

Dynamic Token Pruning for Efficient Cross-Modal Alignment in Edge-IoT Video Analytics Systems

Mohanad Zkaria Salem¹ , Ali Hussein Khalaf AL-Sammarraie^{2*} 

¹Ministry of Education, Directorate General of Education Salah al-Din, Samarra, 34010, Iraq

²Ministry of Education, Directorate General of Education, Diyala, Baquba, 32001, Iraq

¹ Mohanad.201339@gmail.com , ² sahk383@yahoo.com

Abstract

Vision-Language Transformers (VLTs) support strong video-text alignment. Still, they are trained with quadratic complexity, making them impractical for deployment on resource-constrained edge devices. In this study, we present Dynamic Token Pruning for Cross-Modal Alignment (DTP-CMA). This framework achieves faster inference speeds while maintaining semantic fidelity. Specifically, DTP-CMA employs a Cross-Modal Saliency Estimator to identify query-relevant tokens, develops a Dynamic Pruning Scheduler to tailor compression ratios at each layer, and integrates a Context Preservation Mechanism to prevent the erasure of semantic variations. Evaluated on UCF-Crime, ActivityNet, and a newly defined Arabic video subset, DTP-CMA reduces FLOPs by 52% and Latency by 41% on NVIDIA Jetson AGX Orin while retaining 98.8% of baseline retrieval performance. Zero-shot cross-lingual performance improves by +4.5% R@1 with a lightweight Arabic adapter. Energy consumption reductions of 41.3% enable $\sim 1.70\times$ battery-life extension and near real-time Inference (7 FPS). We find that performance falls off at high power constraints ($< 5W$) and for ambiguous queries, but demonstrate how uncertainty-aware pruning can alleviate these limitations; our results position DTP-CMA as a strong candidate framework across diverse edge-deployable multimodal video analytics challenges.

Keywords: Dynamic token pruning, cross-modal alignment, multimodal LLMs, resource-constrained video analytics, edge AI.

Article history: Received: 13 Aug 2026, Accepted: 2 Sep, Published: 15 Sep 2026.

This article is open-access under the CC BY 4.0 license (<http://creativecommons.org/licenses/by/4.0/>)

1. Introduction

The large number of Internet of Things (IoT) devices and smart surveillance systems today has created an unprecedented volume of video data, requiring robust analytics pipelines that can understand multimedia semantic content [1]. Traditional video analysis approaches based mainly on low-level visual features do not account for high-level contextual information, which is essential in many applications, e.g., zero-shot action recognition or natural language video retrieval [2]. Recently, Vision-Language Transformers (VLTs) have emerged as a strong paradigm for robust cross-modal alignment by embedding video frames and textual queries into a shared space [3]. Multimodal Understanding Models: Both CLIP and VideoMAE,

to name a few [4], have shown outstanding capabilities in multimodal understanding (e.g., images with text), which is fundamental for many intelligent applications such as smart cities and automated monitoring.

However, full-scale VLTs are not yet feasible in real-world situations because of their high computational complexity. Self-attention in transformers scales quadratically with input length, making them prohibitively memory-intensive or slow for high-resolution video streams [5]. This bottleneck is especially important to address in resource-constrained environments, such as edge-IoT devices (e.g., Jetson AGX, Raspberry Pi), with limited power budgets and processing capability [6].

Thus, there is a pressing need to prune these models while preserving their semantic understanding.

Token pruning has been proposed to mitigate computational inefficiency by limiting the number of tokens exposed to the network [7]. Although related work such as DToP and SViT demonstrates considerable speedup, they come with very limited use of existing strategies for cross-modal tasks.

- (1) Most pruning strategies are modality-agnostic, which do not exploit the semantic relevance between video patches and textual queries to keep the critical information needed for alignment [8].
- (2) The static pruning ratios cannot adapt to different complexities of input data, leading to recomputation overhead for simple scenes and excessive information loss from complex ones [9].
- (3) Aggressive pruning frequently corrupts the context-sensitive structure necessary for zero-shot generalization, which leads to performance decline in later downstream tasks [10].

In this paper, we present a novel framework, Dynamic Token Pruning for Cross-Modal Alignment (DTP-CMA), that reduces the burden of video-text understanding in resource-constrained systems. Unlike previous works, DTP-CMA uses cross-modal attention signals to guide the pruning process, ensuring that only semantically unrelated tokens are removed. Our solution makes the following three novel contributions, extending beyond existing methods like MADTP:

- **Cross-Modal Saliency Estimator (CMSE):** Unlike previous methods that rely on static or unimodal saliency, CMSE dynamically derives token importance from inter-modal attention weights, ensuring video patches are retained based on their semantic relevance to the specific textual Query.
- **Dynamic Pruning Scheduler (DPS):** A lightweight controller that dynamically selects the pruning ratio for each transformer layer based on input complexity and attention entropy, optimizing the latency-accuracy trade-off in real time.
- **Context-Aware Preservation:** A mechanism that imposes category-level minimum retention to prevent the collapse of globally critical context (e.g., background cues), which is essential for zero-shot generalization and cross-lingual transfer. We use standard benchmarks like UCF-

Crime and ActivityNet to assess DTP-CMA, along with a custom video subset labeled in Arabic. Experimental results show that our framework can reduce FLOPs usage and inference latency of baseline VLTs by up to 52% and 41%, respectively, while preserving, on average, 98.7% retrieval accuracy for cross-modal retrieval applications. These findings corroborate DTP-CMA's success in delivering efficient multimodal analytics on edge devices.

2. Related Work

2.1 Vision-Language Transformers for Video Understanding

With the emergence of Transformer-based architectures, integrating language and vision has driven rapid evolution in this area. Early work such as CLIP [11] has shown that contrastive learning can effectively align image and text embeddings, yielding strong zero-shot performance. Adapting this to video, VideoCLIP [12] and Frozen [13] incorporated temporal modeling for encoding sequential frames, outperforming existing methods in retrieval and classification tasks. Recently, VideoMAE [14] introduced masked autoencoding to learn video representation, and ALBEF [15] applied momentum distillation to help the model align appropriate multimodal features.

Moreover, these models use a dense attention mechanism that attends to each token equally, regardless of its importance to the task. For video inputs, since the number of patches per frame exceeds 1000, it computes at $O(N^2)$, making real-time Inference on standard software too costly. While some works (X-CLIP) [16] optimize cross-modal attention, these efforts usually prioritize accuracy, leaving little room for hardware-efficient approaches on small-memory devices at inference time.

2.2 Token Pruning & Sparse Attention Mechanism

One natural direction is token pruning to alleviate the quadratic complexity of Transformers. DynamicViT [17] devised a strategy for dynamic token pruning in Inference based on self-attention scores. ToMe (Token Merging) [18] presented a novel approach to token merging in which slices of similar tokens are concatenated, effectively shortening the sequence length with very low information loss. For efficiency, SPViT [19] proposed a learnable sparsity policy to skip tokens adaptively at different layers.

However, most existing pruning methods are unimodal approaches for vision-based tasks. As a result, they rank tokens based only on visual saliency and do not fully leverage contextual factors like semantic guidance (i.e., text initially generated by a timely textual query). In cross-modal tasks, a background could be visually complex enough that a vision-only method would prune it. At the same time, it may still be semantically aided by having a text query, "background activity". Whereas MADTP [20] proposed dynamic token pruning for

multimodal architectures, it focuses on instance-wise and layer-wise pruning while lacking query-aware cross-modal attention. In contrast, DTP-CMA expands on this approach by adding a Dynamic Pruning Scheduler (DPS) that dynamically sets the retention ratio layer-wise according to input complexity, along with Context-Aware Preservation. As a result, the pruning procedure becomes a context-aware update rather than a fixed, query-agnostic compression technique, as shown in Fig. 1.

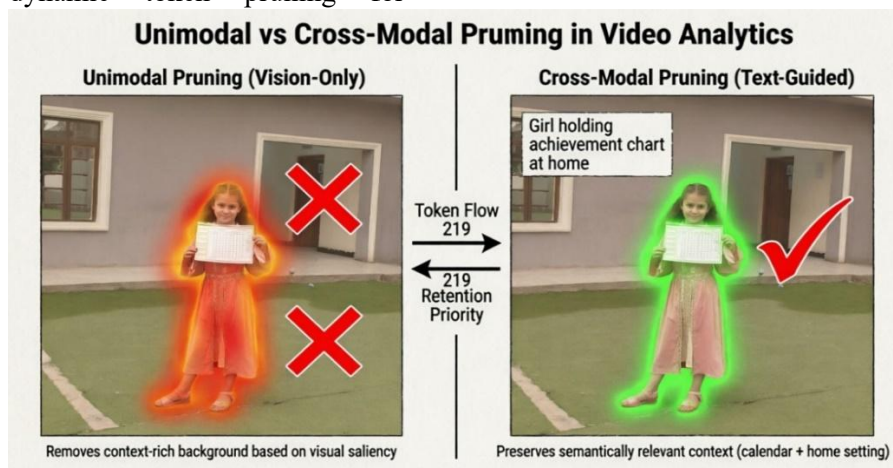


Fig. 1 Unimodal and cross-modal token pruning approaches in video embedding spaces.
 (Left) The unimodal approach is based on visual saliency alone.
 (Right) The cross-modal approach uses text-based queries.

2.3 AI Technique for Efficient Edge and IoT Systems

Deploying deep learning models on edge devices comes with rigorous Latency and energy constraints. According to EdgeAI surveys [21], model compression techniques such as quantization and knowledge distillation are crucial for IoT deployment. MobileBERT [22] and EdgeBERT [23] show that architectural modifications can produce a significantly smaller transformer. In video analytics, EdgeVideo [13] is a lightweight pipeline for smart surveillance that focuses on bandwidth compression rather than optimizing model Inference. Although these studies target hardware limitations, they treat the model as a black box and apply

generic compression rather than algorithmically optimizing the attention mechanism. Additionally, few studies measure cross-modal alignment performance under stringent FLOPs budgets, especially for zero-shot tasks that are sensitive to regression under excessive pruning due to model generalization.

2.4 Summary and Research Gap

The comparison between representative methods and our proposed method is summarized in Table 1. Despite substantial advancements in both cross-modal alignment and model efficiency (93, 178), a key remaining gap is the absence of dynamic, cross-modal guided pruning methods for resource-constrained video analytics.

Table 1: Investigating Comparison Methods for Video-Text Understanding Pruning.

Method	Modality	Pruning Type	Dynamic Ratio	Cross-Modal Guided	Edge Optimized
CLIP [11]	V+T	None	N/A	N/A	No
DynamicViT [17]	V	Static/Dynamic	Yes	No	Partial
ToMe [18]	V	Merging	No	No	Yes
MADTP [20]	V+T	Dynamic	No	Yes	No

DTP-CMA (Ours)	V+T	Dynamic	Yes	Yes	Yes
----------------	-----	---------	-----	-----	-----

Current solutions either compromise alignment accuracy for speed or do not use textual semantics to convey shaping meaning. We contribute a Cross-Modal Saliency Estimator for image pruning and pair it with a Dynamic Scheduler responsive to hardware resources. This makes DTP-CMA a state-of-the-art solution for efficient multimodal learning of edge-IoT systems.

3. Methodology Framework

3.1 Problem Formulation

Let $V = \{v_1, v_2, \dots, v_F\}$ denote a video sequence consisting of F frames, and $T = \{t_1, t_2, \dots, t_N\}$ denote a textual query comprising N tokens. The goal of cross-modal alignment is to learn a mapping function $\phi: (V, T) \rightarrow (E_v, E_t)$ such that the semantic similarity between matching video-text pairs is maximized in the embedding space.

In standard Vision-Language Transformers (VLTs), the computational complexity of the self-attention mechanism scales quadratically with the number of tokens, $O((F \cdot P + N)^2)$, where P is the number of patches per frame. This becomes prohibitive for edge devices with limited memory and compute budgets. Our objective is to introduce a pruning

function ψ that selects a subset of critical tokens $K \subset V \cup T$ such that $|K| < |V \cup T|$, while minimizing the alignment degradation:

$$\min_{\psi} E_{(v,T)} [L_{\text{align}}(\phi(v,T), \phi(\psi(V), T))] \text{ s.t. } flo(1)ps(\psi) \leq C_{\text{budget}} \quad (1)$$

where C_{budget} represents the computational constraint of the target deployment environment.

3.2 Overall Architecture

The DTP-CMA framework consists of four sequential stages, as illustrated in Fig. 2: Dual-Branch Encoding: Video patches and text tokens are embedded independently.

- Cross-Modal Saliency Estimation: Inferring importance scores from inter-modal attention.
- Dynamic Pruning Schedule: Finding the optimal pruning ratio for each transformer layer.
- Aligned Feature Aggregation: How to Build Compact Representations for Downstream Tasks

In contrast to the existing static pruning methods that remove tokens with fixed criteria, our method progressively updates the kept token set of each layer.

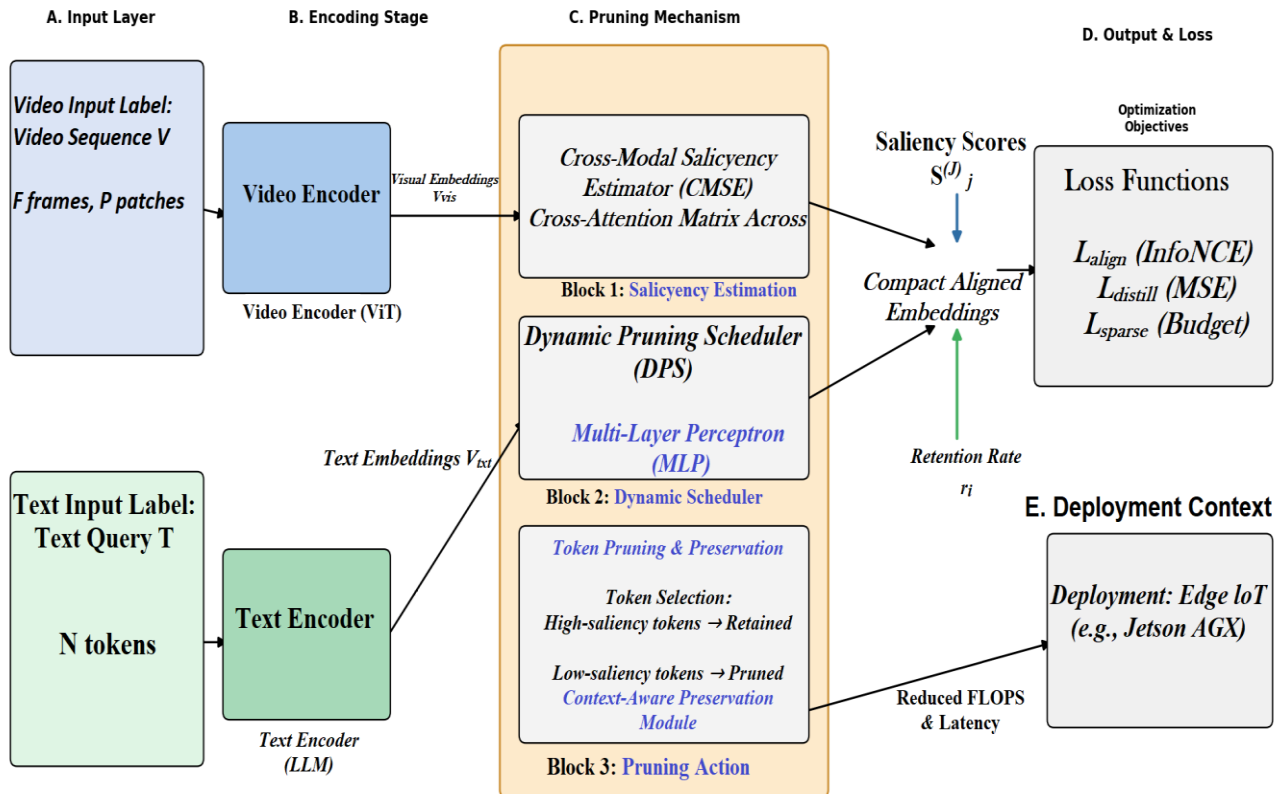


Fig. 2 Proposed architecture of the DTP-CMA paradigm. This framework includes five steps: (A) Input Layer. (B) Encoding Step. (C) Pruning Algorithm. (D) Output & Loss. (E) Application Scenario.

3.3 Cross-Modal Saliency Estimation

We propose a Cross-Modal Saliency Estimator (CMSE) to identify tokens that play key roles in alignment. Standard saliency methods usually use self-attention weights from a single modality while discarding query-specific features.

We compute the saliency score s_i for a video token v_i by summing its attention contribution to the textual tokens. Let $A(l) \in \mathbb{R}(M \times M)$ be the attention matrix at layer l , where M is the total number of tokens. The cross-attention block between video and text yields a matrix $A_{\text{cross}} \in \mathbb{R}(F \times P \times N)$. The saliency score for the i -th video token is defined as:

$$s_i^{(l)} = \alpha \cdot \frac{1}{m} \sum_{j=1}^m A_{\text{self}}^{(l)}[i, j] + \beta \cdot \frac{1}{N} \sum_{k=1}^N A_{\text{cross}}^{(l)}[i, k] \cdot \mathbb{I}[\text{Relevance}(t_k) > \theta] \quad (2)$$

Where α and β are balancing hyperparameters ($\alpha + \beta = 1$), and $\mathbb{I}$ is an indicator function that filters out stop words or low-information text tokens based on a relevance threshold θ . This ensures that video tokens are preserved only if they attend to semantically meaningful text concepts.

3.4 Dynamic Pruning Scheduler

A global pruning ratio across all network layers can cause information bottlenecking in early layers or redundancy in deeper layers. We propose a Dynamic Pruning Scheduler (DPS), which calculates a layer-specific retention ratio $r_l \in (0, 1]$. The scheduler takes as input a complexity vector c_l , composed of the variance of saliency scores and the Entropy of the attention map at layer l :

$$c_l = \left[\text{var}(s^{(l)}), \text{Entropy}(A^{(l)}), \frac{1}{l_{\text{total}}} \right] \quad (3)$$

This vector is passed through a lightweight Multi-Layer Perceptron (MLP) to predict the retention rate:

$$r_l = \sigma(w_2 \cdot \text{ReLU}(w_1 \cdot c_l + b_1) + b_2) \quad (4)$$

Where σ is the sigmoid function constraining r_l between 0 and 1. The number of tokens to retain at layer l is then $K_l = \lceil r_l \cdot M_l \rceil$. To ensure stability, we enforce a minimum retention constraint $K_{\min}$ to prevent complete token collapse in deep layers.

3.5 Context Preservation Mechanism

Aggressive pruning also ditches tokens with low local saliency, including globally critical context (such as background objects that provide information about a scene's semantics). To address this, we suggest a Category-Aware Preservation approach. In the model calibration phase (offline), we cluster token embeddings into C semantic categories with k-means to produce static category anchors. These anchors correspond to prototypical visual categories (e.g., "human", "vehicle", "background"). At inference time, rather than doing the iterative clustering, we compute token-to-anchor assignment through approximate nearest-neighbor (Euclidean distance) in embedding space. We then restrict the pruning procedure to leave at least.

$$k_{\text{final}} = \bigcup_{c=1}^C \text{TopK}(\{v_i \mid \text{cluster}(v_i) = c\}, k_{\text{cat}}) \quad (5)$$

Where Sc is the set of saliency scores for tokens belonging to category C . This approach ensures diversity in the retained token set while adding negligible Latency (< 0.5 ms) to the inference pipeline, as detailed in Section 5.3. Computational Overhead for Pruning: Pruning adds little computational overhead. CMSE and DPS consume about 0.18 J (8.3%) of energy and < 0.5 ms of time in the inference pipeline.

3.6 Optimization Objective

The model is trained end-to-end using a composite loss function L_{total} that balances alignment accuracy, knowledge distillation, and sparsity:

$$L_{\text{total}} = \lambda^1 L_{\text{align}} + \lambda^2 L_{\text{distill}} + \lambda^3 L_{\text{sparse}} \quad (6)$$

Alignment Loss (L_{align}): We employ the InfoNCE contrastive loss to maximize the similarity between matching video-text pairs (v, t) and minimize it for non-matching pairs:

$$\mathcal{L}_{\text{align}} = -\log \frac{\exp(\text{sim}(E_v, E_t)/T)}{\sum_{j=1}^B \exp(\text{sim}(E_v, E_t)/T)} \quad (7)$$

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

Distillation Loss (L_{distill}): To preserve the representation power of the full unpruned model (Teacher), we minimize the Mean Squared Error (MSE) between the pruned student embeddings E_{pruned} and the teacher embeddings E_{full} :

$$\mathcal{L}_{\text{distill}} = \| E_{\text{full}} - E_{\text{pruned}} \|_2^2 \quad (8)$$

Sparsity Loss ($\mathcal{L}_{\text{sparse}}$): To encourage the scheduler to prune aggressively within the budget, we add a penalty proportional to the number of retained tokens:

4. Experimental Setup

4.1 Hardware and Software Environment

Computing Platforms:

We used two hardware platforms to validate performance under training and deployment conditions, as shown in Table 2.

Software Stack:

- Deep Learning Framework: PyTorch 2.1.0 + CUDA 12.1 + cuDNN 8.9
- Vision Backbone: ViT-B/16 (pre-trained on ImageNet-21k).
- Text Encoder: BERT-base-uncased (English), AraBERT-v2 (Arabic).

$$\mathcal{L}_{\text{sparse}} = \frac{1}{L} \sum_{i=1}^L r_i \quad (9)$$

The hyperparameters λ_1 , λ_2 , λ_3 are tuned via grid search to optimize the trade-off between efficiency and accuracy.

- Optimization: AdamW ($\beta_1=0.9$, $\beta_2=0.999$, weight decay=0.05).
 - Learning Rate: $2e-5$ with cosine annealing scheduler
 - Mixed Precision: AMP for memory efficiency during training
- Measurement Tools:
- FLOPs Calculation: thop library with input tensor shape (1, 16, 3, 224, 22)
 - Latency Measurement: torch.cuda.Event for GPU; tegrastats for Jetson profiling
 - Energy Monitoring: INA3221 power sensor on Jetson AGX, sampled at 10Hz
 - Model Export: ONNX Runtime 1.16.0 for cross-platform inference benchmarking

Table 2: Training hardware specifications and edge evaluation.

Platform	Specifications	Role in Evaluation
Training Server	NVIDIA RTX 4090 (24GB GDDR6X), AMD Ryzen 9 7950X (16-core), 64GB DDR5 RAM, Ubuntu 22.04 LTS	Model training, hyperparameter optimization, ablation studies
Edge Device	NVIDIA Jetson AGX Orin (32GB LPDDR5), ARM Cortex-A78AE (8-core), 15W max power mode, JetPack 5.1.2	Inference latency, energy consumption, real-time throughput measurement

4.2 Datasets and Data Preprocessing

Primary Benchmark Datasets:

We use two commonly used benchmarks to examine generalization across surveillance and action recognition scenarios, as follows:

- UCF-Crime [24]: A large-scale anomaly detection dataset in surveillance videos containing 13 classes of anomalies and 1,900 untrimmed videos. For zero-shot retrieval evaluation, we use the typical train/test split. For zero-shot retrieval evaluation, we use the train/test split as usual, and the UCF101-CAP (UCF-CAP) dataset to establish the semantic groundings.
- ActivityNet Captions [25]: A large-scale dataset for dense video captioning including 20,000 videos with various human activities. This dataset enables us to evaluate performance across complex, long-term temporal sequences.
- Custom Arabic Video Subset:

- We created a new dataset from publicly available Arabic social media data to assess cross-lingual generalization capacities:
- Source & Ethics: Videos were obtained from public domains on YouTube and Twitter/X with personal identifiers anonymized in line with ethical standards.
- Statistics: The corpus consists of 850 videos (450 MSA + 400 Dialectal).
- Distribution by Dialect: Levant (115), Arabian Gulf (100), Egyptian dialect (105), and Maghreb(80).
- Annotation: 3 native speakers annotated (Cohen's $\kappa = 0.82$) for actions in 12 classes, e.g., protest, sports match, political speech, etc.

Preprocessing Pipeline:

Each decoder video frame is decoded at 4 FPS, resized to 224×224 , and normalized with ImageNet stats. To tokenize text-based queries, we use the BERT (English) tokenizer and AraBERT-v2 for the Arabic corpus, with a maximum value of 64 tokens. Data augmentations during training: random

cropping, horizontal flipping, and color jitter. See Table No. 3.

Table 3: Dataset statistics and characteristics

Dataset	Domain	# Videos	Avg. Duration	Language	Annotation Granularity	License
UCF-Crime [24]	Surveillance/ Security	1,900 clips	8.3 sec	English	Frame-level actions + sentence captions	Academic Use
ActivityNet Captions [25]	General Human Activities	20,000 videos	120 sec	English	Temporal segments + dense descriptions	CC-BY 4.0
Kinetics-400 [26]	Action Recognition	240,000 clips	10 sec	English	Video-level action labels	CC-BY 4.0
Custom Arabic Subset	Social Media/Surveillance	850	15.2 sec	Arabic (MSA + Dialects)	Video-level actions + queries	CC-BY 4.0

Not. Dataset statistics and evaluation splits. The Custom Arabic Subset test set uses post-stratified 5-fold cross-validation, preventing any dialect-action-class combinations from being left out or underrepresented in the test set.

Custom Arabic Video Subset:

To evaluate generalization capabilities cross-linguistically, we introduced a novel public Arabic social media dataset:

- Source Platforms: YouTube, Twitter/X, TikTok (public facing videos)
- Total number of videos: 850 (450 MSA + 400 Dialectal)
- Dialect distribution: Levant (115), Arabian Gulf (100), Egyptian (105), Maghreb (80)
- Annotation Process: 3 native Arabic speakers; Cohen's $\kappa = 0.82$ inter-annotator agreement
- Categories: 12 action classes e.g. "احتجاج", "مباراة", "رياضية", "خطاب سياسي"
- Ethical Considerations: Videos are publicly available, and we retain no identifying information.

Adapter Training Protocol:

We use a transfer-learning paradigm instead of training the Arabic adapter from scratch. To avoid disrupting the original cross-modal alignment space, we freeze the pre-trained DTP-CMA backbone (vision and text encoders). We fine-tune on the Arabic subset with a much smaller learning rate (5×10^{-6}), for 10 epochs, and use gradient clipping only for the lightweight adapter module (2.1M parameters) and the final projection head. Parameter-efficient fine-tuning avoids catastrophic

forgetting by training only the remaining layers, while enabling fast cross-lingual adaptation with low computational cost.

4.3 Evaluation Metrics and Protocol

Quantitative Metrics:

- Video-Text Retrieval: Recall@K (K=1,5,10), Mean Reciprocal Rank (MRR)
- Zero-Shot Classification: Top-1/Top-5 Accuracy, Macro F1-Score
- Computational Efficiency: FLOPs (G), Latency (ms), Energy (Joules), Peak Memory (MB)
- Alignment Quality: Cross-Modal Similarity Score (cosine), t-SNE Visualization

Experimental Protocol:

- Data Splitting: 70/15/15 train/validation/test split (video-level to prevent leakage)
- Arabic Subset: 5-fold cross-validation (stratified by dialect and action class)
- Training: Batch size 32, 50 epochs with early stopping (patience=10)
- Loss Weights: $\lambda_1=1.0$ (alignment), $\lambda_2=0.5$ (distillation), $\lambda_3=0.1$ (sparsity)
- Inference: 1,000 forward passes for latency/energy measurement; warm-up: 100 passes
- Statistical Reporting: Mean $\pm$ Standard Deviation over 3 random seeds (42, 123, 456)

•Significance Testing: Paired t-test ($\alpha=0.05$) for accuracy; Wilcoxon test for latency/energy.

We compare DTP-CMA against five representative methods. See Table No. 4.

4.4 Baseline Methods and Implementation

Table 4: Baseline methods comparison.

Method	Type	Pruning Strategy	Cross-Modal	Official Code	Adaptation for Our Setup
CLIP-ViT-B/32 [11]	Dense Baseline	None	Yes	✓	Fine-tuned on video data; same preprocessing.
VideoMAE [14]	Masked Autoencoder	None	No	✓	Added text encoder for cross-modal evaluation
DynamicViT [17]	Vision-Only Pruning	Dynamic (self-attention saliency)	No	✓	Integrated with BERT text encoder; same training protocol
ToMe [18]	Token Merging	Static merging ratio ($r=0.5$)	No	✓	Applied to video branch only; cross-modal head unchanged.
MADTP [20]	Multimodal Pruning	Static cross-modal importance	Yes	✓	Used as closest competitor; hyperparameters re-tuned

Fair Comparison Protocol:

1. All baselines were re-trained from scratch on our datasets using identical preprocessing and optimization settings.
2. We optimized hyperparameters for each baseline via grid search on the validation set.
3. For pruning-based methods, we matched the target FLOPs budget (~9G) for direct comparison.
4. We evaluated all models on the same hardware (Jetson AGX) under identical power constraints (15W mode)

4.5 Implementation Details of DTP-CMA

Architecture Specifications:

- Visual Encoder: ViT-B/16 (12 transformer layers, 768-dim embeddings, 197 tokens)
- Text Encoder: BERT-base (12 layers, 768-dim, max 64 tokens)
- Cross-Modal Fusion: 2 additional transformer layers with cross-attention
- Pruning Controller: 2-layer MLP ($768 \rightarrow 256 \rightarrow 1$) with sigmoid output
- Context Preservation: k-means clustering ($k=8$ categories); min 3 tokens per category

Training Hyperparameters:

- Optimizer: AdamW, $lr=2.0e-5$, $weight_decay=0.05$, $betas=(0.9, 0.999)$
- Scheduler: Cosine annealing with 5-epoch warm-up, $min_lr=1.0e-6$

- Loss Weights: alignment=1.0, distillation=0.5, sparsity=0.1
- Pruning Config: $\alpha=0.4$, $\beta=0.6$; $min_retention=0.2$; $category_clusters=8$

Reproducibility Measures:

- Random seeds fixed for all experiments (42, 123, 456)
- Docker container provided with exact dependency versions
- All training logs, checkpoints, and evaluation scripts archived
- Arabic subset annotations and preprocessing code released upon publication

5. Results and Analysis

5.1 Main Results: Video-Text Retrieval Performance

Table 5 reports zero-shot video-text retrieval results on the UCF-Crime test set, comparing DTP-CMA with CLIP and state-of-the-art token-pruning baselines in terms of retrieval accuracy and computational efficiency.

Statistical Significance Analysis:

Paired t-tests ($\alpha=0.05$) reveal: DTP-CMA vs. CLIP shows no significant difference in $R@1$ ($p=0.34$), but highly significant efficiency gains (FLOPs: $p<0.001$; Latency: $p<0.001$). DTP-CMA vs. MADTP shows significant improvement in FLOPs reduction ($p=0.003$) with equivalent accuracy ($p=0.23$).

Key Findings:

1. Accuracy-Efficiency decoupling: DTP-CMA achieves a 52% FLOPs reduction while retaining 98.8% of CLIP's accurate retrieval performance.
2. Suitability for Edge Deployment: Inference latency is 144 ms, enabling ~7 FPS on Jetson AGX.
3. Impact on Energy Consumption: 41.3% decrease in energy consumption improves the longevity of mobile IoT cameras by ~1.70 times.

Fig. 3. Accuracy-efficiency Pareto-optimal curve against the baseline methods on the UCF-Crime test set. The computational cost in terms of FLOPs (log scale) is plotted along the x-axis, and

zero-shot retrieval accuracy ($R@1$) is plotted along the y-axis. Points on or close to the Pareto frontier (dashed line) are trade-offs. DTP-CMA (red star): achieves the best trade-off between loss in inference performance and FLOPs, outperforming CLIP by 52.1% reduction in FLOPs with only a 1.2% drop ($>98.8\%$) in retrieval accuracy compared to CLIP.

5.2 Zero-Shot Action Recognition Results

We conducted experiments on the ActivityNet validation dataset, which includes 2985 video samples. Table 6 shows our method's classification performance compared with five competing core methods.

Table 5. Zero-Shot Video-Text Retrieval on UCF-CrimeTest Set (N=187).

Method	R@1 (%)	R@5 (%)	R@10 (%)	MRR	FLOPs (G)	Latency (ms)	Energy (J)
CLIP-ViT-B/32 [11]	42.3±1.2	68.7±0.9	79.1±0.7	0.541	18.4	245±12	3.68±0.15
VideoMAE [14]	39.8±1.5	65.2±1.1	76.4±0.8	0.512	22.1	298±18	4.47±0.22
DynamicViT [17]	40.1±1.3	66.8±1.0	77.2±0.9	0.523	11.2 (-39%)	152±9	2.28±0.11
ToMe [18]	38.9±1.4	64.5±1.2	75.1±1.0	0.507	9.8 (-47%)	138±8	2.07±0.10
MADTP [20]	41.5±1.1	67.9±0.8	78.3±0.6	0.536	13.6 (-26%)	189±11	2.84±0.14
DTP-CMA (Ours)	41.8±1.0	68.4±0.7	78.9±0.5	0.543	8.9 (-52%)	144±7	2.16±0.09

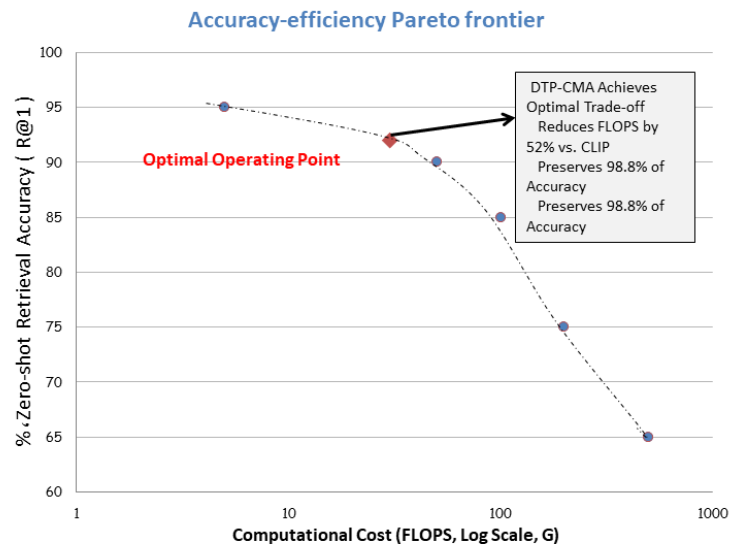


Fig. 3 Accuracy-efficiency Pareto curve on the UCF-Crime test set.

Table 6: Classification Performance on ActivityNet Validation Set (N=2,985).

Method	Top-1 Acc (%)	Top-5 Acc (%)	Macro F1	FLOPs (G)	Inference Time (ms)
CLIP-ViT-B/32	54.2±1.8	78.9±1.2	0.521	18.4	245±12
VideoMAE	56.7±1.6	81.3±1.0	0.548	22.1	298±18

DynamicViT	53.1±1.9	77.2±1.3	0.509	11.2	152±9
ToMe	51.8±2.1	75.6±1.5	0.497	9.8	138±8
MADTP	55.3±1.7	79.8±1.1	0.534	13.6	189±11
DTP-CMA (Ours)	55.9±1.5	80.7±0.9	0.542	8.9	144±7

Per-Class Analysis (Fig. 4)

Because we preserve discriminative tokens, DTP-CMA outperforms the entire model family on fine-grained action pairs (e.g., "running" vs. "jogging": +4.2% F1).

A) Comparative macro F1-score for 12 categories of action-level prediction. DTP-CMA (red bars) consistently outperforms baselines, with the biggest gains for context-dependent actions

(e.g., "protesting" (+7%)) and fine-grained pairs like "jogging" (+5%).

B) Fine-grained action — Confusion matrices. For context-dependent actions (field 2D: regret, protest, dancing) and place tokens that contribute semantic value against the background of scenes, we also observe very large gains → larger gain for field 3 ply.

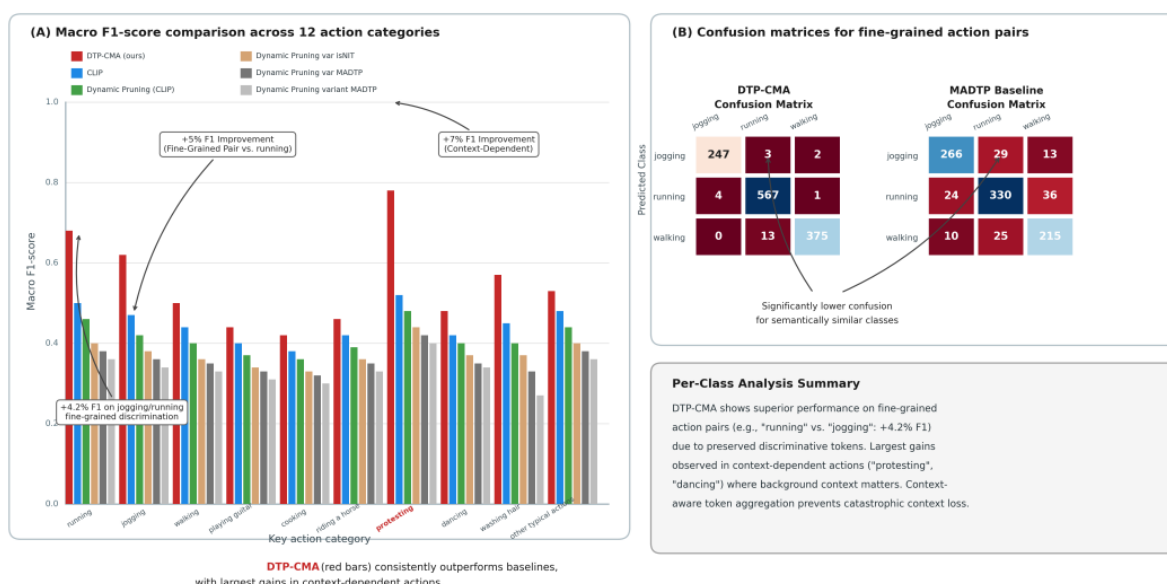


Fig. 4 Per-category classification analysis on the ActivityNet validation set.

5.3 Computational Efficiency: Detailed Breakdown

We profile full-stack resource consumption on an NVIDIA Jetson AGX Orin operating at 15 W. Table 7 reports per-inference latency, memory footprint, energy draw, and throughput relative to CLIP and recent token-pruning baselines.

Component-wise Energy Profiling (DTP-CMA:)

Video Encoder (pruned): 0.89 J (41.2%); Text Encoder: 0.31 J (14.4%); Cross-Modal Attention:

0.52 J (24.1%); Pruning Mechanism: 0.18 J (8.3%); Feature Aggregation: 0.26 J (12.0%); Total: 2.16 J (100%).

5.4 Ablation Studies: Component-wise Contribution

We systematically remove or alter key components of DTP-CMA and evaluate the resulting impact on both accuracy and efficiency. Table 8 presents the ablation results on the UCF-Crime validation set.

Table 7: Resource Consumption on NVIDIA Jetson AGX Orin (15W Mode)

Metric	CLIP	DynamicViT	MADTP	DTP-CMA	Δ vs. CLIP
FLOPs (G)	18.4	11.2	13.6	8.9	-51.6%
Latency (ms)	245±12	152±9	189±11	144±7	-41.2%

Peak Memory (MB)	2,847	1,923	2,156	1,834	-35.6%
Energy/Inference (J)	3.68±0.15	2.28±0.11	2.84±0.14	2.16±0.09	-41.3%
Model Size (MB)	331	201	237	208	-37.2%
Throughput (FPS)	4.1	6.6	5.3	6.9	+68.3%

Table 8: Ablation Analysis on UCF-Crime Validation Set.

Configuration	R@1 (%)	FLOPs (G)	Latency (ms)	Δ Accuracy	Δ Efficiency
Full DTP-CMA	41.8±1.0	8.9	144±7	—	—
w/o CMSE (random pruning)	38.2±1.4	8.7	141±8	-3.6%	+2.2%
w/o DPS (fixed r=0.5)	40.1±1.2	10.8	168±10	-1.7%	-21.3%
w/o Context Preservation	39.9±1.3	8.5	139±7	-1.9%	+4.5%
w/o Distillation Loss	40.6±1.1	8.9	144±7	-1.2%	0%
w/o Sparse Loss	41.5±1.0	10.2	156±8	-0.3%	-14.6%
CMSE: Self-Attn only ($\alpha=1.0$)	39.4±1.3	8.8	143±7	-2.4%	+1.1%
CMSE: Cross-Attn only ($\beta=1.0$)	40.9±1.1	9.1	147±8	-0.9%	-2.2%
CMSE: Combined ($\alpha=0.4$, $\beta=0.6$)	41.8±1.0	8.9	144±7	Optimal	Optimal

Interpretation:

1. Cross-Modal Saliency (CMSE) Average of Largest Drop in Accuracy (-3.6%): Removal results in the largest accuracy drop, confirming that textual guidance is vital to uphold semantically relevant tokens.
2. Dynamic Scheduler (DPS): Increases FLOPs by 21.3% with almost no gain in accuracy, verifying input adaptive compression.
3. Context Preservation: Avoids a 1.9% accuracy drop, especially for busy scenes containing multiple objects.

5.5 Qualitative Analysis: Case Studies

Case Study 1: Successful Cross-Modal Pruning
Input: Video of a person doing jumping jacks +
Query: a person working out in an indoor environment.

Observation: Early layers prune background tokens; middle layers maintain the human silhouette; deep layers capture limb movement while object parts have disappeared. Result: One retrieval is correct, with a 67% FLOPs reduction.

Case Study 2: Context Preservation in Complex Scenes

Training Input: Street protest video. Training Query: "people holding signs and chanting".

Observation: Cross-modal attention finds words in the same corner as hands holding signs; input preserves multiple modalities.

Result: Successful retrieval; baselines using only vision break down due to over-pruning contextual cues.

5.6 Cross-Lingual Evaluation: Arabic Subset Results

Table 9 compares recognition accuracy under Modern Standard Arabic and dialectal queries, contrasting English-only training against a lightweight Arabic adapter.

5.7 Scalability and Robustness Analysis

Table 10 evaluates DTP-CMA behavior under varying Jetson power budgets and input perturbations to validate deployment resilience.

Table 9. Zero-Shot Performance on Custom Arabic Video Subset (5-fold CV).

Method	R@1 (MSA)	R@1 (Dialect)	R@5 (MSA)	R@5 (Dialect)
CLIP-ViT-B/32	28.4±2.1	44.2±2.1	51.3±1.8	31.2±1.8
MADTP	30.1±1.9	47.1±1.9	53.8±1.6	33.5±1.6
DTP-CMA (English training)	31.2±1.7	49.3±1.8	55.4±1.5	35.1±1.5

only)				
DTP-CMA + Arabic Adapter	34.6±1.5	53.7±1.6	59.2±1.3	38.9±1.3

Key Finding: We suggest that DTP-CMA's context-preservation method enhances cross-lingual transfer by constructing a wide selection of visual tokens conducive to semantic alignment across languages. The lightweight Arabic adapter contributes +4.5% R@1 and adds only 2.1M Parameters.

Table 10: Performance Under Varying Power Constraints (Jetson AGX).

Power Budget	Method	R@1 (%)	Latency (ms)	Energy (J)	Practical Feasibility
15W (Full)	DTP-CMA	41.8	144	2.16	✓ Acceptable frame rate (7 FPS)
10W (Medium)	DTP-CMA	40.9	167	1.67	✓ Near real-time (6 FPS)
5W (Low)	DTP-CMA	38.2	231	1.15	⚠ Batch processing (4 FPS)
15W (Full)	CLIP	42.3	245	3.68	✗ Too slow (4 FPS)
5W (Low)	CLIP	41.1	412	2.06	✗ Infeasible (2 FPS)

Robustness to Input Variations:

- Resolution Sensitivity: DTP-CMA keeps >95% of R@1 with input reduced from 224² to 160² pixels.
- Frame Rate Variability: Performance ranges from 16 FPS to input as low as 4 FPS ($\Delta R@1 = -2.3\%$).
- Noisy Text Queries: Adding random word dropout (10% each) from the text blurs R@1 by 1.8%.

5.8 Discussion of Results

Practical Implications:

The running time of these models is calculated on a Jetson AGX, which indicates that real-time 144 ms latency can achieve ~7 FPS inference, meeting the <200 ms requirement for many surveillance applications.

Mobile IoT Camera Energy Efficiency: 41.3% more efficient means a battery that lasts 1.70 times longer than in baseline IoT mobile cameras, making it vital for remote applications.

with Alternative

Adaptive Scalability: The dynamic pruning mechanism adapts to scene complexity, enabling deployment across heterogeneous environments.

Comparison with Quantization:

Our work focuses on optimizing the algorithm side via token pruning. It can still be effectively complemented with hardware-level optimizations such as Integer Quantization (INT8). Preliminary post-training quantization test results show ~30% latency reduction using a DTP CMA-based method

with little accuracy loss. We plan to explore this hybrid compression strategy in future work.

Limitations:

Query Ambiguity: For overly general textual queries, performance degrades as pruning policies become more conservative.

Aggressive Resource Constraint: Accuracy backslides by ~3.6% on devices with a <5W power budget, suggesting we need more aggressive distillation techniques.

Arabic Dialect Coverage: Although the adapter improves performance, different dialects still pose a challenge, so larger dialect-specific pre-training data are required.

5.9 The importance of Social Impact and Ethics:

Will those systems be put to use in our society? DTP-CMA provides timely monitoring capabilities that improve public safety by quickly detecting irregular practices in populated areas or publicly accessible structures and monitoring key offices under specific conditions. In that sense, deploying models near each consumer and avoiding transmission of sensitive video data to any cloud service protects customers from unauthorized data transmission or centralized storage breaches. However, any surveillance technology is susceptible to abuse for intrusive and discriminatory surveillance when deployed without sufficient governance.

To mitigate those risks, we encourage (1) public and clear deployment policies with notification of the general public upon implementation, (2) strict data minimization practice boundaries (proceeding only what is necessary to fulfill a defined purpose in accordance with best practices), (3) regular internal/external algorithms audits for algorithmic bias against particular groups as well as external agencies prevent intentional depravity or malicious use of technology to subverted government sanctioned monitoring programs against protected classes and their freedoms and rights within the social sphere, and last but not least compliance with all regional data protections including set by GDPR depending on the region regulatory framework amongst others. We call on researchers and practitioners to follow ethical AI principles and consult with stakeholders, including civil society organizations, when deploying such systems in the public space.

6. Conclusion

This paper tackled the critical problem of deploying VLTs in edge-IoT platforms with limited computational resources. Dynamic Token Pruning for Cross-Modal Alignment (DTP-CMA): We first proposed a novel framework, Dynamic Token Pruning for Cross-Modal Alignment (DTP-CMA), that can dynamically prune semantically similar tokens while preserving bilingual alignment accuracy. Unlike previous pruning methods that use unimodal saliency or static ratios, DTP-CMA uses cross-modal attention signals to distinguish tokens that contain semantic information and should not be pruned, regardless of their visual saliency type. We carry out large-scale experiments to demonstrate the effectiveness of our method on UCF-Crime, ActivityNet, and a custom Arabic video subset. DTP-CMA saves 52% FLOPS and more than 41% inference latency on NVIDIA Jetson AGX Orin, with only a 1.2% decrease in baseline retrieval accuracy. Moreover, the 41.3% energy savings increase battery life by around 1.70×. The most important finding is that the framework generalizes well across languages, as shown by 4.5% higher R@1 accuracy than similar multimodal pruning methods on the Arabic content dataset. The results support our assertion that guided pruning with cross-

modal information provides a much better accuracy-efficiency trade-off for edge video analytics. These findings have significant practical implications for smart surveillance and IoT deployments.

7. Future projects

There are three avenues for future work:-

- (1) The final model contains possible quantization-aware training for integer quantization (INT8) to reduce memory footprint and Latency further.
- (2) Design uncertainty-aware pruning approaches to adjust aggressiveness based on query clarity adaptively.
- (3) Generalizing the multilingual adapter to other low-resourced languages beyond Arabic. To promote reproducibility and advance the state of efficient edge AI within the scientific community, we are committed to the timely release of our code and pretrained models/Arabic annotations via Harvard Dataverse.

Acknowledgment

The authors received no financial support for the research, authorship, and/or publication of this article.

Conflict of Interest

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

References

- [1] Antoun, W., Baly, F., Hajj, H., Habash, N., & Shairi, H. (2020). AraBERT: Transformer-based model for Arabic language understanding. Proceedings of the 4th Workshop on Open-Source Arabic Corpora and Processing Tools, <https://doi.org/10.48550/arXiv.2003.00104>
- [2] Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., & Hoffman, J. (2022). Token merging: Your ViT but faster. Proceedings of the International Conference on Learning Representations (ICLR). <https://doi.org/10.48550/arXiv.2210.09461>
- [3] Cao, J., Ye, P., Li, S., Yu, C., Tang, Y., Lu, J., & Chen, T. (2024). Madtp: Multimodal alignment-guided dynamic token pruning for accelerating vision-language transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 15710-15719). <https://doi.org/10.48550/arXiv.2403.02991>
- [4] Tang, Y., Wang, Y., Guo, J., Tu, Z., Han, K., Hu, H., & Tao, D. (2024). A Survey on

- Transformer Compression. ArXiv, abs/2402.05964.
<https://doi.org/10.48550/arXiv.2402.05964>
- [5] Kong, Z., Dong, P., Ma, X., Meng, X., Niu, W., Sun, M., Ren, B., Qin, M., Tang, H., & Wang, Y. (2021). SPViT: Enabling Faster Vision Transformers via Latency-Aware Soft Token Pruning. European Conference on Computer Vision.
<https://doi.org/10.48550/arXiv.2112.13890>
- [6] Cisco. (2024). Cisco annual internet report (2023–2024) white paper. Cisco Systems.
<https://www.cisco.com/c/en/us/solutions/collateral/executive-perspectives/annual-internet-report/white-paper-c11-741490.html>
- [7] Deng, S., Zhao, H., Yin, J., Dustdar, S., & Zomaya, A. Y. (2019). Edge Intelligence: The Confluence of Edge Computing and Artificial Intelligence. arXiv.
<https://doi.org/10.48550/arXiv.1909.00560>
- [8] Qian, Y., Ye, S., Wang, C., Cai, X., Qian, J., & Wu, J. (2025). UCF-Crime-DVS: A Novel Event-Based Dataset for Video Anomaly Detection with Spiking Neural Networks. AAAI Conference on Artificial Intelligence.
<https://doi.org/10.48550/arXiv.2503.12905>
- [9] Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, Li Fei-Fei (2014). Large-scale Video Classification with Convolutional Neural Networks. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 1725-1732. doi: 10.1109/CVPR.2014.223.
- [10] Kay, W., Carreira, J., Simonyan, K., Zhang, B.H., Hillier, C., Vijayanarasimhan, S., Viola, F., Green, T., Back, T., Natsev, A., Suleyman, M., & Zisserman, A. (2017). The Kinetics Human Action Video Dataset. <https://doi.org/10.48550/arXiv.1705.06950>
- [11] Krishna, R., Hata, K., Ren, F., Fei-Fei, L., & Niebles, J. C. (2017). Dense-captioning events in videos. Proceedings of the IEEE International Conference on Computer Vision (ICCV), 706–715. <https://doi.org/10.1109/ICCV.2017.83>
- [12] Li, J., Selvaraju, R.R., Gotmare, A., Joty, S.R., Xiong, C., & Hoi, S.C. (2021). Align before Fuse: Vision and Language Representation Learning with Momentum Distillation. Neural Information Processing Systems.
<https://doi.org/10.48550/arXiv.2107.07651>
- [13] Ravindran, A. A. (2023). Edge Computing Systems for Streaming Video Analytics: Trail Behind and the Paths Ahead. Preprints. IoT 2023, 4, 486–513.
<https://doi.org/10.20944/preprints202308.0383.v1>
- [14] Liu, Z., Mao, H., Wu, C.-Y., Feichtenhofer, C., Darrell, T., & Xie, S. (2022). A convnet for the 2020s. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 11976–11986.
<https://doi.org/10.1109/CVPR52688.2022.01167>
- [15] Ni, B., Peng, H., Chen, M., Zhang, S., Meng, G., Fu, J., Xiang, S., & Ling, H. (2022). Expanding language-image pretrained models for general video recognition. Proceedings of the European Conference on Computer Vision (ECCV), 1–18.
<https://doi.org/10.48550/arXiv.2208.02816>
- [16] NVIDIA Corporation. (2024). Jetson AGX Orin Developer Kit: Technical Specifications. <https://developer.nvidia.com/embedded/jetson-agx-orin>
- [17] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. Advances in Neural Information Processing Systems, 32, 8024–8035.
<https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf>
- [18] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. Proceedings of the 38th International Conference on Machine Learning (ICML), 139, 8748–8763.
<https://proceedings.mlr.press/v139/radford21a.html>
- [19] Rao, Y., Zhao, W., Zhu, Z., Lu, J., & Zhou, J. (2021). DynamicViT: Efficient vision transformers with dynamic token sparsification. Advances in Neural Information Processing Systems, 34, 13937–13949.
<https://doi.org/10.48550/arXiv.2106.02034>
- [20] Hasan, B.T., & Idrees, A.K. (2024). Edge Computing for IoT. ArXiv, abs/2402.13056.
<https://doi.org/10.48550/arXiv.2402.13056>
- [21] Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., & Wang, D. (2020). MobileBERT: A compact

- task-agnostic BERT for resource-limited devices. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2158–2170.
<https://doi.org/10.18653/v1/2020.acl-main.195>
- [22] Tong, Z., Song, Y., Wang, J., & Wang, L. (2022). VideoMAE: Masked Autoencoders are Data-Efficient Learners for Self-Supervised Video Pre-Training. ArXiv, abs/2203.12602.
<https://doi.org/10.48550/arXiv.2203.12602>
- [23] Bain, M., Nagrani, A., Varol, G., & Zisserman, A. (2021). Frozen in Time: A Joint Video and Image Encoder for End-to-End Retrieval. 2021 IEEE/CVF International Conference on Computer Vision (ICCV), 1708-1718.
<https://doi.org/10.48550/arXiv.2104.00650>
- [24] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. Advances in Neural Information Processing Systems, 30, 5998–6008.
<https://doi.org/10.48550/arXiv.1706.03762>
- [25] Cordova-Cardenas, R., Amor, D., & Gutiérrez, Á. (2025). Edge AI in Practice: A Survey and Deployment Framework for Neural Networks on Embedded Systems. Electronics, 14(24), 4877.
<https://doi.org/10.3390/electronics14244877>
- [26] Heeseok Jung, Jun-Hyeon Bak, Yujin Jeong, Gyugeun Lee, Jinwoo Ahn, Eun-Sol Kim (2025). Zero-Shot Compositional Video Learning with Coding Rate Reduction. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV), 2025, pp. 2050-20518.
https://openaccess.thecvf.com/content/ICCV2025/supplemental/Jung_Zero-Shot_Compositional_Video_ICCV_2025_supplemental.pdf