

A comparative review of video-based transformer models for modern applications

Ansam Nazar Younis ^{1*}, Fawziya Mahmood Ramo ²

¹Department of Networks, Computer Science and Mathematics, University of Mosul, Mosul, Iraq

²Remote Sensing Center, University of Mosul, Mosul, Iraq

ARTICLE INFO

Received: 31/07/2025
Revised: 26/10/2025
Accepted: 28/10/2025
Available online: 08/08/2026
August Issue
[10.37652/juaps.2025.163476.1571](https://doi.org/10.37652/juaps.2025.163476.1571)

 CITE @ JUAPS

Corresponding author

Ansam Nazar Younis
anyma8@uomosul.edu.iq

ABSTRACT

With advances in technology, handling video files is of paramount importance, especially in the educational, medical, and security fields. Various applications have been developed to understand, present, distinguish, retrieve, extract semantic meaning, or create video using current artificial intelligence technologies. One of the most recent and important technologies is the vision transformer, developed to accurately handle video. In this paper, we discuss the most important developments in this intelligent technology and its main applications over the past few years. In this paper, we present a comprehensive comparative review of the most prominent visual transformer models used in video processing, with an in-depth analysis of their architectures, identification of their applications, and discussion of their advantages and disadvantages. What distinguishes this work from previous reviews is its focus on classifying transformer-based model types, linking them to computational complexity, the type of attention used, and performance across various tasks. This paper also includes analytical tables linking each model to the dataset used and the results achieved, highlights current research gaps, and offers informed future recommendations for developing these models to serve advanced smart applications.

Keywords: Action recognition models, Prediction models, Transformer architecture, Video classification models, ViViT model

1 INTRODUCTION

Video processing and analysis are complex tasks in modern applications, given the abundance of spatial and temporal information in each video file and the multiple local and global scales it spans. Researchers have resorted to various methods to handle this massive amount of data, which has necessitated the provision of substantial resources, especially as the number of video clips increases and their sizes and sources become more diverse [1].

Among the most prominent methods in this field are convolutional neural networks (CNNs) and recurrent neural networks (RNNs), which are able to extract spatial and temporal features. However, these models suffer

from structural limitations, such as sequential processing in RNNs and local attention in CNNs, which limit their ability to represent long-term temporal relationships, leading to a relatively large extrapolation bias [1–3]. The emergence of transformer models, such as ViT [2], has helped overcome these limitations through self-attention mechanisms that allow for global, non-local representation of frames so it helped in longer term temporal relationships, with the potential for parallel training, albeit sometimes at the expense of an increased number of parameters or memory consumption [1, 3]; despite the clear advantages of transformers in processing long sequences, practical applications have shown varying performance depending on the nature of the task and

the experimental environment. Some models, such as TimeSformer [4] for fine-grained motion capture tasks and ViViT [5] for multimedia video classification, have outperformed others, reflecting the importance of choosing the appropriate model based on the application context.

This paper aims to conduct a comprehensive comparative review of the most important and common transformer models for various modern applications, identifying their most prominent advantages and disadvantages, explaining and analyzing the architecture, and reviewing the level of computational complexity of the important and main types, from which most other models are derived, along with the most important data sets they use and the accuracy they have achieved.

A search was conducted for papers that serve the objectives of our research paper in important journals and well-known scientific websites such as "IEEE Xplore," "Google Scholar," and "ScienceDirect. This paper primarily focuses on modern transformer models used to serve video content across various applications and fields. Papers were selected based on two criteria: they must be directly related to video processing using vision transformers, and they must have been published in international peer-reviewed journals within the past few years. What distinguishes this paper from previous reviews is its focus on the applied aspects of the models, linking each model to the dataset used and the results achieved, and analyzing research gaps and providing well-considered future recommendations.

2 BACKGROUND OF TRANSFORMER MODELS

The Transformer Model was introduced in the 2017 paper [6] by Ashish Vaswani et al., which introduced the concept of "self-attention" as an alternative to recurrent layers in machine translation models. This architecture overcomes the structural limitations of sequence models, such as the vanishing gradient problem in recurrent neural networks (RNNs), by processing the entire sequence at once, enabling parallel training capabilities. Transformers also excel at modeling non-local relationships, unlike convolutional networks (CNNs), which rely on local attention and exhibit a locality bias, limiting their ability to represent distant contexts. Although the transformer architecture lacks some useful structural biases that were in convolutional networks, such as spatial symmetry and local attention, this structural flexibility gives it a high

ability to adapt to multiple types of data, including images and videos, provided sufficient training data is provided, or local biases are explicitly introduced [7].

The continued success of that pure transformer led to the development of the vision transformer in 2020 [2]. This extends the transformer framework to the still-image domain and processes the image as a time series within the transformer, without requiring traditional convolution operations. The image is split into patches and converted to one-dimensional vectors.

Because of the need for models that handle video files rich in spatial and temporal information, despite their large sizes, these models led to the design of the ViViT model, a model created especially for video processing and classification that leverages the spatiotemporal characteristics of video frames. It was a necessary structure that efficiently represented these relationships. A team from Google presented this model [5] (see Figure 1), aiming to enable the transformer to train on relatively small data using a model previously trained on large data.

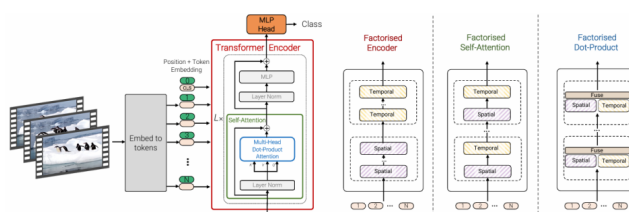


Fig. 1 ViViT architecture shapes [5]

ViViT relies on two basic methods for patch formation. The first, uniform frame sampling, means including each frame in a group of batches in succession until all frames are processed, so they can all be linked together later and pushed to the transformer encoder. Here, time information is added to each frame separately. The second method is Tubelet embedding; the non-overlapping frame set is used to preserve temporal and spatial information during tokenization and embedding, which means using space-time segments (time, height, width) to extract a patch. It has a larger number of patches, and the temporal and spatial information is combined. In the first method, only the temporal information is combined. Whether the first or second method is used, the following steps depend on different shapes of known ViT models:

Shape 1: Spatio-temporal attention in this type: all patches extracted from the video, with temporal and spatial attributes, are sent to the standard transformer

encoder (ViT). All patch pairs are processed together across their layers, so this type requires many operations (Figure 2).

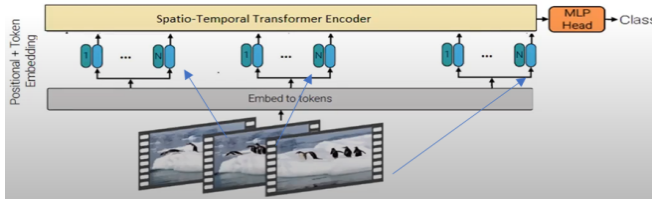


Fig. 2 Spatio-temporal attention [5]

Shape2: Factorized encoder: There are two separate transformers. The first Spatial Transformer Encoder specializes in extracting reliability features from patches at the same time but differing in location, thereby incorporating spatial attentions. Then the second Temporal Transformer Encoder specializes in extracting reliability features from patches from the same location but differing in time, incorporating temporal attentions. This shape requires fewer FLOPs (floating-point operations) because it does not require extracting credentials from all patches in shape1 (Figure 3).

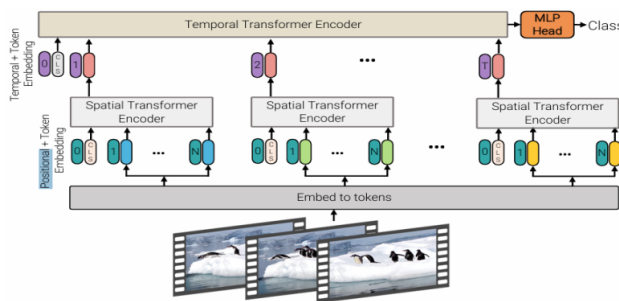


Fig. 3 Factorized encoder(shape2) [5]

Shape3: Factorized self-attention. It is similar to shape1, where a single transformer encoder is used; all patches extracted from the video, carrying temporal and spatial attributes, are sent to the regular transformer encoder. The trick here is to use two blocks: the Spatial Self-Attention Block for spatial attentions and the Temporal Self-Attention Block for temporal attentions, thereby reducing computational complexity, as in shape2 (Figure 4).

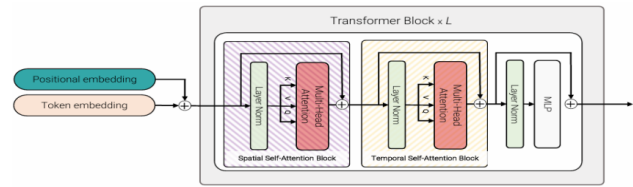


Fig. 4 Factorized self-attention (shape 3) [3]

Shape 4: Factorized dot product. This shape uses the same level of complexity as shapes 2 and 3. It uses a multi-head attention transformer whose operation is similar to shape 3, where the attention heads are divided into two parts: one part includes spatial attention and feeds into the spatial block, and the other includes temporal attention and feeds into the temporal block. A weight value will be used for spatial and temporal attentions separately. Table 1 summarizes the complexity and other properties of the ViViT model's shapes [5].

Table 1 Summary of complexity and properties for the ViViT model's shapes

Model	Name	No. of transformer encoders	components	complexity
Shape 1	Spatio-temporal attention	One	- One Multi-Headed Self-attention	$O((n_t n_h n_w)^2)$
Shape 2	Factorized encoder	Two separately transformer	- Spatial transformer - Temporal transformer	$O((n_h n_w)^2 + n_t^2)$
Shape 3	Factorized self-attention	One	- two self-attention blocks	$O((n_h n_w)^2 + n_t^2)$
Shape 4	Factorized dot-product attention	One	- two self-attention blocks	$O((n_h n_w)^2 + n_t^2)$

3 RESEARCH METHODOLOGY

Previous studies on transformer models in video processing were analyzed using a scientific approach based on a rigorous systematic review. We carefully selected the relevant documents from reliable scientific databases, including IEEE Xplore, Elsevier, Scopus, Web of Science, and Google Scholar. The search process relied on precise keywords to select the appropriate research, such as "ViViT," "Vision Transformer for Video," "Video Recognition Transformer," "Video Classification Transformer," "Video Prediction Transformer," and others. The search was limited to studies published between 2020 and 2025, whether peer-reviewed scientific papers, approved preprints, or reliable university research, with a focus on recent research reflecting the development of transformer models in video processing and the diversity of applications they have seen in recent years.

More than 100 research papers were collected and filtered, and a significant subset was approved. Specific acceptance criteria were adopted for selecting studies, including publication in peer-reviewed journals or at accredited universities, the use of realistic or standard datasets, and the availability of a clear structural description of the model and experimental results. Studies that lacked clear results, did not provide a practical application of the model, or lacked a precise structural description of the model were excluded.

The accepted studies were discussed through two main directions: general reviews and original research. Then, the original research was classified by application type (e.g., video classification, action detection, understanding, and modeling). This paper summarizes the model characteristics using descriptive analysis, and the results are presented in comparative tables that include the number of papers in the particular fields, application fields, datasets, models used, and the achieved accuracy. Based on this classification, models' weaknesses, strengths, and structures were also identified. This information will help guide future research toward more effective models suitable for real-world use.

4 RELATED WORKS

Several comprehensive reviews have addressed the development of transformer models in video processing. The most notable are the review by [1], which classified video transformer models by architecture, performance, and applications, and the review by [8], which discussed the use of transformers in images and videos with diffusion models and analyzed recent trends in this field. Although these reviews focused on general theoretical and classification aspects, the current study features a comparative applied analysis of the models in terms of computational complexity, attention type, and performance accuracy, linking each model to practical applications, the datasets used, and the number of research papers in each field. The study also provides a functional classification of models by task type, identifies current research gaps, and offers well-considered future recommendations for developing more efficient models that serve modern smart applications.

To illustrate the versatility of transformer models in video processing, previous studies are organized by application domain, ranging from video classification and specialized use cases to miscellaneous fields and other important life applications.

1. **Video classification:** In the video classification domain, studies addressed general video content classification; for instance, ViViT was used to classify violent or non-violent videos [9], and it was used to classify traditional dance videos in [10]. HaViT was also used for video classification [11].
2. **Action Recognition and Detection** Research in action-detection models capable of understanding actions and human movements in video. TimeSformer, by Yunhai Han et al., for robot application [4]. COVER [12], STAR [13], ViViT-FE [14], Evaluating Transformers [15], vision transformer [16], and the Recurrent Vision Transformer model [17] are used to recognize actions in videos. [18] RViViT is used for traffic monitoring. ViViT was used to identify action in video clips [19]. DirectFormer [20], COMEDIAN [21], and the RegNet with the ViViT transformer [22] were used for action spotting and understanding. MAiVAR-T [23], VLM [24], DA-ViViT [25], MM-ViT [26], convolutional transformer networks [27], and ViViT [28] were used to recognize motion in videos, such as detecting driver fatigue in real time.

For video recognition, DualFormer [29], Swin Transformer [30], MTV model [31], MViT [32], Vision Transformers [33, 34], , and GestFormer models were used [35]. To detect fake videos, AVTENet [36], (AVTENet), TransAnomaly [37], and ViViT [38] models were used.

For emotion recognition from videos, the ViViX model was used [39]. The SwinAnomaly model [40] was used to detect anomalies in videos, and the MC-ViViT model in [41] analyzed facial features for poor-perception discrimination. The TRecViT model is used for video understanding and modeling [42].
3. **Prediction and Healthcare Fields** In the prediction fields, numerous architectures aim to predict future outcomes or conditions based on video data. Zhou Zhou et al. [43] use the U-Net and ViViT models to detect objects in the frame. In [44], researchers used a TC-Transformer based on ViViT to select the best view from shots captured by multiple cameras for a suitable TV show. The ConvNeXt and ViViT [45] were developed for vehicle monitoring systems to predict accidents in [46],

and transformers were used in self-driving vehicles to predict their motion. ViViT transformer [47] is used to monitor social and economic differences. The Deformable Video Transformer model [48] aimed to predict the important patches that express the video based on its motion data.

Several models have been developed in the last few years for medical video analysis, such as modified ViViT [49] and [50], to assess pain. The ViViT model [51] is also used to predict patients' responses to a specific therapeutic inhibitor for a certain type of cancer. The RTM [52] was used to predict fetal weight from ultrasound video analysis. In [53], 2.5D models are used to efficiently detect early disease progression in multiple medical radiology images. The ViViT-CNN [54] was used to infer vital signs, such as heart rate and oxygen saturation, from different facial expressions in video clips. The DEViViT model [55] was also used to detect eye blinks across different camera angles. The ViTranZheimer model [56] is used to detect early Alzheimer's disease. A hybrid ViViT model with a CNN [57] is used to determine heart health and estimate heart rate, blood oxygen saturation, and blood pressure from facial expressions.

4. **Segmentation, compression, and adaptation fields** Some models focused on segmenting the video into regions of visual or medical significance, such as identifying injury areas or temporal cues (e.g., the Vivim model [58]). Models such as C-ViViT [59] are used to enhance the spatiotemporal representation of frames and overcome limitations in feature extraction accuracy. Other models relied on augmentation techniques and symbol-count reduction strategies to reduce inference and training costs. For example, the LookupViT model [60], which operates on a fixed set of tokens after compressing large amounts of information. Adapter Transformer [61] and LoSA [62] increased the classification accuracy and reduced the parameters. Furthermore, MViT [32] extracted low-level, explicit visual information in the early layers and complex and coarse features in the upper layers to reduce training costs. While Eventful Transformers [63] focuses on the important locations of the frame, i.e., object, to reduce computational operations.

5. **Miscellaneous Topics** Several transformer models have been developed to monitor animal behavior on farms using video files, with the goal of improving safety and productivity, for instance, ViViT [64] to monitor the behavior of pigs. A hybrid Attention-LSTM model [65] was used to identify plankton species in US rivers, facilitating the detection and classification of invasive species such as mussels. Additionally, the use of ViViT to identify the source of sound or images within a video, whether from musical instruments [66] or gas emissions [67], is a source localization method.

Some studies have proposed models for analyzing impactful moments within a video, such as ViViT [68], to identify memorable scenes with the aim of supporting other applications, such as summarization or recommendation. Table 2 summarizes all previous models and applications.

5 RESULT ANALYSIS AND DISCUSSION

The analysis showed that the studied modifications to the transformer models achieved outstanding performance across a number of applications, demonstrating the feasibility of the proposed approach for classifying and analyzing the modified models, as shown in Table 2. However, they still have shortcomings in areas we really need, such as supporting sustainable environments, clean energy, smart cities, fifth-generation networks, video retrieval, and big data. The work is still open for researchers to innovate and add various smart technologies to increase the benefits of the transformer, especially by reducing the number of parameters and the data required for training, and by improving the accuracy of mining and generating data. A direct quantitative comparison between models is not possible due to the variability of datasets, task objectives, the number of models used, and operating conditions. Therefore, this research focuses on a functional qualitative analysis that highlights the strengths and limitations of each model. Although the basic ViViT model was originally designed for video classification, it has been used for several purposes after being modified in various ways. Table 3 shows the modified models classified by function across the various tasks listed in previous studies, along with the number of studies that addressed them in each field.

Table 2 Comparison of video transformer frameworks and applications

Ref.	Framework	Application	Dataset	Result
[4]	TimeSformer	Robotic Grasping	GelSight	85.0
[5]	ViViT	video classification	Kinetics, Epic Kitchens-100, Moments in Time, Something-Something v2	96.5
[9]	Augmentation, ViViT	Violence Detection	Hockey Fight, Violent Crowd	98.46
[10]	ViViT	Dance Classification	IDEEH-10	99.216
[11]	HaViT	Video Classification	Kinetics, Something-Something v2	95.2
[12]	COVER	Action Recognition	Kinetics, SSv2, Moments-in-Time, Standard COVER	87.9
[13]	STAR	Action Localisation	AVA, AVA-Kinetics, UCF101-24, JHMDB51-21	96.3
[14]	SFA_ ViViT	Action Recognition	Kinetics, EPIC-Kitchens, Something-something v2, Moments in Time	84.25
[15]	Multiple ViViT	Action Recognition	EPIC-KITCHENS-100	72.4
[16]	video transformers	Action Recognition	K400, SSV2, EK100	91.5
[17]	RViT	Action Recognition	Kinetics-400, Jester, Something-Something-V2, Charades	92.31
[18]	RViViT	Traffic Estimation	naturalistic vehicle trajectories	91.5
[19]	ViViT, Video Swin	Action Recognition	UCF-101	95
[20]	DirecFormer	Action Recognition	Jester, Kinetics-400, Something-Something-V2	94.86
[21]	COMEDIAN	Action Spotting	SoccerNet-v2	73.1
[22]	Regnet-ViT-Ensemble	Action Detection	AU task	54.93
[23]	MAiVAR-T	Action Recognition	UCF101	91.2
[24]	FTP	Action Understanding	Kinetics-400, Something-Something V2	93.8
[25]	DA-ViViT	Fatigue detection	DMD	96.24
[26]	MM-ViT	Action Recognition	UCF-101, Kinetics-600, Something-Something-v2,	98.9
[27]	CTN	Action recognition	Epic-Kitchens, Diving 48	78.2
[28]	ViViT	Sign Language Recognition	LSA64	90.3
[29]	DualFormer	Video Recognition	Kinetics-400/600	85.2
[30]	Swin	video recognition	Kinetics	85.9
[31]	MTV	Video Recognition	Kinetics, Moments in Time, Epic-Kitchens-100, Something-Something V2,	98.3
[32]	MViT	video and image recognition	Kinetics, Something Something-v2, Charades, AVA	84.8
[33]	VST-LLGG	Video classification	Kinetics-400, SomethingSomethingV2	65.17
[34]	ViViT	Sign Language Recognition	LSA64	99.53
[35]	GestFormer	Hand Gesture Recognition	Nvidia, Briareo	97.42
[36]	AVTENet	Deepfake of Video Detection	FakeAVCeleb	84_100
[37]	TransAnomaly	Anomaly Detection	UCSD Pedestrian, CUHK Avenue	96.1
[38]	ViViT	Deepfake Video Detection	Celeb-DF version 2	92.52
[39]	ViViX and ASX	Emotion recognition	K-EmoCon	82.28
[40]	SwinAnomaly generator	Video Anomaly Detection	CUHK, UCSD	98.2
[41]	MC-ViViT	Detect Mild Cognitive	I-CONNECT	90.63
[42]	TRecViT	Video modelling	SSv2, Kinetics400	78.4
[?]	ViViT and U-Net	Frame Prediction	competition dataset	92
[44]	TC-Transformer	MultiCamera Editing TV	raw videos	outperforms...
[45]	ConvNeXt-ViViT	Predicting Vehicular Crashes	Dashcam Accident	72.80
[46]	ViViT	Change Prediction of Vehicles	PREVENTION	85
[?]	ViViT	Predicting Socio and economic Indicato	SustainBench	15 improvement
[48]	DVT	classification	Something Something-V2, Kinetics-400, EPIC-KITCHENS, Diving-48	96.2
[49]	ViViT	pain detection	UNBC-McMaster	96
[50]	ViViT	Pain Recognition	A14PAIN Challenge, BioVid Pain	79.95
[51]	ViViT	Prediction of treatment response	Collection of 624 serial CT	95
[52]	BabyNet	Birth Weight Prediction	fetal ultrasound videos	best results
[53]	2.5D models	predictive disease progression	DukeAMD	Good results
[54]	Hybrid model	Health monitoring	vital video dataset	predict accurately
[55]	DE ViViT	Blink Detection	HUST-LEBW	85.3

Continued on next page

Table 2 (continued)

Ref.	Framework	Application	Dataset	Result
[56]	ViTranZheimer	Alzheimer Diagnosis	ADNI	98.6
[57]	ViViT	Estimate Vital Signs	VVD-Large	high accuracies
[58]	Vivim	Medical Video Segmentation	BUV2022, VTUS	88.69
[59]	MAGViT-v2	Video tokenizer	ImageNet, Kinetics	96.33
[60]	LookupViT	compresses visual info	ImageNet-1K, ImageNet-21K, COCO-Captions	4 improvement
[61]	Adapter Transformer	Action classification	MM-FITStr, UCF-101	Better 0.006
[62]	LoSA	Visual Adaptation	VTAB	81.3
[63]	Eventful Transformers	Action Spotting	ImageNet VID, EPIC Kitchens 100	82.93
	TimerSformer	Pig aggression classification	Natural videos	72.9
	LSTM with Attention	invasive species detection	quagga mussel	99
[66]	ViViT-based SSL	Sound Source Localization	Dataset-1000ms	94.52
[67]	ViViT	Chemical Gas Source Localization	NBC-RAMS simulator	94.78
[68]	ViViT	Predicting Media Memorability	Memento	66.5

This classification demonstrates the flexibility of the basic models, especially ViT and ViViT, in adapting to the requirements of diverse tasks, including classification, detection, prediction, modeling, and segmentation. It also reflects the variation in performance and accuracy between the models, supporting the study's main objective of analyzing the applicability of these models to real-world smart applications.

Results showed that the ViT and ViViT models were the most widely used in video-based applications, appearing in more than 40 studies across multiple domains. In video classification, the models achieved accuracies ranging from 65% to 99%, with diverse uses such as dance, violence, and animal behavior classification. In motion detection, the models were used in applications such as fatigue detection, deepfake, and pain recognition, with accuracies ranging from 54% to 98%. In action recognition, applications included recognizing signs, gestures, and emotions, with models achieving accuracies of up to 99%. Medical prediction applications, such as fetal weight estimation, treatment response, and accident prediction, also emerged, with accuracies ranging from 66% to 95%. Some models were employed in specialized tasks such as medical video segmentation, visual compression, sound and gas source identification, and Alzheimer's diagnosis, reflecting the breadth of their uses.

Table 4 shows the distribution of research papers that used developed transformer models across different application fields in video processing. According to the analysis, action recognition is the most active field, making up 33.87% of the total research. This covers service applications like facial expression analysis for pain and emotion, as well as sign language recognition. This supports early, intelligent diagnosis and helps to raise the standard of healthcare. Action detection applications come next, accounting for 22.58% of the total, encompassing a variety of important tasks such as identifying signs of impersonation, violence, exhaustion, accidents, and more. Video classification also accounted for a significant share of research, at 11.29% of the total. This research relies on classifying videos based on audio, images, dance videos, and animal behavior monitoring. Prediction applications rank fourth with 9.67%, serving several important applications, such as predicting tire expiration, vehicle modifications in automated driving, disease progression, and responses to various treatments. Given the advanced technologies in the remaining areas such as motion understanding, segmentation, compression, and adaptation, each of them only achieved 1.61% of the total research, despite this, its percentage is 16.1%, which shows a clear expansion in research that supports smart technology in many vital sectors that serve humanity today using transformers, includ-

Table 3 Summary of the modified models and their applications in the related studies

No. of papers	Application's Field	Model used	Results
7	video classification	ViT, ViViT, Adapter, Attention-LSTM, HaViT, Adapter Transformer, VST-LLGG, TimerSformer, DVT	65.17-99.216
14	Action Detection	ViT, DA-ViViT, TransAnomaly, Regnet-ViT-Ensemble, DE ViViT, MC-ViViT, Eventful Transformers, COMEDIAN, AVTENet	54.93-98.46
21	Action Recognition	Network of Multiscale Wavelet and Pooling Transformer, ViT, Swin, SFA_ViViT, Video Swin, CTN, MTV, DirecFormer, GestFormer, ViViX, MM-ViT, RviT, MAIVAR-T, MviT, COVER, DualFormer, SwinAnomaly generator, Multiple VIVIT	72.4-99.53
6	Prediction	Transformer, BabyNet, 2.5D models, ConvNeXt ViViT	66.5-95
1	Segmentation	ViViT, Vivim	88.69
1	compression	LookupViT, LookupViViT	-
1	Adaptation Methods	ViT, LOSA	81.3
1	Action Understanding	ViViT, FTP	93.8
1	Video Modelling	ViT, TRECvIT	78.4
3	Source Localization	ViViT, ViViT-based SSL, STAR	94.5-96.3
2	Health care and Monitoring	ViViT, Hybrid model, ViTranZheimer	–,98.6
1	Video tokenizer	ViViT, MAGVIT-v2	96.33
1	Traffic Estimation	ViViT	91.5
1	MultiCamera Editing TV	ViViT, TC-Transformer	-
1	Robotic Grasping	Transformer, Grasping	85

ing the fields of healthcare, traffic, and multi-camera video processing.

The functional analysis of the modified models showed that the choice of modification type to the ViViT model is closely related to the nature of the applied task, the size of the data, and the available technical constraints. The models were classified by use area, and the most efficient model in each area was identified based on the performance reported in the studies, as shown in Figure 6. This classification supports the research's primary objective: to highlight the flexibility of the modified models and provide a practical reference for selecting the most appropriate model for each task type.

Data collected directly under different natural conditions requires specialized processing. It is affected by a lot of noise, bad weather, shooting conditions, and camera type, which in turn affects the model's overall accuracy. In contrast, standard datasets are organized and pre-processed and subject to systematic processing methods, making comparisons based on them more fair and reliable.

Table 4 Summary of the percentage of research work within a specific field of applications

Application's Field	Percentage
video classification	11.29
Action Detection	22.58
Action Recognition	33.87
Prediction	9.67
Segmentation	1.61
compression	1.61
Adaptation Methods	1.61
Action Understanding	1.61
Video Modeling	1.61
Source Localization	4.83
Health care and Monitoring	3.22
Video tokenizer	1.61
Traffic Estimation	1.61
Multi-Camera Editing TV	1.61
Robotic Grasping	1.61

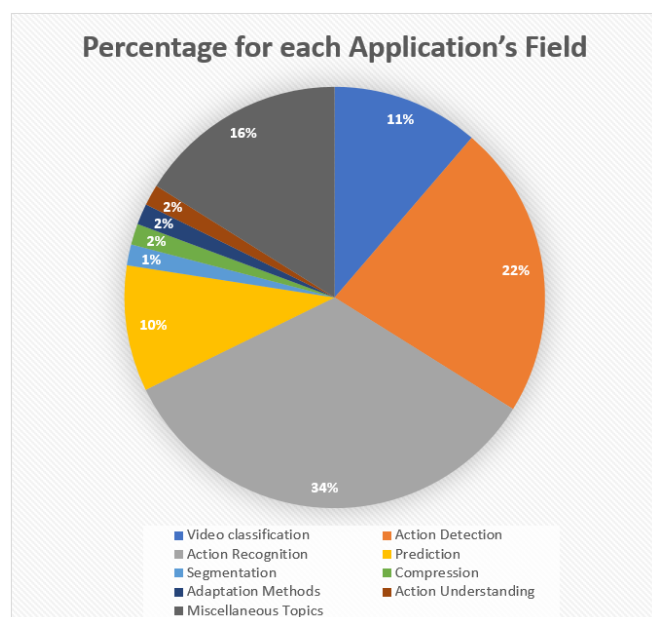


Fig. 5 Percentage of research work within specific fields

From Table 5, a clear disparity in the accuracy of models used in video classification is observed. This is primarily due to differences in the types and sizes of data, as well as the modifications applied to the base model. The DVT model in the study [26] achieved the best results in this field, with some minor limitations. The model includes two modified versions of ViT-B and MViT-B, namely D-ST-A and D-MS-A, respectively, which improve temporal and spatial information representation and reduce computational cost.

Table 5 shows performance variation, with the DVT model achieving the highest accuracy of 96.2%, followed by HaViT at 95.2%. The VST-LLGG model recorded the lowest accuracy of 65.17%, due to the nature of the data used. These results demonstrate that hybridization processes on the original model structure, such as hybrid attention or multi-level representations, play an important role in improving performance, while accounting for the type of data when choosing the appropriate model.

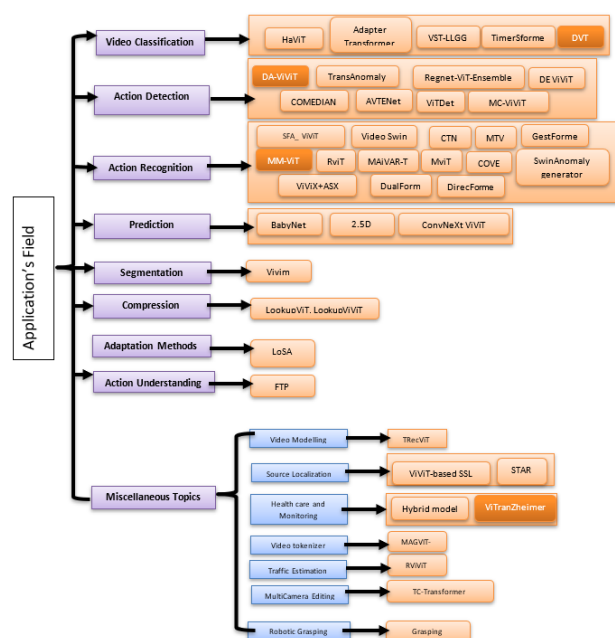


Fig. 6 Types of models related to each type of application field used

Table 6 illustrates the performance of activity detection models using modified transformers. It is noted that most models relied on special data sets prepared specifically for the study, which partly explains the high accuracy of the results. Furthermore, the differences in the applications used led to significant variation in data quality, making quantitative comparisons between models inaccurate. In this context, the DA-ViViT model [25] is among the most prominent, achieving the highest activity detection accuracy of 96.24%. This model requires analyzing two types of real-time videos: facial expressions and body posture within the model structure using an improved triple attention mechanism, which helped to improve the detection of fatigue, drowsiness, and stress while driving.

The models listed in Table 6 show that the DA-ViViT model achieved the highest accuracy in active detection, followed by TransAnomaly at 96.1%. The Regnet-ViT-Ensemble model recorded the lowest accuracy of 54.93%; the researcher stated that this was due to overfitting during the testing phase. The results show that models that combined temporal and spatial attention or used multi-branch networks were

Table 5 Video classification models

Model	Abbreviation of	Advantages	Disadvantages	Dataset	Best Accuracy
HaViT [11]	Hybrid-attention based Vision Transformer	Internal dependencies within the same video and dependencies with previous videos are also handled when following self-attention. -It can be applied to address long term problems such as natural language processing, computer vision, speech, and audio.	-The model requires retraining until the hyperparameters are stable for experiments. -The rate of improvement varied from one type of dataset to another. -It appears to achieve good results when combined with previously known video transformers.	Kinetics, Something-So something v2	95.2%
VST-LL GG [33]	Relaxing the final 2 layers of VST from the local to the global attention	It works on low-label dataset environments such as ImageNet or unsupervised. It reduces the number of tokens used, making self-attention more comprehensive and creating deep hierarchical network representations with global attention in the last two layers to increase accuracy	Using this model improves performance for the K400 dataset with high spatial characteristics, while accuracy decreases for the SSv2 dataset, which is less spatially dominant and has more temporal characteristics.	Kinetics-400, SomethingSo something V2	65.17%
DVT [48]	Deformable Video Transformer	Focuses on dynamically related patches that have information about the motion in the video, reducing the matching process between patches. - Higher accuracy and lower computational cost.	Limits temporal information	Something Something-V2, Kinetics-400, EPIC-KITCHEN S, Diving-48	96.2%
Adapter Transformer [61]	Adapter Transformer	It studies changes between frames to capture human motion information by proposing an expanded attention mechanism for spatial and temporal information. It reduces training costs and time spent due to the presence of projection layers. It uses the knowledge transfer learning method from the original transformer (ViViT). -Reduce learning parameters by 90%	Its classification accuracy is 0.022 lower than that of the original transformer (ViViT).	MM-FiTStr, UCF-101	Better 0.006%
TimeS former [64]	Time-Space Transformer	-The set of patches containing only the information is chosen to shorten training and processing time. -Better in terms of accuracy and precision. -Outperforms long-term videos, such as the research dataset.	With typical datasets, other methods outperform this model.	Natural videos	72.9%

better able to handle complex tasks, although some were affected by data quality or shooting angles.

Table 7 shows the performance of activity recognition models using modified transformers. A wide variety of architectures and techniques are observed, reflecting the development of this field and the diversity of its applications. The MM-ViT model [26] achieved the highest accuracy in recognizing movement within a video and is considered one of the advanced models that rely on integrating visual, auditory, and temporal features simultaneously, making it a multimedia model (Melomedia). The study presents four versions of the MM-ViT model,

as shown in Figure 7. In the first type of MM-ViT I model, the focus is on spatiotemporal attention, and the relationship between the key, value, and query is calculated at each layer.

The second type of model is MM-ViT II, which is based on self-attention and focuses on analyzing features along the temporal and spatial dimensions separately, linking multimodal information at the same location. The third type, the MM-ViT III model, shows a significant improvement in accuracy and low resource consumption after reducing the computational complexity to $O(N+T+S)$ by calculating relationships within non-overlapping windows.

Table 6 Action Detection models

Model	Abbreviation of	Advantages	Disadvantades	Dataset	Best Accuracy
COMEDIAN [21]	COMEDIAN	The model includes unsupervised self-learning and knowledge extraction to detect events using temporal and spatial transformers.	The model requires performance improvement.	SoccerNet -v2	73.1%
Regnet-VIT-Ensemble [22]	Regnet-VIT-Ensemble	The model replaces the VIVIT transformer's feature extraction layer with a RegNet convolutional network to reduce the transformer's large size.	-The model is considered late compared to the models submitted recently. -Low accuracy.	AU task	54.93%
DA-VIVIT [25]	Distributed AttentionVIVIT	-Ensuring high accuracy and significantly reducing computational costs.	-The system uses motion capture cameras on the right side only, which blocks information on the left side for the driver.	DMD	96.24%
TransAnomaly [37]	Transformer Anomaly	Captures more comprehensive time information to predict future frames.	The results are affected by the transformer's depth in the spatial encoder. As for the temporal encoder, a deeper transformer does not improve these results.	UCSD Pedestrian, CUHK Avenue	96.1%
AVTENet [36]	Audio Visual Transformer based on Ensemble Network	The model shows strong performance on a specific dataset, such as FakeAVCeleb, for detecting fake videos using visual and audio information.	It requires multiple techniques to accurately detect fakes on other datasets	FakeAVCeleb	84%-100%
MC-VIVIT [41]	Multibranch Classifier VIVIT	The model handles an easy, hard, and medium-imbalanced dataset to detect cognitive impairment using the HP Loss function while reducing computational complexity.	The model's predictions are strongly influenced by the quality of the videos used, and may yield incorrect results or low prediction scores.	I-CONNECT	90.63%
DE VIVIT [55]	Dual Embedding VIVIT	The model demonstrates efficiency in capturing eye blink movement from video files.	The model's results are affected by the angle at which the camera captures the blinking movement.	HUST-LEB W	85.3%
Eventful Transformers [63]	Eventful Transformers	The model reduces the computational effort of the original transformer by exploiting temporal structure in the input data.	The improvement the model offers comes at the cost of decreased accuracy.	ImageNet VID, EPIC Kitchens 100	82.93%

Finally, MM-ViT IV uses a factorized local-space and time-cross model in which attention relationships are calculated within non-overlapping local windows.

The results show that the MM-ViT model outperformed the motion recognition model with an accuracy of 98.9%, followed by the MTV model with 98.3%, and then SwinAnomaly with 98.2%. In

contrast, the CTN model recorded the lowest accuracy at 78.2%. The results show that models based on multimodal learning or hierarchical structures were better able to handle complex scenes, whereas others were affected by factors such as data size, the number of scenes, or light sensitivity.

Table 7 Action Recognition models

Model	Abbreviation	Advantages	Disadvantages	Dataset	Best Acc.
COVER [12]	Cotraining Transformer using Videos and Images for Action Recognition	It follows a joint training approach on images and videos to extract appearance information only, while incorporating motion features to improve learning.	Relying on images to train the model may lead to poor performance on video clips due to the lack of spatial repetition in images.	Kinetics, SSv2, MiT, Standard COVER	87.9%

Continued on next page

Table 7 (continued)

Model	Abbreviation	Advantages	Disadvantages	Dataset	Best Acc.
SFA_ViViT [14]	Spatial Frozen and Adapter Initialized	- Designed to reduce computational operations and memory by freezing the spatial transform encoder. - Supports research using limited resources.	It may need some improvements to raise accuracy.	Kinetics, EPIC-Kitchens, SSv2, MiT	84.25%
RviT [18]	Recurrent Vision in Transformer	Addresses the problem of processing attention within a single video via an attention gate that calculates dependencies with previous frames.	Requires intensive calculations to achieve good results.	Kinetics-400, Jester, SSv2, Charades	92.31%
Video Swin [19]	Video Swin	Attempts to understand the video and recognize motion with limited available data.	Relies on limited data; has not been tested on live data specific to the model. Relies on base model's original weights.	UCF-101	95%
Direc Former [20]	Directed Attention in the Transformer	Relies on correct temporal ordering of events using directed attention to temporal and spatial information.	The number of parameters is higher.	Jester, Kinetics-400, SSv2	94.86%
MAiVAR-T [23]	Multimodal Audio, Image, and Video Action Recognizer	Multimedia features are extracted into the image domain to effectively enhance motion recognition.	Further testing and evaluation are needed.	UCF101	91.2%
MM-ViT [26]	MultiModal Video Transformer	- Complexity is linear $O(N + T + S)$. - Analyzes temporal, spatial, and dimensional features; works on compressed videos.	- Faces limited scope within local windows (CNN solution adds computation). - Attention order significantly affects performance.	UCF-101, Kinetics-600, SSv2	98.9%
CTN [27]	Convolutional Transformer network	Combines CNN and transformer advantages by replacing linear mapping MHA with a CNN layer.	Slight increase in calculations.	Epic-Kitchens, Diving 48	78.2%
Dual Former [29]	Local-Global Transformer	Hierarchical approach: starts with local dependencies, then calculates global dependencies at higher levels with reduced cost.	Relies on fully pre-trained models.	Kinetics-400/600	85.2%
MTV [31]	Multiview Transformers in Video	Handles video scenes where each scene requires a separate encoder, achieving better performance.	Number of tokens processed increases depending on the number of scenes.	Kinetics, MiT, EK100, SSv2	98.3%
MviT [32]	Multiscale Vision in Transformers	Hierarchical learning uses a progression from simple to complex information to extract multiple spatial features.	5-10% more expensive than other models in terms of calculation and parameters.	Kinetics, SSv2, Charades, AVA	84.8%
Gest Former [35]	Gesture Transformer	Exploits fine-grained, invariant features of the wavelet transform and multiscale features via pooling.	Needs to be tested on other types of datasets.	Nvidia, Briareo	97.42%
ViViX +ASX [39]	ViViT attention mechanism	Recognizes emotions in video chats using two transformers (vision and audio). Offers linear complexity.	Lacks a way to integrate different multimedia information. Limited dataset.	K-EmoCon	82.28%
Swin Anomaly [40]	SwinAnomaly generator	Very effective for detecting anomalies in real-time video flexibly.	Fixed parameter settings for inference. Increased sensitivity to lighting conditions.	CUHK, UCSD	98.2%

Tables 8 show that the number of modified transformer models in specialized scientific fields such as prediction, segmentation, compression, and adaptation remains limited compared to those in classification and detection. Despite the importance of these fields in supporting intelligent applications, development in them remains weak in terms of the number of models and the accuracy of the results achieved. This is likely due to limited data or to

researchers' tendency to design general models that facilitate implementation without tailoring them to a specific purpose or data type.

In the field of prediction models (Table 8), the 2.5D model [53] shows promising performance by combining 2D and 3D techniques to reduce memory consumption. While other models, such as the ConvNeXt ViViT [45] and BabyNet [52] models, showed unstable results due to the narrow data range.

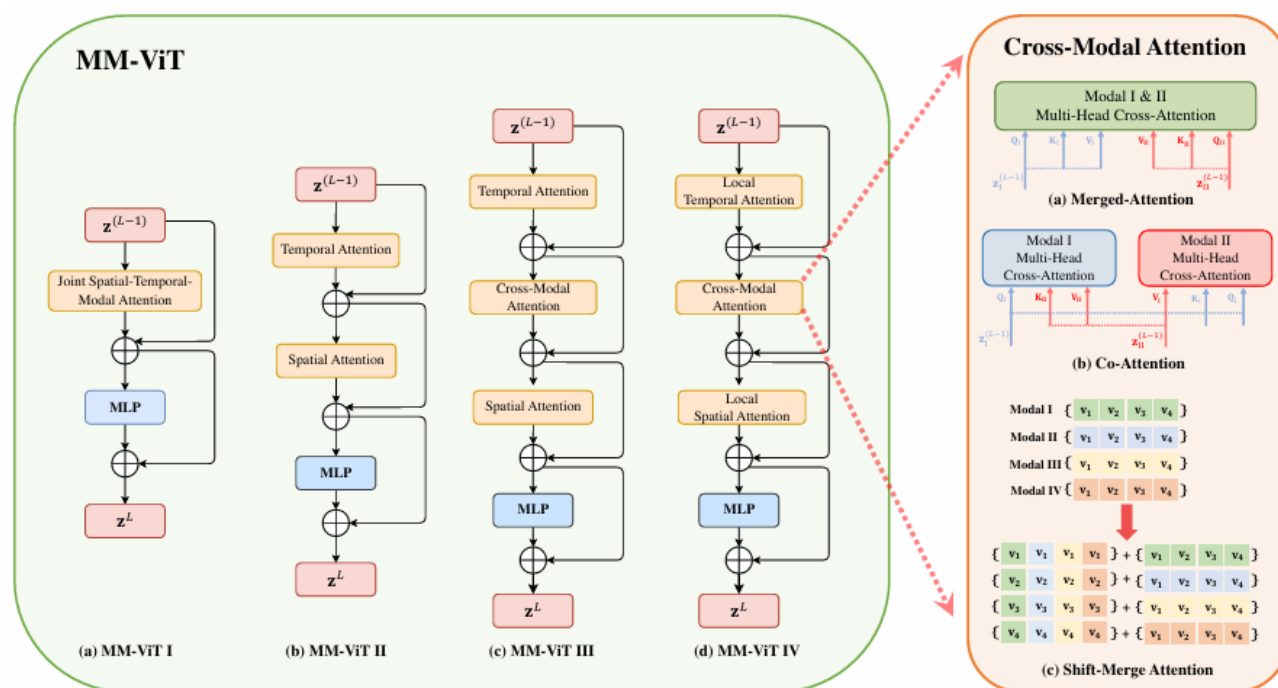


Fig. 7 MM-ViT model [26]

In segmentation (Table 9) and compression (Table 10), models such as Vivim [58] and LookupViViT [60] have demonstrated the ability to reduce computational complexity, but they still require further evaluation.

In the adaptive methods domain (Table 11), the LoSA [62] model is lightweight, reducing training time, but it still suffers from slow inference due to its expanded structure.

Table 12 presents the model of motion understanding, the FTP model [24]; it achieved an accuracy of 93.8% on the Kinetics-400 and Something-

Something V2 datasets, reflecting its high efficiency in understanding complex activities. However, the model's enhanced architecture has a high computational cost at inference, which might limit its applicability in applications with limited resources.

Table 13 presents a variety of models spanning diverse application areas, including motion estimation, source identification, audio localization, biometric prediction, and medical classification. Despite the limited number of models and varying degrees of completeness, the analysis highlights the importance of some advanced models within this category.

Table 8 Prediction models

Model	Abbreviation of	Advantages	Disadvantages	Dataset	Best Accuracy
ConvNeXt VIVIT [45]	ConvNeXt ViViT	-The model relies on selective layer freezing, which facilitates training stability, avoids relativism, optimizes the model, and reduces parameters. -It relies on pooling techniques to reduce processing volume and memory footprint.	-The limited number of video frames and the use of the last clip of the video are used to prevent class imbalance. -Training is limited and results are unreliable.	Dashcam Accident	72.80%
BabyNet [52]	Baby Residual Transformer Net	Accurate in estimating the age of the fetus before birth	The model was tested on a small dataset (225 videos), so its properties cannot be accurately judged.	fetal ultrasound videos	best results
2.5D [53]	2D and 3D techniques	It addresses the problem of training 3D models using 2D models via 2.5D, which combines 3D and 2D techniques, thereby reducing memory requirements.	The disadvantages are not mentioned, but it is clear that increasing the implementation stages may significantly slow down the work.	DukeAMD	Good results

Table 9 Segmentation models

Model	Abbreviation of	Advantages	Disadvantages	Dataset	Best Accuracy
Vivim [58]	Video Vision Mamba	-Professionally compresses spatiotemporal information into specific sequences using Mamba to reduce costs. -The expansion is linear, not quadratic.	Limited Datasets.	VTUS, BUV2022	88.69%

Table 10 Compression models

Model	Abbreviation of	Advantages	Disadvantages	Dataset	Best Accuracy
LookupViT, LookupViViT [60]	LookupViT, LookupViViT	Supports systems with a small dataset. The model compresses data from highly specific tokens into a fixed set for precise processing, achieving a balance between performance and computational complexity.	Retraining was done on a subset of the data.	ImageNet-1K, ImageNet-21K, COCO-Captions, ImageNet-C, R, A, O	4% improvement

Table 11 Adaptation Methods models

Model	Abbreviation of	Advantages	Disadvantages	Dataset	Best Accuracy
LoSA [62]	Low-rank Side-Adaptation	Propose a lightweight network that runs in parallel rather than backpropagation to reduce training time, the number of parameters, and the memory required.	Inference is still slow because its structure is larger.	VTAB	81.3%

Table 12 Action Understanding models

Model	Abbreviation of	Advantages	Disadvantades	Dataset	Best Accuracy
FTP [24]	Four-tiered Prompts	The model leverages the strengths of VLTs in retrieving temporal and spatial file information and obtaining robust representations of them, and its high generalization capabilities from VLMs (visual language model).	It requires a computational cost for the inference process due to the increased size of the structure.	Kinetics-400, Something-Something V2	93.8%

Table 13 Miscellaneous Topics models (Cells surrounded by dark color mean that it belongs to the same application field)

Model	Abbreviation of	Advantages	Disadvantades	Dataset	Best Accuracy
Grasping [4]	Grasping	A system that supports robots to grasp solid objects using tactile and visual information.	It needs to be developed in several aspects, the most important of which is training on other types of fruits and vegetables and hybridizing it with methods that depend on clear models, such as those that depend on physical contact.	GelSight	85.0%
RViViT [18]	Revised Video Vision Transformer	The model provides an efficient traffic estimate.	It requires massive data testing to make accurate predictions.	naturalistic vehicle trajectories	91.5%
TC-Transformer [44]	Temporal Contextual Transformer	The model works with limited datasets, predicting the most important shot and suggesting a good angle for shooting	It lacks footage editing, fade-ins, transition effects, and other features.	raw videos	outperforms existing methods
TRecViT [42]	Recurrent Video-Transformer	The model analyzes temporal and spatial factors in the video with high efficiency, using 3 times fewer parameters and lower computational cost than the original model.	Further experiments on the model on more tasks are needed.	SSv2, Kinetics400	78.4%
ViViT-based SSL [66]	ViViT-based sound source localization	Outperforms ResNet-based SSL for audio source detection.	Limited experiences.	Dataset-1000 ms	94.52%
STAR [13]	Spatio and Temporal Action-Recognition	It does not require pre- or post-processing and determines the spatial or temporal location of the action in the video.	—————	AVA, AVA-Kinetics, UCF101-24, JHMDB51-21	96.3%
Hybrid model [54]	(CNN), (convLSTM), (ViViT)	Extracts temporal and spatial features to predict facial biometrics using a hybrid model.	-Requires a larger dataset to achieve high accuracy and reliability. -Requires higher GPUs to reduce access and inference latency.	A vital video dataset	predict accurately
ViTranZheimer [56]	vision transformer Zheimer	The model outperforms both CNN-BiLSTM and ViT-BiLSTM models for early detection of Alzheimer's and relies on CNNs with recurrent layers, such as LSTMs, to capture dependencies between slices rather than self-attention that captures dependencies within the image.	-There were a few false negatives among the total results. -Limited data used.	ADNI	98.6%
MAGViT-v2 [59]	video tokenizer	Tokenizer is an innovative tokenizer for encoding images and videos that utilizes quantization.	Diffusion models give better performance.	ImageNet, Kinetics	96.33%

The ViTranZheimer model [56] is one of the models in this field. It achieved 98.6% accuracy, showing the importance of advancing healthcare by detecting diseases early. Based on modifications to the ViViT architecture and 3D MRI image processing.

Other models, such as Grasping [4], RViViT [18], SSL-based ViViT [66], and MAGViT-v2 [59] have shown good performance in specialized tasks such as robotic grasping, traffic estimation, audio source identification, and video coding, but they still need

further development, both in data diversity and architectural optimization. On the other hand, models such as Hybrid [54] and TRecViT [42] aim to reduce computational complexity and improve efficiency, but they require additional experiments to evaluate their performance at scale. The STAR model [13] achieved 96.3% accuracy in identifying the spatiotemporal locations of motion without requiring pre- or post-processing.

6 CONCLUSION

This study presents a comparative and systematic analysis of transformer models used in video processing, highlighting their strengths, weaknesses, and structural diversity across classification, detection, prediction, segmentation, compression, adaptation, and motion understanding.

Most researchers have recently tried to balance accuracy and training costs by hybridizing different methods. This study compares about 68 transformer models for processing video data. The study shows that, despite their high accuracy, the various transformer models still suffer from high computational complexity in many cases and poor generalizability due to limited data, limiting their applications in several fields. We recommend focusing on models that rely on linear attention mechanisms, and exploring computationally uncomplicated and resource-intensive architectures that achieve a practical balance between accuracy, resource efficiency, and quality of results, especially in applications that require fast inference or prediction, or operate in environments with limited capabilities or in special conditions, such as satellites, disaster and earthquake detection systems, or sensitive medical systems.

Author contributions

All authors contributed meaningfully to the preparation of this research work. The lead author, A.N.Y formulated the general idea of the study, defined its scope, and compiled and analyzed the models used in video processing based on transformer architectures. She also undertook the literature review, designed comparative tables, and analyzed the advantages and

disadvantages of each model. F.M.R participated in reviewing the scientific content, providing technical comments on training and testing mechanisms, and contributing to the discussion of challenges related to computational complexity and hybrid modeling. She also contributed to reviewing the final results and formulating conclusions and future recommendations. All authors approved the final version of the research and confirm their joint responsibility for the accuracy of the information contained therein.

Funding source

No funding was received.

Data availability

N/A

DECLARATIONS

Conflict of interest

The authors declare no competing interests

Consent to publish

N/A

Ethical approval

N/A

REFERENCES

- [1] Selva J, Johansen AS, Escalera S, Nasrollahi K, Moeslund TB, Clapés A. Video Transformers: A Survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*. 2023;45(11):12922-43. [10.1109/TPAMI.2023.3243465](https://doi.org/10.1109/TPAMI.2023.3243465)
- [2] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al. An Image Is Worth 16x16 Words: Transformers for Image Recognition at Scale. In: *International Conference on Learning Representations*; 2021. Available from: <https://arxiv.org/abs/2010.11929>. [10.48550/arXiv.2010.11929](https://arxiv.org/abs/2010.11929)
- [3] Mashru D, Vora K. Comparative Analysis of CNN, RNN, LSTM, and Transformer Architec-

- tures in Deep Learning. Educational Administration: Theory and Practice. 2023;29(4):5439-43. [10.53555/kuey.v29i4.10364](https://doi.org/10.53555/kuey.v29i4.10364)
- [4] Han Y, Yu K, Batra R, Boyd N, Mehta C, Zhao T, et al. Learning Generalizable Vision-Tactile Robotic Grasping Strategy for Deformable Objects via Transformer. IEEE/ASME Transactions on Mechatronics. 2025;30(1):554-66. [10.1109/TMECH.2024.3400789](https://doi.org/10.1109/TMECH.2024.3400789)
- [5] Arnab A, Dehghani M, Heigold G, Sun C, Lučić M, Schmid C. ViViT: A Video Vision Transformer. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE; 2021. p. 6836-46. [10.1109/ICCV48922.2021.00676](https://doi.org/10.1109/ICCV48922.2021.00676)
- [6] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al. Attention Is All You Need. In: Advances in Neural Information Processing Systems. vol. 30; 2017. p. 5998-6008. [10.48550/arXiv.1706.03762](https://arxiv.org/abs/1706.03762)
- [7] Peruzzo E, Sangineto E, Liu Y, De Nadai M, Bi W, Lepri B, et al. Spatial Entropy as an Inductive Bias for Vision Transformers. Machine Learning. 2024;113:6945-75. [10.1007/s10994-024-06570-7](https://doi.org/10.1007/s10994-024-06570-7)
- [8] Le DPC, Wang D, Le VT. A Comprehensive Survey of Recent Transformers in Image, Video and Diffusion Models. Computers, Materials & Continua. 2024;80(1):37-60. [10.32604/cmc.2024.050790](https://doi.org/10.32604/cmc.2024.050790)
- [9] Singh S, Dewangan S, Krishna GS, Tyagi V, Reddy S, Medi PR. Video Vision Transformers for Violence Detection; 2022. Preprint. Available from: <https://arxiv.org/abs/2209.03561>. [10.48550/arXiv.2209.03561](https://arxiv.org/abs/2209.03561)
- [10] Mulyanto E, Yuniarno EM, Putra OV, Hafidz I, Priyadi A, Purnomo MH. Improvement of Tradition Dance Classification Process Using Video Vision Transformer Based on Tubelet Embedding. International Journal of Intelligent Engineering and Systems. 2024;17(4):530-45. [10.22266/IJIES2024.0831.41](https://doi.org/10.22266/IJIES2024.0831.41)
- [11] Li L, Zhuang L, Gao S, Wang S. HaViT: Hybrid-Attention Based Vision Transformer for Video Classification. In: Wang L, Gall J, Chin TJ, Sato I, Chellappa R, editors. Computer Vision – ACCV 2022. vol. 13844 of Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2023. p. 502-17. [10.1007/978-3-031-26316-3_30](https://doi.org/10.1007/978-3-031-26316-3_30)
- [12] Zhang B, Yu J, Fifty C, Han W, Dai AM, Pang R, et al.. Co-Training Transformer with Videos and Images Improves Action Recognition; 2021. Preprint. Available from: <https://arxiv.org/abs/2112.07175>. [10.48550/arXiv.2112.07175](https://arxiv.org/abs/2112.07175)
- [13] Gritsenko AA, Xiong X, Djolonga J, Dehghani M, Sun C, Lučić M, et al. End-to-End Spatio-Temporal Action Localisation with Video Transformers. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2024. p. 18373-83. [10.1109/CVPR52733.2024.01739](https://doi.org/10.1109/CVPR52733.2024.01739)
- [14] Gowda SN, Arnab A, Huang J. Optimizing Factorized Encoder Models: Time and Memory Reduction for Scalable and Efficient Action Recognition. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer Vision – ECCV 2024. vol. 15068 of Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2025. p. 457-74. [10.1007/978-3-031-72684-2_6](https://doi.org/10.1007/978-3-031-72684-2_6)
- [15] Huang Z, Qing Z, Wang X, Feng Y, Zhang S, Jiang J, et al.. Towards Training Stronger Video Vision Transformers for EPIC-KITCHENS-100 Action Recognition; 2021. Preprint. Available from: <https://arxiv.org/abs/2106.05058>. [10.48550/arXiv.2106.05058](https://arxiv.org/abs/2106.05058)
- [16] Koot R, Hennerbichler M, Lu H. Evaluating Transformers for Lightweight Action Recognition; 2021. Preprint. Available from: <https://arxiv.org/abs/2111.09641>. [10.48550/arXiv.2111.09641](https://arxiv.org/abs/2111.09641)
- [17] Yang J, Dong X, Liu L, Zhang C, Shen J, Yu D. Recurring the Transformer for Video

- Action Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2022. p. 14043-53. [10.1109/CVPR52688.2022.01367](https://doi.org/10.1109/CVPR52688.2022.01367)
- [18] Li D, Lasenby J. A Revised Video Vision Transformer for Traffic Estimation with Fleet Trajectories. IEEE Sensors Journal. 2022;22(17):17103-12. [10.1109/JSEN.2022.3193663](https://doi.org/10.1109/JSEN.2022.3193663)
- [19] Kitškerkin J. In Search of Video Transformer Models for Action Recognition on Limited Data. Tilburg University; 2022. School of Humanities and Digital Sciences, Department of Cognitive Science Artificial Intelligence
- [20] Truong TD, Bui QH, Duong CN, Seo HS, Phung SL, Li X, et al. DirecFormer: A Directed Attention in Transformer Approach to Robust Action Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2022. p. 19998-20008. [10.1109/CVPR52688.2022.01940](https://doi.org/10.1109/CVPR52688.2022.01940)
- [21] Denize J, Liashuha M, Rabarisoa J, Orcesi A, Hérault R. COMEDIAN: Self-Supervised Learning and Knowledge Distillation for Action Spotting Using Transformers. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision Workshops. IEEE; 2024. p. 518-28. [10.1109/WACVW60836.2024.00060](https://doi.org/10.1109/WACVW60836.2024.00060)
- [22] Vu NT, Huynh VT, Nguyen TN, Kim SH. Ensemble Spatial and Temporal Vision Transformer for Action Units Detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. IEEE; 2023. p. 5770-6. [10.1109/CVPRW59228.2023.00612](https://doi.org/10.1109/CVPRW59228.2023.00612)
- [23] Shaikh MB, Chai D, Shamsul Islam SM, Akhtar N. MAiVAR-T: Multimodal Audio-Image and Video Action Recognizer Using Transformers. In: 2023 11th European Workshop on Visual Information Processing. IEEE; 2023. p. 1-6. [10.1109/EUVIP58404.2023.10323051](https://doi.org/10.1109/EUVIP58404.2023.10323051)
- [24] Lu H, Jian H, Poppe R, Salah AA. Enhancing Video Transformers for Action Understanding with VLM-Aided Training; 2024. Preprint. Available from: <https://arxiv.org/abs/2403.16128>. [10.48550/arXiv.2403.16128](https://doi.org/10.48550/arXiv.2403.16128)
- [25] Deng F, Yang C, Guo H, Wang Y, Xu L. DA-ViT: Fatigue Detection Framework Using Joint and Facial Keypoint Features with Dynamic Distributed Attention Video Vision Transformer. Research Square; 2024. Preprint, version 1. Available from: <https://doi.org/10.21203/rs.3.rs-4546491/v1>. [10.21203/rs.3.rs-4546491/v1](https://doi.org/10.21203/rs.3.rs-4546491/v1)
- [26] Chen J, Ho CM. MM-ViT: Multi-Modal Video Transformer for Compressed Video Action Recognition. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. IEEE; 2022. p. 786-97. [10.1109/WACV51458.2022.00086](https://doi.org/10.1109/WACV51458.2022.00086)
- [27] Ma Y, Wang R, Zong M, Ji W, Wang Y, Ye B. Convolutional Transformer Network for Fine-Grained Action Recognition. Neurocomputing. 2024;569:127027. [10.1016/j.neucom.2023.127027](https://doi.org/10.1016/j.neucom.2023.127027)
- [28] Kuznietsova N, Smirnov S. Application of Vision Transformers and 3D Convolutional Neural Networks for Sign Language Cluster Recognition. In: Computer Modeling and Intelligent Systems. vol. 3392 of CEUR Workshop Proceedings; 2023. p. 151-63. [10.32782/cmisi/3392-13](https://doi.org/10.32782/cmisi/3392-13)
- [29] Liang Y, Zhou P, Zimmermann R, Yan S. DualFormer: Local-Global Stratified Transformer for Efficient Video Recognition. In: Avidan S, Brostow G, Cissé M, Farinella GM, Hassner T, editors. Computer Vision – ECCV 2022. vol. 13694 of Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2022. p. 577-95. [10.1007/978-3-031-19830-4_33](https://doi.org/10.1007/978-3-031-19830-4_33)
- [30] Liu Z, Ning J, Cao Y, Wei Y, Zhang Z, Lin S, et al. Video Swin Transformer. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2022. p. 3192-201. [10.1109/CVPR52688.2022.00320](https://doi.org/10.1109/CVPR52688.2022.00320)

- [31] Yan S, Xiong X, Arnab A, Lu Z, Zhang M, Sun C, et al. Multiview Transformers for Video Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2022. p. 3323-33. [10.1109/CVPR52688.2022.00333](https://doi.org/10.1109/CVPR52688.2022.00333)
- [32] Fan H, Xiong B, Mangalam K, Li Y, Yan Z, Malik J, et al. Multiscale Vision Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE; 2021. p. 6804-15. [10.1109/ICCV48922.2021.00675](https://doi.org/10.1109/ICCV48922.2021.00675)
- [33] Rahman F, Mubarek Ö, Kira Z. On the Surprising Effectiveness of Transformers in Low-Labeled Video Recognition; 2022. Preprint. Available from: <https://arxiv.org/abs/2209.07474>. [10.48550/arXiv.2209.07474](https://doi.org/10.48550/arXiv.2209.07474)
- [34] Marais M, Brown D, Connan J, Bobby A. Spatiotemporal Convolutions and Video Vision Transformers for Signer-Independent Sign Language Recognition. In: 2023 International Conference on Artificial Intelligence, Big Data, Computing and Data Communication Systems. IEEE; 2023. p. 1-6. [10.1109/icABCD59051.2023.10220534](https://doi.org/10.1109/icABCD59051.2023.10220534)
- [35] Garg M, Ghosh D, Pradhan PM. GestFormer: Multiscale Wavelet Pooling Transformer Network for Dynamic Hand Gesture Recognition. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. IEEE; 2024. p. 2473-83. [10.1109/CVPRW63382.2024.00254](https://doi.org/10.1109/CVPRW63382.2024.00254)
- [36] Hashmi A, Shahzad SA, Lin CW, Tsao Y, Wang HM. AVTENet: A Human-Cognition-Inspired Audio-Visual Transformer-Based Ensemble Network for Video Deepfake Detection. IEEE Transactions on Cognitive and Developmental Systems. 2025;17(6):1360-76. [10.1109/TCDS.2025.3554477](https://doi.org/10.1109/TCDS.2025.3554477)
- [37] Yuan H, Cai Z, Zhou H, Wang Y, Chen X. TransAnomaly: Video Anomaly Detection Using Video Vision Transformer. IEEE Access. 2021;9:123977-86. [10.1109/ACCESS.2021.3109102](https://doi.org/10.1109/ACCESS.2021.3109102)
- [38] Ramadhani KN, Munir R, Utama NS. Improving Video Vision Transformer for Deepfake Video Detection Using Facial Landmark, Depthwise Separable Convolution and Self Attention. IEEE Access. 2024;12:8932-9. [10.1109/ACCESS.2024.3352890](https://doi.org/10.1109/ACCESS.2024.3352890)
- [39] Poh JQ, See J, El Gayar N, Wong LK. Are You Paying Attention? Multimodal Linear Attention Transformers for Affect Prediction in Video Conversations. In: Proceedings of the 2nd International Workshop on Multimodal and Responsible Affective Computing. Association for Computing Machinery; 2024. p. 15-23. [10.1145/3689092.3689409](https://doi.org/10.1145/3689092.3689409)
- [40] Bajgoti A, Gupta R, Balaji P, Dwivedi R, Siwach M, Gupta D. SwinAnomaly: Real-Time Video Anomaly Detection Using Video Swin Transformer and SORT. IEEE Access. 2023;11:111093-105. [10.1109/ACCESS.2023.3321801](https://doi.org/10.1109/ACCESS.2023.3321801)
- [41] Sun J, Dodge HH, Mahoor MH. MC-ViViT: Multi-Branch Classifier-ViViT to Detect Mild Cognitive Impairment in Older Adults Using Facial Videos. Expert Systems with Applications. 2024;238:121929. [10.1016/j.eswa.2023.121929](https://doi.org/10.1016/j.eswa.2023.121929)
- [42] Pătrăucean V, He XO, Heyward J, Zhang C, Sajjadi MSM, Muraru GC, et al. TRecViT: A Recurrent Video Transformer. Transactions on Machine Learning Research. 2026. Version of record; Google DeepMind publication page dated 9 January 2026. Available from: <https://openreview.net/forum?id=Mmi46Ytb1H>
- [43] Zhou Z, Zhu Y, Tong L. Video Frame Prediction with ViViT and U-Net. ResearchGate; 2024. [10.13140/RG.2.2.36713.10087](https://doi.org/10.13140/RG.2.2.36713.10087)
- [44] Rao A, Jiang X, Wang S, Guo Y, Liu Z, Dai B, et al.. Temporal and Contextual Transformer for Multi-Camera Editing of TV Shows; 2022. Preprint. Available from: <https://arxiv.org/abs/2210.08737>. [10.48550/arXiv.2210.08737](https://doi.org/10.48550/arXiv.2210.08737)
- [45] Naumenka J, Giuffrida V. Evaluating the ConvNeXt-ViViT Hybrid Architecture in Predicting Vehicular Crashes from Video Data.

- Open Science Framework; 2024. Repository preprint. Available from: <https://osf.io/bnkw5/10.17605/OSF.IO/BNKW5>
- [46] Mohamed MAR, Srinath NS, Thomas J, Monica KM. Lane Change Prediction of Surrounding Vehicles Using Video Vision Transformers. *International Journal of Computing and Digital Systems*. 2025;17(1):1-9. [10.12785/ijcds/1571024445](https://doi.org/10.12785/ijcds/1571024445)
- [47] Jarry R, Chaumont M, Berti-Equille L, Subsol G. Predicting Socio-Economic Indicator Variations with Satellite Image Time Series and Transformer. In: *Machine Vision for Earth Observation and Environment Monitoring*. Glasgow, United Kingdom; 2024. MVEO 2024 workshop paper. Available from: <https://hal-lirmm.ccsd.cnrs.fr/lirmm-04895134v2>
- [48] Wang J, Torresani L. Deformable Video Transformer. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE; 2022. p. 14033-42. [10.1109/CVPR52688.2022.01366](https://doi.org/10.1109/CVPR52688.2022.01366)
- [49] Fiorentini G, Önal Ertuğrul I, Salah AA. Fully-Attentive and Interpretable: Vision and Video Vision Transformers for Pain Detection. In: *NeurIPS 2022 Workshop on Vision Transformers: Theory and Applications*; 2022. p. 1-12. Available from: <https://arxiv.org/abs/2210.15769>. [10.48550/arXiv.2210.15769](https://arxiv.org/abs/2210.15769)
- [50] Bargshady G, Joseph C, Hirachan N, Goecke R, Rojas RF. Acute Pain Recognition from Facial Expression Videos Using Vision Transformers. In: *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*. IEEE; 2024. p. 1-4. [10.1109/EMBC53108.2024.1078161](https://doi.org/10.1109/EMBC53108.2024.1078161)
- [51] Wu Y, Guan R, Liang X, Zhang W, Jiang Y, Cui Y, et al.. Prediction of Immune Checkpoint Inhibitors Treatment Response of Non-Small Cell Lung Cancer Patients from Serial Computed Tomography Scans Based on Global Self-Attention Mechanism. Preprints.org; 2024. Preprint, version 1. Available from: <https://www.preprints.org/manuscript/202407.1061/v1>. [10.20944/preprints202407.1061.v1](https://doi.org/10.20944/preprints202407.1061.v1)
- [52] Plotka S, Grzeszczyk MK, Brawura-Biskupski Samaha R, Gutaj P, Lipa M, Trzciński T. BabyNet: Residual Transformer Module for Birth Weight Prediction on Fetal Ultrasound Video. In: Wang L, Dou Q, Fletcher PT, Speidel S, Li S, editors. *Medical Image Computing and Computer Assisted Intervention – MICCAI 2022*. Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2022. p. 350-9. [10.1007/978-3-031-16440-8_34](https://doi.org/10.1007/978-3-031-16440-8_34)
- [53] Emre T, Oghbaie M, Chakravarty A, Rivail A, Riedl S, Mai J, et al. Pretrained Deep 2.5D Models for Efficient Predictive Modeling from Retinal OCT: A PINNACLE Study Report. In: Antony B, Chen H, Fang H, Fu H, Lee CS, Zheng Y, editors. *Ophthalmic Medical Image Analysis*. vol. 14096 of Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2023. p. 132-41. [10.1007/978-3-031-44013-7_4](https://doi.org/10.1007/978-3-031-44013-7_4)
- [54] Zheng Y. Heart Rate and Oxygen Level Estimation from Facial Videos Using a Hybrid Deep Learning Model. In: Agaian SS, DelMarco SP, Asari VK, editors. *Multimodal Image Exploitation and Learning 2024*. vol. 13033 of Proceedings of SPIE. National Harbor, Maryland, USA: SPIE; 2024. Eight-page conference paper. [10.1117/12.3013956](https://doi.org/10.1117/12.3013956)
- [55] Hong J, Shin J, Choi J, Ko M. Robust Eye Blink Detection Using Dual Embedding Video Vision Transformer. In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. IEEE; 2024. p. 6362-72. [10.1109/WACV57701.2024.00625](https://doi.org/10.1109/WACV57701.2024.00625)
- [56] Akan T, Alp S, Bhuiyan MS, Disbrow EA, Conrad SA, Vanchiere JA, et al.. Leveraging Video Vision Transformer for Alzheimer's Disease Diagnosis from 3D Brain MRI; 2025. Preprint. Available from: <https://arxiv.org/abs/2501.15733>. [10.48550/arXiv.2501.15733](https://arxiv.org/abs/2501.15733)

- [57] Zheng Y. Hybrid Neural Network Models to Estimate Vital Signs from Facial Videos. *BioMed-Informatics*. 2025;5(1):6. [10.3390/biomedinformatics5010006](https://doi.org/10.3390/biomedinformatics5010006)
- [58] Yang Y, Xing Z, Yu L, Huang C, Fu H, Zhu L. Vivim: A Video Vision Mamba for Medical Video Segmentation; 2024. Preprint. Available from: <https://arxiv.org/abs/2401.14168>. [10.48550/arXiv.2401.14168](https://arxiv.org/abs/2401.14168)
- [59] Yu L, Lezama J, Gundavarapu NB, Versari L, Sohn K, Minnen D, et al. Language Model Beats Diffusion: Tokenizer Is Key to Visual Generation. In: International Conference on Learning Representations; 2024. Available from: https://proceedings.iclr.cc/paper_files/paper/2024/hash/036912a83bdbb1fd792baf6532f102d8-Abstract-Conference.html
- [60] Koner R, Jain G, Jain P, Tresp V, Paul S. LookupViT: Compressing Visual Information to a Limited Number of Tokens. In: Leonardis A, Ricci E, Roth S, Russakovsky O, Sattler T, Varol G, editors. Computer Vision – ECCV 2024. vol. 15144 of Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2025. p. 322-37. [10.1007/978-3-031-73016-0_9](https://doi.org/10.1007/978-3-031-73016-0_9)
- [61] Zhuang Y. Human Action Classification Task Based on Spatial and Temporal Transformer in Videos. Waseda University; 2022. Available from: <http://hdl.handle.net/2065/00092760>
- [62] Mercea OB, Gritsenko A, Schmid C, Arnab A. Time-, Memory- and Parameter-Efficient Visual Adaptation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2024. p. 5536-45. [10.1109/CVPR52733.2024.00529](https://doi.org/10.1109/CVPR52733.2024.00529)
- [63] Dutson M, Li Y, Gupta M. Eventful Transformers: Leveraging Temporal Redundancy in Vision Transformers. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. IEEE; 2023. p. 16865-77. [10.1109/ICCV51070.2023.01551](https://doi.org/10.1109/ICCV51070.2023.01551)
- [64] Souza JS, Bedin E, Higa GTH, Loebens N, Pistori H. Pig Aggression Classification Using CNN, Transformers and Recurrent Networks; 2024. Preprint. Available from: <https://arxiv.org/abs/2403.08528>. [10.48550/arXiv.2403.08528](https://arxiv.org/abs/2403.08528)
- [65] Chowdhury S, Tisha SN, McGarrity ME, Hamerly G. Video-Based Recognition of Aquatic Invasive Species Larvae Using Attention-LSTM Transformer. In: Bebis G, Ghiasi G, Fang Y, Sharf A, Dong Y, Weaver C, et al., editors. Advances in Visual Computing. vol. 14361 of Lecture Notes in Computer Science. Cham: Springer Nature Switzerland; 2023. p. 224-35. [10.1007/978-3-031-47969-4_8](https://doi.org/10.1007/978-3-031-47969-4_8)
- [66] Yokota H, Bozkurtlar M, Yen B, Itoyama K, Nishida K, Nakadai K. A Video Vision Transformer for Sound Source Localization. In: 2024 32nd European Signal Processing Conference. IEEE; 2024. p. 106-10. [10.23919/EUSIPCO63174.2024.10715427](https://doi.org/10.23919/EUSIPCO63174.2024.10715427)
- [67] Jang HD, Kwon S, Nam H, Chang DE. Chemical Gas Source Localization with Synthetic Time Series Diffusion Data Using Video Vision Transformer. *Applied Sciences*. 2024;14(11):4451. [10.3390/app14114451](https://doi.org/10.3390/app14114451)
- [68] Constantin MG, Ionescu B. AIMultimediaLab at MediaEval 2022: Predicting Media Memorability Using Video Vision Transformers and Augmented Memorable Moments. In: Working Notes Proceedings of the MediaEval 2022 Workshop. vol. 3583 of CEUR Workshop Proceedings. Bergen, Norway and Online; 2023. Workshop held 12–13 January 2023. Available from: <https://ceur-ws.org/Vol-3583/>

How to cite this article

Younis AN, Ramo FM. A comparative review of video-based transformer models for modern applications. *Journal of University of Anbar for Pure Science*. 2026; 20(2):329-349. doi:[10.37652/juaps.2025.158955.1369](https://doi.org/10.37652/juaps.2025.158955.1369)