# Mask Cross-Attention Transformer for Robust Exercise Recognition Under Execution-Speed Domain Shift

Kamil Warchol(✉)[0009−0007−5638−9501] and Bogdan Kwolek

AGH University of Krakow, Al. Mickiewicza 30, Krakow, Poland
{warchol,bkw}@agh.edu.pl

**Abstract.** Human action recognition has achieved remarkable progress on general-purpose benchmarks, yet the subproblem of short-time action recognition, where discriminative motion unfolds over brief and variable-duration intervals, remains underexplored. Video-based exercise recognition faces a critical challenge when deployment conditions differ from training environments. The execution speed variation introduces temporal domain shifts that degrade model performance. We propose the *Mask Cross-Attention Transformer*, a dual-stream architecture for short-time action recognition that conditions temporal reasoning on human-centric spatial priors through cross-attention between appearance features and per-frame human masks. By decoupling semantic motion patterns from execution tempo, the model achieves robust recognition across diverse scenarios. On the public MM-Fit benchmark, the model achieves test accuracy of 95.0% and macro-F1 of 88.8% on 11 exercise classes, surpassing recent multimodal approaches while using only grayscale video with automatically generated masks.

**Keywords:** Exercise Recognition · Video Transformers · Cross-Attention

## 1 Introduction

Video-based human action recognition has advanced rapidly through deep architectures trained on large-scale benchmarks [5, 4, 1]. However, most research targets general activity categories spanning seconds to minutes, while *short-time action recognition*, where discriminative motion occurs over brief, temporally compact intervals, receives comparatively less attention [20]. Exercise recognition is a representative instance of this setting, as it encompasses fitness movements, rehabilitation protocols, and sports drills. Individual repetitions tend to be short, visually similar across classes, and subject to substantial execution variability. Despite its practical relevance to physiotherapy, fitness monitoring, and human performance analysis, dedicated exercise-recognition datasets and methods remain relatively limited [9], and existing approaches often rely on multimodal sensor setups that limit scalability.

Video-based exercise recognition is particularly challenging when deployment conditions differ from training. Recognition models trained on slow, controlled

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

executions fail to generalize to fast, variable-speed movements encountered in realistic deployment. This *execution-speed domain shift* alters motion dynamics without changing semantic identity, causing performance degradation [14, 28, 24, 3]. Standard video models mix semantic patterns with tempo. Three challenges emerge: (1) convolutional filters respond to specific speeds, becoming unreliable under shifts, (2) transformer attention may overfit to tempo-specific patterns, (3) models without human localization are more affected by background noise.

While cross-attention mechanisms have been applied in multimodal video understanding to fuse complementary streams such as RGB and optical flow [2], their use with segmentation masks as structured spatial priors for temporal robustness has not been explored. We propose a dual-stream architecture in which appearance tokens serve as queries and per-frame human mask tokens provide keys and values, enabling the model to ground temporal reasoning in human-centric spatial structure rather than in tempo-dependent motion cues. Dual streams are encoded separately, then fused through this cross-attention layer where appearance features are conditioned on mask-derived spatial layout. On publicly available MM-Fit benchmark [22], our model achieves 88.8% test macro-F1 with automatically generated masks. On a custom 17-class dataset under 2× speed shift (training on slow movements, testing on fast ones), we achieve 85.5% accuracy and 81.5% macro-F1, exceeding concatenation baselines by +9.0 and +15.0 percentage points.

## 2 Related Work

Two-stream architectures [21] established separate processing of appearance and motion as a foundational paradigm for video understanding. 3D CNNs extended this to joint spatiotemporal learning, with I3D [5] inflating 2D filters pre-trained on ImageNet into temporal convolutions, and SlowFast [7] operating two pathways at different temporal resolutions to capture both slow and fast motion cues. To reduce computational cost, several works proposed lightweight temporal modeling on top of 2D backbones. TSN [26] introduced sparse segment-level consensus, while TSM [14] shifted feature channels along the temporal axis at zero additional cost. Although effective on standard benchmarks such as Kinetics [12], these models implicitly encode motion speed into learned filters and pooling operations, making them sensitive to execution-tempo variations at test time.

Transformer architectures [25] have been adapted to video. TimeSformer [4] and ViViT [1] factorize space-time attention to reduce complexity, while hierarchical designs such as Video Swin Transformer [16] and MViT [6] incorporate multiscale feature hierarchies. Self-supervised pre-training through masked reconstruction, including VideoMAE [23] and MaskFeat [27], further improves data efficiency. However, these models learn temporal representations that are tightly coupled to the motion dynamics observed during training, and recent studies on video domain adaptation [28, 24] confirm that temporal distribution shifts remain a persistent source of performance degradation.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

Incorporating human-centric priors into action recognition has been shown to reduce background bias and improve generalization. Attentional pooling [10] re-weights spatial features towards action-relevant regions, while human-centric transformers [15] leverage person detection for domain-adaptive recognition. These approaches demonstrate the value of spatial grounding, but they typically rely on bounding boxes or pose keypoints rather than dense spatial masks, which limits the granularity of the spatial signal available to the temporal model.

Exercise recognition presents additional challenges beyond general action classification. Individual repetitions are short, visually similar across classes, and performed at highly variable speeds depending on the subject and context. The MM-Fit dataset [22] provides a multimodal benchmark for exercise classification, yet most methods evaluated on it depend on inertial sensors or skeleton data rather than video alone, and none explicitly address robustness to execution-speed variation. In contrast, our approach operates on video with automatically generated binary masks and introduces a cross-attention mechanism in which appearance tokens query mask-derived spatial structure, providing dense human-centric conditioning that decouples semantic motion patterns from execution tempo.

## 3 Method: Mask Cross-Attention Transformer

We first formalize the recognition task and the execution-speed domain shift setting. We then present the proposed architecture, covering dual-stream frame encoding, cross-attention conditioning between appearance and mask tokens, and temporal transformer with classification head.

### 3.1 Problem Formulation

Given video clip $\mathbf{X} = \{X_t\}_{t=1}^{T}$ and per-frame binary masks $\mathbf{M} = \{M_t\}_{t=1}^{T}$ representing the subject, the goal is to predict exercise class $y \in \{0, \ldots, C-1\}$. We address execution-speed domain shift, where training clips contain slow, controlled executions ($p_{\text{slow}}(\mathbf{X} \mid y)$) while test clips show fast, variable-speed patterns ($p_{\text{fast}}(\mathbf{X} \mid y)$). The goal is to learn classifier $f_\theta : (\mathbf{X}, \mathbf{M}) \to y$ robust to this temporal distribution shift.

### 3.2 Architecture Overview

The architecture processes clips through four stages (Fig. 1): (1) dual-stream encoding extracts per-frame features from video frames and binary masks, (2) cross-attention fuses streams by querying mask tokens with appearance tokens, (3) temporal transformer captures long-range dependencies, whereas (4) mean pooling and linear classification produce predictions.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

### 3.3 Dual-Stream Frame Encoding

**Frame encoder architecture.** Both RGB and mask streams are processed through structurally identical 2D convolutional encoders. Each encoder consists of three convolutional blocks with batch normalization, ReLU activation, and spatial downsampling:

$$\text{Block}_1: \quad \text{Conv2d}(k{=}3, p{=}1) \to \text{BN} \to \text{ReLU} \to \text{MaxPool2d}(2), \tag{1}$$

$$\text{Block}_2: \quad \text{Conv2d}(k{=}3, p{=}1) \to \text{BN} \to \text{ReLU} \to \text{MaxPool2d}(2), \tag{2}$$

$$\text{Block}_3: \quad \text{Conv2d}(k{=}3, p{=}1) \to \text{BN} \to \text{ReLU} \to \text{AdaptiveAvgPool2d}(1). \tag{3}$$

**Channel progression.** Both encoders use channel progression $1 \to 16 \to 32 \to 64$ (base_channels=16, output $F = 64$), yielding:

$$\mathbf{r}_t = f_{\text{enc}}(X_t) \in \mathbb{R}^{64}, \quad \mathbf{m}_t = f_{\text{enc}}(M_t) \in \mathbb{R}^{64} \tag{4}$$

**Projection and sequences.** Features are projected to a representation of size $d = 256$ via $W_{\text{proj}} \in \mathbb{R}^{256\times 64}$, forming token sequences $\mathbf{Z}^{\text{rgb}}, \mathbf{Z}^{\text{mask}} \in \mathbb{R}^{T\times 256}$ for $T = 16$ frames at $112 \times 112$ resolution.

### 3.4 Cross-Attention Conditioning

Appearance tokens serve as queries while mask tokens provide keys and values (Q from appearance, K and V from mask) for spatial conditioning. Multi-head attention ($H = 8$, $d = 256$) with pre-normalization is calculated as follows:

$$Q = \text{LN}(\mathbf{Z}^{\text{rgb}})W_Q, \tag{5}$$

$$K, V = \text{LN}(\mathbf{Z}^{\text{mask}})W_K, W_V, \tag{6}$$

$$\text{Attn} = \text{softmax}\left(\frac{QK^\top}{\sqrt{d/H}}\right)V \tag{7}$$

Output can be expressed as $\hat{\mathbf{Z}} = \mathbf{Z}^{\text{rgb}} + \text{Attn} + \text{MLP}(\text{LN}(\mathbf{Z}^{\text{rgb}} + \text{Attn}))$ with $4d$ hidden dim, and GELU activation, where LN means layer normalization.
**Interpretation.** The attention weights $\alpha_{t,s}$ measure the relevance of mask token $s$ to appearance token $t$. This allows the model to: (1) dynamically weight appearance features based on spatially relevant human motion patterns, and (2) filter background dynamics by down-weighting regions without human presence. Unlike concatenation, cross-attention enables fine-grained, token-level interaction between modalities [2].

### 3.5 Temporal Transformer and Classification

After cross-attention fusion, the conditioned features $\hat{\mathbf{Z}} \in \mathbb{R}^{T\times 256}$ undergo temporal modeling through three stages:

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

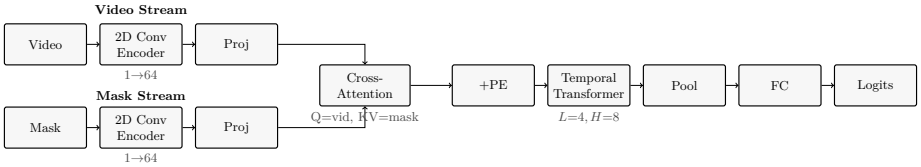


**Fig. 1.** Architecture of the Mask Cross-Attention Transformer. Dual 2D convolutional encoders process video and mask frames independently, projecting to $d$=256 dimensions. Cross-attention fuses modalities, followed by sinusoidal positional encoding, 4-layer temporal transformer, mean pooling, and classification. Input consists of $T$=16 frames at 112×112 resolution.

**Positional encoding.** Sinusoidal positional embeddings are added to preserve temporal ordering:

$$z_t^{(0)} = \hat{z}_t + \mathrm{PE}(t), \quad \text{where} \tag{8}$$

$$\mathrm{PE}(t, 2i) = \sin(t/10000^{2i/d}), \quad \mathrm{PE}(t, 2i+1) = \cos(t/10000^{2i/d}) \tag{9}$$

This inductive bias enables the model to capture temporal relationships without learnable position parameters.

**Temporal transformer blocks.** Four transformer blocks ($L = 4$) with pre-norm architecture (layer normalization applied before each sub-layer) process the sequence. Each block consists of multi-head self-attention ($H = 8$ heads, $d = 256$) followed by a feed-forward network (FFN) with expansion ratio 4.0 (hidden dimension $4d = 1024$) and GELU activation. Dropout rate 0.1 is applied to attention weights and FFN outputs. The blocks capture long-range temporal dependencies across the 16-frame sequence.

**Temporal pooling and classification.** The output sequence $\{z_t^{(L)}\}_{t=1}^{T}$ is aggregated via mean pooling over the temporal dimension:

$$\mathbf{g} = \frac{1}{T} \sum_{t=1}^{T} \mathrm{LN}(z_t^{(L)}) \in \mathbb{R}^{256} \tag{10}$$

A linear classifier $W_c \in \mathbb{R}^{C \times 256}$ maps the pooled representation to class logits: $\mathbf{y} = W_c \mathbf{g} + b_c \in \mathbb{R}^C$, where $C$ is the number of exercise classes (11 for MM-Fit, 17 for our dataset).

## 4 Experimental Setup

We evaluate our approach on two datasets, the public MM-Fit benchmark and a custom dataset [18] designed for execution-speed domain shift assessment. Subsection 4.1 describes both datasets, including the MM-Fit benchmark with automatically generated masks and our custom dataset with controlled train-test partitioning based on execution speed. Subsection 4.2 details the imple-

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

mentation, including preprocessing and sampling strategies, architecture configuration, training procedure with class-weighted loss, evaluation metrics, and baseline models for comparison.

### 4.1 Datasets

**Public Benchmark** MM-Fit [22] is a multimodal benchmark for exercise recognition containing 21 long-form videos (median duration 39 minutes at 30 FPS). Each video includes synchronized RGB frames, skeleton keypoints, and inertial sensor data from wearable devices. Subjects perform structured workout routines with 11 distinct exercise categories.
**Evaluation protocol.** We adopt the standard WITH NULL protocol from the MM-Fit benchmark, consistent with recent work such as MuJo [8], to ensure direct comparability. The dataset contains 11 distinct exercise categories including jumping jacks, push-ups, sit-ups, squats, and variations thereof. We use grayscale video with automatically generated human masks.
**Human mask generation.** MM-Fit does not provide per-frame human masks. We therefore generate pseudo-masks using a YOLO-based person segmentation model pretrained on COCO [29]. These serve as spatial priors without requiring ground-truth annotations. Mask quality statistics show mean coverage of 18.3% frame area with 2.1% empty-mask ratio, indicating reliable segmentation.
**Custom Dataset: Challenging Domain-Shift Evaluation.** To conduct a more challenging evaluation with explicit control over the domain shift variable, we introduce a custom dataset specifically designed to assess robustness under execution-speed domain shift. The dataset consists of 60 long-form videos, each depicting a single subject performing a sequence of physical exercises in a controlled indoor environment with a fixed camera position. Multiple subjects were recorded across sessions, although subject diversity remains limited compared to large-scale action recognition benchmarks. Videos are recorded at 30 FPS with median resolution $456 \times 598$ pixels and median duration 369 seconds (approximately 6 minutes).
**Annotations.** Each frame is annotated with one of 17 class labels: class 0 represents background/transition state, and classes 1–16 represent distinct exercise categories including various push-up variations, plank holds, jumping jacks, burpees, squats, lunges, and upper-body movements. Frame-level annotations enable dense supervision and precise evaluation without segment-level aggregation. Each frame is accompanied by a binary human mask localizing the performing subject. Masks are obtained through manual annotation refinement of automatic segmentation outputs, ensuring high-quality spatial localization. All videos, labels, and masks are temporally aligned at the frame level, enabling consistent sliding-window sampling without interpolation or missing annotations.
**Execution-speed domain shift protocol.** To explicitly evaluate robustness under temporal distribution shift, we partition the dataset into two disjoint subsets based on execution dynamics:

- **Training set (slow)**: 30 videos of slow, controlled, and mostly correct executions. These videos were recorded in instructional or supervised settings

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

where subjects were explicitly asked to perform movements slowly and with attention to form.
- **Test set (fast)**: 30 videos of fast, variable-speed, and more error-prone executions. These videos were recorded in less supervised settings where subjects performed movements at natural or accelerated tempo. Approximately 2× higher motion dynamics compared to the training set.

This protocol explicitly isolates execution-speed domain shift while preserving the semantic label space: both subsets contain all 17 classes with similar class distributions ($D_{KL}(p_{\text{train}} \| p_{\text{test}}) = 0.08$ nats (natural units of information), but differ substantially in motion dynamics. The systematic tempo mismatch forces models to generalize beyond execution-speed cues, reflecting realistic deployment conditions where training data characteristics differ from operational environments.

**Class imbalance.** Both datasets show severe class imbalance. In our dataset, the most frequent exercise (class 8: standard push-ups) accounts for 13.2% of labeled frames, while minority classes (e.g., class 14: diamond push-ups) contribute less than 3% each. Background/transition frames (class 0) make up 22% of the training set and 27% of the test set. In MM-Fit, jumping jacks account for 18% of frames while some exercise variations appear in less than 5%. This imbalance, particularly when combined with execution-speed domain shift, makes both datasets challenging and requires the use of macro-averaged metrics and a class-weighted loss.

### 4.2 Implementation Details

**Preprocessing and sampling.** All models process fixed-length clips consisting of $T = 16$ frames at resolution $112 \times 112$ pixels. RGB frames are converted to grayscale (luminance channel), resized using bilinear interpolation, and normalized to $[0, 1]$ range. Human masks are resized using nearest-neighbor interpolation to preserve binary values, then replicated across the channel dimension for concatenation baselines.

Clips are extracted using a sliding window with stride $s = 8$ frames, resulting in 50% overlap between consecutive windows. This dense sampling ensures sufficient temporal coverage for evaluation. Each window is assigned a single class label using *smart-middle labeling*: the label corresponds to the ground-truth class of the temporal center frame (frame $\lfloor T/2 \rfloor = 8$). This strategy minimizes label noise from temporal boundaries where transitions between exercises occur.

The choice of $T = 16$ frames at 30 FPS yields a temporal window of approximately 0.5 seconds, which is sufficient to capture one or more phases of typical exercise repetitions. A stride of $s = 8$ ensures that consecutive sub-sequences overlap by 50%, preventing gaps in temporal coverage and reducing sensitivity to the exact alignment of action boundaries. The resolution of $112 \times 112$ pixels balances spatial detail with computational efficiency, as higher resolutions increase encoder cost quadratically while providing diminishing returns for silhouette-level spatial priors. Preliminary experiments with $T = 8$ and $T = 32$

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

showed that shorter clips miss multi-phase motion patterns while longer clips increase self-attention cost without improving accuracy under the execution-speed shifts considered in this work.

**Architecture configuration.** The frame encoders use embedding dimension $d = 256$. The temporal transformer uses model dimension $d_m = 256$, $L = 4$ layers, $H = 8$ attention heads, and dropout rate 0.1 applied to attention weights and feed-forward networks. These hyperparameters were selected based on preliminary validation experiments balancing model capacity with overfitting risk.

**Computational considerations.** The dual-stream design introduces only minor overhead compared to single-stream baselines for $T = 16$ frames and $d = 256$ dimensions. Computational costs are dominated by temporal transformer self-attention ($\mathcal{O}(L \cdot T^2 \cdot d)$). The architecture remains suitable for real-time inference on modern GPUs.

**Training procedure.** Models are trained using AdamW optimizer [17] with learning rate $\eta = 3 \times 10^{-4}$, weight decay $\lambda = 10^{-4}$, and $(\beta_1, \beta_2) = (0.9, 0.999)$. Learning rate follows a cosine annealing schedule over 30 epochs with no warmup. Batch size is 32 clips. All experiments use fixed random seeds to ensure reproducibility.

To address class imbalance, we apply class-weighted cross-entropy loss with weights computed using the effective number strategy: $w_c = (1 - \beta)/(1 - \beta^{n_c})$ where $n_c$ is the number of training samples for class $c$ and $\beta = 0.999$. This approach reduces the contribution of majority classes while up-weighting minority classes, promoting balanced learning across all exercise types.

Data augmentation includes: (1) random horizontal flip with probability 0.5, (2) random resized crop with scale $[0.8, 1.0]$ and aspect ratio $[0.9, 1.1]$, and (3) color jittering with brightness, contrast, and saturation factors sampled from $[0.8, 1.2]$. Augmentations are applied consistently to all frames in a clip (geometric transforms to both grayscale and binary masks. Moreover photometric transforms to grayscale only).

**Evaluation metrics.** Performance is evaluated using three metrics:

- **Accuracy**: fraction of correctly classified windows
- **Macro-averaged F1 score**: unweighted mean of per-class F1 scores
- **Macro-averaged ROC-AUC**: unweighted mean of per-class area under ROC curve

Macro metrics are emphasized over weighted or micro-averaged metrics because they better reflect robustness under severe class imbalance: a model that performs well on majority classes but fails on minority classes will achieve high accuracy but low macro-F1, revealing its limitations.

**Baseline models.** We compare results achieved by our model against results achieved by comprehensive temporal baselines:

- **MLP**: frame-wise feature extraction with temporal mean pooling and 2-layer MLP classifier.
- **LSTM** [11]: bidirectional LSTM with hidden size 256 applied to frame features. This baseline represents classical recurrent temporal modeling.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

- **TCN** [13]: temporal convolutional network with kernel size 3, dilation factors $\{1, 2, 4, 8\}$, and 4 layers. It represents efficient convolutional temporal modeling.
- **Transformer** [4]: standard temporal transformer with self-attention.
- **ViViT** [1]: video vision transformer with tubelet embeddings(spatial-temporal patches). It represents joint spatiotemporal modeling.

For each baseline, we evaluate three input configurations:

- **RAW**: grayscale frames only (appearance-only baseline)
- **BOTH**: grayscale frames concatenated with mask channel (2-channel input)
- **MASK**: mask frames only (ablation to assess mask information content)

This experimental design allows us to isolate the effect of architectural integration of mask information (cross-attention) versus simple input concatenation.

## 5 Experimental Results

We report experimental results on two datasets. Subsection 5.1 presents performance on the MM-Fit benchmark, demonstrating competitive results compared to recent multimodal approaches. Subsection 5.2 evaluates robustness under execution-speed domain shift using our custom dataset, comparing the proposed architecture against temporal baselines. Ablation studies in Subsection 5.3 analyze architectural design choices, while qualitative analysis in Subsection 5.4 visualizes learned attention patterns.

### 5.1 Results on MM-Fit Benchmark

Table 1 reports performance on MM-Fit. We follow the WITH NULL protocol (12 classes: 11 exercises + background/transition) for direct comparability with MuJo and standard MM-Fit benchmarks. For brevity, we refer to this protocol as with-null in subsequent text.

**Table 1.** Performance on MM-Fit dataset (grayscale video + masks vs RGB baselines). F1 and AUC are macro-averaged.

| Model | Modality | Protocol | Val Acc (%) | Val F1 (%) | Val AUC (%) | Test Acc (%) | Test F1 (%) | Test AUC (%) |
|---|---|---|---|---|---|---|---|---|
| **Mask Cross-Attn** | **Gray+Mask** | **WITH NULL** | **91.2** | **76.3** | **97.0** | **95.0** | **88.8** | **99.3** |
| MuJo [8] | RGB | WITH NULL | – | – | – | – | 72.9 | – |
| MuJo [8] | RGB | NO NULL | – | – | – | – | 84.8 | – |

Our model achieves 88.8% macro-F1 (with-null) using grayscale video with automatically generated masks, exceeding MuJo's RGB-based result of 72.9%. The with-null protocol includes background frames, making it more challenging than no-null (84.8% for MuJo). We report results achieved for with-null for consistency with the MM-Fit benchmark protocol.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

### 5.2 Results on Custom Dataset Under Challenging Domain Shift

We now evaluate robustness under execution-speed domain shift using our dataset, where training clips contain slow executions and test clips show fast movements. Table 2 reports validation performance on this challenging scenario.

**Table 2.** Validation performance on custom dataset under execution-speed domain shift (17 classes). RAW: grayscale only. BOTH: grayscale + mask. MASK: mask only. F1 and AUC are macro-averaged.

| Model | Input | Val Acc (%) | Val F1 (%) | Val AUC (%) |
|---|---|---|---|---|
| MLP | RAW | 51.1 | 13.4 | 78.5 |
| MLP | BOTH | 58.7 | 30.2 | 91.3 |
| MLP | MASK | 52.3 | 19.8 | 84.2 |
| LSTM | RAW | 50.0 | 14.7 | 80.2 |
| LSTM | BOTH | 70.0 | 55.3 | 96.1 |
| LSTM | MASK | 61.2 | 41.2 | 92.3 |
| TCN | RAW | 57.8 | 28.3 | 89.7 |
| TCN | BOTH | 76.4 | 68.6 | 97.8 |
| TCN | MASK | 69.8 | 58.9 | 95.4 |
| Transformer | RAW | 55.8 | 34.0 | 87.5 |
| Transformer | BOTH | 76.5 | 66.5 | 97.6 |
| Transformer | MASK | 71.2 | 60.1 | 95.8 |
| ViViT | RAW | 51.2 | 35.2 | 48.1 |
| ViViT | BOTH | 54.8 | 48.0 | 88.0 |
| ViViT | MASK | 50.1 | 39.8 | 81.5 |
| **Mask Cross-Attn Transformer** | **RAW** | **65.2** | **51.6** | **92.2** |
| **Mask Cross-Attn Transformer** | **BOTH** | **85.5** | **81.5** | **98.5** |
| **Mask Cross-Attn Transformer** | **MASK** | **80.6** | **74.3** | **98.0** |

**Severe degradation under domain shift.** All baselines show performance degradation under execution-speed mismatch. Appearance-only models perform particularly poorly where Transformer RAW achieves only 55.8% accuracy and 34.0% macro-F1, and MLP and LSTM results are even worse, with accuracies around 50% (barely above chance for 17 classes), confirming that standard temporal models overfit to execution tempo.

**Limited benefit of input concatenation.** Adding masks via simple channel concatenation (BOTH) consistently improves performance across all baselines. However, gains remain limited: the strongest BOTH baseline is Transformer with 76.5% accuracy and 66.5% macro-F1 - a substantial improvement over RAW (+20.7 percentage points accuracy) but still far from robust performance. TCN BOTH demonstrates comparable performance, achieving 76.4% accuracy and 68.6% macro-F1.

Interestingly, ViViT exhibits poor performance even with mask augmentation (54.8% accuracy, 48.0% macro-F1), suggesting that joint spatiotemporal

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

patch embeddings do not effectively leverage mask information when processed via tubelet tokenization. This observation highlights the significant impact of architectural design choices on the integration of spatial information.

**Mask-only ablation.** MASK-only baselines reveal the information content of spatial structure alone. Transformer MASK achieves 71.2% accuracy and 60.1% macro-F1 – better than Transformer RAW, confirming that body pose and spatial distribution provide informative cues independent of appearance. However, MASK remains inferior to BOTH, indicating complementary information.

The strong performance of mask-only models can be attributed to the fact that binary silhouettes implicitly encode body pose and spatial configuration. Different exercises produce distinct silhouette geometries, and the temporal evolution of these shapes captures movement patterns that are largely invariant to execution speed. This explains why masks alone outperform appearance features under tempo shift, as silhouette dynamics change less with speed than pixel-level motion cues. A limitation of this representation is that exercises with similar body configurations may produce nearly identical silhouettes, making fine-grained discrimination more difficult without complementary appearance information.

**Mask Cross-Attention Transformer Performance** The proposed Mask Cross-Attention Transformer with full RGB+mask conditioning (BOTH) achieves **85.5% accuracy** and **81.5% macro-F1**, outperforming all baselines. Compared to the strongest baseline (Transformer BOTH): +9.0, +15.0, and +0.9 percentage points in accuracy, macro-F1, and ROC-AUC, respectively. The gains are particularly pronounced in macro-F1, indicating improved robustness across minority classes.

**Input modality ablation.** We evaluate three configurations:

- **RAW(appearance only)**: 65.2% accuracy, 51.6% macro-F1. Cross-attention architecture alone is insufficient without spatial priors.
- **MASK (spatial structure only)**: 80.6% accuracy, 74.3% macro-F1. Binary spatial structure outperforms all concatenation baselines.
- **BOTH (appearance + mask)**: 85.5% accuracy, 81.5% macro-F1. Full cross-modal conditioning achieves optimal performance.

Spatial structure (MASK: 80.6%) is more informative than appearance alone (RAW: 65.2%) under tempo shift, but full fusion (BOTH: 85.5%) achieves optimal performance by leveraging complementary cues.

### 5.3 Ablation Studies

**Cross-Attention vs Concatenation** Table 3 compares different fusion strategies for combining RGB and mask information. Cross-attention provides larger gains than concatenation (+9.0 vs +20.7 percentage points accuracy over RAW), confirming that architectural integration of spatial priors is crucial for temporal

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

**Table 3.** Ablation study: fusion strategies for RGB and mask integration (our dataset). F1 and AUC are macro-averaged.

| Fusion Strategy | Val Acc (%) | Val F1 (%) | Val AUC (%) |
|---|---|---|---|
| None (appearance only) | 55.8 | 34.0 | 87.5 |
| Channel concatenation | 76.5 | 66.5 | 97.6 |
| Element-wise addition | 74.2 | 62.8 | 96.8 |
| Cross-attention (ours) | **85.5** | **81.5** | **98.5** |

**Table 4.** Ablation study: transformer depth (custom dataset, Mask Cross-Attention Transformer). F1 is macro-averaged.

| Layers $L$ | Val Acc (%) | Val F1 (%) | Params (M) |
|---|---|---|---|
| $L = 2$ | 83.2 | 78.3 | 2.9 |
| $L = 4$ | **85.5** | **81.5** | **4.0** |
| $L = 6$ | 85.2 | 81.1 | 5.2 |
| $L = 8$ | 84.8 | 80.5 | 6.4 |

robustness. Element-wise addition (where RGB and mask tokens are added after separate encoding) performs worse than concatenation, suggesting that the two modalities encode complementary rather than redundant information and benefit from learned interaction.

**Number of Transformer Layers** Table 4 reports performance with varying transformer depth. Accuracy peaks at $L = 4$ layers, deeper models ($L = 6, 8$) show no improvement and slight overfitting. This suggests that the spatially grounded feature space enables efficient temporal modeling without requiring very deep architectures.

### 5.4 Qualitative Analysis

Figure 2 compares activation maps using Grad-CAM [19] across three temporal frames from an exercise sequence. The baseline transformer concentrates activation on prominent local features such as hands, body boundaries, and high-contrast edges-characteristic of models relying on appearance gradients for motion discrimination. In contrast, the proposed Mask Cross-Attention Transformer distributes attention more broadly across the spatial context, including background regions. This suggests that the model develops a holistic scene representation, where spatial context (e.g., floor contact, body-environment spatial relationships) provides complementary cues for exercise recognition beyond isolated body parts. By integrating human-centric spatial priors through cross-attention, the model captures both person-specific features and contextual information, explaining improved robustness under execution-speed domain shift where local motion patterns vary but spatial configurations remain more stable.

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

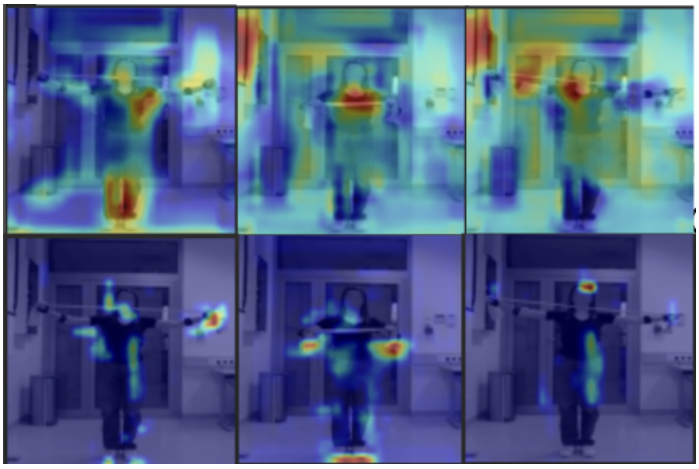

**Fig. 2.** Gradient-based activation maps (Grad-CAM) comparison across three frames from the same exercise sequence. Top row: Mask Cross-Attention Transformer. Bottom row: standard transformer baseline. The proposed model shows broader spatial context awareness including background regions, while the baseline focuses mainly on local edge features and extremities (hands, boundaries).

## 6 Discussion and Limitations

**Limitations.** Our dataset contains 60 videos (1.2M frames) enabling controlled domain shift evaluation, but limited subject diversity may affect generalization. Automatically generated masks on MM-Fit may contain errors when body parts are hidden, however, ablations show gradual performance drop. The architecture assumes single-person clips, multi-person scenarios would require instance-aware encoding. It is also worth noting that performance degrades beyond 3× speed shifts, suggesting limits of spatial conditioning alone.
**Future work.** In future work, we plan to incorporate pose keypoints as additional spatial priors to enrich the human-centric representation, while also exploring model quantization to enable efficient edge deployment for real-time fitness applications.

## 7 Conclusion

We introduced the Mask Cross-Attention Transformer, conditioning temporal representations on human-centric spatial priors to separate semantic motion from execution tempo. On MM-Fit, the model achieves 88.8% test macro-F1 using grayscale video with automatically generated person masks. On our dataset under 2× execution-speed shift, it achieves 85.5% accuracy and 81.5% macro-F1, outperforming concatenation-based baselines by 9.0 and 15.0 percentage points, respectively. Ablations show that cross-attention provides larger performance

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

gains than input concatenation, and that spatial structure is more informative than appearance under tempo shift. Cross-attention between appearance and spatial priors enables temporal robustness through human-centric grounding.

**Acknowledgments.** This work was supported by Polish National Science Center (NCN) under research grant 2024/55/B/ST6/01580.

## References

1. Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., Schmid, C.: ViViT: A video vision transformer. Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 6816–6826 (2021)
2. Baltrušaitis, T., Ahuja, C., Morency, L.P.: Multimodal machine learning: A survey and taxonomy. IEEE Trans. Pattern Anal. Mach. Intell. **41**, 423–443 (2017)
3. Ben-David, S., Blitzer, J., Crammer, K., Kulesza, A., Pereira, F.C., Vaughan, J.W.: A theory of learning from different domains. Mach. Learn. **79**, 151–175 (2010)
4. Bertasius, G., Wang, H., Torresani, L.: Is space–time attention all you need for video understanding? In: Proc. of the 38th Int. Conf. on Machine Learning (2021)
5. Carreira, J., Zisserman, A.: Quo vadis, action recognition? A new model and the Kinetics dataset. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Rec. pp. 4724–4733 (2017)
6. Fan, H., Xiong, B., Mangalam, K., Li, Y., Yan, Z., Malik, J., Feichtenhofer, C.: Multiscale vision transformers. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 6804–6815 (2021)
7. Feichtenhofer, C., Fan, H., Malik, J., He, K.: SlowFast networks for video recognition. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. pp. 6201–6210 (2019)
8. Fritsch, S.G., Oguz, C., Rey, V.F., Ray, L., Kiefer-Emmanouilidis, M., Lukowicz, P.: Mujo: Multimodal joint feature space learning for human activity recognition. In: Proc. IEEE Int. Conf. Pervasive Comput. Commun. pp. 1–12 (2025)
9. Ghosh, I., Ramasamy Ramamurthy, S., Chakma, A., Roy, N.: Sports analytics review: Artificial intelligence applications, emerging technologies, and algorithmic perspective. WIREs Data Min. Knowl. Discov. **13**(5), e1496 (2023)
10. Girdhar, R., Ramanan, D.: Attentional pooling for action recognition. In: Adv. Neural Inf. Process. Syst. vol. 30. Curran Associates, Inc. (2017)
11. Hochreiter, S., Schmidhuber, J.: Long Short-Term Memory. Neural Comput. **9**(8), 1735–1780 (1997)
12. 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.: The Kinetics human action video dataset (2017)
13. Lea, C., Flynn, M.D., Vidal, R., Reiter, A., Hager, G.D.: Temporal convolutional networks for action segmentation and detection. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 1003–1012 (2017)
14. Lin, J., Gan, C., Han, S.: TSM: Temporal shift module for efficient video understanding. In: Proc. IEEE/CVF Int. Conf. Comput. Vis. p. 70827092 (2019)
15. Lin, K.Y., Zhou, J., Zheng, W.S.: Human–centric transformer for domain adaptive action recognition. IEEE Trans. Pattern Anal. Mach. Intell. **47**(2), 679–696 (2025)
16. Liu, Z., Ning, J., Cao, Y., Wei, Y., Zhang, Z., Lin, S., Hu, H.: Video Swin Transformer. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Rec. pp. 3192–3201 (2022)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32

17. Loshchilov, I., Hutter, F.: Decoupled weight decay regularization. In: Proc. Int. Conf. Learn. Represent. (2017)
18. Pelc, E.: Human Action Recognition Based on Sensory Measurements. Master's thesis, AGH University of Science and Technology, Kraków, Poland (2025)
19. Selvaraju, R.R., Das, A., Vedantam, R., Cogswell, M., Parikh, D., Batra, D.: Grad–CAM: Visual explanations from deep networks via gradient–based localization. Int. J. Comput. Vis. **128**, 336–359 (2016)
20. Shin, J., Hassan, N., Miah, A., Nishimura, S.: A comprehensive methodological survey of human activity recognition across diverse data modalities. Sensors **25**(13) (2025)
21. Simonyan, K., Zisserman, A.: Two–stream convolutional networks for action recognition in videos. In: Adv. Neural Inf. Process. Syst. (2014)
22. Strömbäck, D., Huang, S., Radu, V.: MM–Fit: Multimodal deep learning for automatic exercise logging across sensing devices. Proc. ACM Interact. Mob. Wearable Ubiquitous Technol. **4**, 1–22 (12 2020)
23. Tong, Z., Song, Y., Wang, J., Wang, L.: VideoMAE: Masked autoencoders are data–efficient learners for self–supervised video pre–training. In: Adv. Neural Inf. Process. Syst. (2022)
24. Tzeng, E., Hoffman, J., Saenko, K., Darrell, T.: Adversarial discriminative domain adaptation. In: Proc. IEEE/CVF Conf. Comput. Vis. Patt. Rec. pp. 2962–2971 (2017)
25. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., Polosukhin, I.: Attention is all you need. In: Adv. Neural Inf. Process. Syst. (2017)
26. Wang, L., Xiong, Y., Wang, Z., Qiao, Y., Lin, D., Tang, X., Van Gool, L.: Temporal segment networks: Towards good practices for deep action recognition. In: Proc. Eur. Conf. Comput. Vis. pp. 20–36. Springer (2016)
27. Wei, C., Fan, H., Xie, S., Wu, C.Y., Yuille, A., Feichtenhofer, C.: Masked feature prediction for self–supervised visual pre–training. In: Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. pp. 14648–14658 (2022)
28. Xu, Y., Yang, J., Cao, H., Wu, K., Wu, M., Chen, Z.: Source–free video domain adaptation by learning temporal consistency for action recognition. In: Proc. Eur. Conf. Comput. Vis. pp. 147–164. Springer (2022)
29. Yaseen, M.: What is YOLOv8: An in-depth exploration of the internal features of the next-generation object detector **abs/2408.15857** (2024)

ICCS Camera Ready Version 2026
To cite this paper please use the final published version:
DOI: 10.1007/978-3-032-29909-3_32