TCAR-Net: Text-Driven Compressed Action Recognition with Multimodal Fusion

Mengkun Guo, Xinqi Li, Die Tao, Ming Ma*

Inner Mongolia University

Hohhot 010021, China

Abstract

Action recognition directly from compressed video streams (leveraging I-frames, motion vectors, residuals) promises significant efficiency gains over pixel-based methods, but faces inherent challenges in achieving deep semantic understanding, especially when integrating rich textual priors from Vision-Language Models (VLMs). The noisy and sparse nature of motion vectors and residuals, complicates direct fusion with fine-grained semantics. To bridge this gap efficiently, this work investigates adapting cross-modal fusion techniques within a parameter-efficient framework tailored for the compressed domain. We introduce TCAR-Net, a text-driven dual-stream architecture built on largely frozen pre-trained backbones. Its spatial branch processes I-frames and residuals, while the motion branch encodes motion vectors (via a frozen ViT); both streams integrate textual guidance using adapted fusion modules. Crucially, only these lightweight cross-modal fusion components are fine-tuned, minimizing adaptation costs. Experiments on UCF101 and HMDB51 demonstrate TCAR-Net achieves competitive accuracy while operating directly on compressed data, significantly reducing decoding overhead. Our findings validate that adapting existing fusion strategies within a parameter-efficient setup is a feasible and effective approach for enabling semantically rich action recognition directly in the compressed domain, offering a practical pathway for efficient video understanding.

Keywords: Video Action Recognition; Compression-domain Features; Vision-Language Models; Multi-modal Learning

1. Introduction

In recent years, video action recognition technology, as the core for understanding massive video content, has been continuously driving the development of key applications such as intelligent surveillance and autonomous driving[1]. Based on the data domain processing approach, this task is currently divided into two categories: pixel domain and compressed domain action recognition. The first group directly extracts spatiotemporal features from the raw RGB frame sequences[2, 3, 4], offering high potential accuracy but often incurring significant computational and storage costs due to reliance on fully decoded pixels and complex motion estimation (like

*Corresponding author: Ming Ma (Email:csmaming@imu.edu.cn; ORCID:0000-0002-2060-2266)

optical flow), severely limiting real-time performance. To address this efficiency bottleneck, the second group focuses on action recognition directly from compressed video streams, leveraging readily available components like I-frames, motion vectors (MV), and residuals[5, 6, 7, 8, 9]. This approach promises substantial efficiency gains by avoiding full decoding and motion recomputation. However, existing compressed domain methods face significant hurdles in achieving deep semantic understanding. Firstly, they often focus on single-modal modeling of compressed features, neglecting the crucial guiding role text modality can play in reasoning across multiple granularities. Secondly, the inherent nature of compressed data, particularly the noisy and sparse motion vectors derived from block matching, introduces semantic ambiguity when used directly for recognition without stronger semantic constraints.

To overcome these limitations and effectively infuse rich semantic understanding into efficient compressed video analysis, this paper introduces TCAR-Net, a framework designed for efficient text-guided action recognition directly from compressed video streams. Our approach pioneers a synergistic integration of powerful pre-trained vision-language models (specifically CLIP) with a parameter-efficient fine-tuning (PEFT) strategy tailored for the compressed domain. At its core, TCAR-Net leverages frozen, high-capacity backbones: CLIP's visual encoders (pre-adapted on Kinetics-400) process the distinct signals from I-frames/residuals and motion vectors, while a frozen CLIP text encoder provides robust language representations. This strategic freezing preserves invaluable pre-trained knowledge while drastically reducing adaptation costs.

Bridging these potent but frozen priors relies on a minimalist yet effective adaptation strategy: temporal dependencies are captured by a dedicated trainable sequence model, while textual guidance is efficiently infused using a parameter-free mechanism [10] that leverages CLIP's intrinsic cross-modal alignment to highlight semantic relevance, crucial for navigating compressed data characteristics. Therefore, TCAR-Net tackles the core challenge by enabling deep semantic reasoning directly on compressed video through a novel PEFT scheme. Instead of costly full fine-tuning or designing complex new fusion modules from scratch, our innovation lies in demonstrating how to efficiently orchestrate existing powerful components (frozen CLIP variants, a standard trainable sequence model, and a parameter-free saliency mechanism). This specific architecture proves highly effective for adapting large models to the compressed domain, achieving strong performance while maintaining the crucial efficiency benefits inherent to compressed stream processing.

Our main contributions are summarized as follows:

- We demonstrate the application of Parameter-Efficient Fine-Tuning (PEFT) for text-driven action recognition specifically within the compressed video domain. By leveraging frozen CLIP-based backbones and training only minimal components, our approach significantly reduces adaptation costs while enabling the integration of strong semantic priors.
- We propose TCAR-Net, a carefully designed architecture that efficiently adapts powerful pre-trained models (CLIP+K400) to compressed video signals. This involves using a trainable sequence model for temporal modeling followed by a parameter-free mechanism to inject text guidance, effectively handling noise and sparsity in compressed data.
- We achieve competitive action recognition accuracy on standard benchmarks (UCF101, HMDB51) while operating directly on compressed data. Our results validate that the proposed PEFT strategy and architecture offer a practical and effective pathway for semantically rich video understanding without the overhead of full decoding, balancing accuracy and efficiency.

2. Related Work

2.1. Pixel Domain Video Action Recognition

Traditional video action recognition methods primarily focus on spatiotemporal modeling in the pixel domain. Two-Stream Networks [2] pioneered this direction by leveraging RGB frames for static appearance features and optical flow for motion cues. Subsequent works further improved spatiotemporal representation learning through architectures like 3D convolutions [3] and spatiotemporal factorized convolutions [4]. However, the reliance on optical flow estimation introduces a massive computational overhead, severely limiting real-time performance and hindering deployment on resource-constrained edge devices. Moreover, while these methods enhance semantic reasoning, they still suffer from inefficiencies due to frame-wise decoding and motion recomputation. The discrete nature of RGB frame sampling also poses challenges in capturing long-range temporal dependencies, making it difficult to balance accuracy and efficiency in deployment scenarios with limited computational resources. In contrast, our work operates directly on compressed video streams, avoiding the computationally expensive optical flow estimation and full frame decoding, leading to significant efficiency gains.

2.2. Compressed Video Representation Learning

To reduce video decoding overhead, compressed domain methods directly model the inherent features of video encoding standards (e.g., H.264/HEVC). Early work, CoViAR[5], demonstrated the effectiveness of I-frames, motion vectors (MV), and residuals in action recognition by independently encoding compressed features using lightweight networks. Subsequent research optimized compressed domain representations through strategies such as spatiotemporal attention and multi-scale feature fusion[11, 12, 13, 14], but they are still limited by the following bottlenecks: 1) I-frames, due to large keyframe intervals, lead to the loss of continuous motion information, hindering the capture of fine-grained action details; 2) The block-level discretization of MVs makes them vulnerable to encoding noise introduced during compression, leading to inaccurate motion representation; 3) Existing methods often stack single-modal features and lack cross-modal collaboration mechanisms, failing to effectively leverage the complementary information from different compressed features.

Consequently, relying solely on these inherently noisy and potentially sparse compressed features (Points 1-3) makes it challenging to achieve robust fine-grained semantic distinction between similar actions. This motivates leveraging external high-level semantic information, such as textual guidance, to disambiguate representations and improve recognition accuracy in complex scenarios.

These limitations often result in poor performance in complex scenarios with rapid motion or significant lighting variations. Our work addresses these limitations by introducing a text-driven dual-stream architecture that effectively integrates information from I-frames, motion vectors, and residuals, while leveraging textual guidance to mitigate the noise and sparsity inherent in compressed data.

2.3. Cross-Modal Video Semantic Understanding

Recently, the rise of visual-language models (VLMs), such as CLIP[15] and Florence[16], has significantly advanced semantic understanding through cross-modal alignment. This paradigm offers new possibilities for semantic reasoning in video tasks. Pioneering works like ActionCLIP[17] and E-prompt[18] have successfully leveraged VLMs for video comprehension, achieving impressive results in pixel-domain action recognition. These methods typically employ strategies

such as adapting pre-trained models through fine-tuning (*e.g.*, ActionCLIP's "pretrain, prompt and finetune" framework) or optimizing learnable prompts (*e.g.*, E-prompt's approach).

The success of these approaches powerfully demonstrates the effectiveness and potential of using textual semantics to guide video understanding. However, their operation within the pixel domain inherently necessitates full decoding of RGB frame sequences.

Consequently, while the value of text guidance for action recognition has been clearly validated in the pixel domain, effectively and efficiently harnessing this capability directly within the compressed domain remains an open challenge. Existing compressed domain research has largely focused on single-modal feature encoding, lacking mechanisms to deeply integrate textual semantics with compressed features. This gap limits their ability to achieve the level of semantic discrimination demonstrated by state-of-the-art pixel-domain VLM approaches, especially in complex scenarios.

Our framework, TCAR-Net, addresses this gap by pioneering the use of Parameter-Efficient Fine-Tuning (PEFT) to adapt large Vision-Language Models (like CLIP) for effective text-driven semantic understanding directly within the compressed video domain. By employing a carefully designed parameter-efficient fine-tuning (PEFT) strategy, TCAR-Net efficiently adapts powerful pre-trained VLMs (CLIP) to compressed video data, enabling semantically rich action recognition without the computational overhead of full decoding or extensive fine-tuning, thereby bridging the gap between the semantic power validated in the pixel domain and the efficiency required for practical compressed domain applications.

3. The Proposed Approach

3.1. Overview: A PEFT Framework for Compressed Video

Our proposed method, TCAR-Net, introduces text-guided semantic understanding directly within the compressed video domain through a Parameter-Efficient Fine-Tuning (PEFT) strategy that adapts large-scale Vision-Language Models while preserving both efficiency and semantic capabilities.

PEFT Strategy Rationalization. Our approach is grounded in three key principles: *(1) Cross-Modal Alignment Preservation:* CLIP's pre-trained visual-textual alignment remains effective for compressed features through parameter-free text guidance. *(2) Domain-Adaptive Extraction:* The frozen visual encoder maintains representational capacity across compressed modalities, with I-frames leveraging similarity to natural images while residuals and motion vectors benefit from semantic constraints. *(3) Efficient Temporal Adaptation:* Separating spatial extraction (frozen) from temporal modeling (trainable) enables video adaptation without compromising pre-trained spatial knowledge.

TCAR-Net strategically minimizes adaptation costs through:

- **Freezing Backbones:** A frozen CLIP visual encoder ($\mathcal{V}$) processes all compressed signals (I-frames, Residuals, Motion Vectors) and a frozen CLIP text encoder provides semantic embeddings (E_T), preserving cross-modal alignment while reducing trainable parameters.

- **Training Minimal Components:** Only essential components for temporal modeling: trainable sequence models ($\mathcal{T}_{seq}$) and adaptive fusion weights ($\boldsymbol{\beta}$) for spatial and temporal stream combination.

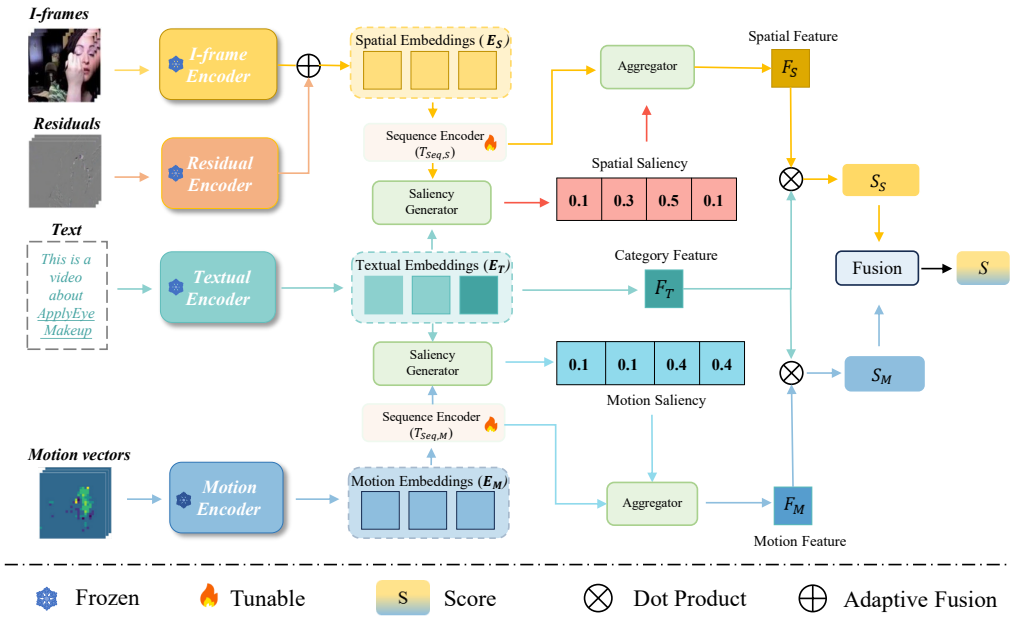

Figure 1: **The overall framework of our proposed method.** The framework integrates I-frames, residuals, motion vectors, and textual descriptions to enhance video understanding. The I-frame Encoder and Residual Encoder extract spatial features, while the Motion Encoder captures motion dynamics from compressed video representations. The Textual Encoder encodes semantic information from textual descriptions. The extracted Spatial Embeddings and Motion Embeddings are processed through a Saliency Generator, which assigns saliency scores to guide feature selection. The Aggregator refines the spatial and motion features, resulting in F_S and F_M. These features are weighted by their respective saliency scores (S_S and S_M) via adaptive fusion, and the final representation S is obtained through the Fusion module. Encoders marked are frozen during training.

- **Employing Parameter-Free Mechanisms:** Cross-modal text guidance is integrated via a parameter-free attention mechanism (Sec 3.3), further enhancing semantics without increasing trainable parameter count.

Built upon this PEFT foundation, TCAR-Net employs a dual-stream architecture (Figure 1) where the spatial stream processes appearance information (I-frames/Residuals) and the motion stream handles temporal information (Motion Vectors), both benefiting from shared frozen representations and unified text guidance for effective compressed domain processing.

3.2. Dual-Stream Feature Extraction and Temporal Modeling

To effectively process the heterogeneous data within compressed streams, we utilize two parallel pathways built upon the shared, frozen visual backbone and processed by the trainable sequence model.

3.2.1. Unified Backbone Feature Extraction

The frozen CLIP visual encoder $\mathcal{V}$ is applied to extract features from the key compressed components:

- I-frames (I) and Residual (R): $E_I = \mathcal{V}(I)$, $E_R = \mathcal{V}(R)$ capture spatial appearance and fine-grained texture changes. $E_I, E_R \in \mathbb{R}^{B\times T\times D}$.

- Motion Vectors (MV): $E_M = \mathcal{V}(MV)$) captures motion patterns. $E_M \in \mathbb{R}^{B\times T_\times D}$.

Using a single $\mathcal{V}$ for all inputs ensures parameter efficiency.

3.2.2. Stream-Specific Processing and Adaptive Spatial Fusion

Before temporal modeling, the streams undergo specific processing:

- **Spatial Stream:** Recognizing the complementarity of I-frames (global context) and Residuals (motion edges/texture), we introduce an adaptive fusion mechanism. Trainable parameters $\boldsymbol{\beta} = [\beta_1, \beta_2]$ generate dynamic weights via softmax:

$$w_i = \frac{\exp(\beta_i)}{\sum_{j=1}^{2} \exp(\beta_j)}, \quad i \in \{1, 2\} \tag{1}$$

  These weights combine the features, allowing the model to emphasize I-frames in low-motion and Residuals in high-motion scenarios:

$$E_S = w_1 E_I + w_2 E_R \tag{2}$$

  This yields the input spatial sequence $E_S \in \mathbb{R}^{B\times T\times D}$.

- **Motion Stream:** The raw extracted motion features $E_M \in \mathbb{R}^{B\times T\times D}$ serve as the input sequence for temporal modeling.

3.2.3. Trainable Sequence Modeling

To efficiently capture temporal dependencies, E_S and E_M are processed by separate trainable sequence models. These are multi-layer (e.g., 6-layers) Transformer encoders with position embeddings, denoted as $\mathcal{T}_{seq,S}$ for the spatial stream and $\mathcal{T}_{seq,M}$ for the motion stream:

$$E'_S = \mathcal{T}_{seq,S}(E_S + P_S) \tag{3}$$

$$E'_M = \mathcal{T}_{seq,M}(E_M + P_M) \tag{4}$$

where P_S, P_M are position embeddings. These encoders model long-range temporal dependencies, yielding temporally enriched features $E'_S, E'_M \in \mathbb{R}^{B\times T\times D}$.

3.3. Parameter-Free Text-Guided Cross-Modal Refinement

A key aspect of TCAR-Net is the integration of high-level semantic guidance from text to actions, especially challenging in the potentially noisy compressed domain. Crucially, this is achieved without introducing additional trainable parameters.

This mechanism operates on the temporally modeled visual features (E'_S, E'_M) from temporal encoders and the frozen textual embeddings ($E_T \in \mathbb{R}^{N_c\times D}$) derived from category descriptions using the frozen CLIP text encoder, where N_c is the number of classes.

3.3.1. Saliency Generation

For each visual stream $X \in \{E'_S, E'_M\}$, we first generate saliency weights through text-visual similarity:

1. **Cross-Modal Similarity Computation:** Calculate similarity between visual features and text embeddings:
$$\Phi_X = \text{sim}(X, E_T) \in \mathbb{R}^{B\times T\times N_c} \tag{5}$$
where $X \in \mathbb{R}^{B\times T\times D}$ represents temporal visual features, and $\text{sim}(\cdot,\cdot)$ denotes cosine similarity.
2. **Saliency Weight Generation:** Normalize similarities to obtain saliency weights:
$$W_X = \text{Softmax}(\Phi_X/\tau) \in \mathbb{R}^{B\times T\times N_c} \tag{6}$$
where τ is a temperature parameter. W_X indicates the relevance of each temporal step for different action categories.

3.3.2. Feature Enhancement via Saliency

The original visual features are enhanced using the generated saliency weights:

$$F_X = \sum_{t=1}^{T} W_X \odot X_t \in \mathbb{R}^{B\times N_c\times D} \tag{7}$$

where $\odot$ denotes element-wise multiplication. This yields enhanced features F_S and F_M for spatial and motion streams respectively.

3.4. Final Prediction via Category Matching and Fusion

3.4.1. Category Similarity Scoring

The enhanced features are compared with text embeddings to generate final scores:

$$S_S = \text{sim}(F_S, E_T), \quad S_M = \text{sim}(F_M, E_T) \in \mathbb{R}^{B\times N_c} \tag{8}$$

where S_S and S_M represent the compatibility scores between video and class for spatial and motion streams respectively.

3.4.2. Weighted Fusion

The final prediction score is obtained by fusing scores from both streams:

$$S = \lambda S_S + (1-\lambda) S_M \in \mathbb{R}^{B\times N_c} \tag{9}$$

The class with the highest score in S for a given video is the final prediction. The hyperparameter λ allows balancing the contribution derived from combined spatial structure and fine-grained changes against the contribution derived from explicit motion fields.

4. Experiments and Analysis

4.1. Datasets and Evaluation Metric

We evaluated our model under supervised learning on two primary datasets: **UCF101** and **HMDB51** . HMDB51 contains 6,766 videos spanning 51 action categories, while UCF101 comprises 13,320 videos from 101 categories. For this typical video understanding task, we selected action recognition and adopted the top-1 accuracy as the evaluation metric across all experiments. Following [5], we used MPEG-4 encoded videos with an average of 11 P-frames per

I-frame, uniformly resizing video resolution to 340×256. Sixteen clips were uniformly sampled from each video, with their spatiotemporal positions determined by the GOP structure: random GOP indices and intra-clip position sampling were employed during training for data augmentation, while deterministic center sampling was applied during testing. Spatial processing included random corner cropping and horizontal flipping during training, and center cropping during testing, with consistent spatial dimension handling across all inputs.

4.2. Implementation Details

For the backbone network, we adopt ViT as the encoder for I-frame/motion vector/residual streams, pretrained on Kinetics-400 using raw video inputs. The text encoder from CLIP ViT-B/32 is employed to encode textual information. To preserve prior knowledge from large-scale datasets, all encoders remain frozen during training, with only the task-specific heads being fine-tuned. The video recognition model is optimized using AdamW with a base learning rate of 5e-5 (scheduled via cosine annealing), $\beta_1 = 0.9$, $\beta_2 = 0.999$ momentum parameters, and 0.2 weight decay. We implement a batch size of 256 across 20 training epochs, incorporating a linear warmup strategy for the first 5 epochs to ensure training stability. The temperature hyperparameter τ is set to 0.01. All experiments are conducted on two NVIDIA 4090 GPUs.

4.3. Comparison with the State-of-the-art Approaches

Table 1: **Comparison of different video action recognition methods.** We report the top-1 accuracy (%) on HMDB51 and UCF101 datasets.

Method	Modality	Pretrain	Frames	GFLOPs	Tunable Params.[M]	HMDB51	UCF101
Two-stream [2]	RGB+Flow	ImageNet1K	8	1600	31.2	59.4	88.0
TDN ResNet101 [19]	RGB	Kinetics	8+16	3240	-	76.3	97.4
ARTNet ResNet18 [20]	RGB	Kinetics	16	5875	-	70.9	94.3
VideoMAE ViT-B [21]	RGB	Kinetics	16	1080	87.0	73.3	96.1
MVFNet ResNet50 [22]	RGB	ImageNet1K	16	1974	-	75.7	96.6
BIKE ViT/L[10]	RGB+text	Kinetics/CLIP	16	3728	230	81.7	97.7
Text4Vis ViT/L[23]	RGB+text	Kinetics/CLIP	16	3829	230.7	81.3	**98.2**
Refined-MV ResNet152 [24]	I-Frame+MV+Res	ImageNet1K	-		142.5	59.7	89.9
IPTSN ResNet152 [25]	I-Frame+MV+Res	ImageNet1K	16+16+16	**215**	130.8	69.1	93.4
SIFT-Net I3D [13]	I-Frame+MV+Res	Kinetics	3+3+3	1971	-	72.3	94
MM-ViT ViT/B [14]	I-Frame+MV+Res	ImageNet1K	8+8+8	820	158.1	-	93.3
CoViAR ResNet152 [5]	I-Frame+MV+Res	ILSVRC 2012-CLS	3+3+3	1222.0	142.5	59.1	90.4
CoViAR ResNet152 [5]	I-Frame+MV+Res+flow	ILSVRC 2012-CLS	-	3970	-	70.2	94.9
DMC-Net I3D [6]	I-Frame+MV+Res+flow	-	3+3+3	401	-	71.8	92.3
CVPT ViT/B [26]	I-Frame+MV+Res	Kinetics	-	772.2	**0.5**	69.7	95.5
Ours ViT/B	I-Frame+MV+Res+text	Kinetics/CLIP	16+16+16	544.85	37.90	74.3	96.0
Ours ViT/L	I-Frame+MV+Res+text	Kinetics/CLIP	16+16+16	3990.65	85.12	**82.4**	97.7

Compressed Video-based Methods. We first compare our method with compressed-domain action recognition models, including CoViAR [5], MM-ViT [14], DMC-Net [6], SIFT-Net [13], and CVPT [26]. Table 1 presents the top-1 accuracy results on HMDB51 and UCF101.

Notably, our parameter-efficient tuning framework achieves comparable or even superior performance to full fine-tuning paradigms while reducing computational demands. When employing the same ViT-B backbone, TCAR achieves a +2.7% accuracy gain over MM-ViT on UCF101 while requiring 50% fewer GFLOPs and 3.17 times fewer tunable parameters. Compared to SIFT-Net, our model demonstrates accuracy improvements of +2.0% and +2.0% on HMDB51 and UCF101 respectively, with computational resources reduced by a factor of 2.6, thereby demonstrating the efficiency of our approach.

Besides the full fine-tuning methods, we also compare with representative efficient fine-tuning models, including CVPT[26]. CVPT leverages Motion Vectors and Residuals to generate conditional prompts.However, since CVPT solely relies on compressed-domain information and lacks textual collaboration to enable efficient cross-modal inference, its achieved accuracy remains limited. In contrast, our method leverages textual prior knowledge to align spatial and temporal branches, outperforming CVPT by significant margins of 4.6% and 0.5% in accuracy on HMDB51 and UCF101 datasets, respectively.These results highlight the effectiveness of our text-augmented dual-branch architecture in enhancing action recognition from compressed video streams.

Raw Video-based Methods. Raw video-based methods generally achieve strong performance due to their reliance on dense RGB frame sampling and large-scale pretraining. However, our method significantly narrows the performance gap between compressed-domain and pixel-domain approaches.

For instance, despite using only 36.8% of the tunable parameters of Text4Vis [23], our method outperforms it by +1.1% on HMDB51. Compared to BIKE [10], our ViT-L model achieves a +0.7% absolute gain on HMDB51, demonstrating the importance of integrating text for enhanced feature representation.

Furthermore, our method maintains a substantially lower computational footprint compared to raw video models such as TDN [19], ARTNet [20] and VideoMAE[21], while delivering competitive or superior performance. These results validate our approach's ability to efficiently utilize compressed video features while leveraging textual semantics for improved temporal modeling and action recognition.

4.4. Ablation Study

To evaluate the effectiveness of components in TCAR-Net, we conduct extensive ablation studies. Without loss of generality, we use a pretrained ViT-B as the backbone and perform experiments on the HMDB51 dataset.

Analysis of Frozen Backbone Effectiveness in Compressed Domain. To address potential concerns about feature extraction bias when applying CLIP's frozen backbone to compressed domain data, we first analyze the effectiveness across different compressed modalities. As shown in Table 2, we evaluate I-frames, residuals, and motion vectors (MV) as individual modalities to verify their information content and compatibility with frozen CLIP encoders.

Table 2: Ablation study on single-modality inputs.

I-frame	**Res**	**MV**	**Top-1** (%)
✓			**71.5**
	✓		53.7
		✓	25.0

Modality-Specific Performance Analysis: The performance hierarchy aligns with the semantic richness and similarity to CLIP's training distribution. I-frames, being closest to natural images, achieve the highest individual performance (71.5%), validating the frozen backbone's effectiveness. Residuals, containing texture and boundary information of moving objects, achieve moderate performance (53.7%) and demonstrate effectiveness for motion localization. Motion vectors, while encoding temporal information, show lower individual performance (25.0%) due

to their block-level discretization and encoding noise, making direct utilization challenging without semantic constraints.

Table 3: Comparison between PEFT and Full Fine-tuning Strategies

Method	Trainable Parameters	Memory	Accuracy(%)
Video Branch Full Fine-tune	296.49M (82.38%)	13.95GB	**74.7**
Video+Text Full Fine-tune	359.91M (100.00%)	14.75GB	74.1
PEFT (Frozen Backbones)	**37.91M (10.53%)**	**7.02GB**	74.3

Complementary Information Integration: The substantial performance gain from multimodal fusion (74.3% vs. individual modalities) demonstrates that our architecture effectively compensates for single-modality limitations through cross-modal collaboration. This validates our hypothesis that frozen CLIP encoders maintain sufficient representational capacity when combined with appropriate fusion mechanisms.

PEFT Strategy Validation: We conduct comprehensive ablation experiments to evaluate the effectiveness of Parameter-Efficient Fine-Tuning (PEFT) compared to full fine-tuning strategies under a fixed architecture. As shown in Table 3, our PEFT approach demonstrates significant advantages in training efficiency while maintaining competitive performance. The experimental results reveal that our PEFT strategy (frozen backbones + fine-tuned components) requires only 37.91M trainable parameters (10.53% of total), achieving a remarkable 7.8× reduction compared to video branch full fine-tuning and 9.5× reduction compared to full model fine-tuning. In terms of GPU memory consumption, PEFT reduces memory usage by 2.1× (from 14.75GB to 7.02GB), making our approach more accessible for resource-constrained environments. This demonstrates that our PEFT strategy effectively balances training efficiency and model performance by selectively fine-tuning only the critical components while keeping the pre-trained backbones frozen.

Table 4: Pretrained Model Comparison

Pretrain	Top-1 (%)
CLIP	68.8
K400	**74.3**

Ablation Study on Pretraining Strategies. As shown in Table 4, we systematically compare the impact of different pretraining approaches on our model's performance. When initialized with CLIP, the model achieves 68.8% Top-1 accuracy. Notably, pretraining on Kinetics-400 (K400) action recognition dataset yields a significant improvement of +5.5%, reaching 74.3% Top-1 accuracy. This demonstrates that video-based pretraining provides stronger temporal modeling capabilities for our action localization task compared to image-text multimodal pretraining. The results validate our hypothesis that domain-specific pretraining strategies are critical for video understanding tasks.

Component Effectiveness and Fusion Strategies. Table 5 systematically evaluates the contribution of key components in our framework. The full configuration with I-frame, motion vectors (MV), spatial-branch residual (S), and feature fusion achieves the best performance (74.3%). Replacing spatial residuals with temporal ones (T) causes a 0.9% accuracy drop (73.4%), indicating spatial modeling is more critical for our task. Removing residual blocks entirely degrades performance to 72.8%, validating their necessity in learning hierarchical representations. The

Table 5: Ablation study on components and fusion strategies

I-frame	Residual	MV	Fusion	Top-1 (%)
✓	(S)	✓	✓	**74.3**
✓	(T)	✓	✓	73.4
✓	×	✓	✓	72.8
✓	(S)	×	×	73.2

- ✓: Component is used; ×: Component is removed.
- (S): Residual in **Spatial** branch; (T): Residual in **Temporal** branch.

configuration without MV and fusion shows a 1.1% performance gap (73.2%), confirming that explicit motion modeling and multi-modal fusion synergistically enhance action understanding.

Table 6: λ Ablation Study

λ	0.2	0.3	**0.4**	0.5	0.6	0.7	0.8
Top-1 (%)	73.2	73.7	**74.3**	73.1	74.2	74.2	73.5

Fusion Weight Analysis. As formulated in Eq. 9, the hyperparameter λ controls the trade-off between spatial and temporal branches. Table 6 reveals a clear performance peak at $\lambda = 0.4$ with 74.3% accuracy. Extreme allocations ($\lambda = 0.2$ or 0.8) degrade accuracy by 1.1-1.2%, suggesting over-reliance on either modality harms representation capability. The symmetrical pattern around $\lambda = 0.4$ confirms that temporal information dominates while requiring complementary spatial cues for optimal effectiveness.

Table 7: Comparison of inference time (ms) per video.

Method	Pre-processing	Model Inference	Full Pipeline
ActionCLIP[17]	41.51	**11.52**	53.03
TCAR-Net (Ours)	**6.11**	26.32	**32.43**

Inference Efficiency Analysis. Table 7 presents comprehensive timing analysis comparing our TCAR-Net with ActionCLIP. Our method achieves 6.8× speedup in pre-processing (6.11ms vs 41.51ms per video) by eliminating full video decoding overhead. While multi-modal encoding introduces modest computational overhead, the substantial preprocessing gains result in 1.6× end-to-end speedup (32.43ms vs 53.03ms per video), validating our core hypothesis that compressed-domain processing substantially improves overall efficiency.

5. Conclusion and Future Work

In this paper, we propose TCAR-Net, which is, to the best of our knowledge, the first text-driven multi-modal action recognition framework in the compressed domain. Our method addresses the gap between semantic reasoning and efficiency in video understanding. Our method introduces a dual-stream collaborative paradigm, where a spatial enhancement branch leverages I-frame and residual fusion with text guidance to improve spatial representation, while a motion

distillation branch refines motion vectors using vision transformers and textual priors to enhance temporal modeling. By aligning these two branches through a text-informed fusion mechanism, TCAR-Net achieves superior semantic-spatiotemporal alignment, significantly improving action recognition in compressed videos.

Extensive experiments on UCF101 and HMDB51 demonstrate that our approach outperforms existing compressed-domain models while achieving competitive performance compared to pixel-domain methods. These results highlight the effectiveness of leveraging text-driven multi-modal interactions to enhance action recognition without incurring the overhead of optical flow computation.

In future work, we aim to explore more fine-grained textual prompts to further enhance cross-modal alignment and investigate the generalization of TCAR-Net to real-world video scenarios, including low-latency edge deployment. We believe that our work provides a scalable and efficient pathway toward semantically enriched action recognition in compressed video streams.

Author Contributions

Mengkun Guo completed the entire main part of the thesis, including the initial conception, design and experimentation of the method, and optimized the model. Ming Ma provided guidance on the research direction and methodology, and led the review and revision of the paper. Xinqi Li and Die Tao assisted in figure preparation, paper review, and experimental investigation. All authors provided key feedback and helped shape the research direction.

Acknowledgements

This work was supported in part by the Natural Science Foundation of China (No.62462049); Research Project on Strengthening the Construction of Important Ecological Security Barrier in Northern China by Higher Education Institutions in Inner Mongolia Autonomous Region under Grant No.STAQZX202321.

References

[1] Oleksandra Poquet, Lisa Lim, Negin Mirriahi, and Shane Dawson. Video and learning: a systematic review (2007–2017). In *Proceedings of the 8th international conference on learning analytics and knowledge*, pages 151–160, 2018.

[2] Karen Simonyan and Andrew Zisserman. Two-stream convolutional networks for action recognition in videos. *Advances in neural information processing systems*, 27, 2014.

[3] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017.

[4] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018.

[5] Chao-Yuan Wu, Manzil Zaheer, Hexiang Hu, R Manmatha, Alexander J Smola, and Philipp Krähenbühl. Compressed video action recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6026–6035, 2018.

[6] Zheng Shou, Xudong Lin, Yannis Kalantidis, Laura Sevilla-Lara, Marcus Rohrbach, Shih-Fu Chang, and Zhicheng Yan. Dmc-net: Generating discriminative motion cues for fast compressed video action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1268–1277, 2019.

[7] Shiyuan Huang, Xudong Lin, Svebor Karaman, and Shih-Fu Chang. Flow-distilled ip two-stream networks for compressed video action recognition. *arXiv preprint arXiv:1912.04462*, 2019.

[8] Barak Battash, Haim Barad, Hanlin Tang, and Amit Bleiweiss. Mimic the raw domain: Accelerating action recognition in the compressed domain. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 684–685, 2020.

[9] Haoyuan Cao, Shining Yu, and Jiashi Feng. Compressed video action recognition with refined motion vector. *arXiv preprint arXiv:1910.02533*, 2019.

[10] Wenhao Wu, Xiaohan Wang, Haipeng Luo, Jingdong Wang, Yi Yang, and Wanli Ouyang. Bidirectional cross-modal knowledge exploration for video recognition with pre-trained vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6620–6630, 2023.

[11] Yuqi Huo, Xiaoli Xu, Yao Lu, Yulei Niu, Zhiwu Lu, and Ji-Rong Wen. Mobile video action recognition. *arXiv preprint arXiv:1908.10155*, 2019.

[12] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[13] Jiapeng Li, Ping Wei, Yongchi Zhang, and Nanning Zheng. A slow-i-fast-p architecture for compressed video action recognition. In *Proceedings of the 28th ACM international conference on multimedia*, pages 2039–2047, 2020.

[14] Jiawei Chen and Chiu Man Ho. Mm-vit: Multi-modal video transformer for compressed video action recognition. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1910–1921, 2022.

[15] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[16] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*, 2021.

[17] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[18] Chen Ju, Tengda Han, Kunhao Zheng, Ya Zhang, and Weidi Xie. Prompting visual-language models for efficient video understanding. In *European Conference on Computer Vision*, pages 105–124. Springer, 2022.

[19] Limin Wang, Zhan Tong, Bin Ji, and Gangshan Wu. Tdn: Temporal difference networks for efficient action recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1895–1904, 2021.

[20] Limin Wang, Wei Li, Wen Li, and Luc Van Gool. Appearance-and-relation networks for video classification. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1430–1439, 2018.

[21] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[22] Wenhao Wu, Dongliang He, Tianwei Lin, Fu Li, Chuang Gan, and Errui Ding. Mvfnet: Multi-view fusion network for efficient video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 35, pages 2943–2951, 2021.

[23] Wenhao Wu, Zhun Sun, and Wanli Ouyang. Revisiting classifier: Transferring vision-language models for video recognition. In *Proceedings of the AAAI conference on artificial intelligence*, volume 37, pages 2847–2855, 2023.

[24] Haoyuan Cao, Shining Yu, and Jiashi Feng. Compressed video action recognition with refined motion vector. *arXiv preprint arXiv:1910.02533*, 2019.

[25] Shiyuan Huang, Xudong Lin, Svebor Karaman, and Shih-Fu Chang. Flow-distilled ip two-stream networks for compressed video action recognition. *arXiv preprint arXiv:1912.04462*, 2019.

[26] Bing Li, Jiaxin Chen, Xiuguo Bao, and Di Huang. Compressed video prompt tuning. *Advances in Neural Information Processing Systems*, 36:31895–31907, 2023.

Author Biography

Mengkun Guo is a postgraduate student at Inner Mongolia University. His research interests include compressed domain video understanding and multimodal information processing.

Die Tao is a postgraduate student at Inner Mongolia University. The main research directions are video understanding in the compressed domain and multimodal fusion.

Xinqi Li is a postgraduate student at Inner Mongolia University. The main research directions are video understanding in the compressed domain and software formalization.

Ming Ma is an associate professor at Inner Mongolia University, mainly engaged in research in the fields of computer vision. He has led and participated in over 10 national and provincial-level scientific research projects, published more than 20 high-quality academic papers, and collaborated with enterprises to develop multiple products, which have been applied in related industries.