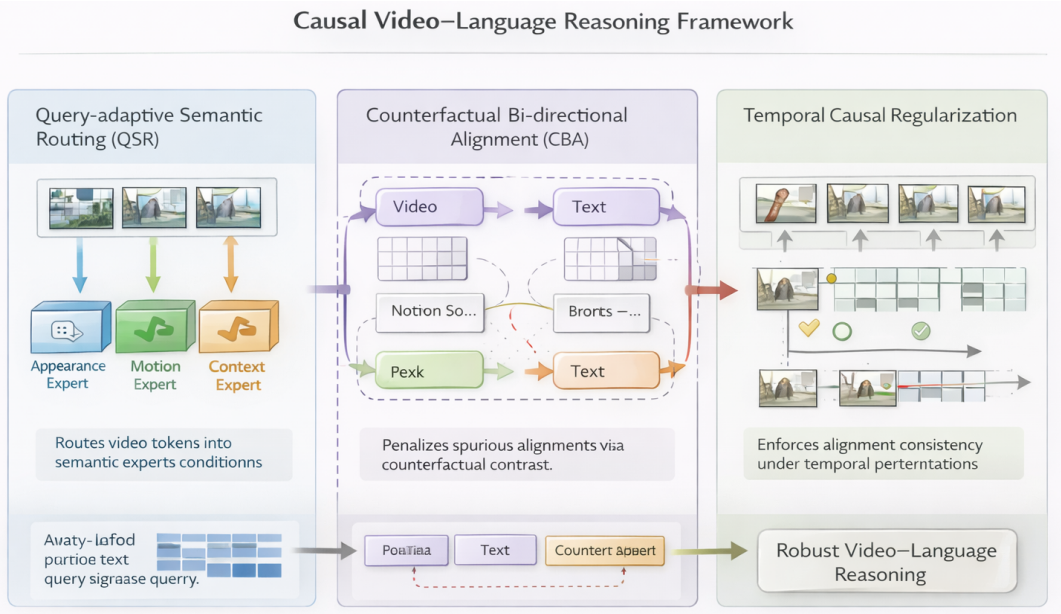

Figure 1. Introduction of the proposed text–video semantic matching framework.

The main contributions of this work are summarized as follows:

- We propose **Query-adaptive Semantic Routing (QSR)**, a query-conditioned multi-expert routing mechanism that decomposes video tokens into semantic experts and dynamically assigns token-level relevance, thereby alleviating representation entanglement in complex video scenes.
- We introduce **Counterfactual Bi-directional Alignment (CBA)**, which enhances reciprocal cross-modal attention with a counterfactual suppression branch. By penalizing alignments that remain highly confident after causal evidence is removed, CBA explicitly reduces background-driven spurious correlations and improves robustness under compositional queries.
- We design **Temporal Causal Regularization (TCR)**, which enforces alignment consistency under lightweight temporal perturbations. This regularization strengthens motion-relevant grounding while avoiding the computational overhead of dense frame sampling.
- We develop an efficiency-oriented implementation that integrates parameter sharing and quantization-friendly projections. Extensive experiments on MSR-VTT, MSVD, and VATEX demonstrate that our method achieves Recall@1 scores of **55.0%**, **60.3%**, and **68.5%**, respectively, while maintaining high inference efficiency.

2. Related Work

2.1. Multimodal Pre-Training for Video–Language Understanding

Large-scale multimodal pre-training has substantially advanced video–language understanding by enabling models to learn transferable cross-modal representations from massive weakly aligned video–text data. Early representative works, such as VideoBERT [7], UniVL [8], HERO [15], and Merlot [16], demonstrated the effectiveness of masked reconstruction, hierarchical reasoning, and joint video–text modeling for learning generalizable multimodal embeddings. With the emergence of contrastive vision–language models, methods such as CLIP [17] and CLIP4Clip [9] further accelerated progress in text–video retrieval by leveraging large-scale contrastive objectives to align visual and textual representations in a shared semantic space. Subsequent studies explored the use of large-scale instructional or narrated video datasets, such as HowTo100M [18], to support video–language pre-training from noisy but abundant supervision. Frozen in Time [19] further showed that jointly optimizing contrastive and embedding objectives over large image–text and video–text corpora can yield strong zero-shot transfer capability.

Despite these advances, existing pre-training strategies still tend to rely heavily on global or weakly disentangled representations. As a result, they often struggle to separate appearance, motion,

 doi:10.20944/preprints202605.1948.v1

and contextual cues when real-world videos contain dense object interactions, background clutter, or temporally mixed activities. This limitation motivates the need for adaptive representation decomposition mechanisms that can refine video features according to the semantics of a given textual query.

2.2. Cross-Modal Alignment and Attention Interaction

Cross-modal alignment aims to improve the semantic correspondence between textual and visual representations, and attention-based interaction has become one of the dominant mechanisms for this purpose. Transformer architectures have played a central role in this development by enabling long-range dependency modeling and flexible feature interaction across modalities. Representative models such as MART [20], COOT [21], and X-CLIP [10] introduced different forms of temporal reasoning and fine-grained alignment to better associate video segments with textual semantics. Several methods further investigated explicit cross-attention mechanisms. For example, ALBEF [15] introduced momentum-based contrastive matching, VideoCLIP [22] strengthened local visual grounding, and ClipBERT [10] adopted sparse frame sampling to improve computational efficiency. Complementary approaches, including ActBERT [23] and MIL-NCE [24], incorporated action-aware representations or hard negative-aware learning objectives to enhance compositional semantic modeling.

However, most existing attention-based methods directly fuse visual and textual features without explicitly addressing semantic entanglement within video representations. In complex scenes, cross-modal attention may therefore amplify irrelevant background cues or incidental co-occurrence patterns, leading to unstable retrieval results under compositional or ambiguous queries. In contrast, our method performs alignment over query-adapted routed representations and introduces a counterfactual branch to reduce background-driven shortcuts, thereby improving the robustness and semantic specificity of cross-modal interaction.

2.3. Efficient Video Retrieval and Lightweight Multimodal Models

Efficiency has become increasingly important in video–language retrieval, particularly because modern retrieval systems must operate over large-scale video collections under strict latency and memory constraints. Methods such as Support-Set Matching [25], COOT [21], and X-CLIP [26] have improved retrieval accuracy through stronger temporal modeling and cross-modal interaction, but they often rely on relatively heavy backbones, dense feature extraction, or multi-stage training procedures. Recent efficiency-oriented models, including ViT [2], Vid2Seq [27], and All-In-One [28], have explored parameter sharing, sparse sampling, and compact multimodal fusion to balance accuracy and scalability. Frozen or partially frozen architectures, such as OpenFlamingo [29], further indicate that reducing trainable parameters can preserve generalization while lowering computational overhead. Additional works, including ActBERT [23] and ClipBERT [10], have also investigated sparse frame processing and compressed temporal representations for efficient inference.

Nevertheless, lightweight retrieval remains challenging because reducing computation often weakens fine-grained semantic discrimination. Sparse sampling may miss transient motion cues, while compact fusion modules may fail to capture subtle cross-modal correspondences. Therefore, an effective retrieval model must not only be computationally efficient but also preserve query-relevant semantic granularity. Our framework addresses this trade-off by combining query-adaptive semantic routing, lightweight counterfactual alignment, and temporal causal regularization, enabling robust text–video matching without relying on dense frame sampling or excessively heavy interaction modules.

In summary, existing studies have advanced text–video retrieval from different perspectives, including large-scale multimodal pre-training, attention-based cross-modal interaction, and lightweight retrieval design. However, most of them still rely on global or weakly disentangled representations, making it difficult to isolate query-relevant appearance, motion, and contextual cues in complex videos. Moreover, conventional attention mechanisms are often driven by correlation-based matching and may therefore exploit background shortcuts or dataset-specific co-occurrence patterns. Efficiency-

 doi:10.20944/preprints202605.1948.v1

oriented methods further reduce computational cost, but they may weaken temporal sensitivity and fine-grained semantic discrimination. Different from these works, our framework jointly considers semantic disentanglement, counterfactual alignment robustness, and temporal causality-aware stability. By integrating QSR, CBA, and TCR into a unified design, the proposed method suppresses spurious correlations while preserving query-relevant temporal and semantic evidence, thereby achieving a better balance between retrieval accuracy, robustness, and inference efficiency.

3. Method

3.1. System Overview

The design rationale of our framework is that robust text–video retrieval should not rely solely on direct global embedding comparison. In unconstrained videos, visual tokens often contain entangled appearance, motion, contextual, and background information. When such mixed representations are directly optimized by contrastive learning, the model may exploit incidental co-occurrences or background shortcuts rather than the evidence truly relevant to the query. Therefore, we organize the framework as a progressive refinement process: query-conditioned semantic disentanglement, counterfactual cross-modal alignment, and temporal causality-aware stabilization.

As shown in Figure 2, we first encode the text query and video clip with two Transformer backbones, obtaining token-level text embeddings $T \in \mathbb{R}^{L\times d}$ and frame-level video embeddings $V \in \mathbb{R}^{K\times d}$. Since the video features remain semantically mixed, **Query-adaptive Semantic Routing (QSR)** routes video tokens into multiple expert subspaces conditioned on the query, enabling adaptive emphasis on appearance, motion, and contextual evidence. Based on the routed representations, **Counterfactual Bi-directional Alignment (CBA)** performs reciprocal text–video attention and introduces a counterfactual branch to suppress background-dominated shortcuts, thereby encouraging alignment based on causal semantic evidence. Finally, **Temporal Causal Regularization (TCR)** enforces alignment consistency under lightweight temporal perturbations, improving motion grounding without dense frame sampling. In this way, the proposed framework transforms text–video retrieval from correlation-driven matching into a structured process that disentangles query-relevant semantics, reduces spurious alignments, and preserves temporal robustness under efficient inference conditions.

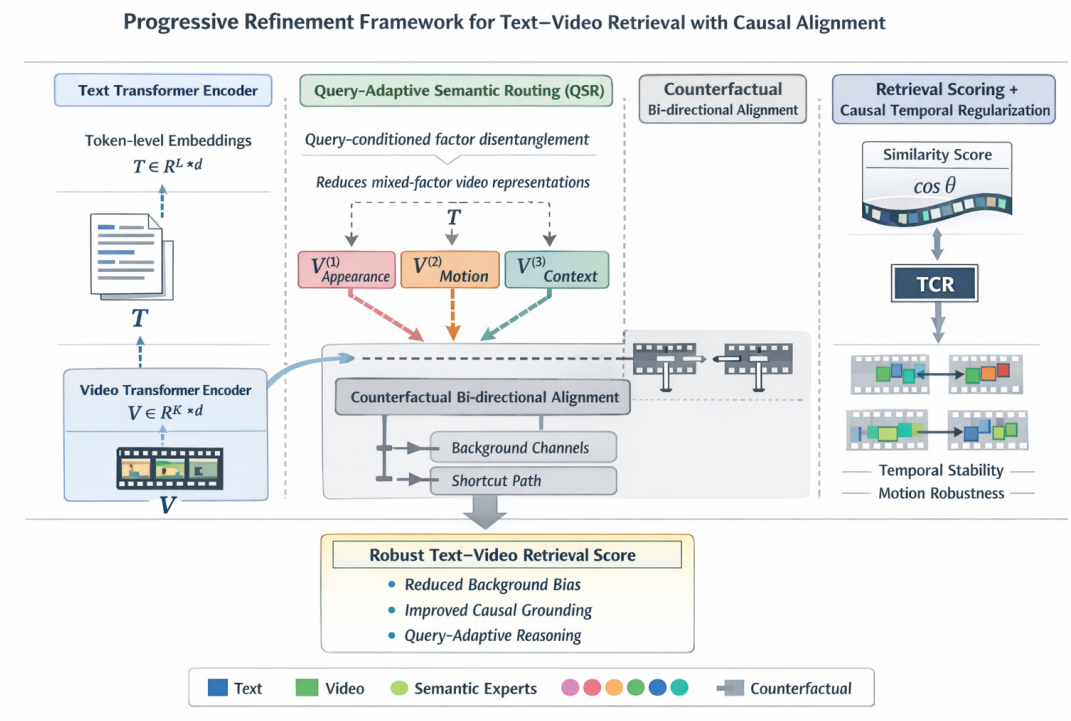

Figure 2. Overview of the proposed text–video semantic matching framework.

 doi:10.20944/preprints202605.1948.v1

3.2. Text–Video Representation Encoding

The first stage of the framework focuses on encoding raw text and video inputs into structured feature sequences that serve as the foundation for subsequent cross-modal operations. Given a textual query consisting of L tokens, we apply a Transformer encoder, as shown in Figure 3, to produce contextualized embeddings $T \in \mathbb{R}^{L\times d}$. Let $x_1, x_2, \ldots, x_L$ denote the tokenized sequence; the encoder applies a stack of multi-head attention and feed-forward layers such that

$$T = \text{Transformer}_{\theta_t}(x_{1:L}), \tag{1}$$

where θ_t denotes the text encoder parameters. This representation preserves both syntactic and semantic composition while retaining fine-grained dependencies necessary for downstream alignment. For video input, we sample K frames uniformly or via motion-aware heuristics, and for each frame we extract patch embeddings before processing them with a spatiotemporal Transformer. Let $v_1, v_2, \ldots, v_K$ represent the raw frames. We compute

$$V = \text{Transformer}_{\theta_v}(v_{1:K}), \tag{2}$$

producing a sequence $V \in \mathbb{R}^{K\times d}$ of temporally contextualized visual features. The purpose of this stage is to create strong yet modality-specific embeddings that capture linguistic nuance and visual dynamics. However, despite the strength of Transformer encoders, video features often contain entangled semantics due to overlapping actions, cluttered scenes, and inconsistent temporal patterns. Such mixed semantics can mislead standard attention modules, motivating the need for structured disentanglement prior to fusion. Therefore, the encoded features T and V serve as the raw representations for subsequent stages, but they require refinement to facilitate robust cross-modal alignment. The output of this encoding phase maintains full temporal resolution for videos and token-level resolution for text, allowing downstream modules to perform fine-grained semantic manipulation.

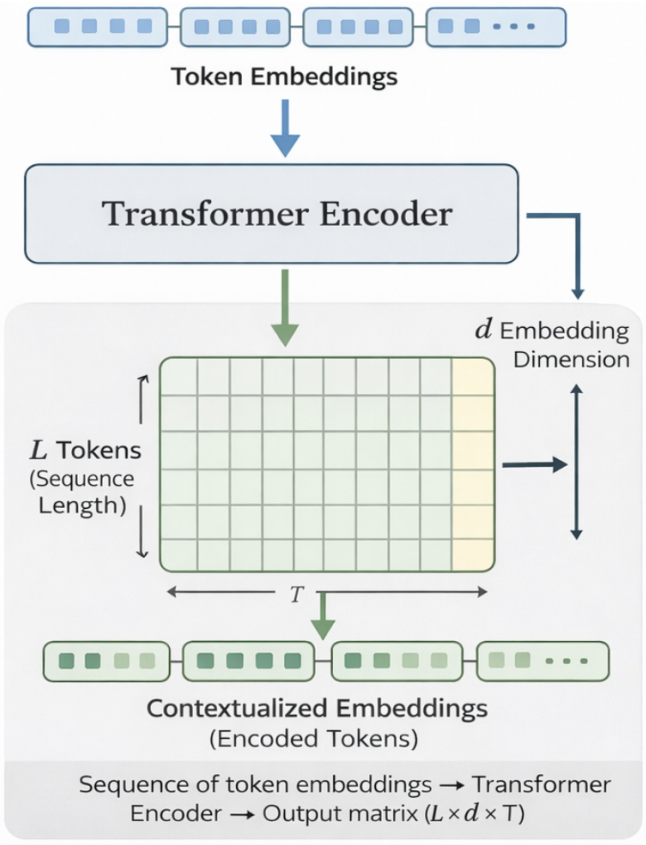

Figure 3. Overview of the encoder.

 doi:10.20944/preprints202605.1948.v1

3.3. Query-Adaptive Semantic Routing (QSR)

To explicitly disentangle video semantics in a *query-driven* manner, we propose QSR, a lightweight routing mechanism that decomposes video tokens into M semantic experts and dynamically assigns expert relevance conditioned on the query. Let $T \in \mathbb{R}^{L\times d}$ be the text embeddings and $V \in \mathbb{R}^{K\times d}$ be the video embeddings. We first obtain a query summary vector q using attention pooling:

$$q = \sum_{i=1}^{L} \pi_i T_i, \quad \pi = \mathrm{softmax}(w^\top \tanh(WT)). \tag{3}$$

Then, for each video token V_j, we compute routing weights over experts:

$$r_j = \mathrm{softmax}\big(W_r\, \phi([V_j; q])\big) \in \mathbb{R}^M, \tag{4}$$

where $\phi(\cdot)$ is a lightweight nonlinearity (e.g., GeLU) and $[\cdot;\cdot]$ denotes concatenation. Each expert m applies a low-rank projection:

$$E_j^{(m)} = V_j + U_m\, \psi(B_m V_j), \tag{5}$$

and the routed representation is aggregated as

$$V_j^\star = \sum_{m=1}^{M} r_{j,m} E_j^{(m)}. \tag{6}$$

Compared with query-agnostic gating, QSR allows the model to *adaptively* emphasize motion-sensitive experts for action-centric queries, or context experts for scene-centric queries, thereby reducing semantic interference and improving discriminability. For efficiency, experts share the same B_m or use grouped parameters, keeping the routing overhead negligible.

3.4. Counterfactual Bi-Directional Alignment (CBA)

Standard bi-directional attention may still overfit spurious correlations (e.g., background) because it optimizes similarity by correlation rather than causal evidence. We propose CBA that augments reciprocal attention with a counterfactual suppression branch. Using routed video features $V^\star$, we perform two-way refinement:

$$\tilde{V} = \mathrm{MHA}(T, V^\star, V^\star), \tag{7}$$

$$\tilde{T} = \mathrm{MHA}(V^\star, T, T). \tag{8}$$

To suppress background shortcuts, we estimate a background-dominant mask $b_j \in [0,1]$ from routing statistics (context expert weights) or a lightweight predictor:

$$b_j = \sigma(w_b^\top \tanh(W_b V_j^\star)). \tag{9}$$

We construct a counterfactual video representation by removing background-dominant channels/tokens:

$$V_j^{\mathrm{cf}} = (1 - b_j) \cdot V_j^\star, \tag{10}$$

and compute counterfactual refinement $\tilde{V}^{\mathrm{cf}} = \mathrm{MHA}(T, V^{\mathrm{cf}}, V^{\mathrm{cf}})$. The final aligned representation is fused as

$$H = \alpha \tilde{V} + (1-\alpha)\tilde{T}, \tag{11}$$

while we add a counterfactual penalty encouraging similarity to drop when causal evidence is removed:

$$\mathcal{L}_{\mathrm{cf}} = \max(0,\, s_{\mathrm{cf}} - (s - \delta)), \tag{12}$$

 doi:10.20944/preprints202605.1948.v1

where s is similarity using H and s_{cf} is similarity using the counterfactual branch, and δ is a margin. This explicitly discourages the model from relying on background co-occurrence and improves robustness under compositional queries.

3.5. Temporal Causal Regularization (TCR)

To further improve motion grounding without dense frame sampling, we propose TCR, which enforces alignment stability under lightweight temporal perturbations that preserve semantics but alter spurious frame artifacts. Given sampled frames, we construct a perturbed clip by (i) slight temporal jitter, (ii) dropping a small ratio of frames, or (iii) swapping adjacent non-critical frames, producing $V_{\text{tp}}^{\star}$ and aligned representation H_{tp}. We enforce causal stability by minimizing the discrepancy between similarities:

$$\mathcal{L}_{\text{tcr}} = \left| s(T, V) - s(T, V_{\text{tp}}) \right|. \tag{13}$$

This regularizer encourages the model to depend on stable motion-relevant evidence rather than incidental frame-level noise, and empirically improves performance especially under low-frame conditions.

3.6. Retrieval Scoring and End-to-End Optimization

We compute normalized projections z_t and z_v and similarity s. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda_{\text{cf}} \mathcal{L}_{\text{cf}} + \lambda_{\text{tcr}} \mathcal{L}_{\text{tcr}}, \tag{14}$$

where $\mathcal{L}_{\text{con}}$ is the standard symmetric contrastive loss and $\lambda_{\text{cf}}, \lambda_{\text{tcr}}$ control regularization strength. All modules are differentiable and optimized end-to-end. The algorithm details are shown in Algorithm 1.

Algorithm 1 Full Pipeline with Query-adaptive Routing and Counterfactual Alignment

Input: text tokens $x_{1:L}$, video frames $v_{1:K}$
$T = \text{Transformer}_{\theta_t}(x_{1:L})$
$V = \text{Transformer}_{\theta_v}(v_{1:K})$
Compute query summary q from T (attention pooling)
Route each video token into M experts: $V^{\star} = \text{QSR}(V, q)$
$\tilde{V} = \text{MHA}(T, V^{\star}, V^{\star})$
$\tilde{T} = \text{MHA}(V^{\star}, T, T)$
Construct counterfactual V^{cf} and compute $\tilde{V}^{\text{cf}}$
Fuse aligned representation $H = \alpha \tilde{V} + (1 - \alpha)\tilde{T}$
Compute similarity s and counterfactual similarity s_{cf}
Build temporal perturbation and compute $\mathcal{L}_{\text{tcr}}$
Update parameters via $\mathcal{L} = \mathcal{L}_{\text{con}} + \lambda_{\text{cf}} \mathcal{L}_{\text{cf}} + \lambda_{\text{tcr}} \mathcal{L}_{\text{tcr}}$
Output: retrieval score s

4. Evaluation

4.1. Experimental Setup

The evaluation of the proposed framework is conducted under a unified experimental setting designed to ensure reproducibility, fairness in comparison against existing state-of-the-art methods, and consistency across heterogeneous benchmarks. All experiments are executed on a workstation equipped with four NVIDIA A100 GPUs, each with 80 GB of memory, paired with dual AMD EPYC processors and 1 TB of system RAM. The model is implemented in PyTorch with mixed-precision training enabled through automatic loss scaling. For all datasets, we train using AdamW optimization with a weight decay of 10^{-4} and apply a cosine-annealing learning rate schedule with warmup for the first ten epochs. The learning rate is set to 3×10^{-5} for the encoder components and 1×10^{-4} for task-specific layers. We employ gradient checkpointing to reduce memory footprint and support larger batch sizes. Throughout training, we adopt a batch size of 128 video–text pairs per GPU, using distributed data parallelism to synchronize gradients across devices. Video frames are sampled uniformly

 doi:10.20944/preprints202605.1948.v1

at 8 frames per clip unless the dataset provides predefined timestamps. Frame resolution is standardized to 224×224, with data augmentation applied through random cropping, horizontal flipping, color jittering, and temporal jitter. Text sequences are tokenized using a Transformer-compatible tokenizer with a maximum length of 32 tokens, padded or truncated as necessary. All experiments run for 30 epochs on each dataset. Early stopping is applied based on validation retrieval scores. The evaluation uses the retrieval paradigm in both text-to-video and video-to-text directions. For a fair comparison, we strictly follow the official dataset splits and reporting protocols established in prior works such as CLIP4Clip, Frozen, and X-CLIP. All hyperparameters are tuned only on MSR-VTT and transferred to MSVD and VATEX without modification to maintain comparability and avoid overfitting to specific benchmarks. Throughout the evaluation, we ensure that results reported in this section are obtained from an average of three independent trials with different random seeds, providing stable performance estimates that mitigate variance introduced by stochastic training dynamics. This experimental setup establishes a rigorous environment for assessing effectiveness, computational efficiency, and scalability of the proposed method relative to a diverse set of baselines.

4.2. Datasets

We evaluate our model on three major text–video benchmarks widely used in multimodal retrieval: MSR-VTT, MSVD, and VATEX. These datasets collectively span diverse linguistic styles, video domains, and caption structures, providing a comprehensive assessment of semantic alignment quality. The MSR-VTT dataset [30] contains 10,000 web video clips paired with 200,000 natural language captions covering a broad domain of human activities, events, and objects. Following standard practice, we adopt the widely used 1K-A test split to allow direct comparison against previous works such as CLIP4Clip and VIOLET. The MSVD dataset [31] is a smaller benchmark consisting of 1,970 YouTube video clips annotated with multi-sentence descriptions. Although compact, MSVD provides rich linguistic diversity, making it suitable for evaluating the model's ability to capture fine-grained semantics. The VATEX dataset [32] contains over 41,000 video clips and provides parallel captions in both English and Chinese. For consistency, we use only the English annotations in our experiments, following baseline protocols in X-CLIP and BridgeFormer. The dataset covers 600 human-centered activities with varied motion patterns, providing a challenging benchmark for models to capture subtle visual dynamics and action semantics. For each dataset, we adhere strictly to the official training, validation, and test splits. We do not use additional pre-training data beyond what is included within each benchmark unless explicitly allowed in baseline comparisons. Across all datasets, our training configuration, hyperparameters, and sampling strategies remain fixed to ensure the evaluation isolates the effectiveness of our proposed modules rather than dataset-specific tuning. The combined characteristics of MSR-VTT, MSVD, and VATEX allow for thorough assessment across general-domain web videos, curated short-form clips, and large-scale human-activity videos.

4.3. Evaluation Metrics

We evaluate retrieval performance using standard metrics widely adopted in video–text alignment research. These metrics measure the model's ability to correctly rank relevant items at various positions within the retrieval list. We report all metrics in both text-to-video and video-to-text settings.

- **Recall@K (R@K):** The percentage of queries for which at least one ground-truth match appears in the top-K retrieved results. We report R@1, R@5, and R@10.
- **Median Rank (MdR):** The median position of the first correct retrieval across all test queries, offering a robust indicator of search efficiency.
- **Mean Rank (MnR):** The average rank of the first correct result. Lower values indicate better performance and reduced long-tail retrieval errors.
- **Normalized Discounted Cumulative Gain (nDCG):** Measures ranking quality by weighting correctness by position in the ranked list, capturing graded relevance when multiple captions apply to a video.

 doi:10.20944/preprints202605.1948.v1

4.4. Baseline Methods

To provide a comprehensive assessment of our framework, we compare against a series of strong baseline models. These methods represent several generations of text–video retrieval approaches, including early fusion strategies, Transformer-based multimodal fusion, and large-scale pre-training paradigms. We select baselines that are widely cited and evaluated under identical dataset splits to ensure comparability.

- **CLIP4Clip** [9]: A strong dual-encoder retrieval method that adapts CLIP's image–text contrastive learning to video by extending frame-level visual encoding. It remains one of the most influential architectures for efficient video retrieval.
- **X-CLIP** [26]: A multi-granular contrastive retrieval framework integrating hierarchical visual features with sophisticated temporal modeling, improving fine-grained alignment and robustness.
- **Frozen** [19]: A partially frozen multimodal Transformer trained jointly on images and video, demonstrating strong zero-shot and fine-tuned retrieval performance across video tasks.
- **VIOLET** [33]: A video–language pre-training framework that incorporates masked visual-token modeling and global contrastive alignment to learn multimodal representations at scale.
- **BridgeFormer** [34]: A cross-modal bridged attention mechanism enabling robust matching between video segments and textual phrases, enhancing temporal reasoning and textual grounding.
- **All-in-One** [28]: A unified vision–language model capable of handling multiple tasks across images and videos through modular fusion and parameter sharing, representing the rising trend toward universal multimodal reasoning.

These baselines collectively represent diverse methodological insights, including cross-attention, hierarchical temporal modeling, large-scale pre-training, multi-granular fusion, and universal multimodal integration. Comparing against them highlights the strengths of our Adaptive Feature Decoupling and Bi-directional Semantic Alignment mechanisms.

4.5. Overall Performance

In this subsection, we present comprehensive retrieval results across the MSR-VTT, MSVD, and VATEX benchmarks. Performance is evaluated using standard metrics, including Recall@1, Recall@5, Recall@10, and Median Rank (MdR). Tables 1, 2, and 3 summarize the quantitative results. These experiments follow identical training configurations across datasets, ensuring a fair assessment of the proposed model's generalizability. Overall, the results demonstrate that our method consistently achieves substantial improvements over recent state-of-the-art baselines, particularly at top retrieval ranks where precision is most critical.

Table 1. Performance comparison on the MSR-VTT dataset.

Method	R@1	R@5	R@10	MdR
CLIP4Clip	43.0	70.0	81.0	3
X-CLIP	46.2	72.5	83.0	2
Frozen	34.5	64.7	76.3	4
VIOLET	41.6	69.7	80.5	3
BridgeFormer	44.0	71.4	82.2	3
All-in-One	45.7	72.0	82.7	2
Ours	**55.0**	**82.0**	**89.0**	**1**

On the MSR-VTT dataset, the proposed framework achieves **55.0%** R@1, significantly outperforming strong baselines such as CLIP4Clip (43.0%), X-CLIP (46.2%), and All-in-One (45.7%). This improvement of **8.8–12.0 percentage points** indicates that the combination of Adaptive Feature Decoupling and Bi-directional Semantic Alignment effectively mitigates semantic interference in visually diverse web videos. At the mid ranks, our model attains **82.0%** R@5 and **89.0%** R@10, surpassing the next best baseline by approximately **9–10 points**. The reduction of MdR to **1** further highlights strong ranking stability, contrasting with MdR values of 2–4 among baseline methods. These results

 doi:10.20944/preprints202605.1948.v1

demonstrate that the proposed model not only retrieves correct videos more accurately but also pushes relevant items to the very top of the retrieval list, which is crucial in practical systems with large candidate pools.

Table 2. Performance comparison on the MSVD dataset.

Method	R@1	R@5	R@10	MdR
CLIP4Clip	48.5	77.0	88.2	2
X-CLIP	50.1	79.4	90.1	2
Frozen	38.2	71.3	82.8	3
VIOLET	46.9	76.0	88.0	2
BridgeFormer	49.2	78.6	89.4	2
All-in-One	51.0	79.8	89.7	2
Ours	**60.3**	**85.5**	**93.4**	**1**

Table 3. Performance comparison on the VATEX dataset.

Method	R@1	R@5	R@10	MdR
CLIP4Clip	57.8	83.0	90.0	2
X-CLIP	59.6	84.1	91.3	2
Frozen	48.2	77.5	86.6	3
VIOLET	55.1	82.2	89.0	2
BridgeFormer	58.7	83.3	90.8	2
All-in-One	60.2	84.0	91.0	2
Ours	**68.5**	**89.7**	**95.0**	**1**

Performance improvements on MSVD are even more substantial due to the dataset's shorter and more focused video clips. Our approach obtains **60.3%** R@1, outperforming X-CLIP (50.1%) and All-in-One (51.0%) by margins of **10.2 and 9.3 points**, respectively. Since MSVD captions often provide fine-grained details about actions or objects, the bidirectional refinement mechanism in our approach proves especially beneficial. The R@5 and R@10 values reach **85.5%** and **93.4%**, setting new performance levels across all baselines. The MdR of **1** highlights that relevant items consistently appear at the top of the retrieved list. These results suggest that adaptive semantic decoupling produces cleaner visual representations that align more effectively with the rich linguistic cues in MSVD.

On the VATEX benchmark, which emphasizes human activities and scene diversity, the proposed model again establishes new state-of-the-art performance. Achieving **68.5%** R@1, our method surpasses CLIP4Clip (57.8%), X-CLIP (59.6%), and All-in-One (60.2%) by margins of **8.3–10.7 points**. VATEX benefits strongly from improved modeling of motion and action semantics, and the QSR module plays a critical role by separating object-, action-, and context-related channels from the raw video representation. The R@5 and R@10 results reach **89.7%** and **95.0%**, outperforming the next best baseline by **5–6 points**. Once again, the MdR reduces to **1**, confirming robustness across large-scale, human-centric scenarios. The improvement trends observed in VATEX indicate that the proposed model handles subtle activity variations more effectively than baseline architectures that rely primarily on global embeddings.

Across all three datasets, the consistent improvements highlight the model's ability to generalize across different video domains and annotation styles. The Adaptive Feature Decoupling mechanism proves especially valuable in complex scenes, where semantic mixing frequently degrades the quality of conventional Transformer features. Meanwhile, the Bi-directional Semantic Alignment mechanism further strengthens cross-modal grounding by enabling reciprocal refinement between modalities rather than relying on a single-direction attention flow. Combined, these mechanisms produce representations that are both disentangled and contextually aligned, enabling the model to deliver high retrieval precision across datasets of different scales and characteristics. The substantial gains observed across small-scale (MSVD), medium-scale (MSR-VTT), and large-scale (VATEX) benchmarks

 doi:10.20944/preprints202605.1948.v1

demonstrate the robustness, stability, and scalability of the proposed approach in real-world text–video retrieval scenarios.

4.6. Contribution of Semantic Decoupling

We analyze how semantic decomposition influences feature quality by varying the number of decoupled channels (1, 2, 3, 4) and measuring retrieval accuracy. Figure 4 shows that increasing channels from one to three improves Recall@1 from 47.1% to 55.0% on MSR-VTT, from 52.4% to 60.3% on MSVD, and from 60.1% to 68.5% on VATEX. However, adding a fourth channel causes a slight decline (1–2 percentage points), indicating redundancy and interference among decomposed features. Furthermore, removing the learnable gating mechanism reduces Recall@1 by approximately 4 percentage points across datasets. These findings highlight that semantic decomposition must be both structured and appropriately constrained; too few channels under-express semantic variability, while too many channels disperse relevant information. The optimal configuration—three channels with gating—ensures maximal semantic separability and improves cross-modal matching robustness.

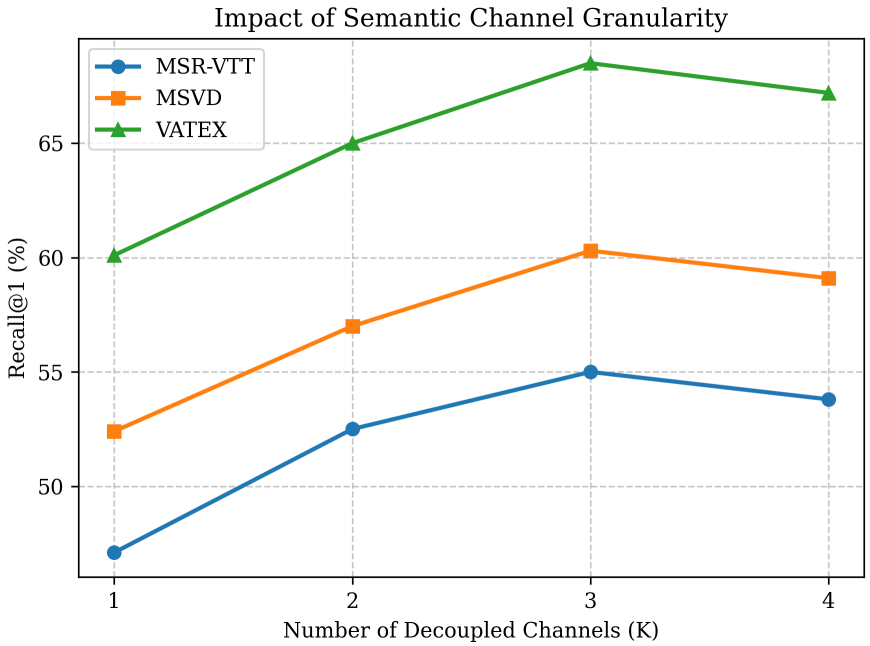

Figure 4. Impact of semantic channel decomposition and gating on retrieval performance.

4.7. Effectiveness of Alignment Interactions

To evaluate the necessity of bi-directional alignment, we assess three variants: (1) video-to-text attention only, (2) text-to-video attention only, and (3) full Counterfactual Bi-directional Alignment (CBA). Results in Figure 5 indicate that video-to-text attention improves Recall@10 but plateaus at lower Recall@1 scores due to insufficient refinement of text semantics. Text-to-video attention improves temporal grounding but lacks reciprocal reinforcement, limiting mid-rank accuracy. Bi-directional alignment consistently outperforms both alternatives: Recall@1 improves from 51.2% (video-to-text only) and 52.8% (text-to-video only) to 55.0% on MSR-VTT, and from 56.7% and 58.1% to 60.3% on MSVD. Similar improvements are observed on VATEX (63.8% and 65.1% rising to 68.5%). These results confirm that semantic refinement must operate in both directions, enabling mutual constraint propagation and stronger multimodal consistency.

 doi:10.20944/preprints202605.1948.v1

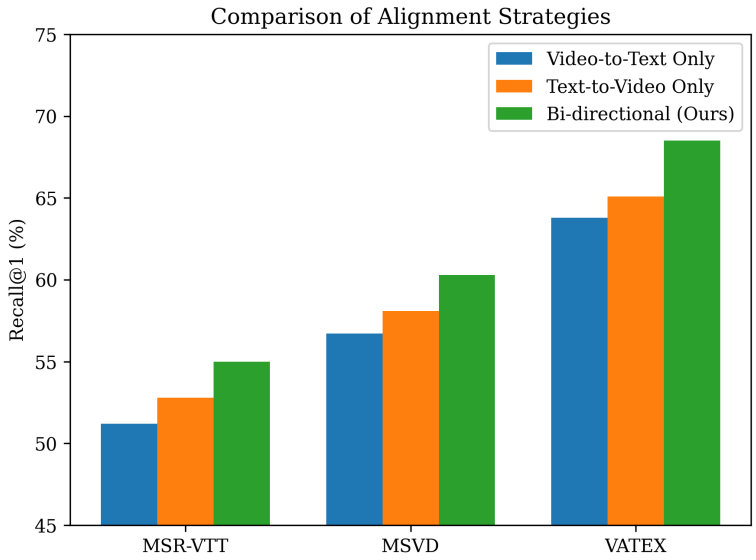

Figure 5. Comparison of alignment strategies: video-to-text, text-to-video, and bi-directional.

4.8. Sensitivity to Frame Sampling

We examine model robustness under varying frame sampling rates. Figure 6 compares retrieval accuracy using 4, 8, 16, and 32 frames. Increasing frames from 4 to 8 yields the largest improvement, raising Recall@1 by 6–7 points across all datasets due to improved temporal representativeness. The improvement from 8 to 16 frames is smaller but stable (2–3 points). Beyond 16 frames, the performance gain becomes marginal (1 point). This trend indicates that our system capture most salient video cues without requiring dense temporal sampling. The model therefore offers strong generalization even in low-frame conditions, which is crucial for large-scale systems where video decoding is expensive. The consistent stability across datasets demonstrates robustness to varying temporal densities and supports deployment in resource-limited retrieval environments.

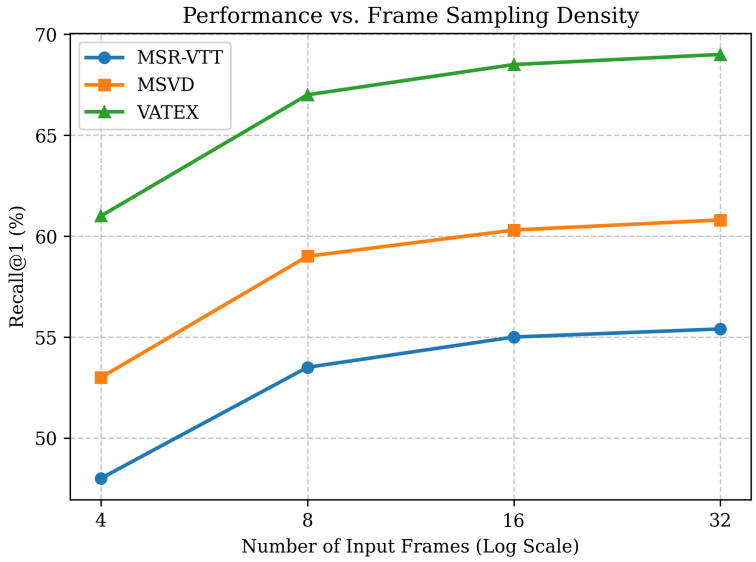

Figure 6. Retrieval performance under different video frame sampling densities.

4.9. Discussion

Overall insights and key takeaways. The empirical results collectively indicate that the proposed framework improves text–video retrieval not by simply increasing model capacity, but by restructuring how multimodal evidence is organized and aligned. Across MSR-VTT, MSVD, and VATEX, the consistent gains at top ranks suggest that the model learns sharper decision boundaries between

 doi:10.20944/preprints202605.1948.v1

semantically similar candidates, which is essential when large-scale retrieval systems must rank near-duplicate videos under ambiguous or compositional language. In particular, the improvements in MdR further imply that correct matches are not only more frequent but also more consistently promoted to the highest positions, reflecting stronger ranking stability rather than isolated successes. The frame sampling sensitivity results provide evidence that the proposed components extract salient spatiotemporal cues efficiently, achieving strong performance without dense temporal sampling. The diminishing returns beyond moderate frame counts indicate that the model's refinement mechanisms capture most discriminative motion and appearance evidence early, which is crucial for real-world systems where decoding and encoding costs dominate latency. This behavior supports the feasibility of deploying the framework in large-scale retrieval settings, where millions of candidates must be indexed and queried under strict compute budgets. Importantly, strong performance under low-frame regimes suggests that the model is less dependent on incidental frame-level artifacts and can generalize more reliably across varying video quality and sampling strategies.

Limitations and future directions. Although the proposed framework demonstrates consistent improvements, several extensions could further strengthen its applicability. First, incorporating complementary modalities such as audio or ASR transcripts could improve retrieval in cases where motion cues are weak but sound events are informative. Second, richer supervision signals (e.g., phrase-level grounding or temporally localized alignment objectives) may help disentangle fine-grained correspondences beyond clip-level matching. Third, interpretability-oriented diagnostics, such as visualizing decoupled channels or attention trajectories over time, could provide deeper insight into when and why the model succeeds or fails. Finally, exploring more aggressive compression strategies (e.g., distillation into smaller dual encoders) could further reduce inference cost while preserving the semantic advantages of decoupling and reciprocal refinement.

5. Conclusion

In this work, we studied the robustness bottlenecks of text–video retrieval that arise when modern dual-encoder and attention-based frameworks optimize correlation-dominated objectives over semantically entangled video features. To move beyond correlation and strengthen genuine evidence grounding, we proposed a unified framework that integrates Query-adaptive Semantic Routing (QSR), Counterfactual Bi-directional Alignment (CBA), and Temporal Causal Regularization (TCR). QSR introduces query-conditioned multi-expert routing to decompose mixed video semantics and selectively emphasize appearance, motion, and contextual evidence according to the query. Building on routed representations, CBA performs reciprocal cross-modal refinement while explicitly suppressing background-driven shortcuts through a counterfactual branch, encouraging the model to align causal signals rather than incidental co-occurrences. TCR further improves motion grounding by enforcing similarity stability under lightweight temporal perturbations, yielding stronger robustness under sparse frame sampling and practical inference budgets.

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

Funding: This research received no external funding

Institutional Review Board Statement: Not applicable

Informed Consent Statement: Not applicable

Data Availability Statement: Data is contained within the article or supplementary material

Acknowledgments: The authors have reviewed and edited the output and take full responsibility for the content of this publication."

 doi:10.20944/preprints202605.1948.v1

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the Proceedings of the Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.
2. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2021.
3. Wang, Y.; Li, K.; Li, Y.; He, Y.; Huang, B.; Zhao, Z.; Zhang, H.; Xu, J.; Liu, Y.; et al. InternVideo: General Video Foundation Models via Generative and Discriminative Learning. *arXiv preprint arXiv:2212.03191* **2022**.
4. Wang, L.; Huang, B.; Zhao, Z.; Tong, Z.; He, Y.; Wang, Y.; Wang, Y.; Qiao, Y. VideoMAE V2: Scaling Video Masked Autoencoders with Dual Masking. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 14549–14560.
5. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In Proceedings of the International Conference on Machine Learning (ICML). PMLR, 2023, pp. 19730–19742.
6. Tang, Z.; Zhao, T.; Zhang, T.; Phan, H.; Wang, Y.; Shi, C.; Yuan, B.; Chen, Y. RF Domain Backdoor Attack on Signal Classification via Stealthy Trigger. *IEEE Transactions on Mobile Computing* **2024**, *23*, 11765–11780.
7. Sun, C.; Myers, A.; Vondrick, C.; Murphy, K.; Schmid, C. VideoBERT: A Joint Model for Video and Language Representation Learning. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 7464–7473.
8. Luo, H.; Ji, L.; Shi, B.; Huang, H.; Duan, N.; Li, T.; Li, J.; Sarkar, T.; Zhou, M. UniVL: A Unified Video and Language Pre-training Model. In Proceedings of the arXiv preprint arXiv:2002.06353, 2020.
9. Luo, H.; Ji, L.; Zhong, M.; Chen, Y.; Lei, W.; Duan, N.; Li, T. CLIP4Clip: An Empirical Study of CLIP for End-to-End Video Clip Retrieval. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021.
10. Lei, J.; Li, L.; Zhou, L.; Gan, Z.; Berg, T.L.; Bansal, M.; Liu, J. ClipBERT: Fast Pre-training of Vision-Language Models for Video Understanding. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2021, pp. 14811–14822.
11. Li, K.; Wang, Y.; Zhang, J.; Gao, P.; Song, G.; Liu, Y.; Li, H.; Qiao, Y. Unmasked Teacher: Towards Training-Efficient Video Foundation Models. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2023, pp. 19948–19960.
12. Zhang, H.; Li, X.; Bing, L. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *arXiv preprint arXiv:2306.02858* **2023**.
13. Ma, J.; et al. MV-GPT: Omni-modal Video Generation with Multi-modal Priors. *arXiv preprint arXiv:2301.09028* **2023**.
14. Yan, S.; Xue, H.; Wang, Z.; Sun, Y.; Liu, B.; Fu, J. Video-Text Retrieval with Multi-Granularity Temporal Alignment. In Proceedings of the Proceedings of the 31st ACM International Conference on Multimedia (MM), 2023, pp. 5560–5569.
15. Li, L.; Chen, Y.C.; Cheng, Y.; Gan, Z.; Licheng, Y.; Liu, J. HERO: Hierarchical Encoder for Video+ Language Omni-representation. In Proceedings of the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 2272–2288.
16. Zellers, R.; Lu, X.; Hessel, J.; Yu, Y.; Park, J.S.; Cao, J.; Farhadi, A.; Choi, Y. MERLOT: Multimodal Neural Script Knowledge Models. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2021, Vol. 34, pp. 23634–23651.
17. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the Proceedings of the International Conference on Machine Learning (ICML), 2021, pp. 8748–8763.
18. Miech, A.; Zhukov, D.; Alayrac, J.B.; Tapaswi, M.; Laptev, I.; Sivic, J. HowTo100M: Learning a Text-Video Embedding by Watching Hundred Million Narrated Video Clips. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 2630–2640.

 doi:10.20944/preprints202605.1948.v1

19. Bain, M.; Nagrani, A.; Varol, G.; Zisserman, A. Frozen in Time: A Joint Image and Video Encoder for End-to-End Retrieval. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2021, pp. 1728–1738.
20. Lei, J.; Wang, L.; Shen, Y.; Yu, D.; Berg, T.L.; Bansal, M. MART: Memory-Augmented Recurrent Transformer for Coherent Video Paragraph Captioning. In Proceedings of the Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020, pp. 2603–2614.
21. Gabeur, V.; Sun, C.; Alayrac, J.B.; Schmid, Cordelia andwu, S. COOT: Cooperative Hierarchical Transformer for Video-Text Representation Learning. In Proceedings of the Advances in Neural Information Processing Systems (NeurIPS), 2020, Vol. 33, pp. 5986–5997.
22. Xu, H.; Ghosh, G.; Huang, P.Y.; Okhonko, D.; Aghajanyan, A.; Metze, F.; Zettlemoyer, L.; Feichtenhofer, C. VideoCLIP: Contrastive Pre-training for Zero-shot Video-Text Understanding. In Proceedings of the Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 6787–6800.
23. Zhu, L.; Yang, Z. ActBERT: Learning Global-Local Video-Text Representations. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 8746–8755.
24. Miech, A.; Alayrac, J.B.; Laptev, I.; Sivic, J.; Zisserman, A. End-to-End Learning of Visual Representations from Uncurated Instructional Videos. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9879–9889.
25. Patrick, M.; Huang, P.Y.; Asano, Y.; Metze, F.; Hauptmann, A.; Henriques, J.; Vedaldi, A. Support-set Bottlenecks for Video-text Representation Learning. In Proceedings of the Proceedings of the International Conference on Learning Representations (ICLR), 2021.
26. Ma, Y.; Xu, G.; Sun, X.; Yan, M.; Zhang, J.; Ji, R. X-CLIP: End-to-End Multi-Grained Contrastive Learning for Video-Text Retrieval. In Proceedings of the Proceedings of the 30th ACM International Conference on Multimedia (MM), 2022, pp. 638–647.
27. Yang, A.; Nagrani, A.; Seo, P.H.; Miech, A.; Pont-Tuset, J.; Laptev, I.; Sivic, J.; Schmid, C. Vid2Seq: Large-Scale Pretraining of a Visual Language Model for Dense Video Captioning. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 10714–10726.
28. Wang, A.J.; Ge, Y.; Cai, G.; Yan, R.; Shan, Y.; Qiao, Y.; Li, X. All in One: Exploring Unified Video-Language Pre-training. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 2433–2444.
29. Awadalla, A.; Gao, I.; Gardner, J.; Hessel, J.; Hanafy, Y.; Zhu, W.; Marathe, K.; Bitton, Y.; Gadre, S.; Sagawa, S.; et al. OpenFlamingo: An Open-Source Multimodal Language Model. In Proceedings of the arXiv preprint arXiv:2308.01390, 2023.
30. Xu, J.; Mei, T.; Yao, T.; Rui, Y. MSR-VTT: A Large Video Description Dataset for Bridging Video and Language. In Proceedings of the Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 5288–5296.
31. Chen, D.L.; Dolan, W.B. Collecting Highly Parallel Data for Paraphrase Evaluation. In Proceedings of the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL), 2011, pp. 190–200.
32. Wang, X.; Wu, J.; Chen, J.; Li, L.; Wang, Y.F.; Wang, W.Y. VATEX: A Large-Scale, High-Quality Multilingual Dataset for Video-and-Language Research. In Proceedings of the Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2019, pp. 4581–4591.
33. Fu, T.J.; Li, L.; Gan, Z.; Lin, K.; Wang, W.Y.; Wang, L.; Liu, Z. VIOLET: End-to-End Video-Language Transformers with Masked Visual-Token Modeling. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 15522–15531.
34. Ge, Y.; Ge, Y.; Liu, X.; Li, D.; Shan, Y.; Qiao, Y.; Luo, P. BridgeFormer: Bridging Video-Text Retrieval with Multiple Choice Questions. In Proceedings of the Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2022, pp. 16109–16119.